Simon Foucart, Holger Rauhut

A Mathematical Introduction to Compressive Sensing

SPIN Springer's internal project number, if known

– Monograph –

May 7, 2012

Springer
Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Dedicated to our families, for their patience and encouragement
Pour Jeanne
Für Daniela, Niels, Paulina und Antonella

Preface

Recent years have seen the explosion of a fascinating new field at the intersection of mathematics, electrical engineering, and computer science called compressive sensing. The name comes from the premise that data acquisition and compression can be performed simultaneously. This book aims at a detailed and self-contained presentation of the mathematical core of compressive sensing.

The basic idea is that many types of signals are sparse in the sense that they can be well-approximated by a small number of non-zero coefficients in a suitable basis. The goal of compressive sensing is to reconstruct such type of vectors from incomplete linear information. This leads to an underdetermined linear system and although this has infinitely many solutions in general, the additional information of sparsity allows to single out the true solution under appropriate conditions. Moreover, efficient algorithms can be used for the reconstruction process, for instance, ℓ_1-minimization — a convex optimization problem. Obviously, one would like to work with a minimal number of measurements. Quite remarkably, it is an open problem to date to come up with optimal explicit measurement matrices, and all known "good" matrix constructions involve randomness. This discovery has a lot of potential applications in signal and image processing. And as if this were not enough, the subject is made even more exciting by the elegance of its underlying theory, which is also appealing to pure mathematicians.

Compressive sensing builds on various branches of mathematics including linear algebra, approximation theory, convex analysis and optimization, probability theory and (in particular) random matrices, Banach space geometry, harmonic analysis, and graph theory. Some of the required background is, of course, much older than the advent of compressive sensing around 2004. This book makes an attempt to introduce to the rich and beautiful mathematical theory of compressive sensing including the necessary background material. Hereby, no special knowledge apart from basic analysis, linear algebra, and probability theory is required.

Despite its strong potential for various applications, we will not go into and only give a teaser on applications. Our perspective is on the mathematical side, given that we are mathematicians by training and by taste. This bias dictated the choice of topics to be covered. Some topics indeed had to be left out, because we intend this volume to be an introduction, not an exhaustive treatise. However, the exposition is complete: we wanted every result to be proved — apart from very basic material from linear algebra, analysis and probability — so that the material should be accessible to master and graduate students in mathematics, but also for engineers, computer scientists and physicists. Another concern was conciseness: we have made efforts to write short, natural proofs that are often simplified versions of the ones found in the literature. Rendering the mathematical foundations of compressive sensing accessible to graduate and master students was our objective, and we both, independently, went through this process when preparing courses at Vanderbilt University, Drexel University, University of Bonn and ETH Zurich. This monograph is the result of a further attempt to clarify the material even more.

We will cover a variety of sparse recovery algorithms together with their theoretical analysis. A practitioner may wonder which algorithm one should choose for a particular purpose. In general, all the algorithms should give reasonable performance, but it is a matter of numerical experiments in a concrete setup to determine which algorithm provides best recovery performance and/or is the fastest. In order not to bias towards a particular algorithm a priori, we made the choice of not presenting numerical comparisons for the simple reason that it is impossible to run such numerical experiments in all possible setups. Nevertheless, we gave some crude hints on the choice of the algorithm in the Notes section of Chapter 3.

We have structured the book as follows. The first chapter gives a quick introduction to the essentials of compressive sensing, describes some motivations and some potential applications and then provides a detailed overview on the whole book. Chapters 2 – 6 cover basically the deterministic theory of compressive sensing. We cover the notion of sparsity, introduce basic algorithms and analyze their performance based on various properties of the measurement matrix. Since the major breakthroughs in showing estimates on the minimal number of required measurements are based on random matrices, we cover the required tools from probability in detail in Chapters 7 and 8. Based on this preparation, Chapters 9 – 13 treat sparse recovery based on random matrices and related topics. Chapter 14 covers sparse recovery using lossless expanders and Chapter 15 introduces algorithms for ℓ_1-minimization. The book is concluded with three Appendices which cover basic material from linear algebra and matrix analysis, convex analysis and various miscellaneous topics.

Each chapter ends with a "Notes" section. This is the place where we provide useful tangential comments which would otherwise disrupt the flow of the text, such as relevant references, further aspects and results, historical remarks, or open questions. We have compiled a selection of exercises for

each chapter. These give the reader an opportunity to work with the material and they provide additional results of interests. There is also a chapter of solutions at the end of the book – in fact, hints of solutions – to the exercises accompanying each chapter.

It was a particular challenge to write a monograph in such a quickly moving field. Some developments in the area appeared during the process of writing and sometimes this resulted in changes or additions to the book. We believe that the present material represents well the foundations of the theory of compressive sensing, and that further developments build on it rather than replace it. But, of course, it is hard to predict the future of a quickly moving field, and maybe an update of the present book will be required in some years.

We greatly acknowledge the help of several colleagues for proofreading and commenting parts of the manuscript in alphabetical order: MANY NAMES TO BE ADDED. Furthermore, we profited from various collaborations and discussions on the subjects covered in this book, in particular, with MANY NAMES TO BE ADDED. We thank our host institutions, the Hausdorff Center for Mathematics and the Institute for Numerical Simulation at the University of Bonn as well as Drexel University, Philadelphia for their support and for providing excellent working conditions. Holger Rauhut acknowledges financial support by the WWTF (Wiener Wissenschafts-, Forschungs- und Technologie-Fonds) as well as by the European Research Council through a Starting Grant. Simon Foucart acknowledges support from the NSF (National Science Foundation) under the grant DMS-1120622.

Finally, we hope that the reader enjoys this book as much as we enjoyed writing it.

Philadelphia, Bonn, *Simon Foucart*
May 2012 *Holger Rauhut*

Contents

1

An Invitation to Compressive Sensing

This first chapter introduces the compressive problem and gives an overview on the book. As the mathematical theory is highly motivated by real-life problems, we briefly describe some of the potential applications.

1.1 What is Compressive Sensing?

In many practical problems of science and technology – especially in signal and image processing – one encounters the task of inferring quantities of interest – signals, images, statistical data, etc. – from measured information. When the information acquisition process is linear then the problem reduces to solving a linear system of equations. In mathematical terms, the observed data $\mathbf{y} \in \mathbb{C}^m$ is connected to the vector (signal) $\mathbf{x} \in \mathbb{C}^N$ of interest via

$$\mathbf{A}\mathbf{x} = \mathbf{y} \, . \tag{1.1}$$

Here the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ models the linear measurement (information) process. One recovers the vector $\mathbf{x} \in \mathbb{C}^N$ of interest by solving the above linear system. Traditional wisdom tells us that the number m of measurements, that is, the amount of measured data, has to be at least as large as the signal length N (the number of components of $\mathbf{x}$). This principle is the basis for most devices of current technology, such as analog to digital conversion, medical imaging, radar, and mobile communication. Indeed, basic linear algebra states that the linear system (1.1) above is underdetermined if $m < N$ and that there are infinitely many solutions (provided, of course, that there exists at least one). In other words, without additional information it is impossible to recover $\mathbf{x}$ from $\mathbf{y}$ in the case that $m < N$. This fact is also related to Shannon's sampling theorem, which states that the sampling rate of a continuous time signal must be twice the highest frequency in order to ensure reconstruction.

It therefore came as a surprise that under certain assumptions it is actually possible to reconstruct signals when the number of available measurements m

(a)

(b)

Fig. 1.1. (a) Original image. (b) Reconstruction using the largest 2% of the wavelet coefficients, that is, 98% of the coefficients are zero.

is smaller than the signal length N. Even more surprisingly, efficient algorithms can be used for the reconstruction. The key assumption is *sparsity*. The research area associated to this phenomenon is called *compressive sensing*, *compressed sensing*, *compressive sampling* or *sparse recovery*. This whole book is devoted to the mathematics of this field.

Sparsity. A signal is called sparse if most of its components are zero. It is an empirical observation that many real-world signals are compressible in the sense that they can be well-approximated by sparse ones — at least after an appropriate change of basis. This is in fact the reason why compression techniques such as JPEG, MPEG, or MP3 work well in practice. For instance, JPEG uses the fact that images are usually sparse in the discrete cosine basis (DCT) or wavelet basis and achieves compression by only storing the largest DCT coefficients. When decompressing the image, the non-stored coefficients are simply set to zero. For an illustration that natural images are sparse in the wavelet domain we refer to Figure 1.1.

Consider again our task of acquiring a signal from measured data. Given the additional knowledge that the signal is sparse or at least compressible the traditional approach of taking at least as many measurements as the signal length seems to waste resources: At first, big efforts are undertaken to measure all entries of a signal and then most coefficients are thrown away in order to arrive at a compressed version. One may ask whether it is possible to acquire "more directly" the compressed version of a signal using significantly fewer measured data than the signal length – knowing that the signal is actually sparse or compressible. In other words, we would like to compressively sense a compressible signal! This is the basic problem of compressive sensing.

Let us emphasize that the main difficulty in the compressive sensing problem lies in the fact that one does not assume a priori knowledge on the locations of the nonzero entries of the unknown vector $\mathbf{x}$. Indeed, if one would know these locations beforehand one could simply reduce the matrix $\mathbf{A}$ to the columns corresponding to this location set. The resulting system of linear equations becomes then overdetermined (provided that the number of nonzero entries in $\mathbf{x}$ is small enough), and we can solve for the nonzero entries in the signal. Not knowing the nonzero locations of the vector $\mathbf{x}$ to be reconstructed leads to a nonlinearity because the set of s-sparse vectors (those having at most s nonzero coefficients) is a nonlinear set. In fact, adding two s-sparse vectors leads to a $2s$-sparse vector in general. Therefore, any successful reconstruction method must necessarily be nonlinear.

Intuitively, the complexity or "intrinsic" information content of compressible signals is much smaller than the signal length (otherwise, compression would not be possible). So one may argue that one only needs an amount of data (number of measurements), which is proportional to this intrinsic information content rather than to the actual signal length. Nevertheless, it does not seem clear at the beginning how to achieve reconstruction in this scenario.

Looking closer at the compressive sensing problem to reconstruct a sparse vector $\mathbf{x} \in \mathbb{C}^N$ from underdetermined measurements $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$, $m < N$, one essentially identifies two questions:

- How should one design the linear measurement process? In other words, what matrices $\mathbf{A} \in \mathbb{C}^{m \times N}$ are suitable?

- How can one reconstruct $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$? What are efficient reconstruction algorithms?

These two questions are not completely independent because the reconstruction algorithm has to take into account $\mathbf{A}$, but we will see that one can often split the analysis of the matrix $\mathbf{A}$ from the analysis of the algorithm.

At first, we notice that compressive sensing will not work for arbitrary matrices $\mathbf{A} \in \mathbb{C}^{m \times N}$. For instance, if $\mathbf{A}$ consists of rows of the identity matrix so that $\mathbf{y} = \mathbf{A}\mathbf{x}$ simply picks some entries of $\mathbf{x}$, then $\mathbf{y}$ contains mostly zero entries. In particular, no information is obtained about the nonzero entries of $\mathbf{x}$ that $\mathbf{y}$ does not catch, and reconstruction is clearly impossible for such choice of $\mathbf{A}$. Therefore, compressive sensing is not only about the recovery algorithm. Also the first question on the design of the measurement matrix $\mathbf{A}$ is important and non-trivial. We emphasize that designing the matrix $\mathbf{A}$ beforehand means that the measurement process is non-adaptive in the sense that we do *not* choose the type of measurements for the next datum y_j, that is, the jth row of $\mathbf{A}$, depending on the previously observed data $y_1, \dots, y_{j-1}$. (Indeed, it turns out that adaptive measurements do not provide better performance in general.)

Algorithms. For practical purposes it is, of course, important that there are reasonably fast reconstruction algorithms. This is arguably the feature of compressive sensing which caused it to catch so much attention. The first algorithmic approach that probably comes to mind is ℓ_0-minimization. Introduce $\|\mathbf{x}\|_0$ to be the number of nonzero entries of a vector $\mathbf{x}$. Then to reconstruct $\mathbf{x}$ it is natural to consider the solution of the combinatorial optimization problem

$$\text{minimize } \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y} \, .$$

In words, we search for the sparsest vector consistent with the measured data $\mathbf{y} = \mathbf{A}\mathbf{x}$. Unfortunately, ℓ_0-minimization turns out to NP hard in general. With this information, it may even seem more surprising that fast and provably effective reconstruction algorithms do exist. A very popular and by now well-understood method is basis pursuit or ℓ_1-minimization, which consists in finding the minimizer of the problem

$$\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y} \, . \tag{1.2}$$

Since the ℓ_1-norm $\|\cdot\|_1$ is convex, this optimization problem can be solved with efficient methods from convex optimization. Basis pursuit can be interpreted as the convex relaxation of ℓ_0-minimization. Alternative reconstruction methods include greedy type methods such as orthogonal matching pursuit, as well as thresholding based methods including iterative hard thresholding. We will see that under suitable assumptions all these methods indeed recover sparse vectors.

Before we continue we invite the reader to look at Figure 1.2, which illustrates the ability of compressive sensing. It shows an example of a signal of

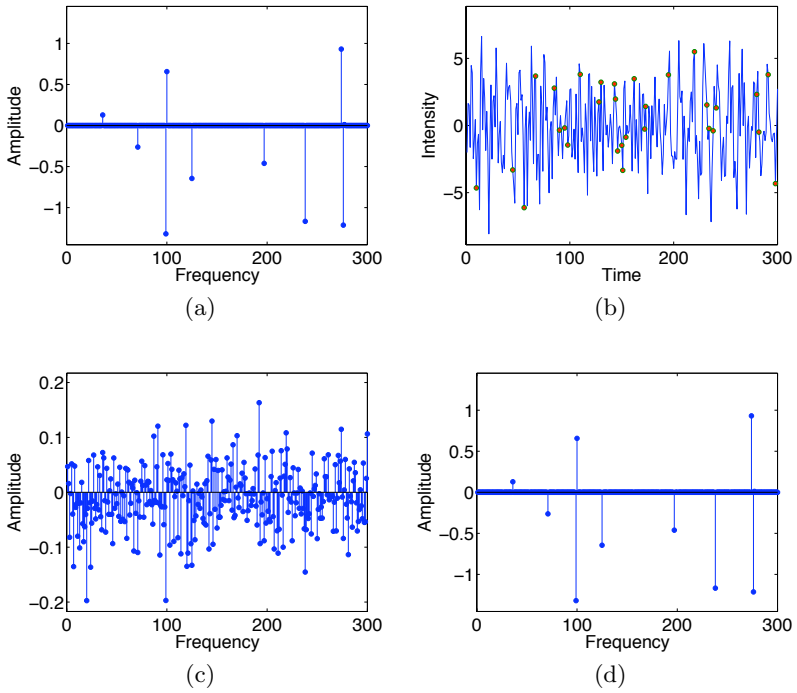

Fig. 1.2. (a) 10-sparse Fourier spectrum, (b) time domain signal of length 300 with 30 samples, (c) reconstruction via ℓ_2-minimization, (d) exact reconstruction via ℓ_1-minimization

length 300, which is 10-sparse in the Fourier domain. It is recovered exactly by the method of basis pursuit (ℓ_1-minimization) from only 30 samples in the time domain. For reference, the traditional linear method of ℓ_2-minimization is also displayed, which clearly fails in reconstructing the original sparse spectrum. (More information on this setup can be found in Chapter 12.)

Random Matrices. The problem of providing provably optimal measurement matrices $\mathbf{A}$ is remarkably intriguing. It is to date an open problem to construct explicit matrices, which behave provably optimal in compressed sensing. Certain matrix constructions from sparse approximation and coding theory (equiangular tight frames, see Chapter 5) provide somewhat reasonable reconstruction guarantees, but these fall considerably short of the optimal achievable bounds. The breakthrough is achieved by passing to *random matrices*, and this discovery can be considered the birth of compressive sensing. A simple model is a Gaussian matrix whose entries consists of independent standard normal distributed random variables, or a Bernoulli matrix where the entries are independent random variables taking the values ± 1 with equal

probability. A key result in compressive sensing states that an s-sparse vector can be reconstructed from $\mathbf{y} = \mathbf{A}\mathbf{x}$ with high probability using a variety of recovery algorithms and a random draw of an $m \times N$ Gaussian or Bernoulli matrix $\mathbf{A}$ provided that

$$m \geq Cs \ln(N/s) \,, \tag{1.3}$$

where $C > 0$ is a universal constant (independent of s, m, N). This bound is optimal.

The bound (1.3) tells us that the amount m of data needed to recover $\mathbf{x}$ scales linearly in the sparsity s of $\mathbf{x}$, while the signal length N has only very mild logarithmic influence. In particular, if the sparsity s is small compared to N then the number of measurements m can also be chosen small in comparison with N, so that we can exactly solve an underdetermined system of linear equations! This fascinating discovery is potentially useful for many applications.

Similar results also hold in the more practical situation of sampling. Assuming that the function of interest has a sparse orthogonal expansion with respect to a suitable system such as the trigonometric monomials one can recover it from a small number of randomly chosen samples via ℓ_1-minimization or several other reconstruction methods. This connection to sampling theory also explains the alternative name compressive sampling.

Stability. It is another important feature of compressive sensing that recovery algorithms are stable. This means that the error of reconstruction remains controlled when passing from sparse to compressible (approximately sparse) vectors and also when the measurements $\mathbf{y}$ are corrupted by noise. In this situation one may, for instance, consider the quadratically constraint ℓ_1-minimization problem

$$\text{minimize} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta \,. \tag{1.4}$$

Without stability compressive sensing would indeed not be very interesting for practical applications where one usually encounters noise, and where signals are only approximated well by sparse ones but are almost never sparse in the strict sense.

1.2 Motivations and Applications

In this section, we present a selection of problems that can be modeled as or that reduce to the standard compressive sensing problem. We hope to thereby convince the reader of its ubiquity. For ease of presentation, an informal style is adopted throughout this section. We do not go into technical details of specific applications, and often describe an idealized mathematical model. In the Notes section at the end of the chapter, we point to references which treat the applications in much more depth.

Sampling

An important task for many applications in technology and science is to reconstruct a continuous-time signal from a discrete set of samples. Examples include, image processing, and sensor technology in general, and analog-to-digital conversion as appearing for instance in audio entertaining systems or mobile communication devices. Currently, most sampling techniques rely on the Shannon sampling theorem, which states that a function of bandwidth B has to sampled at the rate $2B$ in order to ensure reconstruction.

In mathematical terms, the Fourier transform of a continuous time signal $f \in L^1(\mathbb{R})$, that is, $\int_{\mathbb{R}} |f(t)|dt < \infty$, is defined by

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(t) e^{-2\pi i t} dt \ , \quad \xi \in \mathbb{R} \ ,$$

We say that f is bandlimited with bandwidth B if $\widehat{f}$ is supported in $[-B, B]$. Shannon's sampling theorem states that such f can be reconstructed from its discrete sequence of samples $f(k/(2B))$, $k \in \mathbb{Z}$, via the sampling series

$$f(t) = \sum_{k \in \mathbb{Z}} f\left(\frac{k}{2B}\right) \mathrm{sinc}(2\pi B t - \pi k) \ , \tag{1.5}$$

where the sinc function is defined by

$$\mathrm{sinc}(t) = \begin{cases} \dfrac{\sin t}{t} & \text{if } t \neq 0 \ , \\ 1 & \text{if } t = 0 \ . \end{cases}$$

In order to simplify comparison with compressive sensing we formulate Shannon's sampling theorem in a finite dimensional context. We consider trigonometric polynomials of maximal degree M,

$$f(t) = \sum_{k=-M}^{M} x_k e^{2\pi i k t} \ , \quad t \in [0, 1] \ . \tag{1.6}$$

The degree M serves here as a substitute for the bandwidth B. As the space of trigonometric polynomials of degree at most M has dimension $N = 2M+1$, it is reasonable that such f can be reconstructed from $N = 2M + 1$ samples. Indeed, Theorem C.1 in the Appendix states that

$$f(t) = \frac{1}{2M+1} \sum_{j=0}^{2M} f\left(\frac{j}{2M+1}\right) D_M\left(t - \frac{j}{2M+1}\right) \ , \quad t \in [0, 1] \ ,$$

where D_M is the Dirichlet kernel

$$D_M(t) = \sum_{k=-M}^{M} e^{2\pi i k t} = \begin{cases} \dfrac{\sin(\pi(2M+1)t)}{\sin(\pi t)} & \text{if } t \neq 0\,, \\ 2M+1 & \text{if } t = 0\,. \end{cases}$$

Due to dimensionality reasons, it is not possible to reconstruct a general trigonometric polynomial f of degree at most M from fewer than $N = 2M+1$ samples. In practice, however, the required degree M maybe very large, so that also a large number of samples is needed — sometimes significantly more than one is able to take with reasonable effort. So the question arises whether additional assumptions allow to reduce the required number of samples. If the vector $\mathbf{x} \in \mathbb{C}^N$ of Fourier coefficients of f in (1.6) is sparse or compressible then in fact much fewer than N samples may suffice for exact (or approximate) reconstruction. Compressibility of Fourier coefficients is indeed a reasonable assumption in many practical scenarios.

Given a set $\{t_1, \dots, t_m\} \subset [0,1]$ of m sampling points we can write the vector $\mathbf{y} = (f(t_\ell))_{\ell=1}^m$ as

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1.7}$$

where $\mathbf{A} \in \mathbb{C}^{m \times N}$, $N = 2M+1$, is the Fourier type matrix with entries

$$A_{\ell k} = e^{2\pi i k t_\ell}, \quad \ell = 1, \dots, m, \quad k = -M, \dots, M\,.$$

The problem of reconstructing f from its vector $\mathbf{y}$ of m samples reduces to finding the coefficient vector $\mathbf{x}$. This amounts to solving the linear system (1.7), which is underdetermined when $m < N$. Due to the sparsity assumption, we arrive at a compressive sensing problem and related recovery algorithms including ℓ_1-minimization (1.2) apply. A crucial question in this context is how the sampling points should be chosen. As already indicated in the previous section, randomness helps. In fact, we will see in Chapter 12 that choosing the sampling points $t_1, \dots, t_m$ independently and uniformly at random in $[0,1]$ allows to reconstruct f from its m samples $f(t_1), \dots, f(t_m)$ with high probability provided that $m \geq Cs\ln(N)$. Thus, only few samples suffice if s is small. An illustrating example is displayed in Figure 1.2.

Imaging: Single-pixel camera

A device which implements ideas of compressive sensing is the single-pixel camera. The idea is to correlate a real-world image with independent realizations of Bernoulli random vectors in hardware, and measure such correlations, that is, inner products on a single pixel. It turns out that only a rather small number of such measured random inner products suffices for the reconstruction of images.

For the purpose of this exposition, we represent images via gray values of pixels collected in the vector $\mathbf{z} \in \mathbb{R}^N$, where $N = N_1 \times N_2$ and N_1, N_2 denote the width and height of the image in pixels. Usually, images are not sparse in the canonical (pixel) basis, but they are often sparse after a suitable

transformation, for instance, a wavelet transform or discrete cosine transform. This means that we can write $\mathbf{z} = \mathbf{W}\mathbf{x}$, where $\mathbf{x}$ is sparse or compressible and $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a unitary matrix representing the transform.

The crucial ingredient of the single-pixel camera is a micro mirror array consisting of a large number of small mirrors, which can be turned on or off individually. The light of the image is reflected at this mirror array and a lense combines all reflected light on one sensor, the single pixel of the camera. (INCLUDE PICTURE) Depending on whether a small mirror is switched on or off, the light from the corresponding pixel adds up to measured intensity at the sensor or not. In this way, we can realize in hardware the inner product $\langle \mathbf{x}, \mathbf{b}\rangle$ of the image $\mathbf{x}$ with a vector $\mathbf{b}$ containing ones at the locations corresponding to mirrors that are switched on, and zeros for the switched off mirrors. We can also realize inner products with vectors $\mathbf{a}$ containing only $+1$ and -1 as entries by defining two auxiliary vectors $\mathbf{b}^1, \mathbf{b}^2 \in \{0,1\}^N$ via

$$b_j^1 = \begin{cases} 1 & \text{if } a_j = 1 \\ 0 & \text{if } a_j = -1 \end{cases} \qquad b_j^2 = \begin{cases} 1 & \text{if } a_j = -1 \\ 0 & \text{if } a_j = 1 \end{cases}$$

Then $\langle \mathbf{x}, \mathbf{a}\rangle = \langle \mathbf{x}, \mathbf{b}^1\rangle - \langle \mathbf{x}, \mathbf{b}^2\rangle$. Choosing several vectors $\mathbf{a}_\ell$, $\ell = 1, \ldots, m$, independently at random whose entries take the values ± 1 with equal probability, the measured intensities $\mathbf{y}_\ell = \langle \mathbf{z}, \mathbf{a}_\ell\rangle$ become inner products with Bernoulli vectors, and we can write $\mathbf{y} = \mathbf{A}\mathbf{z}$, where $\mathbf{A} \in \mathbb{R}^{m\times N}$ is a (random) Bernoulli matrix. This means that one is able to realize the action of a random matrix on the image $\mathbf{z}$ in hardware. Writing $\mathbf{z} = \mathbf{W}\mathbf{x}$ and $\mathbf{A}' = \mathbf{A}\mathbf{W}$ with a suitable transform matrix $\mathbf{W} \in \mathbb{C}^{N\times N}$, yields the system

$$\mathbf{y} = \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{W}\mathbf{x} = \mathbf{A}'\mathbf{x}\,,$$

where now $\mathbf{x}$ can be assumed to be sparse or compressible. In this context, samples are taken sequentially which may take some time, so that we would clearly like to work with a minimal number m. Therefore, we have arrived at a compressive sensing problem. Once the sparse or compressible $\mathbf{x}$ is reconstructed from $\mathbf{y}$, we obtain back the image as $\mathbf{z} = \mathbf{W}\mathbf{x}$. We will see in Chapter 9 that it is possible to accurately recover an image $\mathbf{z}$, which is (approximately) s-sparse in some transform domain, from $m \geq Cs\ln(N/s)$ samples via efficient algorithms such as ℓ_1-minimization. Figure ?? (INCLUDE PICTURES?) illustrates how the single-pixel camera performs in practice.

While the single pixel camera is more a proof of concept rather than really a new trend in camera design, it is plausible that similar devices may be useful for different imaging tasks. For instance, for certain wavelengths different from the visible light, it is indeed impossible or at least very expensive to build chips that have millions of sensor pixels on an area of several square millimeters. In such context, one would expect that technology based on compressive sensing has the potential to really pay off.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is an important technology for medical imaging, which can be used for various tasks such as brain imaging, angiography (imaging of blood vessels), dynamic heart imaging, and more. In traditional approaches (essentially based on Shannon sampling), the measurement time necessary to achieve high resolution images is very high (up to several minutes or hours depending on the task), and unrealistic in clinical situations. For instance, for heart imaging patients, need to hold their breath, and one cannot expect a patient requiring heart diagnosis to do this for a long time. In such situations, it becomes promising to use compressive sensing in order to achieve high resolution images based on a minimal number of samples.

MRI relies on the interaction of hydrogen nuclei (protons), contained in water molecules in the body, with a strong magnetic field. A static magnetic field polarizes the spin of the protons resulting in a magnetic moment. Applying an additional radio frequency excitation field produces a precessing magnetization transverse to the static field. The precession frequency depends linearly on the strength of the magnetic field. The generated electromagnetic field can be detected by sensors. Imposing further magnetic fields with a spatially dependent strength, the precession frequency depends on the spatial position as well. Exploiting that the transverse magnetization depends on the physical properties of the tissue (for instance, proton density) allows to reconstruct an image of the body from the measured signal.

In mathematical terms, we denote the transverse magnetization by $X(\mathbf{z}) = |X(\mathbf{z})|e^{-i\phi(\mathbf{z})}$, $\mathbf{z} \in \mathbb{R}^3$, where $|X(\mathbf{z})|$ is the magnitude and $\phi(\mathbf{z})$ the phase. The additional possibly time-dependent magnetic field is designed to depend linearly on position and is therefore called gradient field. Denoting by $G \in \mathbb{R}^3$ its gradient the precession frequency as a function of position can be written

$$\omega(\mathbf{z}) = \kappa(B + \langle G, \mathbf{z}\rangle) \ , \quad \mathbf{z} \in \mathbb{R}^3 \ ,$$

where B is the strength of the static field and κ is a physical constant. With a time dependent gradient $G = G(t)$, $t \in [0, T]$, the magnetization phase $\phi(\mathbf{z}) = \phi(\mathbf{z}, t)$ is the integral

$$\phi(\mathbf{z}, t) = 2\pi\kappa \int_0^t G(r) \cdot \mathbf{z} dr \ ,$$

where $t = 0$ corresponds to the time of the radio frequency excitation. We introduce the function $k : [0, T] \to \mathbb{R}^3$ by

$$k(t) = \kappa \int_0^t G(u) du \ .$$

The receiver coil integrates over the whole spatial volume and measures the signal

$$f(t) = \int_{\mathbb{R}^3} |X(\mathbf{z})| e^{-2\pi i k(t)\cdot \mathbf{z}} d\mathbf{z} = \mathcal{F}(|X|)(k(t)) \ ,$$

where $\mathcal{F}(|X|)(\boldsymbol{\xi}) = \int_{\mathbb{R}^3} |X|(\mathbf{z}) e^{-2\pi i \boldsymbol{\xi}\cdot\mathbf{z}} d\mathbf{z}$ denotes the 3-dimensional Fourier transform of the absolute value $|X|$ of the magnetization. It is also possible to measure slices of a body, in which case the 3-dimensional Fourier transform is replaced by a 2-dimensional Fourier transform.

In conclusion, the signal measured by the NMR system is the Fourier transform of the spatially dependent magnetization M (the image), subsampled on the curve $\{k(t) : t \in [0,T]\} \subset \mathbb{R}^3$. By repeating several radio frequency excitations with modified parameters, one obtains samples of the Fourier transform of M along several curves $k_1, \ldots, k_L$ in $\mathbb{R}^3$. The required measurement time is proportional to the number L of such curves, and we would like to work with a minimal such number L.

A natural discretization represents each volume element (or area element in case of 2D imaging of slices) by a single voxel (or pixel), so that the magnitization $|X|$ becomes a finite-dimensional vector $\mathbf{x} \in \mathbb{R}^N$ indexed by $Q := [N_1] \times [N_2] \times [N_3]$ with $\mathrm{card}(Q) = N_1 N_2 N_3 = N$. After discretizing also the curves $k_1, \ldots, k_L$, our measured data become samples of the three-dimensional discrete Fourier transform of $\mathbf{x}$,

$$(\mathcal{F}\mathbf{x})_{\mathbf{k}} = \sum_{\boldsymbol{\ell} \in Q} x_{\boldsymbol{\ell}} e^{-2\pi i \mathbf{k}\cdot\boldsymbol{\ell}/M} \ , \quad \mathbf{k} \in [M]^3 \ .$$

Let $K \subset [M]^3 = Q$ of cardinality $\mathrm{card}(K) = m$ denote a subset of the discretized frequency space Q, which is covered by the trajectories $k_1, \ldots, k_L$. Then the measured data vector $\mathbf{y}$ corresponds to

$$\mathbf{y} = \mathbf{R}_K \mathcal{F}\mathbf{x} = \mathbf{A}\mathbf{x} \ ,$$

$\mathbf{R}_K$ is the linear map that restricts a vector indexed by $Q = [N_1] \times [N_2] \times [N_3]$ to its indices in K. Furthermore, $\mathbf{A} = \mathbf{R}_K \mathcal{F} \in \mathbb{C}^{m \times N}$ is a partial Fourier measurement matrix. In words, the vector $\mathbf{y}$ collects the samples of the three-dimensional Fourier transform of the discretized image $\mathbf{x}$ on the set K. Since we would like to work with a set K of samples with minimal cardinality m, we end up with an underdetermined system of equations.

In certain medical imaging applications such as angiography it is realistic to assume that the image $\mathbf{x}$ is sparse with respect to the canonical basis, so that we immediately arrive at a compressive sensing problem and corresponding reconstruction algorithms apply. In the general scenario, the discretized image $\mathbf{x}$ will be sparse or compressible only after transforming into a suitable domain, for instance, wavelets — in mathematical terms $\mathbf{x} = \mathbf{W}\mathbf{x}'$ for some unitary matrix $\mathbf{W} \in \mathbb{C}^{N \times N}$ and some sparse $\mathbf{x}' \in \mathbb{C}^N$. This leads to the model

$$\mathbf{y} = \mathbf{A}\mathbf{W}\mathbf{x}' = \mathbf{A}'\mathbf{x}' \ ,$$

with the transformed measurement matrix $\mathbf{A}' = \mathbf{A}\mathbf{W} = \mathbf{R}_K \mathcal{F}\mathbf{W} \in \mathbb{C}^{m \times N}$ and a sparse (compressible) vector $\mathbf{x}'$. Again, we arrived at a compressive sensing problem.

The challenge is to determine good sampling sets K of small cardinality m that ensure recovery of sparse images $\mathbf{x}$. The presently available theory predicts that sampling sets K that are chosen uniformly at random among all possible sets of cardinality m work well (at least when $\mathbf{W}$ is the identity matrix). Indeed, the results in Chapter 12 predict that an s-sparse image can be reconstructed if $m \geq Cs \ln N$.

Unfortunately, such type of random sets K are difficult to realize in practice due to the constraints that the trajectories $k_1, \ldots, k_L$ are continuous curves. Therefore, realizable good sets K are investigated empirically. One option that seemingly works well is to choose the trajectories as parallel lines in $\mathbb{R}^3$ whose intersection with a coordinate plane is chosen uniformly at random. This gives some sort of approximation to the case where K is chosen "completely" at random. Other choices such as perturbed spirals are also possible.

(INCLUDE PICTURES FROM SOME EXPERIMENTAL WORK??)

Radar

There are several tasks in radar for which compressive sensing can be potentially applied. Let us describe one of these.

An antenna sends out a suitably designed electromagnetic wave — the radar pulse — which is scattered at objects in the surrounding environment, for instance, airplanes in the sky. A receive antenna measures an electromagnetic signal resulting from the scattered waves. Based on the delay of the received signal, one can determine the distance of an object, and the Doppler effect allows to deduce its speed.

Let us describe a simple finite-dimensional model for this scenario. We denote by $T_k z_j = z_{j-k \mod m}$ the cyclic translation operator on $\mathbb{C}^m$ and by $M_\ell z_j = e^{2\pi i \ell j/N} z_j$ the modulation operator. The map transforming the sent signal to the received signal — also called channel — can be modeled as

$$\mathbf{B} = \sum_{(k,\ell) \in [m]^2} x_{k,\ell} T_k M_\ell \ ,$$

where the translations correspond to delay and the modulations model the Doppler effect. The vector $\mathbf{x} = (x_{k,\ell})$ characterizes the channel. A nonzero entry $x_{k,\ell}$ occurs if there is a scattering object present in the surrounding with distance and speed corresponding to the shift T_k and modulation M_ℓ. Only a limited number of scattering objects are usually present, which results in sparsity of the coefficient vector $\mathbf{x}$. The task is to determine $\mathbf{x}$ and thereby to obtain information about scatterers in the surrounding, by probing the channel with a suitable known radio pulse, modeled in this finite-dimensional setup by a vector $\mathbf{g} \in \mathbb{C}^m$. The received signal $\mathbf{y}$ is then given by

$$\mathbf{y} = \mathbf{B}\mathbf{g} = \sum_{(k,\ell) \in [m]^2} x_{k,\ell} T_k M_\ell \mathbf{g} = \mathbf{A}_\mathbf{g} \mathbf{x} \ ,$$

where the m^2 columns of the measurement matrix $\mathbf{A}_\mathbf{g} \in \mathbb{C}^{m \times m^2}$ are given by $T_k M_\ell \mathbf{g}$, $(k, \ell) \in [m]^2$. Recovering $\mathbf{x} \in \mathbb{C}^{m^2}$ from the measured signal $\mathbf{y}$ amounts to solving an underdetermined linear system. Taking the sparsity of $\mathbf{x}$ into account we arrive at a compressive sensing problem, and associated reconstruction algorithms including ℓ_1-minimization apply.

It remains to find suitable radio pulse sequences $\mathbf{g} \in \mathbb{C}^m$, which ensure that $\mathbf{x}$ can be recovered from $\mathbf{y} = \mathbf{B}\mathbf{g}$ with $\mathbf{B} = \sum_{(k,\ell) \in [m]^2} x_{k,\ell} T_k M_\ell$. A popular choice of $\mathbf{g}$ is the so-called *Alltop* vector, which is defined for prime $m \geq 5$ as

$$g_\ell = e^{2\pi i \ell^3 / m}, \quad \ell \in [m] \, .$$

We refer to Chapter 5 for more details. Although this choice works very well in practice, the theoretical guarantees that are presently available are somewhat limited due to the fact $\mathbf{g}$ is deterministic. As an alternative and in consistence with the general philosophy of compressive sensing, one can choose $\mathbf{g} \in \mathbb{C}^m$ at random, for instance, as a Bernoulli vector with independent ± 1 entries. It is known that an s-sparse vector $\mathbf{x} \in \mathbb{C}^{m^2}$ can be recovered from $\mathbf{y} = \mathbf{B}\mathbf{x} \in \mathbb{C}^m$ provided $s \leq Cm / \ln m$. More information can be found in the Notes section of Chapter 12.

(SOME ILLUSTRATING PICTURES, GRAPHS ?)

Sparse Approximation

Compressive sensing builds on the empirical observation that many types of signals can be approximated by sparse ones. In this sense, compressive sensing can be seen as a subfield of sparse approximation. There is, however, a specific problem in sparse approximation, which leads to a task similar to the compressive sensing problem to recover a sparse vector $\mathbf{x} \in \mathbb{C}^N$ from an incomplete information $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$, where $m < N$.

Suppose that a vector $\mathbf{y} \in \mathbb{C}^m$ (usually a signal or image in applications) is to be represented as a linear combination of given elements $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{C}^m$ such that $\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} = \mathbb{C}^m$. The collection of such elements is called a dictionary. We do not require this system to be linearly independent, in particular, we allow N to be larger than m. Therefore, a representation $\mathbf{y} = \sum_{j=1}^N x_j \mathbf{a}_j$ is not unique, and we are interested in a representation with a small number of terms, i.e., a sparse representation. Redundant systems, where $N > m$, are indeed desired in certain cases, where a linearly independent system is too restrictive. For instance, in time-frequency analysis, bases of time-frequency shifts elements are only possible if the generator has poor time-frequency concentration (this is the Balian-Low theorem). Also unions of several bases are of interest. In such situations, one often wants to remove the drawback of the nonuniqueness of the expansion by considering the sparsest expansion.

We form the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$. Then finding the sparsest representation of $\mathbf{y}$ amounts to solving

$$\text{minimize } \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y} \,. \tag{P_0}$$

If we tolerate a representation error η, then one considers the slightly modified optimization problem

$$\text{minimize } \|\mathbf{z}\|_0 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\| \leq \eta \,. \tag{$\mathrm{P}_{0,\eta}$}$$

Clearly, (P_0) is the same optimization as encountered already in the previous section. Both optimization problems (P_0) and ($\mathrm{P}_{0,\eta}$) are NP-hard in general. To overcome this computational bottleneck, all algorithmic approaches for the compressive sensing problem discussed in this book, including for instance ℓ_1-minimization, are applicable in this context as well. The general conditions on $\mathbf{A}$ ensuring exact or approximate recovery of the sparsest vector $\mathbf{x}$, which will be derived in Chapters 4, 5 and 6 remain valid as well.

There are, however, some differences in the philosophy with respect to the compressive sensing problem. In the latter, one is often free to design the matrix $\mathbf{A}$ with appropriate properties, while the matrix $\mathbf{A}$ is usually given and fixed in the sparse approximation context. In particular, it usually is not very reasonable to assume that it is random, as is often done for compressive sensing. Since it is very hard to verify the appropriate conditions ensuring sparse recovery in the optimal parameter regime (that is, m linear in s up to log-factors), the guarantees that one can usually give fall short of the ones encountered for random matrices. An exception of this rule of thumb will be covered in Chapter 13 where recovery guarantees for randomly chosen signals are treated.

The second difference between sparse approximation and compressive sensing appears in the desired error estimates. While in compressed sensing, one is interested in the error on the coefficient level, that is, $\|\mathbf{x} - \mathbf{x}^\sharp\|$ where $\mathbf{x}$ is the original coefficient vector and $\mathbf{x}^\sharp$ is the reconstruction, in sparse approximation one is rather interested in approximating a given $\mathbf{y}$ with a sparse expansion $\mathbf{y}^\sharp = \sum_j x_j^\sharp \mathbf{a}_j$, so that one is interested in $\|\mathbf{y} - \mathbf{y}^\sharp\|$. With an estimate for $\|\mathbf{x} - \mathbf{x}^\sharp\|$ one is often able to get an estimate on $\|\mathbf{y} - \mathbf{y}^\sharp\| = \|\mathbf{A}(\mathbf{x} - \mathbf{x}^\sharp)\|$, but the converse direction is usually not possible.

Next we briefly describe some signal and image processing applications of sparse approximation.

- **Compression.** Suppose we have found a sparse approximation $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$ of a signal of $\mathbf{y}$ with a sparse vector $\widehat{\mathbf{x}}$. Then storing $\widehat{\mathbf{y}}$ amounts to storing only the nonzero coefficients of $\widehat{\mathbf{x}}$. Since $\widehat{\mathbf{x}}$ is sparse, significantly less memory is required than for storing the entries original signal $\mathbf{y}$.
- **Denoising.** Suppose that we observe a noisy version $\widetilde{\mathbf{y}} = \mathbf{y} + \mathbf{e}$ of the signal $\mathbf{y}$, where $\mathbf{e}$ represents a noise vector with $\|\mathbf{e}\| \leq \eta$. The task is then to remove the noise, and to recover a good approximation of the original signal $\mathbf{y}$. In general, if nothing is known about $\mathbf{y}$, this problem becomes ill-posed. However, assuming that $\mathbf{y}$ can be well-represented by a sparse expansion it is a reasonable approach to take a sparse approximation to

$\widetilde{\mathbf{y}}$. More precisely, we ideally choose the solution $\widehat{\mathbf{x}}$ to the ℓ_0-minimization problem ($\mathrm{P}_{0,\eta}$) with $\mathbf{y}$ replaced by the known signal $\widetilde{\mathbf{y}}$. Then we form $\widehat{\mathbf{y}} = \mathbf{A}\widehat{\mathbf{x}}$ as the denoised version of $\mathbf{y}$. In order to have a computationally tractable approach, one replaces the NP-hard problem ($\mathrm{P}_{0,\eta}$) by one of the compressive sensing (sparse approximation) algorithms, for instance the variant (1.4) of ℓ_1-minimization, which takes the noise into account, or the so-called basis pursuit denoising problem

$$\text{minimize } \|\mathbf{x}\|_1 + \lambda \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \,.$$

- **Data Separation.** Suppose that a vector $\mathbf{y} \in \mathbb{C}^m$ is the composition of two (or more) components, that is, $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$. Given $\mathbf{y}$ we wish to extract the unknown vectors $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{C}^m$. This problem appears in several signal processing tasks. For instance, astronomers would like to separate point structures (stars, galaxy clusters) from filaments in their images. Similarly, a task in audio processing consists in separating harmonic components (pure sinusoids) from short peaks.
  Without additional assumption this separation task is ill-posed. However, if both components $\mathbf{y}_1, \mathbf{y}_2$ have sparse representations in dictionaries $(\mathbf{a}_1, \ldots, \mathbf{a}_{N_1})$ and $(\mathbf{b}_1, \ldots, \mathbf{b}_{N_2})$ (possibly bases) of different nature (for instance, sinusoids and spikes), then the situation changes. We can then write

$$\mathbf{y} = \sum_{j=1}^{N_1} x_{1,j}\mathbf{a}_j + \sum_{j=1}^{N_2} x_{2,j}\mathbf{b}_j = \mathbf{A}\mathbf{x},$$

  where the matrix $\mathbf{A} \in \mathbb{C}^{m \times (N_1+N_2)}$ has columns $\mathbf{a}_1, \ldots, \mathbf{a}_{N_1}, \mathbf{b}_1, \ldots, \mathbf{b}_{N_2}$ and the vector $\mathbf{x} = [x_{1,1}, \ldots, x_{1,N_1}, x_{2,1}, \ldots, x_{2,N_2}]^\top$ is sparse. The compressive sensing methodology then allows — under certain conditions — to determine the coefficient vector $\mathbf{x}$, hence to derive the two components $\mathbf{y}_1 = \sum_{j=1}^{N_1} x_{1,j}\mathbf{a}_j$ and $\mathbf{y}_2 = \sum_{j=1}^{N_2} x_{2,j}\mathbf{b}_j$.

Error Correction

In data transmission, submitted pieces of the data are corrupted from time to time. In order to overcome this unavoidable problem in all realistic data transmission devices, one designs schemes to correct such errors provided they do not occur too often.

Suppose we would like to transmit a vector $\mathbf{z} \in \mathbb{R}^n$. Then a standard strategy is to encode it into a vector $\mathbf{v} = \mathbf{B}\mathbf{z} \in \mathbb{R}^N$ of length $N = n + m$, where $\mathbf{B} \in \mathbb{R}^{N \times n}$. The intuition is that the redundancy in $\mathbf{B}$ (due to $N > n$) should help in identifying transmission errors. The number m reflects the amount of redundancy.

Assume that the receiver measures $\mathbf{w} = \mathbf{v} + \mathbf{x} \in \mathbb{R}^N$, where $\mathbf{x}$ represents transmission error. The assumption that transmission errors occur only occasionally leads to sparsity in $\mathbf{x}$, say $\|\mathbf{x}\|_0 \leq s$. For decoding we construct

a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ — called generalized check sum matrix — such that $\mathbf{AB} = 0$, that is, all rows of $\mathbf{A}$ are orthogonal to all columns of $\mathbf{B}$. We form then the generalized checksum

$$\mathbf{y} = \mathbf{Aw} = \mathbf{A}(\mathbf{v} + \mathbf{x}) = \mathbf{ABz} + \mathbf{Ax} = \mathbf{Ax} \, .$$

We arrived at a standard compressive sensing problem with the matrix $\mathbf{A}$ and the sparse error vector $\mathbf{x}$. The methodology described in this book allows to recover $\mathbf{x}$ under suitable conditions, and thereby the original transmit vector $\mathbf{v} = \mathbf{w} - \mathbf{x}$. Then one solves the overdetermined system $\mathbf{v} = \mathbf{Bz}$ for the data vector $\mathbf{z}$.

In order to make this scheme concrete, we choose a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, which is suitable for compressive sensing, for instance, a Gaussian random matrix. Then we select the matrix $\mathbf{B} \in \mathbb{R}^{N \times n}$ with $n + m = N$ such that its columns span the orthogonal complement of the row space of $\mathbf{A}$. Then $\mathbf{AB} = 0$. With these choice we are able to correct transmission errors as long as the error vector $\mathbf{x}$ is s-sparse and $m \geq Cs \ln(N/s)$.

Statistics and Machine Learning

The goal of statistical regression is to predict an outcome based on certain input data. It is common to choose the linear model

$$\mathbf{y} = \mathbf{Ax} + \mathbf{e} \, ,$$

where $\mathbf{A} \in \mathbb{R}^{m \times N}$ — often called design or predictor matrix in this context — collects the input data, $\mathbf{y}$ the output data, and $\mathbf{e}$ a random noise vector. The vector $\mathbf{x}$ is a parameter that has to be estimated from the data. In the statistical context usually the notation (n, p) instead of (m, N) is used, but for consistency we keep the notation used throughout this book. For instance, in a clinical study the matrix entries $A_{j,k}$ for a fixed row j may refer to data for patient j, such as blood pressure, weight, height, gene data, concentration of certain markers, etc. The corresponding output y_j would be the quantity of interest, for instance, the probability that patient j suffers a certain disease. Having data for m patients, the regression task is to fit the model, that is, to determine the parameter vector $\mathbf{x}$.

In practice, the number N of parameters is often much larger than the number m of observations, so that even without the noise term $\mathbf{e}$ the problem of fitting the parameter $\mathbf{x}$ is ill-posed without further assumption. However, in many cases only a small number of parameters contribute to the effect one would like to predict, but it is not known a priori which of them are important. This leads to sparsity in the vector $\mathbf{x}$, and again we arrive at a compressive sensing type problem. In statistical terms, determining a sparse parameter vector $\mathbf{x}$ corresponds to selecting the relevant explanatory variables, that is, the support of $\mathbf{x}$. Therefore, one also speaks of *model selection*.

The methods described in this book can be applied also in this context. However, there is a slight difference from the usual setup treated in this book due to the randomness of the noise vector $\mathbf{e}$. In particular, rather than the noise-aware ℓ_1-minimization problem (1.4) one usually considers the so called LASSO (least absolute shrinkage and selection operator)

$$\text{minimize } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2 \quad \text{subject to } \|\mathbf{z}\|_1 \leq \tau \tag{1.8}$$

for an appropriate regularization parameter τ, depending on the variance of the noise. Further variants are the *Dantzig selector*

$$\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y})\|_\infty \leq \lambda \,, \tag{1.9}$$

or the ℓ_1-minimization problem (sometimes also called LASSO in the literature)

$$\text{minimize } \|\mathbf{z}\|_1 + \lambda \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2$$

again for appropriate choices of λ. We will not treat the statistical context further in this book, but mention that for both the LASSO and the Dantzig selector near optimal statistical estimation properties can be shown under conditions on $\mathbf{A}$ similar to the ones of the following chapters.

A closely related regression problem arises in **machine learning**. Given random pairs of samples (t_j, y_j), $j \in [m]$, where t_j is some input parameter vector and y_j is a scalar output, one would like to predict the output y corresponding to future input data t. The model relating the input t with the output y is

$$\mathbf{y} = f(t) + \mathbf{e} \,,$$

where $\mathbf{e}$ is random mean-zero noise. The task is to learn the function f based on training samples (t_j, y_j). Without further hypotheses on f, this task is impossible. Therefore, we assume that f is sparse in terms of a given dictionary of functions $\psi_1, \ldots, \psi_N$, that is,

$$f(t) = \sum_{\ell=1}^{N} x_\ell \psi_\ell(t) \,,$$

with a sparse coefficient vector $\mathbf{x}$. Introducing the matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ with entries

$$A_{j,k} = \psi_k(t_j)$$

we arrive at the model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \,,$$

and the task is to estimate the sparse coefficient vector $\mathbf{x}$. This problem is of the same form as the estimation problem described above, and the same estimation procedures including LASSO and the Dantzig selector apply.

Low Rank Matrix Recovery and Matrix Completion

Let us finally describe an extension of compressive sensing and some of its applications. Rather than recovering a sparse vector $\mathbf{x} \in \mathbb{C}^N$, we now aim at recovering a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ from incomplete information. Sparsity is replaced by the assumption that $\mathbf{X}$ has low rank. Indeed, the set of all matrices of a given small rank has much smaller complexity than the set of all matrices, so that recovery of low rank matrices seems plausible.

For a linear map $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ with $m < n_1 n_2$, suppose that we are given the measurement vector

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) \in \mathbb{C}^m \, .$$

Our task is to reconstruct $\mathbf{X}$ from $\mathbf{y}$. In order to have a chance of succeeding we assume that $\mathbf{X}$ has rank at most $r \ll \max\{n_1, n_2\}$. However, the natural approach of solving the optimization problem

$$\text{minimize } \operatorname{rank}(\mathbf{Z}) \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y}$$

is NP-hard. In order to illustrate the analogy with the compressive sensing problem, we consider the singular value decomposition of $\mathbf{X}$,

$$\mathbf{X} = \sum_{\ell=1}^{\min\{n_1,n_2\}} \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^* \, .$$

Here, $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_{\min\{n_1,n_2\}} \geq 0$ are the singular values of $\mathbf{X}$ and $\mathbf{u}_\ell \in \mathbb{C}^{n_1}$, $\mathbf{v}_\ell \in \mathbb{C}^{n_2}$ are the left and right singular values, respectively. We refer to Appendix A.2 for details. The matrix $\mathbf{X}$ is of rank r if and only if the vector $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\mathbf{X})$ of singular values is s-sparse, and $\operatorname{rank}(\mathbf{X}) = \|\boldsymbol{\sigma}(\mathbf{X})\|_0$. Having the ℓ_1-minimization approach for compressive sensing in mind, it is natural to introduce the so-called *nuclear norm* as the ℓ_1-norm of the singular values,

$$\|\mathbf{X}\|_* = \|\boldsymbol{\sigma}(\mathbf{X})\|_1 = \sum_{\ell=1}^{\min\{n_1,n_2\}} \sigma_\ell(\mathbf{X}) \, .$$

Then we consider the nuclear norm minimization problem

$$\text{minimize } \|\mathbf{Z}\|_* \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y} \, . \tag{1.10}$$

This is a convex optimization problem which can be solved efficiently, for instance, via a reformulation as a semidefinite program.

A very similar theory as for the recovery of sparse vectors can be developed, and appropriate conditions on $\mathcal{A}$ ensure exact or approximate recovery via nuclear norm minimization (and other algorithms). Again random maps $\mathcal{A}$ turn out to be optimal, and matrices $\mathbf{X}$ of rank at most r can be recovered from m measurements with high probability provided

$$m \geq Cr \max\{n_1, n_2\}.$$

This bound is optimal since the right hand side corresponds to the number of degrees of freedom required to describe an $n_1 \times n_2$ matrix of rank r. Remarkably, there is no log-factor necessary in contrast to sparse vector recovery.

A popular special case is the **matrix completion** problem, where one seeks to fill in missing entries of a low rank matrix. The measurement map $\mathcal{A}$ samples the entries $\mathcal{A}(\mathbf{X})_\ell = \mathbf{X}_{j,k}$ for some indices j, k depending on ℓ. This setup appears for instance in consumer taste prediction. Assume that an (online) store sells products indexed by the rows of the matrix, and consumers – indexed by the columns – are able to rate these products. Obviously, not every consumer will rate every product, so that only a limited number of the entries of this matrix is at our disposal. For purposes of individualized advertisement, the store is interested in obtaining a prediction of the whole matrix of consumer ratings. Often, if two customers both like some subset of products, they will both also like or dislike other subsets of products (there is essentially only a finite number of "types" of customers). Due to this reason, it can be assumed that the matrix of ratings has (at least approximately) low rank, and indeed, this is observed empirically. Therefore, methods from low rank matrix recovery apply in this setup, in particular, the nuclear norm minimization approach.

Although it is certainly interesting, we will not treat the low rank matrix recovery problem very intensively in this book. Nevertheless, due to the close analogy with compressive sensing for sparse vectors, the main results are covered within exercises, and the reader is invited to work through them.

1.3 Overview of the Book

Let us give an outline on the strategies how to attack the compressive sensing problem and on the basic mathematical results associated with such approaches.

The notion of **sparsity** and **compressibility** are at the core of compressive sensing. A vector $\mathbf{x} \in \mathbb{C}^N$ is called s-sparse if it has at most s non-zero entries, $\|\mathbf{x}\|_0 := \#\{\ell : x_\ell \neq 0\} \leq s$. Note that $\|\mathbf{x}\|_0$ is not a norm, although it has become customary to denote it with this symbol. In practice, one encounters usually vectors that are not exactly s-sparse, but are compressible in the sense that they can be well-approximated by sparse ones. To quantify this notion, one introduces the error of best s-term approximation by

$$\sigma_s(\mathbf{x})_p := \inf_{\|\mathbf{z}\|_0 \leq s} \|\mathbf{x} - \mathbf{z}\|_p \ .$$

Chapter 2 introduces these notions, provides their relation to weak ℓ_p-norms, and shows basic estimates for the error of best s-term approximation including

$$\sigma_s(\mathbf{x})_2 \leq \frac{1}{s^{1/q-1/2}} \|\mathbf{x}\|_q \,, \quad q \leq 2 \,. \tag{1.11}$$

Therefore, unit balls in the ℓ_q-norm for small $q \leq 1$ are good models for compressible vectors. We further study the problem of determining the minimal number of measurements m, which are required (at least in principle) to recover an s-sparse vector $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{C}^{m \times N}$ – namely $m = 2s$. This is remarkable because the actual length N of the vector $\mathbf{x}$ does not play any role. The basic recovery procedure associated to these first recovery guarantees is ℓ_0-minimization

$$\text{minimize } \|\mathbf{x}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y}.$$

We will show in Section 2.3 that ℓ_0-minimization is NP-hard by relating it to the *exact cover by 3-sets problem*, which is known to be NP-complete. Therefore, ℓ_0-minimization is intractable in general, and therefore not useful for practical purposes.

In order to circumvent the computational bottleneck of ℓ_0-minimization, we introduce several tractable alternatives in **Chapter 3**. Rather than going into a detailed analysis at this point, we rather give the intuitive motivation for the respective algorithms, and present only very basic results about them. These recovery methods can be subsumed under roughly three categories: optimization methods, greedy methods and thresholding based methods. Optimization approaches include ℓ_1-minimization (1.2), also called *basis pursuit*, as well as quadratically constrained basis pursuit (also called basis pursuit denoising in the literature), which takes into account potential noise on the measurements. The corresponding optimization problems can be solved with various methods from convex optimization, including interior point methods. We will present specialized numerical methods for ℓ_1-minimization later in Chapter 15.

Orthogonal matching pursuit is a greedy method that builds up the support set of the reconstructed sparse vector in an iterative fashion by adding one element to the current support set in each step. The selection process is greedy in the sense that the element is chosen such that the residual in the next step is minimized. Such residual is then computed by performing an orthogonal projection of $\mathbf{y}$ onto the span of the already selected columns of $\mathbf{A}$. If s iterations are performed then clearly the reconstructed vector is s-sparse. Another greedy method to be presented is *compressive sampling matching pursuit* (CoSaMP), which selects several elements of the support in each step and then iteratively refines this selection.

The very simple recovery procedure of *basic thresholding* determines the support set in one step by choosing the s elements maximizing the correlation $|\langle \mathbf{x}, \mathbf{a}_\ell \rangle|$ of the signal $\mathbf{x}$ with the columns of the matrix $\mathbf{A}$. The reconstruction is then obtained by projecting into the span of the corresponding columns. While this method is very fast, its performance is usually somewhat limited.

A more powerful method is *iterative hard thresholding*. Starting with an initial vector $\mathbf{x}^0$, say $\mathbf{x}^0 = 0$, it iteratively computes

$$\mathbf{x}^{n+1} = H_s(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n) \, ,$$

where H_s denotes the hard-thresholding operator that keeps the s-largest entries of a vector and puts to zero all other entries. Without the operator H_s this would be a classical Landweber iteration, well-known in the area of inverse problems. The application of H_s ensures sparsity of all iterations $\mathbf{x}^n$. If the $\mathbf{x}^n$ converge to some $\mathbf{x}^\sharp$, this vector $\mathbf{x}^\sharp$ is the reconstruction. In practice one stops after a finite number of iterations, of course. Finally, we will present the *hard thresholding pursuit* algorithm, which combines iterative hard thresholding with an orthogonal projection step.

Chapter 4 is devoted to the analysis of basis pursuit (ℓ_1-minimization). We derive conditions that guarantee recovery of sparse vectors. The null space property is a necessary and sufficient condition on the measurement matrix $\mathbf{A}$ that guarantees exact recovery of all s-sparse vectors $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization. It basically requires that all vectors in the kernel of $\mathbf{A}$ are far from being sparse. This is natural because a non-zero sparse vector $\mathbf{v}$ in the kernel of $\mathbf{A}$ would result in the zero measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} = 0$, so that instead of $\mathbf{v}$ the zero vector would be reconstructed. Suitable refinements of the null space property — leading to the notion of stable null space property and robust null space property — ensure stability of reconstruction via basis pursuit under passing from exactly sparse vectors to compressible vectors and they guarantee robustness under adding noise on the measurements. We also derive conditions based on the sparse vector and the measurement matrix, which ensure recovery of this given vector via basis pursuit. While at first glance this does not seem useful because $\mathbf{x}$ is unknown in practice, we will exploit this in later chapters in order to show that a fixed sparse vector is recovered using a randomly chosen measurement matrix $\mathbf{A}$ with high probability. Such results do not ensure simultaneous recovery of all sparse vectors using a single matrix, and therefore are referred to as nonuniform recovery guarantees. Finally, we make a small detour to the low-rank matrix recovery problem, and show that the strategy of nuclear norm minimization (1.10) is successful if and only if a suitable adaptation of the null space property to the matrix recovery setup holds. Further results concerning low-rank matrix recovery are treated within the exercises.

The null space property is usually not easily accessible by a direct computation. The *coherence* is a much simpler concept measuring the quality of a measurement matrix. For $\mathbf{A} \in \mathbb{C}^{m \times N}$ with ℓ_2-normalized columns $\mathbf{a}_j$ it is defined as

$$\mu := \max_{j \neq k} |\langle \mathbf{a}_j, \mathbf{a}_k \rangle| \, .$$

Ideally, μ should be as small as possible for a good measurement matrix. **Chapter 5** provides first some simple consequences on the conditioning of

column submatrices when this is the case. We will also introduce the ℓ_1-coherence function which slightly refines the coherence μ. A fundamental lower bound on how small μ can get is

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}} \, .$$

For large N the right hand side scales like $1/\sqrt{m}$. A similar lower bound will also be provided for the ℓ_1-coherence function. The matrices achieving the lower bound (both for μ and the ℓ_1-coherence functions) are characterized as equiangular tight frames. We will investigate under which conditions on m and N equiangular tight frames exist, and provide an explicit example of an $m \times m^2$ matrix (m being prime) with almost optimal coherence. Finally, based on the coherence, we analyze several recovery algorithms including ℓ_1-minimization, orthogonal matching pursuit and hard thresholding pursuit. For all of these, we obtain a first sufficient condition for the recovery of all s-sparse vectors $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$, namely

$$(2s-1)\mu < 1 \, .$$

In other words, if the sparsity is small enough then all these algorithms are able to perfectly recovery $\mathbf{x}$ from underdetermined linear information. Choosing a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with near-optimal coherence of the order $\mu \leq c/\sqrt{m}$ (which imposes some mild conditions on N), then we obtain an explicit bound on the number m of measurement,

$$m \geq Cs^2 \, . \tag{1.12}$$

If s is very small, $s \ll \sqrt{N}$, then we may in particular choose $m < N$ and still be able to recover from incomplete information. While at first sight one could be satisfied with this result, significantly better estimates are possible as already outlined above. In fact, we will see in later chapters that the optimal (and achievable) estimate is $m \geq Cs \log(N/s)$, so that up to the log-factor the number of measurements scales linearly in the sparsity rather than quadratic as in the above bound. The advantage of the coherence-based approach is, however, its simplicity (indeed, the corresponding analysis of the various recovery algorithms is comparably short) and the fact that explicit (deterministic) constructions of measurement matrices are available.

In order to overcome the so-called quadratic bottleneck in (1.12), the concept of the *restricted isometric property* (RIP) proves to be very powerful. The restricted isometry constant δ_s of a matrix $\mathbf{A}$ is defined as the smallest number such that

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_s)\|\mathbf{x}\|_2^2 \quad \text{for all } s\text{-sparse } \mathbf{x} \, .$$

Informally, $\mathbf{A}$ is said to possess the RIP if δ_s is small for sufficiently large s. Noting that for a vector $\mathbf{x}$ with support S we have $\mathbf{A}\mathbf{x} = \mathbf{A}_S\mathbf{x}_S$, where

$\mathbf{A}_S$ is the column-submatrix corresponding to the indices in S and $\mathbf{x}_S$ is the restriction of $\mathbf{x}$ to S, we observe that the RIP requires that *all* column-submatrices with s columns are well-conditioned.

Chapter 6 starts by providing basic results on the restricted isometry constants. For instance, for $\mathbf{A}$ with ℓ_2-normalized columns, they are related to the coherence via $\delta_2 = \mu$. In this sense, the restricted isometry constants generalize the coherence by taking into account the interaction of s columns of $\mathbf{A}$ at the same time rather than only 2. Gershgorin's disc theorem implies the simple bound $\delta_s \leq (s-1)\mu$, which in relevant cases is, however, very pessimistic.

Then we turn to the analysis of the various recovery algorithms based on the restricted isometry property of $\mathbf{A}$. Under a condition of the type

$$\delta_{\kappa s} \leq \delta^* \tag{1.13}$$

for an appropriate small integer κ and some $\delta^* < 1$ (both depending only on the algorithm) every s-sparse vector $\mathbf{x}$ is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$. For instance, for *basis pursuit* we will show the sufficient condition $\delta_{2s} < 0.4931$. Moreover, the reconstruction is stable under passing from sparse to compressible vector and also when adding noise on the measurement vector $\mathbf{y}$. More precisely, denoting by $\mathbf{x}^\sharp$ the reconstruction from $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta$ using any of the analyzed algorithms then, under (1.13), we have the following error estimates in ℓ_1 and in ℓ_2,

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \leq C_1 \frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + C_2\eta\,, \tag{1.14}$$

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \leq C_1\sigma_s(\mathbf{x})_1 + C_2\sqrt{s}\eta\,, \tag{1.15}$$

for absolute constants $C_1, C_2 > 0$.

As mentioned above, the restricted isometry constants are introduced in order to overcome the quadratic bottleneck (1.12) on the number m of required measurements in terms of the sparsity s. However, it is actually an open problem up-to-date to provide explicit (deterministic) matrices that provably achieve linear scaling of m in s (up to log-factors). The reason may be that the usual tool for deterministic condition number estimates is Gershgorin's disc theorem. The latter, however, essentially boils down to estimating the coherence (or the ℓ_1-coherence function), and then using the already mentioned bound $\delta_s \leq (s-1)\mu$. However, due to the lower bound $\mu \geq 1/\sqrt{m}$ (valid for large $N \geq 2m$, say) this technique only shows that $\delta_s \leq \delta_0$ once $m \geq cs^2$, so we again encounter the quadratic bottleneck. In essence, one must avoid the use of Gershgorin's disc theorem for going beyond, but it is presently not clear which techniques may serve as a replacement in a deterministic setting.

We overcome this problem by passing to random matrices. Then a whole new set of tools from probability theory becomes available, which allow to show that the restricted isometry property (or other conditions ensuring recovery) holds with high probability when $m \geq Cs\log(N/s)$ provided that

A is drawn according to a suitable probability model. **Chapters 7 and 8** introduce to the necessary background from probability theory.

We start in **Chapter 7** by recalling basic concepts and results from probability theory, such as expectation, moments, Gaussian random variables and vectors, Jensen's inequality, etc. Then we treat the relation between the growth of the moments of a random variable and its tail. It will become crucial later to bound the tail of a sum of independent random variables. Cramér's theorem provides a very general estimate using the moment generating functions of the involved random variables. Hoeffding's inequality specializes to the sum of independent bounded mean-zero random variables. Gaussian and Rademacher / Bernoulli variables (the latter taking the values ± 1 with equal probability) fall into the larger class of subgaussian random variables, for which we will present basic results. Finally, Bernstein inequalities refine Hoeffding's inequality by taking into account the variance of the random variables. Further, they extend to possibly unbounded subexponential random variables.

For many results on compressive sensing with Gaussian or Bernoulli random matrices — that is, for large parts of Chapter 9 and 11, including bounds for the restricted isometry property — the relatively simple probabilistic tools of Chapter 7 are already sufficient. Several topics in compressive sensing, however, including for instance the analysis of random partial Fourier matrices, build on more advanced tools from probability theory. **Chapter 8** presents the required material. For instance, we cover Rademacher sums of the form $\sum_j \epsilon_j a_j$, where the $\epsilon_j = \pm 1$ are independent Rademacher variables and the symmetrization technique leading to such sums. Khintchine inequalities bound the moments of Rademacher sums. Decoupling techniques allow to reduce the amount of dependencies by replacing some occurencies of random variables by independent copies in certain expressions. The noncommutative Bernstein inequality provides a tail bound for the operator norm of independent mean-zero random matrices. Dudley's inequality bounds the expected supremum over a family of random variables indexed by some set by a geometric quantity of that set. Slepian's and Gordon's lemma compare expected maxima (minima of maxima) of two families of Gaussian random vectors. Concentration of measure describes the general phenomenon that in high-dimensional spaces functions of random vectors often concentrate around their mean. We present such a result for Lipschitz functions of Gaussian random vectors.

Having the probabilistic tools at hand we are prepared to investigate the use of Gaussian, Bernoulli and, more generally, subgaussian random matrices in **Chapter 9**. A crucial ingredient for the proof of the restricted isometry property is the concentration inequality

$$\mathbb{P}(|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2| \geq t\|\mathbf{x}\|_2^2) \leq 2\exp(-cmt^2)\,, \tag{1.16}$$

valid for any fixed $\mathbf{x} \in \mathbb{R}^N$, $t \in (0,1)$, and a random draw of a suitably scaled subgaussian random matrix $\mathbf{A}$. Using covering arguments — in particular

exploiting covering number bounds in Appendix C.2 — we deduce that the restricted isometry constants of a (suitably scaled) $m \times N$ subgaussian random matrix satisfy $\delta_s \leq \delta$ with high probability provided

$$m \geq C\delta^{-2} s \ln(eN/s) \; . \tag{1.17}$$

Sparsity is often with respect to an orthonormal basis different from the canonical basis. It follows from the invariance of the concentration inequality under orthogonal transformations that subgaussian random matrices are *universal* in the sense that the the signals $\mathbf{x}$ may as well be sparse with respect to an arbitrary (but fixed) orthonormal basis.

The special case of Gaussian random matrices allows to use refined methods that are not available in the subgaussian case, such as Gordon's lemma and concentration of measure. We will deduce explicit and good constants in the nonuniform setting where we only ask for recovery of a fixed s-sparse vector using a random draw of an $m \times N$ Gaussian matrix. For large dimensions we roughly obtain that

$$m \geq 2s \ln(N/s)$$

is sufficient to recover an s-sparse vector using ℓ_1-minimization, see Chapter 9 for precise statements. This is the general rule of thumb for compressive sensing, and reflects well the outcome of empirical tests, even for random matrices different from Gaussian — although the proof of the result above applies only to the Gaussian case. Moreover, in the Gaussian case we can also analyze the null space property directly without passing to the restricted isometry property. Replacing the constant 2 in the above condition by roughly 8 we obtain stable and uniform recovery via ℓ_1-minimization.

We will close Chapter 9 with a detour to the Johnson-Lindenstrauss lemma, which states that a finite point set in a large dimensional space can be mapped to a significantly lower dimensional space by nearly preserving all mutual distances. (No sparsity assumptions are involved here.) Indeed, this is an immediate consequence of the concentration inequality (1.16). In this sense, the Johnson-Lindenstrauss lemma implies the RIP. We will show that a converse is also true: If $\mathbf{A}$ satisfies the RIP then randomizing its column signs yields a Johnson-Lindenstrauss embedding with high probability.

In **Chapter 10** we show that the bound (1.17) on the number of required measurements deduced for subgaussian random matrices is optimal by relating the compressive sensing problem to Gelfand widths of ℓ_1-balls. More precisely, for a subset K of a normed space $X = (\mathbb{R}^N, \|\cdot\|)$ and $m \leq N$, we introduce the quantity

$$E^m(K, X) := \inf \left\{ \sup_{\mathbf{x} \in K} \|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|, \; \mathbf{A} \in \mathbb{R}^{m \times N}, \; \Delta : \mathbb{R}^m \to \mathbb{R}^N \right\} \; ,$$

which provides the worst possible reconstruction error with compressive sensing using the best possible reconstruction map Δ, over all vectors from the set K. The Gelfand widths are defined as

$$d^m(K,X) := \inf\left\{ \sup_{\mathbf{x}\in K\cap\ker \mathbf{A}} \|\mathbf{x}\|,\ \mathbf{A}\in\mathbb{R}^{m\times N}\right\}, \quad m\le N.$$

If $K=-K$ and $K+K\subset aK$ for some constant a, as it is the case with $a=2$ for a unit ball in some norm, then

$$d^m(K,X)\le E^m(K,X)\le 2d^m(K,X)\ .$$

Unit balls $K=B_q^N$ in the N-dimensional ℓ_q-space, $q\le 1$, are good models for compressible vectors by (1.11), so that we are lead to study their Gelfand widths. For ease of exposition we only cover the case $q=1$. An upper bound for $E^m(B_1^N,\ell_2^N)$, and thereby for $d^m(B_1^N,\ell_2^N)$ can be easily derived from the error estimate (1.14) for various recovery algorithms and from the bound for the restricted isometry property of subgaussian random matrices. This gives

$$d^m(B_1^N,\ell_2^N)\le C\left(\frac{\ln(eN/m)}{m}\right)^{1/2}\ .$$

We derive the matching lower bound

$$d^m(B_1^N,\ell_2^N)\ge c\left(\frac{\ln(eN/m)}{m}\right)^{1/2}\ ,$$

and thereby deduce that the bound (1.17) on the number of required measurements is optimal. It is of independent interest that an intermediate step in the proof of the lower bound for $d^m(B_1^N,\ell_2^N)$ shows that a necessary condition on the number of measurements required to recover every s-sparse vector $\mathbf{x}$ from $\mathbf{y}=\mathbf{A}\mathbf{x}$ via ℓ_1-minimization, for an arbitrary $m\times N$ matrix $\mathbf{A}$, is

$$m\ge Cs\ln(eN/s)\ . \tag{1.18}$$

The error bound in (1.14) features the term $\sigma_s(\mathbf{x})_1/\sqrt{s}$, although the error is measured with respect to the ℓ_2-norm. The question arises whether one can also obtain an error bound with rather the best s-term approximation error in ℓ_2, that is, $\sigma_s(\mathbf{x})_2$, on the right hand side of (1.14). **Chapter 11** investigates this question and the more general one whether a pair of measurement matrix $\mathbf{A}\in\mathbb{R}^{m\times N}$ and reconstruction map $\Delta:\mathbb{R}^m\to\mathbb{R}^N$ satisfies

$$\|\mathbf{x}-\Delta(\mathbf{A}\mathbf{x})\|_q\le\frac{C}{s^{1/p-1/q}}\sigma_s(\mathbf{x})_p\quad\text{ for all }\mathbf{x}\in\mathbb{R}^N\ .$$

If $q=p$ this bound is referred to as ℓ_p-*instance optimality* and in the general case as *mixed* (ℓ_p,ℓ_q)-*instance optimality*. The ℓ_1-instance optimality implies the familiar bound $m\ge Cs\ln(eN/s)$. However, ℓ_2-instance optimality necessarily leads to

$$m\ge cN\ ,$$

which is a regime of parameters not interesting for compressive sensing. However, we may ask for less, and require only that the error bound in ℓ_2 holds in a nonuniform setting, that is, for fixed $\mathbf{x}$ with high probability on a random draw of a subgaussian matrix. It turns out that we can have then indeed the error bound

$$\|\mathbf{x} - \Delta_1(\mathbf{A}\mathbf{x})\|_2 \leq C\sigma_s(\mathbf{x})_2 \ ,$$

with high probability under the condition $m \geq Cs\ln(eN/s)$, where Δ_1 refers to reconstruction via ℓ_1-minimization. The corresponding analysis uses the notion of ℓ_1-quotient property, which we will show to hold with high probability for Gaussian random matrices, and with a slight variation also for general subgaussian random matrices.

Chapter 11 investigates also another question. In the setup of noise on the measurements, one may use quadratically constraint ℓ_1-minimization

$$\text{minimize}\,\|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta \ .$$

This, however, requires an accurate guess of the noise level η. Only then, and when the restricted isometry property holds, are the error bounds (1.14), (1.15) valid. (Some of the other algorithms do not require knowledge of the noise level, but in turn they require a good guess of the sparsity level s.) We will see that, somewhat unexpectedly, in the case of Gaussian measurement matrices, one can also use equality constrained ℓ_1-minimization (1.2) even when there is noise on the measurements. The ℓ_1-quotient property allows to deduce the same error bounds (1.14), (1.15) of reconstruction although the noise level may be unknown. For subgaussian random matrices a slight variation of these error bounds hold as well.

From an application point of view, subgaussian random matrices are only of limited use because they do not have any structure. Specific applications, however, may impose certain structure on the measurement matrix. As outlined above, it is presently open to come up with deterministic measurement matrices that provide provable recovery guarantees. This motivates the study of *structured random matrices*. In **Chapter 12** we investigate a particular class of such random matrices arising from sampling problems. This includes random partial Fourier matrices.

Let ψ_j, $j \in [N] := \{1, 2, \ldots, N\}$, be a system of complex-valued functions, which are orthonormal with respect to some probability measure ν on some set $\mathcal{D}$,

$$\int_{\mathcal{D}} \psi_j(t)\overline{\psi_k(t)}d\nu(t) = \delta_{jk} \ .$$

We call $\{\psi_j\}_{j\in[N]}$ a *bounded orthonormal system* if there exists a constant $K \geq 1$ (ideally independent of N) such that

$$\sup_{j\in[N]} \sup_{t\in\mathcal{D}} |\psi_j(t)| \leq K \ .$$

A particular example is the trigonometric system, $\psi_j(t) = e^{2\pi ijt}$, $j \in \Gamma \subset \mathbb{Z}$, $\mathrm{card}(\Gamma) = N$, where $K = 1$. We consider functions that are expanded in this function system,

$$f(t) = \sum_{j=1}^{N} x_j \psi_j(t) \; .$$

We call f sparse if the coefficient sequence $\mathbf{x} \in \mathbb{C}^N$ is sparse. The task is to reconstruct f (or equivalently the coefficient vector $\mathbf{x}$) from sample values at locations $t_1, \ldots, t_m$,

$$y_k = f(t_k) = \sum_{j=1}^{N} x_j \psi_j(t_k) \; .$$

Introducing the *sampling matrix* $\mathbf{A} \in \mathbb{C}^{m \times N}$ with entries

$$A_{jk} = \psi_j(t_k) \; , \tag{1.19}$$

the vector of samples is given by $\mathbf{y} = \mathbf{A}\mathbf{x}$ and we are back to the compressive sensing problem with a matrix $\mathbf{A}$ of this particular form. Randomness enters by choosing the sampling locations t_ℓ, $\ell \in [m]$, independently at random according to the probability measure ν. In this way, $\mathbf{A}$ is a structured random matrix. Before we study its performance in Chapter 12, we relate this sampling setup with discrete uncertainty principle and also provide performance limits. In the context of the Hadamard transform, we show that, in slight contrast to (1.18), at least $m \geq Cs \ln N$ measurements are necessary.

Deriving recovery guarantees for the random sampling matrix $\mathbf{A}$ in (1.19) is more involved than for subgaussian random matrices, where all the entries are independent. In fact, $\mathbf{A}$ has mN entries, but is generated only by m independent random variables. Therefore, we proceed by increasing level of difficulty and start by showing nonuniform sparse recovery guarantees for ℓ_1-minimization. The required number of samples is $m \geq CK^2 s \ln N$ to recover a fixed s-sparse coefficient vector $\mathbf{x}$ with high probability.

The bound of the restricted isometry constants of the random sampling matrix $\mathbf{A}$ in (1.19) is a highlight of the theory of compressive sensing. It states that $\delta_s \leq \delta$ with high probability provided

$$m \geq CK^2 s \ln^3(s) \ln N \; .$$

We close Chapter 12 by illustrating connections to the Λ_1-problem from harmonic analysis.

In **Chapter 13** we follow a slightly different approach to sparse recovery guarantees by considering a fixed (deterministic) matrix $\mathbf{A}$, and rather choose the s-sparse signal $\mathbf{x}$ at random. More precisely, we choose its support set S uniformly at random among all subsets of $[N]$ of cardinality s. The signs of the nonzero coefficients of $\mathbf{x}$ are chosen at random as well, but the magnitudes are

kept arbitrary. Under a very mild condition on the coherence μ of $\mathbf{A} \in \mathbb{C}^{m \times N}$, namely

$$\mu \leq \frac{c}{\ln N} \,, \tag{1.20}$$

and if

$$\frac{s\|\mathbf{A}\|_{2\to 2}}{N} \leq \frac{c}{\ln N} \,, \tag{1.21}$$

then the vector $\mathbf{x}$ is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization with high probability. The (deterministic or random) matrices $\mathbf{A}$ usually encountered in compressive sensing and signal processing, for instance tight frames, satisfy (1.21) provided that

$$m \geq Cs \ln N \,. \tag{1.22}$$

Since (1.20) is satisfied for basically any (deterministic or random) matrix $\mathbf{A}$ that one would reasonably come up in compressive sensing or sparse approximation, we again obtain sparse recovery in the familiar parameter regime (1.22). The crucial ingredient in this analysis is to show that a random column submatrix of $\mathbf{A}$ is well-conditioned under (1.20) and (1.21). We note, however, that this random signal model may not always reflect well the type of signals encountered in practice, so that the theory for random matrices remains very important. Nevertheless, the result for random signals explains well the outcome of numerical experiments where the signals are usually constructed at random.

A further type of measurement matrix for compressive sensing is studied in **Chapter 14**. It arises as adjacency matrix of a certain type of bipartite graphs, called lossless expanders, and therefore, its entries of $\mathbf{A}$ take only the value 0 and 1. The existence of lossless expanders with optimal parameters is shown via probabilistic (combinatorial) arguments. (Again, no explicit construction of an optimal measurement matrix of this form is known.) We show that the $m \times N$ adjacency matrix of such a lossless expander allows for uniform recovery of all s-sparse vectors using ℓ_1-minimization provided that

$$m \geq Cs \ln(N/s) \,.$$

Moreover, we give two iterative reconstruction algorithms. One of them has the remarkable feature that its runtime is *sublinear* in the signal length N, more precisely, its execution requires $\mathcal{O}(s^2 \log^3 N)$ operations. Since one has to report back only the locations and the values of the s nonzero entries of the reconstruction, such super-fast algorithms are no impossibility. Indeed, also in other contexts sublinear algorithms are possible, but they are always designed together with the measurement matrix $\mathbf{A}$.

The ℓ_1-minimization principle (basis pursuit) is one of the most powerful sparse recovery methods — as should have become clear by now. **Chapter 15** presents efficient algorithms to perform this optimization task in practice. The homotopy method applies to the real-valued case $\mathbf{A} \in \mathbb{R}^{m \times N}$, $\mathbf{y} \in \mathbb{R}^m$. For a parameter λ, we consider the functional

$$F_\lambda(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

Its minimizer $\mathbf{x}_\lambda$ converges to the minimizer $\mathbf{x}$ of the equality constrained ℓ_1-minimization problem (1.2). The map $\lambda \mapsto \mathbf{x}_\lambda$ turns out to be piecewise linear. The homotopy method starts with sufficiently large λ, for which $\mathbf{x}_\lambda = 0$, and traces the end points of the linear pieces until $\lambda = 0$ and the solution $\mathbf{x}$ of (1.2). At each step of the algorithm an element is added or removed from the support set of the current minimizer. Since one mostly adds elements to the support, this algorithm is usually very efficient for small sparsity.

As a second method, we treat Chambolle and Pock's primal dual algorithm, which applies to whole class of optimization problems including ℓ_1-minimization. It consists of a simple iteration procedure, which updates a primal, a dual and an auxiliary variable at each step. All of the computations are easy to perform. We show convergence of the sequence of primal variables generated by this algorithm to the minimizer of the given functional, provide an estimate of the convergence rate in terms of a specific primal-dual gap, and outline its specific form for three types of ℓ_1-minimization problems. In contrast to the homotopy method, it applies also in the complex-valued case.

Finally, we discuss a method based on iteratively solving weighted ℓ_2-minimization problems, where the weight is suitably updated in each step based on the solution in the previous iteration. Since weighted ℓ_2-minimization can be performed efficiently (in fact, this is a linear problem), each step of the algorithm can be computed quickly. Although this algorithm is strongly motivated by ℓ_1-minimization, it is not guaranteed that it always converges to the ℓ_1-minimizer. Nevertheless, under the null space property of the matrix $\mathbf{A}$ (equivalent to sparse recovery via ℓ_1-minimization), we show that also iteratively reweighted least-squares recovers every s-sparse vector from $\mathbf{y} = \mathbf{A}\mathbf{x}$. Recovery is stable under passing to compressible vectors. Moreover, we give an estimate of the convergence rate in the exactly sparse case.

The book is concluded with three appendices. **Appendix A** covers background material from linear algebra and matrix analysis including vector and matrix norms, eigenvalues and singular values and matrix functions. Basic concepts and results from convex analysis and convex optimization are presented in **Appendix B**. We also treat matrix convexity and present a proof of Lieb's theorem on the concavity of the matrix function $\mathbf{X} \mapsto \operatorname{tr} \exp(\mathbf{H} + \ln \mathbf{X})$ on the set of positive self-adjoint matrices. **Appendix C** presents miscalleneous material including covering numbers, Fourier transforms, elementary estimates on binomial coefficients, the Gamma-function and Stirling's formula, smoothing of Lipschitz functions via convolution, distributional derivatives and differential inequalities.

Notation is usually introduced when it appears first. Additionally, a collection of used symbols can be found on pp. 543. All the constants in this book are universal unless stated otherwise. This means that they do not depend on

any other quantity. Often, a value can be deduced from the proofs, or is even stated explicitly.

Notes

The field of compressive sensing was initiated with the papers [72] by E. Candès, J. Romberg and T. Tao and [130] by D. Donoho who coined the term *compressed sensing*. Although there have been predecessors on various aspects of the field, these papers seem to be the first to combine the ideas of ℓ_1-minimization and sparse recovery with the concept of choosing the measurement matrix at random, and to realize the effectiveness for solving underdetermined systems of equations. Also, they emphasized the importance for many signal processing tasks.

We give a nonexhaustive list of some of the highlights of the predecessors and earlier developments that are connected to compressive sensing next. Details and references on the development of compressive sensing itself will be given in the Notes of the following chapters.

Arguably the first contribution that can be connected with sparse recovery was made by de Prony [347] already in 1795. He developed a method of identifying the frequencies $\omega_j \in \mathbb{R}$ and the amplitudes $x_j \in \mathbb{C}$ in a non-harmonic trigonometric sum of the form $f(t) = \sum_{j=1}^{s} x_j e^{2\pi i \omega_j t}$. His method takes equidistant samples and solves an eigenvalue problem to compute the ω_j. This method is related to Reed-Solomon decoding covered in the next section, see Theorem 2.15. For more information on the Prony method we refer to [296, 346].

The use of ℓ_1-minimization appeared in 1965 in the Ph.D. thesis of B. Logan [287] in the context of sparse frequency estimation, and an early theoretical work on L_1-minimization is the paper [127] by D. Donoho and B. Logan. Geophysicists observed in the late 1970s that ℓ_1-minimization can be successfully used to compute a sparse reflection function indicating changes between subsurface layers [407, 381]. The use of total-variation minimization, which is closely connected to ℓ_1-minimization, appeared in the 1990's in the work on image processing by L. Rudin, S. Osher and E. Fatemi [377]. The use of ℓ_1-minimization and related greedy methods in statistics was greatly popularized with the work of R. Tibshirani [411] on the LASSO.

The theory of sparse approximation and the associated algorithms began in the 1990s with the papers [294, 310, 94]. The theoretical understanding under which conditions greedy methods and ℓ_1-minimization recover the sparsest solution developed with the work in [136, 158, 134, 206, 194, 187, 414, 417].

Compressive sensing is connected with the area of information-based complexity which considers the general question of how well a function f from a class $\mathcal{F}$ can be recovered from m sample values, or more generally, from the evaluation of m linear or non-linear functionals applied to f [412]. The optimal

recovery error defined as the maximal reconstruction error for the "best" sampling method and "best" recovery method (within a specified class of methods) over all functions in the class $\mathcal{F}$ is closely related to the so-called *Gelfand width* of $\mathcal{F}$ [319], see also Chapter 10. Of particular interest for compressive sensing is the ℓ_1-ball B_1^N in $\mathbb{R}^N$, since its elements can be well-approximated by sparse ones. Famous results due to B. Kashin [259], and E. Gluskin and A. Garnaev [189, 196] sharply bound the Gelfand widths of B_1^N from above and below, see also Chapter 10. While the original interest of Kashin was in the estimate of m-widths of Sobolev classes, these results give precise performance bounds in compressive sensing on how well any method may recover (approximately) sparse vectors from linear measurements. It is remarkable that [259, 189] already used Bernoulli and Gaussian random matrices in a similar ways as they are used in compressive sensing (Chapter 9).

In computer science as well, sparsity appeared before the advent of compressive sensing in the area of sketching. Here, one is not only interested in recovering huge data sets (such as data streams on the internet) from vastly undersampled data, but requires in addition that the associated algorithms have sublinear runtime in the signal length. There is no a-priori contradiction in this desideratum because one needs to report locations and values of nonzero entries. Such algorithms often use ideas from *group testing* [150], which dates back to World War II, when Dorfman [149] invented an efficient method for detecting draftees with syphilis. One usually designs the matrix and the fast algorithm simultaneously [106, 195] in this setup. Lossless expanders as studied in Chapter 14 play a key role in some of the constructions [32]. Quite remarkably sublinear algorithms are also available for sparse Fourier transforms [193, 455, 248, 249, 227, 226].

Applications of compressive sensing. We next provide comments and references on the described applications and motivations in Section 1.2.

Sampling. The classical sampling theorem (1.5) can be associated with the names of Shannon, Nyquist, Whittaker and Kotelnikov. Sampling theory is a broad and well-developed area. We refer to [30, 170, 233, 234, 255] for further information on the classical aspects. The use of sparse recovery techniques in sampling problems appeared very early in the development of theory of compressive sensing [72, 82, 352, 353, 355, 360]. In fact, the alternative name *compressive sampling* indicates that compressive sensing can be viewed as a part of sampling theory – although it draws from quite different mathematical tools than classical sampling theory.

Single pixel camera. The single pixel camera was developed by R. Baraniuk and coworkers [151] as a nice proof of concept that the ideas of compressive sensing can be implemented in hardware.

Magnetic resonance imaging. The initial paper [72] on compressed was motivated by medical imaging – although E. Candès et al. have treated the very similar problem of computerized tomography (CT). The application of compressive sensing techniques to magnetic resonance imaging (MRI) were in-

vestigated, for instance, in [291, 221, 434, 309]. Background on the theoretical foundations of MRI can be found, for instance, in [218, 231, 449]. Applications of compressive sensing to the related problem of *nuclear magnetic resonance spectroscopy* can be found in [240, 387].

Radar. The particular radar application outlined in Section 1.2 is described in more detail in [232]. The same mathematical model appears also in sonar and in the channel estimation problem of wireless communications [333, 356, 332]. The application of compressive sensing to other tasks in radar can be found, for instance, in [161, 164, 343, 394].

Sparse approximation. The theory compressive sensing can certainly be viewed as a part of sparse approximation with roots in signal processing, harmonic analysis [148] and numerical analysis [101]. A general source for background on sparse approximation and their applications are the books [156, 390] as well as the overview paper [60].

The principle to represent a signal by a small number of terms in a suitable basis in order to achieve compression is realized, for instance, in the ubiquitous compression standards JPEG, MPEG and MP3. Wavelets [114] are known to provide a good basis for images, and the analysis of the best (nonlinear) approximation reaches into the area of function spaces, more precisely Besov spaces [445]. Similarly, Gabor expansions [210] may compress audio signals. Since, for instance, good Gabor systems are always redundant systems (frames) and never bases, it is important to have computational tools to compute the sparsest representation of a signal. It was realized in [310, 294] that this problem is in general NP-hard. The greedy approach via orthogonal matching pursuit was then introduced [294] (although it had appeared in different contexts earlier), while basis pursuit (ℓ_1-minimization) was presented in [94].

The use of the uncertainty principle for deducing a positive statement on the data separation problem with respect to the Fourier and canonical basis appeared in [142, 141]. For further information on the separation problem we refer the interested reader to [158, 78, 136, 138, 205, 286, 420]. Background on denoising via sparse representations can be found in [157, 140, 89, 126, 137, 351].

The analysis of conditions under which algorithms such as ℓ_1-minimization or orthogonal matching pursuit can recover the sparsest representation has started with the contributions [133, 135, 136, 139, 194, 414, 417], and these early results are at the basis of the developments in compressive sensing.

Error correction. The idealized setup of error correction and the approach to it via compressive described in Section 1.2 appeared in [81, 145, 372]. For more background on error correction we refer to [244].

Statistics and machine learning. Sparsity has a long history in statistics and, in particular, in linear regression models. The corresponding area is sometimes referred to as high-dimensional statistics or model selection because the support set of the coefficient vector $\mathbf{x}$ determines the relevant explanatory variables and thereby selects a model. Stepwise forward regression methods

are closely related to greedy algorithms such as (orthogonal) matching pursuit. The LASSO, that is the ℓ_1-minimization problem (1.8), was introduced by R. Tibshirani in [411]. E. Candès and T. Tao have introduced the Dantzig selector (1.9) in [83] and realized that the methods of compressive sensing (the restricted isometry property) is very useful for the analysis of sparse regression methods. We refer to [39] and the monograph [63] for details. For more information on machine learning we direct the reader to [14, 108, 109, 385]. Connections of sparsity and machine learning can be found for instance in [19, 122, 450].

Low-rank matrix recovery. The extension of compressive sensing to the recovery of low-rank matrices from incomplete information came up with the papers [76, 84, 362]. The idea of replacing the rank minimization problem by the nuclear norm minimization appeared in the PhD thesis of M. Fazel [165]. The matrix completion problem is treated in [76, 361, 84], and the more general problem of quantum state tomography in [212, 211, 285].

Let us briefly mention further applications and relations to other fields.

In **inverse problems** and methods for their regularization, sparsity has become an important concept as well. Instead of Thikonov regularization with a Hilbert space norm [162], one uses an ℓ_1-norm regularization approach [115, 350]. In many practical applications this improves the recovered solutions. In fact, ill-posed inverse problems appear for instance in geophysics where ℓ_1-norm regularization was already used in [407, 381], however, without rigorous mathematical theory at that time.

Total variation minimization is a classical and successful approach for image denoising and other tasks in image processing [90, 377, 88]. Since the total variation is the ℓ_1-norm of the gradient, the minimization problem is closely related to basis pursuit. In fact, the motivating example for the first contribution of E. Candés, J. Romberg and T. Tao [72] to compressive sensing came from total variation minimization in computer tomography. The restricted isometry property can be used to analyze image recovery via total variation minimization [315]. The primal dual algorithm of A. Chambolle and T. Pock to be presented in Chapter 15 was originally motivated by total variation minimization as well [91].

Further applications of compressive sensing and sparsity in general include imaging (tomography, ultrasound, photoacoustic imaging, hyperspectral imaging etc.), analog-to-digital conversion [426, 304], DNA microarray processing, astronomy [444], wireless communications [23, 406] and more. The website `www.compressedsensing.com` contains a collection of articles on the various applications.

Topics that are not covered in this book. It is impossible to give a detailed account on all the directions in compressive sensing that have emerged so far. This book certainly makes a selection, but we believe that we cover the most important aspects and mathematical techniques. With this basis the reader should be well-equipped to read the original references on further

directions, generalizations and applications. Let us only give a biref account on further topics together with the relevant references. Again, we do not make any claim on completeness of this list.

Joint sparsity, block sparsity. Suppose that we take measurements not only of a single signal but of a collection of signals that are somewhat coupled. Rather than only assuming that each signal is sparse (or compressible) on its own, we assume that the unknown support set is the same for all signals in the collections. In this case, we speak of joint sparsity. A motivating example are color images where each signal corresponds to a color channel of the image, say red, green and blue. Since edges usually appear in all channels at the same location we have joint sparsity in the gradient, for instance. Instead of the usuall ℓ_1-minization problem, one considers mixed ℓ_1/ℓ_2-norm minimization or corresponding greedy algorithm which exploit the joint sparsity structure. A similar setup is described by the block sparsity model, where one groups together certain indices of the sparse vector. Then a signal is block-sparse if most groups (a block) of such coefficients are zero, and when a non-zero coefficient appears then its whole block is non-zero. Recovery algorithms may exploit this preknowledge in order to improve the recovery performance. A similar theory as in the usual sparsity context can be developed [119, 159, 160, 177, 425, 208, 416].

Sublinear algorithms have been developed in computer science for a longer time. The fact that only the locations and values of a sparse vector have to be reported enables one to design recovery algorithms whose runtime is sublinear in the vector length. Corresponding methods are also called streaming algorithms or heavy hitters. We will only cover a "toy sublinear algorithm" in Chapter 14, and refer to [32, 106, 193, 250, 246, 192, 195, 226] for more information.

Analysis of sparse recovery via basis pursuit using random polytope geometry. D. Donoho and J. Tanner [144, 143, 132, 145] approached the analysis sparse recovery via ℓ_1-minimization in connection with Gaussian random matrices through polytope geometry. In fact, recovery of s-sparse vectors via ℓ_1-minimization is equivalent to a geometric property called neighborliness of the projected ℓ_1-ball under the measurement matrix, see also Corollary 4.39. When the measurement matrix is Gaussian, Donoho and Tanner give a precise analysis of so-called phase transitions that predict in which range of (s, m) sparse recovery is successful with high probability. In particular, this analysis provides exact constants and also predicts when sparse recovery starts to fail. We only give a brief account on their work in the Notes of Chapter 9.

Compressive sensing and quantization. If compressive sensing is used for signal acquisition then a realistic sensor has to quantize the measured data. This means that only a finite number of values of the measurements y_ℓ are possible. For instance, if 8 bits are used for each y_ℓ, one has 256 possible values and only an approximation to y_ℓ can be stored. In particular, if quantization is rather course then this additional source of error cannot be ignored, and a theoretical analysis becomes necessary. We refer to [274, 456, 215] for

background information. Let us also mention that even the extreme case of 1-bit compressed sensing is possible, where only the signs of the measurements $\mathbf{y} = \operatorname{sgn}(\mathbf{Ax})$ are taken [251, 340, 341].

Dictionary learning. Sparsity is usually with respect to a suitable basis or redundant dictionary. In practice, it is not always clear which dictionary is good in order to sparsify the signals encountered in a certain application. Dictionary learning tries to identify a good dictionary from training signals. Algorithmic approaches include the K-SVD algorithm [4, 370] and optimization approaches [383]. Since one optimizes over both the dictionary and the coefficients in the expansions one ends up with a nonconvex program even when using ℓ_1-minimization. Therefore, it is notoriously hard to establish a rigorous mathematical theory of dictionary learning although the algorithms work well in practice. Nevertheless, there are a few interesting mathematical results available which are in the spirit of compressive sensing [191, 383].

Hints for preparing a course. Some university teachers may use this book as the basis for a course on compressive sensing. The material in this book probably exceeds the amount that one is able to cover in a reasonable one-semester course, say. Therefore, we give some hints on a possible selection of topics from this book. TO BE COMPLETED.

2

Sparse Solutions of Underdetermined Systems

In this chapter, we define the notions of vector sparsity and compressibility and we establish some related inequalities used throughout the book. We will use basic results on vector and matrix norms, which can be found in Appendix A. We then investigate, in two different settings, the minimal number of linear measurements required to recover sparse vectors. We finally prove that the ideal recovery scheme ℓ_0-minimization is NP-hard in general.

2.1 Sparsity and Compressibility

We start by defining the ideal notion of *sparsity*. Let us before introduce $[N] := \{1, 2, \ldots, N\}$ and $\mathrm{card}(S)$ denoting the cardinality of a set S.

Definition 2.1. *The* support *of a vector $\mathbf{x} \in \mathbb{C}^N$ is the index set of its nonzero entries, i.e.,*

$$\mathrm{supp}(\mathbf{x}) := \{j \in [N] : x_j \neq 0\}.$$

The vector $\mathbf{x} \in \mathbb{C}^N$ is called s-sparse *if at most s of its entries are nonzero, i.e., if*

$$\|\mathbf{x}\|_0 := \mathrm{card}(\mathrm{supp}(\mathbf{x})) \leq s.$$

The customary notation $\|\mathbf{x}\|_0$ — the notation $\|\mathbf{x}\|_0^0$ would in fact be more appropriate — comes from the observation that

$$\|\mathbf{x}\|_p^p := \sum_{j=1}^N |x_j|^p \underset{p \to 0}{\longrightarrow} \sum_{j=1}^N \mathbf{1}_{\{x_j \neq 0\}} = \mathrm{card}(\{j \in [N] : x_j \neq 0\}).$$

Here, we used the notation $\mathbf{1}_{\{x_j \neq 0\}} = 1$ if $x_j \neq 0$, and $\mathbf{1}_{\{x_j \neq 0\}} = 0$ if $x_j = 0$. In other words the quantity $\|\mathbf{x}\|_0$ is the limit as p decreases to zero of the pth power of the ℓ_p-quasinorm of $\mathbf{x}$. It is abusively called the ℓ_0-norm of $\mathbf{x}$, although it is neither a norm nor a quasinorm — see Appendix A for precise definitions of these notions. In practice, sparsity can be a strong constraint to

impose, and we may prefer the weaker concept of *compressibility*. For instance, we may consider vectors that are nearly s-sparse, as measured by the *error of best s-term approximation*. *best s-term approximation*

Definition 2.2. *For $p > 0$, the* ℓ_p-error of best s-term approximation *to a vector $\mathbf{x} \in \mathbb{C}^N$ is defined by*

$$\sigma_s(\mathbf{x})_p := \inf\big\{\|\mathbf{x} - \mathbf{z}\|_p,\ \mathbf{z} \in \mathbb{C}^N \text{ is } s\text{-sparse}\big\}.$$

In the definition of $\sigma_s(\mathbf{x})_p$, the infimum is achieved by an s-sparse vector $\mathbf{z} \in \mathbb{C}^N$ whose nonzero entries equal the s largest absolute entries of $\mathbf{x}$. Hence, although such a vector $\mathbf{z} \in \mathbb{C}^N$ may not be unique, it achieves the infimum independently of $p > 0$.

Informally, we may call $\mathbf{x} \in \mathbb{C}^N$ a *compressible* vector if the error of its best s-term approximation decays quickly in s. According to the following proposition, this happens in particular if $\mathbf{x}$ belongs to the unit ℓ_p-ball for some small $p > 0$, where the unit ℓ_p-ball is defined by

$$B_p^N := \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_p \le 1\}.$$

Consequently, the nonconvex balls B_p^N for $p < 1$ serve as good models for compressible vectors.

Proposition 2.3. *For any $q > p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$,*

$$\sigma_s(\mathbf{x})_q \le \frac{1}{s^{1/p-1/q}} \|\mathbf{x}\|_p.$$

Before proving this proposition, it is useful to introduce the notion of *nonincreasing rearrangement*.

Definition 2.4. *The* nonincreasing rearrangement *of the vector $\mathbf{x} \in \mathbb{C}^N$ is the vector $\mathbf{x}^* \in \mathbb{R}^N$ for which*

$$x_1^* \ge x_2^* \ge \ldots \ge x_N^* \ge 0$$

and there is a permutation $\pi : [N] \to [N]$ with $x_j^* = |x_{\pi(j)}|$ for all $j \in [N]$.

Proof (of Proposition 2.3). If $\mathbf{x}^* \in \mathbb{R}_+^N$ is the nonincreasing rearrangement of $\mathbf{x} \in \mathbb{C}^N$, we have

$$\begin{aligned}\sigma_s(\mathbf{x})_q^q = \sum_{j=s+1}^N (x_j^*)^q &\le (x_s^*)^{q-p} \sum_{j=s+1}^N (x_j^*)^p \le \Big(\frac{1}{s}\sum_{j=1}^s (x_j^*)^p\Big)^{\frac{q-p}{p}} \Big(\sum_{j=s+1}^N (x_j^*)^p\Big) \\ &\le \Big(\frac{1}{s}\,\|\mathbf{x}\|_p^p\Big)^{\frac{q-p}{p}} \|\mathbf{x}\|_p^p = \frac{1}{s^{q/p-1}}\|\mathbf{x}\|_p^q.\end{aligned}$$

The result follows by taking the power $1/q$ in both sides of this inequality. □

We strengthen the previous proposition by finding the smallest possible constant $c_{p,q}$ in the inequality $\sigma_s(\mathbf{x})_q \le c_{p,q} s^{-1/p+1/q} \|\mathbf{x}\|_p$. The proof consists in solving a convex optimization problem by hand.

Theorem 2.5. *For any $q > p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$, the inequality*

$$\sigma_s(\mathbf{x})_q \le \frac{c_{p,q}}{s^{1/p-1/q}} \|\mathbf{x}\|_p$$

holds with

$$c_{p,q} := \Big[\Big(\frac{p}{q}\Big)^{p/q} \Big(1 - \frac{p}{q}\Big)^{1-p/q} \Big]^{1/p} \le 1.$$

Let us point out that the frequent choice $p = 1$ and $q = 2$ gives

$$\sigma_s(\mathbf{x})_2 \le \frac{1}{2\sqrt{s}} \|\mathbf{x}\|_1$$

Proof. Let $\mathbf{x}^* \in \mathbb{R}^N_+$ be the nonincreasing rearrangement of $\mathbf{x} \in \mathbb{C}^N$. Setting $\alpha_j := (x_j^*)^p$, we will prove the equivalent statement

$$\left.\begin{array}{l} \alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_N \ge 0 \\ \alpha_1 + \alpha_2 + \cdots + \alpha_N \le 1 \end{array}\right\} \Longrightarrow \alpha_{s+1}^{q/p} + \alpha_{s+2}^{q/p} + \cdots + \alpha_N^{q/p} \le \frac{c_{p,q}^q}{s^{q/p-1}}.$$

Thus, with $r := q/p > 1$, we aim at maximizing the convex function

$$f(\alpha_1, \alpha_2, \ldots, \alpha_N) := \alpha_{s+1}^r + \alpha_{s+2}^r + \cdots + \alpha_N^r$$

over the convex polygon

$$\mathcal{C} := \{(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N : \alpha_1 \ge \cdots \ge \alpha_N \ge 0 \text{ and } \alpha_1 + \cdots + \alpha_N \le 1\}.$$

According to Theorem B.16, the maximum of f is attained at a vertex of $\mathcal{C}$. The vertices of $\mathcal{C}$ are obtained as intersections of N hyperplanes arising by turning N of the $(N+1)$ inequality constraints into equalities. Thus, we have the following possibilities:

- if $\alpha_1 = \cdots = \alpha_N = 0$, then $f(\alpha_1, \alpha_2, \ldots, \alpha_N) = 0$;
- if $\alpha_1 + \cdots + \alpha_N = 1$ and $\alpha_1 = \cdots = \alpha_k > \alpha_{k+1} = \cdots = \alpha_N = 0$ for some $1 \le k \le s$, then $f(\alpha_1, \alpha_2, \ldots, \alpha_N) = 0$;
- if $\alpha_1 + \cdots + \alpha_N = 1$ and $\alpha_1 = \cdots = \alpha_k > \alpha_{k+1} = \cdots = \alpha_N = 0$ for some $s+1 \le k \le N$, then $\alpha_1 = \cdots = \alpha_k = 1/k$, and consequently $f(\alpha_1, \alpha_2, \ldots, \alpha_N) = (k-s)/k^r$.

It follows that

$$\max_{(\alpha_1,\ldots,\alpha_N) \in \mathcal{C}} f(\alpha_1, \alpha_2, \ldots, \alpha_N) = \max_{s+1 \le k \le N} \frac{k-s}{k^r}.$$

Considering k as a continuous variable, we now observe that the function $g(k) := (k-s)/k^r$ is increasing until the critical point $k^* = (r/(r-1))s$ and decreasing thereafter. We obtain

$$\max_{(\alpha_1,\dots,\alpha_N)\in\mathcal{C}} f(\alpha_1,\alpha_2,\dots,\alpha_N) \le g(k^*) = \frac{1}{r}\Big(1-\frac{1}{r}\Big)^{r-1}\frac{1}{s^{r-1}} = c_{p,q}^q \frac{1}{s^{q/p-1}}.$$

This is the desired result. □

Another possibility to define *compressibility* is to call a vector $\mathbf{x}\in\mathbb{C}^N$ *compressible* if the number

$$\mathrm{card}(\{j\in[N] : |x_j|\ge t\})$$

of its significant — rather than nonzero — components is small. This naturally leads to the introduction of weak ℓ_p-spaces.

Definition 2.6. *For $p>0$, the weak ℓ_p space $w\ell_p^N$ denotes the space $\mathbb{C}^N$ equipped with the quasinorm*

$$\|\mathbf{x}\|_{p,\infty} := \inf\Big\{M\ge 0 : \mathrm{card}(\{j\in[N] : |x_j|\ge t\}) \le \frac{M^p}{t^p} \text{ for all } t>0\Big\}.$$

To verify that the previous quantity indeed defines a quasinorm, we check, for any $\mathbf{x},\mathbf{y}\in\mathbb{C}^N$ and any $\lambda\in\mathbb{C}$, that $\|\mathbf{x}\|=0\Rightarrow\mathbf{x}=0$, $\|\lambda\mathbf{x}\|=|\lambda|\|\mathbf{x}\|$, and $\|\mathbf{x}+\mathbf{y}\|_{p,\infty}\le 2^{\max\{1,1/p\}}\big(\|\mathbf{x}\|_{p,\infty}+\|\mathbf{y}\|_{p,\infty}\big)$. The first two properties are easy, while the third property is a consequence of the more general statement below.

Proposition 2.7. *Let $\mathbf{x}^1,\dots,\mathbf{x}^k\in\mathbb{C}^N$. Then, for $p>0$,*

$$\|\mathbf{x}^1+\cdots+\mathbf{x}^k\|_{p,\infty} \le k^{\max\{1,1/p\}}\big(\|\mathbf{x}^1\|_{p,\infty}+\cdots+\|\mathbf{x}^k\|_{p,\infty}\big)\,.$$

Proof. Let $t>0$. If $|x_j^1+\cdots+x_j^k|\ge t$ for some $j\in[N]$, then we have $|x_j^i|\ge t/k$ for some $i\in[k]$. This means that

$$\{j\in[N] : |x_j^1+\cdots+x_j^k|\ge t\} \subset \bigcup_{i\in[k]}\{j\in[N] : |x_j^i|\ge t/k\}\,.$$

We derive

$$\begin{aligned}\mathrm{card}(\{j\in[N] : |x_j^1+\cdots+x_j^k|\ge t\}) &\le \sum_{i\in[k]}\frac{\|\mathbf{x}^i\|_{p,\infty}^p}{(t/k)^p}\\ &= \frac{k^p\big(\|\mathbf{x}^1\|_{p,\infty}^p+\cdots+\|\mathbf{x}^k\|_{p,\infty}^p\big)}{t^p}\,.\end{aligned}$$

According to the definition of the weak ℓ_p-quasinorm of $\mathbf{x}^1+\cdots+\mathbf{x}^k$, we obtain

$$\|\mathbf{x}^1+\cdots+\mathbf{x}^k\|_{p,\infty} \le k\big(\|\mathbf{x}^1\|_{p,\infty}^p+\cdots+\|\mathbf{x}^k\|_{p,\infty}^p\big)^{1/p}\,.$$

Now, if $p\le 1$, comparing the ℓ_p and ℓ_1 norms in $\mathbb{R}^k$ gives

$$\big(\|\mathbf{x}^1\|_{p,\infty}^p+\cdots+\|\mathbf{x}^k\|_{p,\infty}^p\big)^{1/p} \le k^{1/p-1}(\|\mathbf{x}^1\|_{p,\infty}+\cdots+\|\mathbf{x}^k\|_{p,\infty})\,,$$

and if $p \geq 1$, comparing the ℓ_p and ℓ_1 norms in $\mathbb{R}^k$ gives

$$\left(\|\mathbf{x}^1\|_{p,\infty}^p + \cdots + \|\mathbf{x}^k\|_{p,\infty}^p\right)^{1/p} \leq \|\mathbf{x}^1\|_{p,\infty} + \cdots + \|\mathbf{x}^k\|_{p,\infty} \, .$$

The result immediately follows. □

Remark 2.8. The constant $k^{\max\{1,1/p\}}$ in Proposition 2.7 is sharp, see Exercise (2.2).

It is sometimes preferable to invoke the following alternative expression for the weak ℓ_p-quasinorm of a vector $\mathbf{x} \in \mathbb{C}^N$.

Proposition 2.9. *For $p > 0$, the weak ℓ_p-quasinorm of a vector $\mathbf{x} \in \mathbb{C}^N$ can be expressed as*

$$\|\mathbf{x}\|_{p,\infty} = \max_{k \in [N]} k^{1/p} x_k^* \, ,$$

where $\mathbf{x}^* \in \mathbb{R}_+^N$ denotes the nonincreasing rearrangement of $\mathbf{x} \in \mathbb{C}^N$.

Proof. Given $\mathbf{x} \in \mathbb{C}^N$, in view of $\|\mathbf{x}\|_{p,\infty} = \|\mathbf{x}^*\|_{p,\infty}$, we need to establish that $\|\mathbf{x}\| := \max_{k \in [N]} k^{1/p} x_k^*$ equals $\|\mathbf{x}^*\|_{p,\infty}$. For $t > 0$, we first note that either $\{j \in [N] : x_j^* \geq t\} = [k]$ for some $k \in [N]$ or $\{j \in [N] : x_j^* \geq t\} = \emptyset$. In the former case, $t \leq x_k^* \leq \|\mathbf{x}\|/k^{1/p}$, and hence, $\mathrm{card}(\{j \in [N] : x_j^* \geq t\}) = k \leq \|\mathbf{x}\|^p/t^p$. This inequality holds trivially in the case that $\{j \in [N] : x_j^* \geq t\} = \emptyset$. According to the definition of the weak ℓ_p-quasinorm, we obtain $\|\mathbf{x}^*\|_{p,\infty} \leq \|\mathbf{x}\|$. Let us now suppose that $\|\mathbf{x}\| > \|\mathbf{x}^*\|_{p,\infty}$, so that $\|\mathbf{x}\| \geq (1+\epsilon)\|\mathbf{x}^*\|_{p,\infty}$ for some $\varepsilon > 0$. This means that $k^{1/p} x_k^* \geq (1+\epsilon)\|\mathbf{x}^*\|_{p,\infty}$ for some $k \in [N]$. Therefore, the set

$$\{j \in [N] : x_j^* \geq (1+\epsilon)\|\mathbf{x}^*\|_{p,\infty}/k^{1/p}\}$$

contains the set $[k]$. The definition of the weak ℓ_p-quasinorm yields

$$k \leq \frac{\|\mathbf{x}^*\|_{p,\infty}^p}{\left((1+\epsilon)\|\mathbf{x}^*\|_{p,\infty}/k^{1/p}\right)^p} = \frac{k}{(1+\epsilon)^p} \, ,$$

which is a contradiction. We conclude that $\|\mathbf{x}\| = \|\mathbf{x}^*\|_{p,\infty}$. □

This alternative expression of the weak ℓ_p-quasinorm provides a sligthly easier way to compare it to the ℓ_p-(quasi)norm, as follows.

Proposition 2.10. *For any $p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$,*

$$\|\mathbf{x}\|_{p,\infty} \leq \|\mathbf{x}\|_p \, .$$

Proof. For $k \in [N]$, we write

$$\|\mathbf{x}\|_p^p = \sum_{j=1}^N (x_j^*)^p \geq \sum_{j=1}^k (x_j^*)^p \geq k(x_k^*)^p \, .$$

Raising to the power $1/p$ and taking the maximum over k gives the result. □

The alternative expression of the weak ℓ_p-quasinorm also enables us to easily establish a variation of Proposition 2.3 where weak ℓ_p replaces ℓ_p.

Proposition 2.11. *For any $q > p > 0$ and $\mathbf{x} \in \mathbb{C}^N$, the inequality*

$$\sigma_s(\mathbf{x})_q \le \frac{d_{p,q}}{s^{1/p-1/q}} \|\mathbf{x}\|_{p,\infty}$$

holds with

$$d_{p,q} := \Big(\frac{p}{q-p}\Big)^{1/q} .$$

Proof. We may assume without loss of generality that $\|\mathbf{x}\|_{p,\infty} \le 1$, so that $x_k^* \le 1/k^{1/p}$ for all $k \in [N]$. We then have

$$\begin{aligned} \sigma_s(\mathbf{x})_q^q &= \sum_{k=s+1}^{N} (x_k^*)^q \le \sum_{k=s+1}^{N} \frac{1}{k^{q/p}} \le \int_s^N \frac{1}{t^{q/p}} dt = -\frac{1}{q/p-1} \frac{1}{t^{q/p-1}} \Big|_{t=s}^{t=N} \\ &\le \frac{p}{q-p} \frac{1}{s^{q/p-1}} . \end{aligned}$$

Taking the power $1/q$ yields the desired result. □

Proposition 2.11 shows that vectors $\mathbf{x} \in \mathbb{C}^N$ which are compressible in the sense that $\|\mathbf{x}\|_{p,\infty} \le 1$ for small $p > 0$ are also compressible in the sense that their errors of best s-term approximation decay quickly with s.

We close this section with a technical result on the nonincreasing rearrangement.

Lemma 2.12. *The nonincreasing rearrangement satisfies, for $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$,*

$$\|\mathbf{x}^* - \mathbf{z}^*\|_\infty \le \|\mathbf{x} - \mathbf{z}\|_\infty . \tag{2.1}$$

Moreover, for $s \in [N]$,

$$|\sigma_s(\mathbf{x})_1 - \sigma_s(\mathbf{z})| \le \|\mathbf{x} - \mathbf{z}\|_1 , \tag{2.2}$$

and for $k > s$,

$$(k-s)\mathbf{x}_k^* \le \|\mathbf{x} - \mathbf{z}\|_1 + \sigma_s(\mathbf{z})_1 . \tag{2.3}$$

Proof. For $j \in [N]$ let S be the index set corresponding to the j largest absolute entries of $\mathbf{z}$. Then the nonincreasing rearrangements $\mathbf{x}^*, \mathbf{z}^*$ satisfy

$$x_j^* \le \max_{\ell \in \overline{S}} |x_\ell| \le \max_{\ell \in \overline{S}} |z_\ell| + \|\mathbf{x} - \mathbf{z}\|_\infty = z_j^* + \|\mathbf{x} - \mathbf{z}\|_\infty .$$

Reversing the roles of $\mathbf{x}$ and $\mathbf{z}$ shows (2.1).

Next, let $\mathbf{v} \in \mathbb{C}^N$ be a best s-term approximation to $\mathbf{z}$. Then

$$\sigma_s(\mathbf{x})_1 \le \|\mathbf{x} - \mathbf{v}\|_1 \le \|\mathbf{x} - \mathbf{z}\|_1 + \|\mathbf{z} - \mathbf{v}\|_1 = \|\mathbf{x} - \mathbf{z}\|_1 + \sigma_s(\mathbf{z})_1 ,$$

and again by symmetry this establishes (2.2). Also (2.3) follows from this estimate by noting that

$$(k-s)x_k^* \le \sum_{j=s+1}^{k} x_j^* \le \sum_{j \ge s+1} x_j^* = \sigma_s(\mathbf{x})_1 \, .$$

This completes the proof. □

2.2 Minimal Number of Measurements

The compressive sensing problem consists in reconstructing an s-sparse vector $\mathbf{x}$ from

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{A} \in \mathbb{C}^{m \times N}$ is the so-called measurement matrix, and $m < N$. Then the above system of linear equations is underdetermined, but the hope is that the sparsity assumption helps in identifying the original sparse $\mathbf{x}$.

In this section, we examine the question on the minimal number of linear measurements needed to reconstruct s-sparse vectors from these measurements, regardless of the practicality of the reconstruction scheme. This question can in fact take two meanings, depending on whether we require that the measurement scheme allows the reconstruction of all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$ simultaneously, or whether we require that, given an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$, the measurement scheme allows the reconstruction of this specific vector. While the second scenario seems to be unnatural at first sight because the vector $\mathbf{x}$ is unknown a priori, it will become important later when aiming at recovery guarantees when the matrix $\mathbf{A}$ is chosen at random and the sparse vector $\mathbf{x}$ is fixed (so called nonuniform recovery guarantees).

The minimal number m of measurements depends on the setting considered, namely it equals $2s$ in the first case and $s+1$ in the second case. However, we will see in Chapter 11 that if we also require the reconstruction scheme to be stable (the meaning will be made precise later), then the minimal number of required measurements additionally involves a factor of $\ln(N/s)$, so that recovery will never be stable with only $2s$ measurements.

Before separating the two settings discussed above, it is worth pointing out the equivalence of the following properties.

(a) the s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s-sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{y}$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$, that is, $\{\mathbf{z} \in \mathbb{C}^N : \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}, \|\mathbf{z}\|_0 \le s\} = \{\mathbf{x}\}$,
(b) the s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ can be reconstructed as the unique solution of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_0}$$

Indeed, if an s-sparse $\mathbf{x} \in \mathbb{C}^N$ is the unique s-sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{y}$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$, then a solution $\mathbf{x}^\sharp$ of (P_0) is s-sparse and satisfies $\mathbf{A}\mathbf{x}^\sharp = \mathbf{y}$, so that $\mathbf{x}^\sharp = \mathbf{x}$. This shows $(a) \Rightarrow (b)$. The implication $(b) \Rightarrow (a)$ is clear.

Recovery of all sparse vectors

Before stating the main result for this case, we observe that the uniqueness of sparse solutions of underdetermined linear systems can be reformulated in several ways. For a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ and a subset $S \subset [N]$, we use the notation $\mathbf{A}_S$ to indicate the column submatrix of $\mathbf{A}$ consisting of the columns indexed by S. Similarly, for $\mathbf{x} \in \mathbb{C}^N$ we denote by $\mathbf{x}_S$ either the sub-vector in $\mathbb{C}^S$ consisting of the entries indexed by S, that is, $(\mathbf{x}_S)_\ell = v_\ell$ for $\ell \in S$, or the vector in $\mathbb{C}^N$ which coincides with $\mathbf{x}$ on the entries in S and is zero on the entries outside S, that is,

$$(\mathbf{x}_S)_\ell = \begin{cases} x_\ell & \text{if } \ell \in S\,, \\ 0 & \text{if } \ell \notin S\,. \end{cases} \tag{2.4}$$

It should always become clear from the context, which of the two options apply.

Theorem 2.13. *Given $\mathbf{A} \in \mathbb{C}^{m\times N}$, the following properties are equivalent.*

(a) Every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s-sparse solution of $\mathbf{Az} = \mathbf{Ax}$, that is, if $\mathbf{Ax} = \mathbf{Az}$ and both $\mathbf{x}$ and $\mathbf{z}$ are s-sparse then $\mathbf{x} = \mathbf{z}$.
(b) The null space $\ker \mathbf{A}$ does not contain an $2s$-sparse vector other than the zero vector, that is, $\ker \mathbf{A} \cap \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_0 \le 2s\} = \{\mathbf{0}\}$.
(c) For every $S \subset [N]$ with $\operatorname{card}(S) \le 2s$, the submatrix $\mathbf{A}_S$ is injective.
(d) Every set of $2s$ columns of $\mathbf{A}$ is linearly independent.

Proof. (a)⇔(b) Let $\mathbf{x}$ and $\mathbf{z}$ be s-sparse with $\mathbf{Ax} = \mathbf{Az}$. Then $\mathbf{x} - \mathbf{z}$ is $2s$-sparse and $\mathbf{A}(\mathbf{x} - \mathbf{z}) = \mathbf{0}$. If the kernel does not contain any $2s$-sparse vector different from the zero vector then $\mathbf{x} = \mathbf{z}$.

Conversely, assume that for every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ we have $\{\mathbf{z} \in \mathbb{C}^N : \mathbf{Az} = \mathbf{Ax}, \|\mathbf{z}\|_0 \le s\} = \{\mathbf{x}\}$. Let $\mathbf{v} \in \ker \mathbf{A}$ be $2s$-sparse. We can write $\mathbf{v} = \mathbf{x} - \mathbf{z}$ for s-sparse vectors $\mathbf{x}, \mathbf{z}$ with $\operatorname{supp}\mathbf{x} \cap \operatorname{supp}\mathbf{z} = \emptyset$. Then $\mathbf{Ax} = \mathbf{Az}$, and by assumption $\mathbf{x} = \mathbf{z}$. Since the supports of $\mathbf{x}$ and $\mathbf{z}$ are disjoint it follows that $\mathbf{x} = \mathbf{z} = \mathbf{0}$ and $\mathbf{v} = \mathbf{0}$.

For the equivalence of (b), (c) and (d) we observe that for a $2s$-sparse vector $\mathbf{v}$ with $S = \operatorname{supp}\mathbf{v}$ we have $\mathbf{Av} = \mathbf{A}_S\mathbf{v}_S$. Noting that $S = \operatorname{supp}\mathbf{v}$ ranges through all possible subsets of $[N]$ of cardinality $\operatorname{card}(S) \le 2s$ when $\mathbf{v}$ ranges through all possible $2s$-sparse vectors completes the proof by basic linear algebra. □

We observe, in particular, that if it is possible to reconstruct every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ from the knowledge of its measurement vector $\mathbf{y} = \mathbf{Ax} \in \mathbb{C}^m$, then (i) holds, and consequently so does (iv). This implies $\operatorname{rank}(\mathbf{A}) \ge 2s$. We also have $\operatorname{rank}(\mathbf{A}) \le m$, because the rank is at most equal to the number of rows. Therefore, the number of measurements needed to reconstruct every s-sparse vector always satisfies

$$m \ge 2s.$$

We are now going to see that $m = 2s$ measurements suffice to reconstruct every s-sparse vector — at least in theory.

Theorem 2.14. *For any integer $N \geq 2s$, there exists a measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m = 2s$ rows such that every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ can be recovered from its measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ as a solution of* (P_0).

Proof. Let us fix $t_N > \cdots > t_2 > t_1 > 0$ and consider the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m = 2s$ defined by

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_N \\ \vdots & \vdots & \cdots & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \cdots & t_N^{2s-1} \end{bmatrix}. \tag{2.5}$$

Let $S = \{j_1 < \cdots < j_{2s}\}$ be an index set of cardinality $2s$. The square matrix $\mathbf{A}_S \in \mathbb{C}^{2s \times 2s}$ is (the transpose of) a *Vandermonde* matrix. Theorem A.25 yields

$$\det(\mathbf{A}_S) = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ t_{j_1} & t_{j_2} & \cdots & t_{j_{2s}} \\ \vdots & \vdots & \cdots & \vdots \\ t_{j_1}^{2s-1} & t_{j_2}^{2s-1} & \cdots & t_{j_{2s}}^{2s-1} \end{vmatrix} = \prod_{k<\ell} (t_{j_\ell} - t_{j_k}) > 0.$$

This shows that $\mathbf{A}_S$ is invertible, in particular injective. Since the condition (iii) of Theorem 2.13 is fulfilled, every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s-sparse vector satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, so it can be recovered as the unique solution of (P_0). □

Many other matrices meet the condition (iii) of Theorem 2.13. As an example, the integer powers of $t_1, \ldots, t_N$ in the matrix of (2.5) do not need to be the consecutive integers $0, 1, \ldots, 2s-1$. Instead of the $N \times N$ Vandermonde matrix associated with $t_N > \cdots > t_1 > 0$, we can start with any matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ that is *totally positive*, i.e., that satisfies $\det \mathbf{M}_{I,J} > 0$ for any sets $I, J \subset [N]$ of same cardinality, where $\mathbf{M}_{I,J}$ represents the submatrix of $\mathbf{M}$ with rows indexed by I and columns indexed by J. We then select any $m = 2s$ rows of $\mathbf{M}$, indexed by a set I, say, to form the matrix $\mathbf{A}$. Then, for an index $S \subset [N]$ of cardinality $2s$, the matrix $\mathbf{A}_S$ reduces to $\mathbf{M}_{I,S}$, hence it is invertible. As another example, the numbers $t_N, \ldots, t_1$ do not need to be positive nor real, as long as $\det(\mathbf{A}_S) \neq 0$ instead of $\det(\mathbf{A}_S) > 0$. In particular, with $t_\ell = e^{i2\pi(\ell-1)/N}$ for $\ell \in [N]$, Theorem A.25 guarantees that the partial Fourier matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{i2\pi/N} & e^{i2\pi 2/N} & \cdots & e^{i2\pi(N-1)/N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & e^{i2\pi(2s-1)/N} & e^{i2\pi(2s-1)2/N} & \cdots & e^{i2\pi(2s-1)(N-1)/N} \end{bmatrix}$$

allows the reconstruction of every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^{2s}$. In fact, an argument similar to the one the we will use for Theorem 2.16 below shows that the set of $(2s) \times N$ matrices such that $\det(\mathbf{A}_S) = 0$ for some $S \subset [N]$ with $\mathrm{card}(S) \leq 2s$ has Lebesgue measure zero, hence most $(2s) \times N$ matrices allow the reconstruction of every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^{2s}$. In general, the reconstruction procedure consisting in solving (P_0) is not feasible in practice, as will be shown in Section 2.3. However, in the case of Fourier measurements, a better reconstruction scheme based on the Prony method can be used.

Theorem 2.15. *For any $N \geq 2s$, there exists a practical procedure for the reconstruction of every $2s$-sparse vector from its first $m = 2s$ discrete Fourier measurements.*

Proof. Let $\mathbf{x} \in \mathbb{C}^N$ be an s-sparse vector, which we interpret as a function x from $\{0, 1, \ldots, N-1\}$ into $\mathbb{C}$ supported on an index set $S \subset \{0, 1, \ldots, N-1\}$ of size s. We suppose that this vector is observed via its first $2s$ discrete Fourier coefficients $\hat{x}(0), \ldots, \hat{x}(2s-1)$, where

$$\hat{x}(j) := \sum_{k=0}^{N-1} x(k) e^{-i2\pi jk/N}, \qquad 0 \leq j \leq N-1.$$

We consider the trigonometric polynomial of degree s defined by

$$p(t) := \prod_{k \in S} \big(1 - e^{-i2\pi k/N} e^{i2\pi t/N}\big).$$

This polynomial vanishes exactly for $t \in S$, so we aim at finding the unknown set S by determining p, or equivalently its Fourier transform $\hat{p}$. We note that, since x vanishes on the complementary set $\overline{S}$ of S in $\{0, 1, \ldots, N-1\}$, we have $p(t)x(t) = 0$ for all $0 \leq t \leq N-1$. By discrete convolution, we obtain $\hat{p} * \hat{x} = \widehat{p \cdot x} = 0$, that is to say

$$(\hat{p} * \hat{x})(j) := \sum_{k=0}^{N-1} \hat{p}(k) \cdot \hat{x}(j-k \mod N) = 0 \qquad \text{for all } 0 \leq j \leq N-1. \quad (2.6)$$

We also note that, since $\frac{1}{N}\hat{p}(k)$ is the coefficient of $p(t)$ on the monomial $e^{i2\pi kt/N}$ and since p has degree s, we have $\hat{p}(0) = 1$ and $\hat{p}(k) = 0$ for all $k > s$. It remains to determine the s discrete Fourier coefficients $\hat{p}(1), \ldots, \hat{p}(s)$. For this purpose, we write the s equations (2.6) in the range $s \leq j \leq 2s-1$ in the form

$$\begin{array}{ccccccc}
\hat{x}(s) & + & \hat{p}(1)\hat{x}(s-1) & + \cdots + & \hat{p}(s)\hat{x}(0) & = 0, \\
\hat{x}(s+1) & + & \hat{p}(1)\hat{x}(s) & + \cdots + & \hat{p}(s)\hat{x}(1) & = 0, \\
\vdots & & \vdots & \ddots & \vdots & \vdots \\
\hat{x}(2s-1) & + & \hat{p}(1)\hat{x}(2s-2) & + \cdots + & \hat{p}(s)\hat{x}(s-1) & = 0.
\end{array}$$

This translates into the system

$$\begin{bmatrix} \hat{x}(s-1) & \hat{x}(s-2) & \cdots & \hat{x}(0) \\ \hat{x}(s) & \hat{x}(s-1) & \cdots & \hat{x}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{x}(2s-2) & \hat{x}(2s-3) & \cdots & \hat{x}(s-1) \end{bmatrix} \begin{bmatrix} \hat{p}(1) \\ \hat{p}(2) \\ \vdots \\ \hat{p}(s) \end{bmatrix} = - \begin{bmatrix} \hat{x}(s) \\ \hat{x}(s+1) \\ \vdots \\ \hat{x}(2s-1) \end{bmatrix}.$$

Because $\hat{x}(0), \dots, \hat{x}(2s-1)$ are known, we solve for $\hat{p}(1), \dots, \hat{p}(s)$. Since the Toeplitz matrix above is not always invertible — take e.g. $x = [1, 0, \dots, 0]^\top$, so that $\hat{x} = [1, 1, \dots, 1]^\top$ — we obtain a solution $\hat{q}(1), \dots, \hat{q}(s)$ not guaranteed to be $\hat{p}(1), \dots, \hat{p}(s)$. Appending the values $\hat{q}(0) = 1$ and $\hat{q}(k) = 0$ for all $k > s$, the linear system reads

$$(\hat{q} * \hat{x})(j) = 0 \qquad \text{for all } s \le j \le 2s-1.$$

Therefore, the s-sparse vector $q \cdot x$ has a Fourier transform $\widehat{q \cdot x} = \hat{q} * \hat{x}$ vanishing on a set of s consecutive indices. Writing this in matrix form and using Theorem A.25, we derive that $q \cdot x = 0$, so that the trigonometric polynomial q vanishes on S. Since the degree of q is at most s, the set of zeros of q coincide with the set S, which can thus be found by solving a polynomial equation — or simply by identifying the s smallest values of $|p(j)|$, $0 \le j \le N-1$. Finally, the values of $x(j)$, $j \in S$, are obtained by solving the overdetermined system of $2s$ linear equations imposed by the knowledge of $x(0), \dots, x(2s-1)$. □

Despite its appeal, the reconstruction procedure just described hides some important drawbacks. Namely, it is not stable with respect to sparsity defects nor is it robust with respect to measurement errors. The reader is invited to verify this statement numerically in Exercise 2.8. In fact, we will prove in Chapter 11 that any stable scheme for s-sparse reconstruction requires at least $m \approx c\, s \ln(eN/s)$ linear measurements, where $c > 0$ is a constant depending on the stability requirement.

Recovery of individual sparse vectors

In the next setting, the s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is fixed before the measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is chosen. The conditions for the vector $\mathbf{x}$ to be the unique s-sparse vector consistent with the measurements depend on $\mathbf{A}$ as well as on $\mathbf{x}$ itself. While this seems to be unnatural at first sight because $\mathbf{x}$ is unknown a-priori, the philosophy is that the conditions will be met for *most* $(s+1) \times N$ matrices. This setup is relevant because the measurement matrices are often chosen at random.

Theorem 2.16. *For any $N \ge s+1$, given an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$, there exists a measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m = s+1$ rows such that the vector $\mathbf{x}$ can be reconstructed from its measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ as a solution of* (P_0).

Proof. Let $\mathbf{A} \in \mathbb{C}^{(s+1)\times N}$ be a matrix for which the s-sparse vector $\mathbf{x}$ cannot be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ (via ℓ_0-minimization). This means that there exists a vector $\mathbf{z} \in \mathbb{C}^N$ distinct from $\mathbf{x}$, supported on a set $S = \operatorname{supp}(\mathbf{z}) = \{j_1, \dots, j_s\}$ of size at most s (if $\|\mathbf{z}\|_0 < s$ we fill up S with arbitrary elements $j_\ell \in [N]$), and such that $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. If $\operatorname{supp}(\mathbf{x}) \subset S$, then the equality $\big(\mathbf{A}(\mathbf{z}-\mathbf{x})\big)_{[s]} = 0$ shows that the square matrix $\mathbf{A}_{[s],S}$ is noninvertible, hence

$$f(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N}) := \det(A_{[s],S}) = 0.$$

If $\operatorname{supp}(\mathbf{x}) \not\subset S$, then the space $V := \{\mathbf{u} \in \mathbb{C}^N : \operatorname{supp}(\mathbf{u}) \subset S\} + \mathbb{C}\mathbf{x}$ has dimension $s+1$, and the linear map $G : V \to \mathbb{C}^{s+1}, \mathbf{v} \mapsto \mathbf{A}\mathbf{v}$ is noninvertible, since $G(\mathbf{z}-\mathbf{x}) = 0$. The matrix of the linear map G in the basis $(\mathbf{e}_{j_1}, \dots, \mathbf{e}_{j_s}, \mathbf{x})$ of V takes the form

$$B_{\mathbf{x},S} := \begin{bmatrix} a_{1,j_1} & \cdots & a_{1,j_s} & \sum_{j\in\operatorname{supp}(\mathbf{x})} x_j a_{1,j} \\ \vdots & \ddots & \vdots & \vdots \\ a_{s+1,j_1} & \cdots & a_{s+1,j_s} & \sum_{j\in\operatorname{supp}(\mathbf{x})} x_j a_{s+1,j} \end{bmatrix},$$

and we have

$$g_S(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N}) := \det(B_{\mathbf{x},S}) = 0.$$

This shows that the entries of the matrix $\mathbf{A}$ satisfy

$$(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N}) \in f^{-1}(\{0\}) \cup \bigcup_{\operatorname{card}(S)=s} g_S^{-1}(\{0\}).$$

But since f and all g_S, $\operatorname{card}(S) = s$, are nonzero polynomial functions of the variables $(a_{1,1}, \dots, a_{1,N}, \dots, a_{m,1}, \dots, a_{m,N})$, the sets $f^{-1}(\{0\})$ and $g_S^{-1}(\{0\})$, $\operatorname{card}(S) = s$, have Lebesgue measure zero, and so does their union. It remains to choose the entries of the matrix $\mathbf{A}$ outside of this union of measure zero to ensure that the vector $\mathbf{x}$ can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$. □

2.3 NP-Hardness of ℓ_0-Minimization

As mentioned in Section 2.2, reconstructing an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ from its measurement vector $\mathbf{y} \in \mathbb{C}^m$ amounts to solving the ℓ_0-minimization problem

$$\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_0}$$

Since a minimizer has sparsity at most s, the straightforward approach for finding it consists in solving every rectangular system $\mathbf{A}_S\mathbf{u} = \mathbf{y}$, or rather every square system $\mathbf{A}_S^*\mathbf{A}_S\mathbf{u} = \mathbf{A}_S^*\mathbf{y}$, for $\mathbf{u} \in \mathbb{C}^s$ where S runs through all the possible subsets of $[N]$ with size s. However, since the number $\binom{N}{s}$ of these subsets is prohibitively large, such a straightforward approach is completely

unpractical. By way of illustration, for small problem sizes $N = 1000$ and $s = 10$, we would have to solve $\binom{1000}{10} \geq \left(\frac{1000}{10}\right)^{10} = 10^{20}$ linear systems of size 10×10. Even if each such system could be solved in 10^{-10} seconds, the time required to solve (P_0) with this approach would still be 10^{10} seconds, i.e., more than 300 years. We are going to show that solving (P_0) in fact is intractable for any possible approach. Precisely, for any fixed $\eta \geq 0$, we are going to show that the more general problem

$$\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_0 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta. \tag{$\mathrm{P}_{0,\eta}$}$$

is NP-hard.

We start by introducing the necessary terminology from computational complexity. First, a polynomial-time algorithm is an algorithm performing its task in a number of steps bounded by a polynomial expression in the size of the input. Next, let us describe in a rather informal way a few classes of decision problems.

- The class $\mathfrak{P}$ of P-problems consists of all decision problems for which there exists a polynomial-time algorithm finding a solution.
- The class $\mathfrak{NP}$ of NP-problems consists of all decision problems for which there exists an polynomial-time algorithm certifying a solution. Note that the class $\mathfrak{P}$ is clearly contained in the class $\mathfrak{NP}$.
- The class $\mathfrak{NP}$-hard of NP-hard problems consist of all problems (not necessarily decision problems) for which a solving algorithm could be transformed in polynomial time into a solving algorithm for any NP-problem. Roughly speaking, this is the class of problems at least as hard as any NP-problem. Note that the class $\mathfrak{NP}$-hard is not contained in the class $\mathfrak{NP}$.
- The class $\mathfrak{NP}$-complete of NP-complete problems consist of all problems that are both NP and NP-hard; in other words, it consists of all the NP problems at least as hard as any other NP-problem.

The situation can be summarized with as in Figure 2.3. It is a common belief that $\mathfrak{P}$ is strictly contained in $\mathfrak{NP}$, that is to say that there are problems which can be verified, but not solved, in polynomial time. However, this remains a major open question to this day. There is a vast catalog of NP-complete problems, the most famous of which being perhaps the traveling salesman problem. The one we are going to use is *exact cover by* 3*-sets*.

Exact cover by 3-sets problem:
Given a collection $\{\mathcal{C}_i, i = 1, \ldots, N\}$ of 3-element subsets of $[m]$, does there exist an exact cover (a partition) of $[m]$, i.e., a set $J \in [N]$ such that $\cup_{j\in J}\mathcal{C}_j = [m]$ and $\mathcal{C}_j \cap \mathcal{C}_{j'} = \emptyset$ for all $j, j' \in J$ with $j \neq j'$?

Taking for granted that this problem is NP-complete, we can now prove the main result of this section.

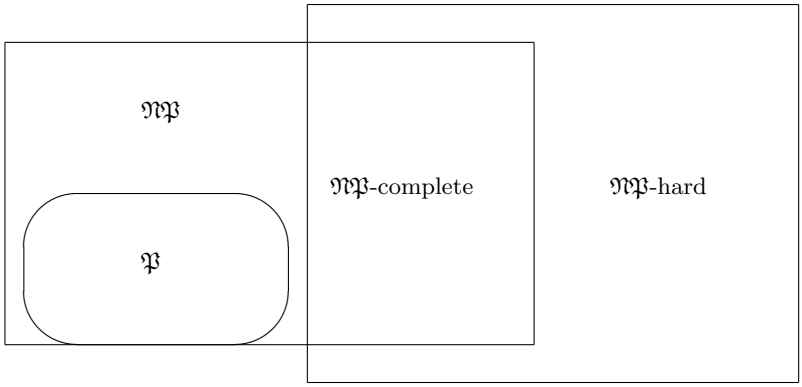

Fig. 2.1. Schematic representation of P, NP, NP-complete and NP-hard problems

Theorem 2.17. *For any $\eta \geq 0$, the ℓ_0-minimization problem $(\mathrm{P}_{0,\eta})$ for general $\mathbf{A} \in \mathbb{C}^{m\times N}$ and $\mathbf{y} \in \mathbb{C}^m$ is NP-hard.*

Proof. By rescaling, we may and do assume that $\eta < 1$. According to the previous considerations, it is enough to show that the exact cover by 3-sets problem can be reduced in polynomial time to the ℓ_0-minimization problem. Let then $\{\mathcal{C}_i, i = 1, \ldots, N\}$ be a collection of 3-element subsets of $[m]$. We define vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N \in \mathbb{C}^m$ by

$$(\mathbf{a}_i)_j = \begin{cases} 1 & \text{if } j \in \mathcal{C}_i, \\ 0 & \text{if } j \notin \mathcal{C}_i. \end{cases}$$

We then define a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ and a vector $\mathbf{y} \in \mathbb{C}^m$ by

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} & & & \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_N \\ & & & \end{array}\right], \qquad \mathbf{y} = [1, 1, \ldots, 1]^\top.$$

Since $N \leq \binom{m}{3}$, this construction can be done in polynomial time. If a vector $\mathbf{z} \in \mathbb{C}^N$ obeys $\|\mathbf{Az} - \mathbf{y}\|_2 \leq \eta$, then all the m components of the vector $\mathbf{Az}$ are distant to 1 by at most η, so they are nonzero and $\|\mathbf{Az}\|_0 = m$. But since each vector $\mathbf{a}_i$ has exactly 3 nonzero components, the vector $\mathbf{Az} = \sum_{j=1}^N z_j \mathbf{a}_j$ has at most $3\|\mathbf{z}\|_0$ nonzero components, $\|\mathbf{Az}\|_0 \leq 3\|\mathbf{z}\|_0$. Therefore, a vector $\mathbf{z} \in \mathbb{C}^N$ obeying $\|\mathbf{Az} - \mathbf{y}\|_2 \leq \eta$ must satisfy $\|\mathbf{z}\|_0 \geq m/3$. Let us now run the ℓ_0-minimization problem, and let $\mathbf{x} \in \mathbb{C}^N$ denote the output. We separate two cases:

1. if $\|\mathbf{x}\|_0 = m/3$, then the collection $\{\mathcal{C}_j, j \in \mathrm{supp}(\mathbf{x})\}$ forms an exact cover of $[m]$, for otherwise the m components of $\mathbf{Ax} = \sum_{j=1}^N x_j \mathbf{a}_j$ would not all be nonzero;

2. if $\|\mathbf{x}\|_0 > m/3$, then no exact cover $\{\mathcal{C}_j, j \in J\}$ can exist, for otherwise the vector $\mathbf{z} \in \mathbb{C}^N$ defined by $z_j = 1$ if $j \in J$ and $z_j = 0$ if $j \notin J$ would satisfy $\mathbf{A}\mathbf{z} = \mathbf{y}$ and $\|\mathbf{z}\|_0 = m/3$, contradicting the ℓ_0-minimality of $\mathbf{x}$.

This shows that solving the ℓ_0-minimization problem enables one to solve the exact cover by 3-sets problem. □

Theorem 2.17 seems rather pessimistic at first sight. However, it concerns the intractability of the problem (P_0) for general matrices $\mathbf{A}$ and vectors $\mathbf{y}$. In other words, any algorithm that is able to solve (P_0) for *any* choice of $\mathbf{A}$ and *any* choice of $\mathbf{y}$ must necessarily be intractable (unless $P = NP$). In compressed sensing, we will rather consider special choices of $\mathbf{A}$ and choose $\mathbf{y} = \mathbf{A}\mathbf{x}$ for some sparse $\mathbf{x}$. We will see that a variety of tractable algorithms will then provably recover $\mathbf{x}$ from $\mathbf{y}$ and thereby solve (P_0) for such specifically designed matrices $\mathbf{A}$. However, to emphasize this point once more, such algorithms will *not* successfully solve the ℓ_0-minimization problem for *all* possible choices of $\mathbf{A}$ and $\mathbf{y}$ due to NP-hardness. A selection of tractable algorithms is introduced in the coming chapter.

Notes

Proposition 2.3 is an observation due to S. Stechkin. In the case $p = 1$ and $q = 2$, the optimal constant $c_{1,2} = 1/2$ was obtained by A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin in [195]. Theorem 2.5 with optimal constants $c_{p,q}$ for all $q > p > 0$ is a particular instance of a more general result, which also contains the *shifting inequality* of Exercise 6.14, see [181].

The weak ℓ_p-spaces are weak L_p-spaces for purely atomic measures. The weak L_p-spaces are also denoted $L_{p,\infty}$ and generalize to Lorentz spaces $L_{p,q}$ [245]. Thus, weak ℓ_p-spaces are a particular instance of more general spaces equipped with the norm

$$\|\mathbf{x}\|_{p,q} = \Big(\sum_{k=1}^{N} k^{q/p-1} (x_k^*)^q \Big)^{1/q}$$

The result of Theorem 2.16 is due to M. Wakin in [440]. Theorem 2.13 can be found in the article by A. Cohen, W. Dahmen, and R. DeVore [102]. One can also add an equivalent proposition expressed in terms of *spark* or in terms of *Kruskal rank*. The spark $\mathrm{sp}(\mathbf{A})$ of a matrix $\mathbf{A}$ was defined by D. Donoho and M. Elad in [133] as the minimal size of a linearly dependent set of columns of $\mathbf{A}$. It is related to the Kruskal rank $\mathrm{kr}(\mathbf{A})$ of $\mathbf{A}$, defined in [271] as the maximal integer k such that any k columns of $\mathbf{A}$ are linearly independent, via $\mathrm{sp}(\mathbf{A}) = \mathrm{kr}(\mathbf{A}) + 1$. Thus, according to Theorem 2.13, every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique s-sparse solution of $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$ if and only if $\mathrm{kr}(\mathbf{A}) \geq 2s$, or if $\mathrm{sp}(\mathbf{A}) > 2s$.

Totally positive matrices were extensively studied by S. Karlin in [258]. One can also consult the more recent book [336] by A. Pinkus.

The reconstruction procedure of Theorem 2.15 based on a discrete version of the Prony method was known long before the development of compressive sensing. It is also related to Reed–Solomon decoding [42, 199]. The general Prony method [347] is designed for recovering a nonharmonic Fourier series of the form

$$f(t) = \sum_{k=1}^{s} x_k e^{2\pi i \omega_k t}$$

from equidistant samples $f(0), f(k/\alpha), f(2k/\alpha), \ldots, f(2s/\alpha)$. Here both the $\omega_k \in \mathbb{R}$ and the x_k are unknown. First the ω_k are found by solving an eigenvalue problem for a Hankel matrix associated to the samples of f. In the second step, the x_k are found by solving a linear system of equations. The difference to the method of Theorem 2.15 is due to the fact that the ω_k are not assumed to lie on a grid anymore. We refer to [308, 296] for more details. The Prony method has the disadvantage of being unstable. Several approaches have been proposed to stabilize it [11, 12, 36, 37, 346], although there seems to be a limit of how stable it can get when the number s of terms gets larger. The recovery methods in so-called theory of *finite rate of innovation* are also related to the Prony method [45].

For an introduction to computational complexity, one can consult [15]. The NP-hardness of the ℓ_0-minimization problem was proved by B. Natarajan in [310]. It was later proved by D. Ge, X. Jiang, and Y. Ye in [190] that the ℓ_p-minimization problem is NP-hard also for any $p < 1$, see Exercise 2.10.

Exercises

2.1. For $0 < p < 1$, prove that the pth power of the ℓ_p-quasinorm satisfies the triangle inequality

$$\|\mathbf{x} + \mathbf{y}\|_p^p \le \|\mathbf{x}\|_p^p + \|\mathbf{y}\|_p^p, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{C}^N.$$

Deduce the inequality

$$\|\mathbf{x}_1 + \cdots + \mathbf{x}_k\|_p \le k^{\max\{0,1/p-1\}} \big(\|\mathbf{x}_1\|_p + \cdots + \|\mathbf{x}_k\|_p\big), \qquad \mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{C}^N.$$

2.2. Show that the constant $k^{\max\{1,1/p\}}$ in Proposition 2.7 is sharp.

2.3. If $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ are disjointly supported, prove that

$$\max(\|\mathbf{u}\|_{1,\infty}, \|\mathbf{v}\|_{1,\infty}) \le \|\mathbf{u} + \mathbf{v}\|_{1,\infty} \le \|\mathbf{u}\|_{1,\infty} + \|\mathbf{v}\|_{1,\infty},$$

and show that these inequalities are sharp.

2.4. As a converse to Proposition 2.10, prove that for any $p > 0$ and any $\mathbf{x} \in \mathbb{C}^N$,

$$\|\mathbf{x}\|_p \le \ln(eN)^{1/p} \|\mathbf{x}\|_{p,\infty}.$$

2.5. Given $q > p > 0$ and $\mathbf{x} \in \mathbb{C}^N$, modify the proof of Proposition 2.3 to obtain $\sigma_s(\mathbf{x})_q \le \|\mathbf{x}\|_{p,\infty}^{1-p/q} \|\mathbf{x}\|_p^{p/q} / s^{1/p-1/q}$.

2.6. Let $(B_0^n, B_1^n, \dots, B_n^n)$ be the *Bernstein polynomials* of degree n defined by

$$B_i^n(x) := \binom{n}{i} x^i (1-x)^{n-i}.$$

For $0 < x_0 < x_1 < \cdots < x_n < 1$, prove that the matrix $[B_i^n(x_j)]_{i,j=0}^n$ is totally positive.

2.7. Prove that the product of two totally positive matrices is totally positive.

2.8. Implement the reconstruction procedure based on $2s$ discrete Fourier measurements as was described in Section 2.2. Test it on a few random examples. Then incorporate small sparsity defect and small measurement error in further tests of the procedure.

2.9. Let us assume that the vectors $\mathbf{x} \in \mathbb{R}^N$ are no longer observed via linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$, but rather via measurements $\mathbf{y} = f(\mathbf{x})$ where $f : \mathbb{R}^N \to \mathbb{R}^m$ is a continuous map satisfying $f(-\mathbf{x}) = -f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^N$. Prove that the minimal number of measurements needed to reconstruct every s-sparse vector equals $2s$.

2.10. NP-Hardness of ℓ_p-minimization for $0 < p < 1$.
Given $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $\mathbf{y} \in \mathbb{C}^m$, the ℓ_p-minimization problem consists in computing a vector $\mathbf{x} \in \mathbb{C}^N$ with minimal ℓ_p-quasinorm subject to $\mathbf{A}\mathbf{x} = \mathbf{y}$. Assuming the NP-completeness of the *partition problem*, which consists, given integers $a_1, \dots, a_n$, in finding two sets $I, J \subset [n]$ such that $\sum_{i \in I} a_i = \sum_{j \in J} a_j$, prove that the ℓ_p-minimization problem is NP-hard. It will be helpful to introduce the matrix $\mathbf{A}$ and the vector $\mathbf{y}$ defined by

$$\mathbf{A} := \begin{bmatrix} a_1 & a_2 & \cdots & a_n & a_1 & a_2 & \cdots & a_n \\ 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & 0 & \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = [0, 1, 1, \dots, 1]^\top.$$

2.11. NP-Hardness of rank minimization.
Show that the rank-minimization problem

$$\underset{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}}{\text{minimize}} \operatorname{rank}(\mathbf{Z}) \quad \text{subject to } \mathcal{A}(\mathbf{X}) = \mathbf{y}\ .$$

is NP-hard on the set of linear measurement maps $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ and vectors $\mathbf{y} \in \mathbb{R}^m$.

3

Basic Algorithms

In this chapter, a selection of popular algorithms used in Compressive Sensing is presented. The algorithms are divided into three categories: optimization methods, greedy methods, and thresholding-based methods. Their rigorous analyses are postponed until later, when appropriate tools such as coherence and restricted isometry constants become available. Only intuitive justification is given for now.

3.1 Optimization Methods

An *optimization problem* is a problem of the type

$$\underset{\mathbf{x}\in\mathbb{R}^N}{\text{minimize}}\ F_0(\mathbf{x}) \quad \text{subject to } F_i(\mathbf{x}) \leq b_i,\ 1 \leq i \leq n,$$

where the function $F_0 : \mathbb{R}^N \to \mathbb{R}$ is called *objective function* and the functions $F_1, \ldots, F_n : \mathbb{R}^N \to \mathbb{R}$ are called *constraint functions* . This general framework also encompasses equality constraints of the type $G_i(\mathbf{x}) = c_i$, since the equality $G_i(\mathbf{x}) = c_i$ is equivalent to the inequalities $G_i(\mathbf{x}) \leq c_i$ and $-G_i(\mathbf{x}) \leq -c_i$. If $F_0, F_1, \ldots, F_n$ are all convex functions, then the problem is called a *convex optimization problem* — see Appendix B.5 for more information. If $F_0, F_1, \ldots, F_n$ are all linear functions, then the problem is called a *linear program*. Our sparse recovery problem is in fact an optimization problem, since it translates into

$$\text{minimize } \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_0}$$

This is a nonconvex problem, and we even have seen in Theorem 2.17 that it is NP-hard in general. However, keeping in mind that $\|\mathbf{z}\|_q^q$ approaches $\|\mathbf{z}\|_0$ as $q > 0$ tends to zero, we can approximate (P_0) by the problem

$$\text{minimize } \|\mathbf{z}\|_q \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_q}$$

For $q > 1$, even 1-sparse vectors are not solutions of (P_q) — see Exercise 3.1. For $0 < q < 1$, (P_q) is again a nonconvex problem, which is also NP-hard in

general — see Exercise 2.10. But for the critical value $q = 1$, it becomes the following convex problem (interpreted as the convex relaxation of (P_0), see Section B.3 for the definition of convex relaxation)

$$\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_1}$$

The associated method is usually called ℓ_1-*minimization* or *basis pursuit*. There are several specific algorithms to solve this optimization problem, and some of them are presented in Chapter 15.

Basis pursuit

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$.
Instruction:

$$\mathbf{x}^\sharp = \operatorname{argmin} \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{BP}$$

Output: the vector $\mathbf{x}^\sharp$.

Let us complement the previous intuitive justification by the observation that ℓ_1-minimizers are sparse, at least in the real setting. In the complex setting, this is not necessarily true, see Exercise 3.2.

Theorem 3.1. *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a measurement matrix with columns $\mathbf{a}_1, \dots, \mathbf{a}_N$. Assuming the uniqueness of a minimizer $\mathbf{x}^\sharp$ of*

$$\underset{\mathbf{z}\in\mathbb{R}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y},$$

the system $\{\mathbf{a}_j, j \in \operatorname{supp} \mathbf{x}^\sharp\}$ is linearly independent, and in particular

$$\|\mathbf{x}^\sharp\|_0 = \operatorname{card}(\operatorname{supp} \mathbf{x}^\sharp) \le m.$$

Proof. By way of contradiction, let us assume that the system $\{\mathbf{a}_j, j \in S\}$ is linearly dependent, where $S = \operatorname{supp} \mathbf{x}^\sharp$. This means that there exists a nonzero vector $\mathbf{v} \in \mathbb{R}^N$ supported on S such that $\mathbf{A}\mathbf{v} = 0$. Then, for any $t \neq 0$,

$$\|\mathbf{x}^\sharp\|_1 < \|\mathbf{x}^\sharp + t\mathbf{v}\|_1 = \sum_{j\in S} |x_j^\sharp + tv_j| = \sum_{j\in S} \operatorname{sgn}(x_j^\sharp + tv_j)(x_j^\sharp + tv_j).$$

If $|t|$ is small enough, namely $|t| < \min_{j\in S} |x_j|/\|\mathbf{v}\|_\infty$, we have

$$\operatorname{sgn}(x_j^\sharp + tv_j) = \operatorname{sgn}(x_j^\sharp) \qquad \text{for all } j \in S.$$

It follows that, for $t \neq 0$ with $|t| < \min_{j\in S} |x_j|/\|\mathbf{v}\|_\infty$,

$$\begin{aligned}\|\mathbf{x}^\sharp\|_1 &< \sum_{j\in S} \operatorname{sgn}(x_j^\sharp)(x_j^\sharp + tv_j) = \sum_{j\in S} \operatorname{sgn}(x_j^\sharp)x_j^\sharp + t\sum_{j\in S} \operatorname{sgn}(x_j^\sharp)v_j \\ &= \|\mathbf{x}^\sharp\|_1 + t\sum_{j\in S} \operatorname{sgn}(x_j^\sharp)v_j.\end{aligned}$$

This is a contradiction, because we can always choose a small $t \neq 0$ such that $t \sum_{j \in S} \operatorname{sgn}(x_j^\sharp) v_j \leq 0$. $\square$

In the real setting, it is also worth pointing out that (P_1) can be recast as a linear program by introducing slack variables $\mathbf{z}^+, \mathbf{z}^- \in \mathbb{R}^N$. Given $\mathbf{z} \in \mathbb{R}^N$, these are defined, for $j \in [N]$, by

$$z_j^+ = \begin{cases} z_j & \text{if } z_j > 0, \\ 0 & \text{if } z_j \leq 0, \end{cases} \qquad z_j^- = \begin{cases} 0 & \text{if } z_j > 0, \\ -z_j & \text{if } z_j \leq 0. \end{cases}$$

The problem (P_1) is thus equivalent to a linear program with optimization variables $\mathbf{z}^+, \mathbf{z}^- \in \mathbb{R}^N$, namely to

$$\underset{\mathbf{z}^+, \mathbf{z}^- \in \mathbb{R}^N}{\text{minimize}} \sum_{j=1}^N (z_j^+ + z_j^-) \quad \text{subject to } \begin{bmatrix} \mathbf{A} \,|\, -\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{z}^+ \\ \mathbf{z}^- \end{bmatrix} = \mathbf{y}, \quad \begin{bmatrix} \mathbf{z}^+ \\ \mathbf{z}^- \end{bmatrix} \geq 0. \quad (\mathrm{P}_1')$$

Given the solution $(\mathbf{x}^+)^*, (\mathbf{x}^-)^*$ of this program, the solution of (P_1) is recovered by $\mathbf{x}^* = (\mathbf{x}^+)^* - (\mathbf{x}^-)^*$.

These considerations do not make sense in the complex setting. In this case, we directly consider a more general ℓ_1-minimization that takes measurement error into account, namely

$$\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta. \qquad (\mathrm{P}_{1,\eta})$$

This variation is natural because in general the measurement vector $\mathbf{y} \in \mathbb{C}^m$ is not exactly equal to $\mathbf{A}\mathbf{x} \in \mathbb{C}^m$, but rather to $\mathbf{A}\mathbf{x} + \mathbf{e}$ for some measurement error $\mathbf{e} \in \mathbb{C}^m$ that can be estimated in ℓ_2-norm, say, by $\|\mathbf{e}\|_2 \leq \eta$ for some $\eta \geq 0$. Then, given a vector $\mathbf{z} \in \mathbb{C}^N$, we introduce its real and imaginary parts $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ and a vector $\mathbf{c} \in \mathbb{R}^N$ such that $c_j \geq |z_j| = \sqrt{u_j^2 + v_j^2}$ for all $j \in [N]$. The problem $(\mathrm{P}_{1,\eta})$ is then equivalent to the following problem with optimization variables $\mathbf{c}, \mathbf{u}, \mathbf{v} \in \mathbb{R}^N$:

$$\underset{\mathbf{c}, \mathbf{u}, \mathbf{v} \in \mathbb{R}^N}{\text{minimize}} \sum_{j=1}^N c_j \quad \text{subject to } \left\| \begin{bmatrix} \operatorname{Re}(\mathbf{A}) & -\operatorname{Im}(\mathbf{A}) \\ \operatorname{Im}(\mathbf{A}) & \operatorname{Re}(\mathbf{A}) \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} - \begin{bmatrix} \operatorname{Re}(\mathbf{y}) \\ \operatorname{Im}(\mathbf{y}) \end{bmatrix} \right\|_2 \leq \eta, \quad (\mathrm{P}_{1,\eta}')$$

$$\sqrt{u_1^2 + v_1^2} \leq c_1,$$
$$\vdots$$
$$\sqrt{u_N^2 + v_N^2} \leq c_N.$$

This is an instance of a *second-order cone program*, see Appendix B.5 for more details. Given its solution $(\mathbf{c}^*, \mathbf{u}^*, \mathbf{v}^*)$, the solution to $(\mathrm{P}_{1,\eta})$ is given by $\mathbf{x}^* = \mathbf{u}^* + i\mathbf{v}^*$. Note that the choice $\eta = 0$ yields the second-order cone formulation of (P_1) in the complex case.

The associated method, called *quadratically-constrained basis pursuit* (or sometimes noise-aware ℓ_1-minimization), reads as follows.

Quadratically-constrained basis pursuit

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, noise level η.
Instruction:

$$\mathbf{x}^\sharp = \operatorname{argmin} \|\mathbf{z}\|_1 \qquad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta. \tag{BP_η}$$

Output: the vector $\mathbf{x}^\sharp$.

The solution $\mathbf{x}^\sharp$ of

$$\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}} \|\mathbf{z}\|_1 \qquad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta \tag{3.1}$$

is strongly linked to the output of the *basis pursuit denoising* algorithm, which consists in solving, for some parameter $\lambda \ge 0$,

$$\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}} \lambda\|\mathbf{z}\|_1 + \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2 \,. \tag{3.2}$$

The solution of (3.1) is also related to the output of the *LASSO*, which consists in solving, for some parameter $\tau \ge 0$,

$$\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}} \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \qquad \text{subject to } \|\mathbf{z}\|_1 \le \tau. \tag{3.3}$$

Precisely, the link between the three algorithms is given in the Proposition below, which follows from Theorem B.28.

Proposition 3.2. *(a) If* $\mathbf{x}$ *is a minimizer of* (3.2)*, then there exists* $\eta = \eta_{\mathbf{x}}$ *such that* $\mathbf{x}$ *is a minimizer of the quadratically constraint basis pursuit problem* (3.1)*.*
(b) If $\mathbf{x}$ *is a minimizer of quadratically constraint basis pursuit* (3.1)*, then there is* $\tau = \tau_{\mathbf{x}}$ *such that* $\mathbf{x}$ *is a minimizer of the LASSO* (3.3)*.*
(c) If $\mathbf{x}$ *is a minimizer of the LASSO* (3.3)*, then there is* $\lambda = \lambda_{\mathbf{x}}$ *such that* $\mathbf{x}$ *is a minimizer of* (3.2)*.*

Another type of ℓ_1-minimization problem is the Dantzig selector,

$$\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}} \|\mathbf{z}\|_1 \qquad \text{subject to } \|\mathbf{A}^*(\mathbf{A}\mathbf{z} - \mathbf{y})\|_\infty \le \tau \,. \tag{3.4}$$

This is again a convex optimization problem. The intuition for the constraint is that the residual $\mathbf{r} = \mathbf{A}\mathbf{z} - \mathbf{y}$ should have small correlation with all columns $\mathbf{a}_j$ of the matrix $\mathbf{A}$ – indeed, $\|\mathbf{A}^*(\mathbf{A}\mathbf{z} - \mathbf{y})\|_\infty = \max_{j\in[N]} |\langle \mathbf{r}, \mathbf{a}_j\rangle|$. A similar theory as will be developed for the ℓ_1-minimization problems (BP) and (BP_η) later in the book is valid for the Dantzig selector as well, but we will not go into details.

3.2 Greedy Methods

In this section, we introduce two iterative greedy algorithms commonly used in compressive sensing. The first algorithm, called *orthogonal matching pursuit*, adds one index to a target support S^n at each iteration, and update a target vector $\mathbf{x}^n$ as the vector supported on the target support S^n that best fits the measurements. The algorithm is formally described as follows.

Orthogonal matching pursuit

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$.
Initialization: $S^0 = \emptyset$, $\mathbf{x}^0 = 0$.
Iteration: repeat the following steps until a stopping criterion is met at $n = \bar{n}$

$$S^{n+1} = S^n \cup \big\{ j_{n+1} := \operatorname{argmax}\{|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_j|, j \in [N]\}\big\}, \quad (\mathrm{OMP}_1)$$
$$\mathbf{x}^{n+1} = \operatorname{argmin}\big\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq S^{n+1}\big\}. \quad (\mathrm{OMP}_2)$$

Output: the $\bar{n}$-sparse vector $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$.

The projection step (OMP_2) is the most costly part of the orthogonal matching pursuit algorithm. It can be accelerated by using the QR-decomposition of $\mathbf{A}_{S_n}$. In fact, efficient methods exist for updating the QR-decomposition when a column is added to the matrix. If available one may alternatively exploit fast matrix-vector multiplications for $\mathbf{A}$ (like the Fast Fourier Transform, see Section C.1). We refer to the discussion at the end of Section A.3 for details. In the case that fast matrix-vector multiplication routines are available for $\mathbf{A}$ and $\mathbf{A}^*$, they should also be used for speed up of the computation of $\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)$.

The choice of the index j_{n+1} is dictated by a greedy strategy where one aims to reduce the ℓ_2-norm of the residual $\mathbf{y} - \mathbf{A}\mathbf{x}^n$ as much as possible at each iteration. The following lemma (refined in Exercise 3.10) applied with $S = S^n$ and $\mathbf{u} = \mathbf{x}^n$ gives some insight as to why an index j maximizing $|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_j|$ is a good candidate for a large decrease of the ℓ_2-norm of the residual.

Lemma 3.3. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be a matrix with ℓ_2-normalized columns. Given $S \subseteq [N]$ and $j \in [N]$, if*

$$\mathbf{v} := \operatorname{argmin}\big\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq S\big\},$$
$$\mathbf{w} := \operatorname{argmin}\big\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq S \cup \{j\}\big\},$$

then

$$\|\mathbf{y} - \mathbf{A}\mathbf{w}\|_2^2 \le \|\mathbf{y} - \mathbf{A}\mathbf{v}\|_2^2 - |(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{v}))_j|^2.$$

Proof. Since any vector of the form $\mathbf{v}+t\mathbf{e}_j$ with $t \in \mathbb{C}$ is supported on $V \cup \{j\}$, we have

$$\|\mathbf{y} - \mathbf{A}\mathbf{w}\|_2^2 \leq \min_{t \in \mathbb{C}} \|\mathbf{y} - \mathbf{A}(\mathbf{v} + t\mathbf{e}_j)\|_2^2.$$

Writing $t = \rho e^{i\theta}$ with $\rho \geq 0$ and $\theta \in [0, 2\pi)$, we compute

$$\begin{aligned}
\|\mathbf{y} - \mathbf{A}(\mathbf{v} + t\mathbf{e}_j)\|_2^2 &= \|\mathbf{y} - \mathbf{A}\mathbf{v} - t\mathbf{A}\mathbf{e}_j\|_2^2 \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{v}\|_2^2 + |t|^2\|\mathbf{A}\mathbf{e}_j\|_2^2 - 2\,\mathrm{Re}(\bar{t}\langle \mathbf{y} - \mathbf{A}\mathbf{v}, \mathbf{A}\mathbf{e}_j\rangle) \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{v}\|_2^2 + \rho^2 - 2\,\mathrm{Re}(\rho e^{-i\theta}(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{v}))_j) \\
&\geq \|\mathbf{y} - \mathbf{A}\mathbf{v}\|_2^2 + \rho^2 - 2\rho|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{v}))_j|,
\end{aligned}$$

with equality for a properly chosen θ. As a quadratic polynomial in ρ, the latter expression is minimized when $\rho = |(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{u}))_j|$. This shows that

$$\min_{t \in \mathbb{C}} \|\mathbf{y} - \mathbf{A}(\mathbf{v} + t\mathbf{e}_j)\|_2^2 = \|\mathbf{y} - \mathbf{A}\mathbf{v}\|_2^2 - |(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{u}))_j|^2,$$

which concludes the proof. □

We point out that step (OMP_2) also reads (with a slight abuse of notations) as

$$\mathbf{x}^{n+1} = \mathbf{A}_{S^{n+1}}^{\dagger}\mathbf{y},$$

where $\mathbf{A}_{S^{n+1}}^{\dagger}$ is the pseudoinverse of $\mathbf{A}_{S^{n+1}}$, see Section A.2 for details. This simply says that $\mathbf{x}^{n+1}$ (to be precise, $\mathbf{x}^{n+1}(S^{n+1})$) is a solution of $\mathbf{A}_{S^{n+1}}^*\mathbf{A}_{S^{n+1}}\mathbf{z} = \mathbf{A}_{S^{n+1}}^*\mathbf{y}$. This fact is justified in the following lemma, which will also be useful for other algorithms containing a step similar to (OMP_2).

Lemma 3.4. *Given an index set $V \subseteq [N]$, if*

$$\mathbf{v} := \operatorname{argmin}\big\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq V\big\},$$

then

$$(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{z}))_V = 0. \tag{3.5}$$

Proof. According the definition of $\mathbf{v}$, the vector $\mathbf{A}\mathbf{v}$ is the orthogonal projection of $\mathbf{y}$ onto the space $\{\mathbf{A}\mathbf{z}, \operatorname{supp}(\mathbf{z}) \subseteq V\}$, hence it is characterized by the orthogonality condition

$$\langle \mathbf{y} - \mathbf{A}\mathbf{v}, \mathbf{A}\mathbf{z}\rangle = 0 \qquad \text{for all } \mathbf{z} \in \mathbb{C}^N \text{ with } \operatorname{supp}(\mathbf{z}) \subseteq V.$$

This means that $\langle \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{v}), \mathbf{z}\rangle = 0$ for all $\mathbf{z} \in \mathbb{C}^N$ with $\operatorname{supp}(\mathbf{z}) \subseteq V$, which holds if and only if (3.5) is satisfied. □

A natural stopping criterion for the orthogonal matching pursuit algorithm is $\mathbf{A}\mathbf{x}^{\bar{n}} = \mathbf{y}$. However, to account for measurement and computation errors, we use instead $\|\mathbf{y} - \mathbf{A}\mathbf{x}^{\bar{n}}\|_2 \leq \varepsilon$ and $\|\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{\bar{n}})\|_\infty \leq \varepsilon$ for some chosen tolerance $\epsilon > 0$. If there is an estimate for the sparsity s of the vector $\mathbf{x} \in \mathbb{C}^N$

to be recovered, another possible stopping criterion can simply be $\bar{n} = s$, since then the target vector $\mathbf{x}^{\bar{n}}$ is s-sparse. For instance, if $\mathbf{A}$ is a square orthogonal matrix, then the algorithm with this stopping criterion successfully recovers an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$, since it can be seen that the vector $\mathbf{x}^n$ produced at the nth iteration equals the n-sparse vector consisting of n largest entries of $\mathbf{x}$. More generally, the success of recovery of s-sparse vectors via s iterations of the orthogonal matching pursuit algorithm is determined by the following result.

Proposition 3.5. *Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$, every nonzero vector $\mathbf{x} \in \mathbb{C}^N$ supported on a set S of size s is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ after at most s iterations of orthogonal matching pursuit if and only if the matrix $\mathbf{A}_S$ is injective and*

$$\max_{j \in S} |(\mathbf{A}^* \mathbf{r})_j| > \max_{\ell \in \overline{S}} |(\mathbf{A}^* \mathbf{r})_\ell| \tag{3.6}$$

for all nonzero $\mathbf{r} \in \big\{\mathbf{A}\mathbf{z}, \operatorname{supp}(\mathbf{z}) \subseteq S\big\}$.

Proof. Let us assume that the orthogonal matching pursuit algorithm recovers all vectors supported on a set S in at most $s = \operatorname{card}(S)$ iterations. Then, since two vectors supported on S which have the same measurement vector must be equal, the matrix $\mathbf{A}_S$ is injective. Moreover, since the index chosen at the first iteration always stays in the target support, if $\mathbf{y} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{C}^N$ exactly supported on S, then an index $\ell \in \overline{S}$ cannot be chosen at the first iteration, i.e., $\max_{j \in S} |(\mathbf{A}^* \mathbf{y})_j| > |(\mathbf{A}^* \mathbf{y})_\ell|$. Therefore, we have $\max_{j \in S} |(\mathbf{A}^* \mathbf{y})_j| > \max_{\ell \in \overline{S}} |(\mathbf{A}^* \mathbf{y})_\ell|$ for all nonzero $\mathbf{y} \in \big\{\mathbf{A}\mathbf{z}, \operatorname{supp}(\mathbf{z}) \subseteq S\big\}$. This shows the necessity of the two conditions given in the proposition.

To prove their sufficiency, assuming that $\mathbf{A}\mathbf{x}^1 \neq \mathbf{y}, \ldots, \mathbf{A}\mathbf{x}^{s-1} \neq \mathbf{y}$ (otherwise there is nothing to do), we are going to prove that S^n is a subset of S of size n for any $0 \leq n \leq s$. This will imply $S^s = S$, hence $\mathbf{A}\mathbf{x}^s = \mathbf{y}$ by (OMP_2), and finally $\mathbf{x}^s = \mathbf{x}$ by the injectivity of $\mathbf{A}_S$. To establish our claim, given $0 \leq n \leq s-1$, we first notice that $S^n \subseteq S$ yields $\mathbf{r}^n := \mathbf{y} - \mathbf{A}\mathbf{x}^n \in \big\{\mathbf{A}\mathbf{z}, \operatorname{supp}(\mathbf{z}) \subseteq S\big\}$, so that the index j_{n+1} lies in S by (3.6), and $S^{n+1} = S^n \cup \{j_{n+1}\} \subseteq S$ by (OMP_1). This inductively proves that S^n is a subset of S for any $0 \leq n \leq s$. Next, given $1 \leq n \leq s-1$, Lemma 3.4 implies that $(\mathbf{A}^* \mathbf{r}^n)_{S^n} = 0$. Therefore, according to its definition in (OMP_1), the index j_{n+1} does not lie in S^n, since this would mean that $\mathbf{A}^* \mathbf{r}^n = 0$, and in turn that $\mathbf{r}^n = 0$ by (3.6). This inductively proves that S^n is a set of size n. The proof is now complete. □

Remark 3.6. A more concise way to formulate the necessary and sufficient condition of Proposition 3.5 is the *exact recovery condition*, which reads

$$\|\mathbf{A}_S^\dagger \mathbf{A}_{\overline{S}}\|_{1 \to 1} < 1, \tag{3.7}$$

see Section A.1 for the definition of matrix norms. Implicitly, the existence of the pseudoinverse $\mathbf{A}_S^\dagger = (\mathbf{A}_S^* \mathbf{A}_S)^{-1} \mathbf{A}_S^*$ is equivalent to the injectivity of $\mathbf{A}_S$. Moreover, (3.6) is then equivalent to

$$\|\mathbf{A}_S^*\mathbf{A}_S\mathbf{u}\|_\infty > \|A_{\overline{S}}^*\mathbf{A}_S\mathbf{u}\|_\infty \quad \text{for all } \mathbf{u} \in \mathbb{C}^s \setminus \{0\}.$$

Making the change $\mathbf{v} = \mathbf{A}_S^*\mathbf{A}_S\mathbf{u}$, this can be written as

$$\|\mathbf{v}\|_\infty > \|\mathbf{A}_{\overline{S}}^*\mathbf{A}_S(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\mathbf{v}\|_\infty = \|\mathbf{A}_{\overline{S}}^*(\mathbf{A}_S^\dagger)^*\mathbf{v}\|_\infty \quad \text{for all } \mathbf{v} \in \mathbb{C}^s \setminus \{0\}.$$

The latter reads $\|\mathbf{A}_{\overline{S}}^*(\mathbf{A}_S^\dagger)^*\|_{\infty\to\infty} < 1$, that is to say $\|\mathbf{A}_S^\dagger\mathbf{A}_{\overline{S}}\|_{1\to 1} < 1$.

A weakness of the orthogonal matching pursuit algorithm is that, once an incorrect index has been selected in a target support S^n, it remains in all the subsequent target supports $S^{n'}$ for $n' \geq n$ — see Section 6.4 where this issue is illustrated on a detailed example. Hence, if an incorrect index has been selected, s iterations of the orthogonal matching pursuit are not enough to recover a vector with sparsity s. A possible way out is to increase the number of iterations. The following algorithm, called *compressive sampling matching pursuit algorithm*, proposes another way out when an estimation of the sparsity s is available. To describe it, it is convenient to introduce the notations $H_s(\mathbf{z})$ for the best s-term approximation to $\mathbf{z} \in \mathbb{C}^N$ and $L_s(\mathbf{z})$ for the support of the latter, i.e.,

$$\begin{aligned} L_s(\mathbf{z}) &:= \text{index set of } s \text{ largest entries of } \mathbf{z} \in \mathbb{C}^N \text{ in modulus}\,, \\ H_s(\mathbf{z}) &:= \mathbf{z}_{L_s(\mathbf{z})}\,. \end{aligned}$$

The nonlinear operator H_s is called *hard thresholding operator* of order s. Given the vector $\mathbf{z} \in \mathbb{C}^N$, the operator H_s keeps its s largest absolute entries and sets the other ones to zero. Note that it may not be uniquely defined. To resolve this issue, we choose the index set $L_s(\mathbf{z})$ out of all possible candidates according to a predefined rule, for instance the lexicographic order.

Compressive sampling matching pursuit

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, sparsity level s.
Initialization: s-sparse vector $\mathbf{x}^0$, typically $\mathbf{x}^0 = 0$.
Iteration: repeat the following steps until a stopping criterion is met at $n = \bar{n}$

$$U^{n+1} = \operatorname{supp}(\mathbf{x}^n) \cup L_{2s}(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))\,, \tag{CoSaMP$_1$}$$

$$\mathbf{u}^{n+1} = \operatorname{argmin}\left\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq U^{n+1}\right\}\,, \tag{CoSaMP$_2$}$$

$$\mathbf{x}^{n+1} = H_s(\mathbf{u}^{n+1}). \tag{CoSaMP$_3$}$$

Output: the s-sparse vector $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$.

3.3 Thresholding-Based Methods

In this section, we describe further algorithms involving the hard thresholding operator H_k. The intuition for these algorithms, which justifies categorizing

them in a different family, relies on the approximate inversion of the action on sparse vectors of the measurement matrix $\mathbf{A}$ by the action of its conjugate $\mathbf{A}^*$. Thus, the *basic thresholding algorithm* consists in determining the support of the s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ to be recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ as the indices of s largest absolute entries of $\mathbf{A}^*\mathbf{y}$, and then to find the vector with this support that best fits the measurement. Formally, the algorithm reads as follows.

Basic thresholding

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, sparsity level s.
Instruction:

$$S^\sharp = L_s(\mathbf{A}^*\mathbf{y}), \tag{BT$_1$}$$

$$\mathbf{x}^\sharp = \operatorname{argmin} \big\{ \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq S^\sharp \big\}. \tag{BT$_2$}$$

Output: the s-sparse vector $\mathbf{x}^\sharp$. □

A necessary and sufficient condition resembling (3.6) can be given for the success of s-sparse recovery using this simple algorithm.

Proposition 3.7. *A vector $\mathbf{x} \in \mathbb{C}^N$ supported on a set S is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via basic thresholding if and only if*

$$\min_{j \in S} |(\mathbf{A}^*\mathbf{y})_j| > \max_{\ell \in \overline{S}} |(\mathbf{A}^*\mathbf{y})_\ell|. \tag{3.8}$$

Proof. It is clear that the vector $\mathbf{x}$ is recovered if and only if the index set $S^\sharp$ defined in (BT$_1$) coincides with the set S, that is to say if and only if any entry of $\mathbf{A}^*\mathbf{y}$ on S is greater than any entry of $\mathbf{A}^*\mathbf{y}$ on $\overline{S}$. This is property (3.8). □

The more elaborate *iterative hard thresholding algorithm* is an iterative algorithm to solve the rectangular system $\mathbf{A}\mathbf{z} = \mathbf{y}$, knowing that the solution is s-sparse. We shall solve the square system $\mathbf{A}^*\mathbf{A}\mathbf{z} = \mathbf{A}^*\mathbf{y}$ instead, which can be interpreted as the fixed-point equation $\mathbf{z} = (\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{z} + \mathbf{A}^*\mathbf{y}$. Classical iterative methods suggest the fixed-point iteration $\mathbf{x}^{n+1} = (\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{x}^n + \mathbf{A}^*\mathbf{y}$. Since we target s-sparse vectors, we only keep the s largest (in modulus) entries of $(\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{x}^n + \mathbf{A}^*\mathbf{y} = \mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)$ at each iteration. The resulting algorithm reads as follows.

Iterative hard thresholding

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, sparsity level s.
Initialization: s-sparse vector $\mathbf{x}^0$, typically $\mathbf{x}^0 = 0$.
Iteration: repeat the following step until a stopping criterion is met at $n = \bar{n}$:

$$\mathbf{x}^{n+1} = H_s(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)). \tag{IHT}$$

Output: the s-sparse vector $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$. □

The iterative hard thresholding algorithm does not require to compute any orthogonal projection. If we are willing to pay the price of the orthogonal projections, like in the greedy methods, it makes sense to look at the vector with the same support as $\mathbf{x}^{n+1}$ that best fits the measurements. This leads to the *hard thresholding pursuit algorithm* defined below.

Hard thresholding pursuit

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, sparsity level s.
Initialization: s-sparse vector $\mathbf{x}^0$, typically $\mathbf{x}^0 = 0$.
Iteration: repeat the following step until a stopping criterion is met at $n = \bar{n}$:

$$S^{n+1} = L_s(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)), \tag{HTP$_1$}$$

$$\mathbf{x}^{n+1} = \operatorname{argmin}\big\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq S^{n+1}\big\}. \tag{HTP$_2$}$$

Output: the s-sparse vector $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$.

Notes

More background on convex optimization can be found in in Appendix B, and in the books [59, 318] by S. Boyd and L. Vandenberghe and by J. Nocedal and S. Wright, respectively.

Basis Pursuit was introduced by S. Chen, D. Donoho and M. Saunders in [94]. The LASSO (Least Absolute Shrinkage and Selection Operator) algorithm is more popular in the statistics literature than the quadratically-constrained basis pursuit or basis pursuit denoising algorithms. It was introduced by R. Tibshirani in [411]. The Dantzig selector (3.4) was introduced by E. Candés and T. Tao in [83]. Like the LASSO, it is more popular in statistics than in signal processing.

A greedy strategy that does not involve any orthogonal projection consists in updating $\mathbf{x}^n$ as $\mathbf{x}^{n+1} = \mathbf{x}^n + t\mathbf{e}_j$, where $t \in \mathbb{C}$ and $j \in [N]$ are chosen to

minimize $\|\mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}\|$. The argument of Lemma 3.3 imposes the choice of j as a maximizer of $|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_j|$ and then $t = (\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_j$. This corresponds to the *matching pursuit* algorithm, introduced in signal processing by S. Mallat and Z. Zhang in [294] and by S. Qian and D. Chen in [349], and in statistics as the projection pursuit regression by J. H. Friedman and W. Stuetzle in [186]. In approximation theory, it is known as pure greedy algorithm, see for instance the surveys [408, 409] and the monograph [410] by V. Temlyakov. There, the orthogonal matching pursuit algorithm is also known as orthogonal greedy algorithm. Just like matching pursuit, it was introduced independently by several researchers in different fields, e.g. by G. Davis, S. Mallat, and Z. Zhang in [121], by Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad in [327], by S. Chen, S. A. Billings, and W. Luo in [93], or in [239] by J. Högborn, where it was called CLEAN in the context of astronomical data processing. The orthogonal matching pursuit algorithm was analyzed in terms of sparse recovery by J. Tropp in [414].

The *compressive sampling matching pursuit* algorithm was devised by D. Needell and J. Tropp in [312]. It was inspired by the earlier *regularized orthogonal matching pursuit* developed and analyzed by D. Needell and R. Vershynin in [313, 314].

The *subspace pursuit* algorithm, introduced by W. Dai and O. Milenkovic in [110], is another algorithm in the greedy family, but it will not be analyzed in this book. It bears some resemblance with compressive sampling matching pursuit, except that, instead of $2s$, only s indices of largest (in modulus) entries of the residual vector are selected, and that an additional orthogonal projection step is performed at each iteration. Its description is given below.

Subspace pursuit

Input: measurement matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, sparsity level s.
Initialization: s-sparse vector $\mathbf{x}^0$, typically $\mathbf{x}^0 = 0$, $S^0 = \mathrm{supp}(\mathbf{x}^0)$.
Iteration: repeat the following steps until a stopping criterion is met at $n = \bar{n}$

$$U^{n+1} = S^n \cup L_s(A^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)), \tag{SP$_1$}$$
$$\mathbf{u}^{n+1} = \mathrm{argmin}\left\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \mathrm{supp}(\mathbf{z}) \subseteq U^{n+1}\right\}, \tag{SP$_2$}$$
$$S^{n+1} = L_s(\mathbf{u}^{n+1}), \tag{SP$_3$}$$
$$\mathbf{x}^{n+1} = \mathrm{argmin}\left\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \mathrm{supp}(\mathbf{z}) \subseteq S^{n+1}\right\}, \tag{SP$_4$}$$

Output: the s-sparse vector $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$.

The thresholding-based family also contains algorithms that do not require an estimation of the sparsity s. In such algorithms, the hard thresholding operator gives way to a *soft thresholding operator* with threshold $\tau > 0$. This

operator, also encountered in (15.20) and (B.17), acts componentwise on a vector $\mathbf{z} \in \mathbb{C}^N$ by sending the entry z_j to

$$S_\tau(z_j) = \begin{cases} \operatorname{sgn}(z_j)(|z_j| - \tau) & \text{if } |z_j| \geq \tau \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

Another important method for sparse recovery is the message-passing algorithm studied by D. Donoho, A. Maleki, and A. Montanari in [128]. The soft thresholding algorithms will not be analyzed in this book.

Which algorithm should one choose? In principle, all the introduced algorithms work reasonably well in practice (with the possible exception of basic thresholding on which we comment below). In the end it depends on the precise situation, that is, on the specific measurement matrix $\mathbf{A}$ and on the values of the parameters s, m, N, which algorithm is the best for the given requirements. The minimal number of needed measurements m in terms of the sparsity s and the signal length N may vary slightly for the different algorithms. It is a matter of numerical tests to compare the recovery rates and to identify the best algorithm for this criterion.

The second criterion is the speed of the algorithm. Comparing this parameter is also a matter of numerical tests, but one can give at least the following rough guidelines. If the sparsity s is very small then orthogonal matching pursuit is very fast because the speed essentially depends on the number of iterations, which is s if the algorithm succeeds. However, if the sparsity s is of rather medium size compared to N then orthogonal matching pursuit may require significant time. The same consideration applies to the homotopy method for ℓ_1-minimization studied in Chapter 15, which builds up the support set of the minimizer iteratively.

Also compressive sampling matching pursuit and hard thresholding pursuit are fast for small s because in each step the orthogonal projections have to be computed for $\mathbf{A}_S$ with small $S \subset [N]$. The number of iterations may rather be independent of s. The runtime of iterative hard thresholding is not very much influenced by the sparsity s at all.

Basis Pursuit is not an algorithm per se, and its runtime depends on the used recovery algorithm. Chambolle and Pock's primal dual algorithm to be studied in Chapter 15 constructs a sequence $\mathbf{x}^n$, which converges to the ℓ_1-minimizer. Here, the sparsity s has no significant influence on the speed. Hence, for mildly large s this algorithm can be significantly faster than for instance orthogonal matching pursuit (we emphasize this point here because one often reads in the literature that greedy algorithms are always faster than ℓ_1-minimization, which however is only true for very small sparsity). Also, the iteratively reweighted least squares method studied in Chapter 15 may be a good alternative for mildly large sparsity.

As an important additional aspect one should consider whether the algorithms allow to easily exploit fast matrix vector multiplication routines if such are available for $\mathbf{A}$ and $\mathbf{A}^*$. In principle, one can speed up any of the

proposed methods in this case, but if an orthogonal projection step is involved then this task may not be completely trivial. For the iterative hard thresholding algorithm and for Chambolle and Pock's primal dual algorithm for ℓ_1-minimization, it is however very easy to exploit fast matrix vector multiplication. The acceleration achieved by the various algorithms in this context may actually differ, and again the fastest algorithm should be determined by numerical tests in the specific scenario.

Finally, basic thresholding is the fastest among all algorithms because it identifies the support in only one step, but its recovery performance is usually significantly worse than for the other algorithms.

Exercises

3.1. Let $q > 1$, and let $\mathbf{A}$ be an $m \times N$ matrix with $m < N$. Prove that there exists a 1-sparse vector which is not a minimizer of (P_q).

3.2. Using the matrix $\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$ and the vector $\mathbf{x} = [1, e^{i2\pi/3}, e^{i4\pi/3}]^\top$, prove that in the complex case a unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{y}$ is not necessarily m-sparse, where m is the number of rows of $\mathbf{A}$.

3.3. Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ and $\mathbf{y} \in \mathbb{R}^m$. Assuming the uniqueness of the minimizer $\mathbf{x}^\sharp$ of

$$\underset{\mathbf{z} \in \mathbb{R}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\| \leq \eta,$$

where $\eta \geq 0$ and $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^m$, prove that $\mathbf{x}^\sharp$ is necessarily m-sparse.

3.4. Given $\mathbf{A} \in \mathbb{R}^{m \times N}$, suppose that every $m \times m$ submatrix of $\mathbf{A}$ is invertible. For $\mathbf{x} \in \mathbb{R}^N$, let $\mathbf{x}^\sharp$ be the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. Prove that either $\mathbf{x}^\sharp = \mathbf{x}$ or $\mathrm{supp}(\mathbf{x}) \not\subseteq \mathrm{supp}(\mathbf{x}^\sharp)$.

3.5. For $\mathbf{A} \in \mathbb{R}^{m \times N}$ and $\mathbf{x} \in \mathbb{R}^N$, prove that there is no ambiguity between $\mathbf{z} \in \mathbb{R}^N$ and $\mathbf{z} \in \mathbb{C}^N$ when one says that the vector $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$.

3.6. Carefully check the equivalences of (P_1) with (P'_1) and $(\mathrm{P}_{1,\eta})$ with $(\mathrm{P}'_{1,\eta})$.

3.7. Given $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $\tau > 0$, show that the solution of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2^2 + \tau \|\mathbf{z}\|_2^2$$

is given by

$$\mathbf{z}^\sharp = (\mathbf{A}^* \mathbf{A} + \tau \mathbf{Id})^{-1} \mathbf{A}^* \mathbf{y}.$$

3.8. Given $\mathbf{A} \in \mathbb{C}^{m\times N}$, suppose that there is a unique minimizer $f(\mathbf{y}) \in \mathbb{C}^N$ of $\|\mathbf{z}\|_1$ subject to $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta$ whenever $\mathbf{y}$ belongs to some set $\mathcal{S}$. Prove that the map f is continuous on $\mathcal{S}$.

3.9. Prove that any 1-sparse vector $\mathbf{x} \in \mathbb{C}^3$ is recovered with one iteration of the orthogonal matching pursuit algorithm for the measurement matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix}.$$

We now add a measurement by appending the row [1 3 3] to $\mathbf{A}$, thus forming the matrix

$$\widehat{A} = \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \\ 1 & 3 & 3 \end{bmatrix}.$$

Prove that the 1-sparse vector $\mathbf{x} = [1\ 0\ 0]^\top$ cannot be recovered via the orthogonal matching pursuit algorithm with the measurement matrix $\widehat{A}$.

3.10. Given a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ with ℓ_2-normalized columns $\mathbf{a}_1, \dots, \mathbf{a}_N$ and given a vector $\mathbf{y} \in \mathbb{C}^m$, we consider an iterative algorithm where the index set S^n is updated via $S^{n+1} = S^n \cup \{j^{n+1}\}$ for an unspecified index j^{n+1} and where the output vector is updated via $\mathbf{x}^{n+1} = \operatorname{argmin}\{\|y - \mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \in S^{n+1}\}$. Prove that the ℓ_2-norm of the residual decreases according to

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}\|_2^2 \le \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \Delta_n,$$

where the quantity Δ_n satisfies

$$\begin{aligned} \Delta_n &= \|\mathbf{A}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|_2^2 = x^{n+1}_{j^{n+1}} (\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{j^{n+1}} \\ &= \frac{|(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{j^{n+1}}|^2}{\operatorname{dist}(\mathbf{a}_{j^{n+1}}, \operatorname{span}\{\mathbf{a}_j, j \in S^n\})^2} \\ &\le |(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{j^{n+1}}|^2. \end{aligned}$$

4

Basis Pursuit

Recall that the intuitive approach to the compressive sensing problem of recovering a sparse vector $\mathbf{x} \in \mathbb{C}^N$ from its measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$, where $m < N$, consists in the ℓ_0-minimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_0 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_0}$$

We have seen in Chapter 2 that this problem is unfortunately NP-hard in general. Chapter 3 has therefore outlined several tractable strategies to solve the compressive sensing problem. In the current chapter, we focus on the basis pursuit (ℓ_1-minimization) strategy, which consists in solving the convex optimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_1}$$

We investigate conditions on the matrix $\mathbf{A}$ which ensure exact or approximate reconstruction of the original sparse or compressible vector $\mathbf{x}$. In Section 4.1, we start with a necessary and sufficient condition for the exact reconstruction of every sparse vector $\mathbf{x} \in \mathbb{C}^N$ as a solution of (P_1) with the vector $\mathbf{y} \in \mathbb{C}^m$ obtained as $\mathbf{y} = \mathbf{A}\mathbf{x}$. This condition is called the null space property. In Sections 4.2 and 4.3, we strengthen this null space property to make the reconstruction via basis pursuit stable with respect to sparsity defect and robust with respect to measurement error. In Section 4.4, we discuss other types of necessary and sufficient conditions for the success of recovery via basis pursuit, which also take into account the given individual sparse vector. While this may seem useless at first sight because the vector $\mathbf{x}$ is unknown a-priori, such conditions will become nevertheless useful later to establish so-called nonuniform recovery guarantees in situations, where the matrix $\mathbf{A}$ is random. We close this Chapter with a short digression to the low-rank recovery problem, and its approach via nuclear norm minimization. Again, a version of the null space property is equivalent to recovery of every low-rank matrix.

4.1 Null Space Property

In this section, we introduce the null space property and we prove that it is a necessary and sufficient condition for exact recovery of sparse vectors via basis pursuit. The arguments are valid in the real and complex settings alike, so we first state the results for a field $\mathbb{K}$ that can either be $\mathbb{R}$ or $\mathbb{C}$. Then we establish the equivalence of the real and complex null space properties. We recall that for a vector $\mathbf{v} \in \mathbb{C}^N$ and a set $S \subset [N]$ we denote by $\mathbf{v}_S$ either the vector in $\mathbb{C}^S$, which is the restriction of $\mathbf{v}$ to the indices in S, or the vector in $\mathbb{C}^N$ which coincides with $\mathbf{v}$ on the indices in S and is extended to zero outside S, see also (2.4). It should always become clear from the context which variant of $\mathbf{v}_S$ is meant (and sometimes both variants lead to the same quantity, such as in expressions like $\|\mathbf{v}_S\|_1$).

Definition 4.1. *A matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$ is said to satisfy the* null space property *relative to a set $S \subset [N]$ if*

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\}. \tag{4.1}$$

It is said to satisfy the null space property of order s if it satisfies the null space property relative to any set $S \subset [N]$ with $\operatorname{card}(S) \le s$.

Remark 4.2. It is important to observe that, for a given $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, the condition $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1$ holds for any set $S \subseteq [N]$ with $\operatorname{card}(S) \le s$ as soon as it holds for an index set of s largest (in modulus) entries of $\mathbf{v}$.

Remark 4.3. There are two convenient reformulations of the null space property. The first one is obtained by adding $\|\mathbf{v}_S\|_1$ to both sides of the inequality $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1$. Thus, the null space property relative to S reads

$$2\,\|\mathbf{v}_S\|_1 < \|\mathbf{v}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\}. \tag{4.2}$$

The second one is obtained by choosing S as an index set of s largest (in modulus) entries of $\mathbf{v}$ and this time by adding $\|\mathbf{v}_{\overline{S}}\|_1$ to both sides of the inequality. Thus, the null space property of order s reads

$$\|\mathbf{v}\|_1 < 2\,\sigma_s(\mathbf{v})_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\}, \tag{4.3}$$

where we recall from Definition 2.2 that, for $p > 0$, the ℓ_p-error of *best s-term approximation* to $\mathbf{x} \in \mathbb{K}^N$ is defined by

$$\sigma_s(\mathbf{x})_p = \inf_{\|\mathbf{z}\|_0 \le s} \|\mathbf{x} - \mathbf{z}\|_p.$$

We now indicate the link between null space property and exact recovery of sparse vectors via basis pursuit.

Theorem 4.4. *Given a matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$, every vector $\mathbf{x} \in \mathbb{K}^N$ supported on a set S is the unique solution of* (P_1) *with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if $\mathbf{A}$ satisfies the null space property relative to S.*

Proof. Given a fixed index set S, let us first assume that every vector $\mathbf{x} \in \mathbb{K}^N$ supported on S is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$. Thus, for any $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, the vector $\mathbf{v}_S$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Av}_S$. But we have $\mathbf{A}(-\mathbf{v}_{\overline{S}}) = \mathbf{Av}_S$ and $-\mathbf{v}_{\overline{S}} \neq \mathbf{v}_S$, because $\mathbf{A}(\mathbf{v}_{\overline{S}} + \mathbf{v}_S) = \mathbf{Av} = 0$ and $\mathbf{v} \neq 0$. We conclude that $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1$. This establishes the null space property relative to S.

Conversely, let us assume that the null space property relative to S holds. Then, given a vector $\mathbf{x} \in \mathbb{K}^N$ supported on S and a vector $\mathbf{z} \in \mathbb{K}^N$, $\mathbf{z} \neq \mathbf{x}$, satisfying $\mathbf{Az} = \mathbf{Ax}$, we consider the vector $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker \mathbf{A} \setminus \{0\}$. In view of the null space property, we obtain

$$\begin{aligned} \|\mathbf{x}\|_1 \le \|\mathbf{x} - \mathbf{z}_S\|_1 + \|\mathbf{z}_S\|_1 &= \|\mathbf{v}_S\|_1 + \|\mathbf{z}_S\|_1 \\ &< \|\mathbf{v}_{\overline{S}}\|_1 + \|\mathbf{z}_S\|_1 = \| - \mathbf{z}_{\overline{S}}\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{z}\|_1. \end{aligned}$$

This establishes the required minimality of $\|\mathbf{x}\|_1$. □

Letting the set S vary, we immediately obtain the following result as a consequence of Theorem 4.4.

Theorem 4.5. *Given a matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$, every s-sparse vector $\mathbf{x} \in \mathbb{K}^N$ is the unique solution of* (P_1) *with $\mathbf{y} = \mathbf{Ax}$ if and only if $\mathbf{A}$ satisfies the null space property of order s.*

Remark 4.6. (a) This theorem shows that for every $\mathbf{y} = \mathbf{Ax}$ with s-sparse $\mathbf{x}$ the ℓ_1-minimization strategy (P_1) actually solves the ℓ_0-minimization problem (P_0) when the null space property of order s holds. Indeed, assume that every s-sparse vector $\mathbf{x}$ is recovered via ℓ_1-minimization from $\mathbf{y} = \mathbf{Ax}$. Let $\mathbf{z}$ be the minimizer of the ℓ_0-minimization problem (P_0) with $\mathbf{y} = \mathbf{Ax}$ then $\|\mathbf{z}\|_0 \le \|\mathbf{x}\|_1$ so that also $\mathbf{z}$ is s-sparse. But since every s-sparse vector is the unique ℓ_1-minimizer it follows that $\mathbf{x} = \mathbf{z}$.

(b) It is desirable for any reconstruction scheme to preserve sparse recovery if some measurements are rescaled, reshuffled, or added. Basis Pursuit actually features such properties. Indeed, mathematically speaking, these operations consist in replacing the original measurement matrix $\mathbf{A}$ by new measurement matrices $\widehat{\mathbf{A}}$, or $\widetilde{\mathbf{A}}$ defined by

$$\widehat{\mathbf{A}} := \mathbf{GA}, \quad \text{where } \mathbf{G} \text{ is some invertible } m \times m \text{ matrix},$$
$$\widetilde{\mathbf{A}} := \begin{bmatrix} \mathbf{A} \\ \hline \mathbf{B} \end{bmatrix}, \quad \text{where } \mathbf{B} \text{ is some } m' \times N \text{ matrix}.$$

We observe that $\ker \widehat{\mathbf{A}} = \ker \mathbf{A}$ and $\ker \widetilde{\mathbf{A}} \subseteq \ker \mathbf{A}$, hence the null space property for the matrices $\widehat{\mathbf{A}}$ $\widetilde{\mathbf{A}}$ remains fulfilled if it is satisfied for the matrix $\mathbf{A}$. It is not true that the null space property remains valid if we multiply on the right by an invertibe matrix — see Exercise 4.2.

We close this section by inspecting the influence of the underlying field. Unifying the arguments by using $\mathbb{K}$ for either $\mathbb{R}$ or $\mathbb{C}$ had the advantage of

brevity, but it results in a potential ambiguity about null space properties. Indeed, we often encounter real-valued measurement matrices, and they can also be regarded as complex-valued matrices. Thus, for such $\mathbf{A} \in \mathbb{R}^{m \times N}$, the distinction between the real null space $\ker_{\mathbb{R}} \mathbf{A}$ and the complex null space $\ker_{\mathbb{C}} \mathbf{A} = \ker_{\mathbb{R}} \mathbf{A} + i \ker_{\mathbb{R}} \mathbf{A}$ leads, on the one hand, to the real null space property relative to a set S, namely

$$\sum_{j \in S} |v_j| < \sum_{\ell \in \overline{S}} |v_\ell| \quad \text{for all } \mathbf{v} \in \ker_{\mathbb{R}} \mathbf{A}, \mathbf{v} \neq 0, \tag{4.4}$$

and on the other hand, to the complex null space property relative to S, namely

$$\sum_{j \in S} \sqrt{v_j^2 + w_j^2} < \sum_{\ell \in \overline{S}} \sqrt{v_\ell^2 + w_\ell^2} \quad \text{for all } \mathbf{v}, \mathbf{w} \in \ker_{\mathbb{R}} \mathbf{A}, (\mathbf{v}, \mathbf{w}) \neq (0, 0). \tag{4.5}$$

We are going to show below that the real and complex versions are in fact equivalent. Therefore, there is no ambiguity when we say that a real measurement matrix allows the exact recovery of all sparse vectors via basis pursuit: these vectors can be interpreted as real or as complex vectors. This explains why we usually work in the complex setting.

Theorem 4.7. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, the real null space property* (4.4) *relative to a set S is equivalent to the complex null space property* (4.5) *relative to this set S.*

In particular, the real null space property of order s is equivalent to the complex null space property of order s.

Proof. We notice first that (4.4) immediately follows from (4.5) by setting $\mathbf{w} = 0$. So let us assume that (4.4) holds. We consider $\mathbf{v}, \mathbf{w} \in \ker_{\mathbb{R}} \mathbf{A}$ with $(\mathbf{v}, \mathbf{w}) \neq (0, 0)$. If $\mathbf{v}$ and $\mathbf{w}$ are linearly dependent, then the inequality $\sum_{j \in S} \sqrt{v_j^2 + w_j^2} < \sum_{\ell \in \overline{S}} \sqrt{v_\ell^2 + w_\ell^2}$ is clear, so we may suppose that they are linearly independent. Then $\mathbf{u} := \cos\theta\, \mathbf{v} + \sin\theta\, \mathbf{w} \in \ker_{\mathbb{R}} \mathbf{A}$ is nonzero, and (4.4) yields, for any $\theta \in \mathbb{R}$,

$$\sum_{j \in S} |\cos\theta\, v_j + \sin\theta\, w_j| < \sum_{\ell \in \overline{S}} |\cos\theta\, v_\ell + \sin\theta\, w_\ell|. \tag{4.6}$$

For each $k \in [N]$, we define $\theta_k \in [-\pi, \pi]$ by the equalities

$$v_k = \sqrt{v_k^2 + w_k^2}\, \cos\theta_k, \quad w_k = \sqrt{v_k^2 + w_k^2}\, \sin\theta_k,$$

so that (4.6) reads

$$\sum_{j \in S} \sqrt{v_j^2 + w_j^2}\, |\cos(\theta - \theta_j)| < \sum_{\ell \in \overline{S}} \sqrt{v_\ell^2 + w_\ell^2}\, |\cos(\theta - \theta_\ell)|.$$

We now integrate over $\theta \in [-\pi, \pi]$ to obtain

$$\sum_{j \in S} \sqrt{v_j^2 + w_j^2} \int_{-\pi}^{\pi} |\cos(\theta - \theta_j)| d\theta \, < \, \sum_{\ell \in \overline{S}} \sqrt{v_\ell^2 + w_\ell^2} \int_{-\pi}^{\pi} |\cos(\theta - \theta_\ell)| d\theta.$$

For the inequality $\sum_{j \in S} \sqrt{v_j^2 + w_j^2} < \sum_{\ell \in \overline{S}} \sqrt{v_\ell^2 + w_\ell^2}$, it remains to observe that

$$\int_{-\pi}^{\pi} |\cos(\theta - \theta')| d\theta$$

is a positive constant independent of $\theta' \in [-\pi, \pi]$ — namely 4. The proof is now complete. □

Nonconvex Minimization

Recall that the number of nonzero entries of a vector $\mathbf{z} \in \mathbb{C}^N$ is approximated by the qth power of its ℓ_q-quasinorm,

$$\sum_{j=0}^{N} |z_j|^q \underset{q \to 0}{\longrightarrow} \sum_{j=1}^{N} \mathbf{1}_{\{z_j \neq 0\}} = \|\mathbf{z}\|_0.$$

This observation suggests to replace the ℓ_0-minimization problem (P_0) by the optimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_q \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y}. \tag{P_q}$$

This optimization problem fails to recover even 1-sparse vectors for $q > 1$, see Exercise 3.1. For $0 < q < 1$, on the other hand, the optimization problem becomes nonconvex, and is even NP-hard, see Exercise 2.10. Thus, the case $q = 1$ might appear as the only important one. Nonetheless, the properties of the ℓ_q-minimization for $0 < q < 1$ can prove useful on theoretical questions. Our goal here is merely to justify the intuitive prediction that the problem (P_q) does not provide a worse approximation of the original problem (P_0) when q gets smaller. For this purpose, we need an analog of the null space property for $0 < q < 1$. The proof of our next result, left as Exercise 4.11, duplicates the proof of Theorem 4.4. It relies on the fact that the qth power of the ℓ_q-quasinorm satisfies the triangle inequality, see Exercise 2.1.

Theorem 4.8. *Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $0 < q \leq 1$, every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique solution of (P_q) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if, for any set $S \subseteq [N]$ with $\mathrm{card}(S) \leq s$,*

$$\|\mathbf{v}_S\|_q < \|\mathbf{v}_{\overline{S}}\|_q \quad \textit{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\}.$$

We can now prove that sparse recovery via ℓ_q-minimization implies sparse recovery via ℓ_p-minimization whenever $0 < p < q \leq 1$.

Theorem 4.9. *Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $0 < p < q \leq 1$, if every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique solution of (P_q) with $\mathbf{y} = \mathbf{A}\mathbf{x}$, then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is also the unique solution of (P_p) with $\mathbf{y} = \mathbf{A}\mathbf{x}$.*

Proof. According to Theorem 4.8, it is enough to prove that, if $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$ and if S is an index set of s largest absolute entries of $\mathbf{v}$, then

$$\sum_{j \in S} |v_j|^p < \sum_{\ell \in \overline{S}} |v_\ell|^p, \tag{4.7}$$

as soon as (4.7) holds with q in place of p. Indeed, if (4.7) holds for p then necessarily $\mathbf{v}_{\overline{S}} \neq 0$ since S is an index of largest absolute entries and $\mathbf{v} \neq 0$. The desired inequality (4.7) can therefore be rewritten as

$$\sum_{j \in S} \frac{1}{\sum\limits_{\ell \in \overline{S}} (|v_\ell|/|v_j|)^p} < 1. \tag{4.8}$$

Now observe that $|v_\ell|/|v_j| \leq 1$ for $\ell \in \overline{S}$ and $j \in S$. This makes the left-hand side of (4.8) a nondecreasing function of $0 < p \leq 1$. Hence, its value at $p < q$ does not exceed its value at q, which is less than one by hypothesis. This shows the validity of (4.7) and concludes the proof. □

4.2 Stability

The vectors we aim to recover via basis pursuit — or other schemes, for that matter — are sparse only in idealized situations. In more realistic scenarios, we can only claim that they are close to sparse vectors. In such cases, we would like to recover a vector $\mathbf{x} \in \mathbb{C}^N$ with an error controlled by its distance to s-sparse vectors. This property is usually referred to as the *stability* of the reconstruction scheme with respect to sparsity defect. We shall prove that the basis pursuit is stable under a slightly strengthened version of the null space property.

Definition 4.10. *A matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is said to satisfy the* stable null space property *with constant $0 < \rho < 1$ relative to a set $S \subset [N]$ if*

$$\|\mathbf{v}_S\|_1 \leq \rho \, \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A}.$$

It is said to satisfy the stable null space property of order s with constant $0 < \rho < 1$ if it satisfies the stable null space property with constant $0 < \rho < 1$ relative to any set $S \subset [N]$ with $\mathrm{card}(S) \leq s$.

The main stability result of this section reads as follows.

Theorem 4.11. *Suppose that a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies the stable null space property of order s with constant $0 < \rho < 1$. Then, for any $\mathbf{x} \in \mathbb{C}^N$, a solution $\mathbf{x}^\sharp$ of (P_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ approximates the vector $\mathbf{x}$ with ℓ_1-error*

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \leq \frac{2(1+\rho)}{(1-\rho)}\, \sigma_s(\mathbf{x})_1. \tag{4.9}$$

Remark 4.12. In contrast to Theorem 4.4 we cannot guarantee uniqueness of the ℓ_1-minimizer anymore – although non-uniqueness is rather pathological. In any case, even when the ℓ_1-minimizer is not unique, the theorem above states that *every* solution $\mathbf{x}^\sharp$ of (P_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ satisfies (4.9).

We are actually going to prove a stronger 'if and only if' theorem below. The result is a statement valid for any index set S in which the vector $\mathbf{x}^\star \in \mathbb{C}^N$ is replaced by any vector $\mathbf{z} \in \mathbb{C}^N$ satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. Apart from improving Theorem 4.11, the result also says that, under the stable null space property relative to S, the distance between a vector $\mathbf{x} \in \mathbb{C}^N$ supported on S and a vector $\mathbf{z} \in \mathbb{C}^N$ satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$ is controlled by the difference between their norms.

Theorem 4.13. *The matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies the stable null space property with constant $0 < \rho < 1$ relative to S if and only if*

$$\|\mathbf{z} - \mathbf{x}\|_1 \leq \frac{1+\rho}{1-\rho}\left(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\,\|\mathbf{x}_{\overline{S}}\|_1\right) \tag{4.10}$$

for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$.

The error bound (4.9) follows from Theorem 4.13 as follows: Take S to be a set of s largest absolute coefficients of $\mathbf{x}$. Then $\|\mathbf{x}_{\overline{S}}\|_1 = \sigma_s(\mathbf{x})_1$. If $\mathbf{x}^\sharp$ is a minimizer of (P_1) then $\|\mathbf{x}^\sharp\|_1 \leq \|\mathbf{x}\|_1$ because $\mathbf{A}\mathbf{x}^\sharp = \mathbf{A}\mathbf{x}$. The right hand side of inequality (4.10) with $\mathbf{z} = \mathbf{x}^\sharp$ can therefore be estimated by the right hand of (4.9).

Before turning to the proof of Theorem 4.13, we isolate the following observation, as it will also be needed later.

Lemma 4.14. *Given a set $S \subset [N]$ and vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$,*

$$\|(\mathbf{x}-\mathbf{z})_{\overline{S}}\|_1 \leq \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \|(\mathbf{x}-\mathbf{z})_S\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1.$$

Proof. The result simply follows from

$$\begin{aligned} \|\mathbf{x}\|_1 = \|\mathbf{x}_{\overline{S}}\|_1 + \|\mathbf{x}_S\|_1 &\leq \|\mathbf{x}_{\overline{S}}\|_1 + \|(\mathbf{x}-\mathbf{z})_S\|_1 + \|\mathbf{z}_S\|_1, \\ \|(\mathbf{x}-\mathbf{z})_{\overline{S}}\|_1 &\leq \|\mathbf{x}_{\overline{S}}\|_1 + \|\mathbf{z}_{\overline{S}}\|_1. \end{aligned}$$

These two inequalities sum up to give

$$\|\mathbf{x}\|_1 + \|(\mathbf{x}-\mathbf{z})_{\overline{S}}\|_1 \leq 2\|\mathbf{x}_{\overline{S}}\|_1 + \|(\mathbf{x}-\mathbf{z})_S\|_1 + \|\mathbf{z}\|_1.$$

This is the desired inequality. □

Proof (of Theorem 4.13). Let us first assume that the matrix $\mathbf{A}$ satisfies (4.10) for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $\mathbf{Az} = \mathbf{Ax}$. Given a vector $\mathbf{v} \in \ker \mathbf{A}$, since $\mathbf{A}\mathbf{v}_{\overline{S}} = \mathbf{A}(-\mathbf{v}_S)$, we can apply (4.10) with $\mathbf{x} = -\mathbf{v}_S$ and $\mathbf{z} = \mathbf{v}_{\overline{S}}$. It yields

$$\|\mathbf{v}\|_1 \le \frac{1+\rho}{1-\rho}\big(\|\mathbf{v}_{\overline{S}}\|_1 - \|\mathbf{v}_S\|_1\big).$$

This can be written as

$$(1-\rho)\big(\|\mathbf{v}_S\|_1 + \|\mathbf{v}_{\overline{S}}\|_1\big) \le (1+\rho)\big(\|\mathbf{v}_{\overline{S}}\|_1 - \|\mathbf{v}_S\|_1\big).$$

After rearranging the terms we obtain

$$2\|\mathbf{v}_S\|_1 \le 2\rho\|\mathbf{v}_{\overline{S}}\|_1,$$

and simplifying by 2, we recognize the stable null space property with constant $0 < \rho < 1$ relative to S.

Conversely, let us now assume that the matrix $\mathbf{A}$ satisfies the stable null space property with constant $0 < \rho < 1$ relative to S. For $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $\mathbf{Az} = \mathbf{Ax}$, since $\mathbf{v} := \mathbf{z} - \mathbf{x} \in \ker \mathbf{A}$, the stable null space property yields

$$\|\mathbf{v}_S\|_1 \le \rho\|\mathbf{v}_{\overline{S}}\|_1. \tag{4.11}$$

Moreover, Lemma 4.14 gives

$$\|\mathbf{v}_{\overline{S}}\|_1 \le \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \|\mathbf{v}_S\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1. \tag{4.12}$$

Substituting (4.11) into (4.12), we obtain

$$\|\mathbf{v}_{\overline{S}}\|_1 \le \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \rho\|\mathbf{v}_{\overline{S}}\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1.$$

Since $\rho < 1$, this can be rewritten as

$$\|\mathbf{v}_{\overline{S}}\|_1 \le \frac{1}{1-\rho}\big(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1\big).$$

Using (4.11) once again, we derive

$$\|\mathbf{v}\|_1 = \|\mathbf{v}_{\overline{S}}\|_1 + \|\mathbf{v}_S\|_1 \le (1+\rho)\|\mathbf{v}_{\overline{S}}\|_1 \le \frac{1+\rho}{1-\rho}\big(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1\big),$$

which is the desired inequality. □

Remark 4.15. Given the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$, let us consider, for each index set $S \subset [N]$ with $\operatorname{card}(S) \le s$, the operator $\mathbf{R}_S$ defined on $\ker \mathbf{A}$ by $\mathbf{R}_S(\mathbf{v}) = \mathbf{v}_S$. The formulation (4.2) of the null space property says that

$$\mu := \max\{\|\mathbf{R}_S\|_{1\to 1} : S \subset [N], \operatorname{card}(S) \le s\} < 1/2.$$

We conclude that $\mathbf{A}$ satisfies then the stable null space property with constant $\rho := \mu/(1-\mu) < 1$. Thus, the stability of the basis pursuit comes for free if sparse vectors are exactly recovered. However, the constant $2(1+\rho)/(1-\rho)$ in (4.9) may be very large if ρ is close to one.

4.3 Robustness

In realistic situations, it is also inconceivable to measure a signal $\mathbf{x} \in \mathbb{C}^N$ with infinite precision. This means that the measurement vector $\mathbf{y} \in \mathbb{C}^m$ is only an approximation of the vector $\mathbf{A}\mathbf{x} \in \mathbb{C}^m$, with

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\| \leq \eta$$

for some $\eta \geq 0$ and for some norm $\|\cdot\|$ on $\mathbb{C}^m$ — usually the ℓ_2-norm, but the ℓ_1-norm will also be considered in Chapter 14. In this case, the reconstruction scheme should be required to output a vector $\mathbf{x}^\star \in \mathbb{C}^N$ whose distance to the original vector $\mathbf{x} \in \mathbb{C}^N$ is controlled by the measurement error $\eta \geq 0$. This property is usually referred to as the *robustness* of the reconstruction scheme with respect to measurement error. We are going to show that if the problem (P_1) is replaced by the convex optimization problem

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\| \leq \eta, \tag{$\mathrm{P}_{1,\eta}$}$$

then the robustness of the basis pursuit algorithm is guaranteed by the following additional strengthening of the null space property.

Definition 4.16. *The matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is said to satisfy the* robust null space property *(with respect to $\|\cdot\|$) with constants $0 < \rho < 1$ and $\tau > 0$ relative to a set $S \subseteq [N]$ if*

$$\|\mathbf{v}_S\|_1 \leq \rho \, \|\mathbf{v}_{\overline{S}}\|_1 + \tau \, \|\mathbf{A}\mathbf{v}\| \quad \text{for all } \mathbf{v} \in \mathbb{C}^N. \tag{4.13}$$

It is said to satisfy the robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ if it satisfies the robust null space property with constants ρ, τ relative to any set $S \subset [N]$ with $\mathrm{card}(S) \leq s$.

Remark 4.17. Observe that the above definition does not require that $\mathbf{v}$ is contained in $\ker \mathbf{A}$. In fact, if $\mathbf{v} \in \ker \mathbf{A}$ then the term $\|\mathbf{A}\mathbf{v}\|$ in (4.13) vanishes, and we see that the robust null space property implies the stable null space property in Definition 4.10.

The following theorem constitutes the first main result of this section. It incorporates Theorem 4.11 as the special case $\eta = 0$. The special case of an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is also worth a separate look.

Theorem 4.18. *Suppose that a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{x} \in \mathbb{C}^N$, a solution $\mathbf{x}^\sharp$ of $(\mathrm{P}_{1,\eta})$ with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ and $\|\mathbf{e}\| \leq \eta$ approximates the vector $\mathbf{x}$ with ℓ_1-error*

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \leq \frac{2(1+\rho)}{(1-\rho)} \, \sigma_s(\mathbf{x})_1 + \frac{4\tau}{1-\rho} \, \eta.$$

In the spirit of Theorem 4.13, we are going to prove a stronger 'if and only if' statement valid for any index set S.

Theorem 4.19. *The matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to S if and only if*

$$\|\mathbf{z} - \mathbf{x}\|_1 \leq \frac{1+\rho}{1-\rho} \left(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\,\|\mathbf{x}_{\overline{S}}\|_1 \right) + \frac{2\tau}{1-\rho} \|\mathbf{A}(\mathbf{z} - \mathbf{x})\| \qquad (4.14)$$

for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$.

Proof. We basically follow the same steps as in the proof of Theorem 4.13. First, we assume that the matrix $\mathbf{A}$ satisfies (4.14) for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$. Thus, for $\mathbf{v} \in \mathbb{C}^N$ taking $\mathbf{x} = -\mathbf{v}_S$ and $\mathbf{z} = \mathbf{v}_{\overline{S}}$ yields

$$\|\mathbf{v}\|_1 \leq \frac{1+\rho}{1-\rho} \big(\|\mathbf{v}_{\overline{S}}\|_1 - \|\mathbf{v}_S\|_1 \big) + \frac{2\tau}{1-\rho} \|\mathbf{A}\mathbf{v}\|.$$

Rearranging the terms gives

$$(1-\rho) \big(\|\mathbf{v}_S\|_1 + \|\mathbf{v}_{\overline{S}}\|_1 \big) \leq (1+\rho) \big(\|\mathbf{v}_{\overline{S}}\|_1 - \|\mathbf{v}_S\|_1 \big) + 2\tau \|\mathbf{A}\mathbf{v}\|,$$

that is to say

$$2\|\mathbf{v}_S\|_1 \leq 2\rho \|\mathbf{v}_{\overline{S}}\|_1 + 2\tau \|\mathbf{A}\mathbf{v}\|.$$

Except for the factor 2, this is the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to S.

Conversely, we assume that the matrix $\mathbf{A}$ satisfies the robust null space property with constant $0 < \rho < 1$ and $\tau > 0$ relative to S. For $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$, setting $\mathbf{v} := \mathbf{z} - \mathbf{x}$, the robust null space property and Lemma 4.14 yield

$$\begin{aligned} \|\mathbf{v}_S\|_1 &\leq \rho \|\mathbf{v}_{\overline{S}}\|_1 + \tau \|\mathbf{A}\mathbf{v}\|, \\ \|\mathbf{v}_{\overline{S}}\|_1 &\leq \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \|\mathbf{v}_S\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1. \end{aligned}$$

Combining these two inequalities gives

$$\|\mathbf{v}_{\overline{S}}\|_1 \leq \frac{1}{1-\rho} \big(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1 + \tau \|\mathbf{A}\mathbf{v}\| \big).$$

Using the robust null space property once again, we derive

$$\begin{aligned} \|\mathbf{v}\|_1 &= \|\mathbf{v}_{\overline{S}}\|_1 + \|\mathbf{v}_S\|_1 \leq (1+\rho) \|\mathbf{v}_{\overline{S}}\|_1 + \tau \|\mathbf{A}\mathbf{v}\|_1 \\ &\leq \frac{1+\rho}{1-\rho} \big(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1 \big) + \frac{2\tau}{1-\rho} \|\mathbf{A}\mathbf{v}\|, \end{aligned}$$

which is the desired inequality. □

We now turn to the second main result of this section. It enhances the previous robustness result by replacing the ℓ_1-error estimate by an ℓ_p-error estimate for $p \geq 1$. A final strengthening of the null space property is required. The corresponding property could be defined relative to any fixed set $S \subset [N]$, but it is not introduced as such because this is not be needed later.

Definition 4.20. *Given $q \geq 1$, the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is said to satisfy the* ℓ_q-robust null space property *of order s (with respect to $\|\cdot\|$) with constants $0 < \rho < 1$ and $\tau > 0$ if, for any set $S \subset [N]$ with* $\mathrm{card}(S) \leq s$,

$$\|\mathbf{v}_S\|_q \leq \frac{\rho}{s^{1-1/q}} \|\mathbf{v}_{\overline{S}}\|_1 + \tau \|\mathbf{A}\mathbf{v}\| \quad \text{for all } \mathbf{v} \in \mathbb{C}^N.$$

In view of the inequality $\|\mathbf{v}_S\|_p \leq s^{1/p-1/q}\|\mathbf{v}_S\|_q$ for $1 \leq p \leq q$, we observe that the ℓ_q-robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ implies that, for any set $S \subset [N]$ with $\mathrm{card}(S) \leq s$,

$$\|\mathbf{v}_S\|_p \leq \frac{\rho}{s^{1-1/p}} \|\mathbf{v}_{\overline{S}}\|_1 + \tau\, s^{1/p-1/q} \|\mathbf{A}\mathbf{v}\| \quad \text{for all } \mathbf{v} \in \mathbb{C}^N.$$

Thus, for $1 \leq p \leq q$, the ℓ_q-robust null space property implies the ℓ_p-robust null space property with identical constants, modulo the change of norms $\|\cdot\|_p \leftarrow s^{1/p-1/q}\|\cdot\|_q$. This justifies in particular that the ℓ_q-robust null space property is a strengthening of the previous robust null space property. In Section 6.2, we will establish the ℓ_2-robust null space property for measurement matrices with small restricted isometry constants. The robustness of the basis pursuit with noise algorithm is then deduced according to the following theorem.

Theorem 4.21. *Suppose that the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{x} \in \mathbb{C}^N$, a solution $\mathbf{x}^\sharp$ of $(\mathrm{P}_{1,\eta})$ with $\|\cdot\| = \|\cdot\|_2$, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, and $\|\mathbf{e}\|_2 \leq \eta$ approximates the vector $\mathbf{x}$ with ℓ_p-error*

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_p \leq \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1 + D\, s^{1/p-1/2}\, \eta, \qquad 1 \leq p \leq 2, \tag{4.15}$$

for some constants $C, D > 0$ depending only on ρ and τ.

The estimates for the extremal values $p = 1$ and $p = 2$ are the most familiar. They read

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^\sharp\|_1 &\leq C\, \sigma_s(\mathbf{x})_1 + D \sqrt{s}\, \eta, \\ \|\mathbf{x} - \mathbf{x}^\sharp\|_2 &\leq \frac{C}{\sqrt{s}}\, \sigma_s(\mathbf{x})_1 + D\, \eta. \end{aligned} \tag{4.16}$$

One should remember that the coefficient of $\sigma_s(\mathbf{x})_1$ is a constant for $p = 1$ and scales like $1/\sqrt{s}$ for $p = 2$, while the coefficient of η scales like $\sqrt{s}$ for $p = 1$ and is a constant for $p = 2$. We then retrieve the correct powers of s appearing in Theorem 4.21 for any $1 \leq p \leq 2$ via interpolating the powers of s with linear functions in $1/p$.

Remark 4.22. Let us shortly comment on the fact that the best s-term approximation error $\sigma_s(\mathbf{x})_1$ is always with respect to the ℓ_1-norm regardless of

the ℓ_p-space in which we measure the error. For instance, one may wonder why the error estimate in ℓ_2 does not feature $\sigma_s(\mathbf{x})_2$ on the right hand side instead of $\sigma_s(\mathbf{x})_1/\sqrt{s}$. In fact, we will see later in Theorem 11.5 that such a type of estimate is impossible in parameter regimes of (m, N) which are interesting for compressive sensing. On the other hand, we have seen in Chapter 2 that unit balls in ℓ_q with $q < 1$ provide good models for compressible vectors by Theorem 2.3 and its refinement Theorem 2.5. Indeed, if $\|\mathbf{x}\|_q \le 1$ for $q < 1$ then, for $p \ge 1$,

$$\sigma_s(\mathbf{x})_p \le s^{1/p-1/q} .$$

Assuming noiseless measurements (that is, $\eta = 0$) the error bound (4.15) reads then

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_p \le \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1 \le C s^{1/p-1/q} , \quad 1 \le p \le 2 .$$

Therefore, the reconstruction error in ℓ_p obeys the same rate of decay in s as the s-term approximation error in ℓ_p for all $p \in [1, 2]$. From this point of view, the term $s^{1/p-1}\sigma_s(\mathbf{x})_1$ is not significantly worse than $\sigma_s(\mathbf{x})_p$.

For the proof of Theorem 4.21, we in fact establish the following stronger result.

Theorem 4.23. *Given $1 \le p \le q$, suppose that the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies the ℓ_q-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$,*

$$\|\mathbf{z} - \mathbf{x}\|_p \le \frac{C}{s^{1-1/p}} \big(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\sigma_s(\mathbf{x})_1 \big) + D\, s^{1/p-1/q} \, \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|,$$

where $C := (1 + \rho)^2/(1 - \rho)$ and $D := (3 + \rho)\tau/(1 - \rho)$.

Proof. Let us first remark that the ℓ_q-robust null space properties implies the ℓ_1-robust and ℓ_p-robust null space property ($p \le q$) in the forms

$$\|\mathbf{v}_S\|_1 \le \rho \, \|\mathbf{v}_{\overline{S}}\|_1 + \tau \, s^{1-1/q} \, \|\mathbf{A}\mathbf{v}\|, \tag{4.17}$$

$$\|\mathbf{v}_S\|_p \le \frac{\rho}{s^{1-1/p}} \, \|\mathbf{v}_{\overline{S}}\|_1 + \tau \, s^{1/p-1/q} \, \|\mathbf{A}\mathbf{v}\|, \tag{4.18}$$

for all $\mathbf{v} \in \mathbb{C}^N$ and all $S \subset [N]$ with $\mathrm{card}(S) \le s$. Thus, in view of (4.17), applying Theorem 4.19 with S chosen as an index set of s largest (in modulus) entries of $\mathbf{x}$ leads to

$$\|\mathbf{z} - \mathbf{x}\|_1 \le \frac{1 + \rho}{1 - \rho} \big(\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\,\sigma_s(\mathbf{x})_1 \big) + \frac{2\tau}{1 - \rho} \, s^{1-1/q} \, \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|. \tag{4.19}$$

Then, choosing S as an index set of s largest (in modulus) entries of $\mathbf{z} - \mathbf{x}$, we use Proposition 2.5 to notice that

$$\|\mathbf{z} - \mathbf{x}\|_p \le \|(\mathbf{z} - \mathbf{x})_{\overline{S}}\|_p + \|(\mathbf{z} - \mathbf{x})_S\|_p \le \frac{1}{s^{1-1/p}} \|\mathbf{z} - \mathbf{x}\|_1 + \|(\mathbf{z} - \mathbf{x})_S\|_p.$$

In view of (4.18), we derive

$$\begin{aligned}\|\mathbf{z}-\mathbf{x}\|_p &\le \frac{1}{s^{1-1/p}}\|\mathbf{z}-\mathbf{x}\|_1 + \frac{\rho}{s^{1-1/p}}\,\|(\mathbf{z}-\mathbf{x})_{\overline{S}}\|_1 + \tau\, s^{1/p-1/q}\,\|\mathbf{A}(\mathbf{z}-\mathbf{x})\| \\ &\le \frac{1+\rho}{s^{1-1/p}}\,\|\mathbf{z}-\mathbf{x}\|_1 + \tau\, s^{1/p-1/q}\,\|\mathbf{A}(\mathbf{z}-\mathbf{x})\|. \end{aligned} \tag{4.20}$$

It remains to substitute (4.19) into the latter to obtain the desired result. □

Remark 4.24. The ℓ_q-robust null space property may seem mysterious at first sight, but it is necessary — save for the condition $\rho < 1$ — to obtain estimates of the type

$$\|\mathbf{x}-\mathbf{x}^\sharp\|_q \le \frac{C}{s^{1-1/q}}\,\sigma_s(\mathbf{x})_1 + D\,\eta, \tag{4.21}$$

where $\mathbf{x}^\sharp$ is a minimizer of $(\mathrm{P}_{1,\eta})$ with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ and $\|\mathbf{e}\| \le \eta$. Indeed, given $\mathbf{v} \in \mathbb{C}^N$ and $S \subset [N]$ with $\mathrm{card}(S) \le s$, we apply (4.21) with $\mathbf{x} = \mathbf{v}$, $\mathbf{e} = -\mathbf{A}\mathbf{v}$, and $\eta = \|\mathbf{A}\mathbf{v}\|$, so that $\mathbf{x}^\sharp = 0$, to obtain

$$\|\mathbf{v}\|_q \le \frac{C}{s^{1-1/q}}\,\|\mathbf{v}_{\overline{S}}\|_1 + D\,\|\mathbf{A}\mathbf{v}\|,$$

and in particular

$$\|\mathbf{v}_S\|_q \le \frac{C}{s^{1-1/q}}\,\|\mathbf{v}_{\overline{S}}\|_1 + D\,\|\mathbf{A}\mathbf{v}\|.$$

4.4 Recovery of Individual Vectors

In some cases, we deal with specific sparse vectors rather than with all vectors supported on a given set or all vectors with a given sparsity. We then require some recovery conditions that are finer then the null space property. This section provides such conditions, with a subtle difference between the real and the complex settings, due to the fact that the *sign* of a number a, defined as

$$\mathrm{sgn}(a) := \begin{cases} \dfrac{a}{|a|} & \text{if } a \neq 0, \\ \\ 0 & \text{if } a = 0, \end{cases}$$

is a discrete quantity when a is real, but a continuous quantity when a is complex. For a vector $\mathbf{x} \in \mathbb{C}^N$ we denote by $\mathrm{sgn}(\mathbf{x}) \in \mathbb{C}^N$ the vector with components $\mathrm{sgn}(x_j)$, $j \in [N]$. Let us start with the complex version of a recovery condition valid for individual sparse vectors.

Theorem 4.25. *Given a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$, a vector $\mathbf{x} \in \mathbb{C}^N$ with support S is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$ if one of the following equivalent conditions holds:*

(a) $\Big|\sum_{j\in S} \overline{\operatorname{sgn}(x_j)} v_j\Big| < \|\mathbf{v}_{\overline{S}}\|_1$ *for all* $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$,
(b) $\mathbf{A}_S$ *is injective and there exists a vector* $\mathbf{h} \in \mathbb{C}^m$ *such that*

$$(\mathbf{A}^*\mathbf{h})_j = \operatorname{sgn}(x_j), \quad j \in S, \qquad\qquad |(\mathbf{A}^*\mathbf{h})_\ell| < 1, \quad \ell \in \overline{S}.$$

Proof. Let us start by proving that (a) implies that $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. For a vector $\mathbf{z} \neq \mathbf{x}$ such that $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, we just have to write, with $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker \mathbf{A} \setminus \{0\}$,

$$\begin{aligned}\|\mathbf{z}\|_1 &= \|\mathbf{z}_S\|_1 + \|\mathbf{z}_{\overline{S}}\|_1 = \|(\mathbf{x}-\mathbf{v})_S\|_1 + \|\mathbf{v}_{\overline{S}}\|_1 \\ &> |\langle \mathbf{x}-\mathbf{v}, \operatorname{sgn}(\mathbf{x})_S\rangle| + |\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x})_S\rangle| \geq |\langle \mathbf{x}, \operatorname{sgn}(\mathbf{x})_S\rangle| = \|\mathbf{x}\|_1.\end{aligned}$$

The implication $(b) \Rightarrow (a)$ is also simple. Indeed, observing that $\mathbf{A}\mathbf{v}_S = -\mathbf{A}\mathbf{v}_{\overline{S}}$ for $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, we write

$$\begin{aligned}\Big|\sum_{j\in S} \overline{\operatorname{sgn}(x_j)} v_j\Big| &= |\langle \mathbf{v}_S, \mathbf{A}^*\mathbf{h}\rangle| = |\langle \mathbf{A}\mathbf{v}_S, \mathbf{h}\rangle| = |\langle \mathbf{A}\mathbf{v}_{\overline{S}}, \mathbf{h}\rangle| \\ &= |\langle \mathbf{v}_{\overline{S}}, \mathbf{A}^*\mathbf{h}\rangle| \leq \max_{\ell\in\overline{S}} |(\mathbf{A}^*\mathbf{h})_\ell| \, \|\mathbf{v}_{\overline{S}}\|_1 < \|\mathbf{v}_{\overline{S}}\|_1.\end{aligned}$$

The strict inequality holds since $\|\mathbf{v}_{\overline{S}}\|_1 > 0$, because otherwise the nonzero vector $\mathbf{v} \in \ker \mathbf{A}$ would be supported on S, contradicting the injectivity of $\mathbf{A}_S$.

The remaining implication $(a) \Rightarrow (b)$ requires more work. We start by noticing that (a) implies $\|\mathbf{v}_{\overline{S}}\|_1 > 0$ for all $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$. It follows that matrix $\mathbf{A}_S$ is injective. Indeed, assume $\mathbf{A}_S\mathbf{v}_S = 0$ for some $\mathbf{v}_S \neq 0$ and complete $\mathbf{v}_S$ to a vector $\mathbf{v} \in \mathbb{C}^N$ by setting $\mathbf{v}_{\overline{S}} = 0$. Then $\mathbf{v}$ is contained in $\ker \mathbf{A} \setminus \{0\}$, which is in contradiction with $\|\mathbf{v}_{\overline{S}}\|_1 > 0$ for all $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$. Next, since the continuous function $\mathbf{v} \mapsto |\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x})_S\rangle| / \|\mathbf{v}_{\overline{S}}\|_1$ takes values less than one on the unit sphere of $\ker \mathbf{A}$, which is compact, its maximum μ satisfies $\mu < 1$. By homogeneity, we deduce

$$|\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x})_S\rangle| \leq \mu \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A}.$$

We then define, for $\mu < \nu < 1$, the convex set $\mathcal{C}$ and the affine set $\mathcal{D}$ by

$$\begin{aligned}\mathcal{C} &:= \big\{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}_S\|_1 + \nu\|\mathbf{z}_{\overline{S}}\|_1 \leq \|\mathbf{x}\|_1\big\}, \\ \mathcal{D} &:= \big\{\mathbf{z} \in \mathbb{C}^N : \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}\big\}.\end{aligned}$$

The intersection $\mathcal{C} \cap \mathcal{D}$ reduces to $\{\mathbf{x}\}$. Indeed, we observe that $\mathbf{x} \in \mathcal{C} \cap \mathcal{D}$, and if $\mathbf{z} \neq \mathbf{x}$ belongs to $\mathcal{C} \cap \mathcal{D}$, setting $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker \mathbf{A} \setminus \{0\}$, we obtain a contradiction from

$$\begin{aligned}\|\mathbf{x}\|_1 &\geq \|\mathbf{z}_S\|_1 + \nu\|\mathbf{z}_{\overline{S}}\|_1 = \|(\mathbf{x}-\mathbf{v})_S\|_1 + \nu\|\mathbf{v}_{\overline{S}}\|_1 \\ &> \|(\mathbf{x}-\mathbf{v})_S\|_1 + \mu\|\mathbf{v}_{\overline{S}}\|_1 \geq |\langle \mathbf{x}-\mathbf{v}, \operatorname{sgn}(\mathbf{x})_S\rangle| + |\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x})_S\rangle| \\ &\geq |\langle \mathbf{x}, \operatorname{sgn}(\mathbf{x})_S\rangle| = \|\mathbf{x}\|_1.\end{aligned}$$

Thus, by the separation of convex sets via hyperplanes, see Theorem B.4 and Remark B.5, there exists a vector $\mathbf{w} \in \mathbb{C}^N$ such that

$$\mathcal{C} \subset \big\{\mathbf{z} \in \mathbb{C}^N : \operatorname{Re}\langle \mathbf{z}, \mathbf{w}\rangle \le \|\mathbf{x}\|_1\big\}, \tag{4.22}$$

$$\mathcal{D} \subset \big\{\mathbf{z} \in \mathbb{C}^N : \operatorname{Re}\langle \mathbf{z}, \mathbf{w}\rangle = \|\mathbf{x}\|_1\big\}. \tag{4.23}$$

In view of (4.22), we have

$$\begin{aligned}
\|\mathbf{x}\|_1 &\ge \max_{\|\mathbf{z}_S + \nu \mathbf{z}_{\overline{S}}\|_1 \le \|\mathbf{x}\|_1} \operatorname{Re}\langle \mathbf{z}, \mathbf{w}\rangle \\
&= \max_{\|\mathbf{z}_S + \nu \mathbf{z}_{\overline{S}}\|_1 \le \|\mathbf{x}\|_1} \operatorname{Re}\left(\sum_{j\in S} z_j \overline{w_j} + \sum_{j \in \overline{S}} \nu z_j \overline{w_j}/\nu\right) \\
&= \max_{\|\mathbf{z}_S + \nu \mathbf{z}_{\overline{S}}\|_1 \le \|\mathbf{x}\|_1} \operatorname{Re}\langle \mathbf{z}_S + \nu \mathbf{z}_{\overline{S}}, \mathbf{w}_S + (1/\nu)\mathbf{w}_{\overline{S}}\rangle \\
&= \|\mathbf{x}\|_1 \, \|\mathbf{w}_S + (1/\nu)\mathbf{w}_{\overline{S}}\|_\infty = \|\mathbf{x}\|_1 \, \max\{\|\mathbf{w}_S\|_\infty, (1/\nu)\|\mathbf{w}_{\overline{S}}\|_\infty\}.
\end{aligned}$$

We derive $\|\mathbf{w}_S\|_\infty \le 1$ and $\|\mathbf{w}_{\overline{S}}\|_\infty \le \nu < 1$. From (4.23), we derive $\operatorname{Re}\langle \mathbf{x}, \mathbf{w}\rangle = \|\mathbf{x}\|_1$, i.e., $w_j = \operatorname{sgn}(x_j)$ for all $j \in S$, and also $\operatorname{Re}\langle \mathbf{v}, \mathbf{w}\rangle = 0$ for all $\mathbf{v} \in \ker \mathbf{A}$, i.e., $\mathbf{w} \in (\ker \mathbf{A})^\perp$. Since $(\ker \mathbf{A})^\perp = \operatorname{ran} \mathbf{A}^*$, we write $\mathbf{w} = \mathbf{A}^*\mathbf{h}$ for some $\mathbf{h} \in \mathbb{C}^m$. This establishes (b). □

Remark 4.26. The previous theorem can be made stable under noise on the measurements and under passing to compressible vectors, see Exercise (4.16) and also compare with Theorem 4.32 below. However, the resulting error bounds are slightly weaker than the ones of Theorem 4.23 under the ℓ_2-robust null space property.

The equalities $(\mathbf{A}^*\mathbf{h})_j = \operatorname{sgn}(x_j)$, $j \in S$, considered in (ii) translate into $\mathbf{A}_S^*\mathbf{h} = \operatorname{sgn}(\mathbf{x}_S)$. This is satisfied for the choice $\mathbf{h} = \big(\mathbf{A}_S^\dagger\big)^* \operatorname{sgn}(\mathbf{x}_S)$, where the expression $\mathbf{A}_S^\dagger := (\mathbf{A}_S^*\mathbf{A}_S)^{-1}\mathbf{A}_S^*$ of the *Moore–Penrose pseudo-inverse* of $\mathbf{A}_S$ is justified by its injectivity, see Appendix (A.24). Since the conditions $|(\mathbf{A}^*\mathbf{h})_\ell| < 1$, $\ell \in \overline{S}$, then read $|\langle \mathbf{a}_\ell, \mathbf{h}\rangle| < 1$, $\ell \in \overline{S}$, where $\mathbf{a_1}, \dots, \mathbf{a}_N$ are the columns of $\mathbf{A}$, we can state the following result.

Corollary 4.27. *Let $\mathbf{a_1}, \dots, \mathbf{a}_N$ be the columns of $\mathbf{A} \in \mathbb{C}^{m\times N}$. For $\mathbf{x} \in \mathbb{C}^N$ with support S, if the matrix $\mathbf{A}_S$ is injective and if*

$$|\langle \mathbf{A}_S^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S)\rangle| < 1 \quad \text{for all } \ell \in \overline{S}, \tag{4.24}$$

then the vector $\mathbf{x}$ is the unique solution of (P_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Remark 4.28. In general, there is no converse to Theorem 4.25. Let us consider, for instance,

$$\mathbf{A} := \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} e^{-i\pi/3} \\ e^{i\pi/3} \\ 0 \end{bmatrix}.$$

We can verify that $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$, see Exercise 4.13. However, (a) fails. Indeed, for a vector $\mathbf{v} = [\zeta, \zeta, \zeta] \in \ker \mathbf{A} \setminus \{0\}$, we have $|\overline{\operatorname{sgn}(x_1)} v_1 + \overline{\operatorname{sgn}(x_2)} v_2| = |(e^{i\pi/3} + e^{-i\pi/3})\zeta| = |\zeta|$, while $\|\mathbf{v}_{\{3\}}\|_1 = |\zeta|$. In contrast, a converse to Theorem 4.25 holds in the real setting.

Theorem 4.29. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, a vector $\mathbf{x} \in \mathbb{R}^N$ with support S is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$ if and only if one of the following equivalent conditions holds:*

(a) $\Big| \sum_{j \in S} \operatorname{sgn}(x_j) v_j \Big| < \|\mathbf{v}_{\overline{S}}\|_1$ *for all* $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$,

(b) $\mathbf{A}_S$ *is injective and there exists a vector* $\mathbf{h} \in \mathbb{R}^m$ *such that*

$$(\mathbf{A}^\top \mathbf{h})_j = \operatorname{sgn}(x_j), \quad j \in S, \qquad\qquad |(\mathbf{A}^\top \mathbf{h})_\ell| < 1, \quad \ell \in \overline{S}.$$

Proof. The arguments given in the proof of Theorem 4.25 still hold in the real setting, hence it is enough to show that (a) holds as soon as $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$. In this situation, for $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, the vector $\mathbf{z} := \mathbf{x} - \mathbf{v}$ satisfies $\mathbf{z} \neq \mathbf{x}$ and $\mathbf{Az} = \mathbf{Ax}$, so that

$$\|\mathbf{x}\|_1 < \|\mathbf{z}\|_1 = \|\mathbf{z}_S\|_1 + \|\mathbf{z}_{\overline{S}}\|_1 = \langle \mathbf{z}, \operatorname{sgn}(\mathbf{z})_S \rangle + \|\mathbf{z}_{\overline{S}}\|_1.$$

Taking $\|\mathbf{x}\|_1 \geq \langle \mathbf{x}, \operatorname{sgn}(\mathbf{z})_S \rangle$ into account, we derive $\langle \mathbf{x} - \mathbf{z}, \operatorname{sgn}(\mathbf{z})_S \rangle < \|\mathbf{z}_{\overline{S}}\|_1$. Hence, we have

$$\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x} - \mathbf{v})_S \rangle < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\}.$$

Writing the latter for $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$ replaced by $t\mathbf{v}$, $t > 0$, and simplifying by t, we obtain

$$\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x} - t\mathbf{v})_S \rangle < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\} \text{ and all } t > 0.$$

Taking $t > 0$ small enough so that $\operatorname{sgn}(x_j - t v_j) = \operatorname{sgn}(x_j)$ — note that it is essential for $\mathbf{x}$ to be *exactly* supported on S — we conclude

$$\langle \mathbf{v}, \operatorname{sgn}(\mathbf{x})_S \rangle < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{0\},$$

which implies (a) by replacing $\mathbf{v}$ by $-\mathbf{v}$ if necessary. □

Remark 4.30. Theorem 4.29 shows that in the real setting the recovery of a given vector via basis pursuit depends only on its sign pattern, but not on the magnitude of its entries. Moreover, it shows that if a vector $\mathbf{x} \in \mathbb{R}^N$ with support S is exactly recovered via basis pursuit, then all vectors $\mathbf{x}' \in \mathbb{R}^N$ with support $S' \subset S$ and $\operatorname{sgn}(\mathbf{x}')_{S'} = \operatorname{sgn}(\mathbf{x})_{S'}$ are also exactly recovered via basis pursuit. Indeed, if (a) holds, then we have, for $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$,

$$\begin{aligned} \Big| \sum_{j \in S'} \operatorname{sgn}(x'_j) v_j \Big| = \Big| \sum_{j \in S} \operatorname{sgn}(x_j) v_j - \sum_{j \in S \setminus S'} \operatorname{sgn}(x_j) v_j \Big| &\leq \Big| \sum_{j \in S} \operatorname{sgn}(x_j) v_j \Big| + \sum_{j \in S \setminus S'} |v_j| \\ &< \|\mathbf{v}_{\overline{S}}\|_1 + \|\mathbf{v}_{\overline{S \setminus S'}}\|_1 = \|\mathbf{v}_{\overline{S'}}\|_1. \end{aligned}$$

It is not always straightforward to construct a "dual vector" $\mathbf{h}$ as described in property (ii) of Theorems 4.25 and 4.29. The following condition based on an "inexact dual vector" is sometimes easier to analyze.

Theorem 4.31. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ with columns $\mathbf{a}_\ell$, $\ell \in [N]$, and $\mathbf{x} \in \mathbb{C}^N$ with support S. Let $\alpha, \beta, \gamma, \eta > 0$, and assume that*

$$\|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\|_{2\to 2} \le \alpha \quad \text{and} \quad \max_{\ell\in\overline{S}} \|\mathbf{A}_S^*\mathbf{a}_\ell\|_2 \le \beta \,. \tag{4.25}$$

Suppose there exists a vector $\mathbf{u} \in \mathbb{C}^N$ of the form $\mathbf{u} = \mathbf{A}^*\mathbf{h}$ with $\mathbf{h} \in \mathbb{C}^m$ such that

$$\|\mathbf{u}_S - \operatorname{sgn}(\mathbf{x}_S)\|_2 \le \gamma \quad \text{and} \quad \|\mathbf{u}_{\overline{S}}\|_\infty \le \theta \,. \tag{4.26}$$

If $\theta + \alpha\beta\gamma < 1$ then $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$.

Proof. Let $\mathbf{x}^\sharp$ be a minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$. Then $\mathbf{v} = \mathbf{x}^\sharp - \mathbf{x}$ satisfies $\mathbf{Av} = 0$. We need to show that $\mathbf{v} = 0$. To this end we first observe that

$$\begin{aligned}\|\mathbf{x}^\sharp\|_1 &= \|\mathbf{x}_S + \mathbf{v}_S\|_1 + \|\mathbf{v}_{\overline{S}}\|_1 = \langle \operatorname{sgn}(\mathbf{x}_S + \mathbf{v}_S), \mathbf{x}_S + \mathbf{v}_S\rangle + \|\mathbf{v}_{\overline{S}}\|_1 \\ &\ge \operatorname{Re}\left(\langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{x}_S + \mathbf{v}_S\rangle\right) + \|\mathbf{v}_{\overline{S}}\|_1 \\ &= \|\mathbf{x}_S\|_1 + \operatorname{Re}\left(\langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{v}_S\rangle\right) + \|\mathbf{v}_{\overline{S}}\|_1 \,. \end{aligned} \tag{4.27}$$

For $\mathbf{u} = \mathbf{A}^*\mathbf{h}$ it holds

$$\langle \mathbf{u}_S, \mathbf{v}_S\rangle = \langle \mathbf{u}, \mathbf{v}\rangle - \langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle = \langle \mathbf{h}, \mathbf{Av}\rangle - \langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle = -\langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle \,.$$

Hence,

$$\begin{aligned}\langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{v}_S\rangle &= \langle \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S, \mathbf{v}_S\rangle + \langle \mathbf{u}_S, \mathbf{v}_S\rangle \\ &= \langle \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S, \mathbf{v}_S\rangle - \langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle \,.\end{aligned}$$

The Cauchy-Schwarz inequality together with (4.26) yields

$$|\operatorname{Re}\left(\langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{v}_S\rangle\right)| \le \|\operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2\|\mathbf{v}_S\|_2 + \|\mathbf{u}_{\overline{S}}\|_\infty\|\mathbf{v}_{\overline{S}}\|_1 \le \gamma\|\mathbf{v}_S\|_2 + \theta\|\mathbf{v}_{\overline{S}}\|_1 \,.$$

Together with (4.27), and since $\operatorname{supp}\mathbf{x} = S$, this gives

$$\|\mathbf{x}^\sharp\|_1 \ge \|\mathbf{x}\|_1 - \gamma\|\mathbf{v}_S\|_2 + (1-\theta)\|\mathbf{v}_{\overline{S}}\|_1 \,.$$

Next we bound $\|\mathbf{v}_S\|_2$. Since $\mathbf{Av} = 0$, we have $\mathbf{A}_S\mathbf{v}_S = -\mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}$ and

$$\begin{aligned}\|\mathbf{v}_S\|_2 &= \|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\mathbf{A}_S^*\mathbf{A}_S\mathbf{v}_S\|_2 = \| - (\mathbf{A}_S^*\mathbf{A}_S)^{-1}\mathbf{A}_S^*\mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}\|_2 \\ &\le \|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\|_{2\to 2}\|\mathbf{A}_S^*\mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}\|_2 \le \alpha\sum_{\ell\in\overline{S}} |\mathbf{v}_\ell|\|\mathbf{A}_S^*\mathbf{a}_\ell\|_2 \\ &\le \alpha\beta\|\mathbf{v}_{\overline{S}}\|_1 \,. \end{aligned} \tag{4.28}$$

Hereby we have used (4.25). We have derived

$$\|\widehat{\mathbf{x}}\|_1 \geq \|\mathbf{x}\|_1 - \gamma\|\mathbf{v}_S\|_2 + (1-\theta)\|\mathbf{v}_{\overline{S}}\|_1 \geq \|\mathbf{x}\|_1 + (1-\theta-\alpha\beta\gamma)\|\mathbf{v}_{\overline{S}}\|_1 \, .$$

Since $1-\theta-\alpha\beta\gamma > 0$ and $\mathbf{x}^\sharp$ is an ℓ_1-minimizer it follows that $\mathbf{v}_{\overline{S}} = 0$. Therefore, $\mathbf{A}_S\mathbf{v}_S = -\mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}} = 0$. Since $\mathbf{A}_S$ is injective (recall that $\mathbf{A}_S^*\mathbf{A}_S$ is invertible) it follows that $\mathbf{v}_S = 0$, so that $\mathbf{v} = 0$. □

The next statement makes the previous result stable robust under noise and under passing from sparse to compressible vectors. Due to the appearance of an additional factor of $\sqrt{s}$ in the error bound, however, is not as sharp as (4.16) obtained under the ℓ_2-robust null space property. Nevertheless, it applies under weaker conditions on $\mathbf{A}$ and is therefore still useful in certain scenarios, especially when the null space property (or the restricted isometry property to be studied in Chapter 6) is not known to hold or harder to prove, see also Chapter 12.

Theorem 4.32. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ and $\mathbf{x} \in \mathbb{C}^N$. Let $S \in [N]$ be the index set of the s largest absolute coefficients of $\mathbf{x}$. Assume that, for positive constants $\delta, \beta, \gamma, \theta \in (0,1)$ with $b := \theta + \beta\gamma/(1-\delta) < 1$ and $\kappa \geq 1$, the columns $\mathbf{a}_j$, $j \in [N]$, of $\mathbf{A}$ satisfy $\|\mathbf{a}_j\|_2 \leq \kappa$ and*

$$\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \leq \delta \tag{4.29}$$

$$\max_{\ell\in\overline{S}} \|\mathbf{A}_S^*\mathbf{a}_\ell\|_2 \leq \beta \, . \tag{4.30}$$

Suppose there exists a vector $\mathbf{u} \in \mathbb{C}^N$ of the form $\mathbf{u} = \mathbf{A}^*\mathbf{h}$ with $\mathbf{h} \in \mathbb{C}^m$ such that

$$\|\mathbf{u}_S - \operatorname{sgn}(\mathbf{x}_S)\|_2 \leq \gamma \, , \tag{4.31}$$

$$\|\mathbf{u}_{\overline{S}}\|_\infty \leq \theta \, , \tag{4.32}$$

$$\|\mathbf{h}\|_2 \leq \tau\sqrt{s} \, . \tag{4.33}$$

Let noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ be given with $\|\mathbf{e}\|_2 \leq \eta$. Then the minimizer $\mathbf{x}^\sharp$ of

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta \tag{4.34}$$

satisfies

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \leq C_1\sigma_s(\mathbf{x})_1 + (C_2 + C_3\sqrt{s})\eta \, ,$$

where

$$C_1 = \left(1 + \frac{\kappa\sqrt{1+\delta}}{1-\delta}\right)\frac{2}{1-b} \, , \quad C_2 = \left(1 + \frac{\kappa\sqrt{1+\delta}}{1-\delta}\right)\frac{2\gamma\sqrt{1+\delta}}{(1-b)(1-\delta)} \, ,$$

$$C_3 = \frac{2\kappa}{1-\delta} + \frac{2\tau}{1-b}\left(1 + \frac{\kappa\sqrt{1+\delta}}{1-\delta}\right) \, .$$

Remark 4.33. The statement of the above theorem can be made more specific by choosing concrete constants, for instance, $\kappa = \tau = 2$, $\delta = \beta = \gamma = 1/2$, $\theta = 1/4$ resulting in $b = 3/4$ and $C_1 \approx 47.19$, $C_2 \approx 57.79$, $C_3 \approx 102.38$.

Proof. The vector $\mathbf{x}$ is feasible for the optimization program (4.34) by the assumption on the noise level. Therefore, $\|\mathbf{x}\|_1 \geq \|\mathbf{x}^\sharp\|_1$ and writing $\mathbf{x}^\sharp = \mathbf{x} + \mathbf{v}$ we have

$$\begin{aligned}\|\mathbf{x}\|_1 &\geq \|\mathbf{x}+\mathbf{v}\|_1 = \|(\mathbf{x}+\mathbf{v})_S\|_1 + \|(\mathbf{x}+\mathbf{v})_{\overline{S}}\|_1 \\ &\geq \operatorname{Re}\left(\langle (\mathbf{x}+\mathbf{v})_S, \operatorname{sgn}(\mathbf{x})_S\rangle\right) + \|\mathbf{v}_{\overline{S}}\|_1 - \|\mathbf{x}_{\overline{S}}\|_1 \\ &= \|\mathbf{x}_S\|_1 + \operatorname{Re}\left(\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x})_S\rangle\right) + \|\mathbf{v}_{\overline{S}}\|_1 - \|\mathbf{x}_{\overline{S}}\|_1 \, .\end{aligned}$$

Rearranging and using that $\|\mathbf{x}\|_1 = \|\mathbf{x}_S\|_1 + \|\mathbf{x}_{\overline{S}}\|_1$ yields

$$\|\mathbf{v}_{\overline{S}}\|_1 \leq |\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x})_S\rangle| + 2\|\mathbf{x}_{\overline{S}}\|_1 \, . \tag{4.35}$$

The triangle inequality and the Cauchy-Schwarz inequality together with (4.31) yield

$$\begin{aligned}|\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x}_S)\rangle| &\leq |\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x})_S - \mathbf{u}_S\rangle| + |\langle \mathbf{v}_S, \mathbf{u}_S\rangle| \\ &\leq \gamma\|\mathbf{v}_S\|_2 + |\langle \mathbf{v}, \mathbf{u}\rangle| + |\langle \mathbf{v}_{\overline{S}}, \mathbf{u}_{\overline{S}}\rangle| \, .\end{aligned}$$

It follows from the assumption on the noise vector $\mathbf{e}$ and from the constraint of the optimization problem (4.34) that

$$\|\mathbf{A}\mathbf{v}\|_2 = \|\mathbf{A}(\mathbf{x}^\sharp - \mathbf{x})\|_2 \leq \|\mathbf{A}\mathbf{x}^\sharp - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq 2\eta \, . \tag{4.36}$$

The well-conditionedness assumption in (4.29) implies that $\|\mathbf{A}_S\|_{2\to 2} \leq \sqrt{1+\delta}$ and $\|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\|_{2\to 2} \leq (1-\delta)^{-1}$, see Theorem A.13 and Proposition A.16, so that

$$\begin{aligned}\|\mathbf{v}_S\|_2 &\leq \frac{1}{1-\delta}\|\mathbf{A}_S^*\mathbf{A}_S\mathbf{v}_S\|_2 \leq \frac{1}{1-\delta}\|\mathbf{A}_S^*\mathbf{A}\mathbf{v}\|_2 + \frac{1}{1-\delta}\|\mathbf{A}_S^*\mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}\|_2 \\ &\leq 2\frac{\sqrt{1+\delta}}{1-\delta}\eta + \frac{\beta}{1-\delta}\|\mathbf{v}_{\overline{S}}\|_1 \, ,\end{aligned}$$

where the inequality $\|\mathbf{A}_S^*\mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}\|_2 \leq \beta\|\mathbf{v}_{\overline{S}}\|_1$ follows from (4.30) in the same way as in (4.28). Condition (4.33) gives

$$|\langle \mathbf{v}, \mathbf{u}\rangle| = |\langle \mathbf{v}, \mathbf{A}^*\mathbf{h}\rangle| = |\langle \mathbf{A}\mathbf{v}, \mathbf{h}\rangle| \leq \|\mathbf{A}\mathbf{v}\|_2\|\mathbf{h}\|_2 \leq 2\tau\eta\sqrt{s} \, , \tag{4.37}$$

while (4.32) implies $|\langle \mathbf{v}_{\overline{S}}, \mathbf{u}_{\overline{S}}\rangle| \leq \theta\|\mathbf{v}_{\overline{S}}\|_1$. We plug these estimates into (4.35) to find that

$$\|\mathbf{v}_{\overline{S}}\|_1 \leq \left(2\gamma\frac{\sqrt{1+\delta}}{1-\delta} + 2\tau\sqrt{s}\right)\eta + \left(\theta + \frac{\beta\gamma}{1-\delta}\right)\|\mathbf{v}_{\overline{S}}\|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1 \, .$$

Since $\theta + \beta\gamma/(1-\delta) = b < 1$ and $\|\mathbf{x}_{\overline{S}}\|_1 = \sigma_s(\mathbf{x})_1$ a rearrangement yields

$$\|\mathbf{v}_{\overline{S}}\|_1 \leq \frac{2\gamma\frac{\sqrt{1+\delta}}{1-\delta}+2\tau\sqrt{s}}{1-b}\eta + \frac{2}{1-b}\sigma_s(\mathbf{x})_1 \ . \tag{4.38}$$

Let us now consider $\mathbf{v}_S$. Due to (4.29) we have

$$(1-\delta)\|\mathbf{v}_S\|_2^2 \leq \|\mathbf{A}_S\mathbf{v}_S\|_2^2 = \langle \mathbf{A}_S\mathbf{v}_S, \mathbf{A}\mathbf{v}\rangle - \langle \mathbf{A}_S\mathbf{v}_S, \mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}\rangle \ . \tag{4.39}$$

The first term above can be estimated as

$$\begin{aligned}|\langle \mathbf{A}_S\mathbf{v}_S, \mathbf{A}\mathbf{v}\rangle| &\leq \|\mathbf{v}_S\|_1\|\mathbf{A}_S^*\mathbf{A}\mathbf{v}\|_\infty \leq \sqrt{s}\|\mathbf{v}\|_2 \max_{j\in S}|\langle \mathbf{a}_j, \mathbf{A}\mathbf{v}\rangle| \\ &\leq \sqrt{s}\|\mathbf{v}\|_2 \max_{j\in S}\|\mathbf{a}_j\|_2\|\mathbf{A}\mathbf{v}\|_2 \leq 2\eta\kappa\sqrt{s}\|\mathbf{v}_S\|_2 \ ,\end{aligned}$$

where we have used (4.36) and the assumption that $\|\mathbf{a}_j\|_2 \leq \kappa$ for all $j \in [N]$. We bound the second term in (4.39) as

$$\begin{aligned}|\langle \mathbf{A}_S\mathbf{v}_S, \mathbf{A}_{\overline{S}}\mathbf{v}_{\overline{S}}\rangle| &\leq \sum_{j\in\overline{S}}|v_j||\langle \mathbf{A}_S\mathbf{v}_S, \mathbf{a}_j\rangle| \leq \sum_{j\in\overline{S}}|v_j|\|\mathbf{A}_S\mathbf{v}_S\|_2\|\mathbf{a}_j\|_2 \\ &\leq \kappa\sqrt{1+\delta}\|\mathbf{v}_{\overline{S}}\|_1\|\mathbf{v}_S\|_2 \ .\end{aligned}$$

Hereby, we have once again applied (4.29) and that $\|\mathbf{a}_j\|_2 \leq \kappa$. Combining these estimates for $\mathbf{v}_S$ we obtain

$$\|\mathbf{v}_S\|_2 \leq \frac{2\kappa}{1-\delta}\sqrt{s}\eta + \frac{\kappa\sqrt{1+\delta}}{1-\delta}\|\mathbf{v}_{\overline{S}}\|_1 \ .$$

Finally, this inequality together with (4.38) yields

$$\begin{aligned}\|\mathbf{v}\|_2 &\leq \|\mathbf{v}_S\|_2 + \|\mathbf{v}_{\overline{S}}\|_2 \leq \|\mathbf{v}_S\|_2 + \|\mathbf{v}_{\overline{S}}\|_1 \\ &\leq \frac{2\kappa}{1-\delta}\sqrt{s}\eta + \left(1+\frac{\kappa\sqrt{1+\delta}}{1-\delta}\right)\|\mathbf{v}_{\overline{S}}\|_1 \\ &\leq \frac{2\kappa}{1-\delta}\sqrt{s}\eta + \left(1+\frac{\kappa\sqrt{1+\delta}}{1-\delta}\right)\left(\frac{2\gamma\frac{\sqrt{1+\delta}}{1-\delta}+2\tau\sqrt{s}}{1-b}\eta + \frac{2}{1-b}\sigma_s(\mathbf{x})_1\right) \\ &= C_1\sigma_s(\mathbf{x})_1 + (C_2 + C_3\sqrt{s})\eta\end{aligned}$$

with the claimed values of the constants. □

Next we give another characterization of exact recovery via ℓ_1-minimization via tangent cones to the ℓ_1-ball. For a vector $\mathbf{x} \in \mathbb{R}^N$, we introduce the convex cone

$$T(\mathbf{x}) = \mathrm{cone}\{\mathbf{z}-\mathbf{x} : \mathbf{z} \in \mathbb{R}^N, \|\mathbf{z}\|_1 \leq \|\mathbf{x}\|_1\} \ , \tag{4.40}$$

where the right hand side is the conic hull of the indicated set, see (B.4).

Theorem 4.34. *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$. A vector $\mathbf{x} \in \mathbb{R}^N$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$ if and only if $\ker\mathbf{A} \cap T(\mathbf{x}) = \{0\}$.*

Proof. Assume that $\ker \mathbf{A} \cap T(\mathbf{x}) = \{0\}$. Let $\mathbf{z}$ be an ℓ_1-minimizer so that $\|\mathbf{z}\|_1 \leq \|\mathbf{x}\|_1$ and $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. This means that $\mathbf{v} := \mathbf{z} - \mathbf{x} \in T(\mathbf{x}) \cap \ker \mathbf{A}$; therefore, by assumption $\mathbf{v} = 0$ and $\mathbf{x}$ is the unique ℓ_1-minimizer. Conversely, assume that $\mathbf{x}$ is the unique ℓ_1-minimizer. Then $\|\mathbf{x} + \mathbf{v}\|_1 > \|\mathbf{x}\|_1$ for all $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, which implies that $\mathbf{v} \notin T(\mathbf{x})$ for such $\mathbf{v}$. It follows that $(\ker \mathbf{A} \setminus \{0\}) \cap T(\mathbf{x}) = \emptyset$ or $\ker \mathbf{A} \cap T(\mathbf{x}) = \{0\}$. □

Remark 4.35. The previous theorem extends literally to the complex case when working with the complex cone $T(\mathbf{x}) = \mathrm{cone}\{\mathbf{z} - \mathbf{x} : \mathbf{z} \in \mathbb{C}^N, \|\mathbf{z}\|_1 \leq \|\mathbf{x}\|_1\}$.

Let us extend the above theorem to stable recovery.

Theorem 4.36. *Let $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{A} \in \mathbb{R}^{m \times N}$. Suppose noisy measurements are given, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta$. Assume that, for some $\tau > 0$,*

$$\inf_{\mathbf{z} \in T(\mathbf{x}), \|\mathbf{z}\|_2 = 1} \|\mathbf{A}\mathbf{z}\|_2 \geq \tau \, .$$

Then the minimizer $\mathbf{x}^\sharp$ of

$$\min \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta \tag{4.41}$$

satisfies

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \leq \frac{2\eta}{\tau} \, . \tag{4.42}$$

Proof. Since $\mathbf{x}$ is feasible for the program (4.41) we have $\|\mathbf{x}^\sharp\|_1 \leq \|\mathbf{x}\|_1$, so that $\mathbf{x}^\sharp - \mathbf{x} \in T(\mathbf{x})$. The triangle inequality gives

$$\|\mathbf{A}(\mathbf{x}^\sharp - \mathbf{x})\|_2 \leq \|\mathbf{A}\mathbf{x}^\sharp - \mathbf{y}\|_2 + \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq 2\eta \, .$$

By assumption $\|\mathbf{A}(\mathbf{x}^\sharp - \mathbf{x})\|_2 \geq \tau \|\mathbf{x}^\sharp - \mathbf{x}\|_2$. This implies (4.42). □

Remark 4.37. Again, the result extends without changes to the complex case.

We conclude this chapter with a geometric interpretation of Theorem 4.29. We first recall that a *convex polytope* K in $\mathbb{R}^n$ can be viewed as either the convex hull of a finite set of points or as a bounded intersection of finitely many half-spaces. For instance, with $(\mathbf{e}_1, \ldots, \mathbf{e}_N)$ denoting the canonical basis of $\mathbb{R}^N$, the unit ball of ℓ_1^N described as

$$B_1^N := \mathrm{conv}\{\mathbf{e}_1, -\mathbf{e}_1, \ldots, \mathbf{e}_N, -\mathbf{e}_N\} = \bigcap_{\boldsymbol{\varepsilon} \in \{-1,1\}^N} \Big\{ \mathbf{z} \in \mathbb{R}^N : \sum_{i=1}^N \varepsilon_i z_i \leq 1 \Big\}$$

is a convex polytope. Its image under a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ is also a convex polytope, since it is the convex hull of $\{\mathbf{A}\mathbf{e}_1, -\mathbf{A}\mathbf{e}_1, \ldots, \mathbf{A}\mathbf{e}_N, -\mathbf{A}\mathbf{e}_N\}$. A *face* of a convex polytope K in $\mathbb{R}^n$ is a set of the form

$$F = \{\mathbf{z} \in K : \langle \mathbf{z}, \mathbf{h} \rangle = c\}$$

for some $\mathbf{h} \in \mathbb{R}^n$ and some $c \in \mathbb{R}$, where $\langle \mathbf{z}, \mathbf{h} \rangle \le c$ holds for all $\mathbf{z} \in C$. Note that $c > 0$ if F is a proper face of a symmetric convex polytope K, so we may always assume $c = 1$ in this case. A face F of K is called a *k-face* if its affine hull has dimension k. The 0-, 1-, $(n-2)$-, and $(n-1)$-faces are called vertices, edges, ridges, and facets, respectively. For $0 \le k \le N-1$, it can be verified that the k-faces of B_1^N are the $2^{k+1}\binom{N}{k+1}$ sets

$$\left\{\mathbf{z} \in B_1^N : \sum_{k \in K} \varepsilon_k z_k = 1\right\} = \operatorname{conv}\{\varepsilon_k \mathbf{e}_k, k \in K\} \,,$$

where K is a subset of $[N]$ with size $k+1$ and $(\varepsilon_k)_{k \in K}$ is a sequence in $\{-1, 1\}$. Thus, if a vector $\mathbf{x} \in \mathbb{R}^N$ with $\|\mathbf{x}\|_1 = 1$ is exactly s-sparse, it is contained in one and only one $(s-1)$-face of B_1^N, namely $\operatorname{conv}\{\operatorname{sgn}(x_j)\mathbf{e_j}, j \in S\}$. We are now ready to give the final necessary and sufficient condition for the recovery of individual vectors via basis pursuit.

Theorem 4.38. *For a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, a vector $\mathbf{x} \in \mathbb{R}^N$ with support S of size $s \ge 1$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$ if and only if the $(s-1)$-face of B_1^N containing $\mathbf{x}/\|\mathbf{x}\|_1$ maps to an $(s-1)$-face of $\mathbf{A}B_1^N$.*

Proof. With $S := \operatorname{supp}(\mathbf{x})$, the desired necessary and sufficient condition says that $F := \mathbf{A}(\operatorname{conv}\{\operatorname{sgn}(x_j)\mathbf{e_j}, j \in S\})$ is an $(s-1)$-face of $\mathbf{A}B_1^N$. We notice that its affine hull $\mathbf{Ax} + V$, where $V := \{\sum_{j \in S} t_j \mathbf{A}\mathbf{e}_j, \sum_{j \in S} t_j = 0\}$, has dimension $s-1$ if and only if the matrix $\mathbf{A}_S$ is injective. This can be seen by considering, for a fixed $j_0 \in S$, the surjective linear map

$$\mathbb{R}^{s-1} \to V : (t_j)_{j \in S \setminus \{j_0\}} \mapsto \sum_{j \in S \setminus \{j_0\}} t_j \mathbf{A}\mathbf{e}_j - \left(\sum_{j \in S \setminus \{j_0\}} t_j\right) \mathbf{A}\mathbf{e}_{j_0} \,.$$

We then notice that F is a face of $\mathbf{A}B_1^N$ if and only if there exists $\mathbf{h} \in \mathbb{R}^m$ such that

$$\mathbf{z} \in B_1^N \Rightarrow \langle \mathbf{Az}, \mathbf{h} \rangle \le 1 \,, \quad \text{with equality if and only if } \mathbf{Az} \in F \,.$$

The latter is equivalent to

$$\mathbf{z} \in B_1^N \Rightarrow \langle \mathbf{z}, \mathbf{A}^\top \mathbf{h} \rangle \le 1$$

with equality if and only if $\mathbf{z} \in \operatorname{conv}\{\operatorname{sgn}(x_j)\mathbf{e}_j, j \in S\}$. This translates into

$$(\mathbf{A}^\top \mathbf{h})_j = \operatorname{sgn}(x_j), \quad j \in S, \qquad\qquad |(\mathbf{A}^\top \mathbf{h})_\ell| < 1, \quad \ell \in \overline{S} \,.$$

The desired necessary and sufficient condition is recognized as Condition (b) of Theorem 4.29, which completes the proof. □

In view of the equivalence between recovery of all vectors $\mathbf{x} \in \mathbb{R}^N$ with sparsity at most s and with sparsity exactly s mentioned in Remark 4.30, we conclude with the following alternative to the null space property.

Corollary 4.39. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique solution of (P_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if every $(s-1)$-face of B_1^N maps to an $(s-1)$-face of $\mathbf{A}B_1^N$.*

4.5 Low-Rank Matrix Recovery

In this section we shortly digress on the problem of recovering matrices of low rank from incomplete linear measurements, which was already mentioned in Section 1.2 (p. 18). In this context, the number of nonzero singular values — the rank of matrix — replaces the number of nonzero entries — the sparsity of a vector.

We suppose that a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ of rank at most r is observed via the measurement vector $\mathbf{y} = \mathcal{A}(\mathbf{X}) \in \mathbb{C}^m$ where $\mathcal{A}$ is a linear map from $\mathbb{C}^{n_1 \times n_2}$ to $\mathbb{C}^m$. As in the vector case the first approach to this problem that probably comes to mind is to solve the rank-minimization problem

$$\underset{\mathbf{Z} \in \mathbb{C}^{n_1 \times n_2}}{\text{minimize}} \ \operatorname{rank}(\mathbf{Z}) \quad \text{subject to } \mathcal{A}(\mathbf{X}) = \mathbf{y} \ .$$

Unfortunately, like ℓ_0-minimization this problem is NP-hard, see Exercise 2.11. Motivated by the vector case where ℓ_1-minimization is a good strategy, we relax the minimization of the rank to the nuclear norm minimization problem

$$\underset{\mathbf{Z} \in \mathbb{C}^{n_1 \times n_2}}{\text{minimize}} \ \|\mathbf{Z}\|_* \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathbf{y} \ . \tag{4.43}$$

This is convex optimization problem. Here, the *nuclear norm*, see (A.28), is defined by

$$\|\mathbf{Z}\|_* := \sum_{j=1}^{n} \sigma_j(\mathbf{Z}) \ , \qquad n := \min\{n_1, n_2\},$$

is the ℓ_1-norm of the vector $[\sigma_1(\mathbf{Z}), \dots, \sigma_n(\mathbf{Z})]^\top$ of singular values of Z, see also (A.28) and Appendix A.2 in general for the fact that $\|\cdot\|_*$ is indeed a norm.

The analysis of the nuclear norm minimization strategy (4.43) is analogous to the vector case. In particular, the success of the strategy is equivalent to a null space property.

Theorem 4.40. *Given a linear map $\mathcal{A}$ from $\mathbb{C}^{n_1 \times n_2}$ to $\mathbb{C}^m$, every matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ of rank at most r is the unique solution of (4.43) with $\mathbf{y} = \mathcal{A}(\mathbf{X})$ if and only if, for all $\mathbf{M} \in \ker \mathcal{A} \setminus \{0\}$ with singular values $\sigma_1(\mathbf{M}) \geq \cdots \geq \sigma_n(\mathbf{M}) \geq 0$, $n := \min\{n_1, n_2\}$,*

$$\sum_{j=1}^{r} \sigma_j(\mathbf{M}) < \sum_{j=r+1}^{n} \sigma_j(\mathbf{M}) \, . \tag{4.44}$$

Proof. Let us first assume that every matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ of rank at most r is the unique solution of (4.43) with $\mathbf{y} = \mathcal{A}\mathbf{X}$. We consider the singular value decomposition of a matrix $\mathbf{M} \in \ker \mathcal{A} \setminus \{0\}$ and write $\mathbf{M} = \mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\mathbf{V}^*$ for $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ and $\mathbf{U} \in \mathbb{C}^{n_1 \times n_1}, \mathbf{V} \in \mathbb{C}^{n_2 \times n_2}$ unitary. Setting $\mathbf{M}_1 = \mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)\mathbf{V}^*$ and $\mathbf{M}_2 = \mathbf{U}\mathrm{diag}(0, \ldots, 0, -\sigma_{r+1}, \ldots, -\sigma_n)\mathbf{V}^*$, we have $\mathbf{M} = \mathbf{M}_1 - \mathbf{M}_2$. Thus, $\mathcal{A}(\mathbf{M}) = 0$ translates into $\mathcal{A}(\mathbf{M}_1) = \mathcal{A}(\mathbf{M}_2)$. Since the rank of $\mathbf{M}_1$ is at most r, its nuclear norm must be smaller than the nuclear norm of $\mathbf{M}_2$. This means that $\sigma_1 + \cdots + \sigma_r < \sigma_{r+1} + \cdots + \sigma_n$, as desired.

Conversely, let us now assume that $\sum_{j=1}^{r} \sigma_j(\mathbf{M}) < \sum_{j=r+1}^{n} \sigma_j(\mathbf{M})$ for every $\mathbf{M} \in \ker \mathcal{A} \setminus \{0\}$ with singular values $\sigma_1(\mathbf{M}) \geq \cdots \geq \sigma_n(\mathbf{M}) \geq 0$. Consider a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ of rank at most r and a matrix $\mathbf{Z} \in \mathbb{C}^{n_1 \times n_2}$, $\mathbf{Z} \neq \mathbf{X}$, satisfying $\mathcal{A}(\mathbf{Z}) = \mathcal{A}(\mathbf{X})$. We aim at proving that $\|\mathbf{Z}\|_* > \|\mathbf{X}\|_*$. Let us set $\mathbf{M} := \mathbf{X} - \mathbf{Z} \in \ker \mathcal{A} \setminus \{0\}$. Lemma A.20 insures that the singular values $\sigma_j(\mathbf{M}), \sigma_j(\mathbf{Z}), \sigma_j(\mathbf{M})$ satisfy

$$\|\mathbf{Z}\|_* = \sum_{j=1}^{n} \sigma_j(\mathbf{X} - \mathbf{M}) \geq \sum_{j=1}^{n} |\sigma_j(\mathbf{X}) - \sigma_j(\mathbf{M})| \, .$$

For $j \in [r]$, we have $|\sigma_j(\mathbf{X}) - \sigma_j(\mathbf{M})| \geq \sigma_j(\mathbf{X}) - \sigma_j(\mathbf{M})$, and for $r+1 \leq j \leq n$, it holds $|\sigma_j(\mathbf{X}) - \sigma_j(\mathbf{M})| = \sigma_j(\mathbf{M})$. In view of our hypothesis, we derive

$$\|\mathbf{Z}\|_* \geq \sum_{j=1}^{r} \sigma_j(\mathbf{X}) - \sum_{j=1}^{r} \sigma_j(\mathbf{M}) + \sum_{j=r+1}^{n} \sigma_j(\mathbf{M}) > \sum_{j=1}^{r} \sigma_j(\mathbf{X}) = \|\mathbf{X}\|_* \, .$$

This establishes the desired inequality. □

Like in the vector case, one can introduce stable and robust versions of the rank null space property (4.44) and show corresponding error estimates for reconstruction via nuclear norm minimization, see Exercises 4.17 and 4.18. Also, recovery conditions for individual low-rank matrices can be shown, analogously to the results for the vector case in Section 4.4, see Exercise 4.19.

In the remainder of the book, the low-rank recovery problem will only be treated via exercises, see e.g. Exercises 6.24 and 9.12. The reader is, of course, very welcome to work through them.

Notes

Throughout the chapter, we have insisted on sparse vectors to be unique solutions of (P_1). If we dropped the uniqueness requirement, then a necessary and sufficient condition for every s-sparse vector to be a solution of (P_1) would

be a weak null space property where the the strict inequality is replaced by a weak inequality sign.

The null space property is somewhat folklore in the compressive sensing literature. It appeared implicitly in works of D. Donoho and M. Elad [133], of D. Donoho and X. Huo [136], and of M. Elad and A. Bruckstein [158]. R. Gribonval and M. Nielsen also isolated the notion in [206]. The name was first used by A. Cohen, W. Dahmen, and R. DeVore in [102], albeit for a property slightly more general than (4.3), namely $\|\mathbf{v}\|_1 \le C\,\sigma_s(\mathbf{v})_1$ for all $\mathbf{v} \in \ker \mathbf{A}$, where $C \ge 1$ is an unspecified constant. We have coined the terms stable and robust null space properties for some notions that are implicit in the literature.

The equivalence between the real and complex null space properties was established by S. Foucart and R. Gribonval in [183] using a different argument than the one of Theorem 4.7. The result was generalized by M.-J. Lai and L. Liu in [273]. The proof of Theorem 4.7 follows their argument.

Given a measurement matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$ and a vector $\mathbf{x} \in \mathbb{K}^N$, one can rephrase the optimization problem (P_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ as the problem of best approximation to $\mathbf{x} \in \mathbb{K}^N$ from the subspace $\ker \mathbf{A}$ of $\mathbb{K}^N$ in the ℓ_1-norm. Some of the results of this chapter can be derived using known characterizations of best ℓ_1-approximation. The book [335] by A. Pinkus is a good source on the subject, although it does not touch the complex setting.

The term *instance optimality* is sometimes also used for what we called stability in this chapter. Chapter 11 gives more details on this topic.

The stability and robustness of sparse reconstruction via basis pursuit, as stated after Theorem 4.21, were established by E. Candès, J. Romberg, and T. Tao in [80] under a restricted isometry property — see Chapter 6 — condition on the measurement matrix.

The fact that sparse recovery via ℓ_q-minimization implies sparse recovery via ℓ_p-minimization whenever $0 < p < q \le 1$ was proved by R. Gribonval and M. Nielsen in [207].

The sufficient condition (ii) of Theorems 4.25 and 4.29, as well as Corollary 4.27, can be found in works of J.-J. Fuchs [187] and of J. Tropp [415]. E. Candès, J. Romberg, and T. Tao stated in [72] that it is also a necessary condition if the measurement matrix is a partial Fourier matrix. The reasoning was slightly incorrect, for it would generalize to other measurement matrices and contradict Remark 4.28.

The success of sparse recovery via basis pursuit was characterized in terms of faces of polytopes by Donoho in [129], where the condition of Corollary 4.39 was also interpreted in terms of *neighborliness* of the polytope $\mathbf{A}B_1^N$ — see Exercise 4.15.

Exercises

4.1. Suppose that $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the null space property of order s. Theorems 4.5 and 2.13 guarantee that $\ker \mathbf{A}$ does not contain any $2s$-sparse vectors other than the zero vector. Give a direct proof of this fact.

4.2. Find a 2×3 matrix $\mathbf{A}$ and a nonsingular 3×3 diagonal matrix D such that $\mathbf{A}$ has the first order null space property, but $\mathbf{A}D$ does not.

4.3. Prove that an individual s-sparse vector can be recovered via basis pursuit with a number of measurements $m < 2s$.

4.4. Suppose that the null space of a real matrix $\mathbf{A}$ is a two-dimensional space with basis $(\mathbf{v}, \mathbf{w})$. Prove that $\mathbf{A}$ has the null space property of order s if and only if

$$\sum_{j \in S} |v_j| < \sum_{\ell \in \overline{S}} |v_\ell|, \quad \sum_{j \in S} |w_j| < \sum_{\ell \in \overline{S}} |w_\ell|, \quad \sum_{j \in S} |v_i w_j - v_j w_i| < \sum_{\ell \in \overline{S}} |v_i w_\ell - v_\ell w_i|,$$

for all $i \in [N]$ and all $S \subset [N]$ with $\operatorname{card}(S) \leq s$.

4.5. Prove the equivalence between the real and complex stable null space properties with constant $0 < \rho < 1$ relative to a set S.

4.6. Given $0 < c < 1$, prove the equivalence of the properties:
(i) $\|\mathbf{v}_S\|_1 \leq \|\mathbf{v}_{\overline{S}}\|_1 - c\,\|\mathbf{v}\|_1$ for all $\mathbf{v} \in \ker \mathbf{A}$ and $S \subseteq [N]$ with $\operatorname{card}(S) \leq s$,
(ii) $\|\mathbf{x}\|_1 \leq \|\mathbf{z}\|_1 - c\,\|\mathbf{x} - \mathbf{z}\|_1$ for all s-sparse $\mathbf{x} \in \mathbb{K}^N$ and $\mathbf{z} \in \mathbb{K}^N$ with $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$.

4.7. Given $S \subset [N]$, prove that a minimizer $\mathbf{x}^\sharp$ of $\|\mathbf{z}\|_1$ subject to a constraint met by $\mathbf{x} \in \mathbb{C}^N$ satisfies

$$\left\| \left(\mathbf{x} - \mathbf{x}^\sharp\right)_{\overline{S}} \right\|_1 \leq \left\| \left(\mathbf{x} - \mathbf{x}^\sharp\right)_S \right\|_1 + 2\, \|\mathbf{x}_{\overline{S}}\|_1 \ .$$

4.8. Given $\mathbf{A} \in \mathbb{R}^{m \times N}$, prove that every nonnegative s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique solution of

$$\underset{\mathbf{z} \in \mathbb{R}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x} \text{ and } \mathbf{z} \geq 0$$

if and only if

$$\mathbf{v}_{\overline{S}} \geq 0 \Longrightarrow \sum_{j=1}^{N} v_j > 0$$

for all $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$ and all $S \subseteq [N]$ with $\operatorname{card}(S) \leq s$.

4.9. Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a matrix for which $\sum_{j=1}^{N} v_j = 0$ whenever $\mathbf{v} \in \ker \mathbf{A}$, and let $S \subset [N]$ be a fixed index set. Suppose that every nonnegative vector supported on S is uniquely recovered by ℓ_1-minimization. Prove that every nonnegative vector $\mathbf{x}$ supported on S is in fact the unique vector in the set $\{\mathbf{z} \in \mathbb{R}^N : \mathbf{z} \geq 0, \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}\}$.

4.10. Given matrices $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $M \in \mathbb{C}^{m \times m}$, suppose that $M\mathbf{A}$ satisfies the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Prove that there exist constants $C, D > 0$ depending only on ρ, τ, and $\|M\|_{2 \to 2}$ such that, for any $\mathbf{x} \in \mathbb{C}^N$,

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_p \le \frac{C}{s^{1-1/p}} \, \sigma_s(\mathbf{x})_1 + D \, s^{1/p-1/2} \, \eta \, , \qquad 1 \le p \le 2 \, ,$$

where $\mathbf{x}^\sharp \in \mathbb{C}^N$ is a solution of $(\mathrm{P}_{1,\eta})$ with $\|\cdot\| = \|\cdot\|_2$, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, and $\|\mathbf{e}\|_2 \le \eta$.

4.11. Prove Theorem 4.8, and generalize other results of Sections 4.1, 4.2, and 4.3 when the ℓ_1-norm is replaced by the ℓ_q-quasinorm for $0 < q < 1$.

4.12. Given an integer $s \ge 1$ and an exponent $q \in (0,1)$, find a measurement matrix that allows reconstruction of s-sparse vectors via ℓ_p-minimization for $p < q$, but not for $p > q$.

4.13. With $\mathbf{A}$ and $\mathbf{x}$ given in Remark 4.28, verify the statement that $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$.

4.14. Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and a vector $\mathbf{x} \in \mathbb{C}^N$ with support S, prove that $\mathbf{x}$ is a minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$ if one of the following equivalent conditions holds:

(i) $\Big| \sum_{j \in S} \overline{\mathrm{sgn}(x_j)} v_j \Big| \le \|\mathbf{v}_{\overline{S}}\|_1$ for all $\mathbf{v} \in \ker \mathbf{A}$,

(ii) there exists a vector $\mathbf{h} \in \mathbb{C}^m$ such that

$$(\mathbf{A}^* \mathbf{h})_j = \mathrm{sgn}(x_j), \quad j \in S \, , \qquad\qquad |(\mathbf{A}^* \mathbf{h})_\ell| \le 1, \quad \ell \in \overline{S} \, .$$

4.15. A symmetric convex polytope K is called *centrally k-neighborly* if any set of $k+1$ of its vertices, not containing an antipodal pair, spans a k-face of K. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, prove that every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is the unique solution of (P_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ if and only if the convex polytope $\mathbf{A}B_1^N$ has $2N$ vertices and is s-neighborly.

4.16. Stable recovery via dual certificate.
Let $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{A} \in \mathbb{C}^{m \times N}$ with ℓ_2-normalized columns, $\|\mathbf{a}_k\|_2 = 1$, $k \in [N]$. For $s \ge 1$, let $S \subset [N]$ be the set of indices of the s largest absolute entries of $\mathbf{x}$. Assume that $\|\mathbf{A}_S\|_{2 \to 2} \le \alpha$, for some $\alpha > 0$ and that there exists a dual certificate $\mathbf{u} = \mathbf{A}^* \mathbf{h} \in \mathbb{C}^N$ with $\mathbf{h} \in \mathbb{C}^m$ such that

$$\mathbf{u}_S = \mathrm{sgn}(\mathbf{x}_S) \, , \quad \|\mathbf{u}_{\overline{S}}\|_\infty \le \beta \, , \quad \|\mathbf{h}\|_2 \le \gamma \sqrt{s} \, .$$

for constants $\beta \in (0,1)$ and $\gamma > 0$. Suppose that noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ are given with $\|\mathbf{e}\|_2 \le \eta$. Show that the solution $\mathbf{x}^\sharp \in \mathbb{C}^N$ of the ℓ_1-minimization problem

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{z}-\mathbf{y}\|_2 \le \eta$$

satisfies

$$\left\|\mathbf{x}-\mathbf{x}^\sharp\right\|_2 \le C_1\sqrt{s}\eta + C_2\sigma_s(\mathbf{x})_1 \ .$$

for appropriate constants $C_1, C_2 > 0$ depending only on α, β, γ.

4.17. Stable rank null space property.
Let $\mathcal{A}: \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ be a linear measurement map. Assume that $\mathcal{A}$ satisfies the stable rank null space property of order r and constant $\rho \in (0,1)$, that is, for all $\mathbf{M} \in \ker\mathcal{A}\setminus\{\mathbf{0}\}$ the singular $\sigma_\ell(\mathbf{M})$ satisfy

$$\sum_{\ell=1}^{r}\sigma_\ell(\mathbf{M}) \le \rho \sum_{\ell=r+1}^{\min\{n_1,n_2\}}\sigma_\ell(\mathbf{M}) \ .$$

Show that, for all $\mathbf{X}, \mathbf{Z} \in \mathbb{C}^{n_1\times n_2}$ with $\mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{Z})$,

$$\|\mathbf{X}-\mathbf{Z}\|_* \le \frac{1+\rho}{1-\rho}\left(\|\mathbf{Z}\|_* - \|\mathbf{X}\|_* + 2\sum_{\ell=r+1}^{\min\{n_1,n_2\}}\sigma_\ell(\mathbf{X})\right) \ . \tag{4.45}$$

For $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ let $\mathbf{X}^\sharp$ be the minimizer of the nuclear norm minimization problem

$$\min_{\mathbf{Z}\in\mathbb{C}^{n_1\times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathcal{A}(\mathbf{X}) \ .$$

Conclude that

$$\|\mathbf{X}-\mathbf{X}^\sharp\|_* \le \frac{2(1+\rho)}{1-\rho}\sum_{\ell=r+1}^{\min\{n_1,n_2\}}\sigma_\ell(\mathbf{X}) \ .$$

Conversely, show that if (4.45) holds for all $\mathbf{X}, \mathbf{Z}$ such that $\mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{Z})$ then $\mathcal{A}$ satisfies the stable rank null space property of order r and constant $\rho \in (0,1)$.

4.18. Robust rank null space property.
Let $\mathcal{A}: \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ be a linear measurement map and $\|\cdot\|$ be some norm on $\mathbb{C}^m$.

(a) We say that $\mathcal{A}$ satisfies the robust rank null space property of order r and constants $\rho \in (0,1)$, $\tau > 0$ if, for all $\mathbf{M} \in \ker\mathcal{A}\setminus\{\mathbf{0}\}$, the singular values $\sigma_\ell(\mathbf{M})$ satisfy

$$\sum_{\ell=1}^{r}\sigma_\ell(\mathbf{M}) \le \rho \sum_{\ell=r+1}^{\min\{n_1,n_2\}}\sigma_\ell(\mathbf{M}) + \tau\|\mathcal{A}(\mathbf{M})\| \ .$$

Show that

$$\|\mathbf{X}-\mathbf{Z}\|_* \le \frac{1+\rho}{1-\rho}\left(\|\mathbf{Z}\|_* - \|\mathbf{X}\|_* + 2\sum_{\ell=r+1}^{\min\{n_1,n_2\}} \sigma_\ell(\mathbf{X})\right) + \frac{2\tau}{1-\rho}\|\mathcal{A}(\mathbf{Z}-\mathbf{X})\|$$

for all $\mathbf{X}, \mathbf{Z} \in \mathbb{C}^{n_1\times n_2}$ if and only if $\mathcal{A}$ satisfies the robust rank null space property of order r with constants ρ, τ with respect to $\|\cdot\|$.

(b) Assume that $\mathcal{A}$ satisfies the Frobenius robust rank null space property of order r and constants $\rho \in (0,1)$, $\tau > 0$ with respect to a norm $\|\cdot\|$, that is, for all $\mathbf{M} \in \ker \mathcal{A} \setminus \{\mathbf{0}\}$,

$$\left(\sum_{\ell=1}^{r} \sigma_\ell(\mathbf{M})^2\right)^{1/2} \le \frac{\rho}{\sqrt{r}} \sum_{\ell=r+1}^{\min\{n_1,n_2\}} \sigma_\ell(\mathbf{M}) + \tau\|\mathcal{A}(\mathbf{M})\|_2 \,.$$

For $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ assume that $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ with $\|\mathbf{e}\|_2 \le \eta$ for some $\eta \ge 0$. Let $\mathbf{X}^\sharp$ be the solution to the quadratically-constrained nuclear norm minimization problem

$$\min_{\mathbf{Z}\in\mathbb{C}^{n_1\times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_2 \le \eta \,.$$

Show that

$$\|\mathbf{X} - \mathbf{X}^\sharp\|_F \le \frac{C}{\sqrt{r}} \sum_{\ell=r+1}^{\min\{n_1,n_2\}} \sigma_\ell(\mathbf{X}) + D\eta$$

for constants $C, D > 0$ only depending on ρ, τ.

4.19. Low-rank matrix recovery via dual certificate.
Let $\|\cdot\|_*$ denote the nuclear norm, see (A.28) and $\langle \mathbf{X}, \mathbf{Y}\rangle_F = \operatorname{tr}(\mathbf{X}\mathbf{Y}^*)$, for matrices $\mathbf{X}, \mathbf{Y}$, be the Frobenius inner product (A.14).

(a) Show that the nuclear norm is the dual norm of the operator norm, that is, for $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$,

$$\|\mathbf{X}\|_* = \sup_{\mathbf{Y}\in\mathbb{C}^{n_1\times n_2}, \|\mathbf{Y}\|_{2\to 2}\le 1} |\langle \mathbf{X}, \mathbf{Y}\rangle_F| \,.$$

(b) Let $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{n_1\times n_2}$ such that $\mathbf{X}\mathbf{Y}^* = \mathbf{0}$ and $\mathbf{X}^*\mathbf{Y} = \mathbf{0}$. Show that

$$\|\mathbf{X}\|_* + \|\mathbf{Y}\|_* = \|\mathbf{X} + \mathbf{Y}\|_* \,.$$

(c) Let $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ of rank r with singular value decomposition $\mathbf{X} = \sum_{\ell=1}^{r} \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^*$, that is, both $\{\mathbf{u}_\ell : \ell \in [r]\}$ and $\{\mathbf{v}_\ell : \ell \in [r]\}$ form orthonormal systems. Let T be the linear space spanned by the vectors

$$\{\mathbf{u}_\ell \mathbf{x}_\ell^*, \mathbf{y}_\ell \mathbf{v}_\ell^* : \mathbf{x}_\ell \in \mathbb{C}^{n_2}, \mathbf{y}_\ell \in \mathbb{C}^{n_1}, \ell \in [r]\} \,. \tag{4.46}$$

Denote by $T^\perp$ the orthogonal complement of T, where orthogonality is with respect to the Frobenius inner product. Let $\mathbf{P}_U \in \mathbb{C}^{n_1\times n_2}$ be the

orthogonal projection onto the span of $\{\mathbf{u}_\ell : \ell \in [r]\}$, that is, $\mathbf{P}_U = \sum_{r=1}^{n} \mathbf{u}_\ell \mathbf{u}_\ell^*$, and $\mathbf{P}_V \in \mathbb{C}^{n_2 \times n_2}$ the orthogonal projection onto the span of $\{\mathbf{v}_\ell : \ell \in [r]\}$. Show that the orthogonal projections $\mathcal{P}_T : \mathbb{C}^{n_1 \times n_2} \to T$, $\mathcal{P}_{T^\perp} : \mathbb{C}^{n_1 \times n_2} \to T^\perp$ are given by

$$\begin{aligned} \mathcal{P}_T(\mathbf{Z}) &= \mathbf{P}_U \mathbf{Z} + \mathbf{Z}\mathbf{P}_V - \mathbf{P}_U \mathbf{Z} \mathbf{P}_V \,, \\ \mathcal{P}_{T^\perp}(\mathbf{Z}) &= (\mathbf{Id} - \mathbf{P}_U)\mathbf{Z}(\mathbf{Id} - \mathbf{P}_V) \,. \end{aligned}$$

(d) Let $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ be a linear map, and let $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ be of rank r with singular value decomposition $\mathbf{X} = \sum_{\ell=1}^{r} \sigma_r \mathbf{u}_\ell \mathbf{v}_\ell^*$. Show that $\mathbf{X}$ is the unique solution of the nuclear norm minimization problem

$$\min_{\mathbf{Z} \in \mathbb{C}^{n_1 \times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to} \quad \mathcal{A}(\mathbf{Z}) = \mathcal{A}(\mathbf{X})$$

if $\mathcal{A}$ restricted to T is injective and if there exists a dual certificate $\mathbf{h} \in \mathbb{C}^m$ such that $\mathbf{M} = \mathcal{A}^* \mathbf{h} \in \mathbb{C}^{n_1 \times n_2}$ satisfies

$$\begin{aligned} \mathcal{P}_T(\mathbf{M}) &= \sum_{\ell=1}^{r} \mathbf{u}_\ell \mathbf{v}_\ell^* \,, \\ \|\mathcal{P}_{T^\perp}(\mathbf{M})\|_{2 \to 2} &< 1 \,, \end{aligned}$$

where T is the span of (4.46) and $T^\perp$ its orthogonal complement.

5

Coherence

In compressive sensing, the analysis of recovery algorithms usually involves a quantity that measures the suitability of the measurement matrix. The coherence is a very simple such measure of quality. In general, the smaller the coherence, the better the recovery algorithms perform. In Section 5.1, we introduce the notion of coherence of a matrix and some of its generalizations. In Section 5.2, we examine how small the coherence can be and we point out some matrices with small coherence. In Sections 5.5, 5.3, and 5.4, we give some sufficient conditions expressed in terms of the coherence that guarantee the success of basic thresholding, orthogonal matching pursuit, and basis pursuit.

5.1 Definitions and Basic Properties

We start with the definition of the coherence of a matrix. We stress that the columns of the matrix are always implicitly understood to be ℓ_2-normalized.

Definition 5.1. *Let* $\mathbf{A} \in \mathbb{C}^{m\times N}$ *be a matrix with* ℓ_2*-normalized columns* $\mathbf{a}_1, \dots, \mathbf{a}_N$*, i.e.,* $\|\mathbf{a}_i\|_2 = 1$ *for all* $1 \le i \le N$*. The* coherence $\mu = \mu(A)$ *of the matrix* $\mathbf{A}$ *is defined as*

$$\mu := \max_{1\le i\ne j\le N} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle| . \tag{5.1}$$

Next we introduce the more general concept of ℓ_1-coherence function, which incorporates the usual coherence as the particular value $s = 1$ of its argument.

Definition 5.2. *Let* $\mathbf{A} \in \mathbb{C}^{m\times N}$ *be a matrix with* ℓ_2*-normalized columns* $\mathbf{a}_1, \dots, \mathbf{a}_N$*. The* ℓ_1-coherence function μ_1 *of the matrix* $\mathbf{A}$ *is defined for* $1 \le s \le N-1$ *by*

$$\mu_1(s) := \max_{i\in[N]} \max\Big\{ \sum_{j\in S} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle| , \, S \subseteq [N], \operatorname{card}(S) = s, \, i \notin S \Big\}.$$

It is straightforward to observe that, for $1 \leq s \leq N-1$,

$$\mu \leq \mu_1(s) \leq s\,\mu, \tag{5.2}$$

and more generally that, for $1 \leq s,t \leq N-1$ with $s+t \leq N-1$,

$$\max\{\mu_1(s),\mu_1(t)\} \leq \mu_1(s+t) \leq \mu_1(s)+\mu_1(t). \tag{5.3}$$

We remark that the coherence, and more generally the ℓ_1-coherence function, is invariant under multiplication on the left by a unitary matrix $\mathbf{U}$, for the columns of $\mathbf{UA}$ are the ℓ_2-normalized vectors $\mathbf{Ua}_1,\ldots,\mathbf{Ua}_N$ and they satisfy $\langle \mathbf{Ua}_i, \mathbf{Ua}_j\rangle = \langle \mathbf{a}_i, \mathbf{a}_j\rangle$. Moreover, because of the Cauchy–Schwarz inequality $|\langle \mathbf{a}_i, \mathbf{a}_j\rangle| \leq \|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2$, it is clear that the coherence of a matrix is bounded above by one, i.e.,

$$\mu \leq 1.$$

Let us consider for a moment a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ with $m \geq N$. We observe that $\mu = 0$ if and only if the columns of $\mathbf{A}$ form an orthonormal system. In particular, in the case of a square matrix, we have $\mu = 0$ if and only if $\mathbf{A}$ is a unitary matrix. From now on, we concentrate on the situation occurring in compressive sensing, i.e., we only consider matrices $\mathbf{A} \in \mathbb{C}^{m\times N}$ with $m < N$. In this case, there are limitations on how small the coherence can be. These limitations are given in Section 5.2. For now, we simply point out that a small coherence implies that column-submatrices of moderate size are well conditioned. Let us recall that the notation $\mathbf{A}_S$ denotes the matrix formed by the columns of $\mathbf{A} \in \mathbb{C}^{m\times N}$ indexed by a subset S of $[N]$.

Theorem 5.3. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns, and let $1 \leq s \leq N$. For all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$,*

$$\big(1-\mu_1(s-1)\big)\,\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq \big(1+\mu_1(s-1)\big)\,\|\mathbf{x}\|_2^2,$$

or equivalently, for each set $S \subseteq [N]$ with $\operatorname{card}(S) \leq s$, the eigenvalues of the matrix $\mathbf{A}_S^*\mathbf{A}_S$ lie in the interval $\big[\,1-\mu_1(s-1)\,,\,1+\mu_1(s-1)\,\big]$. In particular, if $\mu_1(s-1) < 1$, then $\mathbf{A}_S^*\mathbf{A}_S$ is invertible.

Proof. For a set $S \subseteq [N]$ with $\operatorname{card}(S) \leq s$, since the matrix $\mathbf{A}_S^*\mathbf{A}_S$ is positive semidefinite, it has an orthonormal basis of eigenvectors associated with real, positive eigenvalues. We denote the minimal eigenvalue by $\lambda_{\min}$ and the maximal eigenvalue by $\lambda_{\max}$. Then, since $\mathbf{Ax} = \mathbf{A}_S\mathbf{x}_S$ for any $\mathbf{x} \in \mathbb{C}^N$ supported on S, it is easy to see that the maximum of

$$\|\mathbf{Ax}\|_2^2 = \langle \mathbf{A}_S\mathbf{x}_S, \mathbf{A}_S\mathbf{x}_S\rangle = \langle \mathbf{A}_S^*\mathbf{A}_S\mathbf{x}_S, \mathbf{x}_S\rangle$$

over the set $\{\mathbf{x} \in \mathbb{C}^N, \operatorname{supp}\mathbf{x} \subseteq S, \|\mathbf{x}\|_2 = 1\}$ is $\lambda_{\max}$ and that its minimum is $\lambda_{\min}$. This explains the equivalence mentioned in the theorem. Now, due to the normalizations $\|a_j\|_2 = 1$ for all $1 \leq j \leq N$, the diagonal entries of $\mathbf{A}_S^*\mathbf{A}_S$ all equal one. By Gershgorin's disc theorem, see Theorem A.12, the

eigenvalues of $\mathbf{A}_S^*\mathbf{A}_S$ are contained in the union of the discs centered at 1 with radii

$$r_j := \sum_{\ell \in S, \ell \neq j} |(\mathbf{A}_S^*\mathbf{A}_S)_{j,\ell}| = \sum_{\ell \in S, \ell \neq j} |\langle a_\ell, a_j \rangle| \leq \mu_1(s-1), \qquad j \in S.$$

Since these eigenvalues are real, they must lie in $\big[\, 1-\mu_1(s-1)\,,\, 1+\mu_1(s-1) \,\big]$, as announced. □

Corollary 5.4. *Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with ℓ_2-normalized columns and an integer $s \geq 1$, if*

$$\mu_1(s) + \mu_1(s-1) < 1,$$

then, for each set $S \subseteq [N]$ with $\mathrm{card}(S) \leq 2s$, the matrix $\mathbf{A}_S^*\mathbf{A}_S$ is invertible and the matrix $\mathbf{A}_S$ injective. In particular, the conclusion holds if

$$\mu < \frac{1}{2s-1}.$$

Proof. In view of (5.3), the condition $\mu_1(s) + \mu_1(s-1) < 1$ implies that $\mu_1(2s-1) < 1$. For a set $S \subseteq [N]$ with $\mathrm{card}(S) \leq 2s$, according to Theorem 5.3, the smallest eigenvalue of the matrix $\mathbf{A}_S^*\mathbf{A}_S$ satisfies $\lambda_{\min} \geq 1-\mu_1(2s-1) > 0$, which shows that $\mathbf{A}_S^*\mathbf{A}_S$ is invertible. To see that $\mathbf{A}_S$ is injective, we simply observe that $\mathbf{A}_S\mathbf{z} = 0$ yields $\mathbf{A}_S^*\mathbf{A}_S\mathbf{z} = 0$, so that $\mathbf{z} = 0$. This proves the first statement. The second one simply follows from $\mu_1(s)+\mu_1(s-1) \leq (2s-1)\mu < 1$ if $\mu < 1/(2s-1)$. □

5.2 Matrices with Small Coherence

In this section, we give lower bounds for the coherence and for the ℓ_1-coherence function of a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m < N$. We also study the feasibility of achieving these lower bounds. We then give an example of a matrix with an almost minimal coherence. The analysis is carried out for matrices $\mathbf{A} \in \mathbb{K}^{m \times N}$, where the field $\mathbb{K}$ can either be $\mathbb{R}$ or $\mathbb{C}$, because the matrices achieving the lower bounds have different features in the real and complex settings. In both cases, however, their columns are equiangular tight frames, which are defined below.

Definition 5.5. *A system of ℓ_2-normalized vectors $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ in $\mathbb{K}^m$ is called* equiangular *if there is a constant $c \geq 0$ such that*

$$|\langle \mathbf{a}_i, \mathbf{a}_j \rangle| = c \qquad \text{for all } i, j \in [N],\ i \neq j\ .$$

Definition 5.6. *A system of vectors $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ in $\mathbb{K}^m$ is called a* tight frame *if there exists a constant $\lambda > 0$ such that one of the following equivalent conditions holds:*

(a) $\|\mathbf{x}\|_2^2 = \lambda \sum_{j=1}^{N} |\langle \mathbf{x}, \mathbf{a}_j \rangle|^2 \quad$ *for all* $\mathbf{x} \in \mathbb{K}^m$,

(b) $\mathbf{x} = \lambda \sum_{j=1}^{N} \langle \mathbf{x}, \mathbf{a}_j \rangle \mathbf{a}_j \quad$ *for all* $\mathbf{x} \in \mathbb{K}^m$,

(c) $\mathbf{A}\mathbf{A}^* = \dfrac{1}{\lambda} I_m$, *where* $\mathbf{A}$ *is the matrix with columns* $\mathbf{a}_1, \ldots, \mathbf{a}_N$.

Unsurprisingly, a system of ℓ_2-normalized vectors is called an equiangular tight frame if it is both an equiangular system and a tight frame. Such systems are the ones achieving the lower bound given below and known as the *Welch bound*.

Theorem 5.7. *The coherence of a matrix* $\mathbf{A} \in \mathbb{K}^{m \times N}$ *with* ℓ_2*-normalized columns satisfies*

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}} \, . \tag{5.4}$$

Equality holds if and only if the columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$ *of the matrix* $\mathbf{A}$ *form an equiangular tight frame.*

Proof. Let us introduce the *Gram matrix* $\mathbf{G} := \mathbf{A}^* \mathbf{A} \in \mathbb{K}^{N \times N}$ of the system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$, which has entries

$$G_{i,j} = \overline{\langle \mathbf{a}_i, \mathbf{a}_j \rangle} = \langle \mathbf{a}_j, \mathbf{a}_i \rangle \, , \qquad i, j \in [N] \, ,$$

and the matrix $\mathbf{H} := \mathbf{A}\mathbf{A}^* \in \mathbb{K}^{m \times m}$. On the one hand, since the system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ is ℓ_2-normalized, we have

$$\operatorname{tr}(\mathbf{G}) = \sum_{i=1}^{N} \|\mathbf{a}_i\|_2^2 = N. \tag{5.5}$$

On the other hand, since the inner product

$$\langle \mathbf{U}, \mathbf{V} \rangle_F := \operatorname{tr}(\mathbf{U}\mathbf{V}^*) = \sum_{i,j=1}^{n} U_{i,j} \overline{V_{i,j}}$$

induces the so-called *Froebenius norm* $\|\cdot\|_F$ on $\mathbb{K}^{n \times n}$, see (A.15), the Cauchy–Schwarz inequality yields

$$\operatorname{tr}(\mathbf{H}) = \langle \mathbf{H}, \mathbf{Id}_m \rangle_F \leq \|\mathbf{H}\|_F \|\mathbf{Id}_m\|_F = \sqrt{m} \, \sqrt{\operatorname{tr}(\mathbf{H}\mathbf{H}^*)}. \tag{5.6}$$

Let us now observe that

$$\begin{aligned} \operatorname{tr}(\mathbf{H}\mathbf{H}^*) = \operatorname{tr}\,(\mathbf{A}\mathbf{A}^*\mathbf{A}\mathbf{A}^*) = \operatorname{tr}\,(\mathbf{A}^*\mathbf{A}\mathbf{A}^*\mathbf{A}) = \operatorname{tr}\,(\mathbf{G}\mathbf{G}^*) = \sum_{i,j=1}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \\ = \sum_{i=1}^{N} \|\mathbf{a}_i\|_2^2 + \sum_{i,j=1, i \neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 = N + \sum_{i,j=1, i \neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \, . \end{aligned} \tag{5.7}$$

In view of $\operatorname{tr}(\mathbf{G}) = \operatorname{tr}(\mathbf{H})$, combining (5.5), (5.6), and (5.7) yields

$$N^2 \le m\bigg(N + \sum_{i,j=1, i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle|^2\bigg). \tag{5.8}$$

Taking into account that

$$|\langle \mathbf{a}_i, \mathbf{a}_j\rangle| \le \mu \qquad \text{for all } 1 \le i \neq j \le N, \tag{5.9}$$

we obtain

$$N^2 \le m\big(N + (N^2 - N)\mu^2\big),$$

which is a simple rearrangement of (5.4). Moreover, equality holds in (5.4) exactly when equalities hold in (5.6) and in (5.9). Equality in (5.6) says that $\mathbf{H} = \lambda \mathbf{Id}_m$ for some — necessarily nonnegative — constant λ, i.e., that the system $(\mathbf{a}_1, \dots, \mathbf{a}_N)$ is a tight frame. Equality in (5.9) says that this system is equiangular. □

The Welch bound can be extended to the ℓ_1-coherence function for small values of its argument.

Theorem 5.8. *The ℓ_1-coherence function of a matrix $\mathbf{A} \in \mathbb{K}^{m\times N}$ with ℓ_2-normalized columns satisfies*

$$\mu_1(s) \ge s\sqrt{\frac{N-m}{m(N-1)}} \qquad \textit{whenever} \quad s < \sqrt{N-1}. \tag{5.10}$$

Equality holds if and only if the columns $\mathbf{a}_1, \dots, \mathbf{a}_N$ of the matrix $\mathbf{A}$ form an equiangular tight frame.

The proof is based on the following lemma.

Lemma 5.9. *For $k < \sqrt{n}$, if the finite sequence $(\alpha_1, \alpha_2, \dots, \alpha_n)$ satisfies*

$$\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_n \ge 0 \qquad \textit{and} \qquad \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2 \ge \frac{n}{k^2},$$

then

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k \ge 1,$$

with equality if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1/k$.

Proof. We are going to show the equivalent statement

$$\left.\begin{array}{l}\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_n \ge 0\\ \alpha_1 + \alpha_2 + \cdots + \alpha_k \ \le 1\end{array}\right\} \Longrightarrow \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2 \le \frac{n}{k^2},$$

with equality if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1/k$. This is the problem of maximizing the convex function

$$f(\alpha_1, \alpha_2, \ldots, \alpha_n) := \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2$$

over the convex polygon

$$\mathcal{C} := \{(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n : \alpha_1 \geq \cdots \geq \alpha_n \geq 0 \text{ and } \alpha_1 + \cdots + \alpha_k \leq 1\}.$$

Because any point in $\mathcal{C}$ is a convex combination of its vertices (so that the extreme points of $\mathcal{C}$ are vertices) and because the function f is convex, the maximum is attained at a vertex of $\mathcal{C}$ by Theorem B.16. The vertices of $\mathcal{C}$ are obtained as intersections of n hyperplanes arising by turning n of the $(n+1)$ inequality constraints into equalities. Thus, we have the following possibilities:

- if $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$, then $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = 0$;
- if $\alpha_1 + \cdots + \alpha_k = 1$ and $\alpha_1 = \cdots = \alpha_\ell > \alpha_{\ell+1} = \cdots = \alpha_n = 0$ for $1 \leq \ell \leq k$, then $\alpha_1 = \cdots = \alpha_\ell = 1/\ell$, and consequently $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = 1/\ell$;
- if $\alpha_1 + \cdots + \alpha_k = 1$ and $\alpha_1 = \cdots = \alpha_\ell > \alpha_{\ell+1} = \cdots = \alpha_n = 0$ for $k < \ell \leq n$, then $\alpha_1 = \cdots = \alpha_\ell = 1/k$, and consequently $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = \ell/k^2$.

Taking $k < \sqrt{n}$ into account, it follows that

$$\max_{(\alpha_1,\ldots,\alpha_n)\in\mathcal{C}} f(\alpha_1, \ldots, \alpha_n) = \max\Big\{ \max_{1\leq\ell\leq k} \frac{1}{\ell}, \max_{k<\ell\leq n} \frac{\ell}{k^2} \Big\} = \max\Big\{1, \frac{n}{k^2}\Big\} = \frac{n}{k^2},$$

with equality only in the case $\ell = n$ where $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1/k$. □

Proof (of Theorem 5.8). Let us recall from (5.8) that we have

$$\sum_{i,j=1, i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \geq \frac{N^2}{m} - N = \frac{N(N-m)}{m},$$

which yields

$$\max_{i\in[N]} \sum_{j=1, j\neq i}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \geq \frac{1}{N} \sum_{i,j=1, i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \geq \frac{N-m}{m}.$$

For an index $i^* \in [N]$ achieving the latter maximum, we reorder the sequence $(|\langle \mathbf{a}_{i^*}, \mathbf{a}_j \rangle|)_{j=1, j\neq i^*}^{N}$ as $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_{N-1} \geq 0$, so that

$$\beta_1^2 + \beta_2^2 + \cdots + \beta_{N-1}^2 \geq \frac{N-m}{m}.$$

Lemma 5.9 with $n = N-1$, $k = s$, and $\alpha_\ell := \big(\sqrt{m(N-1)/(N-m)}/s\big)\beta_\ell$ gives $\alpha_1 + \alpha_2 + \cdots + \alpha_s \geq 1$. It follows that

$$\mu_1(s) \geq \beta_1 + \beta_2 + \cdots + \beta_s \geq s\sqrt{\frac{N-m}{m(N-1)}},$$

as announced. Let us now assume that equality holds in (5.10), so that all the previous inequalities are in fact equalities. As in the proof of Theorem 5.7, equality in (5.8) implies that the system $(\mathbf{a}_1, \dots, \mathbf{a}_N)$ is a tight frame. The case of equality in Lemma 5.9 implies that $|\langle \mathbf{a}_{i^*}, \mathbf{a}_j \rangle| = \sqrt{(N-m)/(m(N-1))}$ for all $j \in [N]$, $j \neq i^*$. Since the index i^* can be arbitrarily chosen in $[N]$, the system $(\mathbf{a}_1, \dots, \mathbf{a}_N)$ is also equiangular. Conversely, the proof that equiangular tight frames yields equality in (5.10) follows easily from Theorem 5.7 and (5.2). □

In compressive sensing, we are not only interested in small coherence, but also in $m \times N$ matrices where N is much larger than m. This restriction makes it impossible to meet the Welch bound. Indeed, the next theorem shows that the number of vectors in an equiangular tight frame — or in an equiangular system, for that matter — cannot be arbitrarily large.

Theorem 5.10. *The cardinality N of an equiangular system $(\mathbf{a}_1, \dots, \mathbf{a}_N)$ of ℓ_2-normalized vectors in $\mathbb{K}^m$ satisfies*

$$\begin{aligned} N &\leq \frac{m(m+1)}{2} && \text{when } \mathbb{K} = \mathbb{R}\,, \\ N &\leq m^2 && \text{when } \mathbb{K} = \mathbb{C}\,. \end{aligned}$$

If equality is achieved, then the system $(\mathbf{a}_1, \dots, \mathbf{a}_N)$ is a also tight frame.

We will use the following simple lemma twice in the proof of this theorem.

Lemma 5.11. *For any $z \in \mathbb{C}$, the $n \times n$ matrix*

$$\begin{bmatrix} 1 & z & z & \cdots & z \\ z & 1 & z & \cdots & z \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ z & \cdots & z & 1 & z \\ z & \cdots & z & z & 1 \end{bmatrix}$$

admits $1 + (n-1)z$ as a single eigenvalue and $1 - z$ as a multiple eigenvalue of multiplicity $n - 1$.

Proof. Summing the columns of the matrix, we see that the vector $[1, \dots, 1]^\top$ is an eigenvector for the eigenvalue $1 + (n-1)z$. Then, subtracting from the first column each subsequent column, we also see that the $(n-1)$ linearly independent vectors $[1, -1, 0, \dots, 0]^\top$, $[1, 0, -1, 0, \dots, 0]^\top$, …, $[1, 0 \dots, 0, -1]^\top$ are eigenvectors for the eigenvalue $1 - z$. The proof is now complete. □

Proof (of Theorem 5.10). The key point is to lift our considerations from the space $\mathbb{K}^m$ to a subspace $\mathcal{S}_m$ of operators on $\mathbb{K}^m$. In the case $\mathbb{K} = \mathbb{R}$, $\mathcal{S}_m$ is the space of symmetric operators on $\mathbb{R}^m$, and in the case $\mathbb{K} = \mathbb{C}$, $\mathcal{S}_m$ is simply the space of operators on $\mathbb{C}^m$. (it is tempting to consider hermitian operators, but

they do not form a linear space). These spaces are endowed with the Frobenius inner product

$$\langle \mathbf{P}, \mathbf{Q} \rangle_F = \operatorname{tr}(\mathbf{P}\mathbf{Q}^*) , \qquad \mathbf{P}, \mathbf{Q} \in \mathcal{S}_m .$$

Let us introduce the orthogonal projectors $\mathbf{P}_1, \dots, \mathbf{P}_N \in \mathcal{S}_m$ onto the lines spanned by $\mathbf{a}_1, \dots, \mathbf{a}_N$. These operators are defined, for $i \in [N]$, by

$$\mathbf{P}_i(\mathbf{v}) = \langle \mathbf{v}, \mathbf{a}_i \rangle \, \mathbf{a}_i , \qquad \mathbf{v} \in \mathbb{R}^m .$$

Denote by c the common magnitude of the inner products $\langle \mathbf{a}_i, \mathbf{a}_j \rangle$, $i \neq j$, and by $(\mathbf{e}_1, \dots, \mathbf{e}_m)$ the canonical basis of $\mathbb{K}^m$. Using that $\mathbf{P}_i^2 = \mathbf{P}_i = \mathbf{P}_i^*$, we calculate, for $i, j \in [N], i \neq j$,

$$\begin{aligned}
\langle \mathbf{P}_i, \mathbf{P}_i \rangle_F &= \operatorname{tr}(\mathbf{P}_i \mathbf{P}_i^*) = \operatorname{tr}(\mathbf{P}_i) = \sum_{k=1}^m \langle \mathbf{P}_i(\mathbf{e}_k), \mathbf{e}_k \rangle = \sum_{k=1}^m \langle \mathbf{e}_k, \mathbf{a}_i \rangle \langle \mathbf{a}_i, \mathbf{e}_k \rangle \\
&= \sum_{k=1}^m |\langle \mathbf{a}_i, \mathbf{e}_k \rangle|^2 = \|\mathbf{a}_i\|_2^2 = 1, \\
\langle \mathbf{P}_i, \mathbf{P}_j \rangle_F &= \operatorname{tr}(\mathbf{P}_i \mathbf{P}_j^*) = \operatorname{tr}(\mathbf{P}_i \mathbf{P}_j) = \sum_{k=1}^m \langle \mathbf{P}_i \mathbf{P}_j(\mathbf{e}_k), \mathbf{e}_k \rangle = \sum_{k=1}^m \langle \mathbf{P}_j(\mathbf{e}_k), \mathbf{P}_i(\mathbf{e}_k) \rangle \\
&= \sum_{k=1}^m \langle \mathbf{e}_k, \mathbf{a}_j \rangle \overline{\langle \mathbf{e}_k, \mathbf{a}_i \rangle} \langle \mathbf{a}_j, \mathbf{a}_i \rangle = \overline{\langle \mathbf{a}_i, \mathbf{a}_j \rangle} \Big\langle \sum_{k=1}^m \langle \mathbf{a}_i, \mathbf{e}_k \rangle \mathbf{e}_k, \mathbf{a}_j \Big\rangle \\
&= \overline{\langle \mathbf{a}_i, \mathbf{a}_j \rangle} \langle \mathbf{a}_i, \mathbf{a}_j \rangle = |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 = c^2.
\end{aligned}$$

Thus, the Gram matrix of the system $(\mathbf{P}_1, \dots, \mathbf{P}_N)$ is the $N \times N$ matrix

$$\begin{bmatrix}
1 & c^2 & c^2 & \cdots & c^2 \\
c^2 & 1 & c^2 & \cdots & c^2 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
c^2 & \cdots & c^2 & 1 & c^2 \\
c^2 & \cdots & c^2 & c^2 & 1
\end{bmatrix} .$$

In view of $0 \leq c^2 < 1$, Lemma 5.11 implies that this Gram matrix is invertible, which means that the system $(\mathbf{P}_1, \dots, \mathbf{P}_N)$ is linearly independent. But this system lies in the space $\mathcal{S}_m$, which has dimension $m(m+1)/2$ when $\mathbb{K} = \mathbb{R}$ and dimension m^2 when $\mathbb{K} = \mathbb{C}$. Therefore, we obtain

$$\begin{aligned}
N &\leq \frac{m(m+1)}{2} && \text{when } \mathbb{K} = \mathbb{R}, \\
N &\leq m^2 && \text{when } \mathbb{K} = \mathbb{C}.
\end{aligned}$$

Let us now assume that equality holds. Then the system $(\mathbf{Id}_m, \mathbf{P}_1, \dots, \mathbf{P}_N)$ is linearly dependent, hence the determinant of its Gram matrix vanishes. This translates into

$$\begin{vmatrix} m & 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & c^2 & c^2 & \cdots & c^2 \\ 1 & c^2 & 1 & c^2 & \cdots & c^2 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & c^2 & \cdots & c^2 & 1 & c^2 \\ 1 & c^2 & \cdots & c^2 & c^2 & 1 \end{vmatrix} = 0 \, .$$

Subtracting the first row divided by m from all the other rows and expanding with respect to the first column, we derive the $N \times N$ identity

$$\begin{vmatrix} 1 & b & b & \cdots & b \\ b & 1 & b & \cdots & b \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ b & \cdots & b & 1 & b \\ b & \cdots & b & b & 1 \end{vmatrix} = 0, \qquad \text{where} \quad b := \frac{mc^2 - 1}{m-1} \, .$$

Since $1-b = m(1-c^2)/(m-1) \neq 0$, Lemma 5.11 implies that $1+(N-1)b = 0$, which reads after simplification

$$c^2 = \frac{N-m}{m(N-1)}.$$

This shows that the ℓ_2-normalized system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ meets the Welch bound. Thus, according to Theorem 5.7, it is an equiangular tight frame. □

The upper bounds on the number of vectors in an equiangular systems are sharp. For instance, equiangular systems of 6 vectors in $\mathbb{R}^3$ and of 28 vectors in $\mathbb{R}^7$ are given in Exercise 5.5, while equiangular systems of 4 vectors in $\mathbb{C}^2$ and of 9 vectors in $\mathbb{C}^3$ are given in Exercise 5.6. In contrast with $\mathbb{C}^m$, where systems of m^2 equiangular vectors in $\mathbb{C}^m$ seem to exist for all m, systems of $m(m+1)/2$ equiangular vectors in $\mathbb{R}^m$ do not exist for all m, as shown below. They are known to exist when m is equal to 2, 3, 7, and 23, but the cases of other allowed values are not settled.

Theorem 5.12. *For $m \geq 3$, if there is an equiangular system of $m(m+1)/2$ vectors in $\mathbb{R}^m$, then $m+2$ is necessarily the square of an odd integer.*

Proof. Let $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ be a system of $N = m(m+1)/2$ equiangular ℓ_2-normalized vectors. According to Theorem 5.10, this system is a tight frame, hence the matrix $\mathbf{A}$ with columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$ satisfies $\mathbf{A}\mathbf{A}^* = \lambda \mathbf{Id}_m$ for some $\lambda > 0$. Since the matrix $\mathbf{G} := \mathbf{A}^*\mathbf{A}$ has the same nonzero eigenvalues as $\mathbf{A}\mathbf{A}^*$, i.e., λ with multiplicity m, it also has zero as an eigenvalue of multiplicity $N-m$. Moreover, since $\mathbf{G}$ is the Gram matrix of the system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$, its diagonal entries all equal one, while its off-diagonal entries all have the same absolute value c. Consequently, the matrix $\mathbf{B} := (\mathbf{G} - \mathbf{Id}_N)/c$ has the form

$$\mathbf{B} = \begin{bmatrix} 0 & b_{1,2} & \cdots & b_{1,N} \\ b_{2,1} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_{N-1,N} \\ b_{N,1} & \cdots & b_{N,N-1} & 0 \end{bmatrix}, \qquad \text{where } b_{i,j} = \pm 1,$$

and has $-1/c$ as an eigenvalue of multiplicity $N - m$. Thus, its characteristic polynomial $P_{\mathbf{B}}(x) := \sum_{0 \le k \le N} \beta_k (-x)^k$, $\beta_N = 1$, has integer coefficients β_k and vanishes at $x = -1/c$. Given that

$$c = \sqrt{\frac{N-m}{m(N-1)}} = \sqrt{\frac{(m+1)/2 - 1}{m(m+1)/2 - 1}} = \sqrt{\frac{m-1}{m^2+m-2}} = \frac{1}{\sqrt{m+2}},$$

we have $P_{\mathbf{B}}(-\sqrt{m+2}) = 0$, i.e,

$$\left(\sum_{0 \le k \le N/2} b_{2k} (m+2)^k \right) + \sqrt{m+2} \left(\sum_{0 \le k \le (N-1)/2} b_{2k+1} (m+2)^k \right) = 0.$$

Noticing that the two sums above, denoted by Σ_1 and Σ_2, are both integers, we obtain the equality $\Sigma_1^2 = (m+2)\Sigma_2^2$, which shows that $m+2$ is a square, since any prime factor of $m+2$ must appear an even number of times in its prime factor decomposition. We now need to show that $n := \sqrt{m+2}$ is odd. Let us introduce the $N \times N$ matrix $\mathbf{J}_N$ whose entries are all equal to one. Its null space has dimension $N-1$, so it intersects the $(N-m)$-dimensional eigenspace of B corresponding to the eigenvalue $-1/c = -n$, since $N - 1 + N - m > N$ for $m \ge 3$, i.e., $N = m(m+1)/2 > m+1$. Consequently, the matrix $\mathbf{C} := (\mathbf{B} - \mathbf{Id}_N + \mathbf{J}_N)/2$ admits $-(n+1)/2$ as an eigenvalue. Its diagonal entries are all equal to zero, while its off-diagonal entries are all equal to zero or one. Thus, its characteristic polynomial $P_{\mathbf{C}}(x) := \sum_{k=0}^{N} \gamma_k (-x)^k$, $\gamma_N = 1$, has integer coefficients γ_k and vanishes at $x = -(n+1)/2$. The equality $P_{\mathbf{C}}(-(n+1)/2) = 0$ can be rewritten as

$$(n+1)^N = -\sum_{k=0}^{N-1} 2^{N-k} \gamma_k (n+1)^k.$$

This shows that $(n+1)^N$ is an even integer, hence so is $n+1$. This completes the proof that $n = \sqrt{m+2}$ is an odd integer. □

In the complex setting, it seems plausible that equiangular systems of $N = m^2$ vectors exists for all values of m. This would yield $m \times m^2$ matrices with coherence equal to $1/\sqrt{m+1}$, but no construction of such systems is known at the moment. We present below an explicit $m \times m^2$ matrix with coherence equal to $1/\sqrt{m}$ instead. Let us incidentally notice that $1/\sqrt{m}$ is the limit of the Welch bound when N goes to infinity.

Proposition 5.13. *For each prime number $m \geq 5$, there is an explicit $m \times m^2$ complex matrix with coherence $\mu = 1/\sqrt{m}$.*

Proof. Throughout the proof, we identify the set $[m]$ with $\mathbb{Z}/m\mathbb{Z} =: \mathbb{Z}_m$. For $k, \ell \in \mathbb{Z}_m$, we introduce the *translation* and *modulation* operators $\mathbf{T}_k$ and $\mathbf{M}_\ell$ defined, for $\mathbf{z} \in \mathbb{C}^{\mathbb{Z}_m}$ and $j \in \mathbb{Z}_m$, by

$$(\mathbf{T}_k \mathbf{z})_j = z_{j-k}, \qquad (\mathbf{M}_\ell \mathbf{z})_j = e^{i2\pi \ell j/m}\, z_j.$$

These operators are isometries of $\ell_2(\mathbb{Z}_m)$. We also introduce the so-called *Alltop* vector, which is the ℓ_2-normalized vector $\mathbf{x} \in \mathbb{C}^{\mathbb{Z}_m}$ defined by

$$x_j := \frac{1}{\sqrt{m}}\, e^{i2\pi j^3/m}, \qquad j \in \mathbb{Z}_m.$$

The explicit $m \times m^2$ matrix of the proposition is the one with columns $\mathbf{M}_\ell \mathbf{T}_k \mathbf{x}$, $k, \ell \in \mathbb{Z}_m$, i.e., the matrix

$$\Big[\mathbf{M}_1\mathbf{T}_1\mathbf{x}\,\Big|\cdots\Big|\,\mathbf{M}_1\mathbf{T}_m\mathbf{x}\,\Big|\,\mathbf{M}_2\mathbf{T}_1\mathbf{x}\,\Big|\cdots\Big|\,\cdots\,\Big|\,\mathbf{M}_m\mathbf{T}_1\mathbf{x}\,\Big|\cdots\Big|\,\mathbf{M}_m\mathbf{T}_m\mathbf{x}\Big].$$

The inner product of two different columns indexed by (k, ℓ) and (k', ℓ') is

$$\begin{aligned}
\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle &= \sum_{j \in \mathbb{Z}_m} (\mathbf{M}_\ell \mathbf{T}_k \mathbf{x})_j \overline{(\mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x})_j} \\
&= \sum_{j \in \mathbb{Z}_m} e^{i2\pi \ell j/m}\, x_{j-k} e^{-i2\pi \ell' j/m}\, \overline{x_{j-k'}} \\
&= \frac{1}{m} \sum_{j \in \mathbb{Z}_m} e^{i2\pi(\ell-\ell')j/m} e^{i2\pi((j-k)^3-(j-k')^3)/m}.
\end{aligned}$$

Setting $a := \ell - \ell'$ and $b := k - k'$, so that $(a, b) \neq (0, 0)$, we make the change of summation index $h = j - k'$ to obtain

$$\begin{aligned}
\big|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle\big| &= \frac{1}{m} \Big| e^{i2\pi a k'/m} \sum_{h \in \mathbb{Z}_m} e^{i2\pi a h/m} e^{i2\pi((h-b)^3 - h^3)/m} \Big| \\
&= \frac{1}{m} \Big| \sum_{h \in \mathbb{Z}_m} e^{i2\pi a h/m} e^{i2\pi(-3bh^2 + 3b^2 h - b^3)/m} \Big| \\
&= \frac{1}{m} \Big| \sum_{h \in \mathbb{Z}_m} e^{i2\pi(-3bh^2 + (a+3b^2)h)/m} \Big|.
\end{aligned}$$

We now set $c := -3b$ and $d := a + 3b^2$, and we look at the previous modulus squared. We have

$$\begin{aligned}\big|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x}\rangle\big|^2 &= \frac{1}{m^2} \sum_{h\in\mathbb{Z}_m} e^{i2\pi(ch^2+dh)/m} \sum_{h'\in\mathbb{Z}_m} e^{-i2\pi(ch'^2+dh')/m} \\ &= \frac{1}{m^2} \sum_{h,h'\in\mathbb{Z}_m} e^{i2\pi(h-h')(c(h+h')+d)/m} \\ &= \frac{1}{m^2} \sum_{h',h''\in\mathbb{Z}_m} e^{i2\pi h''(c(h''+2h')+d)/m} \\ &= \frac{1}{m^2} \sum_{h''\in\mathbb{Z}_m} e^{i2\pi h''(ch''+d)/m} \Big(\sum_{h'\in\mathbb{Z}_m} e^{i4\pi ch''h'/m} \Big).\end{aligned}$$

For each $h'' \in \mathbb{Z}_m$, we observe that

$$\sum_{h'\in\mathbb{Z}_m} e^{i4\pi ch''h'/m} = \begin{cases} m & \text{if } 2ch'' = 0 \mod m, \\ 0 & \text{if } 2ch'' \neq 0 \mod m. \end{cases}$$

Let us separate two cases:

1. $c = 0 \mod m$:
   since $c = -3b$ and $3 \neq 0 \mod m$, we have $b = 0$, hence $d := a + 3b^2 \neq 0 \mod m$, so that
   $$\big|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x}\rangle\big|^2 = \frac{1}{m} \sum_{h''\in\mathbb{Z}_m} e^{i2\pi dh''/m} = 0;$$
2. $c \neq 0 \mod m$:
   since $2 \neq 0 \mod m$, the equality $2ch'' = 0$ only occurs when $h'' = 0 \mod m$, so that
   $$\big|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x}\rangle\big|^2 = \frac{1}{m}.$$

This allows to conclude that the coherence of the matrix is equal to $1/\sqrt{m}$. □

5.3 Analysis of Orthogonal Matching Pursuit

We claimed at the beginning of this chapter that the performance of sparse recovery algorithms is enhanced by a small coherence. We justify this claim in the remaining sections. For instance, in view of (5.3), Theorems 5.14 and 5.15 guarantee the exact recovery of every s-sparse vector via orthogonal matching pursuit and via basis pursuit when the measurement matrix has a coherence $\mu < 1/(2s-1)$. We focus on the orthogonal matching pursuit algorithm in this section.

Theorem 5.14. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns. If*

$$\mu_1(s) + \mu_1(s-1) < 1, \tag{5.11}$$

then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ after at most s iterations of orthogonal matching pursuit.

Proof. Let $\mathbf{a}_1, \ldots, \mathbf{a}_N$ denote the ℓ_2-normalized columns of $\mathbf{A}$. According to Proposition 3.5, we need to prove that, for any $S \subseteq [N]$ with $\mathrm{card}(S) = s$, the matrix $\mathbf{A}_S$ is injective and that

$$\max_{j \in S} |\langle \mathbf{r}, \mathbf{a}_j \rangle| > \max_{\ell \in \overline{S}} |\langle \mathbf{r}, \mathbf{a}_\ell \rangle| \tag{5.12}$$

for all $\mathbf{r} \in \{\mathbf{A}\mathbf{z}, \mathrm{supp}(\mathbf{z}) \subseteq S\}$. Let then $\mathbf{r} := \sum_{i \in S} r_i \mathbf{a}_i$ be such a vector, and let $k \in S$ be chosen so that $|r_k| = \max_{i \in S} |r_i| > 0$. On the one hand, for $\ell \in \overline{S}$, we have

$$|\langle \mathbf{r}, \mathbf{a}_\ell \rangle| = \Big| \sum_{i \in S} r_i \langle \mathbf{a}_i, \mathbf{a}_\ell \rangle \Big| \le \sum_{i \in S} |r_i| |\langle \mathbf{a}_i, \mathbf{a}_\ell \rangle| \le |r_k| \mu_1(s).$$

On the other hand, we have

$$\begin{aligned} |\langle \mathbf{r}, \mathbf{a}_k \rangle| = \Big| \sum_{i \in S} r_i \langle \mathbf{a}_i, \mathbf{a}_k \rangle \Big| &\ge |r_k| \, |\langle \mathbf{a}_k, \mathbf{a}_k \rangle| - \sum_{i \in S, i \neq k} |r_i| \, |\langle \mathbf{a}_i, \mathbf{a}_k \rangle| \\ &\ge |r_k| - |r_k| \mu_1(s-1). \end{aligned}$$

Thus, (5.12) is fulfilled because $1 - \mu_1(s-1) > \mu_1(s)$ according to (5.11). Finally, the injectivity of $\mathbf{A}_S$ follows from Corollary 5.4. □

5.4 Analysis of Basis Pursuit

In this section, we show that a small coherence also guarantees the success of basis pursuit. As a matter of fact, any condition guaranteeing the unequivocal success of the recovery of all vectors supported on a set S via $\mathrm{card}(S)$ iterations of orthogonal matching pursuit also guarantees the success of the recovery of all vectors supported on S via basis pursuit. This follows from the fact that the exact recovery condition (3.7) implies the null space property (4.1). Indeed, given $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, we have $\mathbf{A}_S \mathbf{v}_S = -\mathbf{A}_{\overline{S}} \mathbf{v}_{\overline{S}}$, and

$$\|\mathbf{v}_S\|_1 = \|\mathbf{A}_S^\dagger \mathbf{A}_S \mathbf{v}_S\|_1 = \|\mathbf{A}_S^\dagger \mathbf{A}_{\overline{S}} \mathbf{v}_{\overline{S}}\|_1 \le \|\mathbf{A}_S^\dagger \mathbf{A}_{\overline{S}}\|_{1 \to 1} \|\mathbf{v}_{\overline{S}}\|_1 < \|\mathbf{v}_{\overline{S}}\|_1.$$

Thus, the following result is immediate. We nonetheless give an alternative self-contained proof.

Theorem 5.15. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be a matrix with ℓ_2-normalized columns. If*

$$\mu_1(s) + \mu_1(s-1) < 1, \tag{5.13}$$

then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ via basis pursuit.

Proof. According to Theorem 4.5, it is necessary and sufficient to prove that the matrix $\mathbf{A}$ satisfies the null space property of order s, i.e., that

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 \tag{5.14}$$

for any nonzero vector $\mathbf{v} \in \ker \mathbf{A}$ and any index set $S \subseteq [N]$ with $\mathrm{card}(S) = s$. If $\mathbf{a}_1, \dots, \mathbf{a}_N$ denote the columns of $\mathbf{A}$, then the condition $\mathbf{v} \in \ker \mathbf{A}$ translates into $\sum_{j=1}^N v_j \mathbf{a}_j = 0$. Thus, taking the inner product with a particular $\mathbf{a}_i$, $i \in S$, and isolating the term in v_i, we obtain

$$v_i = v_i \langle \mathbf{a}_i, \mathbf{a}_i \rangle = - \sum_{j=1, j \neq i}^N v_j \langle \mathbf{a}_j, \mathbf{a}_i \rangle = - \sum_{\ell \in \overline{S}} v_\ell \langle \mathbf{a}_\ell, \mathbf{a}_i \rangle - \sum_{j \in S, j \neq i} v_j \langle \mathbf{a}_j, \mathbf{a}_i \rangle .$$

It follows that

$$|v_i| \le \sum_{\ell \in \overline{S}} |v_\ell| \, |\langle \mathbf{a}_\ell, \mathbf{a}_i \rangle| + \sum_{j \in S, j \neq i} |v_j| \, |\langle \mathbf{a}_j, \mathbf{a}_i \rangle| .$$

Summing over all $i \in S$ and interchanging the summations, we derive

$$\begin{aligned} \|\mathbf{v}_S\|_1 = \sum_{i \in S} |v_i| &\le \sum_{\ell \in \overline{S}} |v_\ell| \sum_{i \in S} |\langle \mathbf{a}_\ell, \mathbf{a}_i \rangle| + \sum_{j \in S} |v_j| \sum_{i \in S, i \neq j} |\langle \mathbf{a}_j, \mathbf{a}_i \rangle| \\ &\le \sum_{\ell \in \overline{S}} |v_\ell| \, \mu_1(s) + \sum_{j \in S} |v_j| \, \mu_1(s-1) = \mu_1(s) \, \|\mathbf{v}_{\overline{S}}\|_1 + \mu_1(s-1) \, \|\mathbf{v}_S\|_1 . \end{aligned}$$

After rearrangement, this reads $(1 - \mu_1(s-1)) \, \|\mathbf{v}_S\|_1 \le \mu_1(s) \, \|\mathbf{v}_{\overline{S}}\|_1$, and (5.14) is fulfilled because $1 - \mu_1(s-1) > \mu_1(s)$, which is a rewriting of (5.13). □

Choosing a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with small coherence $\mu \le c/\sqrt{m}$, for instance the one of Theorem 5.13, we see that the condtion $(2s-1)\mu < 1$ ensuring recovery of s-sparse vectors via orthogonal matching pursuit as well as via ℓ_1-minimization is satisfied once

$$m \ge C s^2 . \tag{5.15}$$

This gives a first estimate of the required number of samples in terms of the sparsity for practical recovery algorithms and a specific matrix $\mathbf{A}$. One could be satisfied with this result at first sight. However, the sparsity s enters quadratically in this bound. Hence, for mildly large s this bound may be very pessimistic. We will see indeed later that a linear scaling of m in s is possible up to log-factors.

Let us point out that it is not possible to overcome the quadratic bottleneck in (5.15) using Theorems 5.14 and 5.15. Indeed, let us assume on the contrary that the sufficient condition $\mu_1(s) + \mu_1(s-1) < 1$ holds with $m \le (2s-1)^2/2$, say. Provided N is large, say $N \ge 2m$, we apply Theorem 5.8 to derive a contradiction from

$$1 > \mu_1(s) + \mu_1(s-1) \geq (2s-1)\sqrt{\frac{N-m}{m(N-1)}} \geq \sqrt{\frac{2(N-m)}{N-1}} \geq \sqrt{\frac{N}{N-1}}.$$

In the following chapters we will reduce the number of required measurements below the order s^2 by introducing new tools for the analysis of sparse recovery algorithms.

5.5 Analysis of Thresholding Algorithms

In this final section, we show that thresholding algorithms can also be analyzed using the coherence. For instance, under the same condition as before, even the basis thresholding algorithm will successfully recover sparse vectors that are flat on their support.

Theorem 5.16. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be a matrix with ℓ_2-normalized columns and let $\mathbf{x} \in \mathbb{C}^N$ be a vector supported on a set S of size s. If*

$$\mu_1(s) + \mu_1(s-1) < \frac{\min_{i \in S} |x_i|}{\max_{i \in S} |x_i|}, \tag{5.16}$$

then the vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ via basic thresholding.

Proof. Let $\mathbf{a}_1, \ldots, \mathbf{a}_N$ denote the ℓ_2-normalized columns of $\mathbf{A}$. According to Proposition 3.7, we need to prove that, for any $j \in S$ and any $\ell \in \overline{S}$,

$$|\langle \mathbf{A}\mathbf{x}, \mathbf{a}_j \rangle| > |\langle \mathbf{A}\mathbf{x}, \mathbf{a}_\ell \rangle|. \tag{5.17}$$

We observe that

$$\begin{aligned}
|\langle \mathbf{A}\mathbf{x}, \mathbf{a}_\ell \rangle| &= |\sum_{i \in S} x_i \langle \mathbf{a}_i, \mathbf{a}_\ell \rangle| \leq \sum_{i \in S} |x_i| |\langle \mathbf{a}_i, \mathbf{a}_\ell \rangle| \leq \mu_1(s) \max_{i \in S} |x_i|, \\
|\langle \mathbf{A}\mathbf{x}, \mathbf{a}_j \rangle| &= |\sum_{i \in S} x_i \langle \mathbf{a}_i, \mathbf{a}_j \rangle| \geq |x_j| - \sum_{i \in S, i \neq j} |x_i| |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| \\
&\geq \min_{i \in S} |x_i| - \mu_1(s-1) \max_{i \in S} |x_i|.
\end{aligned}$$

Thus, taking (5.16) into account, we obtain

$$|\langle \mathbf{A}\mathbf{x}, \mathbf{a}_j \rangle| - |\langle \mathbf{A}\mathbf{x}, \mathbf{a}_\ell \rangle| \geq \min_{i \in S} |x_i| - (\mu(s) + \mu_1(s-1)) \max_{i \in S} |x_i| > 0.$$

This shows (5.17) and concludes the proof. □

We now turn to the more involved hard thresholding pursuit algorithm. Just as for orthogonal matching pursuit, we show that s iterations are enough for the recovery of s-sparse vectors under a condition rather similar to (5.11). In view of (5.2), we observe that the condition in question is met when the coherence of the measurement matrix satisfies $\mu < 1/(3s-1)$.

Theorem 5.17. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns. If*

$$2\mu_1(s) + \mu_1(s-1) < 1,$$

then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ after at most s iterations of hard thresholding pursuit.

Proof. Let us consider indices $j_1, j_2, \dots, j_N$ such that

$$|x_{j_1}| \geq |x_{j_2}| \geq \cdots \geq |x_{j_s}| > |x_{j_{s+1}}| = \cdots = |x_{j_N}| = 0.$$

We are going to prove that, for $0 \leq n \leq s-1$, the set $\{j_1, \dots, j_{n+1}\}$ is included in S^{n+1} defined by (HTP_1) with $\mathbf{y} = \mathbf{A}\mathbf{x}$ as the set of largest absolute entries of

$$\mathbf{z}^{n+1} := \mathbf{x}^n + \mathbf{A}^*\mathbf{A}(\mathbf{x} - \mathbf{x}^n)\,. \tag{5.18}$$

This will imply that $S^s = S = \operatorname{supp}\mathbf{x}$, and consequently that $\mathbf{x}^s = \mathbf{x}$ by (HTP_2). Note that it is sufficient to prove that

$$\min_{1\leq k\leq n+1} |z^{n+1}_{j_k}| > \max_{\ell\in\overline{S}} |z^{n+1}_\ell|\,. \tag{5.19}$$

We notice that, for every $j \in [N]$,

$$z^{n+1}_j = x^n_j + \sum_{i=1}^N (x_i - x^n_i)\langle \mathbf{a}_i, \mathbf{a}_j\rangle = x_j + \sum_{i\neq j} (x_i - x^n_i)\langle \mathbf{a}_i, \mathbf{a}_j\rangle\,.$$

Therefore, we have

$$|z^{n+1}_j - x_j| \leq \sum_{i\in S^n, i\neq j} |x_i - x^n_i||\langle \mathbf{a}_i, \mathbf{a}_j\rangle| + \sum_{i\in S\setminus S^n, i\neq j} |x_i||\langle \mathbf{a}_i, \mathbf{a}_j\rangle|\,. \tag{5.20}$$

We derive, for $1 \leq k \leq n+1$ and $\ell \in \overline{S}$, that

$$|z^{n+1}_{j_k}| \geq |x_{j_k}| - \mu_1(s)\|(\mathbf{x} - \mathbf{x}^n)_{S^n}\|_\infty - \mu_1(s)\|\mathbf{x}_{S\setminus S^n}\|_\infty\,, \tag{5.21}$$

$$|z^{n+1}_\ell| \leq \mu_1(s)\|(\mathbf{x} - \mathbf{x}^n)_{S^n}\|_\infty + \mu_1(s)\|\mathbf{x}_{S\setminus S^n}\|_\infty\,. \tag{5.22}$$

In particular, for n=0, substituting $\|(\mathbf{x} - \mathbf{x}^n)_{S^n}\|_\infty = 0$ into (5.21) and (5.22) gives

$$|z^1_{j_1}| \geq (1 - \mu_1(s))\|\mathbf{x}\|_\infty > \mu_1(s)\|\mathbf{x}\|_\infty \geq |z^1_\ell| \quad \text{for all } \ell \in \overline{S}\,,$$

by virtue of $2\mu_1(s) < 1$. Therefore, the base case of the inductive hypothesis (5.19) holds for $n = 0$. Let us now assume that this hypothesis holds for $n-1$ with $n \geq 1$. This implies that $\{j_1, \dots, j_n\} \subseteq S^n$. We notice that ($\mathrm{HTP}_2$) with n replaced by $n-1$ says that the residual $\mathbf{y} - \mathbf{A}\mathbf{x}^n$ is orthogonal to the space $\{\mathbf{A}\mathbf{z}, \operatorname{supp}(\mathbf{z}) \subseteq S^n\}$, i.e., that $0 = \langle \mathbf{y} - \mathbf{A}\mathbf{x}^n, \mathbf{A}\mathbf{z}\rangle = \langle \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n), \mathbf{z}\rangle$ for any $\mathbf{z} \in \mathbb{C}^N$ supported on S^n. In view of $\mathbf{y} = \mathbf{A}\mathbf{x}$, this means

$$(\mathbf{A}^*\mathbf{A}(\mathbf{x}-\mathbf{x}^n))_{S^n} = 0\,. \tag{5.23}$$

Hence, for any $j \in S^n$, the definition (5.18) of $\mathbf{z}^{n+1}$ implies that $z_j^{n+1} = x_j^n$, and then (5.20) yields

$$|x_j^n - x_j| \le \mu_1(s-1)\|(\mathbf{x}-\mathbf{x}^n)_{S^n}\|_\infty + \mu_1(s-1)\|\mathbf{x}_{S\setminus S^n}\|_\infty\,.$$

Taking the maximum over $j \in S^n$ and rearranging gives

$$\|(\mathbf{x}-\mathbf{x}^n)_{S^n}\|_\infty \le \frac{\mu_1(s-1)}{1-\mu_1(s-1)}\|\mathbf{x}_{S\setminus S^n}\|_\infty\,. \tag{5.24}$$

Substituting the latter into (5.21) and (5.22), we obtain, for $1 \le k \le n+1$ and $\ell \in \overline{S}$,

$$|z_{j_k}^{n+1}| \ge \Big(1 - \frac{\mu_1(s)}{1-\mu_1(s-1)}\Big)|x_{j_{n+1}}|\,,$$
$$|z_\ell^{n+1}| \le \frac{\mu_1(s)}{1-\mu_1(s-1)}\,|x_{j_{n+1}}|\,.$$

Since $\mu_1(s)/(1-\mu_1(s-1)) < 1/2$, this shows that (5.19) holds for n, too. The proof by induction is now complete. □

Notes

The analysis of sparse recovery algorithms could be carried out using merely the coherence. For instance, the conclusion of Theorem 5.15 can be achieved under the sufficient condition $\mu < 1/(2s-1)$, as obtained by R. Gribonval and M. Nielsen in [206]. Theorems 5.14 and 5.15 in their present form were established by J. Tropp in [414]. What we call ℓ_1-coherence function here is called cumulative coherence function there. This concept also appears under the name Babel function. A straightforward extension to any $p > 0$ would be the ℓ_p-coherence function of a matrix $\mathbf{A}$ with ℓ_2-normalized columns $\mathbf{a}_1, \dots, \mathbf{a}_N$ defined by

$$\mu_p(s) := \max_{i\in[N]} \max\Big\{\Big(\sum_{j\in S}|\langle \mathbf{a}_i, \mathbf{a}_j\rangle|^p\Big)^{1/p},\ S \subseteq [N],\ \mathrm{card}(S) = s,\ i \notin S\Big\}\,.$$

Theorem 5.8 on the Welch-type lower bound for the ℓ_1-coherence function appeared in [384]. The matrix considered in Proposition 5.13, with m rows, m^2 columns and whose coherence equal to $1/\sqrt{m}$, is taken from [9, 395]. In [216], S. Gurevich, R. Hadani, and N. Sochen uncovered a matrix with p rows, p being a prime number, roughly p^5 columns, and whose coherence is bounded above by $4/\sqrt{p}$. Another number theoretic construction of $p \times p^k$ matrices, $p > k$ being a prime, with coherence $\mu \le \frac{k-1}{\sqrt{p}}$ can be found, for instance, in [410, Chapter 5.7.4].

There is a vast literature dedicated to frames. The notion is not restricted to the finite-dimensional setting, although this is the only one we considered. Good starting places to learn about the subject are O. Christensen's books [98] and [99]. As mentioned in the text, not everything is known about equiangular tight frames. In particular, whether equiangular systems of m^2 vectors in $\mathbb{C}^m$ exist for all values of m is not known — the numerical experiments performed for $m \leq 45$ by J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves in [363] seem to indicate that they do. More details on the subject of equiangular tight frames, and more generally tight frames in finite dimension, can be found in S. Waldron's book [441].

Exercises

5.1. The *mutual coherence* between two orthonormal bases $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ of $\mathbb{C}^m$ is defined as

$$\mu(\mathbf{U}, \mathbf{V}) := \sqrt{m} \max_{1 \leq i,j \leq m} |\langle \mathbf{u}_i, \mathbf{v}_j \rangle|.$$

Establish the inequalities

$$1 \leq \mu(\mathbf{U}, \mathbf{V}) \leq \sqrt{m}.$$

and prove that they are sharp.

5.2. Prove the equivalence of the three conditions of Definition 5.6, and find the value of the constant λ when the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N$ are ℓ_2-normalized.

5.3. Establish the alternative expressions for the ℓ_1-coherence function

$$\mu_1(s) = \max_{\mathrm{card}(S) \leq s+1} \|\mathbf{A}_S^* \mathbf{A}_S - I\|_{1 \to 1} = \max_{\mathrm{card}(S) \leq s+1} \|\mathbf{A}_S^* \mathbf{A}_S - I\|_{\infty \to \infty}.$$

5.4. Prove that the $m+1$ vertices of a regular simplex in $\mathbb{R}^m$ centered at the origin form an equiangular tight frame.

5.5. With $c := (\sqrt{5} - 1)/2$, prove that the columns of the matrix

$$\begin{bmatrix} 1 & 0 & c & 1 & 0 & -c \\ c & 1 & 0 & -c & 1 & 0 \\ 0 & c & 1 & 0 & -c & 1 \end{bmatrix}$$

form an equiangular system of 6 vectors in $\mathbb{R}^3$. Prove also that the vectors obtained by unit cyclic shifts on four vectors $[1, \pm 1, 0, \pm 1, 0, 0, 0]^\top$ form an equiangular system of 28 vectors in $\mathbb{R}^7$.

5.6. With $c := e^{i\pi/4}\sqrt{2-\sqrt{3}}$, prove that the columns of the matrix

$$\begin{bmatrix} 1 & c & 1 & -c \\ c & 1 & -c & 1 \end{bmatrix}$$

form an equiangular system of 4 vectors in $\mathbb{C}^2$. With $\omega := e^{i2\pi/3}$, prove also that the columns of the matrix

$$\begin{bmatrix} -2 & 1 & 1 & -2 & \omega^2 & \omega & -2 & \omega & \omega^2 \\ 1 & -2 & 1 & \omega & -2 & \omega^2 & \omega^2 & -2 & \omega \\ 1 & 1 & -2 & \omega^2 & \omega & -2 & \omega & \omega^2 & -2 \end{bmatrix}$$

form an equiangular system of 9 vectors in $\mathbb{C}^3$.

5.7. Prove that the columns of the matrix considered in Proposition 5.13 form a tight frame.

5.8. Suppose that a known vector is an s-sparse linear combination of vectors from the canonical and Fourier bases $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_m)$ and $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_m)$, defined as

$$\mathbf{e}_k = [0, \dots, 0, \underbrace{1}_{\text{index } k}, 0, \dots, 0]^\top, \qquad \mathbf{f}_k = \frac{1}{\sqrt{m}}[1, e^{i2\pi k/m}, \dots, e^{i2\pi k(m-1)/m}]^\top.$$

Prove that the unknown coefficients can be found by orthogonal matching pursuit or basis pursuit if $s < (\sqrt{m}+1)/2$.

5.9. Given $\nu < 1/2$, suppose that a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies

$$\mu_1(s) \le \nu$$

Prove that, for any $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \le \eta$, a minimizer $\mathbf{x}^\star$ of $\|\mathbf{z}\|_1$ subject to $\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta$ approximates the vector $\mathbf{x}$ with ℓ_1-error

$$\|\mathbf{x} - \mathbf{x}^\star\|_1 \le C\,\sigma_s(\mathbf{x})_1 + D\,s\,\eta,$$

for some positive constants C and D depending only on ν.

6

Restricted Isometry Constants

The coherence is a simple and useful measure of the quality of a measurement matrix. However, the lower bound on the coherence in Theorem 5.7 limits the performance analysis of recovery algorithms to rather small sparsity levels. A finer measure of the quality of a measurement matrix is needed to overcome this limitation. This is provided by the concept of *restricted isometry property*, also known as *uniform uncertainty principle*. It ensures the success of the sparse recovery algorithms presented in this book. Restricted isometry constants are introduced in Section 6.1. The success of sparse recovery is the established under some conditions on these constants for basis pursuit in Section 6.2, for thresholding-based algorithms in Section 6.3, and for greedy algorithms in Section 6.4.

6.1 Definitions and Basic Properties

Unlike the coherence, which only takes pairs of columns of a matrix into account, the restricted isometry constant of order s involves all s-tuples of columns and is therefore more suited to assess the quality of the matrix. As with the coherence, small restricted isometry constants are desired. Here is their formal definition.

Definition 6.1. *The sth* restricted isometry constant $\delta_s = \delta_s(\mathbf{A})$ *of a matrix* $\mathbf{A} \in \mathbb{C}^{m \times N}$ *is the smallest* $\delta \geq 0$ *such that*

$$(1 - \delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta)\|\mathbf{x}\|_2^2 \tag{6.1}$$

for all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$*. Equivalently, it is given by*

$$\delta_s = \max_{S \subseteq [N], \operatorname{card}(S) \leq s} \|\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}\|_{2 \to 2}. \tag{6.2}$$

We say that $\mathbf{A}$ satisfies the *restricted isometry property* if δ_s is small for reasonably large s — the meaning of small δ_s and large s will be made precise later.

We make a few remarks before establishing the equivalence of these two definitions. The first one is that the sequence of restricted isometry constants is nondecreasing, i.e.,

$$\delta_1 \le \delta_2 \le \cdots \le \delta_s \le \delta_{s+1} \le \cdots \le \delta_N.$$

The second one is that, although $\delta_s \ge 1$ is not forbidden, the relevant situation occurs for $\delta_s < 1$. Indeed, (6.2) says that each column-submatrix $\mathbf{A}_S$, $S \subseteq [N]$ with $\mathrm{card}(S) \le s$, has all its singular values in the interval $[1-\delta_s, 1+\delta_s]$, and is therefore injective when $\delta_s < 1$. In fact, $\delta_{2s} < 1$ is more relevant, since the inequality (6.1) yields $\|\mathbf{A}(\mathbf{x}-\mathbf{x}')\|_2^2 > 0$ for all distinct s-sparse vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{C}^N$, hence distinct s-sparse vectors have distinct measurement vectors. The third and final remark is that, if the entries of the measurement matrix $\mathbf{A}$ are real, then δ_s could also be defined as the smallest $\delta \ge 0$ such that (6.1) holds for all real vectors $\mathbf{x} \in \mathbb{R}^N$. This is because the operator norms of the real symmetric matrix $\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}$ relative $\ell_2(\mathbb{R})$ and to $\ell_2(\mathbb{C})$ are equal — both to its largest eigenvalues in modulus — and because the two definitions of restricted isometry constants would be equivalent in the real setting, too. Here is the argument for (6.2) to be adapted from the complex setting. We start by noticing that (6.1) is equivalent to

$$\big| \|\mathbf{A}_S \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \big| \le \delta \|\mathbf{x}\|_2^2 \quad \text{for all } S \subseteq [N], \ \mathrm{card}(S) \le s, \ \text{and all } \mathbf{x} \in \mathbb{C}^S.$$

We then observe that, for $\mathbf{x} \in \mathbb{C}^S$,

$$\|\mathbf{A}_S \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 = \langle \mathbf{A}_S \mathbf{x}, \mathbf{A}_S \mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{x} \rangle = \langle (\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}) \mathbf{x}, \mathbf{x} \rangle.$$

Since the matrix $(\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id})$ is Hermitian, we have

$$\max_{\mathbf{x} \in \mathbb{C}^S \setminus \{0\}} \frac{\langle (\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}) \mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|_2} = \|\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}\|_{2 \to 2},$$

so that (6.1) is equivalent to

$$\max_{S \subseteq [N], \mathrm{card}(S) \le s} \|\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}\|_{2 \to 2} \le \delta.$$

This proves the identity (6.2), as δ_s is the smallest such δ.

It is now possible to compare the restricted isometry constants of a matrix with its coherence μ and coherence function μ_1, see Definition 5.1 and 5.2.

Proposition 6.2. *If the matrix $\mathbf{A}$ has ℓ_2-normalized columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$, i.e., $\|\mathbf{a}_j\|_2 = 1$ for all $j \in [N]$, then*

$$\delta_1 = 0, \qquad \delta_2 = \mu, \qquad \delta_s \le \mu_1(s-1) \le (s-1)\mu, \quad s \ge 2.$$

Proof. The ℓ_2-normalization of the columns means that $\|\mathbf{A}\mathbf{e}_j\|_2^2 = \|\mathbf{e}_j\|_2^2$ for all $j \in [N]$, that is to say $\delta_1 = 0$. Next, with $\mathbf{a}_1, \ldots, \mathbf{a}_N$ denoting the columns of the matrix $\mathbf{A}$, we have

$$\delta_2 = \max_{1 \le i \ne j \le N} \|\mathbf{A}^*_{\{i,j\}}\mathbf{A}_{\{i,j\}} - \mathbf{Id}\|_{2\to 2}, \quad \text{where } \mathbf{A}^*_{\{i,j\}}\mathbf{A}_{\{i,j\}} = \begin{bmatrix} 1 & \langle \mathbf{a}_j, \mathbf{a}_i \rangle \\ \langle \mathbf{a}_i, \mathbf{a}_j \rangle & 1 \end{bmatrix}.$$

The eigenvalues of the matrix $\mathbf{A}^*_{\{i,j\}}\mathbf{A}_{\{i,j\}} - \mathbf{Id}$ are $|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$ and $-|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$, so its operator norm is $|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$. Taking the maximum over $1 \le i \ne j \le N$ yields the equality $\delta_2 = \mu$. The inequality $\delta_s \le \mu_1(s-1) \le (s-1)\mu$ follows from Theorem 5.3. □

In view of the existence of $m \times m^2$ matrices with coherence μ equal to $1/\sqrt{m}$, see Chapter 5, this already shows the existence of $m \times m^2$ matrices with restricted isometry constant $\delta_s < 1$ for $s \le \sqrt{m}$. We will establish that, given $\delta < 1$, there exist $m \times N$ matrices with restricted isometry constant $\delta_s \le \delta$ for $s \le c\, m/\ln(eN/m)$, where c is a constant depending only on δ, see Chapter 9. This is essentially the largest range possible, see Chapter 10. Matrices with a small restricted isometry constant of this optimal order are informally said to satisfy the *restricted isometry property*, or *uniform uncertainty principle*.

We now make a simple but essential observation, which motivates the related notion of restricted orthogonality constant.

Proposition 6.3. *Let $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ be vectors with $\|\mathbf{u}\|_0 \le s$ and $\|\mathbf{v}\|_0 \le t$. If $\operatorname{supp}(\mathbf{u}) \cap \operatorname{supp}(\mathbf{v}) = \emptyset$, then*

$$|\langle \mathbf{Au}, \mathbf{Av} \rangle| \le \delta_{s+t} \|\mathbf{u}\|_2 \|\mathbf{v}\|_2. \tag{6.3}$$

Proof. Let $S := \operatorname{supp}(\mathbf{u}) \cup \operatorname{supp}(\mathbf{v})$, and let $\mathbf{u}_S, \mathbf{v}_S \in \mathbb{C}^S$ be the restrictions of $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ to S. Since $\mathbf{u}$ and $\mathbf{v}$ have disjoint supports, we have $\langle \mathbf{u}_S, \mathbf{v}_S \rangle = 0$. We derive

$$\begin{aligned} |\langle \mathbf{Au}, \mathbf{Av} \rangle| &= |\langle \mathbf{A}_S\mathbf{u}_S, \mathbf{A}_S\mathbf{v}_S \rangle - \langle \mathbf{u}_S, \mathbf{v}_S \rangle| = |\langle (\mathbf{A}^*_S\mathbf{A}_S - \mathbf{Id})\mathbf{u}_S, \mathbf{v}_S \rangle| \\ &\le \|(\mathbf{A}^*_S\mathbf{A}_S - \mathbf{Id})\mathbf{u}_S\|_2 \|\mathbf{v}_S\|_2 \le \|\mathbf{A}^*_S\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \|\mathbf{u}_S\|_2 \|\mathbf{v}_S\|_2, \end{aligned}$$

and the conclusion follows from (6.2), $\|\mathbf{u}_S\|_2 = \|\mathbf{u}\|_2$, and $\|\mathbf{v}_S\|_2 = \|\mathbf{v}\|_2$. □

Definition 6.4. *The (s,t)-*restricted orthogonality constant *$\theta_{s,t} = \theta_{s,t}(\mathbf{A})$ of a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is the smallest $\theta \ge 0$ such that*

$$|\langle \mathbf{Au}, \mathbf{Av} \rangle| \le \theta \, \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \tag{6.4}$$

for all disjointly supported s-sparse and t-sparse vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$. Equivalently, it is given by

$$\theta_{s,t} = \max \big\{ \big\| \mathbf{A}^*_T \mathbf{A}_S \big\|_{2\to 2},\, S \cap T = \emptyset,\, \operatorname{card}(S) \le s,\, \operatorname{card}(T) \le t \big\}. \tag{6.5}$$

The justification of the equivalence between the two definitions is left as Exercise 6.4. We now give a comparison result between restricted isometry constants and restricted orthogonality constants.

Proposition 6.5. *Restricted isometry constants and restricted orthogonality constants are related by*

$$\theta_{s,t} \le \delta_{s+t} \le \frac{1}{s+t}\big(s\,\delta_s + t\,\delta_t + 2\sqrt{st}\,\theta_{s,t}\big).$$

The special case $t = s$ gives the inequalities

$$\theta_{s,s} \le \delta_{2s} \qquad \text{and} \qquad \delta_{2s} \le \delta_s + \theta_{s,s}.$$

Proof. The first inequality is Proposition 6.3. For the second inequality, given an $(s+t)$-sparse vector $\mathbf{x} \in \mathbb{C}^N$ with $\|\mathbf{x}\|_2 = 1$, we need to show that

$$\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \le \frac{1}{s+t}\big(s\,\delta_s + t\,\delta_t + 2\sqrt{st}\,\theta_{s,t}\big).$$

Let $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ be two disjointly supported vectors such that $\mathbf{u}+\mathbf{v} = \mathbf{x}$, where $\mathbf{u}$ is s-sparse and $\mathbf{v}$ is t-sparse, respectively. We have

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \langle \mathbf{A}(\mathbf{u}+\mathbf{v}), \mathbf{A}(\mathbf{u}+\mathbf{v})\rangle = \|\mathbf{A}\mathbf{u}\|_2^2 + \|\mathbf{A}\mathbf{v}\|_2^2 + 2\,\mathrm{Re}\langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}\rangle.$$

Taking $\|\mathbf{x}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$ into account, we derive

$$\begin{aligned}\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| &\le \big|\|\mathbf{A}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\big| + \big|\|\mathbf{A}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big| + 2\big|\langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}\rangle\big| \\ &\le \delta_s\|\mathbf{u}\|_2^2 + \delta_t\|\mathbf{v}\|_2^2 + 2\theta_{s,t}\|\mathbf{u}\|_2\|\mathbf{v}\|_2 = f\big(\|\mathbf{u}\|_2^2\big),\end{aligned}$$

where we have set, for $\alpha \in [0,1]$,

$$f(\alpha) := \delta_s\alpha + \delta_t(1-\alpha) + 2\theta_{s,t}\sqrt{\alpha(1-\alpha)}. \tag{6.6}$$

It can be shown that there is an $\alpha^* \in [0,1]$ such that this function is nondecreasing on $[0,\alpha^*]$ and then nonincreasing on $[\alpha^*,1]$ — see Exercise 6.5. Depending on the location of this α^* with respect to $s/(s+t)$, the function f is either nondecreasing on $[0, s/(s+t)]$ or nonincreasing on $[s/(s+t), 1]$. By properly choosing the vector $\mathbf{u}$, we can always assume that $\|\mathbf{u}\|_2^2$ is in one of these intervals. Indeed, if $\mathbf{u}$ is made of s smallest modulus components of $\mathbf{x}$ while $\mathbf{v}$ is made of t largest modulus components of $\mathbf{x}$, then we have

$$\frac{\|\mathbf{u}\|_2^2}{s} \le \frac{\|\mathbf{v}\|_2^2}{t} = \frac{1-\|\mathbf{u}\|_2^2}{t}, \qquad \text{so that } \|\mathbf{u}\|_2^2 \le \frac{s}{s+t},$$

and if $\mathbf{u}$ was made of s largest modulus components of $\mathbf{x}$, then we would likewise have $\|\mathbf{u}\|_2^2 \ge s/(s+t)$. This implies

$$\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \le f\Big(\frac{s}{s+t}\Big) = \delta_s\frac{s}{s+t} + \delta_t\frac{t}{s+t} + 2\theta_{s,t}\frac{\sqrt{st}}{s+t}.$$

The proof is complete. □

We close this section by proving that restricted isometry constants and restricted orthogonality constants of high order can be controlled by those of lower order.

Proposition 6.6. *For integers $r, s, t \geq 1$ with $t \geq s$,*

$$\theta_{t,r} \leq \sqrt{\frac{t}{s}}\,\theta_{s,r},$$
$$\delta_t \leq \frac{t-d}{s}\,\delta_{2s} + \frac{d}{s}\,\delta_s, \qquad \text{where } d := \gcd(s,t).$$

The special case $t = c\,s$ gives

$$\delta_{cs} \leq c\,\delta_{2s}.$$

Proof. Given a t-sparse vector $\mathbf{u} \in \mathbb{C}^N$ and an r-sparse vector $\mathbf{v} \in \mathbb{C}^N$ that are disjointly supported, we need to show that

$$|\langle \mathbf{Au}, \mathbf{Av} \rangle| \leq \sqrt{\frac{t}{s}}\,\theta_{s,r}\,\|\mathbf{u}\|_2\,\|\mathbf{v}\|_2, \tag{6.7}$$

$$\big|\|\mathbf{Au}\|_2^2 - \|\mathbf{u}\|_2^2\big| \leq \Big(\frac{t-d}{s}\,\delta_{2s} + \frac{d}{s}\,\delta_s\Big)\,\|\mathbf{u}\|_2^2. \tag{6.8}$$

Let d be a common divisor of s and t. We introduce integers k, n such that

$$s = kd, \qquad t = nd.$$

Let $T = \{j_1, j_2, \ldots, j_t\}$ denote the support of $\mathbf{u}$. We consider the n subsets $S_1, S_2, \ldots, S_n \subseteq T$ of size s defined by

$$S_i = \{j_{(i-1)d+1}, j_{(i-1)d+2}, \ldots, j_{(i-1)d+s}\},$$

where indices are meant modulo t. In this way, each $j \in T$ belongs to exactly $s/d = k$ sets S_i, so that

$$\mathbf{u} = \frac{1}{k}\sum_{i=1}^{n} \mathbf{u}_{S_i}, \qquad \|\mathbf{u}\|_2^2 = \frac{1}{k}\sum_{i=1}^{n} \|\mathbf{u}_{S_i}\|_2^2.$$

We now derive (6.7) from

$$|\langle \mathbf{Au}, \mathbf{Av} \rangle| \leq \frac{1}{k}\sum_{i=1}^{n} |\langle \mathbf{Au}_{S_i}, \mathbf{Av} \rangle| \leq \frac{1}{k}\sum_{i=1}^{n} \theta_{s,r}\|\mathbf{u}_{S_i}\|_2\|\mathbf{v}\|_2$$
$$\leq \theta_{s,r}\frac{\sqrt{n}}{k}\Big(\sum_{i=1}^{n} \|\mathbf{u}_{S_i}\|_2^2\Big)^{1/2}\|\mathbf{v}\|_2 = \theta_{s,r}\Big(\frac{n}{k}\Big)^{1/2}\|\mathbf{u}\|_2\|\mathbf{v}\|_2.$$

Inequality (6.8) follows from

$$\begin{aligned}
\big|\|\mathbf{A}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\big| &= \big|\langle(\mathbf{A}^*\mathbf{A} - \mathbf{Id})\mathbf{u}, \mathbf{u}\rangle\big| \le \frac{1}{k^2} \sum_{1\le i\le n} \sum_{1\le j\le n} \big|\langle(\mathbf{A}^*\mathbf{A} - \mathbf{Id})\mathbf{u}_{S_i}, \mathbf{u}_{S_j}\rangle\big| \\
&= \frac{1}{k^2}\Bigg(\sum_{1\le i\ne j\le n} \big|\langle(\mathbf{A}_{S_i\cup S_j}^*\mathbf{A}_{S_i\cup S_j} - \mathbf{Id})\mathbf{u}_{S_i}, \mathbf{u}_{S_j}\rangle\big| \\
&\qquad + \sum_{1\le i\le n} \big|\langle(\mathbf{A}_{S_i}^*\mathbf{A}_{S_i} - \mathbf{Id})\mathbf{u}_{S_i}, \mathbf{u}_{S_i}\rangle\big| \qquad \Bigg) \\
&\le \frac{1}{k^2}\Bigg(\sum_{1\le i\ne j\le n} \delta_{2s}\|\mathbf{u}_{S_i}\|_2\|\mathbf{u}_{S_j}\|_2 + \sum_{1\le i\le n} \delta_s\|\mathbf{u}_{S_i}\|_2^2\Bigg) \\
&= \frac{\delta_{2s}}{k^2}\Big(\sum_{1\le i\le n} \|\mathbf{u}_{S_i}\|_2\Big)^2 - \frac{\delta_{2s}-\delta_s}{k^2} \sum_{1\le i\le n} \|\mathbf{u}_{S_i}\|_2^2 \\
&\le \Big(\frac{\delta_{2s}\,n}{k^2} - \frac{\delta_{2s}-\delta_s}{k^2}\Big) \sum_{1\le i\le n} \|\mathbf{u}_{S_i}\|_2^2 = \Big(\frac{n}{k}\delta_{2s} - \frac{1}{k}(\delta_{2s}-\delta_s)\Big)\|\mathbf{u}\|_2^2 \\
&= \Big(\frac{t}{s}\delta_{2s} - \frac{1}{k}(\delta_{2s}-\delta_s)\Big)\|\mathbf{u}\|_2^2.
\end{aligned}$$

To make the latter as small as possible, we take k as small as possible, i.e., we choose d as the greatest common divisor of s and t. This completes the proof. □

Just like for the coherence, it is important to know how small the sth restricted isometry constant of a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ can be. In the case $N \ge C\,m$ of relevance in compressive sensing, Theorem 6.7 below states that the restricted isometry constant must satisfy $\delta_s \ge c\,\sqrt{s/m}$. For $s = 2$, this reads $\mu \ge c'/\sqrt{m}$, which is reminiscent of the Welch bound of Theorem 5.7. In fact, the proof below is an adaptation of the proof of this theorem.

Theorem 6.7. *For $\mathbf{A} \in \mathbb{C}^{m\times N}$ and $2 \le s \le N$, one has*

$$m \ge c\,\frac{s}{\delta_s^2}, \tag{6.9}$$

provided $N \ge C\,m$ and $\delta_s \le \delta_*$, where the constants c, C, and δ_* depend only on each other. For instance, the choices $c = 1/162$, $C = 30$, and $\delta_* = 2/3$ are valid.

Proof. We first notice that the statement cannot hold for $s = 1$, as $\delta_1 = 0$ if all the columns of $\mathbf{A}$ have ℓ_2-norm equal to 1. Let us set $t := \lfloor s/2 \rfloor \ge 1$, and let us decompose the matrix $\mathbf{A}$ in blocks of size $m \times t$ — except that possibly the last one may have less columns — as

$$\mathbf{A} = \big[\ \mathbf{A}_1 \mid \mathbf{A}_2 \mid \cdots \mid \mathbf{A}_n\ \big], \qquad N \le nt.$$

From (6.2) and (6.5), we recall that, for all $i, j \in [n], i \ne j$,

$$\|\mathbf{A}_i^*\mathbf{A}_i - \mathbf{Id}\|_{2\to 2} \le \delta_t \le \delta_s, \qquad \|\mathbf{A}_i^*\mathbf{A}_j\|_{2\to 2} \le \theta_{t,t} \le \delta_{2t} \le \delta_s,$$

so that the eigenvalues of $\mathbf{A}_i^*\mathbf{A}_i$ and the singular values of $\mathbf{A}_i^*\mathbf{A}_j$ satisfy

$$1 - \delta_s \le \lambda_k(\mathbf{A}_i^*\mathbf{A}_i) \le 1 + \delta_s, \qquad \sigma_k(\mathbf{A}_i^*\mathbf{A}_j) \le \delta_s.$$

Let us introduce the matrices

$$\mathbf{H} := \mathbf{A}\,\mathbf{A}^* \in \mathbb{C}^{m\times m}, \qquad \mathbf{G} := \mathbf{A}^*\mathbf{A} = \big[\mathbf{A}_i^*\mathbf{A}_j\big]_{1\le i,j\le n} \in \mathbb{C}^{N\times N}.$$

On the one hand, we have the lower bound

$$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathbf{G}) = \sum_{i=1}^{n} \mathrm{tr}(\mathbf{A}_i^*\mathbf{A}_i) = \sum_{i=1}^{n}\sum_{k=1}^{t} \lambda_k(\mathbf{A}_i^*\mathbf{A}_i) \ge n\,t\,(1-\delta_s). \tag{6.10}$$

On the other hand, writing $\langle \mathbf{M}_1, \mathbf{M}_2\rangle_F = \mathrm{tr}(\mathbf{M}_2^*\mathbf{M}_1)$ for the Frobenius inner product of two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, see Appendix A, we have

$$\mathrm{tr}(\mathbf{H})^2 = \langle \mathbf{Id}_m, \mathbf{H}\rangle_F^2 \le \|\mathbf{Id}_m\|_F^2 \|\mathbf{H}\|_F^2 = m\,\mathrm{tr}(\mathbf{H}^*\mathbf{H}).$$

Then, by cyclicity of the trace,

$$\begin{aligned}
\mathrm{tr}(\mathbf{H}^*\mathbf{H}) &= \mathrm{tr}(\mathbf{A}\,\mathbf{A}^*\mathbf{A}\,\mathbf{A}^*) = \mathrm{tr}(\mathbf{A}^*\mathbf{A}\,\mathbf{A}^*\mathbf{A}) = \mathrm{tr}(\mathbf{G}\,\mathbf{G}^*) \\
&= \sum_{i=1}^{n} \mathrm{tr}\Big(\sum_{j=1}^{m} \mathbf{A}_i^*\mathbf{A}_j\mathbf{A}_j^*\mathbf{A}_i\Big) \\
&= \sum_{1\le i\ne j\le n}\ \sum_{k=1}^{t} \sigma_k(\mathbf{A}_i^*\mathbf{A}_j)^2 + \sum_{i=1}^{n}\ \sum_{k=1}^{t} \lambda_k(\mathbf{A}_i^*\mathbf{A}_i)^2 \\
&\le n\,(n-1)\,t\,\delta_s^2 + n\,t\,(1+\delta_s)^2,
\end{aligned}$$

we derive the upper bound

$$\mathrm{tr}(\mathbf{H})^2 \le m\,n\,t\,\big((n-1)\,\delta_s^2 + (1+\delta_s)^2\big). \tag{6.11}$$

Combining the bounds (6.10) and (6.11) yields

$$m \ge \frac{n\,t\,(1-\delta_s)^2}{(n-1)\,\delta_s^2 + (1+\delta_s)^2}.$$

If $(n-1)\,\delta_s^2 < (1+\delta_s)^2/5$, we would obtain, using $\delta_s \le 2/3$,

$$m > \frac{n\,t\,(1-\delta_s)^2}{6(1+\delta_s)^2/5} \ge \frac{5(1-\delta_s)^2}{6(1+\delta_s)^2}\,N \ge \frac{1}{30}\,N,$$

which contradicts our assumption. We therefore have $(n-1)\,\delta_s^2 \ge (1+\delta_s)^2/5$, which yields, using $\delta_s \le 2/3$ again and $s \le 3t$,

$$m \ge \frac{n\,t\,(1-\delta_s)^2}{6(n-1)\,\delta_s^2} \ge \frac{1}{54}\,\frac{t}{\delta_s^2} \ge \frac{1}{162}\,\frac{s}{\delta_s^2}.$$

This is the desired result. □

Let us compare the lower bound

$$\delta_s \geq \sqrt{cs/m} \tag{6.12}$$

of the previous Theorem on the restricted isometry constant with upper bounds that are available at this point. Choose a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with coherence of optimal order $\mu \leq c/\sqrt{m}$, see Chapter 5. Then Proposition 6.2 implies that

$$\delta_s \leq (s-1)\mu \leq cs/\sqrt{m} \, . \tag{6.13}$$

Clearly, there is a large gap between (6.12) and (6.13). In particular, (6.13) requires the quadratic scaling

$$m \geq c' s^2 \tag{6.14}$$

in order to have small δ_s, while (6.12) indicates that the $m \geq cs$ is a necessary condition for the RIP. At this point it is not yet clear, which of the two conditions is closer to the optimal scaling of the restricted isometry constants, but we will see later in Chapter 9 that certain random matrices $\mathbf{A} \in \mathbb{R}^{m \times N}$ satisfy $\delta_s \leq \delta$ with high probability for some $\delta > 0$ provided

$$m \geq C\delta^{-2} s \ln(eN/s) \, . \tag{6.15}$$

We will see in Corollary 10.8 that $\delta_s \leq \delta$ indeed requires $m \geq C_\delta s \ln(eN/s)$, Therefore, the lower bound (6.9) is optimal up to the log-factor. In particular, Theorem 6.7 shows the optimality of the scaling $C_\delta = C\delta^{-2}$.

Difficulty of deterministic constructions of matrices satisfying the RIP of optimal order. As just mentioned, random matrices will be used in order to obtain the restricted isometry property in the optimal scaling (6.15) of the number m of measurements in terms of the sparsity s and the vector length N. It is open to date to provide deterministic, that is, explicit or at least polynomial time, constructions of matrices whose restricted isometry constants provably satisfy $\delta_s \leq \delta$ in the parameter regime (6.15). In fact, basically all available estimates of δ_s for deterministic matrices use the coherence μ combined with Proposition 6.2 at some point (with one notable exception on which we comment in the Notes). This leads then to bounds of the type (6.13) (or even worse, depending on the value of the coherence). In particular, we fall into the quadratic bottleneck. In fact, due to the lower bound on the coherence of Theorem 5.7 this proof technique is not able in principle to arrive at improved bounds.

The intrinsic difficulty of estimating the restricted isometry constants for explicit matrices lies in the fact that the basic tool for estimating the eigenvalues $\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}$ for deterministic $\mathbf{A}$ is Gershgorin's disc Theorem A.12. Assuming ℓ_2-normalization of the columns of $\mathbf{A}$ and taking the supremum over all $S \subset [N]$ with $\mathrm{card}(S) = s$ leads to the ℓ_1-coherence function $\mu_1(s-1)$ in this way. In fact, this is how we showed the bound $\delta_s \leq \mu_1(s-1)$ of Proposition 6.2, see also Theorem 5.3. Therefore, the lower bound for the ℓ_1-coherence

function in Theorem 5.8 tells us that we cannot avoid the quadratic bottleneck (6.14) when using Gershgorin's disc theorem in order to estimate restricted isometry constants. It seems that one should also take into account the signs and not only the absolute values of the entries of the Gramian $\mathbf{A}^*\mathbf{A}$ (as in Gershgorin's theorem) in order to improve estimates for deterministic matrices, but it is to date not very clear which tools can be used to achieve this goal (although we will discuss a slight improvement over the quadratic bottleneck in the Notes). In any case, one may conjecture that some of the matrices with coherence of optimal order, for instance the one of Theorem 5.13, also satisfy the restricted isometry property when m scales linear in s up to log-factors, but this remains an open and probably very difficult problem.

When passing to random matrices, however, a powerful set of tools becomes available that allow to estimate the restricted isometry constants in the optimal regime (6.15).

6.2 Analysis of Basis Pursuit

In this section, we establish the success of sparse recovery via basis pursuit for measurement matrices with small restricted isometry constants. We give two proofs of this fact.. The first proof is simple and quite natural. It shows that the condition $\delta_{2s} < 1/3$ is sufficient to guarantee exact recovery of all s-sparse vectors via ℓ_1-minimization. The second proof is more involved. It shows that the weaker condition $\delta_{2s} < 0.4931$ is actually sufficient to guarantee stable and robust recovery of all s-sparse vectors via ℓ_1-minimization. We start by presenting the simple argument which ignores stability and robustness issues (although such issues can be treated with only a slight extension of the argument).

Theorem 6.8. *Suppose that the $2s$th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies*

$$\delta_{2s} < \frac{1}{3}. \tag{6.16}$$

Then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique solution of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \qquad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}.$$

The following observation is recurring in our argument, so we isolate it from the proof.

Lemma 6.9. *Given $q > p > 0$, if $\mathbf{u} \in \mathbb{C}^s$ and $\mathbf{v} \in \mathbb{C}^t$ satisfy*

$$\max_{i \in [s]} |u_i| \leq \min_{j \in [t]} |v_j|, \tag{6.17}$$

then

$$\|\mathbf{u}\|_q \le \frac{s^{1/q}}{t^{1/p}} \|\mathbf{v}\|_p.$$

The special case $p = 1$, $q = 2$, and $t = s$ gives

$$\|\mathbf{u}\|_2 \le \frac{1}{\sqrt{s}} \|\mathbf{v}\|_1.$$

Proof. We only need to notice that

$$\frac{\|\mathbf{u}\|_q}{s^{1/q}} = \left[\frac{1}{s} \sum_{i=1}^{s} |u_i|^q \right]^{1/q} \le \max_{1 \le i \le s} |u_i|,$$

$$\frac{\|\mathbf{v}\|_p}{t^{1/p}} = \left[\frac{1}{t} \sum_{j=1}^{t} |v_j|^p \right]^{1/p} \ge \min_{1 \le j \le t} |v_j|,$$

to derive (6.17). The second statement is an immediate consequence. □

Proof (of Theorem 6.8). According to Corollary 4.5, it is enough to establish the null space property of order s in the form

$$\|\mathbf{v}_S\|_1 < \frac{1}{2} \|\mathbf{v}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{\mathbf{0}\} \text{ and all } S \subseteq [N] \text{ with } \operatorname{card}(S) = s.$$

This will follow from the stronger statement

$$\|\mathbf{v}_S\|_2 \le \frac{\rho}{2\sqrt{s}} \|\mathbf{v}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \text{ and all } S \subseteq [N] \text{ with } \operatorname{card}(S) = s,$$

where

$$\rho := \frac{2\delta_{2s}}{1 - \delta_{2s}}$$

satisfies $\rho < 1$ whenever $\delta_{2s} < 1/3$. Given $\mathbf{v} \in \ker \mathbf{A}$, we notice that it is enough to consider an index set $S =: S_0$ of s largest entries of the vector $\mathbf{v}$ in modulus. We partition the complement $\overline{S_0}$ of S_0 in $[N]$ as $\overline{S_0} = S_1 \cup S_2 \cup \dots$, where

$$S_1 : \text{index set of } s \text{ largest absolute entries of } \mathbf{v} \text{ in } \overline{S_0},$$
$$S_2 : \text{index set of } s \text{ largest absolute entries of } \mathbf{v} \text{ in } \overline{S_0 \cup S_1},$$

etc. In view of $\mathbf{v} \in \ker \mathbf{A}$, we have $\mathbf{A}(\mathbf{v}_{S_0}) = \mathbf{A}(-\mathbf{v}_{S_1} - \mathbf{v}_{S_2} - \cdots)$, so that

$$\begin{aligned} \|\mathbf{v}_{S_0}\|_2^2 &\le \frac{1}{1 - \delta_{2s}} \|\mathbf{A}(\mathbf{v}_{S_0})\|_2^2 = \frac{1}{1 - \delta_{2s}} \langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_1}) + \mathbf{A}(-\mathbf{v}_{S_2}) + \cdots \rangle \\ &= \frac{1}{1 - \delta_{2s}} \sum_{k \ge 1} \langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_k}) \rangle. \end{aligned} \tag{6.18}$$

According to Proposition 6.3, we also have

$$\langle \mathbf{A}(\mathbf{v}_{S_0}), \mathbf{A}(-\mathbf{v}_{S_k}) \rangle \le \delta_{2s} \|\mathbf{v}_{S_0}\|_2 \|\mathbf{v}_{S_k}\|_2. \tag{6.19}$$

Substituting (6.19) into (6.18) and dividing by $\|\mathbf{v}_{S_0}\|_2 > 0$, we obtain

$$\|\mathbf{v}_{S_0}\|_2 \le \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{k \ge 1} \|\mathbf{v}_{S_k}\|_2 = \frac{\rho}{2} \sum_{k \ge 1} \|\mathbf{v}_{S_k}\|_2.$$

For $k \ge 1$, the s entries of $\mathbf{v}_{S_k}$ do not exceed the s entries of $\mathbf{v}_{S_{k-1}}$, so that Lemma 6.9 yields

$$\|\mathbf{v}_{S_k}\|_2 \le \frac{1}{\sqrt{s}} \|\mathbf{v}_{S_{k-1}}\|_1.$$

We then derive

$$\|\mathbf{v}_{S_0}\|_2 \le \frac{\rho}{2\sqrt{s}} \sum_{k \ge 1} \|\mathbf{v}_{S_{k-1}}\|_1 \le \frac{\rho}{2\sqrt{s}} \|\mathbf{v}\|_1.$$

This is the desired inequality. □

Remark 6.10. In (6.18), the vector $\mathbf{v}_{S_0}$ was interpreted as being $2s$-sparse, although it is in fact s-sparse. The better bound $\|\mathbf{v}_{S_0}\|_2^2 \le \|\mathbf{A}(\mathbf{v}_{S_0})\|_2^2/(1-\delta_s)$ could therefore be invoked. In (6.19), the restricted orthogonality constant $\theta_{s,s}$ could also have been used instead of δ_{2s}. This would yield the sufficient condition $\delta_s + 2\theta_{s,s} < 1$ instead of (6.16).

It is instructive to refine the above proof by establishing stability and robustness. The reader is invited to do so in Exercise 6.11. Here, stability and robustness are incorporated in Theorem 6.11 below, which also improves on Theorem 6.8 by relaxing the sufficient condition (6.16).

Theorem 6.11. *Suppose that the $2s$th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies*

$$\delta_{2s} < \frac{77 - \sqrt{1337}}{82} \approx 0.4931. \tag{6.20}$$

Then, for any $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{y} \in \mathbb{C}^m$ with $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \le \eta$, a solution $\mathbf{x}^\sharp$ of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \qquad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta$$

approximates the vector $\mathbf{x}$ with errors

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \le C\, \sigma_s(\mathbf{x})_1 + D\, \sqrt{s}\, \eta,$$
$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \le \frac{C}{\sqrt{s}}\, \sigma_s(\mathbf{x})_1 + D\, \eta,$$

where the constants $C, D > 0$ depend only on δ_{2s}.

These error estimates — in fact, ℓ_p-error estimates for any $1 \le p \le 2$ — are immediately deduced from Theorem 4.21 and the following result.

Theorem 6.12. *If the $2s$th restricted isometry constant of $\mathbf{A} \in \mathbb{C}^{m \times N}$ obeys* (6.20)*, then the matrix $\mathbf{A}$ satisfies the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ depending only on δ_{2s}.*

The argument makes use of the following lemma, called *square root lifting* inequality. It can be viewed as a counterpart of the inequality $\|\mathbf{a}\|_1 \le \sqrt{s}\|\mathbf{a}\|_2$ for $\mathbf{a} \in \mathbb{C}^s$.

Lemma 6.13. *For $a_1 \ge a_2 \ge \cdots \ge a_s \ge 0$,*

$$\sqrt{a_1^2 + \cdots + a_s^2} \le \frac{a_1 + \cdots + a_s}{\sqrt{s}} + \frac{\sqrt{s}}{4}(a_1 - a_s).$$

Proof. We prove the equivalent statement

$$\left.\begin{array}{r} a_1 \ge a_2 \ge \cdots \ge a_s \ge 0 \\ \dfrac{a_1 + a_2 + \cdots + a_s}{\sqrt{s}} + \dfrac{\sqrt{s}}{4} a_1 \le 1 \end{array}\right\} \Longrightarrow \sqrt{a_1^2 + \cdots + a_s^2} + \frac{\sqrt{s}}{4} a_s \le 1.$$

Thus, we aim at maximizing the convex function

$$f(a_1, a_2, \ldots, a_s) := \sqrt{a_1^2 + \cdots + a_s^2} + \frac{\sqrt{s}}{4} a_s$$

over the convex polygon

$$\mathcal{C} := \{(a_1, \ldots, a_s) \in \mathbb{R}^s : a_1 \ge \cdots \ge a_s \ge 0 \text{ and } \frac{a_1 + \cdots + a_s}{\sqrt{s}} + \frac{\sqrt{s}}{4} a_1 \le 1\}.$$

Because any point in $\mathcal{C}$ is a convex combination of its vertices and because the function f is convex, the maximum is attained at a vertex of $\mathcal{C}$, see Theorem B.16. The vertices of $\mathcal{C}$ are obtained as intersections of s hyperplanes arising by turning s of the $(s+1)$ inequality constraints into equalities. We have the following possibilities.

- If $\alpha_1 = \cdots = \alpha_s = 0$, then $f(\alpha_1, \alpha_2, \ldots, \alpha_s) = 0$.
- If $(a_1 + \cdots + a_s)/\sqrt{s} + \sqrt{s}\, a_1/4 = 1$ and $a_1 = \cdots = a_k > a_{k+1} = \cdots = a_s = 0$ for some $1 \le k \le s-1$, then one has $a_1 = \cdots = a_k = 4\sqrt{s}/(4k+s)$, and consequently $f(a_1, \ldots, a_s) = 4\sqrt{ks}/(4k+s) \le 1$,
- If $(a_1 + \cdots + a_s)/\sqrt{s} + \sqrt{s}\, a_1/4 = 1$ and $a_1 = \cdots = a_s > 0$, then one has $a_1 = \cdots = a_s = 4/(5\sqrt{s})$, and consequently $f(a_1, \ldots, a_s) = 4/5 + 1/5 = 1$.

We have obtained

$$\max_{(a_1, \ldots, a_s) \in \mathcal{C}} f(a_1, a_2, \ldots, a_s) = 1,$$

which is the desired result. □

We are now ready to establish the robust null space property stated in Theorem 6.12. To simplify the initial reading of the proof, the reader may consider only the stable null space property by specifying $\mathbf{v} \in \ker \mathbf{A}$ in the following argument.

Proof (of Theorem 6.12). We need to find constants $0 < \rho < 1$ and $\tau > 0$ such that, for any $\mathbf{v} \in \mathbb{C}^N$ and any $S \subseteq [N]$ with $\mathrm{card}(S) = s$,

$$\|\mathbf{v}_S\|_2 \le \frac{\rho}{\sqrt{s}} \|\mathbf{v}_{\overline{S}}\|_1 + \tau \|\mathbf{A}\mathbf{v}\|_2. \tag{6.21}$$

Given $\mathbf{v} \in \mathbb{C}^N$, it is enough to consider an index set $S =: S_0$ of s largest entries of $\mathbf{v}$ in modulus. As before, we partition the complement of $S =: S_0$ as $\overline{S_0} = S_1 \cup S_2 \cup \ldots$, where

$$\begin{aligned} S_1 &: \text{index set of } s \text{ largest absolute entries of } \mathbf{v} \text{ in } \overline{S_0}, \\ S_2 &: \text{index set of } s \text{ largest absolute entries of } \mathbf{v} \text{ in } \overline{S_0 \cup S_1}, \end{aligned}$$

etc. We start by writing

$$\begin{aligned} (1-\delta_{2s})(\|\mathbf{v}_{S_0} + \mathbf{v}_{S_1}\|_2^2) \le \|\mathbf{A}(\mathbf{v}_{S_0} + \mathbf{v}_{S_1})\|_2^2 &= \|\mathbf{A}\mathbf{v} - \sum_{k \ge 2} \mathbf{A}\mathbf{v}_{S_k}\|_2^2 \\ &= \|\sum_{k \ge 2} \mathbf{A}\mathbf{v}_{S_k}\|_2^2 - 2\,\mathrm{Re}\langle \mathbf{A}\mathbf{v}, \sum_{k \ge 2} \mathbf{A}\mathbf{v}_{S_k} \rangle + \|\mathbf{A}\mathbf{v}\|_2^2. \end{aligned} \tag{6.22}$$

Let us postpone dealing with the term not appearing when $\mathbf{v} \in \ker \mathbf{A}$, namely

$$\lambda := -2\,\mathrm{Re}\langle \mathbf{A}\mathbf{v}, \sum_{k \ge 2} \mathbf{A}\mathbf{v}_{S_k} \rangle + \|\mathbf{A}\mathbf{v}\|_2^2.$$

As for the remaining term, we use (6.1) and (6.3) to derive

$$\begin{aligned} \|\sum_{k \ge 2} \mathbf{A}\mathbf{v}_{S_k}\|_2^2 = \langle \sum_{k \ge 2} \mathbf{A}\mathbf{v}_{S_k}, \sum_{\ell \ge 2} \mathbf{A}\mathbf{v}_{S_\ell} \rangle &= \sum_{k \ge 2} \|\mathbf{A}\mathbf{v}_{S_k}\|_2^2 + \sum_{k,\ell \ge 2, k \ne \ell} \langle \mathbf{A}\mathbf{v}_{S_k}, \mathbf{A}\mathbf{v}_{S_\ell} \rangle \\ &\le (1+\delta_{2s}) \sum_{k \ge 2} \|\mathbf{v}_{S_k}\|_2^2 + \sum_{k,\ell \ge 2, k \ne \ell} \delta_{2s} \|\mathbf{v}_{S_k}\|_2 \|\mathbf{v}_{S_\ell}\|_2 \\ &= \sum_{k \ge 2} \|\mathbf{v}_{S_k}\|_2^2 + \delta_{2s} \Big(\sum_{k \ge 2} \|\mathbf{v}_{S_k}\|_2 \Big)^2. \end{aligned} \tag{6.23}$$

For each $k \ge 0$, let υ_k^- and υ_k^+ denote the smallest and largest entries in modulus of $\mathbf{v}$ on S_k. Let us also set $\Sigma := \sum_{k \ge 2} \|\mathbf{v}_{S_k}\|_1 = \|\mathbf{v}_{\overline{S_0}}\|_1 - \|\mathbf{v}_{S_1}\|_1$. We observe that

$$\sum_{k \ge 2} \|\mathbf{v}_{S_k}\|_2^2 = \sum_{k \ge 2} \sum_{j \in S_k} |v_j|^2 \le \sum_{k \ge 2} \upsilon_2^+ \sum_{j \in S_k} |v_j| = \upsilon_2^+ \, \Sigma. \tag{6.24}$$

Moreover, Lemma 6.13 and $\upsilon_k^- \ge \upsilon_{k+1}^+$ imply

$$\sum_{k \ge 2} \|\mathbf{v}_{S_k}\|_2 \le \sum_{k \ge 2} \Big(\frac{\|\mathbf{v}_{S_k}\|_1}{\sqrt{s}} + \frac{\sqrt{s}}{4} (\upsilon_k^+ - \upsilon_k^-) \Big) \le \frac{\Sigma}{\sqrt{s}} + \frac{\sqrt{s}\,\upsilon_2^+}{4}. \tag{6.25}$$

Substituting (6.24) and (6.25) into (6.23) yields

$$\|\sum_{k\geq 2}\mathbf{A}\mathbf{v}_{S_k}\|_2^2 \leq \upsilon_2^+\,\Sigma + \delta_{2s}\Big(\frac{\Sigma}{\sqrt{s}} + \frac{\sqrt{s}\,\upsilon_2^+}{4}\Big)^2. \tag{6.26}$$

In turn, substituting (6.26) into (6.22) while taking into account the inequality $\|\mathbf{v}_{S_0}+\mathbf{v}_{S_1}\|_2^2 = \|\mathbf{v}_{S_0}\|_2^2 + \|\mathbf{v}_{S_1}\|_2^2 \geq \|\mathbf{v}_{S_0}\|_2^2 + s{\upsilon_2^+}^2$ gives

$$(1-\delta_{2s})\|\mathbf{v}_{S_0}\|_2^2 + (1-\delta_{2s})s{\upsilon_2^+}^2 \leq \upsilon_2^+\,\Sigma + \delta_{2s}\Big(\frac{\Sigma}{\sqrt{s}} + \frac{\sqrt{s}\,\upsilon_2^+}{4}\Big)^2 + \lambda. \tag{6.27}$$

Furthermore, since

$$\Sigma = \|\mathbf{v}_{\overline{S_0}}\|_1 - \|\mathbf{v}_{S_1}\|_1 \leq \|\mathbf{v}_{\overline{S_0}}\|_1 - s\,\upsilon_2^+ = (1-x)\|\mathbf{v}_{\overline{S_0}}\|_1,$$

where we have set $x := s\,\upsilon_2^+/\|\mathbf{v}_{\overline{S_0}}\|_1$, the inequality (6.27) reads

$$\begin{aligned}
&(1-\delta_{2s})\|\mathbf{v}_{S_0}\|_2^2 \\
&\quad\leq -\frac{(1-\delta_{2s})x^2}{s}\|\mathbf{v}_{\overline{S_0}}\|_1^2 + \frac{x(1-x)}{s}\|\mathbf{v}_{\overline{S_0}}\|_1^2 + \delta_{2s}\Big(\frac{1-x}{\sqrt{s}} + \frac{x}{4\sqrt{s}}\Big)^2\|\mathbf{v}_{\overline{S_0}}\|_1^2 + \lambda \\
&\quad= \Big[-(1-\delta_{2s})x^2 + x(1-x) + \delta_{2s}\Big(1-\frac{3x}{4}\Big)^2\Big]\frac{\|\mathbf{v}_{\overline{S_0}}\|_1^2}{s} + \lambda.
\end{aligned} \tag{6.28}$$

We need to prove that the quantity in square brackets divided by $(1-\delta_{2s})$ is smaller than one, or equivalently that this quantity minus $(1-\delta_{2s})$ is negative. This difference is the quadratic expression in x given by

$$q(x) := -\Big(2 - \frac{25}{16}\delta_{2s}\Big)x^2 + \Big(1 - \frac{3}{2}\delta_{2s}\Big)x - 1 + 2\delta_{2s} =: -ax^2 + bx - c.$$

This expression is maximized at $x^* := b/(2a)$, so that

$$q(x) \leq q(x^*) = \frac{b^2}{4a} - c = \frac{(1-3\delta_{2s}/2)^2}{4(2-25\delta_{2s}/16)} - 1 + 2\delta_{2s}.$$

It is easy to check that $q(x^*) < 0$ if and only if $41\delta_{2s}^2 - 77\delta_{2s} + 28 > 0$. Thus $q(x) < 0$ holds as soon as δ_{2s} is smaller than the smallest root of $41t^2 - 77t + 28$. This is exactly Condition (6.20). Under this condition, the quantity in square brackets is at most $\rho^2(1-\delta_{2s})$ for some constant $0 < \rho < 1$ depending only on δ_{2s}. We now derive from (6.28) that

$$\|\mathbf{v}_{S_0}\|_2^2 \leq \frac{\rho^2\|\mathbf{v}_{\overline{S_0}}\|_1^2}{s} + \frac{\lambda}{1-\delta_{2s}}.$$

Turning to the estimation of λ, we write

$$\lambda \leq 2\|\mathbf{A}\mathbf{v}\|_2\sum_{k\geq 2}\|\mathbf{A}\mathbf{v}_{S_k}\|_2 + \|\mathbf{A}\mathbf{v}\|_2^2 \leq 2\sqrt{1+\delta_{2s}}\|\mathbf{A}\mathbf{v}\|_2\sum_{k\geq 2}\|\mathbf{v}_{S_k}\|_2 + \|\mathbf{A}\mathbf{v}\|_2^2.$$

We now call upon Lemma 6.9 to obtain $\|\mathbf{v}_{S_k}\|_2 \le \|\mathbf{v}_{S_{k-1}}\|_1/\sqrt{s}$, hence

$$\sum_{k\ge 2} \|\mathbf{v}_{S_k}\|_2 \le \frac{\|\mathbf{v}_{\overline{S_0}}\|_1}{\sqrt{s}},$$

rather than using the more involved inequality (6.25). Altogether, we deduce that

$$\begin{aligned}
\|\mathbf{v}_{S_0}\|_2^2 &\le \frac{\rho^2 \|\mathbf{v}_{\overline{S_0}}\|_1^2}{s} + 2\frac{\sqrt{1+\delta_{2s}}}{1-\delta_{2s}} \|\mathbf{A}\mathbf{v}\|_2 \frac{\|\mathbf{v}_{\overline{S_0}}\|_1}{\sqrt{s}} + \frac{\|\mathbf{A}\mathbf{v}\|_2^2}{1-\delta_{2s}} \\
&\le \frac{\rho^2 \|\mathbf{v}_{\overline{S_0}}\|_1^2}{s} + 2\frac{\sqrt{1+\delta_{2s}}}{1-\delta_{2s}} \|\mathbf{A}\mathbf{v}\|_2 \frac{\|\mathbf{v}_{\overline{S_0}}\|_1}{\sqrt{s}} + \frac{1+\delta_{2s}}{\rho^2(1-\delta_{2s})^2} \|\mathbf{A}\mathbf{v}\|_2^2 \\
&= \left(\frac{\rho \|\mathbf{v}_{\overline{S_0}}\|_1}{\sqrt{s}} + \frac{\sqrt{1+\delta_{2s}}}{\rho(1-\delta_{2s})} \|\mathbf{A}\mathbf{v}\|_2 \right)^2.
\end{aligned}$$

This is the desired inequality (6.21) with $\tau := \sqrt{1+\delta_{2s}}/(\rho(1-\delta_{2s}))$. □

We close this section by highlighting some limitations of the restricted isometry property in the context of basis pursuit. We recall from Remark 4.6 that s-sparse recovery via basis pursuit is preserved if some measurements are rescaled, reshuffled, or added. However, these operations may deteriorate the restricted isometry constants. Reshuffling measurements corresponds to replacing the measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ by $\mathbf{P}\mathbf{A}$, where $\mathbf{P} \in \mathbb{C}^{m\times m}$ is a permutation matrix. This operation leaves the restricted isometry constants unchanged, since in fact $\delta_s(\mathbf{U}\mathbf{A}) = \delta_s(\mathbf{A})$ for any unitary matrix $\mathbf{U} \in \mathbb{C}^{m\times m}$. Adding a measurement, however, which corresponds to appending a row to the measurement matrix, may increase the restricted isometry constant. Consider for instance a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ with sth order restricted isometry constant $\delta_s(\mathbf{A}) < 1$, and let $\delta > \delta_s(\mathbf{A})$. We construct a matrix $\widetilde{\mathbf{A}}$ by appending the row $\begin{bmatrix} 0 \dots 0 \sqrt{1+\delta} \end{bmatrix}$. With $\mathbf{x} := \begin{bmatrix} 0 \dots 0 \, 1 \end{bmatrix}^\top$, it is easy to see that $\|\mathbf{A}\mathbf{x}\|_2^2 \ge 1+\delta$. This implies that $\delta_1(\widetilde{\mathbf{A}}) \ge \delta$, and consequently that $\delta_s(\widetilde{\mathbf{A}}) > \delta_s(\mathbf{A})$. Likewise, rescaling the measurements, which corresponds to replacing the measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ by $\mathbf{D}\mathbf{A}$, where $\mathbf{D} \in \mathbb{C}^{m\times m}$ is a diagonal matrix, may also increase the restricted isometry constant. This is even the case for scalar rescaling, i.e., replacing $\mathbf{A}$ by $d\mathbf{A}$ for $d \in \mathbb{C}$. For instance, if $\mathbf{A} \in \mathbb{C}^{m\times N}$ has an sth order restricted isometry constant $\delta_s(\mathbf{A}) < 3/5$, then the sth order restricted isometry constant of $2\mathbf{A}$ satisfies $\delta_s(2\mathbf{A}) \ge 3 - 4\delta_s(\mathbf{A}) > \delta_s(\mathbf{A})$. In order to circumvent the issue of scalar rescaling, one can work instead with the sth *restricted isometry ratio* $\gamma_s = \gamma_s(\mathbf{A})$, defined as the ratio

$$\gamma_s := \frac{\beta_s}{\alpha_s} \ge 1,$$

where α_s and β_s are the largest and smallest constants $\alpha, \beta \ge 0$ such that

$$\alpha \|\mathbf{x}\|_2^2 \le \|\mathbf{A}\mathbf{x}\|_2^2 \le \beta \|\mathbf{x}\|_2^2$$

for all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$. Note that this does not settle the issue of general rescaling. Consider indeed the $(2s) \times (2s+1)$ matrix $\mathbf{A}$ and the $(2s) \times (2s)$ diagonal matrix $\mathbf{D}_\varepsilon$ defined by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \ddots & 0 & -1 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{D}_\varepsilon = \operatorname{diag}(\varepsilon, 1/\varepsilon, 1, \ldots, 1).$$

Since $\ker \mathbf{D}_\varepsilon \mathbf{A} = \ker \mathbf{A}$ is spanned by $\big[1, 1, \ldots, 1\big]^\top$, the matrices $\mathbf{D}_\varepsilon \mathbf{A}$ and $\mathbf{A}$ both satisfy the sth order null space property, hence allow s-sparse recovery via basis pursuit. However, the sth order restricted isometry ratio of $\mathbf{D}_\varepsilon \mathbf{A}$ can be made arbitrarily large, since $\gamma_s(\mathbf{D}_\varepsilon \mathbf{A}) \geq 1/\varepsilon^4$. Incidentally, this shows that there are matrices allowing s-sparse recovery via basis pursuit but whose sth order restricted isometry constant are arbitrarily close to 1 — even after scalar renormalization, see Exercise 6.2.

6.3 Analysis of Thresholding Algorithms

In this section, we establish the success of sparse recovery via iterative hard thresholding and via hard thresholding pursuit for measurement matrices with small restricted isometry constants. Again, we start with a simple and quite natural proof of the success of s-sparse recovery via iterative hard thresholding under the condition $\delta_{3s} < 0.5$. This is done in the ideal situation of exactly sparse vectors acquired with perfect accuracy. We then cover the more realistic situation of approximately sparse vectors measured with some errors. The improved result only requires the weaker condition $\delta_{3s} < 0.5773$. It applies to both iterative hard thresholding and hard thresholding pursuit, but its proof is more involved. Before all this, we recall from Section 3 that the iterative hard thresholding algorithm starts with an initial s-sparse vector $\mathbf{x}^0 \in \mathbb{C}^N$, typically $\mathbf{x}^0 = 0$, and produces a sequence $(\mathbf{x}^n)$ defined inductively by

$$\mathbf{x}^{n+1} = H_s(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)). \tag{IHT}$$

The hard thresholding operator H_s keeps the s largest modulus components of a vector, so that $H_s(\mathbf{z})$ is a (not necessarily unique) best s-term approximation to $\mathbf{z} \in \mathbb{C}^N$. For small restricted isometry constants, the success of iterative hard thresholding is intuitively justified by the fact that $\mathbf{A}^*\mathbf{A}$ behaves like the identity when its domain and range are restricted to small support sets. Thus, if $\mathbf{y} = \mathbf{A}\mathbf{x}$ for some sparse $\mathbf{x} \in \mathbb{C}^N$, the contribution to $\mathbf{x}^{n+1}$ of $\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n) = \mathbf{A}^*\mathbf{A}(\mathbf{x} - \mathbf{x}^n)$ is roughly $\mathbf{x} - \mathbf{x}^n$, which sums with $\mathbf{x}^n$ to the desired $\mathbf{x}$. Here is a formal statement of the success of iterative hard thresholding.

Theorem 6.14. *Suppose that the $3s$th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies*

$$\delta_{3s} < \frac{1}{2}. \tag{6.29}$$

Then, for every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$, the sequence $(\mathbf{x}^n)$ defined by (IHT) *with $\mathbf{y} = \mathbf{A}\mathbf{x}$ converges to $\mathbf{x}$.*

The following observation is recurring in our arguments, so we isolate it from the proof.

Lemma 6.15. *Given vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ and an index set $S \subseteq [N]$,*

$$\begin{aligned} |\langle \mathbf{u}, (\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v} \rangle| &\le \delta_t \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \qquad \text{if } \operatorname{card}(\operatorname{supp}(\mathbf{u}) \cup \operatorname{supp}(\mathbf{v})) \le t, \\ \|((\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v})_S\|_2 &\le \delta_t \|\mathbf{v}\|_2 \qquad \text{if } \operatorname{card}(S \cup \operatorname{supp}(\mathbf{v})) \le t. \end{aligned}$$

Proof. For the first inequality, let $T := \operatorname{supp}(\mathbf{u}) \cup \operatorname{supp}(\mathbf{v})$, and let $\mathbf{u}_T$ and $\mathbf{v}_T$ denote the subvectors of $\mathbf{u}$ and $\mathbf{v}$ obtained by only keeping the entries indexed by T as usual. We write

$$\begin{aligned} |\langle \mathbf{u}, (\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v} \rangle| &= |\langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v} \rangle| = |\langle \mathbf{u}_T, \mathbf{v}_T \rangle - \langle \mathbf{A}_T\mathbf{u}_T, \mathbf{A}_T\mathbf{v}_T \rangle| \\ &= |\langle \mathbf{u}_T, (\mathbf{Id} - \mathbf{A}_T^*\mathbf{A}_T)\mathbf{v}_T \rangle| \le \|\mathbf{u}_T\|_2 \, \|(\mathbf{Id} - \mathbf{A}_T^*\mathbf{A}_T)\mathbf{v}_T\|_2 \\ &\le \|\mathbf{u}_T\|_2 \, \|\mathbf{Id} - \mathbf{A}_T^*\mathbf{A}_T\|_{2 \to 2} \|\mathbf{v}_T\|_2 \le \delta_t \|\mathbf{u}\|_2 \|\mathbf{v}\|_2. \end{aligned}$$

The second inequality follows from the first one by observing that

$$\|((\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v})_S\|_2^2 = \langle ((\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v})_S, (\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v} \rangle \le \delta_t \|((\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v})_S\|_2 \|\mathbf{v}\|_2.$$

We divide by $\|((\mathbf{Id} - \mathbf{A}^*\mathbf{A})\mathbf{v})_S\|_2$ to complete the proof. □

Proof (of Theorem 6.14). It is enough to find a constant $0 \le \rho < 1$ such that

$$\|\mathbf{x}^{n+1} - \mathbf{x}\|_2 \le \rho \, \|\mathbf{x}^n - \mathbf{x}\|_2, \qquad n \ge 0, \tag{6.30}$$

since this implies by induction that

$$\|\mathbf{x}^n - \mathbf{x}\|_2 \le \rho^n \|\mathbf{x}^0 - \mathbf{x}\|_2 \underset{n \to \infty}{\longrightarrow} 0.$$

By definition, the s-sparse vector $\mathbf{x}^{n+1}$ is a better (or at least equally good) approximation to

$$\mathbf{u}^n := \mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n) = \mathbf{x}^n + \mathbf{A}^*\mathbf{A}(\mathbf{x} - \mathbf{x}^n)$$

than the s-sparse vector $\mathbf{x}$. This implies

$$\|\mathbf{u}^n - \mathbf{x}^{n+1}\|_2^2 \le \|\mathbf{u}^n - \mathbf{x}\|_2^2.$$

Expanding $\|\mathbf{u}^n - \mathbf{x}^{n+1}\|_2^2 = \|(\mathbf{u}^n - \mathbf{x}) - (\mathbf{x}^{n+1} - \mathbf{x})\|_2^2$ and rearranging yields

$$\|\mathbf{x}^{n+1} - \mathbf{x}\|_2^2 \le 2\,\mathrm{Re}\langle \mathbf{u}^n - \mathbf{x}, \mathbf{x}^{n+1} - \mathbf{x}\rangle. \tag{6.31}$$

We now use Lemma 6.15 to obtain

$$\begin{aligned}\mathrm{Re}\langle \mathbf{u}^n - \mathbf{x}, \mathbf{x}^{n+1} - \mathbf{x}\rangle &= \mathrm{Re}\langle (\mathbf{Id} - \mathbf{A}^*\mathbf{A})(\mathbf{x}^n - \mathbf{x}), \mathbf{x}^{n+1} - \mathbf{x}\rangle \\ &\le \delta_{3s}\|\mathbf{x}^n - \mathbf{x}\|_2 \|\mathbf{x}^{n+1} - \mathbf{x}\|_2. \end{aligned}\tag{6.32}$$

If $\|\mathbf{x}^{n+1} - \mathbf{x}\|_2 > 0$, we derive from (6.31) and (6.32) that

$$\|\mathbf{x}^{n+1} - \mathbf{x}\|_2 \le 2\delta_{3s}\|\mathbf{x}^n - \mathbf{x}\|_2,$$

which is obviously true if $\|\mathbf{x}^{n+1} - \mathbf{x}\|_2 = 0$. Thus, the desired inequality (6.30) holds with $\rho = 2\delta_{3s} < 1$. □

Remark 6.16. Sufficient conditions for the success of s-sparse recovery via basis pursuit were previously given in terms of δ_{2s}. Such sufficient conditions can also be given for iterative hard thresholding. For instance, since $\delta_{3s} \le 2\delta_{2s} + \delta_s \le 3\delta_{2s}$ by Proposition 6.6, it is enough to assume $\delta_{2s} < 1/6$ to guarantee $\delta_{3s} < 1/2$, hence the success of s-sparse recovery via iterative hard thresholding. This condition may be weakened to $\delta_{2s} < 1/4$ by refining the previous argument — see Exercise 6.18. It can be further weakened to $\delta_{2s} < 1/3$ with a slight modification of the algorithm — see Exercise 6.19.

It is again instructive to refine the proof above for approximately sparse vectors measured with some errors, and the reader is invited to do so in Exercise 6.17. Theorem 6.17 below covers this case, while improving on Theorem 6.14 by relaxing the sufficient condition (6.29). As a consequence, we will obtain in Theorem 6.20 error estimates similar to the ones for basis pursuit. We underline that the arguments are valid for both iterative hard thresholding and hard thresholding pursuit. As a reminder, this latter algorithm starts with an initial s-sparse vector $\mathbf{x}^0 \in \mathbb{C}^N$, typically $\mathbf{x}^0 = 0$, and produces a sequence $(\mathbf{x}^n)$ defined inductively by

$$S^{n+1} = L_s(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)), \tag{HTP$_1$}$$

$$\mathbf{x}^{n+1} = \mathrm{argmin}\left\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \mathrm{supp}(\mathbf{z}) \subseteq S^{n+1}\right\}. \tag{HTP$_2$}$$

We recall that $L_s(\mathbf{z})$ denotes an index set of s largest absolute entries of a vector $\mathbf{z} \in \mathbb{C}^N$.

Theorem 6.17. *Suppose that the* $3s$*th restricted isometry constant of the matrix* $\mathbf{A} \in \mathbb{C}^{m\times N}$ *satisfies*

$$\delta_{3s} < \frac{1}{\sqrt{3}} \approx 0.5773. \tag{6.33}$$

Then, for $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{e} \in \mathbb{C}^m$, *and* $S \subseteq [N]$ *with* $\mathrm{card}(S) = s$, *the sequence* $(\mathbf{x}^n)$ *defined by* (IHT) *or by* (HTP) *with* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ *satisfies, for any* $n \ge 0$,

$$\|\mathbf{x}^n - \mathbf{x}_S\|_2 \le \rho^n \|\mathbf{x}^0 - \mathbf{x}_S\|_2 + \tau \|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2, \tag{6.34}$$

where $\rho = \sqrt{3}\,\delta_{3s} < 1$, $\tau \le 2.18/(1-\rho)$ *for* (IHT), *and* $\rho = \sqrt{2\delta_{3s}^2/(1-\delta_{2s}^2)} < 1$, $\tau \le 5.15/(1-\rho)$ *for* (HTP)

Remark 6.18. The intuitive superiority of the hard thresholding pursuit algorithm over the iterative hard thresholding algorithm is not reflected in a weaker sufficient condition in terms of restricted isometry constants, but rather in a faster rate of convergence justified by $\rho_{HTP} < \rho_{IHT}$ when $\delta_{3s} < 1/\sqrt{3}$.

We isolate the following observation from the proof of the theorem.

Lemma 6.19. *Given* $\mathbf{e} \in \mathbb{C}^m$ *and* $S \in [N]$ *with* $\mathrm{card}(S) \le s$,

$$\|(\mathbf{A}^*\mathbf{e})_S\|_2 \le \sqrt{1+\delta_s}\,\|\mathbf{e}\|_2.$$

Proof. We only need to write

$$\begin{aligned}\|(\mathbf{A}^*\mathbf{e})_S\|_2^2 &= \langle \mathbf{A}^*\mathbf{e}, (\mathbf{A}^*\mathbf{e})_S\rangle = \langle \mathbf{e}, \mathbf{A}\big((\mathbf{A}^*\mathbf{e})_S\big)\rangle \le \|\mathbf{e}\|_2\,\|\mathbf{A}\big((\mathbf{A}^*\mathbf{e})_S\big)\|_2 \\ &\le \|\mathbf{e}\|_2\,\sqrt{1+\delta_s}\,\|(\mathbf{A}^*\mathbf{e})_S\|_2,\end{aligned}$$

and to divide by $\|(\mathbf{A}^*\mathbf{e})_S\|_2$. □

Proof (of Theorem 6.17). Given $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{e} \in \mathbb{C}^m$, $S \subseteq [N]$ with $\mathrm{card}(S) = s$, our aim is to prove that, for any $n \ge 0$,

$$\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2 \le \rho \|\mathbf{x}^n - \mathbf{x}_S\|_2 + (1-\rho)\tau \|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2. \tag{6.35}$$

The estimate (6.34) then follows by induction. For both iterative hard thresholding and hard thresholding pursuit, the index set $S^{n+1} := \mathrm{supp}(\mathbf{x}^{n+1})$ consists of s largest absolute entries of $\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)$, so we have

$$\|(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_S\|_2^2 \le \|(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1}}\|_2^2.$$

Eliminating the contribution on $S \cap S^{n+1}$, we derive

$$\|(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S\setminus S^{n+1}}\|_2 \le \|(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1}\setminus S}\|_2.$$

The right-hand side may be written as

$$\|(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1}\setminus S}\|_2 = \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1}\setminus S}\|_2.$$

The left-hand side satisfies

$$\begin{aligned}&\|(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S\setminus S^{n+1}}\|_2 \\ &\quad = \|(\mathbf{x}_S - \mathbf{x}^{n+1} + \mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S\setminus S^{n+1}}\|_2 \\ &\quad \ge \|(\mathbf{x}_S - \mathbf{x}^{n+1})_{S\setminus S^{n+1}}\|_2 - \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S\setminus S^{n+1}}\|_2.\end{aligned}$$

With $S \Delta S^{n+1} = (S \setminus S^{n+1}) \cup (S^{n+1} \setminus S)$ denoting the symmetric difference of the sets S and S^{n+1}, we conclude that

$$\begin{aligned}\|(\mathbf{x}_S - \mathbf{x}^{n+1})_{S \setminus S^{n+1}}\|_2 &\le \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S \setminus S^{n+1}}\|_2 \\ &\quad + \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1} \setminus S}\|_2 \\ &\le \sqrt{2}\, \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S \Delta S^{n+1}}\|_2. \quad (6.36)\end{aligned}$$

Let us first concentrate on iterative hard thresholding. In this case,

$$\mathbf{x}^{n+1} = \big(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)\big)_{S^{n+1}}.$$

It then follows that

$$\begin{aligned}\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2^2 &= \|(\mathbf{x}^{n+1} - \mathbf{x}_S)_{S^{n+1}}\|_2^2 + \|(\mathbf{x}^{n+1} - \mathbf{x}_S)_{\overline{S^{n+1}}}\|_2^2 \\ &= \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1}}\|_2^2 + \|(\mathbf{x}^{n+1} - \mathbf{x}_S)_{S \setminus S^{n+1}}\|_2^2.\end{aligned}$$

Together with (6.36), we obtain

$$\begin{aligned}\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2^2 &\le \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S^{n+1}}\|_2^2 \\ &\quad + 2\, \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S \Delta S^{n+1}}\|_2^2 \\ &\le 3\, \|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{S \cup S^{n+1}}\|_2^2.\end{aligned}$$

We now write $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{A}\mathbf{x}_S + \mathbf{e}'$ with $\mathbf{e}' := \mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}$, and we call upon Lemma 6.15 (noticing that $\mathrm{card}(S \cup S^{n+1} \cup \mathrm{supp}(\mathbf{x}^n - \mathbf{x}_S)) \le 3s$) and Lemma 6.19 to deduce

$$\begin{aligned}\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2 &\le \sqrt{3}\, [\|(\mathbf{x}^n - \mathbf{x}_S + \mathbf{A}^*\mathbf{A}(\mathbf{x}_S - \mathbf{x}^n) + \mathbf{A}^*\mathbf{e}')_{S \cup S^{n+1}}\|_2] \\ &\le \sqrt{3}\, [\|\big((\mathbf{Id} - \mathbf{A}^*\mathbf{A})(\mathbf{x}^n - \mathbf{x}_S)\big)_{S \cup S^{n+1}}\|_2 + \|(\mathbf{A}^*\mathbf{e}')_{S \cup S^{n+1}}\|_2] \\ &\le \sqrt{3}\, [\delta_{3s}\|\mathbf{x}^n - \mathbf{x}_S\|_2 + \sqrt{1 + \delta_{2s}}\, \|\mathbf{e}'\|_2].\end{aligned}$$

This is the desired inequality (6.35) for iterative hard thresholding. We notice that $\rho = \sqrt{3}\, \delta_{3s}$ is indeed smaller than one as soon as $\delta_{3s} < 1/\sqrt{3}$, and that $(1 - \rho)\tau = \sqrt{3}\sqrt{1 + \delta_{2s}} \le \sqrt{3 + \sqrt{3}} \le 2.18$.
Let us now concentrate on hard thresholding pursuit. In this case,

$$\mathbf{x}^{n+1} = \mathrm{argmin}\, \big\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \mathrm{supp}(\mathbf{z}) \subseteq S^{n+1}\big\}.$$

As the best ℓ_2-approximation to $\mathbf{y}$ from the space $\{\mathbf{A}\mathbf{z}, \mathrm{supp}(\mathbf{z}) \subseteq S^{n+1}\}$, the vector $\mathbf{A}\mathbf{x}^{n+1}$ is characterized by

$$\langle \mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}, \mathbf{A}\mathbf{z} \rangle = 0 \qquad \text{whenever } \mathrm{supp}(\mathbf{z}) \subseteq S^{n+1},$$

that is to say, by $\langle \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}), \mathbf{z} \rangle = 0$ whenever $\mathrm{supp}(\mathbf{z}) \subseteq S^{n+1}$, or by

$$(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}))_{S^{n+1}} = \mathbf{0}.$$

Taking this and (6.36) into consideration, we write

$$\begin{aligned}\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2^2 &= \|(\mathbf{x}^{n+1}-\mathbf{x}_S)_{S^{n+1}}\|_2^2 + \|(\mathbf{x}^{n+1}-\mathbf{x}_S)_{S\setminus S^{n+1}}\|_2^2\\ &\le \|(\mathbf{x}^{n+1}-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^{n+1}))_{S^{n+1}}\|_2^2\\ &\quad + 2\,\|(\mathbf{x}^{n}-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^{n}))_{S\Delta S^{n+1}}\|_2^2\\ &\le \big[\|\big((\mathbf{Id}-\mathbf{A}^*\mathbf{A})(\mathbf{x}^{n+1}-\mathbf{x}_S)\big)_{S^{n+1}}\|_2 + \|(\mathbf{A}^*\mathbf{e}')_{S^{n+1}}\|_2\big]^2\\ &\quad + 2\,\big[\|\big((\mathbf{Id}-\mathbf{A}^*\mathbf{A})(\mathbf{x}^{n}-\mathbf{x}_S)\big)_{S\Delta S^{n+1}}\|_2 + \|(\mathbf{A}^*\mathbf{e}')_{S\Delta S^{n+1}}\|_2\big]^2.\end{aligned}$$

Applying Lemma 6.15 and Lemma 6.19 yields

$$\begin{aligned}\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2^2 \le\ & \big[\delta_{2s}\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2 + \sqrt{1+\delta_s}\,\|\mathbf{e}'\|_2\big]^2\\ &+ 2\,\big[\delta_{3s}\|\mathbf{x}^{n}-\mathbf{x}_S\|_2 + \sqrt{1+\delta_{2s}}\,\|\mathbf{e}'\|_2\big]^2.\end{aligned}$$

After rearrangement, this reads

$$\begin{aligned}&2\,\big[\delta_{3s}\|\mathbf{x}^{n}-\mathbf{x}_S\|_2 + \sqrt{1+\delta_{2s}}\,\|\mathbf{e}'\|_2\big]^2\\ &\ge (1-\delta_{2s}^2)\Big(\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2 + \frac{\sqrt{1+\delta_s}}{1+\delta_{2s}}\|\mathbf{e}'\|_2\Big)\Big(\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2 - \frac{\sqrt{1+\delta_s}}{1-\delta_{2s}}\|\mathbf{e}'\|_2\Big).\end{aligned}$$

Since we may assume $\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2 \ge \sqrt{1+\delta_s}\,\|\mathbf{e}'\|_2/(1-\delta_{2s})$ to make the latter expression in parentheses positive — otherwise (6.35) is clear from the value of $(1-\rho)\tau$ given below — we obtain

$$2\,\big[\delta_{3s}\|\mathbf{x}^{n}-\mathbf{x}_S\|_2+\sqrt{1+\delta_{2s}}\,\|\mathbf{e}'\|_2\big]^2 \ge (1-\delta_{2s}^2)\Big(\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2-\frac{\sqrt{1+\delta_s}}{1-\delta_{2s}}\|\mathbf{e}'\|_2\Big)^2.$$

From here, taking the square root and rearranging gives

$$\|\mathbf{x}^{n+1}-\mathbf{x}_S\|_2 \le \frac{\sqrt{2}\,\delta_{3s}}{\sqrt{1-\delta_{2s}^2}}\|\mathbf{x}^{n}-\mathbf{x}_S\|_2 + \Big(\frac{\sqrt{2}}{\sqrt{1-\delta_{2s}}} + \frac{\sqrt{1+\delta_s}}{1-\delta_{2s}}\Big)\|\mathbf{e}'\|_2.$$

This is the desired inequality (6.35) for hard thresholding pursuit. We notice that $\rho := \sqrt{2}\,\delta_{3s}/\sqrt{1-\delta_{2s}^2} \le \sqrt{2}\,\delta_{3s}/\sqrt{1-\delta_{3s}^2}$ is indeed smaller than one as soon as $\delta_{3s} < 1/\sqrt{3}$, and that $(1-\rho)\tau = \sqrt{2}/\sqrt{1-\delta_{2s}} + \sqrt{1+\delta_s}/(1-\delta_{2s}) \le 5.15$. □

Taking the limit as $n \to \infty$ in (6.34) yields $\|\mathbf{x}^\sharp - \mathbf{x}_S\|_2 \le \tau\|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2$ if $\mathbf{x}^\sharp \in \mathbb{C}^N$ is the limit of the sequence $(\mathbf{x}^n)$ or at least one of its cluster points. Note that the existence of this limit is not at all guaranteed by our argument, but at least the existence of cluster points is guaranteed by the boundedness of $\|\mathbf{x}^n\|$ which follows from (6.34). In any case, we have $\|\mathbf{x}-\mathbf{x}^\sharp\|_2 \le \|\mathbf{x}_{\overline{S}}\|_2 + \|\mathbf{x}_S - \mathbf{x}^\sharp\|_2$ by the triangle inequality, so choosing S as an index set of s largest entries of $\mathbf{x}$ in modulus gives

$$\|\mathbf{x}-\mathbf{x}^\sharp\|_2 \le \sigma_s(\mathbf{x})_2 + \tau\|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2. \tag{6.37}$$

This estimate does not resemble the basis pursuit estimates of Theorem 6.11. However, such estimates are available for thresholding algorithms, too, provided we replace the parameter s in (IHT) and (HTP) by $2s$, say. The precise statement is as follows.

Theorem 6.20. *Suppose that the $6s$th order restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies $\delta_{6s} < 1/\sqrt{3}$. Then, for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, the sequence $(\mathbf{x}^n)$ defined by* (IHT) *or by* (HTP_1) *with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, $\mathbf{x}^0 = \mathbf{0}$, and s replaced by $2s$ satisfies, for any $n \geq 0$,*

$$\|\mathbf{x} - \mathbf{x}^n\|_1 \leq C\, \sigma_s(\mathbf{x})_1 + D\, \sqrt{s}\, \|\mathbf{e}\|_2 + 2\, \rho^n \sqrt{s}\, \|\mathbf{x}\|_2 \,,$$
$$\|\mathbf{x} - \mathbf{x}^n\|_2 \leq \frac{C}{\sqrt{s}}\, \sigma_s(\mathbf{x})_1 + D\, \|\mathbf{e}\|_2 + 2\, \rho^n\, \|\mathbf{x}\|_2 \,.$$

where the constants $C, D > 0$ and $0 < \rho < 1$ depend only on δ_{6s}. In particular, if the sequence $(\mathbf{x}^n)$ clusters around some $\mathbf{x}^\sharp \in \mathbb{C}^N$, then

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \leq C\, \sigma_s(\mathbf{x})_1 + D\, \sqrt{s}\, \|\mathbf{e}\|_2 \,, \tag{6.38}$$
$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \leq \frac{C}{\sqrt{s}}\, \sigma_s(\mathbf{x})_1 + D\, \|\mathbf{e}\|_2 \,. \tag{6.39}$$

Remark 6.21. (a) Error estimates of the type (6.38) and (6.39) are not only valid for cluster points $\mathbf{x}^\sharp$, but can be extended to all $\mathbf{x}^n$ for n large enough when $C\sigma_s(\mathbf{x})_1 + D\sqrt{s}\|\mathbf{e}\|_2 > 0$. Indeed, in this case we have, for all $n \geq n_0$ with large enough n_0,

$$2\rho^n \sqrt{s} \|\mathbf{x}\|_2 \leq C\sigma_s(\mathbf{x})_1 + D\sqrt{s}\|\mathbf{e}\|_2 \,.$$

Therefore, the general error estimates in the above theorem imply

$$\|\mathbf{x} - \mathbf{x}^n\|_1 \leq 2C\, \sigma_s(\mathbf{x})_1 + 2D\, \sqrt{s}\, \|\mathbf{e}\|_2 \,,$$
$$\|\mathbf{x} - \mathbf{x}^n\|_2 \leq \frac{2C}{\sqrt{s}}\, \sigma_s(\mathbf{x})_1 + 2D\, \|\mathbf{e}\|_2$$

for all $n \geq n_0$.

(b) A major drawback when running hard thresholding algorithms is that an estimation of the targeted sparsity s is needed. This estimation is not needed for the inequality-constrained ℓ_1-minimization, but an estimation of the measurement error η is (a priori) needed instead. In fact, we will see in Chapter 11 that running the equality-constrained ℓ_1-minimization (P_1) on corrupted measurements may in some cases still have the benefit of stable and robust estimates (6.38)-(6.39).

The auxiliary result below plays a central role when proving statements such as Theorem 6.20.

Lemma 6.22. *Suppose $\mathbf{A} \in \mathbb{C}^{m\times N}$ has restricted isometry constant $\delta_s < 1$. Given $\kappa, \tau > 0$, $\xi \in \mathbb{R}$, and $\mathbf{e} \in \mathbb{C}^m$, assume that two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{C}^N$ satisfy $\|\mathbf{x}'\|_0 \le \kappa s$ and*

$$\|\mathbf{x}_T - \mathbf{x}'\|_2 \le \tau \|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2 + \xi$$

where T denotes an index set of $2s$ largest absolute entries of $\mathbf{x}$. Then, for any $1 \le p \le 2$,

$$\|\mathbf{x} - \mathbf{x}'\|_p \le \frac{1 + c_\kappa\, \tau}{s^{1-1/p}}\, \sigma_s(\mathbf{x})_1 + d_\kappa\, \tau\, s^{1/p-1/2}\|\mathbf{e}\|_2 + d_\kappa\, s^{1/p-1/2}\xi, \tag{6.40}$$

where the constants $c_\kappa, d_\kappa > 0$ depends only on κ.

Proof. We first use the fact that the vector $\mathbf{x}_T - \mathbf{x}'$ is $(2+\kappa)s$-sparse to write

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}'\|_p &\le \|\mathbf{x}_{\overline{T}}\|_p + \|\mathbf{x}_T - \mathbf{x}'\|_p \le \|\mathbf{x}_{\overline{T}}\|_p + ((2+\kappa)s)^{1/p-1/2}\|\mathbf{x}_T - \mathbf{x}'\|_2 \\ &\le \|\mathbf{x}_{\overline{T}}\|_p + \sqrt{2+\kappa}\, s^{1/p-1/2}(\tau\|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2 + \xi). \end{aligned} \tag{6.41}$$

Let now $S \subseteq T$ denote index sets of s largest entries of $\mathbf{x}$ in modulus. We observe that, according to Proposition 2.3,

$$\|\mathbf{x}_{\overline{T}}\|_p = \sigma_s(\mathbf{x}_{\overline{S}})_p \le \frac{1}{s^{1-1/p}}\|\mathbf{x}_{\overline{S}}\|_1 = \frac{1}{s^{1-1/p}}\sigma_s(\mathbf{x})_1. \tag{6.42}$$

Let us partition the complement of T as $\overline{T} = S_2 \cup S_3 \cup \ldots$, where

$$\begin{aligned} S_2 &: \text{index set of } s \text{ largest modulus entries of } \mathbf{x} \text{ in } \overline{T}, \\ S_3 &: \text{index set of } s \text{ largest modulus entries of } \mathbf{x} \text{ in } \overline{T \cup S_2}, \end{aligned}$$

etc. In this way, we have

$$\begin{aligned}\|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2 &\le \sum_{k\ge 2} \|\mathbf{A}\mathbf{x}_{S_k}\|_2 + \|\mathbf{e}\|_2 \le \sum_{k\ge 2} \sqrt{1+\delta_s}\|\mathbf{x}_{S_k}\|_2 + \|\mathbf{e}\|_2 \\ &\le \sqrt{2}\sum_{k\ge 2} \|\mathbf{x}_{S_k}\|_2 + \|\mathbf{e}\|_2. \end{aligned}$$

Using Lemma 6.9, it has become usual to derive

$$\sum_{k\ge 2} \|\mathbf{x}_{S_k}\|_2 \le \frac{1}{s^{1/2}}\|\mathbf{x}_{\overline{S}}\|_1 = \frac{1}{s^{1/2}}\sigma_s(\mathbf{x})_1,$$

hence we obtain

$$\|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2 \le \frac{\sqrt{2}}{s^{1/2}}\sigma_s(\mathbf{x})_1 + \|\mathbf{e}\|_2. \tag{6.43}$$

Substituting (6.42) and (6.43) into (6.41), we obtain the estimate (6.40) with $c_\kappa = \sqrt{4+2\kappa}$ and $d_\kappa = \sqrt{2+\kappa}$. □

Proof (of Theorem 6.20). Given $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, under the present hypotheses, Theorem 6.17 implies that there exist $0 < \rho < 1$ and $\tau > 0$ depending only on δ_{6s} such that, for any $n \geq 0$,

$$\|\mathbf{x}_T - \mathbf{x}^n\|_2 \leq \tau \|\mathbf{A}\mathbf{x}_T + \mathbf{e}\|_2 + \rho^n \|\mathbf{x}_T\|_2,$$

where T denotes an index set of $2s$ largest entries of $\mathbf{x}$ in modulus. Then Lemma 6.22 with $\mathbf{x}' = \mathbf{x}^n$ and $\xi = \rho^n \|\mathbf{x}_T\|_2$ implies that, for any $1 \leq p \leq 2$,

$$\|\mathbf{x} - \mathbf{x}^n\|_p \leq \frac{C}{s^{1-1/p}} \, \sigma_s(\mathbf{x})_1 + D \, s^{1/p-1/2} \, \|\mathbf{e}\|_2 + \rho^n \, s^{1/p-1/2} \, \|\mathbf{x}_T\|_2,$$

where $C, D > 0$ depend only on τ, hence only on δ_{6s}. The desired estimates are the particular cases $p = 1$ and $p = 2$. □

6.4 Analysis of Greedy Algorithms

In this final section, we establish the success of sparse recovery via the greedy algorithms presented in Section 3.2, namely orthogonal matching pursuit and compressive sampling matching pursuit. For the orthogonal matching pursuit algorithm, we first remark that standard restricted isometry conditions are not enough to guarantee the recovery of all s-sparse vectors in at most s iterations. Indeed, for a fixed $1 < \eta < \sqrt{s}$, consider the $(s+1) \times (s+1)$ matrix with ℓ_2-normalized columns defined by

$$A := \left[\begin{array}{c|c} \mathbf{Id} & \begin{matrix} \frac{\eta}{s} \\ \vdots \\ \frac{\eta}{s} \end{matrix} \\ \hline 0 \cdots 0 & \sqrt{\frac{s-\eta^2}{s}} \end{array}\right]. \tag{6.44}$$

We calculate

$$\mathbf{A}^*\mathbf{A} - \mathbf{Id} = \left[\begin{array}{c|c} 0 & \begin{matrix} \frac{\eta}{s} \\ \vdots \\ \frac{\eta}{s} \end{matrix} \\ \hline \frac{\eta}{s} \cdots \frac{\eta}{s} & 0 \end{array}\right].$$

This matrix has eigenvalues $-\eta/\sqrt{s}$, $\eta/\sqrt{s}$, and 0 with multiplicity $s-1$. Thus,

$$\delta_{s+1} = \|\mathbf{A}^*\mathbf{A} - \mathbf{Id}\|_{2\to 2} = \frac{\eta}{\sqrt{s}.}$$

However, the s-sparse vector $\mathbf{x} = [1, \ldots, 1, 0]^\top$ is not recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ after s iterations, since the wrong index $s+1$ is picked at the first iteration. Indeed,

$$\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^0)=\mathbf{A}^*\mathbf{A}\mathbf{x}=\left[\begin{array}{c|c} \multicolumn{1}{c|}{\multirow{2}{*}{$\mathbf{Id}$}} & \frac{\eta}{s} \\ & \vdots \\ \hline \frac{\eta}{s} \ \dots \ \frac{\eta}{s} & 1 \end{array}\right]\begin{bmatrix}1\\ \vdots \\ 1 \\ \hline 0\end{bmatrix}=\begin{bmatrix}1\\ \vdots \\ 1 \\ \hline \eta\end{bmatrix}.$$

There are two possibilities to bypass this issue: either perform more than s iterations, or find a way to reject the wrong indices by modifying the orthogonal matching pursuit, which is the rationale behind compressive sampling matching pursuit. In both cases, sparse recovery will be established under restricted isometry conditions. In what follows, we do not separate the ideal situation of exactly sparse vectors measured with perfect accuracy, but we directly give the more cumbersome proofs for stable and robust s-sparse recovery under the condition $\delta_{10s} < 0.1666$ for $6s$ iterations of orthogonal matching pursuit and $\delta_{4s} < 0.4782$ for compressive sampling matching pursuit. Although the argument for compressive sampling matching pursuit are close to the argument used in the previous section, we start with the orthogonal matching pursuit algorithm.

Orthogonal Matching Pursuit

For the purpose of proving the main result, we consider the slightly more general algorithm starting with an index set S^0 and with

$$\mathbf{x}^0 := \operatorname{argmin}\{\|\mathbf{y}-\mathbf{A}\mathbf{z}\|, \operatorname{supp}(\mathbf{z}) \subseteq S^0\}, \tag{6.45}$$

and iterating the scheme

$$S^{n+1} = S^n \cup L_1(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)), \tag{OMP'$_1$}$$

$$\mathbf{x}^{n+1} = \operatorname{argmin}\left\{\|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2, \operatorname{supp}(\mathbf{z}) \subseteq S^{n+1}\right\}. \tag{OMP'$_2$}$$

The usual orthogonal matching pursuit algorithm corresponds to the default choice of $S^0 = \emptyset$ and $\mathbf{x}^0 = \mathbf{0}$. The following proposition is the key.

Proposition 6.23. *Suppose $\mathbf{A} \in \mathbb{C}^{m\times N}$ has restricted isometry constant $\delta_{10s} < 1/6$. Then there is a constant $C > 0$ depending only on δ_{10s} such that the sequence $(\mathbf{x}^n)$ defined by* (OMP') *with $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{e}$ for some s-sparse $\mathbf{x} \in \mathbb{C}^N$ and some $\mathbf{e} \in \mathbb{C}^m$ satisfies*

$$\|\mathbf{y}-\mathbf{A}\mathbf{x}^{\bar{n}}\|_2 \le C\|\mathbf{e}\|_2, \qquad \bar{n} := 6\ \operatorname{card}(\operatorname{supp}(\mathbf{x}) \setminus S^0).$$

Note that if $\mathbf{e} = 0$ and $S^0 = \emptyset$, this proposition implies exact s-sparse recovery via (OMP) in $6s$ iterations. Indeed, we have $\mathbf{A}(\mathbf{x}-\mathbf{x}^{6s}) = \mathbf{0}$, which implies $\mathbf{x}-\mathbf{x}^{6s} = \mathbf{0}$ since $\|\mathbf{x}-\mathbf{x}^{6s}\|_0 \le 7s$ and $\delta_{7s} \le \delta_{10s} < 1$. Proposition 6.23 also implies stability and robustness results stated in a familiar form.

Theorem 6.24. *Suppose $\mathbf{A} \in \mathbb{C}^{m\times N}$ has restricted isometry constant $\delta_{10s} < 1/6$. Then there is a constant $C > 0$ depending only on δ_{10s} such that, for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, the sequence $(\mathbf{x}^n)$ defined by* (OMP) *with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ satisfies*

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{6s}\|_2 \le C\|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2$$

for any $S \subseteq [N]$ with $\operatorname{card}(S) = s$. Furthermore, if $\delta_{20s} < 1/6$, then there are constants $C, D > 0$ depending only on δ_{20s} such that, for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, the sequence $(\mathbf{x}^n)$ defined by (OMP) *with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ satisfies, for any $1 \le p \le 2$,*

$$\|\mathbf{x} - \mathbf{x}^{12s}\|_p \le \frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + Ds^{1/p-1/2}\|\mathbf{e}\|_2.$$

Proof. Given $S \subseteq [N]$ with $\operatorname{card}(S) = s$, we can write $\mathbf{y} = \mathbf{A}\mathbf{x}_S + \mathbf{e}'$ where $\mathbf{e}' := \mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}$. Applying Proposition 6.23 with $S^0 = \emptyset$ then gives the desired inequality

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{6s}\|_2 \le C\|\mathbf{e}'\|_2 = C\|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2$$

for some constant $C > 0$ depending only on δ_{10s}. For the second inequality, we choose T to be an index set of $2s$ largest absolute entries of $\mathbf{x}$, so the previous argument yields

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{12s}\|_2 \le C'\|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2$$

for some constant $C' > 0$ depending only on δ_{20s}. Now, in view of

$$\begin{aligned}\|\mathbf{y} - \mathbf{A}\mathbf{x}^{12s}\|_2 &= \|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^{12s}) + \mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2 \ge \|\mathbf{A}(\mathbf{x}_T - \mathbf{x}^{12s})\|_2 - \|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2 \\ &\ge \sqrt{1-\delta_{20s}}\,\|\mathbf{x}^{12s} - \mathbf{x}_T\|_2 - \|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2,\end{aligned}$$

we derive

$$\|\mathbf{x}^{12s} - \mathbf{x}_T\|_2 \le \frac{C'+1}{\sqrt{1-\delta_{20s}}}\|\mathbf{A}\mathbf{x}_{\overline{T}} + \mathbf{e}\|_2.$$

An application of Lemma 6.22 with $\xi = 0$ gives the desired result. □

It remains to establish the crucial Proposition 6.23. As a matter of fact, this proposition is valid for any sequence satisfying the conclusion of the following lemma.

Lemma 6.25. *Let $(\mathbf{x}^n)$ be the sequence defined by* (OMP') *with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ for some s-sparse $\mathbf{x} \in \mathbb{C}^N$ and for some $\mathbf{e} \in \mathbb{C}^m$. Then, for $n \ge 0$, $T \subseteq [N]$ not included in S^n, and $\mathbf{z} \in \mathbb{C}^N$ supported on T,*

$$\begin{aligned}\|\mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}\|_2^2 &\le \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \frac{\|\mathbf{A}(\mathbf{z} - \mathbf{x}^n)\|_2^2}{\|\mathbf{z}_{T\setminus S^n}\|_1^2}\max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2\} \\ &\le \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \frac{1-\delta}{\operatorname{card}(T\setminus S^n)}\max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2\},\end{aligned}$$

where $\delta := \delta_{\operatorname{card}(T\cup S^n)}$.

Proof. The second inequality follows from the first one by noticing that

$$\|\mathbf{A}(\mathbf{x}^n - \mathbf{z})\|_2^2 \geq (1-\delta)\|\mathbf{x}^n - \mathbf{z}\|_2^2 \geq (1-\delta)\|(\mathbf{x}^n - \mathbf{z})_{T\setminus S^n}\|_2^2,$$
$$\|\mathbf{z}_{T\setminus S^n}\|_1^2 \leq \mathrm{card}(T\setminus S^n)\|\mathbf{z}_{T\setminus S^n}\|_2^2 = \mathrm{card}(T\setminus S^n)\|(\mathbf{x}^n - \mathbf{z})_{T\setminus S^n}\|_2^2.$$

We recall from Lemma 3.3 that the decrease in the squared ℓ_2-norm of the residual is at least $|(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{j^{n+1}}|^2$, where j^{n+1} denote an index of largest absolute entry of $\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)$. Thus, it is enough to prove that

$$|(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{j^{n+1}}|^2 \geq \frac{\|\mathbf{A}(\mathbf{z}-\mathbf{x}^n)\|_2^2}{\|\mathbf{z}_{T\setminus S^n}\|_1^2}\big(\|\mathbf{y}-\mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2^2\big) \quad (6.46)$$

when $\|\mathbf{y}-\mathbf{A}\mathbf{x}^n\|_2^2 \geq \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2^2$. Let us also recall from Lemma 3.4 that $(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{S^n} = \mathbf{0}$ to observe on the one hand that

$$\begin{aligned}
&\mathrm{Re}\langle \mathbf{A}(\mathbf{z}-\mathbf{x}^n), \mathbf{y}-\mathbf{A}\mathbf{x}^n\rangle \\
&\qquad = \mathrm{Re}\langle \mathbf{z}-\mathbf{x}^n, \mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\rangle = \mathrm{Re}\langle \mathbf{z}-\mathbf{x}^n, (\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{\overline{S^n}}\rangle \\
&\qquad = \mathrm{Re}\langle (\mathbf{z}-\mathbf{x}^n)_{T\setminus S^n}, (\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{T\setminus S^n}\rangle \\
&\qquad \leq \|(\mathbf{z}-\mathbf{x}^n)_{T\setminus S^n}\|_1 \|\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\|_\infty \\
&\qquad = \|\mathbf{z}_{T\setminus S^n}\|_1 |(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{j^{n+1}}|. \qquad (6.47)
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&2\,\mathrm{Re}\langle \mathbf{A}(\mathbf{z}-\mathbf{x}^n), \mathbf{y}-\mathbf{A}\mathbf{x}^n\rangle \\
&\qquad = \|\mathbf{A}(\mathbf{z}-\mathbf{x}^n)\|_2^2 + \|\mathbf{y}-\mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{A}(\mathbf{z}-\mathbf{x}^n) - (\mathbf{y}-\mathbf{A}\mathbf{x}^n)\|_2^2 \\
&\qquad = \|\mathbf{A}(\mathbf{z}-\mathbf{x}^n)\|_2^2 + \big(\|\mathbf{y}-\mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2^2\big) \\
&\qquad \geq 2\|\mathbf{A}(\mathbf{z}-\mathbf{x}^n)\|_2 \sqrt{\|\mathbf{y}-\mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2^2}, \qquad (6.48)
\end{aligned}$$

where we have used the inequality between arithmetic and geometric means in the last step. Combining the squared versions of (6.47) and (6.48), we arrive at

$$\|\mathbf{A}(\mathbf{z}-\mathbf{x}^n)\|_2^2 \big(\|\mathbf{y}-\mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{y}-\mathbf{A}\mathbf{z}\|_2^2\big) \leq \|\mathbf{z}_{T\setminus S^n}\|_1^2 \,|(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n))_{j^{n+1}}|^2.$$

The desired inequality (6.46) follows from here. □

We are now ready for the proof of the key proposition.

Proof (of Proposition 6.23). Let $S = \mathrm{supp}(\mathbf{x})$. The proof proceeds by induction on $\mathrm{card}(S\setminus S^0)$. If it is zero, i.e., if $S \subseteq S^0$, then the definition of $\mathbf{x}^0$ implies

$$\|\mathbf{y}-\mathbf{A}\mathbf{x}^0\|_2 \leq \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{e}\|_2,$$

and the result holds with $C = 1$. Let us now assume that the result holds for all S and S^0 such that $\mathrm{card}(S\setminus S^0) \leq s'-1$, $s' \geq 1$, and let us show that it

holds when $\mathrm{card}(S \setminus S^0) = s'$. We consider subsets of $S \setminus S^0$ defined by $T^0 = \emptyset$ and

$$T^\ell = \{\text{indices of } 2^{\ell-1} \text{ largest absolute entries of } \mathbf{x}_{\overline{S^0}} \} \text{ for } \ell \geq 1,$$

to which we associate the vectors

$$\tilde{\mathbf{x}}^\ell := \mathbf{x}_{\overline{S^0 \cup T^\ell}}, \quad \ell \geq 0.$$

Note that the last T^ℓ, namely $T^{\lceil \log_2(s') \rceil + 1}$, is taken to be the whole set $S \setminus S^0$ (and may have less than $2^{\ell-1}$ elements), so that $\tilde{\mathbf{x}}^\ell = \mathbf{0}$. For a constant $\mu > 0$ to be chosen later, since $\|\tilde{\mathbf{x}}^{\ell-1}\|_2^2 \geq \mu \|\tilde{\mathbf{x}}^\ell\|_2^2 = 0$ for this last index, we can consider the smallest integer $1 \leq L \leq \lceil \log_2(s') \rceil + 1$ such that

$$\|\tilde{\mathbf{x}}^{L-1}\|_2^2 \geq \mu \|\tilde{\mathbf{x}}^L\|_2^2.$$

This definition implies the (possibly empty) list of inequalities

$$\|\tilde{\mathbf{x}}^0\|_2^2 < \mu \|\tilde{\mathbf{x}}^1\|_2^2, \ldots, \|\tilde{\mathbf{x}}^{L-2}\|_2^2 < \mu \|\tilde{\mathbf{x}}^{L-1}\|_2^2.$$

For each $\ell \in [L]$, we apply Lemma 6.25 to the vector $\mathbf{z} = \mathbf{x} - \tilde{\mathbf{x}}^\ell$, which is supported on $S^0 \cup T^\ell$. Taking into account that $(S^0 \cup T^\ell) \cup S^n \subseteq S \cup S^n$ and that $(S^0 \cup T^\ell) \setminus S^n \subseteq (S^0 \cup T^\ell) \setminus S^0 = T^\ell$, we obtain, after subtracting $\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 = \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2$ from both sides,

$$\begin{aligned}
&\max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^{n+1}\|_2^2 - \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2\} \\
&\qquad \leq \left(1 - \frac{1 - \delta_{s+n}}{\mathrm{card}(T^\ell)}\right) \max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2\} \\
&\qquad \leq \exp\left(-\frac{1 - \delta_{s+n}}{\mathrm{card}(T^\ell)}\right) \max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2\}.
\end{aligned}$$

For any $K \geq 0$ and any $n, k \geq 0$ satisfying $n + k \leq K$, we derive by induction that

$$\begin{aligned}
&\max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^{n+k}\|_2^2 - \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2\} \\
&\qquad \leq \exp\left(-\frac{k(1 - \delta_{s+K})}{\mathrm{card}(T^\ell)}\right) \max\{0, \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 - \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2\}.
\end{aligned}$$

By separating cases in the rightmost maximum, we easily deduce

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{n+k}\|_2^2 \leq \exp\left(-\frac{k(1 - \delta_{s+K})}{\mathrm{card}(T^\ell)}\right) \|\mathbf{y} - \mathbf{A}\mathbf{x}^n\|_2^2 + \|\mathbf{A}\tilde{\mathbf{x}}^\ell + \mathbf{e}\|_2^2.$$

For some positive integer κ to be chosen later, applying this successively with

$$k_1 := \kappa\, \mathrm{card}(T^1), \ldots, \; k_L := \kappa\, \mathrm{card}(T^L), \quad \text{and} \quad K := k_1 + \cdots + k_L,$$

yields, with $\nu := \exp(\kappa(1 - \delta_{s+K}))$,

$$\begin{aligned}
\|\mathbf{y}-\mathbf{A}\mathbf{x}^{k_1}\|_2^2 &\le \frac{1}{\nu}\|\mathbf{y}-\mathbf{A}\mathbf{x}^0\|_2^2 + \|\mathbf{A}\tilde{\mathbf{x}}^1+\mathbf{e}\|_2^2\\
\|\mathbf{y}-\mathbf{A}\mathbf{x}^{k_1+k_2}\|_2^2 &\le \frac{1}{\nu}\|\mathbf{y}-\mathbf{A}\mathbf{x}^{k_1}\|_2^2 + \|\mathbf{A}\tilde{\mathbf{x}}^2+\mathbf{e}\|_2^2\\
&\vdots\\
\|\mathbf{y}-\mathbf{A}\mathbf{x}^{k_1+\cdots+k_{L-1}+k_L}\|_2^2 &\le \frac{1}{\nu}\|\mathbf{y}-\mathbf{A}\mathbf{x}^{k_1+\cdots+k_{L-1}}\|_2^2 + \|\mathbf{A}\tilde{\mathbf{x}}^L+\mathbf{e}\|_2^2.
\end{aligned}$$

By combining these inequalities, we obtain

$$\|\mathbf{y}-\mathbf{A}\mathbf{x}^K\|_2^2 \le \frac{\|\mathbf{y}-\mathbf{A}\mathbf{x}^0\|_2^2}{\nu^L} + \frac{\|\mathbf{A}\tilde{\mathbf{x}}^1+\mathbf{e}\|_2^2}{\nu^{L-1}} + \cdots + \frac{\|\mathbf{A}\tilde{\mathbf{x}}^{L-1}+\mathbf{e}\|_2^2}{\nu} + \|\mathbf{A}\tilde{\mathbf{x}}^L+\mathbf{e}\|_2^2.$$

Taking into account that $\mathbf{x}-\tilde{\mathbf{x}}^0$ is supported on $S^0\cup T^0 = S^0$, the definition (6.45) of $\mathbf{x}^0$ implies that $\|\mathbf{y}-\mathbf{A}\mathbf{x}^0\|_2^2 \le \|\mathbf{y}-\mathbf{A}(\mathbf{x}-\tilde{\mathbf{x}}^0)\|_2^2 = \|\mathbf{A}\tilde{\mathbf{x}}^0+\mathbf{e}\|_2^2$, hence

$$\|\mathbf{y}-\mathbf{A}\mathbf{x}^K\|_2^2 \le \sum_{\ell=0}^{L} \frac{\|\mathbf{A}\tilde{\mathbf{x}}^\ell+\mathbf{e}\|_2^2}{\nu^{L-\ell}} \le \sum_{\ell=0}^{L} \frac{2(\|\mathbf{A}\tilde{\mathbf{x}}^\ell\|_2^2+\|\mathbf{e}\|_2^2)}{\nu^{L-\ell}}.$$

Let us remark that, for $\ell \le L-1$ and also for $\ell = L$,

$$\|\mathbf{A}\tilde{\mathbf{x}}^\ell\|_2^2 \le (1+\delta_s)\|\tilde{\mathbf{x}}^\ell\|_2^2 \le (1+\delta_s)\mu^{L-1-\ell}\|\tilde{\mathbf{x}}^{L-1}\|_2^2.$$

As a result, we have

$$\begin{aligned}
\|\mathbf{y}-\mathbf{A}\mathbf{x}^K\|_2^2 &\le \frac{2(1+\delta_s)\|\tilde{\mathbf{x}}^{L-1}\|_2^2}{\mu}\sum_{\ell=0}^{L}\left(\frac{\mu}{\nu}\right)^{L-\ell} + 2\|\mathbf{e}\|_2^2\sum_{\ell=0}^{L}\frac{1}{\nu^{L-\ell}}\\
&\le \frac{2(1+\delta_s)\|\tilde{\mathbf{x}}^{L-1}\|_2^2}{\mu(1-\mu/\nu)} + \frac{2\|\mathbf{e}\|_2^2}{1-\nu}.
\end{aligned}$$

We choose $\mu = \nu/2$ so that $\mu(1-\mu/\nu)$ takes its maximal value $\nu/4$. It follows that, with $\alpha := \sqrt{8(1+\delta_s)/\nu}$ and $\beta := \sqrt{2/(1-\nu)}$,

$$\|\mathbf{y}-\mathbf{A}\mathbf{x}^K\|_2 \le \alpha\,\|\tilde{\mathbf{x}}^{L-1}\|_2 + \beta\,\|\mathbf{e}\|_2. \tag{6.49}$$

On the other hand, with $\gamma := \sqrt{1-\delta_{s+K}}$, we have

$$\begin{aligned}
\|\mathbf{y}-\mathbf{A}\mathbf{x}^K\|_2 &= \|\mathbf{A}(\mathbf{x}-\mathbf{x}^K)+\mathbf{e}\|_2 \ge \|\mathbf{A}(\mathbf{x}-\mathbf{x}^K)\|_2 - \|\mathbf{e}\|_2\\
&\ge \gamma\,\|\mathbf{x}-\mathbf{x}^K\|_2 - \|\mathbf{e}\|_2 \ge \gamma\,\|\mathbf{x}_{\overline{S^K}}\|_2 - \|\mathbf{e}\|_2.
\end{aligned}$$

We deduce that

$$\|\mathbf{x}_{\overline{S^K}}\|_2 \le \frac{\alpha}{\gamma}\|\tilde{\mathbf{x}}^{L-1}\|_2 + \frac{\beta+1}{\gamma}\|\mathbf{e}\|_2. \tag{6.50}$$

Let us now choose $\kappa = 3$, which guarantees that

$$\frac{\alpha}{\gamma} = \sqrt{\frac{8(1+\delta_s)}{(1-\delta_{s+K})\exp(\kappa(1-\delta_{s+K}))}} \le 0.92 < 1,$$

since $\delta_s \le \delta_{s+K} \le \delta_{10s} < 1/6$. Hereby, we have used the fact that $L \le \lceil \log_2(s') \rceil + 1$ implies

$$K = \kappa(1 + \cdots + 2^{L-2} + \mathrm{card}(T^L)) < \kappa(2^{L-1} + s') \le 3\kappa s' \le 9s.$$

Thus, in the case $((\beta+1)/\gamma)\|\mathbf{e}\|_2 < (1-\alpha/\gamma)\|\tilde{\mathbf{x}}^{L-1}\|_2$, we derive from (6.50) that

$$\|\mathbf{x}_{\overline{S^K}}\|_2 < \|\tilde{\mathbf{x}}^{L-1}\|_2, \quad \text{i.e.,} \quad \|(\mathbf{x}_{\overline{S^0}})_{S\setminus S^K}\|_2 < \|(\mathbf{x}_{\overline{S^0}})_{(S\setminus S^0)\setminus T^{L-1}}\|_2.$$

But since T^{L-1} lists the 2^{L-1} largest absolute entries of $\mathbf{x}_{\overline{S^0}}$, this yields

$$\mathrm{card}(S \setminus S^K) < \mathrm{card}((S \setminus S^0) \setminus T^{L-1}) = s' - 2^{L-1}.$$

Continuing the algorithm from iteration K amounts to starting it from iteration 0 with $\mathbf{x}^0$ replaced by $\mathbf{x}^K$, therefore the induction hypothesis implies that

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^{K+\bar{n}}\|_2 \le C\|\mathbf{e}\|_2, \qquad \bar{n} := 6(s' - 2^{L-1}).$$

Thus, since we also have the bound $K \le \kappa(1 + \cdots + 2^{L-2} + 2^{L-1}) < 3 \cdot 2^L$, the number of required iterations satisfies $K + \bar{n} \le 6s'$, as desired. In the alternative case where $((\beta+1)/\gamma)\|\mathbf{e}\|_2 \ge (1-\alpha/\gamma)\|\tilde{\mathbf{x}}^{L-1}\|_2$, the situation is easier, since (6.49) yields

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}^K\|_2 \le \frac{\alpha(\beta+1)}{\gamma-\alpha}\,\|\mathbf{e}\|_2 + \beta\,\|\mathbf{e}\|_2 =: C\|\mathbf{e}\|_2,$$

where the constant $C \ge 1$ depends only on δ_{10s}. This shows that the induction hypothesis holds when $\mathrm{card}(S \setminus S^0) = s'$. □

Compressive Sampling Matching Pursuit

As a reminder, we recall that the compressive sampling matching pursuit algorithm (CoSaMP) starts with an initial s-sparse vector $\mathbf{x}^0 \in \mathbb{C}^N$, typically $\mathbf{x}^0 = \mathbf{0}$, and produces a sequence $(\mathbf{x}^n)$ defined inductively by

$$U^{n+1} = \mathrm{supp}(\mathbf{x}^n) \cup L_{2s}(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)), \tag{CoSaMP$_1$}$$

$$\mathbf{u}^{n+1} = \mathrm{argmin}\left\{\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2, \mathrm{supp}(\mathbf{z}) \subseteq U^{n+1}\right\}, \tag{CoSaMP$_2$}$$

$$\mathbf{x}^{n+1} = H_s(\mathbf{u}^{n+1}). \tag{CoSaMP$_3$}$$

Here are the main results for this algorithm.

Theorem 6.26. *Suppose that the $4s$th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies*

$$\delta_{4s} < \frac{\sqrt{\sqrt{11/3}-1}}{2} \approx 0.4782. \tag{6.51}$$

Then, for $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{e} \in \mathbb{C}^m$, and $S \subseteq [N]$ with $\operatorname{card}(S) = s$, the sequence $(\mathbf{x}^n)$ defined by (CoSaMP) *with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ satisfies*

$$\|\mathbf{x}^n - \mathbf{x}_S\|_2 \le \rho^n \|\mathbf{x}^0 - \mathbf{x}_S\|_2 + \tau \|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2, \tag{6.52}$$

where the constant $0 < \rho < 1$ and $\tau > 0$ depend only on δ_{4s}.

Note that, if $\mathbf{x}$ is s-sparse and if $\mathbf{e} = \mathbf{0}$, then $\mathbf{x}$ is recovered as the limit of the sequence $(\mathbf{x}^n)$. In a more general situation, there is no guarantee that the sequence $(\mathbf{x}^n)$ converges. But (6.52) implies at least boundedness of the sequence $\|\mathbf{x}^n\|_2$ so that existence of cluster points is guaranteed. Stability and robustness results can then be stated as follows.

Theorem 6.27. *Suppose that the $8s$th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies $\delta_{8s} < 0.4782$. Then, for $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, the sequence $(\mathbf{x}^n)$ defined by* (CoSaMP) *with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, $\mathbf{x}^0 = \mathbf{0}$ and s replaced by $2s$ satisfies, for any $n \ge 0$,*

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}^n\|_1 &\le C\,\sigma_s(\mathbf{x})_1 + D\,\sqrt{s}\,\|\mathbf{e}\|_2 + 2\,\rho^n\sqrt{s}\,\|\mathbf{x}\|_2\,, \\
\|\mathbf{x} - \mathbf{x}^n\|_2 &\le \frac{C}{\sqrt{s}}\,\sigma_s(\mathbf{x})_1 + D\,\|\mathbf{e}\|_2 + 2\,\rho^n\,\|\mathbf{x}\|_2\,.
\end{aligned}$$

where the constants $C, D > 0$ and $0 < \rho < 1$ depend only on δ_{8s}. In particular, if $\mathbf{x}^\sharp \in \mathbb{C}^N$ denotes a cluster point of the sequence $(\mathbf{x}^n)$, then

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}^\sharp\|_1 &\le C\,\sigma_s(\mathbf{x})_1 + D\,\sqrt{s}\,\|\mathbf{e}\|_2\,, \\
\|\mathbf{x} - \mathbf{x}^\sharp\|_2 &\le \frac{C}{\sqrt{s}}\,\sigma_s(\mathbf{x})_1 + D\,\|\mathbf{e}\|_2\,.
\end{aligned}$$

Remark 6.28. Similarly as in Remark 6.21(a), error estimates as for cluster points $\mathbf{x}^\sharp$ apply actually to all iterates $\mathbf{x}^n$ for n large enough provided that the right hand side is non-trivial.

Theorem 6.27 follows from Theorem 6.26 via Lemma 6.22 in the same way as Theorem 6.20 follows from Theorem 6.17 for thresholding algorithms. We therefore concentrate on establishing Theorem 6.26.

Proof (of Theorem 6.26). As in the proof of Theorem 6.17, we establish that for any $n \ge 0$,

$$\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2 \le \rho\|\mathbf{x}^n - \mathbf{x}_S\|_2 + (1-\rho)\tau\|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_2 \tag{6.53}$$

with $0 < \rho < 1$ and $\tau > 0$ to be determined. This implies the estimate (6.52) by induction. Our strategy for proving (6.53) consists in inferring a consequence of each (CoSaMP) step — namely, discarding $\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}$, (CoSaMP$_1$) yields an estimate for $\|\big(\mathbf{x}_S - \mathbf{u}^{n+1}\big)_{\overline{U^{n+1}}}\|_2$ in terms of $\|\mathbf{x}^n - \mathbf{x}_S\|_2$, (CoSaMP$_2$) yields an estimate for $\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2$ in terms of $\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2$, and (CoSaMP$_3$) yields an estimate for $\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2$ in terms of $\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2$ and $\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2$, so overall an estimate for $\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_2$ in terms of $\|\mathbf{x}^n - \mathbf{x}_S\|_2$ is deduced.

We start with (CoSaMP$_3$). Specifically, we observe that $\mathbf{x}^{n+1}$ is a better (or at least equally good) s-term approximation to $\mathbf{u}^{n+1}$ than $\mathbf{x}_{S \cap U^{n+1}}$. Denoting $S^{n+1} = \operatorname{supp}(\mathbf{x}^{n+1})$ and observing that $S^{n+1} \subseteq U^{n+1}$, we conclude that

$$\begin{aligned}\|(\mathbf{x}_S - \mathbf{x}^{n+1})_{U^{n+1}}\|_2 &= \|\mathbf{x}_{S \cap \mathbf{U}^{n+1}} - \mathbf{x}^{n+1}\|_2 \\ &\le \|\mathbf{u}^{n+1} - \mathbf{x}^{n+1}\|_2 + \|\mathbf{u}^{n+1} - \mathbf{x}_{S \cap U^{n+1}}\|_2 \\ &\le 2\|\mathbf{u}^{n+1} - \mathbf{x}_{S \cap U^{n+1}}\|_2 = 2\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2.\end{aligned}$$

Then, using $(\mathbf{x}^{n+1})_{\overline{U^{n+1}}} = \mathbf{0}$ and $(\mathbf{u}^{n+1})_{\overline{U^{n+1}}} = \mathbf{0}$, it follows that

$$\begin{aligned}\|\mathbf{x}_S - \mathbf{x}^{n+1}\|_2^2 &= \|(\mathbf{x}_S - \mathbf{x}^{n+1})_{\overline{U^{n+1}}}\|_2^2 + \|(\mathbf{x}_S - \mathbf{x}^{n+1})_{U^{n+1}}\|_2^2 \\ &\le \|(\mathbf{x}_S - \mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2^2 + 4\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2^2. \qquad (6.54)\end{aligned}$$

Now, as a consequence of (CoSaMP$_2$), the vector $\mathbf{A}\mathbf{u}^{n+1}$ is characterized by

$$\langle \mathbf{y} - \mathbf{A}\mathbf{u}^{n+1}, \mathbf{A}\mathbf{z} \rangle = 0 \qquad \text{whenever } \operatorname{supp}(\mathbf{z}) \subseteq U^{n+1}.$$

This is equivalent to $\langle \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{u}^{n+1}), \mathbf{z} \rangle = 0$ whenever $\operatorname{supp}(\mathbf{z}) \subseteq U^{n+1}$, or to $(\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{u}^{n+1}))_{U^{n+1}} = \mathbf{0}$. Since $\mathbf{y} = \mathbf{A}\mathbf{x}_S + \mathbf{e}'$ with $\mathbf{e}' := \mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}$, this means

$$(\mathbf{A}^*\mathbf{A}(\mathbf{x}_S - \mathbf{u}^{n+1}))_{U^{n+1}} = -(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}.$$

We make use of this fact to obtain

$$\begin{aligned}\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 &\le \|\big((\mathbf{Id} - \mathbf{A}^*\mathbf{A})(\mathbf{x}_S - \mathbf{u}^{n+1})\big)_{U^{n+1}}\|_2 + \|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2 \\ &\le \delta_{4s}\|\mathbf{x}_S - \mathbf{u}^{n+1}\|_2 + \|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2,\end{aligned}$$

where the last inequality follows from Lemma 6.15. In other words, we have

$$\begin{aligned}&\big[\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 - \|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2\big]^2 \\ &\qquad \le \delta_{4s}^2\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2^2 + \delta_{4s}^2\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2^2.\end{aligned}$$

Using the identity $a^2 - b^2 = (a+b)(a-b)$, we derive

$$\begin{aligned}\delta_{4s}^2\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2^2 &\ge (1 - \delta_{4s}^2) \\ &\quad \times \Big(\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 - \frac{1}{1+\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2\Big) \\ &\quad \times \Big(\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 - \frac{1}{1-\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2\Big). \qquad (6.55)\end{aligned}$$

We may assume $\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 > \|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2/(1-\delta_{4s})$ to make the bottom term positive — otherwise (6.54) and (6.57) below imply the desired estimate (6.53), see Exercise 6.23. Thus, bounding the middle term from below by the bottom term, we obtain

$$\frac{\delta_{4s}^2}{1-\delta_{4s}^2}\|(\mathbf{x}_S-\mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2^2 \geq \Big(\|(\mathbf{x}_S-\mathbf{u}^{n+1})_{U^{n+1}}\|_2-\frac{1}{1-\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2\Big)^2.$$

Taking the square root and rearranging gives

$$\begin{aligned}\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 \leq \frac{\delta_{4s}}{\sqrt{1-\delta_{4s}^2}}\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2 \\ + \frac{1}{1-\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2. \qquad (6.56)\end{aligned}$$

Next if S^n denotes the support of $\mathbf{x}^n$ and if T^{n+1} denotes a set of $2s$ largest entries of $\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)$, we have

$$\|\big(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{S\cup S^n}\|_2^2 \leq \|\big(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{T^{n+1}}\|_2^2.$$

Eliminating the contribution on $(S\cup S^n)\cap T^{n+1}$, we derive

$$\|\big(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{(S\cup S^n)\setminus T^{n+1}}\|_2 \leq \|\big(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{T^{n+1}\setminus(S\cup S^n)}\|_2.$$

The right-hand side may be written as

$$\|\big(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{T^{n+1}\setminus(S\cup S^n)}\|_2 = \|\big(\mathbf{x}^n-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{T^{n+1}\setminus(S\cup S^n)}\|_2.$$

The left-hand side satisfies

$$\begin{aligned}\|\big(\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{(S\cup S^n)\setminus T^{n+1}}\|_2 \geq \|\big(\mathbf{x}_S-\mathbf{x}^n\big)_{\overline{T^{n+1}}}\|_2 \\ - \|\big(\mathbf{x}^n-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{(S\cup S^n)\setminus T^{n+1}}\|_2.\end{aligned}$$

These observations imply that

$$\begin{aligned}\|\big(\mathbf{x}_S-\mathbf{x}^n\big)_{\overline{T^{n+1}}}\|_2 &\leq \|\big(\mathbf{x}^n-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{(S\cup S^n)\setminus T^{n+1}}\|_2 \\ &+ \|\big(\mathbf{x}^n-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{T^{n+1}\setminus(S\cup S^n)}\|_2 \\ &\leq \sqrt{2}\,\|\big(\mathbf{x}^n-\mathbf{x}_S+\mathbf{A}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^n)\big)_{(S\cup S^n)\Delta T^{n+1}}\|_2 \\ &\leq \sqrt{2}\,\|\big((\mathbf{Id}-\mathbf{A}^*\mathbf{A})(\mathbf{x}^n-\mathbf{x}_S)\big)_{(S\cup S^n)\Delta T^{n+1}}\|_2 \\ &+ \sqrt{2}\,\|\big(\mathbf{A}^*\mathbf{e}'\big)_{(S\cup S^n)\Delta T^{n+1}}\|_2,\end{aligned}$$

where $(S\cup S^n)\Delta T^{n+1}$ denotes the symmetric difference of the sets $S\cup S^n$ and T^{n+1} and where $\mathbf{y}=\mathbf{A}\mathbf{x}_S+\mathbf{e}'$ has been used. Since $T^{n+1}\subseteq U^{n+1}$ by (CoSaMP$_1$) and $S^n\subseteq U^{n+1}$ by (CoSaMP$_3$), the left-hand side can be bounded from below as

$$\|\big(\mathbf{x}_S-\mathbf{x}^n\big)_{\overline{T^{n+1}}}\|_2 \geq \|\big(\mathbf{x}_S-\mathbf{x}^n\big)_{\overline{U^{n+1}}}\|_2 = \|(\mathbf{x}_S)_{\overline{U^{n+1}}}\|_2 = \|\big(\mathbf{x}_S-\mathbf{u}^{n+1}\big)_{\overline{U^{n+1}}}\|_2.$$

Since the right-hand side can be bounded from above using Lemma 6.15, we derive accordingly

$$\begin{aligned}\|\big(\mathbf{x}_S-\mathbf{u}^{n+1}\big)_{\overline{U^{n+1}}}\|_2 &\leq \sqrt{2}\,\delta_{4s}\,\|\mathbf{x}^n-\mathbf{x}_S\|_2 \\ &\quad + \sqrt{2}\,\|\big(\mathbf{A}^*\mathbf{e}'\big)_{(S\cup S^n)\Delta T^{n+1}}\|_2. \end{aligned} \tag{6.57}$$

It remains to put (6.54), (6.56), and (6.57) together. First combining (6.54) and (6.56), and using the inequality $a^2+(b+c)^2 \leq (\sqrt{a^2+b^2}+c)^2$, gives

$$\begin{aligned}\|\mathbf{x}_S-\mathbf{x}^{n+1}\|_2^2 &\leq \|(\mathbf{x}_S-\mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2^2 \\ &\quad + 4\Big(\frac{\delta_{4s}}{\sqrt{1-\delta_{4s}^2}}\|(\mathbf{x}_S-\mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2 + \frac{1}{1-\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2\Big)^2 \\ &\leq \Big(\sqrt{\frac{1+3\delta_{4s}^2}{1-\delta_{4s}^2}}\|(\mathbf{x}_S-\mathbf{u}^{n+1})_{\overline{U^{n+1}}}\|_2 + \frac{2}{1-\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2\Big)^2.\end{aligned}$$

Next, taking (6.57) into account we obtain

$$\begin{aligned}\|\mathbf{x}_S-\mathbf{x}^{n+1}\|_2 &\leq \sqrt{\frac{2\delta_{4s}^2(1+3\delta_{4s}^2)}{1-\delta_{4s}^2}}\|\mathbf{x}^n-\mathbf{x}_S\|_2 \\ &\quad + \sqrt{\frac{2(1+3\delta_{4s}^2)}{1-\delta_{4s}^2}}\|\big(\mathbf{A}^*\mathbf{e}'\big)_{(S\cup S^n)\Delta T^{n+1}}\|_2 + \frac{2}{1-\delta_{4s}}\|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2.\end{aligned}$$

In view of Lemma 6.19, we conclude that the desired inequality (6.53) holds with

$$\rho = \sqrt{\frac{2\delta_{4s}^2(1+3\delta_{4s}^2)}{1-\delta_{4s}^2}}, \quad (1-\rho)\tau = \sqrt{\frac{2(1+3\delta_{4s}^2)}{1-\delta_{4s}}} + \frac{2\sqrt{1+\delta_{4s}}}{1-\delta_{4s}}.$$

The constant ρ is less than one if and only if $6\delta_{4s}^4+3\delta_{4s}^2-1<0$. This occurs as soon as δ_{4s}^2 is smaller than the largest root of $6t^2+3t-1$, i.e., as soon as $\delta_{4s}^2 < (\sqrt{11/3}-1)/4$, which is Condition (6.51). □

Notes

E. Candès and T. Tao introduced the concept of uniform uncertainty principle in [82], which they refined by defining the restricted isometry constants and the restricted orthogonality constants in [81]. In the latter, they proved the inequality $\delta_{s+t} \leq \max(\delta_s,\delta_t)+\theta_{s,t}$. The slightly improved inequality in Proposition 6.5 is believed to be new. Some authors define the restricted isometry constants 'without squares'. For instance, A. Cohen, W. Dahmen, and R. DeVore considered in [102] the smallest $\delta \geq 0$ such that the inequality

$$(1-\delta)\|\mathbf{x}\|_2 \le \|\mathbf{A}\mathbf{x}\|_2 \le (1+\delta)\|\mathbf{x}\|_2$$

holds for all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$. Up to transformation of the constants, this is of course essentially equivalent to our definition.

E. Candès and T. Tao showed in [81] that the condition $\delta_{2s}+\delta_{3s} < 1$ guarantees exact s-sparse recovery via ℓ_1-minimization. E. Candès, J. Romberg, and T. Tao further showed in [80] that the condition $\delta_{3s} + 3\delta_{4s} < 2$ guarantees stable and robust s-sparse recovery via ℓ_1-minimization. Later, a sufficient condition for stable and robust s-sparse recovery involving only δ_{2s} was obtained by E. Candès in [70], namely $\delta_{2s} < \sqrt{2}-1 \approx 0.414$. This sufficient condition was improved several times, see [184, 69, 179, 68, 305]. Exercises 6.12 through 6.15 retrace some of these improvements. Central to some of these improvements is the *shifting inequality* — see Exercise 6.14 — put forward by T. Cai, L. Wang, and G. Xu in [69]. They also introduced the square root lifting inequality of Lemma 6.13 in [68]. The condition $\delta_{2s} < 0.4931$ of Theorem 6.11 is the best available so far. It is due to Q. Mo and S. Li in [305]. On the other hand, M. Davies and R. Gribonval constructed in [120] matrices with restricted isometry constant δ_{2s} arbitrarily close to $1/\sqrt{2} \approx 0.707$ for which some s-sparse vectors are not recovered via ℓ_1-minimization. The natural proof of Theorem 6.8, with the sufficient condition $\delta_{2s} < 1/3$, does not seem to have appeared before. We point out that other sufficient conditions involving δ_k with $k \neq 2s$ can also be found, see for instance Exercises 6.13, 6.15, and 6.16. As a matter of fact, J. Blanchard and A. Thompson argue that the parameter $2s$ is not the best choice for Gaussian random matrices, see [44]. Theorem 6.7, which appeared in [180], has to be kept in mind when assessing such conditions.

The use of the iterative hard thresholding algorithm in the context of Compressive Sensing was initiated by T. Blumensath and M. Davies in [46]. In [47], they established stable and robust estimates under the sufficient condition $\delta_{3s} < 1/\sqrt{8}$. The weaker condition $\delta_{3s} < 1/2$ of Theorem 6.14 appeared in [181]. The improved condition $\delta_{3s} < 1/\sqrt{3}$ of Theorem 6.17 was established in the paper [180] dedicated to the analysis of the hard thresholding pursuit algorithm. There, Theorem 6.17 was in fact established for a family of thresholding algorithms indexed by an integer k, with iterative hard thresholding and hard thresholding pursuit corresponding to the cases $k = 0$ and $k = \infty$, respectively. Exercise 6.19, which considers a variation of the iterative hard thresholding algorithm where a factor $\mu \neq 1$ is introduced in front of $\mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)$, is inspired by the paper [188] by R. Garg and R. Khandekar. This factor μ may be dependent on n in some algorithms, notably in the normalized iterative hard thresholding algorithm of T. Blumensath and M. Davies [48].

The impossibility of s-sparse recovery via s iterations of Orthogonal Matching Pursuit under a standard restricted isometry condition was first observed in [131, Section 7], see also [353]. The example given at the beginning of Section 6.4 is taken from the article [306] by Q. Mo and Y. Shen,

who also established the result of Exercise 6.21. The possibility of s-sparse recovery via a number of iterations of orthogonal matching pursuit that is proportional to s was shown in [451] by T. Zhang, who also proved the stability and robustness of the recovery by establishing Proposition 6.23 with $\bar{n} = 30\,\mathrm{card}(\mathrm{supp}(\mathbf{x}) \setminus S^0)$ under the condition $\delta_{31s} < 1/3$. Our proof follows his argument, which is also valid in more general settings.

In the original article [312] of D. Needell and J. Tropp introducing the compressive sampling matching pursuit algorithm, stability and robustness were stated under the condition $\delta_{4s} \leq 0.1$, although the arguments actually yield the condition $\delta_{4s} < 0.17157$. Theorem 6.26, which gives the condition $\delta_{4s} < 0.4782$, appears here for the first time. The first analysis of a greedy algorithm under the restricted isometry property appeared in [313, 314] for the regularized orthogonal matching pursuit algorithm where, however, an additional $\ln(N)$-factor appeared in the condition on the restricted isometry constant. The Subspace Pursuit algorithm of W. Dai and O. Milenkovic was also proved to be stable and robust under some restricted isometry conditions. We refer to the original paper [110] for details.

We mentioned at the end of Section 6.1 that the most of the available bounds of the restricted isometry property for explicit (deterministic) matrix constructions are based on the coherence and therefore the number m of required samples scales quadratically in the sparsity s, see also (6.14). A notable exception is a sophisticated explicit matrix construction by J. Bourgain, S. Dilworth, K. Ford, S. Konyagin and D. Kutzarova [53], see also [54]. The authors showed that their matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ has small restricted isometry constants δ_s once $m \geq Cs^{2-\varepsilon}$ and when $s^{2-\varepsilon} <\leq N \leq s^{2+\varepsilon}$ for some $\varepsilon > 0$. While this slightly overcomes the quadratic bottleneck, this range of s, m, N is too limited in order to make it relevant for practical purposes. Nevertheless, it is certainly a very important contribution to the theory.

Exercises

6.1. Suppose that $\mathbf{A} \in \mathbb{C}^{m \times N}$ has an sth order restricted isometry constant satisfying $\delta_s < 1$. Prove that, for any $S \subseteq [N]$ with $\mathrm{card}(S) \leq s$,

$$\frac{1}{1+\delta_s} \leq \big\|(\mathbf{A}_S^* \mathbf{A}_S)^{-1}\big\|_{2\to 2} \leq \frac{1}{1-\delta_s} \quad \text{and} \quad \frac{1}{\sqrt{1+\delta_s}} \leq \big\|\mathbf{A}_S^\dagger\big\|_{2\to 2} \leq \frac{1}{\sqrt{1-\delta_s}}.$$

6.2. Given $\mathbf{A} \in \mathbb{C}^{m \times N}$, let α_s and β_s be the largest and smallest positive constants α and β such that

$$\alpha \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq \beta \|\mathbf{x}\|_2^2$$

for all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$. Find the scaling factor $t > 0$ for which $\delta_s(t\mathbf{A})$ takes its minimal value, and prove that this value equals $(\beta - \alpha)/(\beta + \alpha)$.

6.3. Find a matrix $\mathbf{A} \in \mathbb{R}^{2\times 3}$ with minimal 2nd order restricted isometry constant.

6.4. Prove the equivalence of the two definitions (6.4) and (6.5) of restricted orthogonality constants.

6.5. Verify in details that the function f defined on $[0,1]$ as in (6.6) is first nondecreasing, then nonincreasing.

6.6. Given $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{A} \in \mathbb{C}^{m\times N}$ with sth restricted isometry constant δ_s, prove that

$$\|\mathbf{A}\mathbf{x}\|_2 \le \sqrt{1+\delta_s}\Big(\|\mathbf{x}\|_2 + \frac{\|\mathbf{x}\|_1}{\sqrt{s}}\Big).$$

6.7. Let $D_{s,N} = \{\mathbf{x} \in \mathbb{C}^N : \|\mathbf{x}\|_2 \le 1, \|\mathbf{x}\|_0 \le s\}$ be the Euclidean unit ball restricted to the s-sparse vectors. Show that

$$D_{s,N} \subset \operatorname{conv}(D_{s,N}) \subset \sqrt{s}B_1^N \cap B_2^N \subset 2\operatorname{conv}(D_{s,N}),$$

where $B_p^N = \{\mathbf{x} \in \mathbb{C}^N : \|\mathbf{x}\|_p \le 1\}$ is the unit ball in ℓ_p and conv denotes the convex hull, see Definition B.2.

6.8. Prove Proposition 6.3 directly from (6.1), without using (6.2) but rather with the help of the *polarization formula*

$$\operatorname{Re}\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{4}\big(\|\mathbf{x}+\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\big).$$

6.9. In the case $t = ns$ where t is a multiple of s, improve the second inequality of Proposition 6.6 by showing that

$$\delta_{ns} \le (n-1)\theta_{s,s} + \delta_s.$$

6.10. Suppose that the columns of the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ are ℓ_2-normalized. Under the assumption $N > s^2 + 1$, derive the result of Theorem 6.7 with constants $c = 1/2$, $C = 2$, and without restriction on δ_*. Use Theorem 5.8, Exercise 5.3, and compare the matrix norms induced by the ℓ_2 and the ℓ_1 norms.

6.11. Refine the proof of Theorem 6.8 in order to establish the stability and robustness of s-sparse recovery via basis pursuit when $\delta_{2s} < 1/3$.

6.12. Let $\mathbf{A} \in \mathbb{C}^{m\times N}$, and let $S_0, S_1, S_2, \ldots$ denote index sets of size s ordered by decreasing modulus of entries of a vector $\mathbf{v} \in \ker \mathbf{A}$. Prove that

$$\|\mathbf{v}_{S_0}\|_2^2 + \|\mathbf{v}_{S_1}\|_2^2 \le \frac{2\delta_{2s}}{1-\delta_{2s}} \sum_{k\ge 2} \|\mathbf{v}_{S_k}\|_2 \, (\|\mathbf{v}_{S_0}\|_2 + \|\mathbf{v}_{S_1}\|_2).$$

By interpreting this as the equation of a disk or by completing squares, deduce that

$$\|\mathbf{v}_{S_0}\|_2 \le \frac{\rho}{\sqrt{s}}\|\mathbf{v}_{\overline{S_0}}\|_1, \qquad \text{where } \rho := \frac{1+\sqrt{2}}{2}\frac{\delta_{2s}}{1-\delta_{2s}}.$$

Conclude that s-sparse recovery via basis pursuit is guaranteed if $\delta_{2s} < 0.453$.

6.13. For an integer $k \geq 1$, suppose that $\mathbf{A} \in \mathbb{C}^{m \times N}$ has restricted isometry constant $\delta_{(2k+1)s} < 1 - 1/\sqrt{2k}$. Prove that every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ via ℓ_1-minimization. [Hint: to establish the null space property, partition $[N]$ as $S \cup T_1 \cup T_2 \cup \ldots$, where S has size s and $T_1, T_2, \ldots$ have size ks.]

6.14. Given $a_1 \geq a_2 \geq \cdots \geq a_{k+\ell} \geq 0$, prove the *shifting inequality*

$$\sqrt{a_{\ell+1}^2 + \cdots + a_{\ell+k}^2} \leq c_{k,\ell}(a_1 + \cdots + a_k), \qquad \text{where } c_{k,\ell} := \max\Big(\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{4\ell}}\Big).$$

6.15. Suppose that $s =: 4r$ is a multiple of 4. For a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$, establish the success of s-sparse recovery via basis pursuit if $\delta_{5r} + \theta_{5r,s} < 1$. Show in particular that this holds if $\delta_{9s/4} < 0.5$, $\delta_{2s} < 1/(1 + \sqrt{5/4}) \approx 0.472$, or $\delta_{5s/4} < 1/(1 + \sqrt{10/3}) \approx 0.353$.

6.16. Using the square root lifting inequality of Lemma 6.13, find a condition on δ_s that guarantees the exact recovery of every s-sparse vector via basis pursuit.

6.17. Refine the proof of Theorem 6.14 in order to establish the stability and robustness of s-sparse recovery via iterative hard thresholding when $\delta_{3s} < 1/3$.

6.18. Given $\mathbf{A} \in \mathbb{C}^{m \times N}$, prove that every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ via iterative hard thresholding if $\delta_{2s} < 1/4$. To do so, return to the proof of Theorem 6.14 , precisely to (6.32), and separate the contributions to the inner product from the index sets of size $2s$ given by

$$(S \cup S^n) \cap (S \cup S^{n+1}), \quad (S \cup S^n) \setminus (S \cup S^{n+1}), \quad (S \cup S^{n+1}) \setminus (S \cup S^n),$$

where $S := \operatorname{supp}(\mathbf{x})$, $S^n := \operatorname{supp}(\mathbf{x}^n)$, and $S^{n+1} := \operatorname{supp}(\mathbf{x}^{n+1})$.

6.19. Given $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ for some s-sparse $\mathbf{x} \in \mathbb{C}^N$, we define a sequence $(\mathbf{x}^n)$ inductively, starting with an initial s-sparse vector $\mathbf{x}^0 \in \mathbb{C}^N$, by

$$\mathbf{x}^{n+1} = H_s(\mathbf{x}^n + \mu\, \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)), \qquad n \geq 0,$$

where the constant μ is to be determined later. Establish the identity

$$\begin{aligned} &\|\mathbf{A}(\mathbf{x}^{n+1} - \mathbf{x})\|_2^2 - \|\mathbf{A}(\mathbf{x}^n - \mathbf{x})\|_2^2 \\ &\qquad = \|\mathbf{A}(\mathbf{x}^{n+1} - \mathbf{x}^n)\|_2^2 + 2\langle \mathbf{x}^n - \mathbf{x}^{n+1}, \mathbf{A}^*\mathbf{A}(\mathbf{x} - \mathbf{x}^n)\rangle. \end{aligned}$$

Prove also the inequality

$$\begin{aligned} &2\mu\langle \mathbf{x}^n - \mathbf{x}^{n+1}, \mathbf{A}^*\mathbf{A}(\mathbf{x} - \mathbf{x}^n)\rangle \\ &\qquad \leq \|\mathbf{x}^n - \mathbf{x}\|_2^2 - 2\mu\|\mathbf{A}(\mathbf{x}^n - \mathbf{x})\|_2^2 - \|\mathbf{x}^{n+1} - \mathbf{x}^n\|_2^2. \end{aligned}$$

With δ_{2s} denoting the $2s$th order restricted isometry constant of $\mathbf{A}$, derive the inequality

$$\|\mathbf{A}(\mathbf{x}^{n+1}-\mathbf{x})\|_2^2 \le \Big(1-\frac{1}{\mu(1+\delta_{2s})}\Big)\|\mathbf{A}(\mathbf{x}^{n+1}-\mathbf{x}^n)\|_2^2 + \Big(\frac{1}{\mu(1-\delta_{2s})}-1\Big)\|\mathbf{A}(\mathbf{x}^n-\mathbf{x})\|_2^2.$$

Deduce that the sequence $(\mathbf{x}^n)$ converges to $\mathbf{x}$ when $1+\delta_{2s} < 1/\mu < 2(1-\delta_{2s})$. Conclude by justifying the choice $\mu = 3/4$ under the condition $\delta_{2s} < 1/3$.

6.20. Verify the claims made at the beginning of Section 6.4 about the matrix $\mathbf{A}$ defined in (6.44).

6.21. Prove that every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ via s iterations of Orthogonal Matching Pursuit provided the restricted isometry constant of $\mathbf{A}$ satisfies

$$\delta_{s+1} < \frac{1}{\sqrt{s}+1}.$$

6.22. Improve Lemma 6.23 in the case $\mathbf{e} = 0$ by reducing the number of required iterations and by weakening the restricted isometry condition.

6.23. Verify that $\|(\mathbf{x}_S - \mathbf{u}^{n+1})_{U^{n+1}}\|_2 > \|(\mathbf{A}^*\mathbf{e}')_{U^{n+1}}\|_2/(1-\delta_{4s})$ can indeed be assumed after (6.55) in the proof of Theorem 6.26.

6.24. Rank Restricted Isometry Property.
Let $\mathcal{A} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ be a linear map. For $r \le \min\{n_1, n_2\}$ the rank restricted isometry constant $\delta_r = \delta_r(\mathcal{A})$ is the defined as the smallest number such that

$$(1-\delta_r)\|\mathbf{X}\|_F^2 \le \|\mathcal{A}(\mathbf{X})\|_2^2 \le (1+\delta_r)\|\mathbf{X}\|_F^2$$

for all matrix $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ of rank at most r.

(a) Let $\mathbf{X}, \mathbf{Z} \in \mathbb{C}^{n_1\times n_2}$ with $\langle \mathbf{X}, \mathbf{Z}\rangle_F = \mathrm{tr}\,(\mathbf{X}\mathbf{Z}^*) = 0$ and $\mathrm{rank}(\mathbf{X}) + \mathrm{rank}(\mathbf{Z}) \le r$. Show that

$$|\langle \mathcal{A}(\mathbf{X}), \mathcal{A}(\mathbf{Z})\rangle| \le \delta_r \|\mathbf{X}\|_F \|\mathbf{Z}\|_F\ .$$

(b) Assume that $\delta_{2r} < 1/3$. Show that $\mathcal{A}$ possesses the rank null space property of order r defined by (4.44). In particular, every $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ of rank at most r is the unique solution to the nuclear norm minimization problem (see also Section 4.5)

$$\min_{\mathbf{Z}\in\mathbb{C}^{n_1\times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } \mathcal{A}(\mathbf{Z}) = \mathcal{A}(\mathbf{X})\ .$$

(c) Assume that $\delta_{2r} < 0.4931$. Let $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ and $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ with $\|\mathbf{e}\|_2 \le 1$. Let $\mathbf{X}^\sharp$ be the solution to the quadratically constraint nuclear norm minimization problem

$$\min_{\mathbf{Z}\in\mathbb{C}^{n_1\times n_2}} \|\mathbf{Z}\|_* \quad \text{subject to } \|\mathcal{A}(\mathbf{Z})-\mathbf{y}\|_2 \le \eta\,.$$

Show that

$$\|\mathbf{X}-\mathbf{X}^\sharp\|_F \le \frac{C_1}{\sqrt{r}} \sum_{\ell=r+1}^{\min\{n_1,n_2\}} \sigma_\ell(\mathbf{X}) + C_2\eta$$

for appropriate constants $C_1, C_2 > 0$ depending only on δ_{2r}.

7

Basic Tools from Probability Theory

The major breakthrough in proving recovery results in compressive sensing is obtained using random matrices. Most parts of the remainder of this book indeed requires tools from probability theory. This and the next chapter are therefore somewhat exceptional in the sense that they do not deal directly with compressive sensing. Instead, we rather collect the necessary background material from probability theory. In this chapter we introduce a first set of tools that will be sufficient to understand a large part of the theory in connection with sparse recovery and random matrices. More advanced tools that will be used only in parts of the remainder of the book are postponed to Chapter 8.

We only assume that the reader has basic knowledge of probability theory as can be found in most introductory textbooks on the subject. We recall the most basic facts of probability in Section 7.1. The relation of moments of random variables to their tails is presented in Section 7.2. Then in Section 7.3 we study deviation inequalities for sums of independent random variables by means of moment generating function. Cramér's theorem gives a very general estimate from which we deduce Hoeffding's inequality, and later in Section 7.5 Bernstein's inequality for bounded and subgaussian random variables. We introduce the latter in Section 7.4.

The theory presented in this chapter will be sufficient to follow Sections 9.1 and 9.2, Chapter 11, and Chapter 14. For the remaining parts of Chapter 9 as well as for Chapters 12 and 13 more advanced tools from probability theory will be required, which will be introduced in Chapter 8.

7.1 Essentials from Probability

In this section we recall some important facts from basic probability theory, and prove simple statements that might not be found in all basic textbooks.

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, where Σ denotes a σ-algebra on the sample space Ω and $\mathbb{P}$ a probability measure on (Ω, Σ). The probability of an event $B \in \Sigma$ is denoted by

$$\mathbb{P}(B) = \int_B d\mathbb{P}(\omega) = \int_\Omega I_B(\omega) d\mathbb{P}(\omega) \ ,$$

where the characteristic function $I_B(\omega)$ takes the value 1 if $\omega \in B$ and 0 otherwise. The *union bound* (or Bonferroni's inequality, or Boole's inequality) states that for a collection of events $B_\ell \in \Sigma$, $\ell = 1, \ldots, n$, we have

$$\mathbb{P}\left(\bigcup_{\ell=1}^n B_\ell\right) \le \sum_{\ell=1}^n \mathbb{P}(B_\ell) \ . \tag{7.1}$$

A *random variable* X is a real-valued measurable function on (Ω, Σ). Recall that X is called measurable if the preimage $X^{-1}(A) = \{\omega \in \Omega : X(\Omega) \in A\}$ is contained in Σ for all Borel measurable subsets $A \subset \mathbb{R}$. Usually, every reasonable function X will be measurable; in particular, all functions appearing in this book. In what follows we will usually not mention the underlying probability space $(\Omega, \Sigma, \mathbb{P})$ when speaking about random variables. The *distribution function* $F = F_X$ of X is defined as

$$F(t) = \mathbb{P}(X \le t), \quad t \in \mathbb{R} \ .$$

A random variable X possesses a *probability density function* $\phi : \mathbb{R} \to \mathbb{R}_+$ if

$$\mathbb{P}(a < X \le b) = \int_a^b \phi(t) dt \quad \text{for all } a < b \in \mathbb{R} \ . \tag{7.2}$$

Then $\phi(t) = \frac{d}{dt} F(t)$. The *expectation* or mean of a random variable will be denoted by

$$\mathbb{E}X = \int_\Omega X(\omega) d\mathbb{P}(\omega) \ .$$

If X has probability density function ϕ then for a function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}g(X) = \int_{-\infty}^\infty g(t)\phi(t) dt \tag{7.3}$$

whenever the integral exists. The quantities $\mathbb{E}X^p$, $p > 0$ are called *moments* of X, while $\mathbb{E}|X|^p$ are called *absolute moments*. (Sometimes we may omit "absolute".) The quantity $\mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ is called *variance*. For $1 \le p < \infty$, $(\mathbb{E}|X|^p)^{1/p}$ defines a norm on the $L^p(\Omega, \mathbb{P})$-space of all p-integrable random variables, in particular, the triangle inequality

$$(\mathbb{E}|X + Y|^p)^{1/p} \le (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p} \tag{7.4}$$

holds for all p-integrable random variables X, Y on $(\Omega, \Sigma, \mathbb{P})$.

Hölder's inequality states that, for random variables X, Y on a common probability space and $p, q \ge 1$ with $1/p + 1/q = 1$, we have

$$|\mathbb{E}XY| \le (\mathbb{E}|X|^p)^{1/q} \, (\mathbb{E}|Y|^q)^{1/q} \ .$$

The special case $p = q = 2$ is the *Cauchy-Schwarz inequality*,

$$|\mathbb{E}XY| \leq \sqrt{\mathbb{E}|X|^2 \mathbb{E}|Y|^2} \, .$$

Since the constant (deterministic) random variable 1 has expectation $\mathbb{E}1 = 1$, Hölder's inequality shows that $\mathbb{E}|X|^p = \mathbb{E}[1 \times |X|^p] \leq (\mathbb{E}|X|^{pr})^{1/r}$ for all $p > 0, r \geq 1$ and therefore, for all $0 < p \leq q < \infty$,

$$(\mathbb{E}|X|^p)^{1/p} \leq (\mathbb{E}|X|^q)^{1/q} \, . \tag{7.5}$$

Let X_n, $n \in \mathbb{N}$, be a sequence of random variables such that X_n converges to X as $n \to \infty$ in the sense that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all ω. *Lebesgue's dominated convergence theorem* states that if there exists a random variable Y with $\mathbb{E}|Y| < \infty$ such that $|X_n| \leq |Y|$ a.s. then $\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X$. Lebesgue's dominated convergence theorem has as well an obvious formulation for integrals of sequences of functions.

Fubini's theorem on the integration of functions of two variables can be formulated as follows. Let $f : A \times B \to \mathbb{C}$ be measurable, where (A, ν) and (B, μ) are measurable spaces. If $\int_{A\times B} |f(x, y)| d(\nu \otimes \mu)(x, y) < \infty$ then

$$\int_A \left(\int_B f(x, y) d\mu(y) \right) d\nu(x) = \int_B \left(\int_A f(x, y) d\mu(x) \right) d\mu(y) \, .$$

A formulation for expectations of functions of independent random vectors is provided below in (7.15).

Absolute moments can be computed by means of the following formula.

Proposition 7.1. *The absolute moments of a random variable X can be expressed as*

$$\mathbb{E}|X|^p = p \int_0^\infty \mathbb{P}(|X| \geq t) t^{p-1} dt, \quad p > 0 \, .$$

Proof. Recall that $I_{\{|X|^p \geq x\}}$ is the random variable that takes the value 1 on the event $|X|^p \geq x$ and 0 otherwise. Using Fubini's theorem we derive

$$\begin{aligned} \mathbb{E}|X|^p &= \int_\Omega |X|^p d\mathbb{P} = \int_\Omega \int_0^{|X|^p} 1 dx d\mathbb{P} = \int_\Omega \int_0^\infty I_{\{|X|^p \geq x\}} dx d\mathbb{P} \\ &= \int_0^\infty \int_\Omega I_{\{|X|^p \geq x\}} d\mathbb{P} dx = \int_0^\infty \mathbb{P}(|X|^p \geq x) dx \\ &= p \int_0^\infty \mathbb{P}(|X|^p \geq t^p) t^{p-1} dt = p \int_0^\infty \mathbb{P}(|X| \geq t) t^{p-1} dt \, , \end{aligned}$$

where we also applied a change of variables. □

Corollary 7.2. *For a random variable X the expectation satisfies*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t) dt - \int_0^\infty \mathbb{P}(X \leq -t) dt \, .$$

Proof. We can write $X = XI_{\{X\in[0,\infty)\}} + XI_{\{X\in(-\infty,0)\}}$ so that

$$\mathbb{E}X = \mathbb{E}XI_{\{X\in[0,\infty)\}} - \mathbb{E}(-XI_{\{-X\in(0,\infty)\}}).$$

Both $XI_{\{X\in[0,\infty)\}}$ and $-XI_{\{-X\in(0,\infty)\}}$ are positive random variables, so that an application of Proposition 7.1 shows the statement.

The function $t \mapsto \mathbb{P}(|X| \geq t)$ is called the *tail* of X. A simple but often effective tool to estimate the tail by expectations and moments is the *Markov inequality*.

Theorem 7.3. *Let X be a random variable. Then*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t} \quad \textit{for all } t > 0 \,.$$

Proof. Note that $\mathbb{P}(|X| \geq t) = \mathbb{E}I_{\{|X|\geq t\}}$ and $tI_{\{|X|\geq t\}} \leq |X|$. Hence, $t\mathbb{P}(|X| \geq t) = \mathbb{E}tI_{\{|X|\geq t\}} \leq \mathbb{E}|X|$ and the proof is complete. □

Remark 7.4. As an important consequence we note that for $p > 0$

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(|X|^p \geq t^p) \leq t^{-p}\mathbb{E}|X|^p, \quad \text{for all } t > 0 \,.$$

The special case $p = 2$ is referred to as the Chebyshev inequality. Similarly, for $\theta > 0$ we obtain

$$\mathbb{P}(X \geq t) = \mathbb{P}(\exp(\theta X) \geq \exp(\theta t)) \leq \exp(-\theta t)\mathbb{E}\exp(\theta X), \quad \text{for all } t \in \mathbb{R} \,.$$

The function $\theta \mapsto \mathbb{E}\exp(\theta X)$ is usually called the *Laplace transform* or the *moment generating function* of X.

The *median* of a random variable X is a number M such that

$$\mathbb{P}(X \geq M) \geq 1/2 \quad \text{and} \quad \mathbb{P}(X \leq M) \geq 1/2 \,.$$

The *binomial distribution* is the discrete probability distribution counting the number of successes in a sequence of N independent experiments where the probability of each individual success is p. If X has the binomial distribution then

$$\mathbb{P}(X = k) = \binom{N}{k} p^k (1-p)^{N-k} \,.$$

The expectation of X is given by $\mathbb{E}X = pN$. If pN is an integer then the median $M = M(X)$ coincides with the expection,

$$M(X) = pN \,. \tag{7.6}$$

A *normal distributed* random variable or *Gaussian random variable* X has probability density function

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) . \tag{7.7}$$

It has mean $\mathbb{E}X = \mu$ and variance $\mathbb{E}(X-\mu)^2 = \sigma^2$. A *standard Gaussian* random variable (or standard normal or simply standard Gaussian), usually denoted g, is a Gaussian random variable with $\mathbb{E}g = 0$ and $\mathbb{E}g^2 = 1$. Its tail satisfies the following simple estimates.

Proposition 7.5. *Let g be a standard Gaussian random variable. Then, for all $u > 0$,*

$$\mathbb{P}(|g| \geq u) \leq \exp(-u^2/2) , \tag{7.8}$$

$$\mathbb{P}(|g| \geq u) \leq \sqrt{\frac{2}{\pi}} \frac{1}{u} \exp\left(-\frac{u^2}{2}\right) , \tag{7.9}$$

$$\mathbb{P}(|g| \geq u) \geq \sqrt{\frac{2}{\pi}} \frac{1}{u} \left(1 - \frac{1}{u^2}\right) \exp\left(-\frac{u^2}{2}\right) ,$$

$$\mathbb{P}(|g| \geq u) \geq \left(1 - \sqrt{\frac{2}{\pi}} u\right) \exp\left(-\frac{u^2}{2}\right) .$$

Proof. By (7.7) we have

$$\mathbb{P}(|g| \geq u) = \frac{2}{\sqrt{2\pi}} \int_u^\infty e^{-t^2/2} dt . \tag{7.10}$$

Therefore, the stated estimates follow from Lemma C.8 and C.9. □

Let us compute the moment generating function of a standard Gaussian.

Lemma 7.6. *Let g be a standard Gaussian random variable. Then, for $\theta \in \mathbb{R}$,*

$$\mathbb{E} \exp(\theta g) = \exp\left(\theta^2/2\right) , \tag{7.11}$$

and more generally, for $\theta \in \mathbb{R}, a < 1/2$,

$$\mathbb{E} \exp(ag^2 + \theta g) = \frac{1}{\sqrt{1-2a}} \exp\left(\frac{\theta^2}{2(1-2a)}\right) .$$

Proof. For $\theta, a \in \mathbb{R}$, we have

$$\mathbb{E}\left(\exp(ag^2 + \theta g)\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ax^2 + \theta x) \exp\left(-x^2/2\right) dx$$

Noting the identity

$$ax^2 - x^2/2 + \theta x = -\frac{1-2a}{2} \left(x - \frac{\theta}{1-2a}\right)^2 - \frac{\theta}{2(1-2a)} .$$

After a change of variable, the latter integral reduces to the integral of the normal probability density function, so it equals one. We deduce that

$$\mathbb{E}\left(\exp(\theta g)\right) = \exp\left(\frac{\theta^2}{2}\right) \qquad \text{for all } \theta \in \mathbb{R}\,, \tag{7.12}$$

which proves the statement. □

The proof of the next result should explain the terminology *moment generating function*.

Corollary 7.7. *The even moments of a standard Gaussian random variable g are given by*

$$\mathbb{E}g^{2n} = \frac{(2n)!}{2^n n!}\,, \quad n \in \mathbb{N}\,.$$

Proof. On the one hand, by Taylor expansion we can write the moment generating function as

$$\mathbb{E}\exp(\theta g) = \sum_{j=0}^{\infty} \frac{\theta^j \mathbb{E}[g^j]}{j!} = \sum_{n=0}^{\infty} \frac{\theta^{2n}\mathbb{E}g^{2n}}{(2n)!}\,,$$

where we have used that $\mathbb{E}g^j = 0$ for all odd j. On the other hand, Lemma 7.6 gives

$$\mathbb{E}\exp(\theta g) = \exp(\theta^2/2) = \sum_{n=0}^{\infty} \frac{\theta^{2n}}{2^n n!}\,.$$

Comparing coefficients gives

$$\frac{\mathbb{E}g^{2n}}{(2n)!} = \frac{1}{2^n n!}\,,$$

which is equivalent to the claim. □

A *random vector* $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^n$ is a collection of n random variables on a common probability space $(\Omega, \Sigma, \mathbb{P})$. Its expectation is the vector $\mathbb{E}\mathbf{X} = [\mathbb{E}X_1, \ldots, \mathbb{E}X_n]^\top \in \mathbb{R}^n$, while its *joint distribution function* is defined as

$$F(t_1, \ldots, t_n) = \mathbb{P}(X_1 \le t_1, \ldots, X_n \le t_n), \quad t_1, \ldots, t_n \in \mathbb{R}\,.$$

Similarly to the univariate case, the random vector $\mathbf{X}$ has a *joint probability density* if there exists a function $\phi : \mathbb{R}^n \to [0, 1]$ such that for any measurable domain $D \subset \mathbb{R}^n$

$$\mathbb{P}(\mathbf{X} \in D) = \int_D \phi(t_1, \ldots, t_n) dt_1 \cdots dt_n\,.$$

A *complex random vector* $\mathbf{Z} = \mathbf{X} + i\mathbf{Y} \in \mathbb{C}^n$ is a special case of a $2n$-dimensional real random vector $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{2n}$.

A collection of random variables $X_1, \ldots, X_n$ is (stochastically) independent if, for all $t_1, \ldots, t_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n) = \prod_{\ell=1}^{n} \mathbb{P}(X_\ell \leq t_\ell) \; .$$

For independent random variables, we have

$$\mathbb{E}\left[\prod_{\ell=1}^{n} X_\ell\right] = \prod_{\ell=1}^{n} \mathbb{E}\left[X_\ell\right] \; . \tag{7.13}$$

If they have a joint probability density function ϕ then the latter factorizes as

$$\phi(t_1, \ldots, t_n) = \phi_1(t_1) \times \cdots \times \phi_n(t_n)$$

where the $\phi_1, \ldots, \phi_n$ are the probability density functions of $X_1, \ldots, X_n$.

In generalization, a collection $\mathbf{X}_1 \in \mathbb{R}^{n_1}, \ldots, \mathbf{X}_m \in \mathbb{R}^{n_m}$ of random vectors are independent if for any collection of measurable sets $A_\ell \subset \mathbb{R}^{n_\ell}$, $\ell \in [m]$,

$$\mathbb{P}(\mathbf{X}_1 \in A_1, \ldots, \mathbf{X}_m \in A_m) = \prod_{\ell=1}^{m} \mathbb{P}(\mathbf{X}_\ell \in A_\ell) \; .$$

If furthermore $f_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{N_\ell}$, $\ell = 1, \ldots, m$, are measurable functions then also the random vectors $f_1(\mathbf{X}_1), \ldots, f_m(\mathbf{X}_m)$ are independent. A collection $\mathbf{X}_1, \ldots, \mathbf{X}_m \in \mathbb{R}^n$ of independent random vectors that all have the same distribution is called *independent identically distributed* (i.i.d.).

A random vector $\mathbf{X}'$ will be called an independent copy of $\mathbf{X}$ if $\mathbf{X}$ and $\mathbf{X}'$ are independent and have the same distribution.

The sum $X + Y$ of two independent random variables X, Y having probability density functions ϕ_X, ϕ_Y, has probability density function ϕ_{X+Y} given by the convolution

$$\phi_{X+Y}(t) = (\phi_X * \phi_Y)(t) = \int_{-\infty}^{\infty} \phi_X(u)\phi_Y(t-u)du \; . \tag{7.14}$$

Fubini's theorem for expectations takes the following form. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^n$ be two independent random vectors (or simply random variables) and $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a measurable function such that $\mathbb{E}|f(\mathbf{X}, \mathbf{Y})| < \infty$. Then the functions

$$f_1 : \mathbb{R}^n \to \mathbb{R}, \; f_1(\mathbf{x}) = \mathbb{E}f(\mathbf{x}, \mathbf{Y}), \qquad f_2 : \mathbb{R}^n \to \mathbb{R}, \; f_2(\mathbf{y}) = \mathbb{E}f(\mathbf{X}, \mathbf{y})$$

are measurable, $\mathbb{E}|f_1(\mathbf{X})| < \infty$ and $\mathbb{E}|f_2(\mathbf{Y})| < \infty$ and

$$\mathbb{E}f_1(\mathbf{X}) = \mathbb{E}f_2(\mathbf{Y}) = \mathbb{E}f(\mathbf{X}, \mathbf{Y}) \; . \tag{7.15}$$

The random variable $f_1(\mathbf{X})$ is also called conditional expectation or expectation conditional on $\mathbf{X}$ and will sometimes be denoted by $\mathbb{E}_Y f(\mathbf{X}, \mathbf{Y})$.

A random vector $\mathbf{g} \in \mathbb{R}^n$ is called a standard Gaussian vector if its components are independent standard normal distributed random variables. More generally, a random vector $\mathbf{X} \in \mathbb{R}^n$ is said to be a Gaussian vector or multivariate normal distributed if there exists a matrix $\mathbf{A} \in \mathbb{R}^{n\times k}$ such that $\mathbf{X} = \mathbf{A}\mathbf{g} + \boldsymbol{\mu}$, where $\mathbf{g} \in \mathbb{R}^k$ is a standard Gaussian vector and $\boldsymbol{\mu} \in \mathbb{R}^n$ is the mean of $\mathbf{X}$. The matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^*$ is then the covariance matrix of $\mathbf{X}$, i.e., $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top$. If $\boldsymbol{\Sigma}$ is non-degenerate, i.e., $\boldsymbol{\Sigma}$ is positive definite, then $\mathbf{X}$ has a joint probability density function of the form

$$\psi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\langle \mathbf{x} - \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\rangle\right), \quad \mathbf{x} \in \mathbb{R}^n .$$

In the degenerate case when $\boldsymbol{\Sigma}$ is not invertible $\mathbf{X}$ does not have a density. It is easily deduced from the density that a rotated standard Gaussian $\mathbf{U}\mathbf{g}$, where $\mathbf{U}$ is an orthogonal matrix, has the same distribution as $\mathbf{g}$ itself.

If $X_1, \ldots, X_n$ are independent and normal distributed random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ then $\mathbf{X} = [X_1, \ldots, X_n]^\top$ has a multivariate normal distribution and its sum $Z = \sum_{\ell=1}^n X_\ell$ has the univariate normal distribution with mean $\mu = \sum_{\ell=1}^n \mu_\ell$ and variance $\sigma^2 = \sum_{\ell=1}^n \sigma_\ell^2$, as can be calculated from (7.14).

The next statement is concerned with another important distribution derived from the normal distribution.

Lemma 7.8. *Let $\mathbf{g} = [g_1, \ldots, g_n]^\top$ be a standard Gaussian vector. Then the random variable*

$$Z = \sum_{\ell=1}^n g_\ell^2$$

has the $\chi^2(n)$-distribution whose probability density function ϕ_n is given by

$$\phi_n(u) = \frac{1}{2^{n/2}\Gamma(n/2)} u^{(n/2)-1} \exp(-u/2) I_{(0,\infty)}(u) , \quad \text{for all } u \in \mathbb{R} , \quad (7.16)$$

where Γ is the Gamma-function, see Appendix C.3.

Proof. We proceed by induction on n. The distribution function of a scalar squared standard Gaussian g^2 is given by $\mathbb{P}(g^2 \le u) = 0$ for $u < 0$ and $\mathbb{P}(g^2 \le u) = \mathbb{P}(-\sqrt{u} \le g \le \sqrt{u}) = F(\sqrt{u}) - F(-\sqrt{u})$ for $u \ge 0$, where F is the distribution function of g. If ψ denotes the corresponding probability density function, it follows that the probability density ϕ_1 of the random variable g^2 is given for $u < 0$ by $\phi_1 = 0$ and, for $u \ge 0$ by

$$\begin{aligned}\phi_1(u) &= \frac{d}{du}\left(F(\sqrt{u}) - F(-\sqrt{u})\right) = \frac{1}{2}u^{-1/2}\psi(\sqrt{u}) + \frac{1}{2}u^{-1/2}\psi(-\sqrt{u}) \\ &= \frac{1}{\sqrt{2\pi}} u^{-1/2} e^{-u/2} .\end{aligned}$$

Hence, for $n = 1$, (7.16) is established since $\Gamma(1/2) = \sqrt{\pi}$.

Now assume that the formula 7.16 has already been established for $n \geq 1$. For $u \leq 0$ we have $\phi_{n+1}(u) = 0$, and for $u > 0$, since by (7.14), the probability density function of the sum of independent random variables is the convolution of their probability density functions, we have

$$\begin{aligned}
\phi_{n+1}(u) &= \phi_n * \phi_1(u) = \int_{-\infty}^{\infty} \phi_n(t)\phi_1(u-t)dt \\
&= \frac{1}{2^{n/2+1/2}\Gamma(n/2)\Gamma(1/2)} \int_0^{\infty} t^{(n/2)-1} e^{-t/2} (u-t)^{-1/2} e^{-(u-t)/2} I_{(0,\infty)}(u-t)dt \\
&= \frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)} e^{-u/2} \int_0^{u} t^{(n/2)-1} (u-t)^{-1/2} dt \\
&= \frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)} e^{-u/2} u^{(n/2)-1/2} \int_0^{1} t^{(n/2)-1} (1-t)^{-1/2} dt \\
&= \frac{1}{2^{(n+1)/2}\Gamma(1/2)\Gamma(n/2)} e^{-u/2} u^{(n+1)/2-1} B(n/2, 1/2) \\
&= \frac{1}{2^{(n+1)/2}\Gamma((n+1)/2)} u^{(n+1)/2-1} e^{-u/2} ,
\end{aligned}$$

where we used that the Beta function B satisfies

$$B(x,y) := \int_0^1 u^{x-1}(1-u)^{y-1}du = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \qquad x, y > 0 , \tag{7.17}$$

see Exercise 7.1. Thus we proved the formula (7.16) for $n+1$. This completes the proof by induction. □

Jensen's inequality reads as follows.

Theorem 7.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and let $\mathbf{X} \in \mathbb{R}^n$ be a random vector. Then*

$$f(\mathbb{E}\mathbf{X}) \leq \mathbb{E}f(\mathbf{X}) .$$

Proof. Let $\mathbf{v}$ be an element of the subdifferential $\partial f(\mathbb{E}\mathbf{X})$, see Definition B.20. (Note that the subdifferential of a convex function is always non-empty at every point.) By definition of ∂f we have, for any realization of $\mathbf{X}$,

$$f(\mathbb{E}\mathbf{X}) \leq f(\mathbf{X}) + \langle \mathbf{v}, \mathbb{E}\mathbf{X} - \mathbf{X} \rangle .$$

Taking expectations on both sides of this inequality gives the statement by noting that $\mathbb{E}[\mathbb{E}\mathbf{X} - \mathbf{X}] = 0$ □

Note that $-f$ is convex if f is concave, so that that for concave functions f, Jensen's inequality reads

$$\mathbb{E}f(\mathbf{X}) \leq f(\mathbb{E}\mathbf{X}) . \tag{7.18}$$

Finally, we state the Borel-Cantelli lemma.

Lemma 7.10. *Let $A_1, A_2, \ldots \in \Sigma$ be events and let*

$$A^* = \limsup_{n\to\infty} A_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m.$$

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A^*) = 0$.

Proof. Since $A^* \subset \bigcup_{m=n}^{\infty} A_m$ for all n, it holds $\mathbb{P}(A^*) \leq \sum_{m=n}^{\infty} \mathbb{P}(A_m) \to 0$ as $n \to \infty$ whenever $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$.

This concludes our outline of basic facts of probability theory.

7.2 Moments and Tails

Moment and tail estimates of random variables are intimately related. We start with a simple statement in this direction.

Proposition 7.11. *Suppose that Z is a random variable satisfying*

$$(\mathbb{E}|Z|^p)^{1/p} \leq \alpha\beta^{1/p} p^{1/\gamma} \quad \text{for all } p_0 \leq p \leq p_1 \tag{7.19}$$

for some constants $\alpha, \beta, \gamma, p_1 > p_0 > 0$. Then

$$\mathbb{P}(|Z| \geq e^{1/\gamma}\alpha u) \leq \beta e^{-u^\gamma/\gamma}$$

for all $u \in [p_0^{1/\gamma}, p_1^{1/\gamma}]$.

Proof. By Markov's inequality, Theorem 7.3, we obtain for an arbitrary $\kappa > 0$

$$\mathbb{P}(|Z| \geq e^{\kappa}\alpha u) \leq \frac{\mathbb{E}|Z|^p}{(e^{\kappa}\alpha u)^p} \leq \beta \left(\frac{\alpha p^{1/\gamma}}{e^{\kappa}\alpha u}\right)^p.$$

Choosing $p = u^\gamma$ yields $\mathbb{P}(|Z| \geq e^{\kappa}\alpha u) \leq \beta e^{-\kappa u^\gamma}$ and further setting $\kappa = 1/\gamma$ yields the claim. □

Remark 7.12. Important special cases of Proposition 7.11 are $\gamma = 1, 2$. Indeed, if $(\mathbb{E}|Z|^p)^{1/p} \leq \alpha\beta^{1/p}\sqrt{p}$ for all $p \geq 2$ then

$$\mathbb{P}(|Z| \geq e^{1/2}\alpha u) \leq \beta e^{-u^2/2} \quad \text{for all } u \geq \sqrt{2}\,; \tag{7.20}$$

while if $(\mathbb{E}|Z|^p)^{1/p} \leq \alpha\beta^{1/p} p$ for all $p \geq 2$ then

$$\mathbb{P}(|Z| \geq e\alpha u) \leq \beta e^{-u} \quad \text{for all } u \geq 2\,. \tag{7.21}$$

If one replaces β by $\beta' = \max\{\beta, e^{2/\gamma}\}$, $\gamma = 1, 2$, on the right hand sides of (7.21) and (7.20), then the inequalities hold for all $u \geq 0$, since for $u < \sqrt{2}$ they become trivial, that is, the right hand sides becomes larger than 1.

Also a converse to Proposition 7.11 holds, which involves the Gamma function Γ, see Appendix C.3.

Proposition 7.13. *Suppose that a random variable Z satisfies, for some $\gamma > 0$,*

$$\mathbb{P}(|Z| \geq e^{1/\gamma}\alpha u) \leq \beta e^{-u^\gamma/\gamma}, \quad \textit{for all } u > 0 \,.$$

Then, for $p > 0$,

$$\mathbb{E}|Z|^p \leq \beta\alpha^p(e\gamma)^{p/\gamma}\Gamma\left(\frac{p}{\gamma}+1\right) \,. \tag{7.22}$$

As a consequence, for $p \geq 1$,

$$(\mathbb{E}|Z|^p)^{1/p} \leq C_1\alpha(C_{2,\gamma}\beta)^{1/p}p^{1/\gamma} \quad \textit{for all } p \geq 1 \,, \tag{7.23}$$

where $C_1 = e^{1/(2e)} \approx 1.2019$ and $C_{2,\gamma} = \sqrt{\frac{2\pi}{\gamma}}e^{\gamma/12}$. In particular, one has $C_{2,1} \approx 2.7245$, $C_{2,2} \approx 2.0939$.

Proof. Using Proposition 7.1 and two changes of variables we obtain

$$\begin{aligned}
\mathbb{E}|Z|^p &= p\int_0^\infty \mathbb{P}(|Z| > t)t^{p-1}dt = p\alpha^p e^{p/\gamma}\int_0^\infty \mathbb{P}(|Z| \geq e^{1/\gamma}\alpha u)u^{p-1}du \\
&\leq p\alpha^p e^{p/\gamma}\int_0^\infty \beta e^{-u^\gamma/\gamma}u^{p-1}du = p\beta\alpha^p e^{p/\gamma}\int_0^\infty e^{-v}(\gamma v)^{p/\gamma-1}dv \\
&= \beta\alpha^p(e\gamma)^{p/\gamma}\frac{p}{\gamma}\Gamma\left(\frac{p}{\gamma}\right) \,.
\end{aligned} \tag{7.24}$$

This shows (7.22) taking into account the functional equation for the Gamma function. Applying Stirling's formula (C.12) yields

$$\begin{aligned}
\mathbb{E}|Z|^p &\leq \beta\alpha^p(e\gamma)^{p/\gamma}\sqrt{2\pi}\left(\frac{p}{\gamma}\right)^{p/\gamma+1/2}e^{-p/\gamma}e^{\gamma/(12p)} \\
&= \sqrt{2\pi}\beta\alpha^p e^{\gamma/(12p)}p^{p/\gamma+1/2}\gamma^{-1/2} \,.
\end{aligned}$$

Using the assumption $p \geq 1$ we obtain

$$(\mathbb{E}|Z|^p)^{1/p} \leq \left(\frac{\sqrt{2\pi}e^{\gamma/12}}{\sqrt{\gamma}}\beta\right)^{1/p}\alpha p^{1/\gamma}p^{1/(2p)} \,.$$

Finally, $p^{1/(2p)}$ takes its maximum value for $p = e$, i.e., $p^{1/(2p)} \leq e^{1/(2e)}$. This yields the statement of the proposition. □

Next we consider the expectation $\mathbb{E}|Z|$ of a random variable Z satisfying a subgaussian tail estimate (see (7.32) below), and improve on the general estimate (7.23) for $p = 1$.

Proposition 7.14. *Let Z be a random variable satisfying*

$$\mathbb{P}(|Z| \geq \alpha u) \leq \beta e^{-u^2/2} \quad \text{for all } u \geq \sqrt{2\ln(\beta)}\,,$$

for some constants $\alpha > 0, \beta \geq 2$. Then

$$\mathbb{E}|Z| \leq C_\beta \alpha \sqrt{\ln(4\beta)}$$

with $C_\beta = \sqrt{2} + \frac{1}{4\sqrt{2}\ln(4\beta)} \leq \sqrt{2} + \frac{1}{4\sqrt{2}\ln(8)} \approx 1.499 < 3/2$.

Proof. Let $\kappa \geq \sqrt{2\ln(\beta)}$ be some number to be chosen later. By Proposition 7.11 the expectation can be expressed as

$$\begin{aligned}\mathbb{E}|Z| &= \int_0^\infty \mathbb{P}(|Z| \geq u)du = \alpha \int_0^\infty \mathbb{P}(|Z| \geq \alpha u)du \\ &\leq \alpha \left(\int_0^\kappa 1 du + \beta \int_\kappa^\infty e^{-u^2/2} du \right) \leq \alpha \left(\kappa + \frac{\beta}{\kappa} e^{-\kappa^2/2} \right)\,.\end{aligned}$$

In the second line we used that any probability is bounded by 1 and in the last step we applied Lemma C.8. Choosing $\kappa = \sqrt{2\ln(4\beta)}$ completes the proof. □

Let us also provide a slight variation on Proposition 7.11.

Proposition 7.15. *Suppose Z is a random variable satisfying*

$$(\mathbb{E}|Z|^p)^{1/p} \leq \beta^{1/p}(\alpha_1 p + \alpha_2\sqrt{p} + \alpha_3) \quad \text{for all } p \geq p_0\,.$$

Then, for $u \geq p_0$,

$$\mathbb{P}\big(|Z| \geq e(\alpha_1 u + \alpha_2\sqrt{u} + \alpha_3)\big) \leq \beta e^{-u}\,.$$

Proof. The proof is basically the same as the one of Proposition 7.11 and left as Exercise 7.15. □

Tail probabilities can also be bounded from *below* using moments. We start with the classical *Paley–Zygmund* inequality.

Lemma 7.16. *If a nonnegative random variable Z has finite second moment then*

$$\mathbb{P}(Z > t) \geq \frac{(\mathbb{E}Z - t)^2}{\mathbb{E}Z^2}, \qquad 0 \leq t \leq \mathbb{E}Z.$$

Proof. For $t \geq 0$, the Cauchy-Schwarz inequality yields

$$\begin{aligned}\mathbb{E}Z &= \mathbb{E}[ZI_{\{Z>t\}}] + \mathbb{E}[ZI_{\{Z\leq t\}}] \\ &\leq (\mathbb{E}Z^2)^{1/2}\,\mathbb{E}(I_{\{Z>t\}})^{1/2} + t = (\mathbb{E}Z^2)^{1/2}\,\mathbb{P}(Z>t)^{1/2} + t\,.\end{aligned}$$

With $t \leq \mathbb{E}Z$, this is a rearrangement of the claim. □

Lemma 7.17. *If $X_1, \dots, X_n$ are independent mean zero random variables with variance σ^2 and fourth moment bounded from above by μ^4, then, for all $\mathbf{a} \in \mathbb{R}^n$,*

$$\mathbb{P}\Big(\Big|\sum_{\ell=1}^{n} a_\ell X_\ell\Big| > t\|\mathbf{a}\|_2\Big) \geq \frac{(\sigma^2 - t^2)^2}{\mu^4}, \qquad 0 \leq t \leq \sigma \, .$$

Proof. Setting $Z := \big(\sum_{\ell=1}^{n} a_\ell X_\ell\big)^2$, independence and the mean zero assumption yield

$$\begin{aligned} \mathbb{E}Z &= \mathbb{E}\Big(\sum_{j=1}^{n} a_\ell X_\ell\Big)^2 = \sum_{\ell=1}^{n} a_\ell^2 \, \mathbb{E}X_\ell^2 = \|\mathbf{a}\|_2^2 \, \sigma^2, \qquad (7.25) \\ \mathbb{E}Z^2 &= \mathbb{E}\Big(\sum_{\ell=1}^{n} a_\ell X_\ell\Big)^4 = \sum_{i,j,k,\ell \in [n]} a_i a_j a_k a_\ell \mathbb{E}(X_i X_j X_k X_\ell) \\ &= \sum_{i,j \in [n]} a_i^2 a_j^2 \mathbb{E}(X_i^2 X_j^2) \, , \end{aligned}$$

because if a random variable X_i is not repeated in the product $X_i X_j X_k X_\ell$, then the independence of X_i, X_j, X_k, and X_ℓ yields $\mathbb{E}(X_i X_j X_k X_\ell) = \mathbb{E}(X_i)\mathbb{E}(X_j X_k X_\ell) = 0$. Moreover, using the Cauchy–Schwarz inequality, we have, for $i, j \in [n]$,

$$\mathbb{E}(X_i^2 X_j^2) \leq \mathbb{E}(X_i^4)^{1/2} \mathbb{E}(X_j^4)^{1/2} \leq \mu^4 \, .$$

We deduce that

$$\mathbb{E}Z^2 \leq \sum_{i,j \in [n]} a_i^2 a_j^2 \mu^4 = \|\mathbf{a}\|_2^4 \, \mu^4 \, . \qquad (7.26)$$

Substituting (7.25) and (7.26) into Lemma 7.16, we obtain, for $0 \leq t \leq \sigma$,

$$\mathbb{P}\Big(\Big|\sum_{\ell=1}^{n} a_\ell X_\ell\Big| > t\|\mathbf{a}\|_2\Big) = \mathbb{P}\big(Z > t^2 \|\mathbf{a}\|_2^2\big) \geq \frac{\big(\sigma^2 - t^2\big)^2}{\mu^4} \, ,$$

which is the desired result. □

7.3 Cramér's Theorem and Hoeffding's Inequality

We often encounter sums of independent mean zero random variables. Deviation inequalities bound the tail of such sums.

We recall that the *moment generating function* of a (real-valued) random variable X is defined by

$$\theta \mapsto \mathbb{E} \exp(\theta X) \, ,$$

for all $\theta \in \mathbb{R}$ whenever the expectation on the right hand side is well-defined. Its logarithm is the *cumulant generating function*

$$C_X(\theta) = \ln \mathbb{E} \exp(\theta X) \, .$$

With the help of these definitions we can formulate Cramér's theorem.

Theorem 7.18. *Let $X_1, \ldots, X_M$ be a sequence of independent (real-valued) random variables, with cumulant generating functions C_{X_ℓ}, $\ell \in [M]$. Then, for $t > 0$,*

$$\mathbb{P}\big(\sum_{\ell=1}^{M} X_\ell \geq t\big) \leq \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta)\right\}\right) \, .$$

Proof. For $\theta > 0$, Markov's inequality (Theorem 7.3) and independence yield

$$\begin{aligned}
\mathbb{P}\big(\sum_{\ell=1}^{M} X_\ell \geq t\big) &= \mathbb{P}\big(\exp(\theta \sum_{\ell=1}^{M} X_\ell) \geq \exp(\theta t)\big) \leq e^{-\theta t}\mathbb{E}[\exp(\theta \sum_{\ell=1}^{M} X_j)] \\
&= e^{-\theta t}\mathbb{E}[\prod_{\ell=1}^{M} \exp(\theta X_j)] = e^{-\theta t} \prod_{\ell=1}^{M} \mathbb{E}[\exp(\theta X_j)] \\
&= e^{-\theta t} \prod_{\ell=1}^{M} \exp(C_{X_\ell}(\theta)) = \exp\left(-\theta t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta)\right) \, .
\end{aligned}$$

Taking the infimum over $\theta > 0$ concludes the proof. □

Remark 7.19. The function

$$t \mapsto \inf_{\theta>0}\left\{-\theta t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta)\right\}$$

appearing in the exponential is closely connected to a convex conjugate function appearing in convex analysis, see Section B.3.

We will use this theorem several times later on. Let us state Hoeffdings' inequality for the sum of almost surely bounded random variables as a first consequence.

Theorem 7.20. *Let $X_1, \ldots, X_M$ be a sequence of independent random variables such that $\mathbb{E}X_\ell = 0$ and $|X_\ell| \leq B_\ell$ almost surely, $\ell \in [M]$. Then*

$$\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \leq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{\ell=1}^{M} B_\ell^2}\right) \, ,$$

and consequently,

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \leq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{\ell=1}^{M} B_\ell^2}\right) \, . \tag{7.27}$$

Proof. Cramér's theorem suggests to estimate the moment generating function of X_ℓ. Since (except possibly for an event of measure zero) $X_\ell \in [-B_\ell, B_\ell]$, we can write

$$X_\ell = t(-B_\ell) + (1-t)B_\ell \ ,$$

where $t = \frac{B_\ell - X_\ell}{2B_\ell} \in [0,1]$. Since $f(x) = \exp(\theta x)$ is convex we have

$$\begin{aligned} \exp(\theta X_\ell) = f(X_\ell) = f(t(-B_\ell)) + (1-t)B_\ell) &\le tf(-B_\ell) + (1-t)f(B_\ell) \\ &= \frac{B_\ell - X_\ell}{2B_\ell} e^{-\theta B_\ell} + \frac{B_\ell + X_\ell}{2B_\ell} e^{\theta B_\ell} \ . \end{aligned} \tag{7.28}$$

Taking expectation and using that $\mathbb{E}X_\ell = 0$ we arrive at

$$\begin{aligned} \mathbb{E}\exp(\theta X_\ell) &\le \frac{1}{2}(\exp(-\theta B_\ell) + \exp(\theta B_\ell)) = \frac{1}{2}\left(\sum_{k=0}^{\infty} \frac{(-\theta B_\ell)^k}{k!} + \sum_{k=0}^{\infty} \frac{(\theta B_\ell)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \frac{(\theta B_\ell)^{2k}}{(2k)!} \le \sum_{k=0}^{\infty} \frac{(\theta B_\ell)^{2k}}{2^k k!} = \exp(\theta^2 B_\ell^2/2) \ . \end{aligned} \tag{7.29}$$

Therefore, the cumulant generating function of X_ℓ satisfies

$$C_{X_\ell}(\theta) \le B_\ell^2 \theta^2/2 \ .$$

It follows from Cramér's theorem 7.18 that

$$\begin{aligned} \mathbb{P}(\sum_{\ell=1}^{M} X_\ell \ge t) &\le \exp\left(\inf_{\theta>0}\{-\theta t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta)\}\right) \\ &\le \exp\left(\inf_{\theta>0}\left\{-\theta t + \frac{\theta^2}{2}\sum_{\ell=1}^{M} B_\ell^2\right\}\right) \ . \end{aligned}$$

The optimal choice $\theta = t/(\sum_{\ell=1}^{M} B_\ell^2)$ in the above infimum yields

$$\mathbb{P}(\sum_{\ell=1}^{M} X_\ell \ge t) \le \exp\left(-\frac{t^2}{2\sum_{\ell=1}^{M} B_\ell^2}\right) \ .$$

Replacing X_ℓ by $-X_\ell$ gives the same bound, and an application of the union bound (7.1) then shows (7.27). □

A Rademacher variable (sometimes also called Bernoulli variable) is a random variable ϵ that takes the values $+1$ and -1 with equal probability. A Rademacher sequence $\boldsymbol{\epsilon}$ is a vector of independent Rademacher variables. We obtain the following version of Hoeffding's inequality for Rademacher sums.

Corollary 7.21. *Let* $\mathbf{a} \in \mathbb{R}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ *be a Rademacher sequence. Then, for* $u > 0$,

$$\mathbb{P}\left(|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell| \ge \|\mathbf{a}\|_2 u\right) \le 2\exp(-u^2/2) \ . \tag{7.30}$$

Proof. The random variable $a_\ell \epsilon_\ell$ has mean zero and is bounded in absolute value by $|a_\ell|$. Therefore, the stated inequality follows immediately from Hoeffding's inequality 7.20. □

Remark 7.22. Note that specializing (7.29) to a Rademacher variable ϵ shows that its moment generating function satisfies

$$\mathbb{E}\exp(\theta\epsilon) \le \exp(\theta^2/2) \; . \tag{7.31}$$

7.4 Subgaussian Random Variables

A random variable X is called *subgaussian* if there exist constants $\beta, \kappa > 0$ such that

$$\mathbb{P}(|X| \ge t) \le \beta e^{-\kappa t^2} \quad \text{for all } t > 0 \; . \tag{7.32}$$

It is called *subexponential* if

$$\mathbb{P}(|X| \ge t) \le \beta e^{-\kappa t} \quad \text{for all } t > 0 \; .$$

According to Proposition 7.5 a standard Gaussian random variable is subgaussian with $\beta = 1$ and $\kappa = 1/2$. Furthermore, Bernoulli and bounded random variables are subgaussian. According to Theorem 7.20, Rademacher sums are subgaussian random variables as well.

Clearly, a random variable X is subgaussian if and only if X^2 is subexponential. Setting $\alpha = (2e\kappa)^{-1/2}$ and $\gamma = 2$ in Proposition 7.13 shows that the moments of a subgaussian variable X satisfy

$$(\mathbb{E}|X|^p)^{1/p} \le \widetilde{C}\kappa^{-1/2}\beta^{1/p}p^{1/2} \quad \text{for all } p \ge 1 \tag{7.33}$$

with $\widetilde{C} = e^{1/(2e)}C_{2,2}/\sqrt{2e} = e^{1/(2e)+1/6}\sqrt{\pi/(2e)} \approx 1.0282$, while the moments of a subexponential variable X satisfy (setting $\alpha = (e\kappa)^{-1}$ and $\gamma = 1$ in Proposition 7.13)

$$(\mathbb{E}|X|^p)^{1/p} \le \widehat{C}\kappa^{-1}\beta^{1/p}p \quad \text{for all } p \ge 1$$

with $\widehat{C} = e^{1/(2e)}C_{2,1}e^{-1} = e^{1/(2e)+1/12}\sqrt{2\pi} \approx 3.1193$. Proposition 7.11 provides a statement in the converse direction. Let us give an equivalent characterization of subgaussian random variables.

Proposition 7.23. *Let X be a random variable.*

(a) If X is subgaussian, then there exist constants $c > 0, C > 1$ such that $\mathbb{E}[\exp(cX^2)] \le C$.
(b) If $\mathbb{E}[\exp(cX^2)] \le C$ for some constants $c, C > 0$ then X is subgaussian. More precisely, we have $\mathbb{P}(|X| \ge t) \le Ce^{-ct^2}$.

Proof. (a) The moment estimate (7.22) with $\kappa = 1/(2e\alpha^2)$ yields

$$\mathbb{E}X^{2n} \le \beta\kappa^{-n} n! \, .$$

Expanding the exponential function into its Taylor series and using Fubini's theorem shows that

$$\mathbb{E}[\exp(cX^2)] = 1 + \sum_{n=1}^{\infty} \frac{c^n \mathbb{E}[X^{2n}]}{n!} \le 1 + \beta \sum_{n=1}^{\infty} \frac{c^n \kappa^{-n} n!}{n!} = 1 + \frac{\beta c \kappa^{-1}}{1 - c\kappa^{-1}} \, .$$

provided $c < \kappa$.

(b) This statement follows from Markov's inequality, Theorem 7.3,

$$\mathbb{P}(|X| \ge t) = \mathbb{P}(\exp(cX^2) \ge \exp(ct^2)) \le \mathbb{E}[\exp(cX^2)] e^{-ct^2} \le C e^{-ct^2} \, .$$

This completes the proof. □

Exercise 7.6 refines the statement of Proposition 7.23(a).

Let us study the Laplace transform (or moment generating function) of a mean zero subgaussian random variable.

Proposition 7.24. *Let X be a random variable.*

(a) If X is subgaussian with $\mathbb{E}X = 0$ then there exists a constant c (depending only on β and κ) such that

$$\mathbb{E}[\exp(\theta X)] \le \exp(c\theta^2) \quad \text{for all } \theta \in \mathbb{R} \, . \tag{7.34}$$

(b) Conversely, if (7.34) holds then $\mathbb{E}X = 0$ and X is subgaussian with parameters $\beta = 2$ and $\kappa = \frac{1}{4c}$.

Remark 7.25. Any valid constant c in (7.34) is called *subgaussian parameter* of X. Of course, one preferably chooses the minimal possible c.

Proof. For the easier part (b) we take $\theta, t > 0$ and apply Markov's inequality, Theorem 7.3, to get

$$\mathbb{P}(X \ge t) = \mathbb{P}(\exp(\theta X) \ge \exp(\theta t)) \le \mathbb{E}[\exp(\theta X)] e^{-\theta t} \le e^{c\theta^2 - \theta t} \, .$$

The optimal choice $\theta = t/(2c)$ yields

$$\mathbb{P}(X \ge t) \le e^{-t^2/(4c)} \, .$$

Repeating the above computation with $-X$ instead of X shows that

$$\mathbb{P}(-X \ge t) \le e^{-t^2/(4c)} \, ,$$

and the union bound yields the desired estimate $\mathbb{P}(|X| \ge t) \le 2e^{-t^2/(4c)}$. In order to deduce that X has mean zero, we take the expectation in the inequality $1 + \theta X \le \exp(\theta X)$ to deduce, for $|\theta| < 1$,

$$1 + \theta\mathbb{E}(X) \le \mathbb{E}[\exp(\theta X)] \le \exp(c\theta^2) \le 1 + (c/2)\theta^2 + \mathcal{O}(\theta^4) \,.$$

Letting $\theta \to 0$ shows that $\mathbb{E}X = 0$.

Let us now turn to the converse implication (a). We note that it is enough to consider $\theta \ge 0$, as the statement for $\theta < 0$ follows from exchanging X with $-X$. Expanding the exponential function into its Taylor series yields (together with Fubini's theorem),

$$\mathbb{E}[\exp(\theta X)] = 1 + \theta\mathbb{E}(X) + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}X^n}{n!} = 1 + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}|X|^n}{n!} \,,$$

where we used the mean zero assumption. First suppose that $0 \le \theta \le \theta_0$ for some θ_0 to be determined below. Then the moment estimate (7.33) and Stirling's formula (C.13), $n! \ge \sqrt{2\pi} n^n e^{-n}$, yield

$$\begin{aligned}
\mathbb{E}[\exp(\theta X)] &\le 1 + \beta \sum_{n=2}^{\infty} \frac{\theta^n \widetilde{C}^n \kappa^{-n/2} n^{n/2}}{n!} \le 1 + \frac{\beta}{\sqrt{2\pi}} \sum_{n=2}^{\infty} \frac{\widetilde{C}^n \theta^n \kappa^{-n/2} n^{n/2}}{n^n e^{-n}} \\
&\le 1 + \theta^2 \frac{\beta(\widetilde{C}e)^2}{\sqrt{2\pi}\kappa} \sum_{n=0}^{\infty} (\widetilde{C}e\theta_0 \kappa^{-1/2})^n \\
&= 1 + \theta^2 \frac{\beta(\widetilde{C}e)^2}{\sqrt{2\pi}\kappa} \frac{1}{1 - \widetilde{C}e\theta_0\kappa^{-1/2}} \\
&= 1 + c_1\theta^2 \le \exp(c_1\theta^2) \,,
\end{aligned}$$

provided $\widetilde{C}e\theta_0\kappa^{-1/2} < 1$. The latter is satisfied by setting

$$\theta_0 = (2\widetilde{C}e)^{-1}\sqrt{\kappa} \,,$$

which gives $c_1 = \sqrt{2}\beta\kappa^{-1}((\widetilde{C}e)^2/\sqrt{\pi})$.

Let us now assume that $\theta > \theta_0$. We aim at proving $\mathbb{E}[\exp(\theta X - c_2\theta^2)] \le 1$. Observe that

$$\theta X - c_2\theta^2 = -\big(\sqrt{c_2}\theta - \frac{X}{2\sqrt{c_2}}\big)^2 + \frac{X^2}{4c_2} \le \frac{X^2}{4c_2} \,.$$

Let $\tilde{c} > 0$, $\tilde{C} \ge 1$ be the constants from Proposition 7.23(a), and choose $c_2 = 1/(4\tilde{c})$. Then

$$\mathbb{E}[\exp(\theta X - c_2\theta^2)] \le \mathbb{E}[\exp(\tilde{c}X^2)] \le \tilde{C} \,.$$

Defining $\rho = \ln(\tilde{C})\theta_0^{-2}$ yields

$$\begin{aligned}
\mathbb{E}[\exp(\theta X)] &\le \tilde{C}\exp(c_2\theta^2) = \tilde{C}\exp(-\rho\theta^2)\exp((c_2+\rho)\theta^2) \\
&\le \tilde{C}\exp(-\rho\theta_0^2)e^{(c_2+\rho)\theta^2} \le e^{(c_2+\rho)\theta^2} \,.
\end{aligned}$$

Setting $c = \max\{c_1, c_2 + \rho\}$ completes the proof. □

Remark 7.26. For Rademacher and standard Gaussian random variables, the constant in (7.34) satisfies $c = 1/2$ by (7.11) and (7.31). Furthermore, for mean zero random variables X with $|X| \le K$ almost surely, $c = K^2/2$ is a valid choice of the subgaussian parameter by (7.29).

The sum of independent mean zero subgaussian variables is again subgaussian by the next statement.

Theorem 7.27. *Let $X_1, \dots, X_M$ be a sequence of independent mean zero subgaussian random variables with subgaussian parameter c in* (7.34)*. Let $\mathbf{a} \in \mathbb{R}^M$ be some vector. Then $Z := \sum_{\ell=1}^M a_\ell X_\ell$ is subgaussian, that is,*

$$\mathbb{E} \exp(\theta Z) \le \exp(c\|\mathbf{a}\|_2^2 \theta^2) \,, \tag{7.35}$$

and

$$\mathbb{P}\left(\left| \sum_{\ell=1}^M a_\ell X_\ell \right| \ge t \right) \le 2 \exp\left(-\frac{t^2}{4c\|\mathbf{a}\|_2^2} \right) \quad \text{for all } t > 0 \,. \tag{7.36}$$

Proof. By independence we have

$$\begin{aligned} \mathbb{E} \exp(\theta \sum_{\ell=1}^M a_\ell X_\ell) = \mathbb{E} \prod_{\ell=1}^M \exp(\theta a_\ell X_\ell) = \prod_{\ell=1}^M \mathbb{E} \exp(\theta a_\ell X_\ell) \le \prod_{\ell=1}^M \exp(c\theta^2 a_\ell^2) \\ = \exp(c\|\mathbf{a}\|_2^2 \theta^2) \,. \end{aligned}$$

This proves (7.35). The second inequality (7.36) follows then from Proposition 7.24(b). □

Remark 7.28. In particular, if $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ is a Rademacher sequence, and $Z = \sum_{\ell=1}^M a_\ell \epsilon_\ell$ then

$$\mathbb{E} \exp(\theta Z) \le \exp(\theta^2 \|\mathbf{a}\|_2^2 / 2) \,.$$

The expected maximum of a finite number of subgaussian random variables can be estimated as follows.

Proposition 7.29. *Let $X_1, \dots, X_M$, be a sequence of (not necessarily independent) mean zero subgaussian random variables satisfying $\mathbb{E}[\exp(\theta X_\ell)] \le \exp(c_\ell \theta^2)$, $\ell \in [M]$. Then, with $c = \max_{\ell=1,\dots,M} c_\ell$,*

$$\mathbb{E} \max_{\ell \in [M]} X_\ell \le \sqrt{4c \, \ln(M)} \,, \tag{7.37}$$

$$\mathbb{E} \max_{\ell \in [M]} |X_\ell| \le \sqrt{4c \, \ln(2M)} \,. \tag{7.38}$$

Proof. Since (7.37) is obvious for $M = 1$, we assume $M \geq 2$. Let $\beta > 0$ be a number to be chosen later. Using concavity of the logarithm in connection with Jensen's inequality we obtain

$$\begin{aligned}\beta\mathbb{E}\max_{\ell\in[M]} X_\ell &= \mathbb{E}\ln\max_{\ell\in[M]}\exp(\beta X_\ell) \leq \mathbb{E}\ln\left(\sum_{\ell=1}^{M}\exp(\beta X_\ell)\right) \leq \ln(\sum_{\ell=1}^{M}\mathbb{E}\exp(\beta X_\ell)) \\ &\leq \ln(M\exp(c\beta^2)) = c\beta^2 + \ln(M)\ .\end{aligned}$$

Choosing $\beta = \sqrt{c^{-1}\ln(M)}$ yields

$$\sqrt{c^{-1}\ln(M)}\mathbb{E}\max_{\ell\in[M]} X_\ell \leq \ln(M) + \ln(M)$$

so that $\mathbb{E}\max_{\ell\in[M]} \leq \sqrt{4c\ln(M)}$.

For (7.38) we write $\mathbb{E}\max_{\ell\in[m]}|X_\ell| = \mathbb{E}\max\{X_1,\ldots,X_m,-X_1,\ldots,-X_m\}$ and apply (7.37). □

The example of a sequence of standard Gaussian random variables shows that the estimates in the previous Proposition are optimal up to possibly the constants, see Proposition 8.1(c) below.

7.5 Bernstein Inequalities

Bernstein's inequality provides a useful generalization of Hoeffding's inequality (7.30) to sums of bounded or even unbounded independent random variables, which also takes into account the variance or higher moments. We start with the version below, and then derive variations as consequences.

Theorem 7.30. *Let $X_1,\ldots,X_M$ be independent mean zero random variables such that, for all integers $n \geq 2$,*

$$\mathbb{E}|X_\ell|^n \leq n!R^{n-2}\sigma_\ell^2/2 \quad \textit{for all } \ell \in [M] \tag{7.39}$$

for some constants $R > 0$ and $\sigma_\ell > 0$, $\ell \in [M]$. Then, for all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Rt}\right)\ , \tag{7.40}$$

where $\sigma^2 := \sum_{\ell=1}^{M}\sigma_\ell^2$.

Before providing the proof we give two consequences. The first is the Bernstein inequality for bounded random variables.

Corollary 7.31. *Let $X_1, \dots, X_M$ be independent random variables with zero mean such that $|X_\ell| \le K$ almost surely, for $\ell \in [M]$ and some constant $K > 0$. Further assume $\mathbb{E}|X_\ell|^2 \le \sigma_\ell^2$ for constants $\sigma_\ell > 0$, $\ell \in [M]$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \ge t\right) \le 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right) , \tag{7.41}$$

where $\sigma^2 := \sum_{\ell=1}^{M} \sigma_\ell^2$.

Proof. For $n = 2$, condition (7.39) is clearly satisfied. So let $n \in \mathbb{N}$, $n \ge 3$. Since then $n! \ge 3 \cdot 2^{n-2}$, we obtain

$$\mathbb{E}|X_\ell|^n = \mathbb{E}[|X_\ell|^{n-2} X_\ell^2] \le K^{n-2}\sigma_\ell^2 \le \frac{n!K^{n-2}}{n!}\sigma_\ell^2 \le \frac{n!K^{n-2}}{2 \cdot 3^{n-2}}\sigma_\ell^2 . \tag{7.42}$$

In other words, condition (7.39) holds for all $n \ge 2$ with constants $R = K/3$ and σ_ℓ. Hence, the statement follows from Theorem 7.30. □

As a second consequence, we present the Bernstein inequality for subexponential random variables.

Corollary 7.32. *Let $X_1, \dots, X_M$ be independent mean zero subexponential random variables, that is, $\mathbb{P}(|X_\ell| \ge t) \le \beta e^{-\kappa t}$ for some constants $\beta, \kappa > 0$ for all $t > 0$, $\ell \in [M]$. Then*

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \ge t\right) \le 2\exp\left(-\frac{(\kappa t)^2/2}{2\beta M + \kappa t}\right) . \tag{7.43}$$

Proof. Similarly to the proof of Proposition 7.13 we estimate, for $n \in \mathbb{N}, n \ge 2$,

$$\begin{aligned}\mathbb{E}|X_\ell|^n &= n\int_0^\infty \mathbb{P}(|X_\ell| \ge t)t^{n-1}dt \le \beta n \int_0^\infty e^{-\kappa t}t^{n-1}dt \\ &= \beta n\kappa^{-n}\int_0^\infty e^{-u}u^{n-1}du = \beta n!\kappa^{-n} = n!\kappa^{-(n-2)}\frac{2\beta\kappa^{-2}}{2} .\end{aligned}$$

Hereby, we have used that the integral in the second line equals $\Gamma(n) = (n-1)!$. Hence, condition (7.39) holds with $R = \kappa^{-1}$ and $\sigma_\ell^2 = 2\beta\kappa^{-2}$. The claim follows therefore from Theorem 7.30. □

Let us now turn to the proof of the Bernstein inequality in Theorem 7.30.

Proof (of Theorem 7.30). Cramér's theorem suggests to estimate the moment generating function of the X_ℓ. Expanding the exponential function into its series expansion and using Fubini's theorem in order to interchange expectation and summation yields

$$\mathbb{E}[\exp(\theta X_\ell)] = 1 + \theta\mathbb{E}[X_\ell] + \sum_{n=2}^{\infty}\frac{\theta^n\mathbb{E}[X_\ell^n]}{n!} = 1 + \frac{\theta^2\sigma_\ell^2}{2}\sum_{n=2}^{\infty}\frac{\theta^{n-2}\mathbb{E}[X_\ell^n]}{n!\sigma_\ell^2/2} ,$$

where we additionally used that $\mathbb{E}[X_\ell] = 0$. Defining

$$F_\ell(\theta) = \sum_{n=2}^{\infty} \frac{\theta^{n-2}\mathbb{E}[X_\ell^n]}{n!\sigma_\ell^2/2}$$

we obtain

$$\mathbb{E}[\exp(\theta X_\ell)] = 1 + \theta^2\sigma_\ell^2 F_\ell(\theta)/2 \le \exp(\theta^2\sigma_\ell^2 F_\ell(\theta)/2) \; .$$

Introducing $F(\theta) = \max_{\ell\in[M]} F_\ell(\theta)$ and recalling that $\sigma^2 = \sum_{\ell=1}^M \sigma_\ell^2$ we obtain from Cramér's theorem

$$\mathbb{P}\big(\sum_{\ell=1}^M X_\ell \ge t\big) \le \inf_{\theta>0} \exp(\theta^2\sigma^2 F(\theta)/2 - \theta t) \le \inf_{0<R\theta<1} \exp(\theta^2\sigma^2 F(\theta)/2 - \theta t).$$

Since $\mathbb{E}[X_\ell^n] \le \mathbb{E}[|X_\ell|^n]$ the assumption (7.39) yields

$$F_\ell(\theta) \le \sum_{n=2}^{\infty} \frac{\theta^{n-2}\mathbb{E}[|X_\ell|^n]}{n!\sigma_\ell^2/2} \le \sum_{n=2}^{\infty} (R\theta)^{n-2} = \frac{1}{1-R\theta}$$

provided $R\theta < 1$. Therefore, $F(\theta) \le (1-R\theta)^{-1}$ and

$$\mathbb{P}\big(\sum_{\ell=1}^M X_\ell \ge t\big) \le \inf_{0<\theta R<1} \exp\left(\frac{\theta^2\sigma^2}{2(1-R\theta)} - \theta t\right) \; .$$

Now we choose $\theta = t/(\sigma^2 + Rt)$, which clearly satisfies $R\theta < 1$. This yields

$$\begin{aligned}\mathbb{P}\big(\sum_{\ell=1}^M X_\ell \ge t\big) &\le \exp\left(\frac{t^2\sigma^2}{2(\sigma^2+Rt)^2}\frac{1}{1-\frac{Rt}{\sigma^2+Rt}} - \frac{t^2}{\sigma^2+Rt}\right) \\ &= \exp\left(-\frac{t^2/2}{\sigma^2+Rt}\right) \; .\end{aligned}$$

Exchanging X_ℓ with $-X_\ell$ yields the same estimate, and applying the union bound completes the proof. □

Notes

Good sources for background on basic probability theory are for instance the monographs [209, 369]. The relation of tails and moments is well-known, see e.g. [280], although the refinement with the parameter β in (7.19) seems to have appeared only recently [355]. Cramér proved the theorem named after him in [107]. We refer to [432] for more information on large deviation results in this spirit. Hoeffding's inequality (7.30) was derived in [238]. In the special case

of random variables that take only values in $\{0,1\}$ with probabilities p and $1-p$, so-called Chernoff bounds refine the Hoeffding inequalities, see for instance [95, 220]. Bernstein's inequality was first proved in [34, 35], and refined later by Bennett [31]. For further reading on scalar deviation inequalities the reader is referred to [300, 432].

The notion of subgaussian random variables may be refined to *strictly* subgaussian random variables, for which the constant in (7.34) satisfies $c = \mathbb{E}|X|^2/2$. Gaussian and Bernoulli random variables, as well as random variables that are uniformly distributed on $[-1,1]$ are strictly subgaussian, see Exercise 7.5. More information on subgaussian random variables can be found, for instance, in [65, 438].

Exercises

7.1. Show the relation (7.17) of the Beta function B to the Gamma function.

7.2. Prove Proposition 7.15.

7.3. Let $p > 1$. Generalize Lemma 7.16 by showing that any nonnegative random variable Z with finite pth moment satisfies

$$\mathbb{P}(Z > t) \geq \frac{(\mathbb{E}Z - t)^{p/(p-1)}}{(\mathbb{E}Z^p)^{1/(p-1)}}, \qquad 0 \leq t \leq \mathbb{E}Z.$$

Prove also that if $X_1, \ldots, X_M$ are independent mean zero random variables with variance σ^2 and $2p$th absolute moment bounded above by μ^{2p}, then, for all $\mathbf{a} \in \mathbb{R}^M$,

$$\mathbb{P}\Big(\Big|\sum_{\ell=1}^{M} a_\ell X_\ell\Big| > t\|\mathbf{a}\|_2\Big) \geq c_p \frac{(\sigma^2 - t^2)^2}{\mu^{2p/(p-1)}}, \qquad 0 \leq t \leq \sigma,$$

for some constant c_p to be determined.

7.4. Let X be a subgaussian random variable with $\mathbb{E}\exp(\theta X) \leq \exp(c\theta^2)$ for some constant $c > 0$. Show that its variance satisfies $\mathbb{E}X^2 \leq 2c$. (A subgaussian variable for which equality holds, is called strictly subgaussian).

7.5. Let X be a random variable that is uniformly distributed on $[-1,1]$. Show that $\mathbb{E}|X|^2 = 1/3$ and that

$$\mathbb{E}\exp(\theta X) \leq \exp(\theta^2/6) = \exp(\theta^2 \mathbb{E}|X|^2/2)\ ,$$

so that X is strictly subgaussian.

7.6. Let X be a subgaussian random variable with parameter $c > 0$, that is, $\mathbb{E}\exp(\theta X) \leq \exp(c\theta^2)$ for all $\theta \in \mathbb{R}$. Show that, for $t \in [0, 1/2]$,

$$\mathbb{E}\exp(tX^2/c^2) \leq \frac{1}{\sqrt{1-2t}}\ .$$

8

Advanced Tools from Probability Theory

This chapter introduces further probabilistic tools that will be required for some of the more advanced results in the remainder of the book.

In Section 8.1, we compute the expectation of the ℓ_p-norm of a standard Gaussian vector for $p = 1, 2, \infty$ (required in Section 9.3). Section 8.2 presents simple results for Rademacher sums as well as the symmetrization technique, which randomizes a sum of random vectors by introducing additional (random) Rademacher vectors. This simple technique turns out to be powerful in various setups and will be needed in Section 12.5 and Chapter 13. Khintchine inequalities, treated in Section 8.3, estimate the moments of a Rademacher sum and allow to deduce Hoeffding type inequalities for Rademacher sums in a different way than via moment generating functions (required for Section 12.5 and Chapter 13). Decoupling inequalities to be introduced in Section 8.6 replace one sequence of random variables in a double sum by an independent copy (required for Section 9.4 and Chapter 13). The scalar Bernstein inequality for bounded random variables (Corollary 7.31) will be extended in Section 8.5 to a powerful deviation inequality for the operator norm of sums of random matrices (required for Sections 12.3, 12.4, and 13.1). Section 8.6 deals with Dudley's inequality, which is a crucial tool to estimate the expectation of a supremum of a subgaussian process by an integral over covering numbers of the index set of the process (required for the estimate of the restricted isometry constants in Section 12.5). The Slepian and Gordon lemmas compare expectations of functions of two Gaussian random vectors in terms of the covariances of the two vectors. In particular, maxima as well as minima of maxima are important choices of such functions. These will be treated in Section 8.7 and will be used in Sections 9.2 and 9.3. Section 8.8 treats the concentration of measure phenomenon which states that a Lipschitz function of a Gaussian random vector concentrates around its mean (required in Sections 9.2 and 9.3). The final section of this chapter deals with a deviation inequality for the supremum of an empirical process, which is sometimes called Talagrand's inequality. It will be required in Chapter 12.

8.1 Expectation of Standard Gaussians in Norm

We state simple results on the expectation of the norms of standard Gaussian random vectors in ℓ_1, ℓ_2 and ℓ_∞.

Proposition 8.1. *Let $\mathbf{g} = (g_1, \dots, g_n)$ be a vector of (not necessarily independent) standard Gaussian random variables. Then*

(a) $\mathbb{E}\|\mathbf{g}\|_1 = \sqrt{\frac{2}{\pi}}n$;

(b) $\mathbb{E}\|\mathbf{g}\|_2^2 = n$ *and* $\sqrt{\frac{2}{\pi}}\sqrt{n} \le \mathbb{E}\|\mathbf{g}\|_2 \le \sqrt{n}$.

If the entries of $\mathbf{g}$ *are independent then*

$$\frac{n}{\sqrt{n+1}} \le \mathbb{E}\|\mathbf{g}\|_2 = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \le \sqrt{n}\,, \tag{8.1}$$

and consequently $\mathbb{E}\|\mathbf{g}\|_2 \sim \sqrt{n}$ *as* $n \to \infty$.

(c) It holds

$$\mathbb{E}\max_{\ell\in[n]} g_\ell \le \sqrt{2\ln(n)}\,, \quad \textit{and} \quad \mathbb{E}\|\mathbf{g}\|_\infty \le \sqrt{2\ln(2n)}\,. \tag{8.2}$$

If the entries of $\mathbf{g}$ *are independent then, for* $n \ge 2$,

$$\mathbb{E}\|\mathbf{g}\|_\infty \ge C\sqrt{\ln(n)} \tag{8.3}$$

with $C \approx 0.265$.

Proof. (a) By the formula for the density of a standard Gaussian random variable, we have

$$\mathbb{E}|g_\ell| = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |u|\exp(-u^2/2)du = \sqrt{\frac{2}{\pi}}\int_0^\infty u\exp(-u^2/2)du = \sqrt{\frac{2}{\pi}}\,.$$

By linearity of expectation $\mathbb{E}\|\mathbf{g}\|_1 = \sum_{\ell=1}^n \mathbb{E}|g_\ell| = \sqrt{2/\pi}\,n$.

(b) Clearly, $\mathbb{E}\|\mathbf{g}\|_2^2 = \sum_{\ell=1}^n \mathbb{E}g_\ell^2 = n$ for standard Gaussian random variables g_ℓ. The Cauchy-Schwarz inequality for expectations (or Jensen's inequality) yields $\mathbb{E}\|\mathbf{g}\|_2 \le \sqrt{\mathbb{E}\|\mathbf{g}\|_2^2} = \sqrt{n}$, while the Cauchy-Schwarz inequality for the inner product on $\mathbb{R}^n$ gives $\mathbb{E}\|\mathbf{g}\|_2 \ge \mathbb{E}\frac{1}{\sqrt{n}}\|\mathbf{g}\|_1 = \sqrt{2/\pi}\,\sqrt{n}$.

If the entries of $\mathbf{g}$ are independent, then $\|\mathbf{g}\|_2^2$ has the $\chi^2(n)$-distribution with probability density function given by (7.16). Therefore,

$$\begin{aligned}
\mathbb{E}\|\mathbf{g}\|_2 = \mathbb{E}\left(\sum_{\ell=1}^n g_\ell^2\right)^{1/2} &= \int_0^\infty u^{1/2}\phi_n(u)du \\
&= \frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty u^{1/2}u^{(n/2)-1}e^{-u/2}du \\
&= \frac{2^{n/2+1/2}}{2^{n/2}\Gamma(n/2)}\int_0^\infty t^{(n/2)-1/2}e^{-t}dt = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\,,
\end{aligned}$$

where we used the definition of the Gamma function in (C.9). The estimate $E_n := \mathbb{E}\|\mathbf{g}\|_2 \le \sqrt{n}$ for Gaussian vector $\mathbf{g}$ of length n was already shown above. Furthermore,

$$E_{n+1}E_n = 2\frac{\Gamma(n/2+1)}{\Gamma(n/2)} = n\ ,$$

by the functional equation (C.11) for the Gamma function so that $E_n = n/E_{n+1} \ge n/\sqrt{n+1}$ (compare also Lemma C.4).

(c) The inequalities in (8.2) follow from Proposition 7.29 by noting that due to Lemma 7.6 $\mathbb{E}\exp(\beta g) = \exp(\beta^2/2)$ so that the subgaussian parameter $c = 1/2$ for Gaussian random variables.

If the g_ℓ are independent then by Corollary 7.2

$$\begin{aligned}\mathbb{E}\|\mathbf{g}\|_\infty &= \int_0^\infty \mathbb{P}\left(\max_{\ell\in[n]}|g_\ell| > u\right)du = \int_0^\infty \left(1-\mathbb{P}(\max_{\ell\in[n]}|g_\ell| \le u)\right)du\\ &= \int_0^\infty \left(1-\prod_{\ell=1}^n \mathbb{P}(|g_\ell|\le u)\right)du \ge \int_0^\delta \left(1-(1-\mathbb{P}(|g|>u))^n\right)du\\ &\ge \delta\left(1-(1-\mathbb{P}(|g|>\delta))^n\right)\ .\end{aligned}$$

Further,

$$\mathbb{P}(|g|>\delta) = \sqrt{\frac{2}{\pi}}\int_\delta^\infty e^{-t^2/2}dt \ge \sqrt{\frac{2}{\pi}}\int_\delta^{2\delta} e^{-t^2/2}dt \ge \sqrt{\frac{2}{\pi}}\delta e^{-2\delta^2}\ .$$

Now, we choose $\delta = \sqrt{\ln n/2}$. Then, for $n \ge 2$,

$$\begin{aligned}\mathbb{E}\|\mathbf{g}\|_\infty &\ge \sqrt{\frac{\ln n}{2}}\left(1-\left(1-\sqrt{\frac{\ln n}{\pi}}\frac{1}{n}\right)^n\right) \ge \sqrt{\frac{\ln n}{2}}\left(1-\exp\left(-\sqrt{\frac{\ln n}{\pi}}\right)\right)\\ &\ge \frac{1-\exp(-\sqrt{(\ln 2)/\pi})}{\sqrt{2}}\sqrt{\ln n}\ ,\end{aligned}$$

which establishes the claim with $C = (1-\exp(-\sqrt{(\ln 2)/\pi}))/\sqrt{2} \approx 0.265$. □

Next we extend part (c) of the previous proposition to the maximum squared ℓ_2-norm of a sequence of standard Gaussian random vectors.

Proposition 8.2. *Let $\mathbf{g}_1,\dots,\mathbf{g}_M \in \mathbb{R}^n$ be a sequence of (not necessarily independent) standard Gaussian random vectors. Then, for any $\kappa > 0$,*

$$\mathbb{E}\max_{\ell\in[M]}\|\mathbf{g}_\ell\|_2^2 \le (2+2\kappa)\ln(M) + n(1+\kappa)\ln(1+\kappa^{-1})\ .$$

Consequently,

$$\mathbb{E}\max_{\ell\in[M]}\|\mathbf{g}_\ell\|_2^2 \le (\sqrt{2\ln(M)}+\sqrt{n})^2\ .$$

Proof. By concavity of the logarithm and Jensen's inequality we have, for $\theta > 0$,

$$\begin{aligned}\mathbb{E} \max_{\ell\in[M]} \|\mathbf{g}_\ell\|_2^2 &= \theta^{-1}\mathbb{E} \ln \max_{\ell\in[M]} \exp\left(\theta\|\mathbf{g}_\ell\|_2^2\right) \le \theta^{-1} \ln \mathbb{E} \max_{\ell\in[M]} \exp\left(\theta\|\mathbf{g}_\ell\|_2^2\right) \\ &\le \theta^{-1} \ln\left(M\mathbb{E}\exp(\theta\|\mathbf{g}\|_2^2)\right) ,\end{aligned}$$

where $\mathbf{g}$ denotes a standard Gaussian random vector in $\mathbb{R}^n$. In the last step we have used that $\max_{\ell\in[M]} \exp\left(\theta\|\mathbf{g}_\ell\|_2^2 \le \sum_{\ell=1}^M \exp\left(\theta\|\mathbf{g}_\ell\|_2^2\right.\right.$. By the independence of the components of $\mathbf{g}$ and Lemma 7.6,

$$\begin{aligned}\mathbb{E}\exp(\theta\|\mathbf{g}\|_2^2) &= \mathbb{E}\exp(\theta\sum_{j=1}^n g_j^2) = \mathbb{E}\prod_{j=1}^n \exp(\theta g_j^2) = \prod_{j=1}^n \mathbb{E}\exp(\theta g_j^2) \\ &= (1-2\theta)^{-n/2} ,\end{aligned}$$

provided that $\theta < 1/2$. Therefore,

$$\mathbb{E} \max_{\ell\in[M]} \|\mathbf{g}_\ell\|_2^2 \le \inf_{0<\theta<1/2} \theta^{-1}\left(\ln M + \frac{n}{2}\ln\left((1-2\theta)^{-1}\right)\right) .$$

Substituting $\theta = (2+2\kappa)^{-1}$ yields the first claim. Using that $\ln(1+\kappa^{-1}) \le \kappa^{-1}$, we further get

$$\mathbb{E} \max_{\ell\in[M]} \|\mathbf{g}_\ell\|_2^2 \le 2(1+\kappa)\ln(M) + n(1+\kappa^{-1}) . \tag{8.4}$$

Choosing $\kappa = \sqrt{n/(2\ln(M))}$ gives

$$\mathbb{E} \max_{\ell\in[M]} \|\mathbf{g}_\ell\|_2^2 \le 2\ln(M) + 2\sqrt{2n\ln(M)} + n = (\sqrt{2\ln(M)} + \sqrt{n})^2 .$$

This concludes the proof. □

8.2 Rademacher Sums and Symmetrization

A Rademacher variable (sometimes called Bernoulli random variable) is presumably the simplest random variable. It takes the values $+1$ or -1, each with probability $1/2$. A sequence $\boldsymbol{\epsilon}$ of independent Rademacher variables $\epsilon_\ell, \ell \in [M]$, is called a Rademacher sequence. In the sequel we will often consider Rademacher sums of the form

$$\sum_{\ell=1}^M \boldsymbol{\epsilon}_\ell \mathbf{x}_\ell ,$$

where the $\mathbf{x}_\ell$ are scalars, vectors or matrices.

Below we present the contraction principle for Rademacher sums and the symmetrization principle, which allows to replace a sum of independent random vectors by its randomized Rademacher sum in moment estimates. Although rather simple, this tool will prove very effective later.

Let us first present the contraction principle .

Theorem 8.3. *Let $\mathbf{x}_\ell, \ell \in [M]$, be vectors in a (finite-dimensional) vector space endowed with a norm $\|\cdot\|$ and $\alpha_\ell \in \mathbb{R}$, $\ell \in [M]$, be scalars satisfying $|\alpha_\ell| \le 1$. If $\boldsymbol{\epsilon} \in \mathbb{R}^M$ is a Rademacher sequence, then for any $1 \le p < \infty$,*

$$\mathbb{E}\|\sum_{\ell=1}^{M} \alpha_\ell \epsilon_\ell \mathbf{x}_\ell\|^p \le \mathbb{E}\|\sum_{\ell=1}^{M} \epsilon_\ell \mathbf{x}_\ell\|^p. \tag{8.5}$$

Proof. The function $(\alpha_1, \dots, \alpha_M) \mapsto \mathbb{E}\|\sum_{\ell=1}^{M} \alpha_\ell \epsilon_\ell \mathbf{x}_\ell\|^p$ is convex. Therefore, on $[-1,1]^M$ it attains its maximum at an extreme point, i.e., a point $\boldsymbol{\alpha} = (\alpha_\ell)_{\ell=1}^M$ such that $\alpha_\ell = \pm 1$, see Theorem B.16. For such values of α_ℓ, both $\alpha_\ell \epsilon_\ell$ and ϵ_ℓ have the same distribution and in this case both terms in (8.5) are equal. □

Symmetrization is a simple yet powerful technique to pass from a sum of arbitrary independent random variables to a Rademacher sum. A random vector $\mathbf{X} \in \mathbb{C}^n$ is called symmetric, if $\mathbf{X}$ and $-\mathbf{X}$ have the same distribution. Clearly, $\mathbb{E}\mathbf{X} = 0$ for a symmetric random vector $\mathbf{X}$. The crucial observation for symmetrization is that a symmetric random vector $\mathbf{X}$ and the random vector $\epsilon\mathbf{X}$, where ϵ is a Rademacher random variable independent of $\mathbf{X}$, have the same distribution.

Lemma 8.4. *Assume that $\boldsymbol{\xi} = (\boldsymbol{\xi}_\ell)_{\ell=1}^M$ is a sequence of independent random vectors in a finite-dimensional vector space X with norm $\|\cdot\|$. Let $F : X \to \mathbb{R}$ be a convex function. Then, with $\mathbf{x}_\ell = \mathbb{E}\boldsymbol{\xi}_\ell$,*

$$\mathbb{E}F\big(\sum_{\ell=1}^{M}(\boldsymbol{\xi}_\ell - \mathbf{x}_\ell)\big) \le \mathbb{E}F\big(2\sum_{\ell=1}^{M}\epsilon_\ell \boldsymbol{\xi}_\ell\big)\,, \tag{8.6}$$

where $\boldsymbol{\epsilon} = (\epsilon_\ell)_{\ell=1}^N$ is a Rademacher sequence independent of $\boldsymbol{\xi}$. In particular, for $1 \le p < \infty$,

$$\big(\mathbb{E}\|\sum_{\ell=1}^{M}(\boldsymbol{\xi}_\ell - \mathbf{x}_\ell)\|^p\big)^{1/p} \le 2\big(\mathbb{E}\|\sum_{\ell=1}^{M}\epsilon_\ell \boldsymbol{\xi}_\ell\|^p\big)^{1/p}\,, \tag{8.7}$$

Proof. Let $\boldsymbol{\xi}' = (\boldsymbol{\xi}'_1, \dots, \boldsymbol{\xi}'_M)$ denote an independent copy of the sequence of random vectors $(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_M)$. An application of Jensen's inequality yields

$$E := \mathbb{E}F\big(\sum_{\ell=1}^{M}(\boldsymbol{\xi}_\ell - \mathbf{x}_\ell)\big) = \mathbb{E}F\big(\sum_{\ell=1}^{M}(\boldsymbol{\xi}_\ell - \mathbb{E}\boldsymbol{\xi}'_\ell)\big) \le \mathbb{E}F\big(\sum_{\ell=1}^{M}(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)\big)\,.$$

Now observe that $(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell)_\ell$ is a sequence of independent symmetric random variables; hence, it has the same distribution as $(\epsilon_\ell(\boldsymbol{\xi}_\ell - \boldsymbol{\xi}'_\ell))_\ell$. Convexity of F gives

$$E \le \mathbb{E}F\big(\sum_{\ell=1}^{M} \epsilon_\ell(\xi_\ell - \xi'_\ell)\big) \le \mathbb{E}\left(\frac{1}{2}F\big(2\sum_{\ell=1}^{M}\epsilon_\ell \boldsymbol{\xi}_\ell\big) + \frac{1}{2}F\big(2\sum_{\ell=1}^{M}(-\epsilon_\ell)\boldsymbol{\xi}'_\ell\big)\right)$$
$$= \mathbb{E}F\big(2\sum_{\ell=1}^{M}\epsilon_\ell \boldsymbol{\xi}_\ell\big)$$

because $\boldsymbol{\epsilon}$ is symmetric and $\boldsymbol{\xi}'$ has the same distribution as $\boldsymbol{\xi}$. Inequality (8.7) follows from taking the convex function $F(\mathbf{x}) = \|\mathbf{x}\|^p$ for $p \in [1, \infty)$. □

The lemma will be very useful because there are powerful techniques for estimating Rademacher sums as we will see in the next section.

8.3 Khintchine Inequalities

Khintchine inequalities provide estimates of the moments of Rademacher and related sums.

Theorem 8.5. *Let* $\mathbf{a} \in \mathbb{C}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ *be a Rademacher sequence. Then, for all* $n \in \mathbb{N}$*,*

$$\mathbb{E}|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|^{2n} \le \frac{(2n)!}{2^n n!}\|\mathbf{a}\|_2^{2n} . \tag{8.8}$$

Proof. First assume that the a_ℓ are real-valued. Expanding the expectation on the left hand side of (8.8) with the multinomial theorem, see Appendix C.4, yields

$$E := \mathbb{E}|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|^{2n}$$
$$= \sum_{\substack{j_1+\cdots+j_M=n \\ j_i\ge 0}} \frac{(2n)!}{(2j_1)!\cdots(2j_M)!}|a_1|^{2j_1}\cdots|a_M|^{2j_M}\mathbb{E}\epsilon_1^{2j_1}\cdots\mathbb{E}\epsilon_M^{2j_M}$$
$$= \sum_{\substack{j_1+\cdots+j_M=n \\ j_i\ge 0}} \frac{(2n)!}{(2j_1)!\cdots(2j_M)!}|a_1|^{2j_1}\cdots|a_M|^{2j_M} .$$

Hereby we used the independence of the ϵ_ℓ and the fact that $\mathbb{E}\epsilon_\ell^k = 0$ if k is an odd integer. For integers satisfying $j_1 + \cdots + j_M = n$ we have

$$2^n j_1! \times \cdots \times j_M! = 2^{j_1} j_1! \times \cdots \times 2^{j_M} j_M! \le (2j_1)! \times \cdots \times (2j_M)! .$$

This implies

$$E \le \frac{(2n)!}{2^n n!} \sum_{\substack{j_1+\cdots+j_M=n \\ j_i \ge 0}} \frac{n!}{j_1! \cdots j_n!} |a_1|^{2j_1} \cdots |a_M|^{2j_M}$$
$$= \frac{(2n)!}{2^n n!} \big(\sum_{j=1}^{M} |a_j|^2 \big)^n = \frac{(2n)!}{2^n n!} \|\mathbf{a}\|_2^{2n} .$$

The complex case is derived by splitting into real and imaginary parts and applying the triangle inequality as follows,

$$\big(\mathbb{E}| \sum_{\ell=1}^{M} \epsilon_\ell (\mathrm{Re}(a_\ell) + i\,\mathrm{Im}(a_\ell))|^{2n}\big)^{1/2n}$$
$$= \big(\mathbb{E}\big[| \sum_{\ell=1}^{M} \epsilon_\ell \,\mathrm{Re}(a_\ell)|^2 + | \sum_{\ell=1}^{M} \epsilon_\ell \,\mathrm{Im}(a_\ell)|^2\big]^n\big)^{1/2n}$$
$$\le \big(\big(\mathbb{E}| \sum_{\ell=1}^{M} \epsilon_\ell \,\mathrm{Re}(a_\ell)|^{2n}\big)^{1/n} + \big(\mathbb{E}| \sum_{\ell=1}^{M} \epsilon_\ell \,\mathrm{Im}(a_\ell)|^{2n}\big)^{1/n}\big)^{1/2}$$
$$\le \left(\left(\frac{(2n)!}{2^n n!} \right)^{1/n} (\|\,\mathrm{Re}(\mathbf{a})\|_2^2 + \|\,\mathrm{Im}(\mathbf{a})\|_2^2) \right)^{1/2} = \left(\frac{(2n)!}{2^n n!} \right)^{1/2n} \|\mathbf{a}\|_2 .$$

This concludes the proof. □

Remark 8.6. (a) The constant in Khintchine's inequality can be expressed as a double factorial,

$$\frac{(2n)!}{2^n n!} = (2n-1)!! := 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-1) .$$

(b) If $\mathbf{g} = (g_1, \ldots, g_M)$ is a standard Gaussian random vector, then the sum $\sum_{\ell=1}^{M} a_\ell g_\ell$ with real a_ℓ is a Gaussian random variable with mean zero and variance $\|\mathbf{a}\|_2^2$. By Corollary 7.7 its moments are given by

$$\mathbb{E}| \sum_{\ell=1}^{M} a_\ell g_\ell|^{2n} = \frac{(2n)!}{2^n n!} \|\mathbf{a}\|_2^{2n} .$$

In other words, if the Rademacher sequence is replaced by independent standard normal variables then (8.8) holds with equality. Therefore, the central limit theorem shows that the constants in (8.8) are optimal. Moreover, it also follows that $\mathbb{E}| \sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^{2n} \le \mathbb{E}| \sum_{\ell=1}^{M} g_\ell a_\ell|^{2n}$, compare also Exercise 8.2.

Based on Theorem 8.5, we can also estimate the general absolute pth moment of a Rademacher sum.

Corollary 8.7. *Let* $\mathbf{a} \in \mathbb{C}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ *be a Rademacher sequence. Then, for all* $p > 0$,

$$\big(\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^p\big)^{1/p} \leq 2^{3/(4p)} e^{-1/2} \sqrt{p} \|\mathbf{a}\|_2 \,. \tag{8.9}$$

Proof. We first assume that $p \geq 2$. Stirling's formula (C.13) for the factorial gives

$$\frac{(2n)!}{2^n n!} = \frac{\sqrt{2\pi 2n}(2n/e)^{2n} e^{R_{2n}}}{2^n \sqrt{2\pi n}(n/e)^n e^{R_n}} \leq \sqrt{2}\,(2/e)^n n^n \,. \tag{8.10}$$

where $1/(12n+1) \leq R_n \leq 1/(12n)$. An application of Hölder's inequality yields, for $\theta \in [0,1]$, and an arbitrary random variable Z,

$$\mathbb{E}|Z|^{2n+2\theta} = \mathbb{E}[|Z|^{(1-\theta)2n} |Z|^{\theta(2n+2)}] \leq (\mathbb{E}|Z|^{2n})^{1-\theta} (\mathbb{E}|Z|^{2n+2})^{\theta} \,. \tag{8.11}$$

Without loss of generality we may assume $\|\mathbf{a}\|_2 = 1$. Combining the two estimates above yields

$$\begin{aligned}
\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^{2n+2\theta} &\leq (\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^{2n})^{1-\theta} (\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^{2n+2})^{\theta} \\
&\leq (\sqrt{2}(2/e)^n n^n)^{1-\theta} (\sqrt{2}(2/e)^{n+1}(n+1)^{n+1})^{\theta} \\
&= \sqrt{2}(2/e)^{n+\theta} n^{(1-\theta)n} (n+1)^{\theta(n+1)} \\
&= \sqrt{2}(2/e)^{n+\theta} (n^{1-\theta}(n+1)^{\theta})^{n+\theta} \left(\frac{n+1}{n}\right)^{\theta(1-\theta)} \\
&\leq \sqrt{2}(2/e)^{n+\theta} (n+\theta)^{n+\theta} \left(\frac{n+1}{n}\right)^{\theta(1-\theta)} \\
&\leq 2^{3/4} (2/e)^{n+\theta} (n+\theta)^{n+\theta} \,.
\end{aligned} \tag{8.12}$$

In the second line from below the inequality between the geometric and arithmetic mean was applied. The last step used that $(n+1)/n \leq 2$ and $\theta(1-\theta) \leq 1/4$. Replacing $n+\theta$ by $p/2$ completes the proof of (8.9) for $p \geq 2$.

For the case $0 < p \leq 2$ we observe that Hölder's inequality gives

$$(\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^p)^{1/p} \leq (\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^2)^{1/2} = 1 \,.$$

It is an elementary exercise to show that the function $f(p) = 2^{3/(4p)} e^{-1/2} \sqrt{p}$ takes its minimum at the point $p_0 = (3(\ln 2)/2)^{2/3}$ and $f(p_0) \approx 1.0197 > 1$. Therefore, we have (8.9) also for $p < 2$. □

We obtain the following version of Hoeffding's inequality for complex Rademacher sums.

Corollary 8.8. *Let* $\mathbf{a} \in \mathbb{C}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ *be a Rademacher sequence. Then, for* $u > 0$,

$$\mathbb{P}\big(|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell| \geq \|\mathbf{a}\|_2 u\big) \leq 2\exp(-u^2/2) \,. \tag{8.13}$$

Proof. We combine (8.9) with Proposition 7.11 to obtain

$$\mathbb{P}(|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell| \geq \|\mathbf{a}\|_2 u) \leq 2^{3/4}\exp(-u^2/2) \,, \quad u > 0 \,,$$

which is even slightly better (but less appealing) than the claimed estimate. □

A complex random variable which is uniformly distributed on the torus $\mathbb{T} = \{z \in \mathbb{C}, |z| = 1\}$ is called a Steinhaus variable. A sequence $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)$ of independent Steinhaus variables is called a *Steinhaus sequence*. There is also a version of Khintchine's inequality for Steinhaus sequences.

Theorem 8.9. *Let* $\mathbf{a} \in \mathbb{C}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ *be a Steinhaus sequence. Then, for all* $n \in \mathbb{N}$,

$$\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^{2n} \leq n!\|\mathbf{a}\|_2^{2n} \,.$$

Proof. We expand the moments of the Steinhaus sum using the multinomial theorem,

$$\begin{aligned}
\mathbb{E}|\sum_{\ell=1}^{M} \epsilon_\ell a_\ell|^{2n} &= \mathbb{E}\left[\big(\sum_{\ell=1}^{M} \epsilon_\ell a_\ell\big)^n \big(\sum_{\ell=1}^{M} \overline{\epsilon_\ell a_\ell}\big)^n\right] \\
&= \mathbb{E}\left[\sum_{\substack{j_1+\cdots+j_M=n \\ j_\ell \geq 0}} \frac{n!}{j_1!\cdots j_M!} a_1^{j_1}\cdots a_M^{j_M}\epsilon_1^{j_1}\cdots\epsilon_M^{j_M}\right. \\
&\quad \left. \times \sum_{\substack{k_1+\cdots k_M=n \\ k_\ell \geq 0}} \frac{n!}{k_1!\cdots k_M!} \overline{a_1^{k_1}\cdots a_M^{k_M}\epsilon_1^{k_1}\cdots\epsilon_M^{k_M}}\right] \\
&= \sum_{\substack{j_1+\cdots+j_M=n \\ k_1+\cdots+k_M=n \\ j_\ell,k_\ell \geq 0}} \frac{n!}{j_1!\cdots j_M!}\frac{n!}{k_1!\cdots k_M!} a_1^{j_1}\overline{a_1^{k_1}}\cdots a_M^{j_M}\overline{a_M^{k_M}}\,\mathbb{E}[\epsilon_1^{j_1}\overline{\epsilon_1^{k_1}}\cdots\epsilon_M^{j_M}\overline{\epsilon_M^{k_M}}] \,.
\end{aligned}$$

Since the ϵ_j are independent and uniformly distributed on the torus it holds

$$\mathbb{E}[\epsilon_1^{j_1}\overline{\epsilon_1^{k_1}}\cdots\epsilon_M^{j_M}\overline{\epsilon_M^{k_M}}] = \mathbb{E}[\epsilon_1^{j_1}\overline{\epsilon_1^{k_1}}] \times \cdots \times \mathbb{E}[\epsilon_M^{j_M}\overline{\epsilon_M^{k_M}}] = \delta_{j_1,k_1} \times \cdots \times \delta_{j_M,k_M} \,.$$

This yields

$$\begin{aligned}\mathbb{E}|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|^{2n} &= \sum_{\substack{k_1+\cdots k_M=n\\ k_\ell\geq 0}}\left(\frac{n!}{k_1!\cdots k_M!}\right)^2 |a_1|^{2k_1}\cdots|a_M|^{2k_M}\\ &\leq n!\sum_{\substack{k_1+\cdots+k_M=n\\ k_\ell\geq 0}}\frac{n!}{k_1!\cdots k_M!}|a_1|^{2k_1}\cdots|a_M|^{2k_M}\\ &= n!\big(\sum_{\ell=1}^{M}|a_\ell|^2\big)^{2n}\,,\end{aligned}$$

where the multinomial theorem was applied in the last step. □

The above moment estimate leads to a Hoeffding type inequality for Steinhaus sums.

Corollary 8.10. *Let* $\mathbf{a}\in\mathbb{C}^M$ *and* $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_M)$ *be a Steinhaus sequence. Assume* $0<\lambda<1$*. Then*

$$\mathbb{P}(|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|\geq u\|\mathbf{a}\|_2)\leq\frac{1}{1-\lambda}e^{-\lambda u^2}\quad\text{ for all } u>0\,. \tag{8.14}$$

In particular, using the optimal choice $\lambda=1-u^{-2}$*,*

$$\mathbb{P}\big(|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|\geq u\|\mathbf{a}\|_2\big)\leq\exp(-u^2+\ln(u^2)+1)\quad\text{ for all } u\geq 1\,. \tag{8.15}$$

Proof. Without loss of generality we assume that $\|\mathbf{a}\|_2=1$. Markov's inequality gives

$$\begin{aligned}&\mathbb{P}\big(|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|\geq u\big)=\mathbb{P}\big(\exp(\lambda|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|^2)\geq\exp(\lambda u^2)\big)\\ &\leq\mathbb{E}[\exp(\lambda|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|^2)]\exp(-\lambda u^2)=\exp(-\lambda u^2)\sum_{n=0}^{\infty}\frac{\lambda^n\mathbb{E}|\sum_{\ell=1}^{M}\epsilon_\ell a_\ell|^{2n}}{n!}\\ &\leq\exp(-\lambda u^2)\sum_{n=0}^{\infty}\lambda^n=\frac{1}{1-\lambda}e^{-\lambda u^2}\,.\end{aligned}$$

In the second line Fubini's theorem and in the third line Theorem 8.9 was applied. □

8.4 Decoupling

Decoupling is a technique that reduces stochastic dependencies in certain sums of random variables, called chaos variables. A typical example is a sum of the form $\sum_{j\neq k} \epsilon_j \epsilon_k \mathbf{x}_{j,k}$ where the $\mathbf{x}_{j,k}$ are some vectors and $\boldsymbol{\epsilon} = (\epsilon_j)$ is a Rademacher sequence. Such a sum is called a homogeneous Rademacher chaos of order 2. The term homogeneous refers to the fact that the diagonal terms in this double sum are missing so that its expectation is zero. The following statement provides a way of "decoupling" the sum.

Theorem 8.11. *Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$ be a sequence of independent random variables with $\mathbb{E}\xi_j = 0$ for all $j \in [M]$. Let $\mathbf{x}_{j,k}$, $j,k \in [M]$, be a double sequence of elements in a finite-dimensional vector space X. If $F : X \to \mathbb{R}$ is a convex function, then*

$$\mathbb{E}F\left(\sum_{\substack{j,k=1\\ j\neq k}}^{M} \xi_j \xi_k \mathbf{x}_{j,k}\right) \leq \mathbb{E}F\left(4\sum_{j,k=1}^{M} \xi_j \xi'_k \mathbf{x}_{j,k}\right) , \tag{8.16}$$

where $\boldsymbol{\xi}'$ denotes an independent copy of $\boldsymbol{\xi}$.

Proof. Introduce a sequence $\boldsymbol{\delta} = (\delta_j)_{j=1}^M$ of independent random variables δ_j taking the values 0 and 1 with probability 1/2. Then, for $j \neq k$,

$$\mathbb{E}\delta_k(1-\delta_j) = 1/4 . \tag{8.17}$$

This gives

$$\begin{aligned} E := \mathbb{E}F\left(\sum_{j\neq k}^{M} \xi_j \xi_k \mathbf{x}_{j,k}\right) &= \mathbb{E}_{\boldsymbol{\xi}} F\left(4\sum_{j\neq k}^{M} \mathbb{E}_\delta[\delta_j(1-\delta_k)]\xi_j\xi_k\mathbf{x}_{j,k}\right) \\ &\leq \mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{\boldsymbol{\delta}} F\left(4\sum_{j\neq k}^{M} \delta_j(1-\delta_k)\xi_j\xi_k\mathbf{x}_{j,k}\right) , \end{aligned}$$

where Jensen's inequality was applied in the last step. Now let

$$\sigma(\boldsymbol{\delta}) := \{j \in [M] : \delta_j = 1\}.$$

Then, by Fubini's theorem,

$$E \leq \mathbb{E}_{\boldsymbol{\delta}}\mathbb{E}_{\boldsymbol{\xi}} F\left(4\sum_{j\in\sigma(\boldsymbol{\delta})}\sum_{k\notin\sigma(\boldsymbol{\delta})} \xi_j\xi_k\mathbf{x}_{j,k}\right) .$$

For fixed $\boldsymbol{\delta}$ the sequences $(\xi_j)_{j\in\sigma(\delta)}$ and $(\xi_k)_{k\notin\sigma(\delta)}$ are independent, hence, we can replace ξ_k, $k \notin \sigma(\delta)$, by an independent copy ξ'_k and obtain

$$E \le \mathbb{E}_{\boldsymbol{\delta}} \mathbb{E}_{\boldsymbol{\xi}} \mathbb{E}_{\boldsymbol{\xi}'} F \left(4 \sum_{j \in \sigma(\delta)} \sum_{k \notin \sigma(\delta)} \xi_j \xi'_k \mathbf{x}_{j,k} \right) .$$

This implies the existence of a $\boldsymbol{\delta}^* \in \{0,1\}^M$, and hence a $\sigma = \sigma(\boldsymbol{\delta}^*)$ such that

$$E \le \mathbb{E}_{\boldsymbol{\xi}} \mathbb{E}_{\boldsymbol{\xi}'} F \left(4 \sum_{j \in \sigma} \sum_{k \notin \sigma} \xi_j \xi'_k \mathbf{x}_{j,k} \right) .$$

Since $\mathbb{E}\xi_j = \mathbb{E}\xi'_j = 0$, an application of Jensen's inequality yields

$$\begin{aligned} E &\le \mathbb{E} F \left(4 \sum_{j \in \sigma} \left(\sum_{k \notin \sigma} \xi_j \xi'_k \mathbf{x}_{j,k} + \sum_{k \in \sigma} \xi_j \mathbb{E}[\xi'_k] \mathbf{x}_{j,k} \right) + 4 \sum_{j \notin \sigma} \mathbb{E}[\xi_j] \sum_{k=1}^{M} \xi'_k \mathbf{x}_{j,k} \right) \\ &\le \mathbb{E} F \left(4 \sum_{j=1}^{M} \sum_{k=1}^{M} \xi_j \xi'_k \mathbf{x}_{j,k} \right) , \end{aligned}$$

and the proof is complete. □

The sum $\sum_{j,k} \xi_j \xi'_k \mathbf{x}_{j,k}$ on the right hand side of (8.16) is called a decoupled chaos. It is important that the double sum on the left hand side of (8.16) runs only over indices $j \neq k$. Moreover, since the left hand side of (8.16) is independent of the diagonal entries $\mathbf{x}_{j,j}$, they can be chosen arbitrarily on the right hand side. Sometimes it is convenient to choose them as $\mathbf{x}_{j,j} = 0$, but other choices may simplify computations.

An important special case of the above theorem is $F(\mathbf{x}) = \|\mathbf{x}\|^p$ with $p \ge 1$ and some (semi-)norm $\|\cdot\|$. Then (8.16) implies

$$\left(\mathbb{E} \left\| \sum_{j \neq k} \xi_j \xi_k \mathbf{x}_{j,k} \right\|^p \right)^{1/p} \le 4 \left(\mathbb{E} \left\| \sum_{j,k} \xi_j \xi'_k \mathbf{x}_{j,k} \right\|^p \right)^{1/p} .$$

The mean-zero assumption above for the random variables ξ_j can be removed after possibly adjusting constants. We will exemplify this for the following special case involving the operator norm where, additionally, the constant can be improved.

Theorem 8.12. *Let $\widehat{\mathbf{H}} \in \mathbb{C}^{M \times M}$ be self-adjoint, and $\mathbf{H}$ the matrix $\widehat{\mathbf{H}}$ with the diagonal entries put to zero. Let ξ_j, $j \in [M]$, be a sequence of independent random variables. Introduce the random diagonal matrix $\mathbf{D}_{\boldsymbol{\xi}} = \operatorname{diag}(\xi_j, j \in [M])$. If $F : \mathbb{R}_+ \to \mathbb{R}$ is a convex nondecreasing function, then*

$$\mathbb{E} F(\|\mathbf{D}_{\boldsymbol{\xi}} \mathbf{H} \mathbf{D}_{\boldsymbol{\xi}}\|_{2\to 2}) \le \mathbb{E} F(2\|\mathbf{D}_{\boldsymbol{\xi}} \widehat{\mathbf{H}} \mathbf{D}_{\boldsymbol{\xi}'}\|_{2\to 2}) , \tag{8.18}$$

where $\boldsymbol{\xi}'$ denotes an independent copy of $\boldsymbol{\xi}$.

Proof. Let $\mathbf{H}_{jk} \in \mathbb{C}^{M\times M}$ be the matrix with entry $\widehat{H}_{jk}$ in position (j,k) and zero elsewhere. Let δ_j, $j \in [M]$, be independent Bernoulli random variables taking the values 0 and 1 both with probability 1/2. The function $\mathbf{x} \mapsto F(\|\mathbf{x}\|_{2\to 2})$ is convex by Proposition B.10(b) so that Jensen's inequality and (8.17) yield

$$\begin{aligned}
\mathbb{E}F\big(\|\mathbf{D}_{\boldsymbol{\xi}}\mathbf{H}\mathbf{D}_{\boldsymbol{\xi}}\|_{2\to 2}\big) &= \mathbb{E}F\big(\|\sum_{j<k}\xi_j\xi_k(\mathbf{H}_{jk}+\mathbf{H}_{kj})\|_{2\to 2}\big)\\
&= \mathbb{E}_{\boldsymbol{\xi}}F\big(2\|\mathbb{E}_{\boldsymbol{\delta}}\sum_{j<k}[\delta_j(1-\delta_k)+\delta_k(1-\delta_j)]\xi_j\xi_k(\mathbf{H}_{jk}+\mathbf{H}_{kj})\|_{2\to 2}\big)\\
&\le \mathbb{E}_{\boldsymbol{\xi}}\mathbb{E}_{\boldsymbol{\delta}}F\big(2\|\sum_{j<k}[\delta_j(1-\delta_k)+\delta_k(1-\delta_j)]\xi_j\xi_k(\mathbf{H}_{jk}+\mathbf{H}_{kj})\|_{2\to 2}\big)\,. \qquad (8.19)
\end{aligned}$$

Therefore, there exists a vector $\boldsymbol{\delta}^*$ with entries in $\{0,1\}$ such that

$$\mathbb{E}F\big(\|\mathbf{D}_{\boldsymbol{\xi}}\mathbf{H}\mathbf{D}_{\boldsymbol{\xi}}\|_{2\to 2}\big) \le \mathbb{E}F\big(2\|\sum_{j<k}[\delta_j^*(1-\delta_k^*)+\delta_k^*(1-\delta_j^*)]\xi_j\xi_k(\mathbf{H}_{jk}+\mathbf{H}_{kj})\|_{2\to 2}\big)\,.$$

Let $\sigma = \sigma(\boldsymbol{\delta}^*) = \{j \in [M], \delta_j^* = 1\}$. Then

$$\mathbb{E}F\big(\|\mathbf{D}_{\boldsymbol{\xi}}\mathbf{H}\mathbf{D}_{\boldsymbol{\xi}}\|_{2\to 2}\big) \le \mathbb{E}F\big(2\|\sum_{j\in\sigma,k\in\overline{\sigma}}\xi_j\xi_k(\mathbf{H}_{jk}+\mathbf{H}_{kj})\|_{2\to 2}\big)\,.$$

By rearranging the index set, we may assume that $\sigma = \{1,\dots,\mathrm{card}(\sigma)\}$ and $\overline{\sigma} = \{\mathrm{card}(\sigma)+1,\dots,M\}$. Then we can write

$$\sum_{j\in\sigma,k\in\overline{\sigma}}\xi_j\xi_k(\mathbf{H}_{jk}+\mathbf{H}_{kj}) = \begin{pmatrix}\mathbf{0} & \mathbf{B}\\ \mathbf{B}^* & \mathbf{0}\end{pmatrix}$$

with $\mathbf{B} \in \mathbb{C}^{\mathrm{card}(\sigma)\times\mathrm{card}(\overline{\sigma})}$ being the restriction of $\sum_{j\in\sigma,k\in\overline{\sigma}} = \xi_j\xi_k\mathbf{H}_{jk}$ to the indices in $\sigma\times\overline{\sigma}$. Using

$$\left\|\begin{pmatrix}\mathbf{0} & \mathbf{B}\\ \mathbf{B}^* & \mathbf{0}\end{pmatrix}\right\|_{2\to 2} = \|\mathbf{B}\|_{2\to 2}$$

we arrive at

$$\begin{aligned}
\mathbb{E}F\big(\|\mathbf{D}_{\boldsymbol{\xi}}\mathbf{H}\mathbf{D}_{\boldsymbol{\xi}}\|_{2\to 2}\big) &\le \mathbb{E}F\big(2\|\sum_{j\in\sigma,k\in\overline{\sigma}}\xi_j\xi_k\mathbf{H}_{jk}\|_{2\to 2}\big)\\
&= \mathbb{E}F\big(2\|\sum_{j\in\sigma,k\in\overline{\sigma}}\xi_j\xi_k'\mathbf{H}_{jk}\|_{2\to 2}\big)\,,
\end{aligned}$$

where $\boldsymbol{\xi}'$ is an independent copy of $\boldsymbol{\xi}$. Since the operator norm of a submatrix is bounded by the operator norm of the full matrix, see Lemma A.10, we reinsert the missing entries to get

$$\mathbb{E}F\big(\|\mathbf{D}_{\boldsymbol{\xi}}\mathbf{H}\mathbf{D}_{\boldsymbol{\xi}}\|_{2\to 2}\big) \le \mathbb{E}F\big(2\|\sum_{j,k}\xi_j\xi_k'\mathbf{H}_{jk}\|_{2\to 2}\big) = \mathbb{E}F\big(2\|\mathbf{D}_{\boldsymbol{\xi}}\widehat{\mathbf{H}}\mathbf{D}_{\boldsymbol{\xi}'}\|_{2\to 2}\big)\,,$$

where we used that F is nondecreasing. This completes the argument. □

We finish this section with an application to tail bounds for scalar Rademacher chaos. Let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ be a Rademacher vector. For a self-adjoint matrix $\mathbf{A} \in \mathbb{C}^{M\times M}$ with zero diagonal we consider the homogeneous Rademacher chaos

$$X := \boldsymbol{\epsilon}^* \mathbf{A} \boldsymbol{\epsilon} = \sum_{j\neq k} \epsilon_j \epsilon_k A_{jk} \; . \tag{8.20}$$

Note that by self-adjointness, X is real-valued even if $\mathbf{A}$ is complex-valued. This fact allows to reduce our considerations to real-valued symmetric matrices $\mathbf{A} \in \mathbb{R}^{M\times M}$ since $X = \operatorname{Re}(X) = \boldsymbol{\epsilon}^* \operatorname{Re}(\mathbf{A})\boldsymbol{\epsilon}$. The next result states that a homogeneous Rademacher chaos obeys a mixture of subgaussian and subexponential tail behavior, similar to Bernstein inequalities. The subgaussian part is determined by the Frobenius norm $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{A})$, see (A.15), while the operator norm $\|\mathbf{A}\|_{2\to 2}$ controls the subexponential part.

Proposition 8.13. *Let $\mathbf{A} \in \mathbb{R}^{M\times M}$ be a symmetric matrix with zero diagonal, and $\boldsymbol{\epsilon}$ a Rademacher vector. Then the homogeneous Rademacher chaos X defined in* (8.20) *satisfies, for $t > 0$,*

$$\begin{aligned}\mathbb{P}\left(|\sum_{j\neq k} \epsilon_j \epsilon_k A_{jk}| \geq t \right) &\leq 2\exp\left(-\min\left\{ \frac{3t^2}{128\|\mathbf{A}\|_F^2}, \frac{t}{32\|\mathbf{A}\|_{2\to 2}} \right\} \right) \\ &= \begin{cases} 2\exp\left(-\frac{3t^2}{128\|\mathbf{A}\|_F^2}\right) & \text{if } 0 < t \leq \frac{4\|\mathbf{A}\|_F^2}{3\|\mathbf{A}\|_{2\to 2}} \\ 2\exp\left(-\frac{t}{32\|\mathbf{A}\|_{2\to 2}}\right) & \text{if } t > \frac{4\|\mathbf{A}\|_F^2}{3\|\mathbf{A}\|_{2\to 2}} \; . \end{cases}\end{aligned}$$

Proof. The proof is based on an estimate of the moment generating function of X. For $\theta > 0$, convexity of $x \mapsto \exp(\theta x)$ combined with the decoupling inequality (8.16) yields

$$\begin{aligned}&\mathbb{E}\exp(\theta X) = \mathbb{E}\exp(\theta \sum_{j\neq k} \epsilon_j \epsilon_k A_{jk}) \leq \mathbb{E}\exp(4\theta \sum_{j,k} \epsilon_j \epsilon'_k A_{jk}) \\ &= \mathbb{E}_{\boldsymbol{\epsilon}} \mathbb{E}_{\boldsymbol{\epsilon}'} \exp\left(4\theta \sum_k \epsilon'_k \sum_j \epsilon_j A_{jk}\right) \leq \mathbb{E}\exp\left(8\theta^2 \sum_k (\sum_j \epsilon_j A_{jk})^2\right) \; . \end{aligned} \tag{8.21}$$

In the last step we have applied Theorem 7.27 conditionally on $\boldsymbol{\epsilon}$, using that $c = 1/2$ for Rademacher variables, see Remark 7.26. Observe that by symmetry of $\mathbf{A}$

$$\sum_k (\sum_j \epsilon_j A_{jk})^2 = \sum_k \sum_j \epsilon_j A_{jk} \sum_\ell \epsilon_\ell A_{\ell k} = \sum_{j,\ell} \epsilon_j \epsilon_\ell \sum_k A_{jk} A_{k\ell} = \boldsymbol{\epsilon}^* \mathbf{A}^2 \boldsymbol{\epsilon} \; .$$

Set $\mathbf{B} = \mathbf{A}^2$. The moment generating function of the positive semidefinite chaos $\boldsymbol{\epsilon}^* \mathbf{B} \boldsymbol{\epsilon}$ can be estimated by

$$\begin{aligned}\mathbb{E}\exp(\kappa\boldsymbol{\epsilon}^*\mathbf{B}\boldsymbol{\epsilon}) &= \mathbb{E}\exp(\kappa\sum_j B_{jj} + \kappa\sum_{j\neq k}\epsilon_j\epsilon_k B_{jk}) \\ &\leq \exp(\kappa\mathrm{tr}\,(\mathbf{B}))\mathbb{E}\exp\left(4\kappa\sum_{j,k}\epsilon_j\epsilon'_k B_{jk}\right) \\ &\leq \exp(\kappa\mathrm{tr}\,(\mathbf{B}))\mathbb{E}\exp\left(8\kappa^2\sum_k(\sum_j\epsilon_j B_{j,k})^2\right)\,,\end{aligned}$$

where we have again applied the decoupling inequality (8.16) together with Theorem 7.27 conditionally on $\boldsymbol{\epsilon}$. Now, positive semidefiniteness of $\mathbf{B} = \mathbf{A}^*\mathbf{A}$ allows to take the square root of $\mathbf{B}$ so that

$$\sum_k(\sum_j\epsilon_j B_{j,k})^2 = \boldsymbol{\epsilon}^*\mathbf{B}^2\boldsymbol{\epsilon} = (\mathbf{B}^{1/2}\boldsymbol{\epsilon})^*\mathbf{B}(\mathbf{B}^{1/2}\boldsymbol{\epsilon}) \leq \|\mathbf{B}\|_{2\to 2}\boldsymbol{\epsilon}^*\mathbf{B}\boldsymbol{\epsilon}\,.$$

If $8\kappa\|\mathbf{B}\|_{2\to 2} < 1$ then Hölder's (or Jensen's) inequality yields

$$\begin{aligned}\mathbb{E}\exp(\kappa\boldsymbol{\epsilon}^*\mathbf{B}\boldsymbol{\epsilon}) &\leq \exp(\kappa\mathrm{tr}\,(\mathbf{B}))\mathbb{E}\exp(8\kappa^2\|\mathbf{B}\|_{2\to 2}\boldsymbol{\epsilon}^*\mathbf{B}\boldsymbol{\epsilon}) \\ &\leq \exp(\kappa\mathrm{tr}\,(\mathbf{B}))\left(\mathbb{E}\exp(\kappa\boldsymbol{\epsilon}^*\mathbf{B}\boldsymbol{\epsilon})\right)^{8\kappa\|\mathbf{B}\|_{2\to 2}}\,.\end{aligned}$$

After rearranging we deduce that

$$\mathbb{E}\exp(\kappa\boldsymbol{\epsilon}^*\mathbf{B}\boldsymbol{\epsilon}) \leq \exp\left(\frac{\kappa\mathrm{tr}\,(\mathbf{B})}{1-8\kappa\|\mathbf{B}\|_{2\to 2}}\right)\,,\quad 0<\kappa<(8\|\mathbf{B}\|_{2\to 2})^{-1}\,. \tag{8.22}$$

Setting $\kappa = 8\theta^2$ and plugging into (8.21) yields, for $0 < \theta < (8\|\mathbf{A}\|_{2\to 2})^{-1}$,

$$\mathbb{E}\exp(\theta X) \leq \exp\left(\frac{8\theta^2\mathrm{tr}\,(\mathbf{A}^2)}{1-64\theta^2\|\mathbf{A}^2\|_{2\to 2}}\right) = \exp\left(\frac{8\theta^2\|\mathbf{A}\|_F^2}{1-64\theta^2\|\mathbf{A}\|_{2\to 2}^2}\right)\,.$$

Next we use Markov's inequality to deduce, for $0 < \theta \leq (16\|\mathbf{A}\|_{2\to 2})^{-1}$

$$\begin{aligned}\mathbb{P}(X\geq t) &= \mathbb{P}(\exp(\theta X)\geq\exp(\theta t)) \leq \exp(-\theta t)\mathbb{E}\exp(\theta X) \\ &\leq \exp\left(-\theta t + \frac{8\theta^2\|\mathbf{A}\|_F^2}{1-64\theta^2\|\mathbf{A}\|_{2\to 2}^2}\right) \leq \exp\left(-\theta t + \frac{8\theta^2\|\mathbf{A}\|_F^2}{1-1/4}\right) \\ &= \exp\left(-\theta t + 32\theta^2\|\mathbf{A}\|_F^2/3\right)\,.\end{aligned}$$

The optimal choice $\theta = 3t/(64\|\mathbf{A}\|_F^2)$ satisfies $\theta \leq (16\|\mathbf{A}\|_{2\to 2})^{-1}$ provided that $t \leq 4\|\mathbf{A}\|_F^2/(3\|\mathbf{A}\|_{2\to 2})$. In this regime, we therefore obtain

$$\mathbb{P}(X\geq t) \leq \exp\left(-\frac{3t^2}{128\|\mathbf{A}\|_F^2}\right)\,.$$

In the other regime where $t > 4\|\mathbf{A}\|_F^2/(3\|\mathbf{A}\|_{2\to 2})$ we set $\theta = (16\|\mathbf{A}\|_{2\to 2})^{-1}$ so that $\theta < 3t/(64\|\mathbf{A}\|_F^2)$. Then

$$\begin{aligned}\mathbb{P}(X\geq t) &\leq \exp\left(-\theta t + 32\theta^2\|\mathbf{A}\|_F^2/3\right) \leq \exp\left(-\theta t + \theta t/2\right) = \exp(-\theta t/2) \\ &= \exp(-t/(32\|\mathbf{A}\|_{2\to 2}))\,.\end{aligned}$$

Since X has the same distribution as $-X$, we get the same bounds for $\mathbb{P}(X \leq -t)$, and the union bound completes the proof. □

8.5 Noncommutative Bernstein Inequality

The scalar Bernstein inequalities from the previous section have a powerful extension to sums of random matrices. We present one version below. Another version is treated in Exercise 8.8. We denote by $\lambda_{\max}(\mathbf{X})$ the maximal eigenvalue of a selfadjoint square matrix $\mathbf{X}$. Furthermore, we introduce the function

$$h(x) := (1+x)\ln(1+x) - x \; . \tag{8.23}$$

Theorem 8.14. *Let $\mathbf{X}_1, \dots, \mathbf{X}_M \in \mathbb{C}^{d\times d}$ be independent mean-zero self-adjoint random matrices. Assume that the largest eigenvalue of $\mathbf{X}_\ell$ satisfies*

$$\lambda_{\max}(\mathbf{X}_\ell) \le K \quad \text{almost surely for all } \ell \in [M] \; , \tag{8.24}$$

and set

$$\sigma^2 := \left\| \sum_{\ell=1}^M \mathbb{E}(\mathbf{X}_\ell^2) \right\|_{2\to 2} \; .$$

Then, for $t > 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^M \mathbf{X}_\ell \right) \ge t \right) \le d \exp\left(-\frac{\sigma^2}{K^2} h\left(\frac{Kt}{\sigma^2} \right) \right) \tag{8.25}$$

$$\le d \exp\left(-\frac{t^2/2}{\sigma^2 + Kt} \right) \; . \tag{8.26}$$

The inequality (8.25) may also be referred to as matrix Bennett inequality. Although it is slightly stronger than the matrix Bernstein inequality (8.26), the latter is usually more convenient to use. Clearly, the difference with respect to the scalar Bernstein inequalities of the previous chapter is only the appearance of the dimensional factor d in front of the exponential. In general, this factor cannot be avoided.

Since for a self-adjoint matrix $\|\mathbf{A}\|_{2\to 2} = \max\{\lambda_{\max}(\mathbf{A}), \lambda_{\max}(-\mathbf{A})\}$, we obtain the next statement as a simple consequence.

Corollary 8.15. *Let $\mathbf{X}_1, \dots, \mathbf{X}_M \in \mathbb{C}^{d\times d}$ be independent mean-zero self-adjoint random matrices. Assume that*

$$\|\mathbf{X}_\ell\|_{2\to 2} \le K \quad \text{almost surely,} \quad \ell \in [M] \; , \tag{8.27}$$

and set

$$\sigma^2 := \left\| \sum_{\ell=1}^M \mathbb{E}(\mathbf{X}_\ell^2) \right\|_{2\to 2} \; . \tag{8.28}$$

Then, for $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{\ell=1}^{M}\mathbf{X}_\ell\right\| \geq t\right) \leq 2d\exp\left(-\frac{\sigma^2}{K^2}h\left(\frac{Kt}{\sigma^2}\right)\right) \tag{8.29}$$

$$\leq 2d\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right) . \tag{8.30}$$

An extension to rectangular (and not necessarily self-adjoint) matrices is developed in Exercise 8.7.

The essential steps of the proof proceed in the same way as the ones of the scalar Bernstein inequality, but since we are dealing with matrices, we encounter some additional complications. We will use an extension of the Laplace transform method (or moment generating function method) to matrices. A crucial ingredient is Lieb's concavity Theorem B.31.

We start with a simple consequence of the Markov inequality. It uses the matrix exponential $\mathbf{A} \mapsto \exp(\mathbf{A})$ defined in (A.45). We refer to Appendix A.5 for basic facts on matrix functions.

Proposition 8.16. *Let $\mathbf{Y} \in \mathbb{C}^{d\times d}$ be a self-adjoint random matrix. Then, for $t \in \mathbb{R}$,*

$$\mathbb{P}(\lambda_{\max}(\mathbf{Y}) \geq t) \leq \inf_{\theta>0}\left\{e^{-\theta t}\mathbb{E}\operatorname{tr}\exp(\theta\mathbf{Y})\right\} \tag{8.31}$$

Proof. For any $\theta > 0$ Markov's inequality, Theorem 7.3, yields

$$\mathbb{P}(\lambda_{\max}(\mathbf{Y}) \geq t) = \mathbb{P}\left(e^{\lambda_{\max}(\theta\mathbf{Y})} \geq e^{\theta t}\right) \leq e^{-\theta t}\mathbb{E}\left[e^{\lambda_{\max}(\theta\mathbf{Y})}\right] . \tag{8.32}$$

By the spectral mapping theorem (A.42) (or by the definition of a matrix function), and positivity of the exponential function, we have

$$e^{\lambda_{\max}(\theta\mathbf{Y})} = \lambda_{\max}(e^{\theta\mathbf{Y}}) \leq \sum_{j=1}^{d}\lambda_j(e^{\theta\mathbf{Y}}) = \operatorname{tr} e^{\theta\mathbf{Y}} ,$$

where $\lambda_j(e^{\theta\mathbf{Y}}) \geq 0$, $j \in [d]$, are the eigenvalues of $e^{\theta\mathbf{Y}}$ (possibly with repetitions). Combined with the previous estimate we reach

$$\mathbb{P}(\lambda_{\max}(\mathbf{Y}) \geq t) \leq e^{-\theta t}\mathbb{E}\operatorname{tr} e^{\theta\mathbf{Y}} .$$

Taking the infimum over all positive θ concludes the proof. □

The previous proposition suggests to study the expectation of the trace exponential $\theta \mapsto \mathbb{E}\operatorname{tr} e^{\theta\mathbf{Y}}$. The next result provides a useful tool for analyzing it, and is a consequence of Lieb's theorem B.31. We will use the matrix logarithm introduced in Appendix A.5, see (A.50).

Proposition 8.17. *Let $\mathbf{H} \in \mathbb{C}^{d\times d}$ be a fixed self-adjoint matrix, and let $\mathbf{Y} \in \mathbb{C}^{d\times d}$ be a self-adjoint random matrix. Then*

$$\mathbb{E}\operatorname{tr}\exp(\mathbf{H}+\mathbf{Y}) \leq \operatorname{tr}\exp\left(\mathbf{H} + \ln\left(\mathbb{E}e^{\mathbf{Y}}\right)\right) . \tag{8.33}$$

Proof. With $\mathbf{X} = e^{\mathbf{Y}}$ we have $\mathbf{Y} = \ln(\mathbf{X})$ by (A.50). By Lieb's Theorem B.31 the function $\mathbf{X} \mapsto \operatorname{tr} \exp(\mathbf{H} + \ln(\mathbf{X}))$ is concave. Jensen's inequality (7.18) therefore gives

$$\begin{aligned}\mathbb{E}\operatorname{tr} \exp(\mathbf{H} + \mathbf{Y}) &= \mathbb{E}\operatorname{tr} \exp(\mathbf{H} + \ln(\mathbf{X})) \leq \operatorname{tr} \exp(\mathbf{H} + \ln(\mathbb{E}\mathbf{X})) \\ &= \operatorname{tr} \exp\left(\mathbf{H} + \ln\left(\mathbb{E}e^{\mathbf{Y}}\right)\right) \ .\end{aligned}$$

This concludes the proof. □

The next tool extends the previous inequality to a sequence of independent random matrices.

Proposition 8.18. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_M \in \mathbb{C}^{d\times d}$ be independent, self-adjoint random matrices. Then, for $\theta \in \mathbb{R}$,*

$$\mathbb{E}\operatorname{tr} \exp\left(\theta \sum_{\ell=1}^{M} \mathbf{X}_\ell\right) \leq \operatorname{tr} \exp\left(\sum_{\ell=1}^{M} \ln \mathbb{E} \exp(\theta \mathbf{X}_\ell)\right) \ . \tag{8.34}$$

Proof. Without loss of generality we may assume that $\theta = 1$. We denote

$$\mathbf{Z}_\ell := \ln \mathbb{E} \exp(\mathbf{X}_\ell) \ .$$

Since the $\mathbf{X}_\ell$ are independent, we are in the position to write $\mathbb{E}_{\mathbf{X}_\ell}$ for the expectation with respect to $\mathbf{X}_\ell$ (or in other words, the expectation conditional on $\mathbf{X}_1, \ldots, \mathbf{X}_{\ell-1}, \mathbf{X}_\ell, \ldots, \mathbf{X}_M$). Using Fubini's theorem and Proposition 8.17 we arrive at

$$\begin{aligned}\mathbb{E}\operatorname{tr} \exp\left(\sum_{\ell=1}^{M} \mathbf{X}_\ell\right) &= \mathbb{E}_{\mathbf{X}_1} \cdots \mathbb{E}_{\mathbf{X}_M} \operatorname{tr} \exp\left(\sum_{\ell=1}^{M-1} \mathbf{X}_\ell + \mathbf{X}_M\right) \\ &\leq \mathbb{E}_{\mathbf{X}_1} \cdots \mathbb{E}_{\mathbf{X}_{M-1}} \operatorname{tr} \exp\left(\sum_{\ell=1}^{M-1} \mathbf{X}_\ell + \ln \mathbb{E} \exp(\mathbf{X}_M)\right) \\ &= \mathbb{E}_{\mathbf{X}_1} \cdots \mathbb{E}_{\mathbf{X}_{M-1}} \operatorname{tr} \exp\left(\sum_{\ell=1}^{M-2} \mathbf{X}_\ell + \mathbf{Z}_M + \mathbf{X}_{M-1}\right) \\ &\leq \mathbb{E}_{\mathbf{X}_1} \cdots \mathbb{E}_{\mathbf{X}_{M-2}} \operatorname{tr} \exp\left(\sum_{\ell=1}^{M-2} \mathbf{X}_\ell + \mathbf{Z}_M + \mathbf{Z}_{M-1}\right) \\ \cdots &\leq \operatorname{tr} \exp\left(\sum_{\ell=1}^{M} \mathbf{Z}_\ell\right) \ .\end{aligned}$$

The application of Proposition 8.17 at step $k \in [M]$ with the matrices

$$\mathbf{H}_k = \sum_{\ell=1}^{k-1} \mathbf{X}_\ell + \sum_{\ell=k+1}^{M} \mathbf{Z}_\ell$$

is permitted since $\mathbf{H}_k$ does not depend on $\mathbf{X}_k$. □

Before giving the next intermediate result, we recall that a self-adjoint square matrix $\mathbf{A} \in \mathbb{C}^{d\times d}$ is called positive semidefinite if $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \geq 0$ for all $\mathbf{x} \in \mathbb{C}^d$. Equivalently, all eigenvalues of a positive semidefinite matrix $\mathbf{A}$ are non-negative. Furthermore, we write $\mathbf{A} \preccurlyeq \mathbf{B}$ for two self-adjoint matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{d\times d}$ if $\mathbf{B} - \mathbf{A}$ is positive semidefinite.

Next we provide a matrix version of Cramér's Theorem 7.18.

Proposition 8.19. *Let $\mathbf{X}_1, \dots, \mathbf{X}_M \in \mathbb{C}^{d\times d}$ be independent, self-adjoint random matrices. Assume that there exist a function $g : (0,\infty) \to [0,\infty)$ and fixed self-adjoint matrices $\mathbf{A}_1, \dots, \mathbf{A}_k$ such that*

$$\mathbb{E}\exp(\theta \mathbf{X}_k) \preccurlyeq \exp(g(\theta)\mathbf{A}_k), \quad \text{for all } \theta > 0\,. \tag{8.35}$$

Then, with $\rho := \lambda_{\max}\left(\sum_{\ell=1}^M \mathbf{A}_\ell\right)$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^M \mathbf{X}_\ell\right) \geq t\right) \leq d \inf_{\theta>0} e^{-\theta t + g(\theta)\rho}\,, \quad t \in \mathbb{R}\,.$$

Proof. Plugging (8.34) into (8.31) yields

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^M \mathbf{X}_\ell\right) \geq t\right) \leq \inf_{\theta>0}\left\{e^{-\theta t}\mathrm{tr}\,\exp\left(\sum_{\ell=1}^M \ln \mathbb{E}\exp(\theta \mathbf{X}_\ell)\right)\right\}\,.$$

By Proposition A.35, the matrix logarithm is matrix monotone, so that (8.35) implies

$$\ln \mathbb{E}\exp(\theta \mathbf{X}_\ell) \preccurlyeq g(\theta)\mathbf{A}_\ell \quad \text{for all } \theta > 0\,.$$

Since the trace exponential is monotone, see (A.48), a combination of the above facts yields, for each $\theta > 0$,

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^M \mathbf{X}_\ell\right) \geq t\right) &\leq e^{-\theta t}\mathrm{tr}\,\exp\left(g(\theta)\sum_{\ell=1}^M \mathbf{A}_\ell\right)\\ &\leq e^{-\theta t} d\,\lambda_{\max}\left(\exp\left(g(\theta)\sum_{\ell=1}^M \mathbf{A}_\ell\right)\right) = d\,e^{-\theta t}\exp\left(g(\theta)\lambda_{\max}\left(\sum_{\ell=1}^M \mathbf{A}_\ell\right)\right)\,.\end{aligned}$$

The second inequality is valid because, for a positive definite $d \times d$ matrix $\mathbf{B}$, we have $\mathrm{tr}\,\mathbf{B} = \sum_{j=1}^d \lambda_j(\mathbf{B}) \leq d\,\lambda_{\max}(\mathbf{B})$, where $\lambda_j(\mathbf{B})$, $j \in [d]$, denote the eigenvalues of $\mathbf{B}$ (with possible repetitions). Taking the infimum over all positive θ and using the definition of ρ, we arrive at the statement of the proposition. □

Before we pass to the proof of the noncommutative Bernstein inequality, we note the following deviation inequality for matrix-valued Rademacher sums, i.e., the matrix-valued analog of Hoeffding's inequality for scalar Rademacher sums in Corollaries 7.21 and 8.8.

Proposition 8.20. *Let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ be a Rademacher sequence, and $\mathbf{B}_1, \dots, \mathbf{B}_M \in \mathbb{C}^{d\times d}$ be self-adjoint matrices. Set*

$$\sigma^2 := \|\sum_{\ell=1}^{M} \mathbf{B}_\ell^2\|_{2\to 2} \, .$$

Then, for $t > 0$,

$$\mathbb{P}\big(\|\sum_{\ell=1}^{M} \epsilon_\ell \mathbf{B}_\ell\|_{2\to 2} \geq t\big) \leq 2d \exp\big(-t^2/(2\sigma^2)\big) \, . \tag{8.36}$$

Proof. Proposition 8.19 requires to estimate $\mathbb{E}\exp(\theta\epsilon\mathbf{B})$ for a Rademacher variable ϵ and a self-adjoint matrix $\mathbf{B}$. Similarly to the scalar case in (7.29) we get

$$\begin{aligned}\mathbb{E}\exp(\theta\epsilon\mathbf{B}) &= \frac{1}{2}\big(\exp(\theta\mathbf{B}) + \exp(-\theta\mathbf{B})\big) = \sum_{k=0}^{\infty} \frac{(\theta\mathbf{B})^{2k}}{(2k)!} \\ &\preccurlyeq \sum_{k=0}^{\infty} \frac{(\theta\mathbf{B})^{2k}}{2^k k!} = \exp(\theta^2\mathbf{B}^2/2) \, ,\end{aligned}$$

because $\mathbf{B}^2$ is positive semidefinite. Therefore, (8.35) holds with $g(\theta) = \theta^2/2$ and $\mathbf{A}_\ell = \mathbf{B}_\ell^2$. The paramter ρ in Proposition 8.19 is given by

$$\rho = \|\sum_{\ell=1}^{M} \mathbf{B}_\ell^2\|_{2\to 2} = \sigma^2$$

because $\sum_{\ell=1}^{M} \mathbf{B}_\ell^2$ is positive semidefinite. Therefore,

$$\begin{aligned}&\mathbb{P}\big(\|\sum_{\ell=1}^{M} \epsilon_\ell \mathbf{B}_\ell\|_{2\to 2} \geq t\big) \leq \mathbb{P}\big(\lambda_{\max}(\sum_{\ell=1}^{M} \epsilon_\ell \mathbf{B}_\ell\big) \geq t) + \mathbb{P}\big(\lambda_{\max}(-\sum_{\ell=1}^{M} \epsilon_\ell \mathbf{B}_\ell) \geq t\big) \\ &\leq 2d \inf_{\theta>0} e^{-\theta t + \theta^2\sigma^2/2} = 2d\, e^{-t^2/(2\sigma^2)} \, .\end{aligned}$$

The optimal choice of θ above was $\theta = t/\sigma^2$. □

The case $d = 1$ reduces to the Hoeffding type inequality of Corollary 8.8. The same deviation inequality holds also for matrix-valued Gaussian sums, see Exercise 8.6.

Proof (of Theorem 8.14). Proposition 8.19 requires to establish (8.35) for an appropriate function g and appropriate matrices $\mathbf{A}_k$. We may assume that the bound K on the maximal eigenvalue of $\mathbf{X}_\ell$, $\ell \in [M]$, satisfies $K = 1$. The general case follows then from applying the result to the rescaled matrices $\tilde{\mathbf{X}}_\ell = \mathbf{X}_\ell/K$.

We fix $\theta > 0$, and define the smooth function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = x^{-2}(e^{\theta x} - \theta x - 1) \quad \text{for } x \neq 0, \quad \text{and } f(0) = \theta^2/2 \,.$$

Clearly, $f(x) = \theta^2 \sum_{k=2}^{\infty} \frac{(\theta x)^{k-2}}{k!}$. The derivative is given by

$$f'(x) = \theta^2 \sum_{k=3}^{\infty} \frac{\theta^{k-2}(k-2)x^{k-3}}{k!} = \frac{(\theta x - 2)e^{\theta x} + (\theta x + 2)}{x^3} \,.$$

We claim that $f'(x) \geq 0$ for all $x \in \mathbb{R}$, so that f is nondecreasing. Indeed, for $x \geq 0$ this follows from the power series expansion of f' as all coefficients are positive. For $x \in (-2/\theta, 0)$ one verifies that the absolute values $\theta^{k-2}(k-2)|x|^{k-3}/k!$, $k \geq 3$, of the terms in the power series of f' are monotonically decreasing in k, and the term for $k = 3$ is positive. Since the power series is alternating, $f'(x) \geq 0$ holds also in this case. For $x \leq -2/\theta$ the nonnegativity of f' follows from the explicit formula above, where both the nominator and denominator are easily seen to be negative.

In particular, we have proven that $f(x) \leq f(1)$ whenever $x \leq 1$. All the eigenvalues of $\mathbf{X}_\ell$ are bounded by 1, so by the definition of the extension of f to matrices (A.42) and by the rule (A.43), it follows that

$$f(\mathbf{X}_\ell) \preccurlyeq f(1)\mathbf{Id} \,.$$

The identity $\exp(\theta x) = 1 + \theta x + x^2 f(x)$ and the fact that $f(\mathbf{X})$ commutes with $\mathbf{X}$ yield together with (A.43) that

$$\exp(\theta \mathbf{X}_\ell) = \mathbf{Id} + \theta \mathbf{X}_\ell + \mathbf{X}_\ell f(\mathbf{X}_\ell)\mathbf{X}_\ell \preccurlyeq \mathbf{Id} + \theta \mathbf{X}_\ell + f(1)\mathbf{X}_\ell^2 \,.$$

Hereby, we used additionally the elementary fact that $\mathbf{A} \preccurlyeq \mathbf{B}$ implies $\mathbf{HAH}^* \preccurlyeq \mathbf{HBH}^*$ for any matrix $\mathbf{H}$ of matching dimension (Lemma A.32), together with the self-adjointness of $\mathbf{X}_\ell$.

Taking expectations in the above semidefinite bound, and using $\mathbb{E}\mathbf{X}_k = 0$ we obtain

$$\mathbb{E}\exp(\theta \mathbf{X}_\ell) \preccurlyeq \mathbf{Id} + f(1)\mathbb{E}\mathbf{X}_\ell^2 \preccurlyeq \exp(f(1)\mathbb{E}\mathbf{X}_\ell^2) = \exp\left((e^\theta - \theta - 1)\mathbb{E}\mathbf{X}_\ell^2\right) \,.$$

The second semidefinite bound follows from the general bound (A.46) for the matrix exponential. Setting $g(\theta) = e^\theta - \theta - 1$, it follows from Proposition 8.19 that, for $t \in \mathbb{R}$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^{M} \mathbf{X}_\ell\right) \geq t\right) \leq d \inf_{\theta > 0} \left\{e^{-\theta t + g(\theta)\sigma^2}\right\} \,, \tag{8.37}$$

where we have used that $\lambda_{\max}\left(\sum_{\ell=1}^{M} \mathbb{E}\mathbf{X}_\ell^2\right) = \sigma^2$ by positive semidefiniteness of $\sum_{\ell=1}^{M} \mathbb{E}\mathbf{X}_\ell^2$. Then both the Bennett type inequality (8.25) and the Bernstein type inequality (8.26) follow from Lemma (8.21) below. □

Lemma 8.21. *Let* $h(x) := (1+x)\ln(1+x) - x$ *be the function defined in* (8.23) *and* $g(\theta) = e^\theta - \theta - 1$. *Then, for* $a > 0$,

$$\inf_{\theta>0}\{-\theta x + g(\theta)a\} = -ah(x/a)\,, \quad x \geq 0\,,$$

and

$$h(x) \geq \frac{x^2/2}{1+x/3} \quad \textit{for all } x \geq 0\,.$$

Proof. The function $r(\theta) := g(\theta)a - \theta x$ attains its minimal value for $\theta = \ln(1+x/a)$, and

$$r(\ln(1+x/a)) = (x/a - \ln(1+x/a))a - x\ln(1+x/a) = -ah(x/a).$$

For the second statement we first note that

$$g(\theta) = e^\theta - \theta - 1 = \sum_{k=2}^{\infty}\frac{\theta^k}{k!} = \frac{\theta^2}{2}\sum_{k=2}^{\infty}\frac{2\theta^{k-2}}{k!}\,.$$

By induction, it follows that $2/k! \leq (1/3)^{k-2}$ for all $k \geq 2$. Therefore, for $\theta < 3$,

$$g(\theta) \leq \frac{\theta^2}{2}\sum_{k=0}^{\infty}(\theta/3)^k = \frac{\theta^2/2}{1-\theta/3}\,.$$

Making the specific choice $\theta = \dfrac{x}{1+x/3} < 3$ below shows that

$$\begin{aligned} -h(x) = \inf_{\theta>0}(g(\theta)-\theta x) &\leq \inf_{\theta\in(0,3)}\left(\frac{\theta^2/2}{1-\theta/3} - \theta x\right) \\ &\leq \frac{x^2/2}{(1+x/3)^2\left(1-\frac{x/3}{1+x/3}\right)} - \frac{x^2}{1+x/3} = -\frac{x^2/2}{1+x/3}\,. \end{aligned} \tag{8.38}$$

This point completes the proof. □

8.6 Dudley's Inequality

A stochastic process is a collection X_t, $t \in T$, of random variables indexed by some set T. We are interested in bounding the expectation of the supremum of a real-valued stochastic process. In order to avoid measurability issues (in general, the supremum of an uncountable number of random variables might not be measurable) we define the so called lattice supremum

$$\mathbb{E}\sup_{t\in T} X_t := \sup\{\mathbb{E}\sup_{t\in F} X_t, F \subset T, F \text{ finite}\}\,. \tag{8.39}$$

Note that for a countable index set T, where no measurability problems can arise, $\mathbb{E}(\sup_{t\in T} X_t)$ equals the right hand side above (see Exercise 8.9), so that this definition is consistent. Also, if $t \mapsto X_t$ is continuous on T for each realization of X_t (as will always be the case in the situations we encounter), and T is separable, then $\sup_{t\in T} X_t$ coincides with the supremum over a dense countable subset of T, so that in this case the lattice supremum coincides with $\mathbb{E}(\sup_{t\in T} X_t)$ as well.

We always assume that the process is centered, that is,

$$\mathbb{E}X_t = 0 \quad \text{for all } t \in T \,. \tag{8.40}$$

Associated to the process X_t, $t \in T$, we define the pseudo-metric

$$d(s,t) := \left(\mathbb{E}|X_s - X_t|^2\right)^{1/2} \,, \quad s,t \in T \,. \tag{8.41}$$

We refer to Definition A.2 for the notion of pseudo-metric.

Definition 8.22. *A centered stochastic process X_t, $t \in T$, is called subgaussian if*

$$\mathbb{E}\exp(\theta(X_s - X_t)) \le \exp(\theta^2 d(s,t)^2/2) \,, \quad s,t \in T \,, \theta > 0 \,, \tag{8.42}$$

with d being the pseudo-metric defined in (8.41).

Clearly, one may replace the constant $1/2$ in (8.42) by a general constant c, but for our purposes it is enough to consider $c = 1/2$.

Examples of subgaussian processes include Gaussian and Rademacher processes. A process X_t is called centered Gaussian process if for each finite collection $t_1, \dots, t_n \in T$ the random vector $(X_{t_1}, \dots, X_{t_n})$ is a mean zero Gaussian random vector. This implies in particular that $X_t - X_s$ is a univariate Gaussian with $\mathbb{E}(X_t - X_s) = 0$ by (8.40) and variance $\mathbb{E}|X_s - X_t|^2$. It follows from (7.11) (or Remark 7.26 and Theorem 7.27) that a Gaussian process is a subgaussian process in the sense of Definition 8.22. A typical example of a Gaussian process takes the form

$$X_t = \sum_{j=1}^{M} g_j x_j(t) \,,$$

where $\mathbf{g} = (g_1, \dots, g_M)$ is a standard Gaussian random vector and $x_j : T \to \mathbb{R}$, $j \in [M]$, are some functions.

A Rademacher process has the form

$$X_t = \sum_{j=1}^{M} \epsilon_j x_j(t) \,, \tag{8.43}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ is a Rademacher sequence. Clearly, such a process satisfies (8.40). By Remark 7.26 and Theorem 7.27, it is a subgaussian process. Observe that

$$\mathbb{E}|X_t - X_s|^2 = \mathbb{E}|\sum_{j=1}^{M} \epsilon_j (x_j(t) - x_j(s))|^2 = \sum_{j=1}^{M} (x_j(t) - x_j(s))^2 = \|x(t) - x(s)\|_2^2,$$

where $x(t)$ denotes the vector with components $x_j(t), j \in [M]$. Therefore, the pseudo-metric associated to X_t is given by

$$d(s,t) = \left(\mathbb{E}|X_t - X_s|^2\right)^{1/2} = \|x(t) - x(s)\|_2 \,. \tag{8.44}$$

It follows from Theorem 7.27 that the increments of a subgaussian process X_t satisfy the tail estimate

$$\mathbb{P}(|X_s - X_t| \geq u d(s,t)) \leq 2\exp(-u^2/2) \,. \tag{8.45}$$

By Proposition (7.24)(b) this inequality could as well by taken for the definition of subgaussian processes.

Dudley's inequality below relates the stochastic quantity of the lattice supremum (8.39) to the geometric concept of covering numbers. We recall from Section C.2 that the covering number $N(T,d,\varepsilon)$ is defined as the smallest integer N such that there exists a subset F of T with $\mathrm{card}(F) = N$ and $\min_{s\in F} d(t,s) \leq \varepsilon$ for all $t \in T$. We denote the diameter of T by

$$\Delta(T) = \sup_{s,t\in T} d(s,t) \,. \tag{8.46}$$

Dudley's inequality for subgaussian processes reads as follows.

Theorem 8.23. *Let X_t, $t \in T$, be a centered subgaussian processes with associated pseudo-metric d. Then, for any $t_0 \in T$,*

$$\mathbb{E}\sup_{t\in T} X_t \leq 12 \int_0^{\Delta(T)/2} \sqrt{\ln(N(T,d,u))}du \,, \tag{8.47}$$

$$\mathbb{E}\sup_{t\in T} |X_t - X_{t_0}| \leq 12 \int_0^{\Delta(T)/2} \sqrt{\ln(\sqrt{2}N(T,d,u))}du \,. \tag{8.48}$$

Remark 8.24. Inequality (8.48) with constant 24 (but without the factor $\sqrt{2}$ inside the logarithm) follows also directly from (8.47) in the case of symmetric processes, see Exercise 8.10. It is known that these inequalities are sharp up to log-factors if T is a subset of a finite-dimensional space, and d is induced by a norm, see also the notes section.

Proof. We write $\Delta = \Delta(T)$ for convenience. Let F be a finite subset of T. We set $\varepsilon_n := 2^{-n}\Delta$ and $N_n := N(T,d,\varepsilon_n)$. By definition of the covering numbers, we can find subsets $T_n \subset T$ of cardinality at most N_n such that for all $t \in F \subset T$ there exists $s \in T_n$ such that $d(t,s) \leq \varepsilon_n$. We write $s = \phi_n(t)$ for this particular s and set $X_t^n := X_{\phi_n(t)}$. Note that T_0 consists only of one point by definition of the diameter, and we can choose $\phi_0(t) = t_0$, so that $X_t^0 = X_{t_0}$ for all $t \in T$. By construction

$$\begin{aligned}\big(\mathbb{E}(X_t^n - X_t^{n-1})^2\big)^{1/2} &= d(\phi_n(t), \phi_{n-1}(t)) \le d(\phi_n(t), t) + d(t, \phi_{n-1}(t)) \\ &\le (2^{-n} + 2^{-(n-1)})\Delta = 3 \cdot 2^{-n}\Delta \,. \end{aligned} \tag{8.49}$$

We claim that the following *chaining identity* holds almost surely,

$$X_t = X_t^0 + \sum_{n=1}^{\infty} (X_t^n - X_t^{n-1}) \,. \tag{8.50}$$

Indeed, by (8.45) we have

$$\begin{aligned}&\mathbb{P}(|X_t^n - X_t^{n-1}| \ge 2^{-n/2}) \\ &\le \mathbb{P}\left(|X_t^n - X_t^{n-1}| \ge \frac{2^{n/2}}{3} d(\phi_n(t), \phi_{n-1}(t))\right) \le 2\exp\left(-\frac{1}{18}2^n\right).\end{aligned}$$

This implies that $\sum_{n=1}^{\infty} \mathbb{P}(|X_t^n - X_t^{n-1}| \ge 2^{-n/2}) < \infty$. It follows from the Borel-Cantelli Lemma 7.10 that for almost all $\omega \in \Omega$ there exists $n_0(\omega)$ such that for all $n \ge n_0(\omega)$ we have $|X_t^n - X_t^{n-1}| < 2^{-n/2}$. Consequently, the series on the right hand side of (8.50) converges almost surely. Therefore,

$$\begin{aligned}\sup_{t\in F} X_t &\le X_{t_0} + \sum_{n\ge 1} \sup_{t\in F} (X_t^n - X_t^{n-1}) \\ \sup_{t\in F} |X_t - X_{t_0}| &\le \sum_{n\ge 1} \sup_{t\in F} |X_t^n - X_t^{n-1}| \,.\end{aligned}$$

Since X_t is centered, that is, $\mathbb{E}X_{t_0} = 0$, we obtain

$$\begin{aligned}\mathbb{E}\sup_{t\in F} X_t &\le \sum_{n\ge 1} \mathbb{E}\sup_{t\in F} (X_t^n - X_t^{n-1}) \,, \\ \mathbb{E}\sup_{t\in F} |X_t - X_{t_0}| &\le \sum_{n\ge 1} \mathbb{E}\sup_{t\in F} |X_t^n - X_t^{n-1}| \,.\end{aligned}$$

Observe that $X_t^n - X_t^{n-1}$ is a subgaussian random variable. Further, by definition of the pseudo-metric and by (8.49)

$$\big(\mathbb{E}(X_t^n - X_t^{n-1})^2\big)^{1/2} \le 3 \cdot 2^{-n}\Delta \,.$$

Note that $\sup_{t\in F}(X_t^n - X_t^{n-1})$ is the supremum over at most $N_n \cdot N_{n-1}$ subgaussian random variables satisfying $\mathbb{E}\exp(\theta(X_t^n - X_t^{n-1})) \le \exp((3 \cdot 2^{-n}\Delta)^2\theta^2/2)$. It follows from Proposition 7.29 and (8.49) that

$$\begin{aligned}\mathbb{E}\sup_{t\in F}(X_t^n - X_t^{n-1}) &\le 3 \cdot 2^{-n}\Delta\sqrt{2\ln(N_n \cdot N_{n-1})} \\ &\le 3 \cdot 2^{-n}\Delta\sqrt{2\ln(N_n^2)} = 12 \cdot 2^{-n-1}\Delta\sqrt{\ln(N_n)} \,,\end{aligned}$$

where we have used that $N_{n-1} \le N_n$ by elementary properties of the covering numbers. Similarly, we get from (7.38) that

$$\mathbb{E}\sup_{t\in F}|X_t^n - X_t^{n-1}| \le 3\cdot 2^{-n}\Delta\sqrt{2\ln(2N_{n-1}N_n)}$$
$$= 12\cdot 2^{-n-1}\Delta\sqrt{\ln(\sqrt{2}N_n)}\,. \tag{8.51}$$

We finally obtain

$$\mathbb{E}\sup_{t\in F} X_t \le \sum_{n\ge 1} 12\cdot 2^{-n-1}\Delta\sqrt{\ln(N(T,d,2^{-n}\Delta)}$$
$$\le 12\sum_{n\ge 1}\int_{2^{-n-1}\Delta}^{2^{-n}\Delta}\sqrt{\ln(N(T,d,u\cdot\Delta))}du$$
$$= 12\int_0^{\Delta/2}\sqrt{\ln(N(T,d,u))}du\,.$$

Hereby, we have applied that $N(T,d,2^{-n}\Delta) \le N(T,d,u\cdot\Delta)$ for all $u\in[2^{-n-1},2^{-n}]$. Taking the supremum over all finite subsets of T completes the proof of (8.47) by definition of the lattice supremum in (8.39). Inequality (8.48) follows in the same way from (8.51). □

Remark 8.25. If $2N(T,d,2t)\le N(T,d,t)$ for all $t\le\Delta(T)$ then (8.51) can be improved to

$$\mathbb{E}\sup_{t\in F}|X_t^n - X_t^{n-1}| \le 3\cdot 2^{-n+1}\Delta(T)\sqrt{\ln(N_n)}\,,$$

and consequently the factor $\sqrt{2}$ can be removed from (8.48).

8.7 Slepian and Gordon Lemmas

The Slepian lemma and its generalization due to Gordon compare extrema of two families of Gaussian random variables. The basic idea is that the distribution of a mean-zero Gaussian vector is completely determined by its covariance structure. This suggests to compare expectations of functions of the two families by means of comparing the covariances.

Slepian's lemma reads as follows.

Lemma 8.26. *Let* $\mathbf{X},\mathbf{Y}$ *be mean-zero Gaussian random vectors on* $\mathbb{R}^m$*. If*

$$\mathbb{E}|X_i - X_j|^2 \le \mathbb{E}|Y_i - Y_j|^2 \quad \textit{for all } i,j\in[m]\,, \tag{8.52}$$

then

$$\mathbb{E}\max_{j\in[m]} X_j \le \mathbb{E}\max_{j\in[m]} Y_j\,.$$

Remark 8.27. The L_2-distances above can be written in terms of the covariances,

$$\mathbb{E}|X_i - X_j|^2 = \mathbb{E}X_i^2 - 2\mathbb{E}X_iX_j + \mathbb{E}X_j^2 \ .$$

Under the additional assumption that $\mathbb{E}X_j^2 = \mathbb{E}Y_j^2$, condition (8.52) reads therefore $\mathbb{E}X_jX_k \le \mathbb{E}Y_jY_k$. In particular, comparison of the covariance structures of $\mathbf{X}$ and $\mathbf{Y}$ allows to compare the expected maxima of the two Gaussian vectors as claimed above.

Gordon's lemma stated next compares expected minima of maxima of Gaussian vectors. Slepian's lemma is the special case $n = 1$.

Lemma 8.28. *Let $X_{i,j}$, $Y_{i,j}$, $i \in [n], j \in [m]$, be two finite families of mean-zero Gaussian random variables. If*

$$\mathbb{E}|X_{i,j} - X_{k,\ell}|^2 \le \mathbb{E}|Y_{i,j} - Y_{k,\ell}|^2 \text{ for all } i \ne k \text{ and } j, \ell \ , \tag{8.53}$$

$$\mathbb{E}|X_{i,j} - X_{i,\ell}|^2 \ge \mathbb{E}|Y_{i,j} - Y_{i,\ell}|^2 \text{ for } i, j, \ell \ , \tag{8.54}$$

then

$$\mathbb{E}\min_{i\in[n]}\max_{j\in[m]} X_{i,j} \ge \mathbb{E}\min_{i\in[n]}\max_{j\in[m]} Y_{i,j} \ .$$

Remark 8.29. Both Slepian's and Gordon's lemma extend to Gaussian processes indexed by possibly infinite sets. In particular, if $\mathbf{X} = (X_t)_{t\in T}, \mathbf{Y} = (Y_t)_{t\in T}$, are Gaussian processes (which by definition means that any restriction $\mathbf{X}_{T_0} = (X_t)_{t\in T_0}$ to a finite subset $T_0 \subset T$ yields a Gaussian random vector) and if $\mathbb{E}|X_s - X_t|^2 \le \mathbb{E}|Y_s - Y_t|^2$ for all $s, t \in T$, then Slepian's lemma states that

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t \ ,$$

where the supremum is understood in the sense of a lattice supremum (8.39). Indeed, by the finite-dimensional version in Lemma 8.26 this relation holds for the restriction to any finite subset T_0 so that the above inequality holds.

In a similar sense, Gordon's lemma extends to doubly indexed Gaussian processes.

The proof of Slepian and Gordon's lemma requires some preparation. We say that a function $F : \mathbb{R}^m \to \mathbb{R}$ is of *moderate growth* if for each $\beta > 0$

$$\lim_{\|\mathbf{x}\|_2\to\infty} F(\mathbf{x})\exp(-\beta\|\mathbf{x}\|_2^2) = 0 \ . \tag{8.55}$$

Our first technical tool is the Gaussian integration by parts formula and its generalization to higher dimensions.

Proposition 8.30. *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function such that F together with its partial derivative is of moderate growth.*

(a) Let g be a mean-zero Gaussian random variable and $m = 1$. Then

$$\mathbb{E}[gF(g)] = \mathbb{E}g^2 \mathbb{E}F'(g) \,. \tag{8.56}$$

(b) Let $\mathbf{g} = (g_1, \ldots, g_m)$ be a Gaussian random vector and $\tilde{g}$ be a Gaussian random variable (not necessarily independent of $\mathbf{g}$). Then

$$\mathbb{E}\tilde{g}F(\mathbf{g}) = \sum_{j=1}^{m} \mathbb{E}(\tilde{g}g_j)\mathbb{E}\left[\frac{\partial F}{\partial x_j}(\mathbf{g})\right] \,. \tag{8.57}$$

Proof. (a) Setting $\tau^2 = \mathbb{E}g^2$ and using integration by parts yields

$$\begin{aligned}\mathbb{E}gF(g) &= \frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty} t\exp(-t^2/(2\tau^2))F(t)dt \\ &= \frac{\tau^2}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty} \exp(-t^2/(2\tau^2))F'(t)dt = \mathbb{E}g^2\mathbb{E}F'(g) \,.\end{aligned}$$

The moderate growth condition ensures that all integrals are well-defined and that $\exp(-\tau^2/2)F(t)\big|_{-\infty}^{\infty} = 0$

(b) Consider the random variables $g_j' = g_j - \tilde{g}\frac{\mathbb{E}g_j\tilde{g}}{\mathbb{E}\tilde{g}^2}$. They satisfy $\mathbb{E}g_j'\tilde{g} = 0$, and therefore, the Gaussian random vector $\mathbf{g}' = (g_1', \ldots, g_m')$ is independent of $\tilde{g}$. (In fact, in the Gaussian case, independence follows from the random variables being uncorrelated.) Using Fubini's theorem and applying (8.56) conditional on $\mathbf{g}'$ yields

$$\begin{aligned}\mathbb{E}\tilde{g}F(\mathbf{g}) &= \mathbb{E}\tilde{g}F\left(g_1' + \tilde{g}\frac{\mathbb{E}\tilde{g}g_1}{\mathbb{E}\tilde{g}^2}, \ldots, g_m' + \tilde{g}\frac{\mathbb{E}\tilde{g}g_m}{\mathbb{E}\tilde{g}^2}\right) \\ &= \mathbb{E}\tilde{g}^2\sum_{j=1}^{m}\frac{\mathbb{E}\tilde{g}g_1}{\mathbb{E}\tilde{g}^2}\mathbb{E}\frac{\partial F}{\partial x_j}\left(g_1' + \tilde{g}\frac{\mathbb{E}\tilde{g}g_1}{\mathbb{E}\tilde{g}^2}, \ldots, g_m' + \tilde{g}\frac{\mathbb{E}\tilde{g}g_m}{\mathbb{E}\tilde{g}^2}\right) \\ &= \sum_{j=1}^{m}\mathbb{E}[\tilde{g}g_j]\mathbb{E}\left[\frac{\partial F}{\partial x_j}(\mathbf{g})\right] \,.\end{aligned}$$

This completes the proof. □

We will also require the following standard result in integration theory.

Proposition 8.31. *Let $\psi : J \times \Omega \to \mathbb{R}$ be a (random) function on an open interval $J \subset \mathbb{R}$. Let X be a random variable (or vector) such that $t \mapsto \psi(t, X)$ is continuously differentiable in J for each realization of X. Assume that for each compact subinterval $I \subset J$*

$$\mathbb{E}\sup_{t\in I}|\psi'(t, X)| < \infty \,. \tag{8.58}$$

Then the function $t \mapsto \phi(t) = \mathbb{E}\psi(t, X)$ is continuously differentiable and

$$\phi'(t) = \mathbb{E}\psi'(t, X). \tag{8.59}$$

Proof. Let t be in the interior of J and consider a compact subinterval $I \subset J$ containing t in its interior. For $h \in \mathbb{R} \setminus \{0\}$ such that $t + h \in I$ we consider the difference quotients

$$\phi_h(t) := \frac{\phi(t+h) - \phi(t)}{h} , \quad \psi_h(t, X) := \frac{\psi(t+h, X) - \psi(t, X)}{h} .$$

By the mean value theorem there exists $\xi \in [t, t+h]$ such that $\psi'(\xi, X) = \psi_h(t, X)$. Therefore, $|\psi_h(t, X)| \le \sup_{t \in I} |\psi'(t, X)|$ and by (8.58) $\psi_h(t, X)$ has an integrable majorant. By Lebesgue's dominated convergence theorem we have

$$\lim_{h \to 0} \phi_h(t) = \mathbb{E} \lim_{h \to 0} \psi_h(t, X) = \mathbb{E}\psi'(t, X),$$

so that ϕ is continuously differentiable and (8.59) holds. □

The crucial tool in the proof of Slepian's and Gordon's lemma is stated next.

Proposition 8.32. *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function such that F together with all its partial derivatives of first order are of moderate growth. Let $\mathbf{X} = (X_1, \dots, X_m)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ be two independent mean-zero Gaussian vectors. For $t \in [0, 1]$ we define the new random vector $\mathbf{U}(t) = (U_1(t), \dots, U_m(t))$ with components*

$$U_i(t) = \sqrt{t} X_i + \sqrt{1-t} Y_i, \quad i = 1, \dots, m . \tag{8.60}$$

Then the function

$$\phi(t) = \mathbb{E}F(\mathbf{U}(t))$$

has derivative

$$\phi'(t) = \sum_{i=1}^{m} \mathbb{E}\left[U_i'(t) \frac{\partial F}{\partial x_i}(\mathbf{U}(t)) \right] . \tag{8.61}$$

If, in addition, F is twice differentiable with all partial derivatives of second order of moderate growth then

$$\phi'(t) = \frac{1}{2} \sum_{i,j=1}^{m} (\mathbb{E}X_i X_j - \mathbb{E}Y_i Y_j) \mathbb{E}\left[\frac{\partial^2 F}{\partial x_i \partial x_j}(\mathbf{U}(t)) \right] . \tag{8.62}$$

Proof. We note that

$$\frac{d}{dt} F(\mathbf{U}(t)) = \sum_{i=1}^{m} U_i'(t) \frac{\partial F}{\partial x_i}(\mathbf{U}(t)) ,$$

where clearly

$$U_i'(t) = \frac{d}{dt} U_i(t) = \frac{1}{2\sqrt{t}} X_i - \frac{1}{2\sqrt{1-t}} Y_i .$$

By Proposition 8.31 it therefore suffices to verify (8.58). For a compact subinterval $I = [a, b] \subset (0, 1)$, we have

$$\begin{aligned}\mathbb{E} \sup_{t\in I} |U_i'(t) \frac{\partial F}{\partial x_i}(\mathbf{U}(t))| &\le \mathbb{E} \sup_{t\in I} |U_i'(t)| \sup_{t\in I} |\frac{\partial F}{\partial x_i}(\mathbf{U}(t))| \\ &\le \sqrt{\mathbb{E} \sup_{t\in I} |U_i'(t)|^2} \sqrt{\mathbb{E} \sup_{t\in I} |\frac{\partial F}{\partial x_i}(\mathbf{U}(t))|^2} \,,\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality. We treat both expectations above separately. The triangle inequality gives

$$\sqrt{\mathbb{E} \sup_{t\in I} |U_i'(t)|^2} \le \sqrt{\mathbb{E} \frac{1}{4a} X_i^2} + \sqrt{\mathbb{E} \frac{1}{4(1-b)} Y_i^2} < \infty \,.$$

For the second expectation choose $\beta > 0$. Since $\frac{\partial F}{\partial x_i}$ is of moderate growth there exists $A > 0$ such that

$$\left| \frac{\partial F}{\partial x_i}(x) \right| \le A \exp(\beta \|x\|_2^2) \quad \text{for all } x \in \mathbb{R}^m \,.$$

Furthermore,

$$\|\mathbf{U}(t)\|_2 \le \sqrt{t} \|\mathbf{X}\|_2 + \sqrt{1-t} \|\mathbf{Y}\|_2 \le 2 \max\{\|\mathbf{X}\|_2, \|\mathbf{Y}\|_2\} \,,$$

and hence,

$$\sup_{t\in I} \left| \frac{\partial F}{\partial x_i}(\mathbf{U}(t)) \right| \le A \max\{\exp(4\beta \|\mathbf{X}\|_2^2), \exp(4\beta \|\mathbf{Y}\|_2^2)\} \,.$$

Since $\mathbf{X}$ and $\mathbf{Y}$ are mean-zero Gaussian vectors, there exist matrices $\mathbf{\Gamma}, \mathbf{\Gamma}'$ such that $\mathbf{X} = \mathbf{\Gamma}\mathbf{g}$ and $\mathbf{Y} = \mathbf{\Gamma}'\mathbf{g}'$ where $\mathbf{g}, \mathbf{g}'$ are independent standard Gaussian vectors. Therefore,

$$\begin{aligned}&\mathbb{E} \sup_{t\in I} \left| \frac{\partial F}{\partial x_i}(\mathbf{U}(t)) \right| \le A \mathbb{E} \left[\exp\left(4\beta \|\mathbf{\Gamma}\|_{2\to 2}^2 \|\mathbf{g}\|_2^2 + 4\beta \|\mathbf{\Gamma}'\|_{2\to 2}^2 \|\mathbf{g}'\|_2^2 \right) \right] \\ &= A \prod_{i=1}^m \mathbb{E} \left[\exp(4\beta \|\mathbf{\Gamma}\|_{2\to 2}^2 g_i^2) \right] \prod_{j=1}^m \mathbb{E} \left[\exp(4\beta \|\mathbf{\Gamma}'\|_{2\to 2}^2 (g_j')^2) \right] \\ &= A (1 - 8\beta \|\mathbf{\Gamma}\|_{2\to 2}^2)^{-m/2} (1 - 8\beta \|\mathbf{\Gamma}'\|_{2\to 2}^2)^{-m/2} < \infty \,.\end{aligned}$$

The last equality follows from Lemma 7.6 with $\theta = 0$ and a choice of $\beta > 0$ such that $8\beta \max\{\|\mathbf{\Gamma}\|_{2\to 2}^2, \|\mathbf{\Gamma}'\|_{2\to 2}^2\} < 1$. (Recall that $\beta > 0$ can be chosen arbitrarily and influences only the constant A). This completes the proof of (8.61).

For (8.62) we observe that $\mathbb{E} U_i'(t) U_j(t) = \frac{1}{2}(\mathbb{E} X_i X_j - \mathbb{E} Y_i Y_j)$. The Gaussian integration by parts formula (8.57) yields

$$\mathbb{E}\left[U_i'(t)\frac{\partial F}{\partial x_i}(\mathbf{U}(t))\right] = \frac{1}{2}\sum_{j=1}^{m}(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j)\mathbb{E}\frac{\partial^2 F}{\partial x_i \partial x_j}(\mathbf{U}(t)) \ .$$

This completes the proof. □

The next result is a generalized version of Gordon's lemma. Since we will require it also for not necessarily differentiable functions F, we work with the distributional derivative, see Section C.9. In particular, we say that a function F has positive distributional derivatives and write $\frac{\partial^2 F}{\partial x_i \partial x_j} \geq 0$ if, for all nonnegative twice differentiable functions g with compact support,

$$\int_{\mathbb{R}^d} F(x)\frac{\partial^2 g}{\partial x_i \partial x_j}(x)dx \geq 0 \ .$$

Integration by parts shows that this definition is consistent with positivity of $\frac{\partial^2 F}{\partial x_i \partial x_j}$ when F is twice differentiable.

Lemma 8.33. *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz function, $|F(\mathbf{x}) - F(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and some constant $L > 0$. Let $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ be two mean-zero Gaussian vectors. Assume that (in the distributional sense)*

$$(\mathbb{E}|X_i - X_j|^2 - \mathbb{E}|Y_i - Y_j|^2)\frac{\partial^2 F}{\partial x_i \partial x_j} \geq 0 \quad \text{for all } i, j \in [m] \ , \tag{8.63}$$

and

$$F(\mathbf{x} + t\mathbf{e}) = F(\mathbf{x}) + ct \quad \text{for all } \mathbf{x} \in \mathbb{R}^m \tag{8.64}$$

where $\mathbf{e} = (1, 1, \ldots, 1) \in \mathbb{R}^m$ and c is some constant. Then

$$\mathbb{E}F(\mathbf{X}) \leq \mathbb{E}F(\mathbf{Y}) \ .$$

Proof. Observe that the Lipschitz assumption implies

$$|F(\mathbf{x})| \leq L\|\mathbf{x}\|_2, \quad \mathbf{x} \in \mathbb{R}^m, \tag{8.65}$$

so that F is of moderate growth.

We first assume that F is twice continuously differentiable such that its derivatives up to second order are of moderate growth. We note that (8.64) implies

$$\sum_{j=1}^{m}\frac{\partial^2 F}{\partial x_i \partial x_j}(\mathbf{x}) = 0 \quad \text{for all } i \in [m], \mathbf{x} \in \mathbb{R}^m \ . \tag{8.66}$$

(In fact, (8.66) and (8.64) are equivalent.) With this observation we write

$$\begin{aligned}
&\sum_{i,j=1}^{m} (\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j)\frac{\partial^2 F}{\partial x_i \partial x_j} \\
&= -\sum_{i=1}^{m}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2) \sum_{j=1, j\neq i}^{m} \frac{\partial^2 F}{\partial x_i \partial x_j} + \sum_{i\neq j}(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j)\frac{\partial^2 F}{\partial x_i \partial x_j} \\
&= -\frac{1}{2}\sum_{i\neq j}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2 + \mathbb{E}X_j^2 - \mathbb{E}Y_j^2 - 2(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j))\frac{\partial^2 F}{\partial x_i \partial x_j} \\
&= -\frac{1}{2}\sum_{i,j=1}^{m}(\mathbb{E}|X_i - X_j|^2 - \mathbb{E}|Y_i - Y_j|^2)\frac{\partial^2 F}{\partial x_i \partial x_j} \leq 0
\end{aligned}$$

by (8.63). Therefore, the function ϕ of Proposition 8.32 has nonpositive derivative and therefore $\mathbb{E}F(\mathbf{X}) \leq \mathbb{E}F(\mathbf{Y})$ (noting that we can assume without loss of generality that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent).

In the general case that F is not necessarily twice continuously differentiable, we approximate F by twice continuously differentiable functions. To this end we choose a nonnegative twice continuously differentiable function ψ with support in $B_1 = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq 1\}$ such that $\int_{\mathbb{R}^m} \psi(\mathbf{x})d\mathbf{x} = 1$. Let $\psi_h = h^{-m}\psi(\mathbf{x}/h)$, $h > 0$, which also satisfies $\int_{\mathbb{R}^m} \psi_h(\mathbf{x})d\mathbf{x} = 1$. We introduce smoothed versions F_h of the function F via convolution,

$$F_h(\mathbf{x}) = F * \psi_h(\mathbf{x}) = \int_{\mathbb{R}^m} F(\mathbf{y})\psi_h(\mathbf{x} - \mathbf{y})d\mathbf{y} \,. \tag{8.67}$$

Since $\int_{\mathbb{R}^m} \psi_h(\mathbf{x})d\mathbf{x} = 1$ and $\operatorname{supp}\psi_h \subset B(0,h) = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq h\}$ we have

$$\begin{aligned}
&|F_h(\mathbf{x}) - F(\mathbf{x})| = \Big| \int_{\mathbb{R}^m} (F(\mathbf{y}) - F(\mathbf{x}))\psi_h(\mathbf{x} - \mathbf{y})d\mathbf{y}\Big| \\
&\leq \int_{B(\mathbf{y},h)} |F(\mathbf{y}) - F(\mathbf{x})|\psi_h(\mathbf{x} - \mathbf{y})d\mathbf{y} \leq \int_{B(\mathbf{y},h)} L\|\mathbf{y} - \mathbf{x}\|_2\psi_h(\mathbf{x} - \mathbf{y})d\mathbf{y} \leq Lh \,,
\end{aligned}$$

where we have also used the Lipschitz assumption. In particular, F_h converges uniformly to F when $h \to 0$. Moreover, Lebesgue's dominated convergence theorem allows to interchange the integral and derivatives, so that F_h is twice continuously differentiable, and

$$\frac{\partial F_h}{\partial x_i} = F * \left(\frac{\partial \psi_h}{\partial x_i}\right), \quad \text{and} \quad \frac{\partial^2 F_h}{\partial x_i \partial x_j} = F * \left(\frac{\partial^2 \psi_h}{\partial x_i \partial x_j}\right) .$$

By (8.65) and since ψ_h has compact support and is twice continuously differentiable, it is straightforward to verify from the definition of the convolution (8.67) that the partial derivatives of F_h up to second order are of moderate growth. Furthermore, for any nonnegative twice continuously differentiable function g on $\mathbb{R}^m$ with compact support, it follows from Fubini's theorem that

$$\int_{\mathbb{R}^m} F_h(x)\frac{\partial^2 g}{\partial x_i \partial x_j}(x)dx = \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} F(y)\psi_h(y-x)dy\frac{\partial^2 g}{\partial x_i \partial x_j}(x)dx$$
$$= \int_{\mathbb{R}^m} F(y)\int_{\mathbb{R}^m}\psi_h(y-x)\frac{\partial^2 g}{\partial x_i \partial x_j}(x)dxdy$$
$$= \int_{\mathbb{R}^m} F(y)\frac{\partial^2}{\partial x_i \partial x_j}(\psi_h * g)(y)dy\ . \tag{8.68}$$

The last identity, that is, the interchange of taking derivatives and convolution, is justified again by Lebesgue's dominated convergence theorem. Since both ψ_h and g are nonnegative, the function $\psi_h * g$ is nonnegative as well. It follows from (8.68) and from the assumption (8.63) on the distributional derivative of F that (8.63) is valid also for F_h in place of F. Also the property (8.64) extends to F_h by the following calculation,

$$F_h(\mathbf{x}+t\mathbf{e}) = \int_{\mathbb{R}^m} F(\mathbf{x}+t\mathbf{e}-\mathbf{y})\psi_h(\mathbf{y})d\mathbf{y} = \int_{\mathbb{R}^m}(F(\mathbf{x}-\mathbf{y})+ct)\psi_h(\mathbf{y})d\mathbf{y}$$
$$= F_h(\mathbf{x}) + ct\int_{\mathbb{R}^m}\psi_h(\mathbf{x})d\mathbf{x} = F_h(\mathbf{x}) + ct\ .$$

From the already proven statement for twice continuously differentiable functions it follows that $\mathbb{E}F_h(\mathbf{X}) \le \mathbb{E}F_h(\mathbf{Y})$ for all $h > 0$. By uniform convergence of F_h to F we have

$$\mathbb{E}F(\mathbf{X}) = \lim_{h\to 0} F_h(\mathbf{X}) \le \lim_{h\to 0} F_h(\mathbf{Y}) = \mathbb{E}F(\mathbf{Y})\ .$$

This completes the proof. □

Remark 8.34. The Lipschitz assumption in the previous Lemma is not essential but simplifies the proof. The result can also be shown under other conditions on F – in particular, as used in the proof, for twice differentiable F such that F together with all its derivatives up to second order are of moderate growth.

Now we are prepared for the proof of Gordon's lemma, which in turn implies Slepian's lemma as a special case.

Proof (of Lemma 8.28). Let

$$F(\mathbf{x}) = \min_{i\in[n]}\max_{j\in[m]} x_{ij}\ ,$$

where $\mathbf{x} = (x_{ij})_{i\in[n],j\in[m]}$ is a doubly indexed vector. Then F is a Lipschitz function (with Lipschitz constant 1). We first aim at verifying (8.63). Since this condition involves only derivatives in two variables at a time, we can fix the other variables for the moment, which simplifies the notational burden. Setting $t = x_{ij}$ and $s = x_{k\ell}$ and fixing all other variables we realize that F takes the form

$$F(\mathbf{x}) = A(t,s) := \max\{\alpha(t), \beta(s)\} \quad \text{if } i = k \,,$$

or

$$F(\mathbf{x}) = B(t,s) := \min\{\alpha(t), \beta(s)\} \quad \text{if } i \neq k \,,$$

where both α and β are functions of the form

$$g(t) = \begin{cases} a & \text{if } t < a \,, \\ t & \text{if } a \le t \le b \,, \\ b & \text{if } t > b \,. \end{cases} \tag{8.69}$$

Here $a \le b$ are some numbers that may possibly take the values $a = -\infty$ and $b = +\infty$. We claim that the distributional derivatives of A, B are nonnegative. To prove this for A we note that

$$A(t,s) = \frac{1}{2}(\alpha(t) + \beta(s) + |\alpha(t) - \beta(s)|) \,.$$

Therefore, a partial weak derivative of A is given by (see Exercise 8.12(a))

$$\begin{aligned} \frac{\partial}{\partial t} A(t,s) &= \frac{1}{2}(\alpha'(t) + \alpha'(t)\mathrm{sgn}(\alpha(t) - \beta(s)) \\ &= \begin{cases} 0 & \text{if } t \notin [a,b] \,, \\ \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(t - \beta(s)) & \text{if } t \in [a,b] \,. \end{cases} \end{aligned} \tag{8.70}$$

where α' is a weak derivative of α, see Exercise 8.12(b), and a, b are the numbers defining α, see (8.69). The function $s \mapsto \mathrm{sgn}(t - \beta(s))$ is nonincreasing in s and therefore the distributional derivative $\frac{\partial^2}{\partial s \partial t} A$ is nonpositive as claimed, see also Exercise 8.12(c).

Nonnegativity of $\frac{\partial^2}{\partial s \partial t} B$ follows similarly by writing

$$B(s,t) = \min\{\alpha(t), \beta(s)\} = (\alpha(t) + \beta(s) - |\alpha(t) - \beta(s)|)/2 \,.$$

Therefore, we showed that (in the sense of distributional derivatives)

$$\begin{aligned} \frac{\partial^2 F}{\partial x_{ij} \partial x_{k\ell}} &\le 0 \quad \text{if } i = k \,, \\ \frac{\partial^2 F}{\partial x_{ij} \partial x_{k\ell}} &\ge 0 \quad \text{if } i \neq k \,. \end{aligned}$$

It follows from Assumptions (8.53), (8.54) that

$$(\mathbb{E}|X_{i,j} - X_{k,\ell}|^2 - \mathbb{E}|Y_{i,j} - Y_{k,\ell}|^2) \frac{\partial^2 F}{\partial x_{ij} \partial x_{k\ell}} \ge 0 \quad \text{for all } i, j, k, \ell \,. \tag{8.71}$$

Moreover, the function F satisfies $F(\mathbf{x} + t\mathbf{e}) = F(\mathbf{x}) + t$. The conditions of Lemma 8.33 are therefore satisfied and we conclude that $\mathbb{E}F(\mathbf{X}) \le \mathbb{E}F(\mathbf{Y})$.

□

8.8 Concentration of Measure

Concentration of measure describes the phenomenon that Lipschitz functions on high-dimensional probability spaces concentrate well around their expectation. We present a precise statement for Gaussian measures. The proof of our first theorem uses the auxiliary tools developed in the previous section and is rather short, but only gives the non-optimal constant 4 in the probability decay, see (8.73). With a somewhat more sophisticated technique using semi-group tools we provide the optimal constant 2 in Theorem 8.38 below.

Theorem 8.35. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function, that is,*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \tag{8.72}$$

for a constant $L > 0$. Let $\mathbf{g} = (g_1, \ldots, g_n)$ be a standard Gaussian random vector. Then for all $t > 0$

$$\mathbb{P}(f(\mathbf{g}) - \mathbb{E}[f(\mathbf{g})] > t) \leq \exp\left(-\frac{t^2}{4L^2}\right) , \tag{8.73}$$

and consequently

$$\mathbb{P}(|f(\mathbf{g}) - \mathbb{E}f(\mathbf{g})| \geq t) \leq 2\exp(-t^2/(4L^2)) .$$

Proof. We first assume that f is differentiable. Let $\mathbf{X}, \mathbf{Y}$ be independent copies of $\mathbf{g}$. We use the Laplace transform method which, for a parameter $\lambda \in \mathbb{R}$, requires to bound

$$\psi(\lambda) := \mathbb{E}\exp(\lambda(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{Y})])) ,$$

where $\mathbf{Y}$ denotes an independent copy of $\mathbf{X}$. Using convexity of $t \mapsto \exp(-\lambda t)$ and Jensen's inequality yields

$$\psi(\lambda) \leq \mathbb{E}\exp(\lambda(f(\mathbf{X}) - f(\mathbf{Y}))) = \mathbb{E}G_\lambda(\mathbf{X}, \mathbf{Y}) ,$$

where we have set $G_\lambda(\mathbf{x}, \mathbf{y}) = \exp(\lambda(f(\mathbf{x}) - f(\mathbf{y})))$. The concatenated vector $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ is a standard Gaussian vector of length $2n$. Let $\mathbf{X}'$ denote an independent copy of $\mathbf{X}$ and put $\mathbf{W} = (\mathbf{X}', \mathbf{X}')$. For $0 \leq t \leq 1$ define $\mathbf{U}(t) = \sqrt{t}\mathbf{Z} + \sqrt{1-t}\mathbf{W}$ and $\phi(t) = \mathbb{E}G_\lambda(\mathbf{U}(t))$. Clearly, $\phi(0) = \mathbb{E}G_\lambda(\mathbf{X}', \mathbf{X}') = \mathbb{E}\exp(\lambda(f(\mathbf{X}') - f(\mathbf{X}'))) = 1$. As the next step, we use Proposition 8.32 to compute the derivative of ϕ. To this end we note that $\mathbb{E}X_iX_j = \mathbb{E}X_i'X_j' = \delta_{ij}$ and $\mathbb{E}X_iY_j = 0$ for all i, j. Furthermore, it follows from the Lipschitz assumption (8.72) that G_λ is of moderate growth, see (8.55). Therefore, (8.62) yields

$$\begin{aligned}\phi'(t) &= \frac{1}{2}\sum_{i,j\in[2n]} (\mathbb{E}W_iW_j - \mathbb{E}Z_iZ_j)\mathbb{E}\left[\frac{\partial^2 G_\lambda}{\partial z_i \partial z_j}(\mathbf{U}(t))\right] \\ &= -\mathbb{E}\sum_{i=1}^n \frac{\partial^2 G_\lambda}{\partial x_i \partial y_i}(\mathbf{U}(t)) .\end{aligned}$$

The partial derivatives of G_λ are given by

$$\frac{\partial^2 G_\lambda}{\partial x_i \partial y_i}(\mathbf{x},\mathbf{y}) = -\lambda^2 \frac{\partial f}{\partial x_i}(\mathbf{x}) \frac{\partial f}{\partial y_i}(\mathbf{y})\, G_\lambda(\mathbf{x},\mathbf{y}), \quad \mathbf{x},\mathbf{y} \in \mathbb{R}^n \ .$$

Since we assumed f to be differentiable it follows from the Lipschitz assumption (8.72) that

$$\|\nabla f(\mathbf{x})\|_2^2 = \sum_{i=1}^m \left| \frac{\partial f}{\partial x_i}(\mathbf{x}) \right|^2 \le L^2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n \ ,$$

so that the Cauchy-Schwarz inequality yields

$$\begin{aligned} \phi'(t) &= \lambda^2 \mathbb{E} \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{X}) \frac{\partial f}{\partial y_i}(\mathbf{Y})\, G_\lambda(\mathbf{U}(t)) \\ &\le \lambda^2 \mathbb{E} \|\nabla f(\mathbf{X})\|_2 \|\nabla f(\mathbf{Y})\|_2 G_\lambda(\mathbf{U}(t)) \le \lambda^2 L^2 \mathbb{E} G_\lambda(\mathbf{U}(t)) = \lambda^2 L^2 \phi(t) \ . \end{aligned}$$

Since $\phi(t) > 0$ we may divide by it, and setting $\tau(t) := \ln \phi(t)$ shows that

$$\tau'(t) \le \lambda^2 L^2 \ .$$

Together with $\phi(0) = 1$ this differential inequality implies by integration that

$$\tau(1) \le \int_0^1 \lambda^2 L^2 dt = \lambda^2 L^2 \ ,$$

and consequently,

$$\psi(\lambda) \le \phi(1) = \exp(\tau(1)) \le \exp(\lambda^2 L^2) \ .$$

For $t, \lambda > 0$, Markov's inequality yields

$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E} f(\mathbf{X}) \ge t) \le \psi(\lambda) e^{-\lambda t} \le \exp(\lambda^2 L^2 - \lambda t) \ .$$

Choosing $\lambda = t/(2L^2)$ yields the claimed inequality (8.73).

In the general case, where f is not necessarily differentiable, we can find for each $\varepsilon > 0$ a differentiable Lipschitz function g with the same Lipschitz constant L, such that $|f(\mathbf{x}) - g(\mathbf{x})| \le \epsilon$ for all $\mathbf{x} \in \mathbb{R}^n$, see Theorem C.11. It follows then that

$$\begin{aligned} \mathbb{P}(f(\mathbf{X}) - \mathbb{E} f(\mathbf{X}) > t) &\le \mathbb{P}(g(\mathbf{X}) - \mathbb{E} g(\mathbf{X}) \ge t - 2\varepsilon) \\ &\le \exp(-(t-2\varepsilon)^2/(4L^2)) \ . \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, (8.73) follows also for general, not necessarily differentiable Lipschitz functions. □

In order to improve on the constant 4 in (8.73) we will use an alternative approach based on the Ornstein–Uhlenbeck semigroup P_t. For $t \geq 0$ and a measurable function $f : \mathbb{R}^n \to \mathbb{R}$ of moderate growth it is defined as

$$\begin{aligned}(P_t f)(x) &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(e^{-t}x + (1-e^{-2t})^{1/2}y) e^{-\|y\|_2^2/2} dy \\ &= \mathbb{E} f(e^{-t}x + (1-e^{-2t})^{1/2}\mathbf{Y}) \,, \end{aligned} \tag{8.74}$$

where $\mathbf{Y}$ is a standard Gaussian vector in $\mathbb{R}^n$. We also require the L^p-space with respect to the Gaussian measure γ,

$$L^p(\gamma) =: \{f \text{ measurable }, \|f\|_{L^p(\gamma)} := (\mathbb{E}|f(\mathbf{Y})|^p)^{1/p} < \infty\} \,, 1 \leq p < \infty \,,$$

and the obvious modification for the space $L^\infty(\gamma)$. We summarize some basic properties of the Ornstein–Uhlenbeck semigroup. Below when we speak of smooth functions f, we mean that that f should have sufficiently many continuous derivatives and that f together with these derivatives should be bounded.

Proposition 8.36. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be of moderate growth and P_t, $t \geq 0$, the Ornstein–Uhlenbeck semigroup.*

(a) (Positivity) If $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ then $P_t f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.
(b) (Semigroup property) For $t, s \geq 0$ we have $P_t P_s f = P_{t+s} f$.
(c) (Boundedness) For $f \in L^p(\gamma)$, $1 \leq p \leq \infty$, it holds $\|P_t f\|_{L^p(\gamma)} \leq \|f\|_{L^p(\gamma)}$, that is P_t is bounded on $L^p(\gamma)$.
(d) (Continuity) The mapping $t \mapsto P_t f$ is strongly continuous in $L^p(\gamma)$, $1 \leq p < \infty$, i.e., $\lim_{t\to 0} \|P_t f - f\|_p = 0$ for all $f \in L^p(\gamma)$.
(e) It holds $\lim_{t\to\infty} P_t f(\mathbf{x}) = \mathbb{E} f(\mathbf{Y})$ for all $\mathbf{x} \in \mathbb{R}^n$, and for all $f \in L^1(\gamma)$.
(f) (Infinitesimal generator) The differential operator L defined, for smooth enough f, via

$$(Lf)(\mathbf{x}) = \Delta f(\mathbf{x}) - \mathbf{x} \cdot \nabla f(x) = \sum_{j=1}^n \left(\frac{\partial^2 f}{\partial x_j^2}(\mathbf{x}) - x_j \frac{\partial f}{\partial x_j}(\mathbf{x}) \right), \quad \mathbf{x} \in \mathbb{R}^n \,,$$

is the infinitesimal generator of the semigroup P_t, that is,

$$\lim_{t\to 0} \frac{P_t f(\mathbf{x}) - f(\mathbf{x})}{t} = Lf(\mathbf{x}) \,, \mathbf{x} \in \mathbb{R}^n \,. \tag{8.75}$$

(g) (Heat equation) Given a smooth enough f, the function

$$u(\mathbf{x}, t) = P_t f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, t \geq 0 \,,$$

is the solution of the partial differential (heat) equation

$$\frac{du}{dt} = Lu$$

with initial condition $u(\mathbf{x}, 0) = f(\mathbf{x})$.

(h) (Integration by parts) For smooth enough f, g it holds

$$\int_{\mathbb{R}^n} f(\mathbf{x})(-Lg)(\mathbf{x})\frac{e^{-\|\mathbf{x}\|_2^2/2}}{(2\pi)^{n/2}}d\mathbf{x} = \int_{\mathbb{R}^n} \nabla f(\mathbf{x})\cdot\nabla g(\mathbf{x})\frac{e^{-\|\mathbf{x}\|_2^2/2}}{(2\pi)^{n/2}}d\mathbf{x}\ . \quad (8.76)$$

(i) For smooth enough f it holds

$$\frac{1}{2}L(\|\nabla f\|_2^2) - \nabla f\cdot\nabla(Lf) \geq \|\nabla f\|_2^2 \quad \textit{pointwise}\ .$$

(j) For every $t \geq 0$, it holds

$$\|\nabla(P_t f)\|_2^2 \leq e^{-2t}P_t(\|\nabla f\|_2^2) \quad \textit{pointwise}\ .$$

Proof. (a) Positivity follows immediately from the definition (8.74).

(b) Let $\mathbf{Y}, \mathbf{Z}$ be two independent standard Gaussian vectors on $\mathbb{R}^n$. For $t, s \geq 0$ we have

$$\begin{aligned} P_tP_sf(\mathbf{x}) &= \mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\mathbf{Y}} f\left(e^{-s}(e^{-t}\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y}) + (1-e^{-2s})^{1/2}\mathbf{Z}\right) \\ &= \mathbb{E}f(e^{-(t+s)}\mathbf{x} + e^{-s}(1-e^{-2t})^{1/2}\mathbf{Y} + (1-e^{-2s})^{1/2}\mathbf{Z})\ . \end{aligned}$$

The vector $e^{-s}(1-e^{-2t})^{1/2}\mathbf{Y}+(1-e^{-2s})^{1/2}\mathbf{Z}$ is a Gaussian random vector with independent mean zero entries of variance $\sigma^2 = e^{-2s}(1-e^{-2t}) + (1-e^{-2s}) = 1-e^{-2(t+s)}$, and therefore has the same distribution as the random vector $\sqrt{1-e^{-2(t+s)}}\mathbf{X}$, where $\mathbf{X}$ is a standard Gaussian vector. It follows that

$$P_tP_sf(\mathbf{x}) = \mathbb{E}f(e^{-(t+s)}\mathbf{x} + \sqrt{1-e^{-2(t+s)}}\mathbf{X}) = P_{t+s}f(\mathbf{x})\ .$$

This shows the semigroup property.

(c) Denote by $\mathbf{X}, \mathbf{Y}$ two independent standard Gaussian random variables. For $1 \leq p < \infty$ and $f \in L^p(\gamma)$ we have

$$\mathbb{E}|P_tf(\mathbf{X})|^p = \mathbb{E}|\mathbb{E}f(e^{-t}\mathbf{X}+(1-e^{-2t})^{1/2}\mathbf{Y})|^p \leq \mathbb{E}|f(e^{-t}\mathbf{X}+(1-e^{-2t})^{1/2}\mathbf{Y})|^p\ .$$

Observe that for all $t \geq 0$ the entries of the random vector $\mathbf{W} = e^{-t}\mathbf{X} + (1-e^{-2t})^{1/2}\mathbf{Y}$ have mean zero and variance 1, so that it is again a standard Gaussian random vector. Therefore,

$$\mathbb{E}|P_tf(\mathbf{X})|^p \leq \mathbb{E}|f(\mathbf{X})|^p\ .$$

The case $p = \infty$ is even easier.

(d) Assume first that $f \in L^p(\gamma)$ is continuous and bounded. Then it follows from Lebesgue's dominated convergence theorem that $\lim_{t\to 0}\|P_tf - f\|_p^p = 0$. In the general case, we can find, for each $\epsilon > 0$ a bounded and continuous function g such that $\|f-g\|_{L^p(\gamma)} \leq \epsilon$. Let further t such that $\|P_tg - g\|_{L^p\gamma} \leq \epsilon$. Then the triangle inequality together with (c) yields

$$\|P_tf - f\| \leq \|P_tf - P_tg\| + \|P_tg - g\| + \|g - f\| \leq 2\|f-g\| + \|P_tg - g\| \leq 3\epsilon\ .$$

This shows the claim.

(e) For continuous and bounded f it follows from Lebesgue's dominated convergence theorem that

$$\begin{aligned}\lim_{t\to\infty} P_t f(\mathbf{x}) &= \lim_{t\to\infty} \mathbb{E} f(e^{-t}\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y}) \\ &= \mathbb{E} \lim_{t\to\infty} f(e^{-t}\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y}) = \mathbb{E} f(\mathbf{Y}) \,.\end{aligned}$$

The general case follows from density of the continuous and bounded functions in $L^1(\gamma)$, similarly as in the proof of (d).

(f) We use the Taylor expansion of f in $\mathbf{x}$ up to third order in the form

$$f(\mathbf{z}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (\mathbf{z}-\mathbf{x}) + \frac{1}{2}(\mathbf{z}-\mathbf{x})^\top (\mathbf{H}f)(\mathbf{x})(\mathbf{z}-\mathbf{x}) + Rf(\mathbf{z},\mathbf{x}) \,,$$

where $\mathbf{H}f(\mathbf{x})$ denotes the Hessian matrix of f in $\mathbf{x}$ and the remainder satisfies $|Rf(\mathbf{x},\mathbf{z})| \le C\|\mathbf{z}-\mathbf{x}\|_2^3$ due to the boundedness of the third order partial derivates of f. (Recall that we agreed to call f smooth when it has sufficiently many bounded derivatives.) Denoting by $\mathbf{Y}$ a standard Gaussian vector we obtain, for $t > 0$,

$$\frac{P_t f(\mathbf{x}) - f(\mathbf{x})}{t} = G_1(\mathbf{x},t) + G_2(\mathbf{x},t) + G_3(\mathbf{x},t) \,,$$

where

$$\begin{aligned}&G_1(\mathbf{x},t) = t^{-1}\mathbb{E}\left[\nabla f(x) \cdot ((e^{-t}-1)\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y})\right] \,, \\ &G_2(\mathbf{x},t) \qquad\qquad (8.77) \\ &= \frac{1}{2t}\mathbb{E}\left[((e^{-t}-1)\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y})^\top \mathbf{H}f(\mathbf{x})((e^{-t}-1)\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y})\right] \,, \\ &G_3(\mathbf{x},t) = t^{-1}\mathbb{E}\left[Rf(e^{-t}\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y}, \mathbf{x})\right] \,.\end{aligned}$$

The third term satisfies

$$|\lim_{t\to 0} G_3(\mathbf{x},t)| \le C \lim_{t\to 0} t^{-1}\mathbb{E}\|(e^{-t}-1)\mathbf{x} + (1-e^{-2t})^{1/2}\mathbf{Y}\|_2^3 = 0 \,.$$

Lebesgue's dominated convergence theorem justifies interchange of the limit $t \to 0$ and the expectation, which yields

$$\lim_{t\to 0} t^{-1} G_1(\mathbf{x},t) = -\nabla f(\mathbf{x}) \cdot \mathbf{x} \,.$$

For the term in (8.77) we similarly obtain

$$\begin{aligned}\lim_{t\to 0} G_2(\mathbf{x},t) &= \mathbb{E}\left[(\lim_{t\to 0} \frac{1-e^{-2t}}{2t}\mathbf{Y})^\top \mathbf{H}f(\mathbf{x})\mathbf{Y}\right] = \mathbb{E}\mathbf{Y}^\top \mathbf{H}f(\mathbf{x})\mathbf{Y} \\ &= \sum_{j,k=1}^n \mathbb{E}[Y_j Y_k] \frac{\partial f^2}{\partial x_j \partial x_k}(\mathbf{x}) = \sum_{j=1}^n \frac{\partial f^2}{\partial x_j^2}(\mathbf{x}) = \Delta f(\mathbf{x}) \,.\end{aligned}$$

This shows the claimed relation (8.75).

(g) Clearly, $P_0 f = f = u(\cdot, 0)$. It follows from (b) and (c) that

$$\begin{aligned}\frac{du}{dt}(\mathbf{x}, s) &= \lim_{r \to s} \frac{P_r f(\mathbf{x}) - P_s f(\mathbf{x})}{r - s} = \lim_{t \to 0} \frac{P_{s+t} f(\mathbf{x}) - P_s f(\mathbf{x})}{t} \\ &= \lim_{t \to 0} \frac{P_t (P_s f) - P_s f(\mathbf{x})}{t} = L P_s f(\mathbf{x}) = (Lu)(\mathbf{x}, s).\end{aligned}$$

This establishes the validity of the heat equation for $u(\mathbf{x}, t) = P_t f(\mathbf{x})$.

(h) We start with the case $n = 1$ and assume that f, g are smooth with compact support. Observe that the function $h(x) = e^{-x^2/2} f'(x)$ has derivative $h'(x) = (-x f'(x) + f''(x)) e^{-x^2/2}$. Therefore, it follows from integration by parts that

$$\begin{aligned}&\int_{\mathbb{R}} (-Lf)(x) g(x) e^{-x^2/2} dx = \int_{\mathbb{R}} (-f''(x) + x f'(x)) g(x) e^{-x^2/2} dx \\ &= \int_{\mathbb{R}} (-h'(x)) g(x) dx = \int_{\mathbb{R}} h(x) g'(x) dx = \int_{\mathbb{R}} f'(x) g'(x) e^{-x/2} dx.\end{aligned}$$

This establishes the claim for $n = 1$. For general n and smooth f, g with compact support we observe that

$$\begin{aligned}&\int_{\mathbb{R}^n} (-Lf)(\mathbf{x}) g(\mathbf{x}) e^{-\|\mathbf{x}\|_2^2/2} d\mathbf{x} \\ &= \sum_{j=1}^n \int_{\mathbb{R}^n} \left(-\frac{\partial f^2}{\partial x_j^2}(\mathbf{x}) + x_j \frac{\partial f}{\partial x_j}(\mathbf{x}) \right) g(\mathbf{x}) \prod_{\ell=1}^n e^{-x_\ell^2/2} d\mathbf{x} \\ &= \sum_{j=1}^n \int_{\mathbb{R}^n} \frac{\partial f}{\partial x_j}(\mathbf{x}) \frac{\partial g}{\partial x_j}(\mathbf{x}) e^{-\|\mathbf{x}\|_2^2/2} d\mathbf{x} = \int_{\mathbb{R}^n} \nabla f(\mathbf{x}) \cdot \nabla g(\mathbf{x}) e^{-\|\mathbf{x}\|_2^2/2} d\mathbf{x}\,,\end{aligned}$$

where the second equality follows from the case $n = 1$. General smooth functions f with $\mathbb{E}\|\nabla f(\mathbf{Y})\|_2^2 < \infty$ can be approximated arbitrarily well by smooth functions with compact support in the sense that for given ϵ one can find a smooth function $\tilde{f}$ with compact support such that $\mathbb{E}\|\nabla f(\mathbf{Y}) - \nabla \tilde{f}(\mathbf{Y})\|_2^2 < \epsilon$. This extends the relation (8.76) for general smooth functions f for which both sides of (8.76) are well-defined.

(i) It is straightforward to verify the following identities,

$$\begin{aligned}
\Delta\|\nabla f\|_2^2 &= 2\sum_{i,j=1}^n \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)^2 + 2\sum_{i,j=1}^n \frac{\partial^3 f}{\partial x_i \partial x_j^2}\frac{\partial f}{\partial x_i}\,, \\
\mathbf{x}\cdot\nabla\|\nabla f\|_2^2(\mathbf{x}) &= 2\sum_{i,j=1}^n x_i \frac{\partial^2 f}{\partial x_i \partial x_j}\frac{\partial f}{\partial x_j}\,, \\
\nabla f\cdot\nabla(\Delta f) &= \sum_{i,j=1}^n \frac{\partial f}{\partial x_i}\frac{\partial^3 f}{\partial x_i \partial x_j^2}\,, \\
\nabla f(\mathbf{x})\cdot(\nabla(\mathbf{x}\cdot f))(\mathbf{x}) &= \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}\right)^2 + \sum_{i,j=1}^n x_j \frac{\partial f}{\partial x_i}\frac{\partial^2 f}{\partial x_i \partial x_j}\,.
\end{aligned}$$

It follows that the pointwise inequality

$$\frac{1}{2}L(\|\nabla f\|_2^2) - \nabla f\cdot\nabla(Lf) = \|Hf\|_F^2 + \|\nabla f\|_2^2 \geq \|\nabla f\|_2^2$$

holds, where Hf is the Hessian matrix of f.

(j) For fixed $r \geq 0$ we set $\psi(s) = e^{-2s}P_s(\|\nabla P_{r-s}f\|_2^2)$ for $0 \leq s \leq r$. The derivative of ψ is given as

$$\psi'(s) = -2e^{-2s}P_s(\|\nabla P_{r-s}f\|_2^2) + e^{-2s}\frac{\partial}{\partial s}P_s(\|\nabla P_{r-s}f\|_2^2). \tag{8.78}$$

Using the semigroup property (b), continuity (d), the infinitesimal generator L in (e), and the heat equation for P_t in (f) we obtain for the second term,

$$\begin{aligned}
&\frac{\partial}{\partial s}P_s(\|\nabla P_{r-s}f\|_2^2) = \lim_{t\to 0} t^{-1}\left(P_{s+t}(\|\nabla P_{r-s-t}f\|_2^2) - P_s(\|\nabla P_{r-s}f\|_2^2)\right) \\
&= \lim_{t\to 0}\frac{P_{s+t}-P_s}{t}(\|\nabla P_{r-s-t}f\|_2^2) + \lim_{t\to 0} P_s\left(\frac{\|\nabla P_{r-s-t}f\|_2^2 - \|\nabla P_{r-s}f\|_2^2}{t}\right) \\
&= P_s L(\|\nabla P_{r-s}f\|_2^2) + 2P_s\left(\nabla P_{r-s}f\cdot\nabla\left(\frac{d}{ds}P_{r-s}f\right)\right) \\
&= P_s(L(\|\nabla P_{r-s}f\|_2^2) - 2\nabla P_{r-s}f\cdot\nabla L(P_{r-s}f))\,.
\end{aligned}$$

Using positivity (a) together with (h) applied to $P_{r-s}f$ shows that

$$\begin{aligned}
\psi'(s) &= -2e^{-2s}P_s\left(\|\nabla P_{r-s}f\|_2^2 - \frac{1}{2}L(\|\nabla P_{r-s}f\|_2^2) + \nabla P_{r-s}f\cdot L(\nabla P_{r-s}f)\right) \\
&\geq -2e^{-2s}\left(\|\nabla P_{r-s}f\|_2^2 - \frac{1}{2}L(\|\nabla P_{r-s}f\|_2^2) + \nabla P_{r-s}f\cdot L(\nabla P_{r-s}f)\right) \\
&\geq 0\,.
\end{aligned}$$

This implies that

$$\|P_r f\|_2^2 = \psi(0) \leq \psi(r) = e^{-2r}P_r(\|\nabla f\|_2^2)\,,$$

which is the claimed inequality. □

Remark 8.37. The semigroup property of P_t and L being the infinitesimal generator allows to write $P_t = e^{tL}$, where the latter is an operator valued exponential function.

With this preparation we are ready to provide an alternative proof of concentration of measure for Lipschitz functions with an improved constant.

Theorem 8.38. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant L, see (8.72). Let $\mathbf{X} = (X_1, \dots, X_n)$ be a vector of independent standard Gaussian random variables. Then for all $t > 0$*

$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] > t) \leq \exp\left(-\frac{t^2}{2L^2}\right) , \tag{8.79}$$

and consequently

$$\mathbb{P}(|f(\mathbf{X}) - \mathbb{E}f(\mathbf{X})| \geq t) \leq 2\exp(-t^2/(2L^2)) .$$

Proof. We may assume that f is differentiable, so that the Lipschitz condition implies $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$. The general case follows then with the same approximation argument as in the proof of Theorem (8.35). The Lipschitz condition implies as well that f is of moderate growth, see (8.55), in particular $f \in L^1(\gamma)$. We may furthermore assume that $\mathbb{E}f(\mathbf{X}) = 0$ – otherwise, we subtract the mean. For a parameter $\lambda \in \mathbb{R}$, we set

$$\psi_\lambda(t) = \mathbb{E}\exp(\lambda P_t f(\mathbf{X})), \quad t \geq 0 .$$

It follows from the moderate growth condition and from the contraction property of P_t on $L^1(\gamma)$ in Proposition 8.36(c) that the expectation defining ψ_λ exists for all $t \geq 0$ and $\lambda \in \mathbb{R}$. Furthermore, it follows from Proposition 8.36(e) that

$$\lim_{t\to\infty} \psi_\lambda(t) = \mathbb{E}\exp(\lambda \lim_{t\to\infty} P_t f(\mathbf{X})) = \mathbb{E}(\exp(\lambda \mathbb{E}f(\mathbf{Y}))) = 1$$

by the mean-zero assumption on f. Now parts (f), (h) and (j) of Proposition 8.36 yield, for $t \geq 0$,

$$\begin{aligned}
\psi_\lambda(t) &= 1 - \int_t^\infty \psi'(s)ds = 1 - \lambda \int_t^\infty \mathbb{E}L(P_s f)(\mathbf{X}) \exp(\lambda P_s f(\mathbf{Y}))ds \\
&= 1 + \lambda^2 \int_t^\infty \mathbb{E}\left[\nabla(P_s f)(\mathbf{Y}) \cdot \nabla \exp(\lambda P_s f(\mathbf{Y}))\right] ds \\
&= 1 + \lambda^2 \int_t^\infty \mathbb{E}\left[\|\nabla(P_s f)\mathbf{Y}\|_2^2 \exp(\lambda P_s f(\mathbf{Y}))\right] ds \\
&\leq 1 + \lambda^2 \int_t^\infty e^{-2s}\mathbb{E}\left[\|\nabla f(\mathbf{Y})\|_2^2 \exp(\lambda P_s f(\mathbf{Y}))\right] ds \\
&\leq 1 + \lambda^2 L^2 \int_t^\infty e^{-2s}\psi_\lambda(s)ds .
\end{aligned} \tag{8.80}$$

Set $H(t)$ as the logarithm of the last term above, that is,

$$H(t) = \ln\left(1 + \lambda^2 L^2 \int_t^\infty e^{-2s}\psi_\lambda(s)ds\right) .$$

Then the estimate (8.80) yields

$$H'(t) = \frac{-\lambda^2 L^2 e^{-2t}\psi_\lambda(t)}{\exp(H(t))} \geq \frac{-\lambda^2 L^2 e^{-2t}\exp(H(t))}{\exp(H(t))} = -\lambda^2 L^2 e^{-2t} .$$

Therefore,

$$\ln\psi_\lambda(0) \leq H(0) = -\int_0^\infty H'(s)ds \leq \lambda^2 L^2 \int_0^\infty e^{-2t}dt = \frac{1}{2}\lambda^2 L^2 ,$$

and we deduced that $\psi_\lambda(0) = \mathbb{E}\exp(\lambda f(\mathbf{Y})) \leq \exp(\lambda^2 L^2/2)$. It follows from Markov's inequality that, for $t, \lambda > 0$,

$$\mathbb{P}(f(\mathbf{Y}) \geq t) \leq \exp(-\lambda t)\exp(\lambda^2 L^2/2) .$$

Choosing $\lambda = t/L^2$ completes the proof. □

We close this section with the useful special case of the Lipschitz function $\|\cdot\|_2$, which has Liptschitz constant 1. If $\mathbf{g} \in \mathbb{R}^n$ is a standard Gaussian vector then it follows from Theorem 8.38 and Proposition 8.1 that

$$\mathbb{P}(\|\mathbf{g}\|_2 \geq \sqrt{n} + t) \leq \mathbb{P}(\|\mathbf{g}\|_2 \geq \mathbb{E}\|\mathbf{g}\|_2 + t) \leq e^{-t^2/2} . \tag{8.81}$$

8.9 Bernstein Inequality for Suprema of Empirical Processes

In this section we present a deviation inequality for suprema of empirical processes above their mean, which will become very useful in Chapter 12. Let $Y_1, \ldots, Y_M$ be independent random vectors in $\mathbb{C}^n$ and let $\mathcal{F}$ be a countable collection of functions from $\mathbb{C}^n$ into $\mathbb{R}$. We are interested in the random variable $Z = \sup_{\mathsf{F}\in\mathcal{F}} \sum_{\ell=1}^M \mathsf{F}(Y_\ell)$, that is, the supremum of an empirical process. In particular, we study its deviation from its mean $\mathbb{E}Z$.

Theorem 8.39. *Let $\mathcal{F}$ be a countable set of functions $\mathsf{F} : \mathbb{C}^n \to \mathbb{R}$. Let $Y_1, \ldots, Y_M$ be independent random vectors on $\mathbb{C}^n$ such that $\mathbb{E}\mathsf{F}(Y_\ell) = 0$ and $\mathsf{F}(Y_\ell) \leq K$ for $\ell \in [M]$ and for all $\mathsf{F} \in \mathcal{F}$ for some constant $K > 0$. Introduce*

$$Z = \sup_{\mathsf{F}\in\mathcal{F}} \sum_{\ell=1}^M \mathsf{F}(Y_\ell) . \tag{8.82}$$

Let $\sigma_\ell^2 > 0$ such that $\mathbb{E}\left[\mathsf{F}(Y_\ell)^2\right] \leq \sigma_\ell^2$ for all $\mathsf{F} \in \mathcal{F}$ and $\ell \in [M]$. Then, for all $t > 0$,

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left(-\frac{t^2/2}{\sigma^2 + 2K\mathbb{E}Z + tK/3}\right) , \tag{8.83}$$

where $\sigma^2 = \sum_{\ell=1}^M \sigma_\ell^2$.

Remark 8.40. (a) If $\mathcal{F}$ consists only of a single function, then inequality (8.83) reduces to the standard Bernstein inequality in Corollary 7.31. It is remarkable that Theorem 8.39 reproduces the same constants in this more general setting.

(b) The deviation inequality (8.83) can be extended to a concentration inequality, which is sometimes refered to as Talagrand's inequality, see the Notes section.

(c) Theorem 8.39 holds without change if Z is replaced by

$$\widetilde{Z} = \sup_{\mathsf{F}\in\mathcal{F}} \left| \sum_{\ell=1}^M \mathsf{F}(Y_\ell) \right| .$$

Before turning to the proof of the theorem, we present the following Bernstein type inequality for the sum of independent mean zero random vectors in a normed space. Its formulation uses the dual norm, see Definition A.4 and in particular (A.5).

Corollary 8.41. *Let* $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ *be independent copies of a random vector* $\mathbf{Y}$ *on* $\mathbb{C}^n$ *satisfying* $\mathbb{E}\mathbf{Y} = 0$. *Assume* $\|\mathbf{Y}\| \leq K$ *for some* $K > 0$ *and some norm* $\|\cdot\|$ *on* $\mathbb{C}^n$. *Let*

$$Z = \left\| \sum_{\ell=1}^M \mathbf{Y}_\ell \right\|$$

and

$$\sigma^2 = \sup_{\mathbf{x}\in B^*} \mathbb{E}|\langle \mathbf{x}, \mathbf{Y}\rangle|^2 , \tag{8.84}$$

where $B^* = \{\mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_* \leq 1\}$ *denotes the unit ball in the dual norm* $\|\cdot\|_*$. *Then, for* $t > 0$,

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left(-\frac{t^2/2}{M\sigma^2 + 2K\mathbb{E}Z + tK/3}\right) . \tag{8.85}$$

Proof. Introduce the random functions $\mathsf{F}_\mathbf{x}(\mathbf{Y}) := \mathrm{Re}(\langle \mathbf{x}, \mathbf{Y}\rangle)$, $\mathbf{x} \in \widetilde{B^*}$. By the characterization (A.5) of a norm by its dual norm we have

$$Z = \sup_{\mathbf{x}\in B^*} \mathrm{Re}\left(\left\langle \mathbf{x}, \sum_{\ell=1}^M \mathbf{Y}_\ell \right\rangle\right) = \sup_{\mathbf{x}\in B^*} \sum_{\ell=1}^M \mathrm{Re}\left(\langle \mathbf{x}, \mathbf{Y}_\ell\rangle\right) = \sup_{\mathbf{x}\in B^*} \sum_{\ell=1}^M \mathsf{F}_\mathbf{x}(\mathbf{Y}_\ell) .$$

Let $\widetilde{B^*}$ be a dense countable subset of B^*. Then $Z = \sup_{\mathbf{x}\in\widetilde{B^*}} \sum_{\ell=1}^M \mathsf{F}_\mathbf{x}(\mathbf{Y}_\ell)$ and

$$\sup_{\mathbf{x}\in\widetilde{B^*}} \mathbb{E}\mathsf{F}_{\mathbf{x}}(\mathbf{Y}_\ell)^2 = \sup_{\mathbf{x}\in B^*} \mathbb{E}|\langle \mathbf{x}, \mathbf{Y}\rangle|^2 = \sigma^2 \,.$$

The random variables $\mathsf{F}_{\mathbf{x}}(\mathbf{Y}) := \mathrm{Re}(\langle \mathbf{x}, \mathbf{Y}\rangle)$, $\mathbf{x} \in \widetilde{B^*}$ satisfy $\mathbb{E}\mathsf{F}_{\mathbf{x}}(\mathbf{Y}) = 0$, and are almost surely bounded, $|\mathsf{F}_{\mathbf{x}}(\mathbf{Y})| \le \|\mathbf{x}\|_* \|\mathbf{Y}\| \le K$. The conclusion follows therefore from Theorem 8.39. □

We specialize to the case of the ℓ_2-norm in the next statement.

Corollary 8.42. *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_M$ be independent copies of a random vector $\mathbf{Y}$ on $\mathbb{C}^n$ satisfying $\mathbb{E}\mathbf{Y} = 0$. Assume $\|\mathbf{Y}\|_2 \le K$ for some $K > 0$. Let*

$$Z = \left\| \sum_{\ell=1}^M \mathbf{Y}_\ell \right\|_2 , \qquad \mathbb{E}Z^2 = M\mathbb{E}\|\mathbf{Y}\|_2^2, \tag{8.86}$$

and

$$\sigma^2 = \sup_{\|\mathbf{x}\|_2 \le 1} \mathbb{E}|\langle \mathbf{x}, \mathbf{Y}\rangle|^2 \,.$$

Then, for $t > 0$,

$$\mathbb{P}(Z \ge \sqrt{\mathbb{E}Z^2} + t) \le \exp\left(-\frac{t^2/2}{M\sigma^2 + 2K\sqrt{\mathbb{E}Z^2} + tK/3} \right) . \tag{8.87}$$

Proof. The formula for $\mathbb{E}Z^2$ in (8.86) follows from independence and since $\mathbb{E}\mathbf{Y}_\ell = 0$,

$$\mathbb{E}Z^2 = \sum_{\ell,k=1}^M \mathbb{E}\langle \mathbf{Y}_\ell, \mathbf{Y}_k\rangle = \sum_{\ell=1}^M \mathbb{E}\|\mathbf{Y}_\ell\|_2^2 = M\mathbb{E}\|\mathbf{Y}\|_2^2 \,.$$

By Hölder's inequality $\mathbb{E}Z \le \sqrt{\mathbb{E}Z^2}$. Therefore, the claim is a consequence of Corollary 8.41. □

The so-called weak variance σ^2 in (8.84) can be estimated by

$$\sigma^2 = \sup_{\mathbf{x}\in B^*} \mathbb{E}|\langle \mathbf{x}, \mathbf{Y}\rangle|^2 \le \mathbb{E} \sup_{\mathbf{x}\in B^*} |\langle \mathbf{x}, \mathbf{Y}\rangle|^2 = \mathbb{E}\|\mathbf{Y}\|^2 \,. \tag{8.88}$$

Hence, the variance term σ^2 can be replaced by $\mathbb{E}\|\mathbf{Y}\|^2$ in Theorem 8.39 and Corollaries 8.41 and 8.42. Usually, however, σ^2 provides better estimates that $\mathbb{E}\|\mathbf{Y}\|^2$. In any case, noting that $\|\mathbf{Y}\| \le K$ almost surely implies $\sigma^2 \le \mathbb{E}\|\mathbf{Y}\|^2 \le K^2$ yields the next statement.

Corollary 8.43. *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_M$ be independent copies of a random vector $\mathbf{Y}$ on $\mathbb{C}^n$ satisfying $\mathbb{E}\mathbf{Y} = 0$. Assume $\|\mathbf{Y}\| \le K$ for some constant $K > 0$ and some norm $\|\cdot\|$ on $\mathbb{C}^n$. Let $Z = \left\|\sum_{\ell=1}^M \mathbf{Y}_\ell\right\|$. Then, for $t > 0$,*

$$\mathbb{P}(Z \ge \mathbb{E}Z + t) \le \exp\left(-\frac{t^2/2}{MK^2 + 2K\mathbb{E}Z + Kt/3} \right) . \tag{8.89}$$

We will derive the Bernstein type inequality for suprema of empirical processes as a consequence of a more general deviation inequality for functions in independent random variables. Its formulation needs some notation.

For a sequence $\mathbf{X} = (X_1, \ldots, X_n)$ of independent random variables (or random vectors) we will write $\mathbf{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. We recall the notation

$$\mathbb{E}_{X_i} f(\mathbf{X}) = \mathbb{E}_{X_i} \left[f(X_1, \ldots, X_i, \ldots, X_n) \right] := \mathbb{E} \left[f(\mathbf{X}) | \mathbf{X}^{(i)} \right] \tag{8.90}$$

for the conditional expectation, which is still a function of the random variables $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$. In other words, $\mathbb{E}_{X_i} f(\mathbf{X})$ "integrates out" the dependence in X_i, and is constant with respect to X_i. Further, we recall the function h defined in (8.23), that is,

$$h(x) := (1+x)\ln(1+x) - x.$$

Then the Bernstein type inequality for functions in independent random variables reads as follows.

Theorem 8.44. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of independent random variables (or vectors). Let f, g_i, $i = 1, \ldots, n$, be measurable functions of $\mathbf{X}$ and f_i, $i \in [n]$, be measurable functions of $\mathbf{X}^{(i)}$. Assume that*

$$g_i(\mathbf{X}) \le f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}) \le 1 \,, \quad i \in [n] \,, \tag{8.91}$$

$$\text{and} \quad \mathbb{E}_{X_i} \left[g_i(\mathbf{X}) \right] \ge 0 \,, \quad i \in [n] \,, \tag{8.92}$$

as well as

$$\sum_{i=1}^{n} (f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})) \le f(\mathbf{X}) \,. \tag{8.93}$$

Suppose further that there exists $B, \sigma > 0$ such that

$$g_i(\mathbf{X}) \le B \,, i \in [n] \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{X_i} \left[g_i(\mathbf{X})^2 \right] \le \sigma^2 \,. \tag{8.94}$$

Set $v = (1+B)\mathbb{E}\left[f(\mathbf{X}) \right] + n\sigma^2$. Then, for all $\lambda > 0$,

$$\ln \mathbb{E} \left[e^{\lambda (f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})])} \right] \le v(e^\lambda - \lambda - 1) \,. \tag{8.95}$$

As a consequence, for $t > 0$,

$$\mathbb{P} \left(f(\mathbf{X}) \ge \mathbb{E}[f(\mathbf{X})] + t \right) \le \exp\left(-v h\left(\frac{t}{v} \right) \right) \le \exp\left(-\frac{t^2}{2v + 2t/3} \right) \,. \tag{8.96}$$

Before we prove this theorem, we show how it implies the Bernstein type inequality (8.83) for suprema of empirical processes.

Proof (of Theorem 8.39). We assume that $K = 1$. The general case is deduced via replacing F by F/K.

Suppose first that $\mathcal{F}$ is a finite set. Let $\mathbf{Y} = (Y_1, \ldots, Y_M)$. We define

$$f(\mathbf{Y}) := \sup_{\mathsf{F} \in \mathcal{F}} \sum_{\ell=1}^{M} \mathsf{F}(Y_\ell) = Z$$

and, for $i \in [M]$ we set

$$\begin{aligned} f_i(\mathbf{Y}^{(i)}) &:= \sup_{\mathsf{F} \in \mathcal{F}} \sum_{\ell \neq i} \mathsf{F}(Y_\ell) \ , \\ \text{and} \quad g_i(\mathbf{Y}) &:= \left(\sum_{\ell=1}^{M} \mathsf{F}_i(Y_\ell) \right) - f_i(\mathbf{Y}^{(i)}) = \mathsf{F}_i(Y_i) \ , \end{aligned}$$

where F_i is the function for which the supremum is attained in the definition of f_i (recall that $\mathcal{F}$ is assumed to be finite). Note that F_i may depend on $\mathbf{Y}^{(i)}$, but not on Y_i. Further, F_0 denotes the function for which the supremum is attained in the definition of f. We obtain

$$g_i(\mathbf{Y}) \leq f(\mathbf{Y}) - f_i(\mathbf{Y}) \leq \sum_{\ell=1}^{M} \mathsf{F}_0(Y_\ell) - \sum_{\ell \neq i} \mathsf{F}_0(Y_\ell) = \mathsf{F}_0(Y_i) \leq 1 \ .$$

This verifies Condition (8.91) and the first condition in (8.94) with $B = 1$. Moreover, since F_i is independent of Y_i and $\mathbb{E}[\mathsf{F}_i(Y_i)] = 0$

$$\mathbb{E}_{\mathbf{Y}_i} g_i(\mathbf{Y}) = \mathbb{E}_{\mathbf{Y}_i} \left[\sum_{\ell=1}^{M} \mathsf{F}_i(\mathbf{Y}_\ell) - f_i(\mathbf{Y}_i) \right] = \sum_{\ell \neq i} \mathsf{F}_i(\mathbf{Y}_\ell) - f_i(\mathbf{Y}^{(i)}) = 0 \ ,$$

which shows (8.92). Moreover,

$$(M-1) f(\mathbf{X}) = \sum_{i=1}^{M} \sum_{k \neq i} \mathsf{F}_0(\mathbf{Y}_k) \leq \sum_{i=1}^{M} f_i(\mathbf{Y}^{(i)}),$$

so that also (8.93) is satisfied. Finally,

$$\sum_{i=1}^{M} \mathbb{E}_{\mathbf{Y}_i}[g_i(\mathbf{Y})^2] = \sum_{i=1}^{M} \mathbb{E}_{\mathbf{Y}_i}[\mathsf{F}_i(\mathbf{Y}_i)^2] \leq \sum_{i=1}^{M} \sigma_i^2 \ ,$$

which shows that we can choose σ as desired noting that $B = 1$ in (8.94). An application of Theorem 8.44 yields (8.83) for finite $\mathcal{F}$.

To conclude the proof for countably infinite $\mathcal{F}$, we let $G_n \subset \mathcal{F}$, $n \in \mathbb{N}$ be a sequence of finite subsets, such that $G_n \subset G_{n+1}$ and $\cup_{n \in N} G_n = \mathcal{F}$. Introduce the random variables

$$Z_n := \sup_{\mathsf{F}\in G_n} \sum_{\ell=1}^{M} \mathsf{F}(\mathbf{Y}_\ell)$$

and, for $t > 0$, the characteristic random variables $\chi_n := I_{\{Z_n - \mathbb{E}Z_n > t\}}$. We have the pointwise limit

$$\lim_{n\to\infty} \chi_n = \chi,$$

where χ is the characteristic random variable of the event $\{Z - \mathbb{E}Z > t\}$. Clearly, $\chi_n \le 1$, so that the sequence χ_n has the integrable majorant 1. It follows from Lebesgue's dominated convergence theorem that

$$\begin{aligned}\mathbb{P}(Z > \mathbb{E}Z + t) &= \mathbb{P}(\sup_n (Z_n - \mathbb{E}Z_n) > t) = \mathbb{E}\left[\lim_{n\to\infty} \chi_n\right] = \lim_{n\to\infty} \mathbb{E}\chi_n \\ &= \lim_{n\to\infty} \mathbb{P}\left(\sup_{\mathsf{F}\in G_n} \sum_{\ell=1}^{M} \mathsf{F}(\mathbf{Y}_\ell) > \mathbb{E}\left[\sup_{\mathsf{F}\in G_n} \sum_{\ell=1}^{M} \mathsf{F}(\mathbf{Y}_\ell)\right] + t\right) \\ &\le \exp\left(-\frac{t^2/2}{v_M + tK/3}\right) ,\end{aligned}$$

where we have used the just established estimate for finite sets of functions in the last step. □

The proof of Theorem 8.44 uses the concept of *entropy* (not to be confused with the entropy numbers defined in Section C.2). We introduce the convex function

$$\phi(x) := x\ln(x) , \quad x > 0 .$$

For a nonnegative random variable X on some probability space $(\Omega, \Sigma, \mathbb{P})$ we then define the entropy as

$$\mathscr{E}(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}X) = \mathbb{E}[X\ln X] - \mathbb{E}X\ln(\mathbb{E}X) . \tag{8.97}$$

If the first term is infinite then we set $\mathscr{E}(X) = \infty$. By convexity of ϕ, it follows from Jensen's inequality that $\mathscr{E}(X) \ge 0$. The entropy is homogeneous, that is, for a scalar $t > 0$,

$$\begin{aligned}\mathscr{E}(tX) &= \mathbb{E}[tX\ln(tX)] - \mathbb{E}[tX]\ln(t\mathbb{E}X) \\ &= t\mathbb{E}[X\ln X] + t\mathbb{E}[X\ln t] - t\mathbb{E}X\ln t + \mathbb{E}X\ln(\mathbb{E}X) = t\mathscr{E}(X) .\end{aligned}$$

The basic idea of the entropy method is to derive a bound on the entropy of the random variable $e^{\lambda X}$, for $\lambda > 0$, of the form

$$\mathscr{E}(e^{\lambda X}) \le g(\lambda)\mathbb{E}[e^{\lambda X}]$$

for some appropriate function g. Setting $F(\lambda) := \mathbb{E}[e^{\lambda X}]$ such an inequality is equivalent to

$$\mathscr{E}(e^{\lambda X}) = \lambda F'(\lambda) - F(\lambda)\ln F(\lambda) \le g(\lambda)F(\lambda).$$

Setting further $G(\lambda) = \lambda^{-1} \ln F(\lambda)$ yields then

$$G'(\lambda) \leq \lambda^2 g(\lambda) \; .$$

Noting that $G(0) = \lim_{\lambda \to 0} \lambda^{-1} \ln F(\lambda) = F'(0)/F(0) = \mathbb{E}[X]$ this shows by integration that $G(\lambda) - \mathbb{E}[X] \leq \int_0^\lambda t^2 g(t) dt$, or

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \exp\left(\lambda \int_0^\lambda t^2 g(t) dt \right) \; , \quad \lambda > 0 \; . \tag{8.98}$$

Then one uses Markov's inequality to derive a tail bound.

Below, a slight variation of this idea is worked out in our specific situation. To this end we first provide the following dual characterizations of entropy.

Lemma 8.45. *Let X be a strictly positive and integrable random variable. Then*

$$\mathscr{E}(X) = \sup \left\{ \mathbb{E}(XY) : \mathbb{E}[\exp(Y)] \leq 1 \right\} \; . \tag{8.99}$$

Proof. By homogeneity of the entropy we may and do assume $\mathbb{E}X = 1$. Young's inequality (B.10) yields for Y satisfying $\mathbb{E}[\exp(Y)] \leq 1$,

$$\mathbb{E}[XY] \leq \mathbb{E}[X \ln X] - \mathbb{E}[X] + \mathbb{E}[\exp(Y)] \leq \mathbb{E}[X \ln X] = \mathscr{E}(X) \; .$$

This shows that the right hand side in (8.99) is smaller or equal to the left hand side. For the converse direction choose $Y = \ln X - \mathbb{E}[\ln X]$, so that $\mathscr{E}(X) = \mathbb{E}[XY]$. This choice satisfies

$$\mathbb{E} \exp(Y) = \mathbb{E}[X] \exp\left(-\mathbb{E} \ln X\right) = 1 \; ,$$

by Jensen's inequality. Therefore, the right hand side in (8.99) majorizes $\mathscr{E}(X)$. □

Remark 8.46. Substituting $Y = \ln(Z/\mathbb{E}Z)$ for a positive random variable Z in (8.99) shows that

$$\mathscr{E}(X) = \sup \left\{ \mathbb{E}[X \ln(Z)] - \mathbb{E}[X] \ln(\mathbb{E}[Z]) : Z > 0 \right\} \; , \tag{8.100}$$

where the supremum is taken over all positive integrable random variables Z.

Next, we provide another characterization of entropy.

Lemma 8.47. *Let X be a strictly positive and integrable random variable. Then*

$$\mathscr{E}(X) = \inf_{u > 0} \mathbb{E}[\phi(X) - \phi(u) - (X - u)\phi'(u)] \; ,$$

where $\phi'(x) = \ln(x) + 1$.

Proof. Convexity of ϕ implies that, for $u > 0$,

$$\phi(\mathbb{E}X) \geq \phi(u) + \phi'(u)(\mathbb{E}X - u) .$$

By definition of the entropy this yields

$$\begin{aligned} \mathscr{E}(X) &= \mathbb{E}[\phi(X)] - \phi(\mathbb{E}X)] \leq \mathbb{E}[\phi(X)] - \phi(u) - \phi'(u)(\mathbb{E}X - u) \\ &= \mathbb{E}[\phi(X) - \phi(u) - \phi'(u)(X - u)] . \end{aligned} \tag{8.101}$$

Choosing $u = \mathbb{E}X$ yields an equality above, which proves the claim. □

For a sequence $\mathbf{X} = (X_1, \ldots, X_n)$ and a function f on $\mathbf{X}$ we recall the conditional expectation $\mathbb{E}_{X_i} f(\mathbf{X})$ in (8.90). Then we define the conditional entropy of $f(\mathbf{X})$, for any $i \in [n]$, as

$$\begin{aligned} \mathscr{E}_{X_i}(f(\mathbf{X})) &:= \mathscr{E}\left(f(\mathbf{X})|\mathbf{X}^{(i)}\right) := \mathbb{E}_{X_i}\left(\phi(f(\mathbf{X}))\right) - \phi\left(\mathbb{E}_{X_i}(f(\mathbf{X}))\right) \\ &= \mathbb{E}_{X_i}[f(\mathbf{X}) \ln f(\mathbf{X})] - \mathbb{E}_{X_i}[f(\mathbf{X})] \ln\left(\mathbb{E}_{X_i}[f(\mathbf{X})]\right) . \end{aligned}$$

Clearly, $\mathscr{E}_{X_i}(f(\mathbf{X}))$ is still a random variable that depends on $X^{(i)}$, that is, entropy is taken only with respect to X_i. The tensorization inequality for entropy reads as follows.

Proposition 8.48. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables and let f be an integrable function of $\mathbf{X}$. Then*

$$\mathscr{E}(f(\mathbf{X})) \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathscr{E}_{X_i}(f(\mathbf{X}))\right] . \tag{8.102}$$

Proof. We introduce the conditional expectation operator $\mathbb{E}^i$,

$$\mathbb{E}^i[f(\mathbf{X})] := \mathbb{E}_{X_1,\ldots,X_{i-1}}[f(X)] = \mathbb{E}[f(\mathbf{X})|X_i, \ldots, X_n],$$

which "integrates out" the dependence on the first $i-1$ random variables $X_1, \ldots, X_{i-1}$. Clearly, $\mathbb{E}^1[f(\mathbf{X})] = f(\mathbf{X})$ and $\mathbb{E}^{n+1}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{X})]$. We have the following decomposition by a telescoping sum,

$$\ln(f(\mathbf{X})) - \ln(\mathbb{E}[f(\mathbf{X})] = \sum_{i=1}^{m} (\ln(\mathbb{E}^i[f(\mathbf{X})]) - \ln(\mathbb{E}^{i+1}[f(\mathbf{X})])) . \tag{8.103}$$

Multiplying by $f(\mathbf{X})$, the duality formula (8.100) with $Z = \mathbb{E}^i[f(\mathbf{X})]$ yields

$$\mathbb{E}_{X_i}\left[f(\mathbf{X})\left(\ln(\mathbb{E}^i[f(\mathbf{X})]) - \ln(\mathbb{E}_{X_i}[\mathbb{E}^i[f(\mathbf{X})]])\right)\right] \leq \mathscr{E}_{X_i}\left(f(\mathbf{X})\right) .$$

Observe that by independence and Fubini's theorem

$$\mathbb{E}_{X_i}[\mathbb{E}^i[f(\mathbf{X})]] = \mathbb{E}_{X_i}\mathbb{E}_{X_1,\ldots,X_{i-1}}[f(\mathbf{X})] = \mathbb{E}^{i+1}[f(\mathbf{X})] .$$

Taking expectations on both sides of (8.103) yields

$$\begin{aligned}\mathscr{E}(f(\mathbf{X})) &= \mathbb{E}[f(\mathbf{X})(\ln(f(\mathbf{X})) - \ln(\mathbb{E}[f(\mathbf{X})])] \\ &= \sum_{i=1}^{m} \mathbb{E}\left[\mathbb{E}_{X_i}\left[f(\mathbf{X})(\ln(\mathbb{E}^i[f(\mathbf{X})]) - \ln(\mathbb{E}_{X_i}[\mathbb{E}^i[f(\mathbf{X})]])\right]\right] \\ &\le \sum_{i=1}^{m} \mathbb{E}\left[\mathscr{E}_{X_i}(f(\mathbf{X}))\right] \ .\end{aligned}$$

This completes the proof. □

We will need the following consequence of the tensorization inequality.

Corollary 8.49. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of independent random vectors. Let f be a measurable function of $\mathbf{X}$ and f_i, $i \in [n]$, be measurable functions of $\mathbf{X}^{(i)}$ (that is, constant in X_i). Then, for any $\lambda \in \mathbb{R}$ such that $\mathbb{E}[\exp(\lambda f(\mathbf{X}))] < \infty$,*

$$\lambda \mathbb{E}\left[f(\mathbf{X})e^{\lambda f(\mathbf{X})}\right] - \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\psi\big(-\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}))\big)\right],$$

where $\psi(x) := e^x - x - 1$.

Proof. Suppose g is a positive measurable function of X and g_i are positive measurable functions of $\mathbf{X}^{(i)}$, $i \in [n]$. Taking the entropy conditionally with respect to X_i in Lemma 8.47 (i.e., choosing $u = g_i(X^{(i)})$ in (8.101), so that u does not depend on X_i) yields

$$\begin{aligned}&\mathscr{E}_{X_i}(g(\mathbf{X})) \le \mathbb{E}_{X_i}\left[\phi(g(\mathbf{X})) - \phi(g_i(\mathbf{X}^{(i)})) - (g(\mathbf{X}) - g_i(\mathbf{X}^{(i)}))\phi'(g_i(\mathbf{X}^{(i)}))\right] \\ &= \mathbb{E}_{X_i}\left[g(\mathbf{X})(\ln(g(\mathbf{X})) - \ln(g_i(\mathbf{X}^{(i)}))) - (g(\mathbf{X}) - g_i(\mathbf{X}^{(i)})\right] \ . \qquad (8.104)\end{aligned}$$

We apply the above inequality to $g(\mathbf{X}) = e^{\lambda f(\mathbf{X})}$ and $g_i(\mathbf{X}) = e^{\lambda f_i(\mathbf{X}^{(i)})}$ to obtain

$$\begin{aligned}\mathscr{E}_{X_i}(g(\mathbf{X})) &= \lambda \mathbb{E}_{X_i}\left[f(\mathbf{X})e^{\lambda f(\mathbf{X})}\right] - \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}\right] \ln \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}\right] \\ &\le \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}(\lambda f(\mathbf{X}) - \lambda f_i(\mathbf{X}^{(i)})) - (e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})})\right] \\ &= \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}\psi(-\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}))\right] \ .\end{aligned}$$

An application of the tensorization inequality (8.102) shows that

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{X})e^{\lambda f(\mathbf{X})}\right] - \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}\right]\ln \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}\right] \\
&= \mathscr{E}(g(\mathbf{X})) \leq \mathbb{E}\left[\sum_{i=1}^{m} \mathscr{E}_{X_i}(g(\mathbf{X}))\right] \\
&\leq \mathbb{E}\left[\sum_{i=1} \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})}\psi(-\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})))\right]\right] \\
&= \sum_{i=1} \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\psi(-\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})))\right] .
\end{aligned}$$

This completes the proof. □

As the next auxiliary tool we need the following decoupling inequality.

Lemma 8.50. *Let Y, Z be random variables on a probability space $(\Omega, \Sigma, \mathbb{P})$ and $\lambda > 0$ such that $e^{\lambda Y}, e^{\lambda Z}$ are $\mathbb{P}$-integrable. Then,*

$$\lambda\mathbb{E}\left[Ye^{\lambda Z}\right] \leq \lambda\mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\ln\mathbb{E}\left[e^{\lambda Z}\right] + \mathbb{E}\left[e^{\lambda Z}\right]\ln\mathbb{E}\left[e^{\lambda Y}\right] .$$

Proof. Let $\mathbb{Q}$ be the probability measure defined via $d\mathbb{Q} = \frac{e^{\lambda Y}}{\mathbb{E}[e^{\lambda Y}]}d\mathbb{P}$, and associated expectation given by

$$\mathbb{E}_{\mathbb{Q}}[X] := \frac{\mathbb{E}[Xe^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} ,$$

where $\mathbb{E}$ is the expectation with respect to $\mathbb{P}$. Jensen's inequality yields

$$\lambda\mathbb{E}_{\mathbb{Q}}[Y - Z] = \mathbb{E}_{\mathbb{Q}}\left[\ln(e^{\lambda(Y-Z)})\right] \leq \ln\mathbb{E}_{\mathbb{Q}}\left[e^{\lambda(Y-Z)}\right] .$$

By definition of $\mathbb{E}_{\mathbb{Q}}$ this translates into

$$\frac{\lambda\mathbb{E}[(Y - Z)e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \leq \ln\mathbb{E}[e^{\lambda Y}] - \ln\mathbb{E}[e^{\lambda Z}] ,$$

which is equivalent to the claim. □

The next statement is a consequence of Lemma 8.50 and Corollary 8.49.

Lemma 8.51. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of independent random variables (vectors). Let f be a measurable function of $\mathbf{X}$ and f_i, $i \in [n]$, be measurable functions of $\mathbf{X}^{(i)}$. Let further g be a measurable function of $\mathbf{X}$ such that*

$$\sum_{i=1}^{n}\left(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})\right) \leq g(\mathbf{X}) . \tag{8.105}$$

Then, for all $\lambda > 0$,

$$\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})}\right] \leq \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\right]\ln\mathbb{E}\left[e^{\lambda g(\mathbf{X})}\right] .$$

Proof. Denote $F(\lambda) = \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\right]$ and $G(\lambda) = \mathbb{E}\left[e^{\lambda g(\mathbf{X})}\right]$. Observe that $F'(\lambda) = \mathbb{E}\left[f(\mathbf{X})e^{\lambda f(\mathbf{X})}\right]$. We apply Corollary 8.49 to $f(\mathbf{X})$ and $\widetilde{f}_i(\mathbf{X}^{(i)}) = f_i(\mathbf{X}^{(i)}) + \frac{1}{n\lambda}\ln G(\lambda)$, $i = 1, \ldots, n$, to obtain

$$
\begin{aligned}
&\lambda F'(\lambda) - F(\lambda)\ln F(\lambda) \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\psi\left(-\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}) - \ln(G(\lambda))/(n\lambda)\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[G(\lambda)^{1/n}e^{\lambda f_i(\mathbf{X}^{(i)})} - e^{\lambda f(\mathbf{X})} + e^{\lambda f(\mathbf{X})}(\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})) - \frac{1}{n}\ln G(\lambda)\right] \\
&\leq G(\lambda)^{1/n}\left(\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})}\right]\right) - nF(\lambda) + \lambda\mathbb{E}\left[e^{\lambda f(\mathbf{X})}\sum_{i=1}^{n}(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}))\right] \\
&\quad - F(\lambda)\ln G(\lambda) \\
&\leq G(\lambda)^{1/n}\left(\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})}\right]\right) - nF(\lambda) + \lambda\mathbb{E}\left[e^{\lambda f(\mathbf{X})}g(\mathbf{X})\right] - F(\lambda)\ln G(\lambda) \\
&\leq G(\lambda)^{1/n}\left(\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})}\right]\right) - nF(\lambda) + \lambda\mathbb{E}\left[f(\mathbf{X})e^{\lambda f(\mathbf{X})}\right] \\
&\quad - \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\right]\ln\mathbb{E}\left[e^{\lambda f(\mathbf{X})}\right] + \mathbb{E}\left[e^{\lambda f(\mathbf{X})}\right]\ln\mathbb{E}\left[e^{\lambda g(\mathbf{X})}\right] - F(\lambda)\ln G(\lambda) \\
&= G(\lambda)^{1/n}\left(\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})}\right]\right) - nF(\lambda) + \lambda F'(\lambda) - F(\lambda)\ln F(\lambda)\,.
\end{aligned}
$$

Hereby, we used the assumption (8.105) in the sixth line, and Lemma 8.50 in the last inequality. We rewrite this as

$$
nF(\lambda) \leq G(\lambda)^{1/n}\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})}\right]\,,
$$

which in turn is equivalent to

$$
\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})}\right] \leq nF(\lambda)(1 - G(\lambda)^{-1/n})\,.
$$

The inequality $e^x \geq 1 + x$ implies then that $n(1 - G(\lambda)^{-1/n}) = n(1 - e^{-\frac{1}{n}\ln G(\lambda)}) \leq \ln G(\lambda)$, so that

$$
\sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})}\right] \leq F(\lambda)\ln G(\lambda)\,.
$$

This completes the proof. □

Based on this preparation we can now prove Theorem 8.44.

Proof (of Theorem 8.44). We define $\alpha(x) := 1-(1+x)e^{-x}$, $\beta(x) := e^{-x}-1+x$, and for $\tau > 0$ to be specified later,

$$\gamma(x) := \frac{\alpha(-x)}{\beta(-x)+\lambda\tau} .$$

Step 1: We prove that, for $x \le 1$, $\lambda, \tau > 0$,

$$\beta(\lambda x) \le \gamma(x)\left(\alpha(\lambda x) + \lambda\tau x^2 e^{-\lambda x}\right) . \tag{8.106}$$

To this end we introduce the function

$$b(x) := \beta(\lambda x) - \gamma(x)\left(\alpha(\lambda x) + \lambda\tau x^2 e^{-\lambda x}\right) .$$

Note that $\alpha(0) = \beta(0) = \alpha'(0) = \beta'(0) = 0$ so that $b(0) = b'(0) = 0$. Furthermore,

$$\alpha(-\lambda) + \lambda\tau = e^{\lambda}(1 - e^{-\lambda} - \lambda e^{-\lambda} + \lambda\tau e^{-\lambda}) = e^{\lambda}(\beta(\lambda) + \tau\lambda e^{-\lambda}) ,$$

which implies that $b(1) = 0$. Furthermore,

$$b'(x) = \lambda\left(1 - e^{-\lambda x} - f(\lambda)(\lambda x e^{-\lambda x} + 2\tau x e^{-\lambda x} - \tau\lambda x^2 e^{-\lambda x})\right) .$$

Therefore, $\lim_{x\to+\infty} b'(x) = \lambda$ and $\lim_{x\to-\infty} b'(x) = +\infty$. Next, observe that we can write $b''(x) = e^{-\lambda x}p(x)$ with a second degree polynomial p with leading term $-\lambda^3\gamma(\lambda)\tau$. If follows that $b''(x) = 0$ has at most two solutions. If there is no solution then b' is decreasing, which is a contradiction to $\lim_{x\to-\infty} b'(x) = +\infty$, $b'(0) = 0$ and $\lim_{x\to+\infty} b'(x) = \lambda$. So let x_1, x_2 with $x_1 \le x_2$ be the (possibly equal) solutions. Then b' is decreasing in $(-\infty, x_1) \cup (x_2, \infty)$ and increasing in (x_1, x_2). Since $\lim_{x\to+\infty} b'(x) = \lambda > 0$, the equation $b'(x) = 0$ can have at most two solutions, one in $(-\infty, x_1)$ and one in $[x_1, x_2)$. Recall that $b'(0) = 0$, so denote by x_3 the other solution to $b'(x) = 0$. If $x_3 \le 0$ then b is increasing in $(0, \infty)$, which is a contradiction to $b(0) = b(1) = 0$ and $\lambda > 0$. Therefore, $x_3 > 0$ and b is increasing in $(-\infty, 0)$, decreasing in $(0, x_3)$ and increasing in (x_3, ∞). Since $b(0) = b(1) = 0$ this shows that $b(x) \le 0$ for $x \le 1$, which implies the claim inequality (8.106).

Step 2: Next we use (8.106) with $x = f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})$ to obtain

$$\begin{aligned}
&\beta(\lambda(f(\mathbf{X}) - f(\mathbf{X}^{(i)}))e^{\lambda f(\mathbf{X})} \\
&\le \gamma(\lambda)\left(\beta(-\lambda(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})) + \lambda\tau(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}))^2\right) \\
&= \gamma(\lambda)\left(e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})}\right) \\
&\quad + \lambda\gamma(\lambda)e^{\lambda f(\mathbf{X}^{(i)})}\left(\tau(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}))^2 - (f(\mathbf{X}) - f(\mathbf{X}^{(i)}))\right) .
\end{aligned} \tag{8.107}$$

Now we choose $\tau = 1/(1+B)$. Note that if $y \le x \le 1$ and $y \le B$ then

$$\tau x^2 - x \le \tau y^2 - y . \tag{8.108}$$

Indeed, under these assumptions,

$$\tau(x^2 - y^2) = \tau(x+y)(x-y) \le \tau(1+B)(x-y) = x - y \; .$$

Using the assumption $g_i(\mathbf{X}) \le f(\mathbf{X}) - f_i(\mathbf{X}^{(i)}) \le 1$ and $g_i(\mathbf{X}) \le B$ in (8.107) and exploiting (8.108) we get

$$\begin{aligned} &\beta \left(\lambda (f(\mathbf{X}) - f(\mathbf{X}^{(i)})) \right) e^{\lambda f(\mathbf{X})} \le \gamma(\lambda) \left(e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})} \right) \\ &+ \lambda \gamma(\lambda) e^{\lambda f_i(\mathbf{X}^{(i)})} \left(\tau g_i^2(\mathbf{X}) - g_i(\mathbf{X}) \right) \; . \end{aligned} \tag{8.109}$$

Since $f_i(\mathbf{X}^{(i)})$ does not depend on X_i, the assumption $\mathbb{E}_{X_i} g_i(\mathbf{X}) \ge 0$ yields

$$\begin{aligned} \mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})} g_i(\mathbf{X}) \right] &= \mathbb{E}\left[\mathbb{E}_{X_i} e^{\lambda f_i(\mathbf{X}^{(i)})} g_i(\mathbf{X}) \right] = \mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})} \mathbb{E}_{X_i} g_i(\mathbf{X}) \right] \\ &\ge \mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})} \right] \; . \end{aligned}$$

Further note that (8.91) and (8.92) imply that

$$\mathbb{E}_{X_i}[f(\mathbf{X})] \ge \mathbb{E}_{X_i}[f_i(\mathbf{X}^{(i)}) - g_i(\mathbf{X})] \ge f_i(\mathbf{X}^{(i)}) \; ,$$

and by Jensen's inequality this yields

$$e^{\lambda f_i(\mathbf{X}^{(i)})} \le e^{\lambda \mathbb{E}_{X_i} f(\mathbf{X})} \le \mathbb{E}_{X_i}\left[e^{\lambda f(\mathbf{X})} \right] \; .$$

By taking expectations in (8.109) we therefore reach

$$\begin{aligned} &\mathbb{E}\left[\beta \left(\lambda (f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})) \right) e^{\lambda f(\mathbf{X})} \right] \\ &\le \gamma(\lambda) \mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})} \right] + \frac{\lambda \gamma(\lambda)}{1+B} \mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})} g_i^2(\mathbf{X}) \right] \\ &= \gamma(\lambda) \mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})} \right] + \frac{\lambda \gamma(\lambda)}{1+B} \mathbb{E}\left[e^{\lambda f_i(\mathbf{X}^{(i)})} \mathbb{E}_{X_i}\left[g_i^2(\mathbf{X}) \right] \right] \\ &\le \gamma(\lambda) \mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})} \right] + \frac{\lambda \gamma(\lambda)}{1+B} \mathbb{E}\left[e^{\lambda f(\mathbf{X})} \mathbb{E}_{X_i}\left[g_i^2(\mathbf{X}) \right] \right] \; . \end{aligned}$$

Hereby, we used twice that $\mathbb{E}[\cdot] = \mathbb{E}\mathbb{E}_{X_i}[\cdot]$. Now denote $F(\lambda) = \mathbb{E}[e^{\lambda f(\mathbf{X})}]$. Then Corollary 8.49 together with (8.94) implies that

$$\begin{aligned} &\lambda F'(\lambda) - F(\lambda) \ln F(\lambda) \le \sum_{i=1}^n \mathbb{E}\left[\beta \left(\lambda (f(\mathbf{X}) - f(\mathbf{X}^{(i)})) \right) e^{\lambda f(\mathbf{X})} \right] \\ &\le \gamma(\lambda) \sum_{i=1}^n \mathbb{E}\left[e^{\lambda f(\mathbf{X})} - e^{\lambda f_i(\mathbf{X}^{(i)})} \right] + \frac{\lambda \gamma(\lambda)}{1+B} \mathbb{E}\left[e^{\lambda f(\mathbf{X})} \sum_{i=1}^n \mathbb{E}_{X_i}\left[g_i^2(\mathbf{X}) \right] \right] \\ &\le \gamma(\lambda) \mathbb{E}\left[e^{\lambda f(\mathbf{X})} \right] \ln \mathbb{E}\left[e^{\lambda g(\mathbf{X})} \right] + \frac{\lambda \gamma(\lambda) n \sigma^2}{1+B} \mathbb{E}\left[e^{\lambda f(\mathbf{X})} \right] \; , \\ &= \gamma(\lambda) F(\lambda) \ln F(\lambda) + \frac{\lambda \gamma(\lambda) n \sigma^2}{1+B} F(\lambda) \; , \end{aligned} \tag{8.110}$$

where we used in the last inequality that $\sum_{i=1}^{n}(f(\mathbf{X}) - f_i(\mathbf{X}^{(i)})) \leq f(\mathbf{X})$ in combination with Lemma 8.51.

Step 3: Set $G(\lambda) = \mathbb{E}\left[e^{\lambda(f(\mathbf{X})-\mathbb{E}[f(\mathbf{X})])}\right] = F(\lambda)e^{-\lambda\mathbb{E}[f(\mathbf{X})]}$. Then

$$\begin{aligned} G'(\lambda) &= e^{-\lambda\mathbb{E}[f(\mathbf{X})]}\left(F'(\lambda) - \mathbb{E}[f(\mathbf{X})]F(\lambda)\right) , \\ \ln G(\lambda) &= \ln F(\lambda) - \lambda\mathbb{E}[f(\mathbf{X})] , \\ \text{and} \quad \frac{G'(\lambda)}{G(\lambda)} &= \frac{F'(\lambda)}{F(\lambda)} - \mathbb{E}[f(\mathbf{X})] . \end{aligned}$$

Therefore, (8.110) can be rewritten as

$$\lambda\frac{G'(\lambda)}{G(\lambda)} - \ln G(\lambda) \leq \gamma(\lambda)\left(\ln G(\lambda) + \lambda\mathbb{E}[f(\mathbf{X})]\right) + \frac{n\sigma^2\lambda\gamma(\lambda)}{1+B} .$$

Introducing $L(\lambda) = \ln G(\lambda)$ the above inequality is in turn equivalent to

$$\lambda L'(\lambda) - (1+\gamma(\lambda))L(\lambda) \leq \frac{n\sigma^2 + (1+B)\mathbb{E}[f(\mathbf{X})]}{1+B}\lambda\gamma(\lambda) = \frac{v}{1+B}\lambda\gamma(\lambda) .$$

Recall that we have set $\tau = 1/(1+B)$ in the definition of the function γ, so that

$$\gamma(\lambda) = \frac{\alpha(-\lambda)}{\beta(-\lambda) + \lambda/(1+B)}$$

We claim that $L_0(\lambda) := v\beta(-\lambda) = v(e^{\lambda} - 1 - \lambda)$ is a solution to the associated differential equation

$$\lambda L'(\lambda) - (1+\gamma(\lambda))L(\lambda) = \frac{v}{1+B}\lambda\gamma(\lambda) ,$$

with initial conditions $L_0(0) = L_0'(0) = 0$. Indeed,

$$\begin{aligned} &v^{-1}\left(\lambda L_0'(\lambda) - (1+\gamma(\lambda))L_0(\lambda)\right) \\ &= \lambda(e^{\lambda} - 1) - e^{\lambda} + \lambda + 1 - \frac{\alpha(-\lambda)\beta(-\lambda)}{\beta(-\lambda) + \lambda/(1+B)} \\ &= \alpha(-\lambda) - \frac{\alpha(-\lambda)(\beta(-\lambda) + \lambda/(1+B)}{\beta(-\lambda) + \lambda/(1+B))} + \frac{\alpha(-\lambda)\lambda/(1+B)}{\beta(-\lambda) + \lambda/(1+B)} \\ &= \frac{\lambda\gamma(\lambda)}{1+B} . \end{aligned}$$

It follows from Lemma (C.12) that $L(\lambda) \leq L_0(\lambda)$, that is,

$$\ln \mathbb{E}[e^{\lambda(f(\mathbf{X})-\mathbb{E}[f(\mathbf{X})])}] \leq v(e^{\lambda} - 1 - \lambda) .$$

This completes the proof of (8.95).

Step 4: To deduce the tail inequalities in (8.96) we use Markov's inequality (Theorem 7.3) to obtain, for $\lambda > 0$,

$$\begin{aligned}\mathbb{P}(f(\mathbf{X}) \geq \mathbb{E}[f(\mathbf{X})] + x) &= \mathbb{P}\left(e^{\lambda(f(\mathbf{X})-\mathbb{E}[f(\mathbf{X})])} \geq e^{\lambda x}\right) \\ &\leq e^{-\lambda x}\mathbb{E}[e^{\lambda(f(\mathbf{X})-\mathbb{E}[f(\mathbf{X})])}] \leq e^{-\lambda x} e^{v(e^\lambda - 1 - \lambda)} \\ &= e^{v(e^\lambda - 1 - \lambda) - \lambda x} \,. \end{aligned} \tag{8.111}$$

It follows from Lemma 8.21 that

$$\inf_{\lambda > 0} (v(e^\lambda - \lambda - 1) - \lambda x) = -vh(x/v) \,,$$

where we recall that $h(x) = (1+x)\ln(1+x) - x$. Together with (8.111) this shows the first estimate in (8.96). The second part of Lemma 8.21 implies that $vh(x/v) \geq \frac{x^2}{2v + 2x/3}$, which yields the second inequality in (8.96). □

Notes

Many results of this chapter also hold in infinite dimensional Banach spaces. Introducing random vectors in general Banach spaces, however, requires additional technicalities that we prefered to avoid here. For such details and many more results on probability in Banach spaces we refer to the monograph [280] by M. Ledoux and M. Talagrand, and to the collection of articles in [257]. In particular, the relation between moments and tails as well as an introduction to Rademacher sums and to symmetrization are contained in [280].

The Khintchine inequalities are named after the Russian mathematician A. Khintchine (also spelled Khinchin) who was the first to show Theorem 8.5 in [260]. Our proof essentially follows his ideas. We have only provided estimates from above for the absolute moments of a Rademacher sum. Estimates from below have also been investigated and the optimal constants for both lower and upper estimates for all $p > 0$ have been derived in [219], see also [311] for simplified proofs. We have already noted that, for $p = 2n$, $n \in \mathbb{N}$, the constant $C_{2n} = (2n)!/(2^n n!)$ for the upper estimate provided in Theorem 8.5 is optimal. In case of general $p \geq 2$ (which is much harder than the even integer case) the best constant is $C_p = 2^{\frac{p-1}{2}} \Gamma(p/2)/\Gamma(3/2)$. This value is very close to the estimate in (8.9).

The proof of Khintchine's inequality for Steinhaus sums in Theorem 8.9 is slightly shorter than the one given in [328]. The technique for the proof of Corollary 8.10 for Steinhaus sums was taken from [328, 419]. An overview on (scalar) Khintchine and related inequalites can be found in [329]. An extension of the Khintchine inequalities to sums of independent random vectors that are uniformly distributed on spheres is provided in [264]. Using a similar technique as in Corollary 8.10 the following Hoeffding type inequality has been deduced in [160] for $\mathbf{X}_1, \ldots, \mathbf{X}_M \in \mathbb{R}^n$ being independent random vectors, uniformly distributed on the unit sphere $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 = 1\}$,

$$\mathbb{P}\left(\|\sum_{\ell=1}^M a_\ell \mathbf{X}_\ell\|_2 \geq \|\mathbf{a}\|_2 u\right) \leq \exp\left(-\frac{n}{2}(u^2 - \log(u^2) - 1)\right) \quad \text{for all } u > 1 \,.$$

The noncommutative version of Bernstein's inequality was proven by J. Tropp in [424] by refining an approach to the Laplace transform method for matrices due to R. Ahlswede and A. Winter [5], see also [322, 323]. Based on the method of exchangeable pairs, a different approach to its proof, which does not require Lieb's concavity theorem (nor a similar result on matrix convexity), is presented in [292]. The more traditional approach for studying tail bounds for random matrices uses the noncommutative Khintchine inequality, which first appeared in the work of F. Lust-Piquard [289], see also [290]. These inequalities work with the Schatten $2n$-norms $\|A\|_{S_{2n}} = \|\sigma(A)\|_{2n} = (\operatorname{tr}((A^*A)^n))^{1/(2n)}$, $n \in \mathbb{N}$, where $\sigma(A)$ is the vector of singular values of A, and provide bounds for matrix-valued Rademacher sums,

$$\begin{aligned}&\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\mathbf{B}_j\|_{S_{2n}}^{2n}\\&\leq\frac{(2n)!}{2^n n!}\max\left\{\left\|\left(\sum_{j=1}^{M}\mathbf{B}_j\mathbf{B}_j^*\right)^{1/2}\right\|_{S_{2n}}^{2n},\left\|\left(\sum_{j=1}^{M}\mathbf{B}_j^*\mathbf{B}_j\right)^{1/2}\right\|_{S_{2n}}^{2n}\right\}.\end{aligned}\tag{8.112}$$

The optimal constants for these inequalities for $p = 2n$ match the scalar case in Theorem 8.5, and were derived by A. Buchholz in [61, 62], see also [355]. As a consequence of the noncommutative Khintchine inequality, M. Rudelson showed a lemma now named after him in [371], see also [323, 355], which allows to derive tail bounds and moment bounds for sums of random rank-one matrices. The approach to random matrices via the noncommutative Khintchine inequality has the drawback that one needs significant practice in order to apply them, see for instance [437, 355]. In contrast, the noncommutative Bernstein inequality of Theorem 8.14 is easy to apply and provides very good constants.

The decoupling inequality of Theorem 8.11, including its proof, is essentially taken from [55]. The variant Theorem 8.12 for the operator norm was shown by J.Tropp in [418, 419]. Decoupling techniques can be extended to higher order chaos and also to sums of the form $\sum_{j\neq k} h_{j,k}(\mathbf{X}_j, \mathbf{X}_k)$, where the $\mathbf{X}_k$ are independent random vectors and the $h_{j,k}$ are vector-valued functions. Moreover, decoupling inequalities do not only apply for expectations and moments. Also, a probability estimate of the form

$$\mathbb{P}(\|\sum_{j\neq k} h_{j,k}(\mathbf{X}_j, \mathbf{X}_k)\| \geq t) \leq C\mathbb{P}(\|\sum_{j\neq k} h_{j,k}(\mathbf{X}_j, \mathbf{X}'_k)\| \geq t/C)\ ,$$

can be shown, where $\mathbf{X}'_k$ is an independent copy of $\mathbf{X}_k$ and $C > 1$ is an appropriate constant. We refer the interested reader to [123] for further information.

The tail bounds for Rademacher chaos (Theorem 8.13) and quadratic forms in more general subgaussian random vectors have first been obtained

by D. Hanson and F. Wright in [225]. The proof given here follows arguments from a not yet published work of Rauhut and Tropp. For Gaussian chaos better constants are available in [27], and yet another proof of the tail inequality appears in [404, Section 2.5].

Dudley's Theorem 8.23 is named after R. Dudley who proved his inequality in [152]. The proof in Section 8.6 follows the argument in Pisier's book [339]. Further proofs can be found in [18, 168, 169, 355, 404]. In particular for the Gaussian case, X. Fernique's book [169] contains the better constant $4\sqrt{2}$ instead of 12 in (8.47). The nice exposition in M. Talagrand's book [404] leads to more powerful generic chaining inequalities, also called majorizing measure inequalities. These use the so called γ_2-functional of a metric space (T, d), which is defined as

$$\gamma_2(T,d) = \inf \sup_{t\in T} \sum_{r=0}^{\infty} 2^{r/2} d(t, T_r),$$

where the infimum is taken over all sequences T_r, $r \in \mathbb{N}_0$, of subsets of T with cardinalities $\mathrm{card}(T_0) = 1$, $\mathrm{card}(T_r) \leq 2^{2^r}$, $r \geq 1$. Further, $d(t, T_r) = \inf_{s\in T_r} d(t, s)$. Given a subgaussian processes X_t, $t \in T$, with associated pseudo-metric d defined by (8.41), Talagrand's majorizing measures [400, 403, 404] theorem states that

$$C_1\gamma_2(T,d) \leq \mathbb{E} \sup_{t\in T} X_t \leq C_2\gamma_2(T,d) \ ,$$

for universal constants $C_1, C_2 > 0$. In particular, the lower bound is remarkable. Since $\gamma_2(T, d)$ is bounded by a constant times the Dudley type integral in (8.47), see [404], the above inequality implies also Dudley's inequality (with possibly a different constant). In general, $\gamma_2(T, d)$ may provide sharper bounds than Dudley's integral. However, if T is a subset of $\mathbb{R}^N$ and d is induced by a norm, then one looses at most a factor of $\ln(N)$ when passing from the $\gamma_2(T, d)$ functional to Dudley's integral. The latter has the advantage that it is usually easier to estimate. Another type of lower bound for Gaussian processes is Sudakov's minoration, see e.g. [280, 299].

Dudley's inequality extends to moments, see for instance [355]. Indeed, one also has the following inequality (using the same notation as Theorem 8.23)

$$\left(\mathbb{E} \sup_{t\in T} |X_t|^p\right)^{1/p} \leq C\sqrt{p} \int_0^{\Delta(T)/2} \sqrt{\ln(N(T,d,u))}du \ .$$

Estimates for suprema of Gaussian chaos processes of the form $X_t = \sum_{j\neq k} g_j g_k x_{j,k}(t)$, where $\mathbf{g} = (g_1, \ldots, g_N)$ is a standard Gaussian vector, can be found in [404].

A generalization of Dudley's inequality [265, 338, 280] holds in the framework of Orlicz spaces. A Young function is a positive convex function ψ that satisfies $\psi(0) = 0$ and $\lim_{x\to\infty} \psi(x) = \infty$. The Orlicz space L_ψ consists of all random variables X for which $\mathbb{E}\psi(|X|/c) < \infty$ for some $c > 0$. The norm

$$\|X\|_\psi = \inf\{c > 0, \mathbb{E}\psi(|X|/c) \leq 1\}$$

turns L_ψ into a Banach space [270]. Suppose that X_t, $t \in T$, is a stochastic process indexed by a (pseudo-)metric d of diameter Δ such that

$$\|X_s - X_t\|_\psi \leq d(s,t) \, .$$

Then the generalization of Dudley's inequality [280, Theorem 11.1] states that

$$\mathbb{E} \sup_{s,t} |X_s - X_t| \leq 8 \int_0^\Delta \psi^{-1}(N(T,d,u))du \, ,$$

where ψ^{-1} is the inverse function of ψ. Taking $\psi(x) = \exp(x^2) - 1$ yields Theorem 8.23 (up to the constant). Further important special cases are $\psi(x) = \exp(x) - 1$ (exponential tail of the increments) and $\psi(x) = x^p$ (resulting in L^p-spaces of random variables).

In slightly different form, Slepian's lemma appeared for the first time in [388], see also X. Fernique's notes [168]. Other references on Slepian's lemma include [280, Corollary 3.14], [299, Theorem 3.14], [295]. Gordon's Lemma 8.28 appeared in [200, 201].

Many more details and references on the general theory of concentration of measure such as connections to isoperimetric inequalities are provided in the expositions by A. Barvinok [26] and M. Ledoux [279]. The proof of Theorem 8.38 follows [279], while the proof of Theorem 8.35 follows [405, Theorem 1.3.4]. An alternative proof of Theorem 8.38 based on the concept of entropy, see Section 8.9, can be found in [299, Chapter 3]. Indeed, L. Gross' logarithmic Sobolev inequality, which may be derived using the tensorization inequality (8.102) for entropy, states that

$$\mathscr{E}(u^2(\mathbf{X})) \leq 2\mathbb{E}[\|\nabla u(\mathbf{X})\|^2],$$

for a standard Gaussian random vector $\mathbf{X} \in \mathbb{R}^n$, and any continuously differentiable function $u : \mathbb{R}^n \to \mathbb{R}$ [299, Theorem 3.9], [213]. Setting $u = e^{\lambda F}$ shows the inequality $\mathscr{E}[e^{\lambda F(\mathbf{X})}] \leq \lambda^2 L^2/2\mathbb{E}[e^{\lambda F(\mathbf{X})}]$. Following the arguments leading to (8.98) (the so-called Herbst argument [299, Proposition 2.14]), and applying Markov's inequality leads then to Theorem 8.38. By using rotation invariance of the Gaussian distribution, concentration of measure for the uniform distribution on the sphere (or on the ball) can be deduced from the Gaussian case (and vice versa), see for instance [26, 280, 279].

Concentration of measure inequalities are valid also for independent random variables $X_1, \ldots, X_n$ with values in $[-1, 1]$. However, one has to impose the assumption that the function $F : [0,1]^n \to \mathbb{R}$ is convex, in addition to being L-Lipschitz. Denoting by M a median, that is, a number such that $\mathbb{P}(F(X_1, \ldots, X_n) \geq M) = 1/2$, then [279, 399]

$$\mathbb{P}(|F(X_1, \ldots, X_n) - M| \geq t) \leq 4\exp(-t^2/(4L)^2) \, .$$

The median can replace the mean via general principles outlined in [279].

Deviation inequalities for suprema of empirical processes were already investigated in the 1980ies by P. Massart and others, see e.g. [297, 8]. M. Talagrand achieved major breakthroughs in [398, 401]. In particular, he showed a concentration inequality similar to (8.83) in [401], see also [279, Theorem 7.6]. M. Ledoux noticed in [278] that deviation and concentration inequalities may be deduced using entropy. The constants in the deviation and concentration inequalities were successfully improved in [298, 364, 365, 57, 58, 263]. The proof of Theorem 8.39 follows [57], see also [58]. Background on the entropy method can be found e.g. in [299]. Concentration below the expected supremum of an empirical can be shown as well [401, 278, 58]. A version for not necessarily identically distributed random vectors is presented in [263], and collections $\mathcal{F}$ of unbounded functions are treated in [2, 277].

Versions of Corollary 8.43 can already be found in the monograph by M. Ledoux and M. Talagrand [280, Theorems 6.17 and 6.19] , however, with non-optimal constants. More general deviation and concentration inequalities for suprema of empirical processes and other functions of independent variables are derived for instance in [50, 49, 279], in particular, a version for Rademacher chaos processes is stated in [50].

Exercises

8.1. Let $X = (X_1, \dots, X_n)$ be a vector of mean zero Gaussians with variances $\sigma_\ell^2 = \mathbb{E} g_\ell^2$, $\ell \in [n]$. Show that

$$\mathbb{E} \max_{\ell \in [n]} X_\ell \le \sqrt{2 \ln(n)} \max_{\ell \in [n]} \sigma_\ell^2 \, .$$

8.2. Comparison principle.
Let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ be a Rademacher sequence and $\mathbf{g} = (g_1, \dots, g_N)$ be a standard Gaussian vector. Let $\mathbf{x}_1, \dots, \mathbf{x}_M$ be vectors in a normed space.

(a) Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$ be a sequence of independent and symmetric real-valued random variables with $\mathbb{E}|\xi_\ell| < \infty$ for all $\ell \in [M]$. Show that, for $p \in [1, \infty)$,

$$\left(\min_{\ell=1,\dots,M} \mathbb{E}|\xi_\ell| \right) \left(\mathbb{E} \| \sum_{\ell=1}^M \epsilon_\ell x_\ell \|^p \right)^{1/p} \le \left(\mathbb{E} \| \sum_{\ell=1}^M \xi_\ell x_\ell \|^p \right)^{1/p} \, .$$

Conclude that

$$\mathbb{E} \| \sum_{\ell=1}^N \epsilon_\ell \mathbf{x}_\ell \| \le \sqrt{\frac{\pi}{2}} \, \mathbb{E} \| \sum_{\ell=1}^N g_\ell \mathbf{x}_\ell \| \, .$$

(b) Show that

$$\mathbb{E}\|\sum_{\ell=1}^{M} g_\ell \mathbf{x}_\ell\| \le \sqrt{2\log(2M)}\,\mathbb{E}\|\sum_{\ell=1}^{M} \epsilon_\ell \mathbf{x}_\ell\|.$$

Find an example which shows that the log-factor above cannot be removed in general.

8.3. Let $\mathbf{a} \in \mathbb{C}^N$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)$ be a Steinhaus sequence. Show a moment estimate of the form

$$\left(\mathbb{E}|\sum_{\ell=1}^{N} \epsilon_\ell a_\ell|^p\right)^{1/p} \le \alpha\beta^{1/p}\sqrt{p}\|\mathbf{a}\|_2, \quad p \ge 2,$$

in two ways; (a) by using the method of Corollary 8.7; (b) by using Proposition 7.13. Provide small values of α and β.

8.4. Hoeffdings's inequality for complex random variables.
Let $\mathbf{X} = (X_1, \dots, X_N)$ be a vector of complex-valued mean-zero symmetric random variables, that is, X_ℓ has the same distribution as $-X_\ell$. Assume that $|X_\ell| \le 1$, $\ell \in [M]$, almost surely. Let $\mathbf{a} \in \mathbb{C}^M$ be a complex vector. Show that, for $u > 0$,

$$\mathbb{P}\big(|\sum_{j=1}^{M} a_j X_j| \ge u\big) \le 2\exp(-u^2/2)\ .$$

Provide a version of this inequality when the symmetry assumption is removed.

8.5. Let $\mathbf{A} \in \mathbb{C}^{m\times N}$.

(a) Let $\mathbf{g}$ be a standard Gaussian random vector. Show that, for $t > 0$,

$$\mathbb{P}(\|\mathbf{A}\mathbf{g}\|_2 \ge \|\mathbf{A}\|_F + t\|\mathbf{A}\|_{2\to 2}) \le e^{-t^2/2}\ .$$

(b) Let $\boldsymbol{\epsilon}$ be a Rademacher vector. Show that, for $t > 0$,

$$\mathbb{P}(\|\mathbf{A}\boldsymbol{\epsilon}\|_2 \ge c_1\|\mathbf{A}\|_F + c_2 t\|\mathbf{A}\|_{2\to 2}) \le e^{-t^2/2}\ .$$

Provide appropriate values of the constants $c_1, c_2 > 0$.

8.6. Deviation for matrix-valued Gaussian sums.

(a) Let g be a standard Gaussian variable and $\mathbf{B} \in \mathbb{C}^{d\times d}$ a self-adjoint matrix. Show that $\mathbb{E}\exp(g\theta\mathbf{B}) = \exp(\theta^2\mathbf{B}^2/2)$.

(b) Let $\mathbf{g} = (g_1, \dots, g_M)$ be a vector of independent standard Gaussian variables, and $\mathbf{B}_1, \dots, \mathbf{B}_M \in \mathbb{C}^{d\times d}$ be self-adjoint matrices. Introduce $\sigma^2 = \|\sum_{j=1}^{M} \mathbf{B}_j^2\|_{2\to 2}$. Show that

$$\mathbb{E}\exp(\theta\|\sum_{j=1}^{M} g_j\mathbf{B}_j\|_{2\to 2}) \le 2d\exp(\theta^2\sigma^2/2) \quad \text{for } \theta > 0$$

and

$$\mathbb{P}\big(\|\sum_{j=1}^{M} g_j \mathbf{B}_j\|_{2\to 2} \geq t\big) \leq 2d \exp(\frac{-t^2}{2\sigma^2}) \;, \quad t > 0 \;.$$

(c) For a random variable X, show that $\mathbb{E}X \leq \inf_{\theta>0} \theta^{-1} \ln \mathbb{E}[\exp(\theta X)]$.
(d) Show that

$$\mathbb{E}\|\sum_{j=1}^{M} g_j \mathbf{B}_j\|_{2\to 2} \leq \sqrt{2\ln(2d)}\|\sum_{j=1}^{M} \mathbf{B}_j^2\|_{2\to 2}^{1/2} \;, \tag{8.113}$$

and, for a Rademacher sequence $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$,

$$\mathbb{E}\|\sum_{j=1}^{M} \epsilon_j \mathbf{B}_j\|_{2\to 2} \leq \sqrt{2\ln(2d)}\|\sum_{j=1}^{M} \mathbf{B}_j^2\|_{2\to 2}^{1/2} \;. \tag{8.114}$$

(e) Give an example that shows that the factor $\sqrt{\ln(2d)}$ cannot be removed from (8.113) in general.

8.7. Deviation inequalities for sums of rectangular random matrices.

(a) The self-adjoint dilation of a matrix $\mathbf{A} \in \mathbb{C}^{d_1 \times d_2}$ is defined as

$$S(\mathbf{A}) = \begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{pmatrix} \;.$$

Then $S(\mathbf{A}) \in \mathbb{C}^{(d_1+d_2)\times(d_1+d_2)}$ is self-adjoint and $\|S(\mathbf{A})\|_{2\to 2} = \|\mathbf{A}\|_{2\to 2}$.
(b) Let $\mathbf{X}_1, \ldots, \mathbf{X}_M$ be a sequence of $d_1 \times d_2$ random matrices with

$$\|\mathbf{X}_\ell\|_{2\to 2} \leq K \quad \text{ for all } \ell \in [M] \;,$$

and set

$$\sigma^2 := \max\big\{\|\sum_{\ell=1}^{M} \mathbb{E}(\mathbf{X}_\ell \mathbf{X}_\ell^*)\|_{2\to 2}, \|\sum_{\ell=1}^{M} \mathbb{E}(\mathbf{X}_\ell^* \mathbf{X}_\ell)\|_{2\to 2}\big\} \;. \tag{8.115}$$

Show that, for $t > 0$,

$$\mathbb{P}(\|\sum_{\ell=1}^{M} \mathbf{X}_\ell\|_{2\to 2} \geq t) \leq 2(d_1+d_2) \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right) \;. \tag{8.116}$$

8.8. Noncommutative Bernstein inequality, subexponential version. Let $\mathbf{X}_1, \ldots, \mathbf{X}_M \in \mathbb{C}^{d\times d}$ be independent mean-zero self-adjoint random matrices. Assume that

$$\mathbb{E}[\mathbf{X}_\ell^n] \preccurlyeq n! R^{n-2} \sigma_\ell^2 \mathbf{B}_\ell^2/2 \;, \quad \ell \in [M]$$

for some self-adjoint matrices $\mathbf{B}_\ell$ and set

$$\sigma^2 := \left\| \sum_{\ell=1}^{M} \mathbf{B}_\ell^2 \right\|_{2\to 2} .$$

Show that, for $t > 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^{M} \mathbf{X}_\ell \right) \geq t \right) \leq d \exp\left(-\frac{t^2/2}{\sigma^2 + Rt} \right) .$$

8.9. Let T be a countable index set. Show the consistency of the definition (8.39) of the lattice supremum in this case, that is, show that

$$\mathbb{E}(\sup_{t\in T} X_t) = \sup\{\mathbb{E}(\sup_{t\in F} X_t), F \subset T, F \text{ finite}\} .$$

8.10. Let X_t, $t \in T$, be a symmetric random process, i.e., X_t has the same distribution as $-X_t$ for all $t \in T$. Show that, for an arbitrary $t_0 \in T$,

$$\mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\sup_{t\in T} |X_t - X_{t_0}| \leq 2\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{s,t\in T} |X_s - X_t| .$$

8.11. Derive the following generalization of Dudley's inequality: Let X_t, $t \in T$, be a subgaussian process with associated psuedo-metric d, i.e.,

$$\mathbb{P}(|X_s - X_t| \geq u d(s,t)) \leq 2e^{-cu^2}.$$

Then, for some arbitrary $t_0 \in T$ and $p \geq 1$,

$$\left(\mathbb{E}\sup_{t\in T} |X_t - X_{t_0}|^p \right)^{1/p} \leq C\sqrt{p} \int_0^\infty \sqrt{\log(N(T,d,u))}\, du,$$

for some appropriate constant $C > 0$ depending only on c.

8.12. Weak and distributional derivatives.
Recall the notion of weak and distributional derivative in Section C.9.

(a) Show that the function $f(t) = |t|$ has weak derivative

$$f'(t) = \operatorname{sgn}(t) = \begin{cases} -1 & \text{if } t < 0 , \\ 0 & \text{if } t = 0 , \\ 1 & \text{if } t > 0 . \end{cases}$$

(b) Let g be a function of the form (8.69). Show that a weak derivative is given by

$$g'(t) = \chi_{[a,b]}(t) = \begin{cases} 1 & \text{if } t \in [a,b] , \\ 0 & \text{if } t \notin [a,b] . \end{cases}$$

(c) Let f be nondecreasing and differentiable except at possibly a finite number of points. Show that f has a positive distributional derivative.

(d) Assume that f has a positive weak derivative. Show that f is nondecreasing.

9

Sparse Recovery with Random Matrices

It was shown in Chapter 6 that recovery of s-sparse vectors by various recovery algorithms including ℓ_1-minimization is guaranteed if the restricted isometry constants of the measurement matrix satisfy $\delta_{\kappa s} \leq \delta^*$ for an appropriate small integer κ and some $\delta^* \in (0,1)$ both depending only on the algorithm. The derived condition for ℓ_1-minimization is, for instance, $\delta_{2s} < 0.4931$. In Chapter 5 we have seen explicit $m \times m^2$ matrices that satisfy such a condition once $m \geq Cs^2$, see also the discussion at the end of Section 6.1. But it is not clear at this point whether $m \times N$ matrices exist that have small δ_s when m is significantly smaller than Cs^2. The purpose of this chapter is to show the existence of $m \times N$ matrices with $\delta_s \leq \delta$ provided $m \geq C_\delta s \ln(N/s)$ using probabilistic arguments. We use subgaussian random matrices, where all entries are drawn independently according to a subgaussian distribution. This includes Gaussian, Bernoulli, and random variables that are uniformly distributed on $[-1,1]$. For such matrices, the restricted isometry property holds with high probability in the stated parameter regime. We refer to Theorem 9.11 for an exact statement.

For ℓ_1-minimization, we also show that a fixed s-sparse vector $\mathbf{x}$ can be recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization using a random draw of a subgaussian matrix. This nonuniform setting has the advantage of a simple proof that provides good constants (although however, the term $\ln(N/s)$ is replaced by $\ln N$ in our first result). Then we restrict our considerations to Gaussian matrices. Using the Slepian and Gordon lemma as well as concentration of measure, we derive in the nonuniform setting "roughly" (that is, for large dimensions) the sufficient condition

$$m \geq 2s \ln(N/s) \, .$$

We further obtain bounds for the conditioning of Gaussian random matrices, and as a consequence for the restricted isometry property. Again, constants are given explicitly. The Gaussian case also allows to directly show the null space property without passing to the restricted isometry property.

Finally, we make a small detour to the Johnson-Lindenstrauss lemma, which states that a finite set of points in a high-dimensional space can be mapped to a lower-dimensional space via a linear map without significantly perturbing their mutual distances. A subgaussian random matrix can be chosen as this linear mapping. This fact follows immediately from a concentration inequality that is crucial for the proof of the restricted isometry property for subgaussian matrices, see (9.6). In this sense, the Johnson-Lindenstrauss lemma implies the restricted isometry property. We will also show the converse statement that a matrix satisfying the restricted isometry property provides a Johnson-Lindenstrauss mapping when the column signs are randomized.

9.1 Restricted Isometry Property for Subgaussian Matrices

We consider a matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$ having random variables as their entries. Such $\mathbf{A}$ is called *random matrix* or random matrix ensemble.

Definition 9.1. *Let $\mathbf{A}$ be an $m \times N$ random matrix.*

(a) If the entries of $\mathbf{A}$ are independent Rademacher variables (i.e., taking values ± 1 with equal probability) then $\mathbf{A}$ is called a Bernoulli random matrix.

(b) If the entries of $\mathbf{A}$ are independent standard normal distributed random variables then $\mathbf{A}$ is called a Gaussian random matrix.

(c) If all entries of $\mathbf{A}$ are independent mean-zero subgaussian random variables of variance 1 with the same constants β, θ in the definition (7.32) of subgaussian random variables, that is,

$$\mathbb{P}(|A_{j,k}| \geq t) \leq \beta e^{-\kappa t^2} \quad \text{for all } t > 0, \quad j \in [m], k \in [N] \,, \tag{9.1}$$

than $\mathbf{A}$ is called a subgaussian random matrix.

Clearly, Gaussian and Bernoulli random matrices are subgaussian. Also note that the entries of a subgaussian matrix do not necessarily have to be identically distributed. Equivalently to (9.1) we may require that

$$\mathbb{E}[\exp(\theta A_{j,k})] \leq \exp(c\theta^2), \quad \text{for all } \theta \in \mathbb{R}, \quad j \in [m], k \in [N] \,, \tag{9.2}$$

for some constant c that is independent of j, k and N, see Proposition 7.24.

We start with our main result on the restricted isometry property of subgaussian random matrices.

Theorem 9.2. *Let $\mathbf{A}$ be an $m \times N$ subgaussian random matrix. Then there exists a constant $C > 0$ (depending only on the subgaussian parameters β, κ) such that the restricted isometry constant of $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies $\delta_s \leq \delta$ with probability at least $1 - \varepsilon$ provided*

$$m \geq C\delta^{-2}\big(s \ln(eN/s) + \ln(2\varepsilon^{-1})\big) \,. \tag{9.3}$$

Setting $\varepsilon = \exp(-\delta^2 m/(2C))$ yields the condition

$$m \geq 2C\delta^{-2} s \ln(eN/s) \ ,$$

which guarantees that $\delta_s \leq \delta$ with probability at least $1-2\exp\big(-\delta^2 m/(2C)\big)$. This is the statement often found in the literature.

The normalization $\frac{1}{\sqrt{m}}\mathbf{A}$ is natural because $\mathbb{E}\|\frac{1}{\sqrt{m}}\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$ for a fixed vector $\mathbf{x}$ and a subgaussian random matrix $\mathbf{A}$ (where by convention all entries have variance 1). Therefore, the restricted isometry constant δ_s measures the deviation of $\|\frac{1}{\sqrt{m}}\mathbf{A}\mathbf{x}\|_2^2$ from its mean, uniformly over all s-sparse vectors $\mathbf{x}$.

As the entries of Gaussian and Bernoulli random matrices are subgaussian (see Proposition 7.5) with variance 1, we obtain as an immediate consequence that they satisfy the restricted isometry property under Condition (9.3).

Corollary 9.3. *Let $\mathbf{A}$ be an $m \times N$ Gaussian or Bernoulli random matrix. Then there exists a universal constants $C > 0$ such that the restricted isometry constant of $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies $\delta_s \leq \delta$ with probability at least $1-\varepsilon$ provided*

$$m \geq C\delta^{-2}\big(s\ln(eN/s) + \ln(2\varepsilon^{-1})\big) \ . \tag{9.4}$$

For Gaussian matrices we will slightly improve on (9.4) in Section 9.3 by making the constants explicit.

Subgaussian matrices fall into an even larger class of random matrices that we introduce now. Theorem 9.2 will then follow from its generalization to this larger class. We start with some definitions.

Definition 9.4. *Let $\mathbf{Y}$ be a random vector on $\mathbb{R}^N$.*

(a) If $\mathbb{E}|\langle \mathbf{Y}, \mathbf{x}\rangle|^2 = \|\mathbf{x}\|_2^2$ for all $\mathbf{x} \in \mathbb{R}^N$ then $\mathbf{Y}$ is called isotropic.

(b) If, for all $\mathbf{x} \in \mathbb{R}^N$ with $\|\mathbf{x}\|_2 = 1$, the random variable $\langle \mathbf{Y}, \mathbf{x}\rangle$ is subgaussian with subgaussian parameter c being independent of $\mathbf{x}$ (and ideally independent of N), that is,

$$\mathbb{E}[\exp(\lambda\langle \mathbf{Y}, \mathbf{x}\rangle)] \leq \exp(c\lambda^2) \ , \quad \text{for all } \lambda \in \mathbb{R}, \quad \|\mathbf{x}\|_2 = 1 \ , \tag{9.5}$$

then $\mathbf{Y}$ is called a subgaussian random vector.

Note that isotropic subgaussian random vectors do not necessarily have independent entries. We consider random matrices $\mathbf{A} \in \mathbb{R}^{m\times N}$ with independent subgaussian and isotropic rows $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, that is, matrices of the form.

$$\mathbf{A} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix} \ .$$

The following result settles the restricted isometry property for such matrices.

Theorem 9.5. *Let* $\mathbf{A}$ *be an* $m \times N$ *random matrix with independent, isotropic, and subgaussian rows with the same subgaussian parameter* c *in* (9.5)*. If*

$$m \geq C\delta^{-2}\big(s \ln(eN/s) + \ln(2\varepsilon^{-1})\big)$$

then the restricted isometry constant of $\frac{1}{\sqrt{m}}\mathbf{A}$ *satisfies* $\delta_s \leq \delta$ *with probability at least* $1 - \varepsilon$*.*

The proof of this theorem is given in the next section. Theorem 9.2 follows then from a combination with the following lemma.

Lemma 9.6. *Let* $\mathbf{Y} \in \mathbb{R}^N$ *be a random vector with independent, mean-zero and subgaussian entries with variance* 1 *and the same subgaussian parameter* c *in* (9.5)*. Then* $\mathbf{Y}$ *is an isotropic and subgaussian random vector with subgaussian parameters independent of* N*.*

Proof. Let $\mathbf{x} \in \mathbb{R}^N$ with $\|\mathbf{x}\|_2 = 1$. Since the Y_ℓ are independent, zero-mean and of variance 1 we have

$$\mathbb{E}|\langle \mathbf{Y}, \mathbf{x}\rangle|^2 = \sum_{\ell,\ell'=1}^{N} x_\ell x_{\ell'} \mathbb{E} Y_\ell Y_{\ell'} = \sum_{\ell=1}^{N} x_\ell^2 = \|\mathbf{x}\|_2^2 \,.$$

Therefore, $\mathbf{Y}$ is isotropic. Furthermore, according to Theorem 7.27 the random variable $Z = \langle \mathbf{Y}, \mathbf{x}\rangle = \sum_{\ell=1}^{N} x_\ell Y_\ell$ is subgaussian with parameters $\beta = 2$ and $\theta = 1/(4c\|\mathbf{x}\|_2^2) = 1/(4c)$. Hence, $\mathbf{Y}$ is a subgaussian random vector with parameters independent of N. □

Concentration Inequality

The proof of Theorem 9.5 on the restricted isometry property of random matrices relies heavily on the following concentration inequality. The latter in turn is a consequence of Bernstein's inequality for subexponential random variables, which arise when forming the ℓ_2-norm by summing up squares of subgaussian random variables.

Lemma 9.7. *Let* $\mathbf{A}$ *be an* $m \times N$ *random matrix with independent, isotropic, and subgaussian rows with the same subgaussian parameter* c *in* (9.5)*. Then, for all* $\mathbf{x} \in \mathbb{R}^N$ *and every* $t \in (0, 1)$*,*

$$\mathbb{P}\big(\,\big|m^{-1}\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq t\|\mathbf{x}\|_2^2\big) \leq 2\exp(-\tilde{c}t^2 m)\,, \tag{9.6}$$

where $\tilde{c}$ *depends only on* c*.*

Proof. Let $\mathbf{x} \in \mathbb{R}^N$. Without loss of generality we may assume that $\|\mathbf{x}\|_2 = 1$. Denote the rows of $\mathbf{A}$ by $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \in \mathbb{R}^N$ and consider the random variables

$$Z_\ell = |\langle \mathbf{Y}_\ell, \mathbf{x}\rangle|^2 - \|\mathbf{x}\|_2^2\,, \ell \in [m]\,.$$

Since $\mathbf{Y}_\ell$ is isotropic we have $\mathbb{E}Z_\ell = 0$. Further, Z_ℓ is subexponential because $\langle \mathbf{Y}_\ell, \mathbf{x}\rangle$ is subgaussian, that is, $\mathbb{P}(|Z_\ell| \geq t) \leq \beta \exp(-\kappa t)$ for some parameters β, κ depending only on c. Observe now that

$$m^{-1}\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 = \frac{1}{m}\sum_{\ell=1}^m \left(|\langle \mathbf{Y}_\ell, \mathbf{x}\rangle|^2 - \|\mathbf{x}\|_2^2\right) = \frac{1}{m}\sum_{\ell=1}^m Z_\ell \, .$$

By assumption the Z_ℓ are independent. Therefore, it follows from Bernstein's inequality for subexponential random variables, Corollary 7.32, that

$$\begin{aligned}\mathbb{P}\left(|m^{-1}\sum_{\ell=1}^m Z_\ell| \geq t\right) = \mathbb{P}\left(|\sum_{\ell=1}^m Z_\ell| \geq tm\right) &\leq 2\exp\left(-\frac{\kappa^2 m^2 t^2/2}{2\beta m + \kappa t}\right) \\ &\leq 2\exp\left(-\frac{\kappa^2}{4\beta + 2\kappa} m t^2\right) \, ,\end{aligned}$$

where we used that $t \in (0,1)$ in the last step. Hence, the claim follows with $\tilde{c} = \frac{\kappa^2}{4\beta+2\kappa}$. □

We note that the normalized random matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$, with $\mathbf{A}$ satisfying the assumptions of the previous lemma, satisfies

$$\mathbb{P}\Big(\Big|\|\widetilde{\mathbf{A}}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\Big| \geq t\|\mathbf{x}\|_2^2\Big) \leq 2\exp(-\tilde{c}t^2 m) \, . \tag{9.7}$$

This will be the starting point of the proof of the restricted isometry property.

Proof of the RIP

Next we show that a random matrix satisfying the concentration inequality (9.7) also satisfies the sth order restricted isometry property, provided that its number of rows scales at least like s times a log factor. We first show that a single column submatrix of a random matrix is well-conditioned under an appropriate condition on its size.

Theorem 9.8. *Let $S \subset [N]$ with $\operatorname{card}(S) = s$. Suppose that an $m \times N$ random matrix $\mathbf{A}$ is drawn according to a probability distribution for which the concentration inequality (9.7) holds, that is, for $t \in (0,1)$,*

$$\mathbb{P}\left(\left|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| > t\|\mathbf{x}\|_2^2\right) \;\leq\; 2\,\exp\left(-\tilde{c}t^2\, m\right) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^N \, . \tag{9.8}$$

If, for $\varepsilon, \delta \in (0,1)$,

$$m \geq C\delta^{-2}(7s + 2\ln(2\varepsilon^{-1})) \, , \tag{9.9}$$

where $C = 2/(3\tilde{c})$, then with probability at least $1-\varepsilon$

$$\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \leq \delta \, .$$

Proof. According to Proposition C.3, for $\rho \in (0, 1/2)$, there exists a subset $\mathcal{U}$ of the unit sphere $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^N, \operatorname{supp} \mathbf{x} \subset S, \|\mathbf{x}\|_2 = 1\}$ which satisfies

$$\operatorname{card}(\mathcal{U}) \le \left(1 + \frac{2}{\rho}\right)^s \qquad \text{and} \qquad \min_{\mathbf{u} \in \mathcal{U}} \|\mathbf{z} - \mathbf{u}\|_2 \le \rho \quad \text{for all } \mathbf{z} \in \mathcal{S}\,.$$

The concentration inequality (9.8) gives, for $t \in (0, 1)$ depending on δ and ρ to be determined later,

$$\begin{aligned}
&\mathbb{P}\left(\left|\|\mathbf{A}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| > t\,\|\mathbf{u}\|_2^2 \quad \text{for some } \mathbf{u} \in \mathcal{U}\right) \\
&\le \sum_{\mathbf{u} \in \mathcal{U}} \mathbb{P}\left(\left|\|\mathbf{A}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| > t\,\|\mathbf{u}\|_2^2\right) \le 2\operatorname{card}(\mathcal{U})\,\exp\left(-\tilde{c}t^2\,m\right) \\
&\le 2\left(1 + \frac{2}{\rho}\right)^s \exp\left(-\tilde{c}t^2\,m\right).
\end{aligned}$$

Let us assume now that the realization of the random matrix $\mathbf{A}$ yields

$$\left|\|\mathbf{A}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| \le t \qquad \text{for all } \mathbf{u} \in \mathcal{U}. \tag{9.10}$$

By the above, this occurs with probability exceeding

$$1 \;-\; 2\left(1 + \frac{2}{\rho}\right)^s \exp\left(-\tilde{c}t^2\,m\right). \tag{9.11}$$

We are going to prove that (9.10) implies $\left|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \le \delta$ for all $\mathbf{x} \in \mathcal{S}$, i.e., $\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \le \delta$ once ρ, t are chosen appropriately. Let $\mathbf{B} = \mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}$. Then (9.10) means that $|\langle \mathbf{B}\mathbf{u}, \mathbf{u}\rangle| \le t$ for all $\mathbf{u} \in \mathcal{U}$. Now consider a vector $\mathbf{x} \in \mathcal{S}$, for which we choose another vector $\mathbf{u} \in \mathcal{U}$ satisfying $\|\mathbf{x} - \mathbf{u}\|_2 \le \rho < 1/2$. We obtain

$$\begin{aligned}
|\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle| &= |\langle \mathbf{B}\mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{B}(\mathbf{x}+\mathbf{u}), \mathbf{x}-\mathbf{u}\rangle| \le |\langle \mathbf{B}\mathbf{u}, \mathbf{u}\rangle| + |\langle \mathbf{B}(\mathbf{x}+\mathbf{u}), \mathbf{x}-\mathbf{u}\rangle| \\
&\le t + \|\mathbf{B}\|_{2\to 2}\,\|\mathbf{x}+\mathbf{u}\|_2\|\mathbf{x}-\mathbf{u}\|_2 \le t + 2\,\|\mathbf{B}\|_{2\to 2}\,\rho.
\end{aligned}$$

Taking the supremum over all $\mathbf{x} \in \mathcal{S}$, we deduce that

$$\|\mathbf{B}\|_{2\to 2} \le t + 2\,\|\mathbf{B}\|_{2\to 2}\,\rho, \qquad \text{i.e.,} \qquad \|\mathbf{B}\|_{2\to 2} \le \frac{t}{1-2\rho}.$$

We therefore choose $t := (1-2\rho)\delta$, so that $\|\mathbf{B}\|_{2\to 2} \le \delta$. By (9.11) we conclude that

$$\mathbb{P}\left(\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} > \delta\right) \le 2\left(1 + \frac{2}{\rho}\right)^s \exp\left(-\tilde{c}(1-2\rho)^2\delta^2 m\right). \tag{9.12}$$

Yet another reformulation states that $\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \le \delta$ with probability at least $1 - \varepsilon$ provided

$$m \ge \frac{1}{\tilde{c}(1-2\rho)^2}\delta^{-2}\left(\ln(1 + 2/\rho)s + \ln(2\varepsilon^{-1})\right). \tag{9.13}$$

We now choose $\rho = 2/(e^{3/2}-1)$, so that $\ln(1+2/\rho)/(1-2\rho)^2 \le 14/3$ and $1/(1-2\rho)^2 \le 4/3$. Thus, (9.13) is fulfilled when

$$m \ge \frac{2}{3\tilde{c}}\delta^{-2}\left(7s + 2\ln(2\varepsilon^{-1})\right). \tag{9.14}$$

This concludes the proof. □

Remark 9.9. (a) The attentive reader may have noticed that the above proof applies without changes if one passes from coordinate subspaces indexed by S to restrictions of $\mathbf{A}$ to arbitrary s-dimensional subspaces of $\mathbb{R}^N$.

(b) On a similar note, the statement (and proof) does not depend on the columns of $\mathbf{A}$ outside S. Therefore, one could as well state the previous theorem for a $m \times s$ subgaussian random matrix $\mathbf{B}$. Indeed, for such a matrix,

$$\|\frac{1}{m}\mathbf{B}^*\mathbf{B} - \mathbf{Id}\|_{2\to 2} \le \delta$$

with probability at least $1-\varepsilon$ provided that (9.9) holds, or equivalently,

$$\mathbb{P}(\|m^{-1}\mathbf{B}^*\mathbf{B} - \mathbf{Id}\|_{2\to 2} \ge \delta) \le 2\exp\left(-\frac{3\tilde{c}}{4}\delta^2 m - \frac{7}{2}s\right). \tag{9.15}$$

We now turn to the main result of this section.

Theorem 9.10. *Suppose that an $m \times N$ random matrix $\mathbf{A}$ is drawn according to a probability distribution for which the concentration inequality*

$$\mathbb{P}\left(\left|\|\mathbf{Ax}\|_2^2 - \|\mathbf{x}\|_2^2\right| > t\|\mathbf{x}\|_2^2\right) \;\le\; 2\,\exp\left(-\tilde{c}t^2\,m\right)$$

holds for all $t \in (0,1)$ and $\mathbf{x} \in \mathbb{R}^N$. If, for $\delta, \varepsilon \in (0,1)$,

$$m \ge C\delta^{-2}\left[s\,(9 + 2\ln(N/s)) + 2\ln(2\varepsilon^{-1})\right]$$

where $C = 2/(3\tilde{c})$, then with probability at least $1-\varepsilon$ the restricted isometry constant δ_s of $\mathbf{A}$ satisfies $\delta_s \le \delta$.

Proof. The event that a single submatrix $\mathbf{A}_S$ with $\mathrm{card}(S) = s$ is well-conditioned is investigated in Theorem 9.8. We use the same notation as in its proof. Recall that $\delta_s = \sup_{S\subset[N],\mathrm{card}(S)=s} \|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2}$, see (6.2). Taking the union bound over all $\binom{N}{s}$ subsets $S \subset [N]$ of cardinality s and using (9.12) yields

$$\begin{aligned}
\mathbb{P}(\delta_s > \delta) &\le \sum_{S\subset[N],\mathrm{card}(S)=s} \mathbb{P}\left(\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \ge \delta\right) \\
&\le 2\binom{N}{s}\left(1+\frac{2}{\rho}\right)^s \exp\left(-\tilde{c}\delta^2(1-2\rho)^2 m\right) \\
&\le 2\left(\frac{eN}{s}\right)^s \left(1+\frac{2}{\rho}\right)^s \exp\left(-\tilde{c}\delta^2(1-2\rho)^2 m\right),
\end{aligned}$$

where we have additionally applied Lemma C.5 in the last step. Making the choice $\rho = 2/(e^{3/2} - 1)$ as before yields that $\delta_s \leq \delta$ with probability at least $1 - \varepsilon$ provided

$$m \geq \frac{1}{\widetilde{c}\delta^2} \left(\frac{4}{3} s \ln(eN/s) + \frac{14}{3} s + \frac{4}{3} \ln(2\varepsilon^{-1}) \right)$$

which is a reformulation of the desired condition. □

By possibly adjusting constants, the above theorem in combination with Lemma 9.7 and 9.6 clearly implies Theorem 9.2, Corollary 9.3 and Theorem 9.5.

We now gather the results of this chapter to conclude with the major theorem about sparse reconstruction by ℓ_1-minimization from random measurements.

Theorem 9.11. *Let* $\mathbf{A}$ *be an* $m \times N$ *subgaussian random matrix. Let* $s < N, \varepsilon \in (0,1)$ *such that*

$$m \geq C_1 s \ln(eN/s) + C_2 \ln(2\varepsilon^{-1})$$

for some constants $C_1, C_2 > 0$ *only depending on the subgaussian parameters* β, θ*. Then with probability at least* $1 - \varepsilon$ *every* s*-sparse vector* $\mathbf{x}$ *is recovered from* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *via* ℓ_1*-minimization.*

Proof. The statement follows from a combination of Theorem 9.2 and Theorem 6.8 (or alternatively, Theorem 6.11) by additionally noting that exact sparse recovery is independent of the normalization of the matrix. □

Remark 9.12. Setting $\varepsilon = 2\exp(-m/(2C_2))$ shows recovery of all s-sparse vectors via ℓ_1-minimization with probability at least $1 - 2\exp(-m/(2C_2))$ using a subgaussian random provided that

$$m \geq 2C_1 s \ln(eN/s) \, .$$

We will see in Chapter 10 that this condition on the required number of measurement cannot be improved.

Once the restricted isometry property is established, recovery via ℓ_1-minimization is also stable under sparsity defect and robust under noise on the measurements, see Theorem 6.11. We obtain the same type of uniform recovery results also for the other algorithms which are guaranteed to succeed under conditions on the restricted isometry property, see Chapter 6. This includes iterative hard thresholding, iterative hard thresholding pursuit, orthogonal matching pursuit, and compressive sampling matching pursuit. Moreover, such recovery guarantees hold as well for general random matrices satisfying the concentration inequality (9.8), such as random matrices with independent isotropic subgaussian rows.

Universality

Often sparsity is not with respect to the canonical basis, but rather with respect to some other orthonormal basis. This means that the vector of interest can be written as $\mathbf{z} = \mathbf{U}\mathbf{x}$ with an $N \times N$ orthogonal matrix $\mathbf{U}$ and an s-sparse vector $\mathbf{x} \in \mathbb{R}^N$. Taking measurements of $\mathbf{z}$ with a random matrix $\mathbf{A}$ can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{U}\mathbf{x}.$$

In order to recover $\mathbf{z}$, it clearly suffices to first recover the sparse vector $\mathbf{x}$ and then forming $\mathbf{z} = \mathbf{U}\mathbf{x}$. Therefore, this more general problem reduces to the standard compressive sensing problem with measurement matrix $\mathbf{A}' = \mathbf{A}\mathbf{U}$. We therefore consider this model with a random $m \times N$ matrix $\mathbf{A}$ and a fixed (deterministic) orthogonal matrix $\mathbf{U} \in \mathbb{R}^{N \times N}$ as a new measurement matrix of interest in this context. It turns out that the analysis in the preceding sections can easily be applied to this more general situation.

Theorem 9.13. *Let $\mathbf{U} \in \mathbb{R}^{N \times N}$ be a (fixed) orthogonal matrix. Suppose that an $m \times N$ random matrix $\mathbf{A}$ is drawn according to a probability distribution for which the concentration inequality*

$$\mathbb{P}\left(\left|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| > t\|\mathbf{x}\|_2^2\right) \;\le\; 2\,\exp\left(-\tilde{c}t^2\,m\right) \tag{9.16}$$

holds for all $t \in (0,1)$ and $\mathbf{x} \in \mathbb{R}^N$. Let $\delta, \varepsilon \in (0,1)$. Then the restricted isometry constant δ_s of $\mathbf{A}\mathbf{U}$ satisfies $\delta_s \le \delta$ with probability at least $1-\varepsilon$ provided

$$m \ge C\delta^{-2}\left[s\,(9 + 2\ln(N/s)) + 2\ln(2\varepsilon^{-1})\right]$$

with $C = 2/(3\tilde{c})$.

Proof. The crucial point of the proof is that the concentration inequality (9.16) holds also with $\mathbf{A}$ replaced by $\mathbf{A}\mathbf{U}$. Indeed, let $\mathbf{x} \in \mathbb{R}^N$ and set $\mathbf{x}' = \mathbf{U}\mathbf{x}$. Orthogonality of $\mathbf{U}$ yields

$$\begin{aligned}&\mathbb{P}\left(\left|\|\mathbf{A}\mathbf{U}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| > t\|\mathbf{x}\|_2^2\right) = \mathbb{P}\left(\left|\|\mathbf{A}\mathbf{x}'\|_2^2 - \|\mathbf{U}^{-1}\mathbf{x}'\|_2^2\right| > t\|\mathbf{U}^{-1}\mathbf{x}'\|_2^2\right)\\ &= \mathbb{P}\left(\left|\|\mathbf{A}\mathbf{x}'\|_2^2 - \|\mathbf{x}'\|_2^2\right| > t\|\mathbf{x}'\|_2^2\right) \le 2\,\exp\left(-\tilde{c}t^2\,m\right).\end{aligned}$$

Therefore, the statement follows from Theorem 9.10. □

In particular, the above theorem implies that sparse recovery with subgaussian matrices is universal with respect to the orthogonal basis in which the signal is sparse. Indeed, the matrix $\mathbf{U}$ is arbitrary in the above theorem. It means even that at the encoding step when measurements $\mathbf{y} = \mathbf{A}\mathbf{U}\mathbf{x}$ are taken, the orthogonal matrix $\mathbf{U}$ does not need to be known. Only at the decoding stage when the ℓ_1-minimization principle is applied it has to be used.

We emphasize, however, that universality does not mean that a single (fixed) measurement matrix $\mathbf{A}$ is able to deal with sparsity in any basis. It is actually straightforward to see that this is impossible because once $\mathbf{A}$ is given, one may construct a basis $\mathbf{U}$ for which sparse recovery is not possible. The theorem only states that for a fixed orthogonal $\mathbf{U}$, a random choice of $\mathbf{A}$ will work well with high probability.

9.2 Nonuniform Recovery

In this section we consider the probability that a fixed sparse vector $\mathbf{x}$ is recovered via ℓ_1-minimization from $\mathbf{y} = \mathbf{A}\mathbf{x}$ using a random draw of a subgaussian matrix $\mathbf{A}$. We first discuss differences between uniform and nonuniform recovery. Then we give a first simple estimate for subgaussian matrices with good constants and then an improved version for the special case of Gaussian matrices.

Uniform versus Nonuniform Recovery

One may pursue different strategies in order to come up with rigorous recovery results. We distinguish between uniform and nonuniform recovery guarantees. A uniform recovery guarantee means that once the random matrix is chosen, then with high probability all sparse signals can be recovered using the same matrix. The bounds for the restricted isometry property that we have just derived, indeed imply *uniform recovery* for subgaussian random matrices. A nonuniform recovery result only states that a fixed sparse signal can be recovered with high probability using a random draw of the matrix. In particular, such weaker nonuniform results allow in principle that the small exceptional set of matrices for which recovery is not necessarily guaranteed may depend on the signal, in contrast to a uniform statement. Clearly, uniform recovery implies nonuniform recovery, but the converse is not true. In mathematical terms, a uniform recovery guarantee provides a lower probability estimate of the form

$$\mathbb{P}(\forall s\text{-sparse } \mathbf{x} \text{ recovery is sucessful using } \mathbf{A}) \geq 1 - \varepsilon \ ,$$

while non-uniform recovery provides a statement of the form

$$\forall s\text{-sparse } \mathbf{x} : \mathbb{P}(\text{recovery of } \mathbf{x} \text{ using } \mathbf{A} \text{ succeeds}) \geq 1 - \varepsilon \ ,$$

where in both cases the probability is over the random draw of $\mathbf{A}$. Due to the appearance of the quantifier $\forall \mathbf{x}$ at different places, the two types of statements are clearly different.

For subgaussian random matrices, nonuniform analysis is able to provide explicit and good constants – although the asymptotic analysis is essentially the same in both type of recovery guarantees. The advantage of the nonuniform approach will become more apparent later in Chapter 12, where we will see that such type of results will be easier to prove for structured random matrices and will provide better estimates *both* in terms of the constants and the asymptotic behavior.

Subgaussian Random Matrices

Our first nonuniform recovery results for ℓ_1-minimization applies to subgaussian random matrices.

Theorem 9.14. *Let $\mathbf{x} \in \mathbb{C}^N$ be an s-sparse vector. Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a randomly drawn subgaussian matrix with parameter c in* (9.5)*. If, for some $\varepsilon \in (0,1)$,*

$$m \geq s\left[\sqrt{4c\ln(4N/\varepsilon)} + \sqrt{C(7+2\ln(2/(\varepsilon s))}\right]^2 , \tag{9.17}$$

then with probability at least $1-\varepsilon$ the vector $\mathbf{x}$ is the unique solution to the ℓ_1-minimization problem $\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$.

The constant $C = 2/(3\tilde{c})$ depends only on the subgaussian parameter through $\tilde{c}$ in (9.6)*.*

Remark 9.15. The term $\sqrt{C(7+2\ln(2/(\varepsilon s))}$ in (9.17) becomes negligible for large N and mildly large s, so that roughly speaking sparse recovery is successful provided $m \geq 4cs\ln(4N/\varepsilon)$. In the Gaussian and Bernoulli case where $c = 1/2$, we roughly obtain the sufficient condition

$$m \geq 2s\ln(4N/\varepsilon) . \tag{9.18}$$

Below we will replace the $\ln N$-factor by $\ln(N/s)$ in the Gaussian case.

Proof. Set $S := \operatorname{supp}\mathbf{x}$ and note that $\operatorname{card}(S) = s$. By Corollary 4.27 it is sufficient to show that

$$|\langle (\mathbf{A}_S)^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S)\rangle| = |\langle \mathbf{a}_\ell, (\mathbf{A}_S^\dagger)^* \operatorname{sgn}(\mathbf{x}_S)\rangle| < 1 \qquad \text{for all } \ \ell \in \overline{S} .$$

Therefore, the probability of failure of recovery is bounded by

$$\begin{aligned} P &:= \mathbb{P}\left(\exists \ell \notin S : |\langle \mathbf{a}_\ell, (\mathbf{A}_S^\dagger)^* \operatorname{sgn}(\mathbf{x}_S)\rangle| \geq 1\right) \\ &\leq \mathbb{P}\left(\exists \ell \notin S : |\langle (\mathbf{A}_S)^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S)\rangle| \geq 1 \,\middle|\, \|(\mathbf{A}_S^\dagger)^* \operatorname{sgn}(\mathbf{x}_S)\|_2 < \alpha\right) \qquad (9.19) \\ &+ \mathbb{P}(\|(\mathbf{A}_S^\dagger)^* \operatorname{sgn}(\mathbf{x}_S)\|_2 \geq \alpha) . \qquad (9.20) \end{aligned}$$

The first term above is estimated using Theorem 7.27. Hereby, we additionally use the independence of all the entries of $\mathbf{A}$ so that, in particular, $\mathbf{a}_\ell$ and $\mathbf{A}_S$ are independent for $\ell \notin S$. Conditioning on the event that $\|(\mathbf{A}_S^\dagger)^* \operatorname{sgn}(\mathbf{x}_S)\|_2 < \alpha$ we obtain

$$\begin{aligned} &\mathbb{P}\left(|\langle (\mathbf{A}_S)^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S)\rangle| \geq 1\right) = \mathbb{P}\left(\left|\sum_{j=1}^m (\mathbf{a}_\ell)_j [(\mathbf{A}_S^\dagger)^* \operatorname{sgn}(\mathbf{x}_S)]_j\right| \geq 1\right) \\ &\leq 2\exp\left(-\frac{1}{4c\alpha^2}\right) . \end{aligned}$$

By the union bound the term in (9.19) can be estimated by $2N\exp(-1/(4c\alpha^2))$, which in turn is no larger than $\varepsilon/2$ provided

$$\alpha \leq \sqrt{1/(4c\ln(4N/\varepsilon))} . \tag{9.21}$$

For the term in (9.20), we observe that

$$\|(\mathbf{A}_S^\dagger)^* \mathrm{sgn}(\mathbf{x}_S)\|_2^2 \le \sigma_{\min}^{-2}(\mathbf{A}_S)\|\mathrm{sgn}(\mathbf{x}_S)\|_2^2 = \sigma_{\min}^{-2}(\mathbf{A}_S)\, s\ ,$$

where $\sigma_{\min}$ denotes the smallest singular value, see also (A.22). Therefore,

$$\mathbb{P}(\|(\mathbf{A}_S^\dagger)^* \mathrm{sgn}(\mathbf{x}_S)\|_2 \ge \alpha) \le \mathbb{P}\left(\sigma_{\min}(\mathbf{A}_S/\sqrt{m}) \le \frac{1}{\sqrt{m}}\frac{\sqrt{s}}{\alpha}\right)\ .$$

By Theorem 9.8 the matrix $\mathbf{B} = \mathbf{A}_S/\sqrt{m}$ satisfies, for $\delta \in (0,1)$,

$$\mathbb{P}(\sigma_{\min}(\mathbf{B}) < 1-\delta) < \mathbb{P}(\sigma_{\min}(\mathbf{B}) < \sqrt{1-\delta}) < \mathbb{P}(\|\mathbf{B}^*\mathbf{B} - \mathbf{Id}\|_{2\to 2} \ge \delta) \le \varepsilon/2$$

provided $m \ge C\delta^{-2}(7s + 2\ln(2\varepsilon^{-1}))$ with $C = 2/(3\tilde{c})$. We next choose $\delta = 1 - \frac{\sqrt{s}}{\alpha\sqrt{m}}$ and α such that equality holds in (9.21). Combining the above arguments, we can bound the failure probability by ε provided

$$m \ge C\left(1 - \frac{\sqrt{4cs\ln(4N/\varepsilon)}}{\sqrt{m}}\right)^{-2}(7s + 2\ln(2\varepsilon^{-1}))\ . \tag{9.22}$$

Solving for m yields the condition

$$m \ge s\left[\sqrt{4c\ln(4N/\varepsilon)} + \sqrt{C(7 + 2\ln(2\varepsilon^{-1})/s)}\right]^2.$$

This condition also implies $\delta \in (0,1)$. □

Gaussian Random Matrices

Next we improve on the log-factor in (9.17), and make recovery also stable under noise. For technical reasons we restrict to recovery of real vectors, but note that extensions to the complex case are possible.

Theorem 9.16. *Let $\mathbf{x} \in \mathbb{R}^N$ be an s-sparse vector, and $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a random drawn from the Gaussian matrix ensemble. Let $\varepsilon \in (0,1)$. If*

$$\frac{m^2}{m+1} \ge 2s\left(\sqrt{\ln(2.34\,N/s)} + \sqrt{\frac{\ln(\varepsilon^{-1})}{s}}\right)^2 \tag{9.23}$$

then with probability at least $1-\varepsilon$, $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Ax} = \mathbf{Az}$.

Remark 9.17. The proof actually allows to deduce even a slightly more precise (but more complicated) condition, see Remark 9.23. Roughly speaking, for mildly large N, s, condition (9.23) requires

$$m \ge 2s\ln\left(\frac{2.34\,N}{s}\right)\ . \tag{9.24}$$

The previous result can be extended to robust recovery.

Theorem 9.18. *Let $\mathbf{x} \in \mathbb{R}^N$ be an s-sparse vector, and $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a random drawn of a Gaussian matrix. Assume that noisy measurements are taken, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \le \eta$. If, for $\varepsilon \in (0,1)$, $\tau > 0$,*

$$\frac{m^2}{m+1} \ge 2s \left(\sqrt{\ln\left(2.34\, N/s\right)} + \sqrt{\frac{\ln(\varepsilon^{-1})}{s}} + \tau \right)^2 ,$$

then with probability at least $1-\varepsilon$, every minimizer $\mathbf{x}^\sharp$ of

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \le \eta$$

satisfies

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \le \frac{2\eta}{\tau} .$$

We develop the proof of these theorems in several steps. Our basic ingredients are the recovery conditions of Theorem 4.34 and 4.36 based on the tangent cone $T(\mathbf{x})$ of the ℓ_1-norm defined in (4.40).

We start our analysis with a general concentration of measure result for Gaussian random matrices. We recall from Proposition 8.1(b) that for a standard Gaussian random vector $\mathbf{g} \in \mathbb{R}^m$

$$E_m := \mathbb{E}\|\mathbf{g}\|_2 = \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)}$$

with $m/\sqrt{m+1} \le E_m \le \sqrt{m}$. For a set $T \subset \mathbb{R}^N$ we introduce its *Gaussian width* by

$$\ell(T) := \mathbb{E} \sup_{\mathbf{x} \in T} \langle \mathbf{x}, \mathbf{g} \rangle , \tag{9.25}$$

where $\mathbf{g} \in \mathbb{R}^N$ is a standard Gaussian random vector. The following result is known as Gordon's escape through the mesh theorem.

Theorem 9.19. *Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ be a Gaussian random matrix, and T be a subset of the unit sphere $S^{N-1} = \{\mathbf{x} \in \mathbb{R}^N, \|\mathbf{x}\|_2 = 1\}$. Then, for $t > 0$,*

$$\mathbb{P}\left(\inf_{\mathbf{x} \in T} \|\mathbf{A}\mathbf{x}\|_2 \le E_m - \ell(T) - t \right) \le e^{-t^2/2} .$$

Proof. Our first aim is to estimate the expectation $\mathbb{E}\inf_{x \in T} \|\mathbf{A}\mathbf{x}\|_2$ via Gordon's lemma (Lemma 8.28). For $\mathbf{x} \in T$ and $\mathbf{y} \in S^{m-1}$ we define the Gaussian process

$$X_{\mathbf{x},\mathbf{y}} := \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = \operatorname{tr}\left(\mathbf{A}\mathbf{x}\mathbf{y}^*\right) .$$

Then $\inf_{\mathbf{x} \in T} \|\mathbf{A}\mathbf{x}\|_2 = \inf_{\mathbf{x} \in T} \max_{\mathbf{y} \in S^{m-1}} X_{\mathbf{x},\mathbf{y}}$. The key idea is to compare $X_{\mathbf{x},\mathbf{y}}$ to another Gaussian process $Y_{\mathbf{x},\mathbf{y}}$. To this end, we let $\mathbf{g} \in \mathbb{R}^N$, $\mathbf{h} \in \mathbb{R}^m$ be independent standard Gaussian vectors and introduce

$$Y_{\mathbf{x},\mathbf{y}} := \sum_{j=1}^{N} g_j x_j + \sum_{k=1}^{m} h_j y_j = \langle \mathbf{x}, \mathbf{g} \rangle + \langle \mathbf{y}, \mathbf{h} \rangle \ .$$

Note that, for $\mathbf{x}, \mathbf{x}' \in T$, $\mathbf{y}, \mathbf{y}' \in S^{N-1}$,

$$\mathbb{E}|X_{\mathbf{x},\mathbf{y}} - X_{\mathbf{x}',\mathbf{y}'}|^2 = \mathbb{E}\mathrm{tr}\,(\mathbf{A}(\mathbf{x}\mathbf{y}^* - \mathbf{x}'(\mathbf{y}')^*)^2 = \|\mathbf{x}\mathbf{y}^* - \mathbf{x}'(\mathbf{y}')^*\|_F^2 \ .$$

and, by independence,

$$\mathbb{E}|Y_{\mathbf{x},\mathbf{y}} - Y_{\mathbf{x}',\mathbf{y}'}|^2 = \mathbb{E}\langle \mathbf{g}, \mathbf{x} - \mathbf{x}' \rangle^2 + \mathbb{E}\langle \mathbf{h}, \mathbf{y} - \mathbf{y}' \rangle^2 = \|\mathbf{x} - \mathbf{x}'\|_2^2 + \|\mathbf{y} - \mathbf{y}'\|_2^2 \ .$$

For $\mathbf{x} \in S^{N-1}$, $\mathbf{y}, \mathbf{y}' \in S^{m-1}$, we have

$$\|\mathbf{x}\mathbf{y}^* - \mathbf{x}(\mathbf{y}')^*\|_F^2 = \sum_{k,\ell} x_\ell^2 (y_k - y_k')^2 = \|\mathbf{x}\|_2^2 \|\mathbf{y} - \mathbf{y}'\|_2^2 = \|\mathbf{y} - \mathbf{y}'\|_2^2 \ ,$$

so that

$$\mathbb{E}|X_{\mathbf{x},\mathbf{y}} - X_{\mathbf{x},\mathbf{y}'}|^2 = \mathbb{E}|Y_{\mathbf{x},\mathbf{y}} - Y_{\mathbf{x},\mathbf{y}'}|^2 \ . \tag{9.26}$$

Furthermore, for arbitrary $\mathbf{x}, \mathbf{x}' \in S^{N-1}$, $\mathbf{y}, \mathbf{y}' \in S^{m-1}$, we have by cyclicity of the trace

$$\begin{aligned}
&\|\mathbf{x}\mathbf{y}^* - \mathbf{x}'(\mathbf{y}')^*\|_F^2 = \|(\mathbf{x} - \mathbf{x}')\mathbf{y}^* + \mathbf{x}'(\mathbf{y} - \mathbf{y}')^*\|_F^2 \\
&= \|(\mathbf{x} - \mathbf{x}')\mathbf{y}^*\|_F^2 + \|\mathbf{x}'(\mathbf{y} - \mathbf{y}')^*\|_F^2 + 2\langle (\mathbf{x} - \mathbf{x}')\mathbf{y}^*, \mathbf{x}'(\mathbf{y} - \mathbf{y}')^* \rangle_F \\
&= \|(\mathbf{x} - \mathbf{x}')\mathbf{y}^*\|_F^2 + \|\mathbf{x}'(\mathbf{y} - \mathbf{y}')^*\|_F^2 + 2\mathrm{tr}\,((\mathbf{x} - \mathbf{x}')\mathbf{y}^*(\mathbf{y} - \mathbf{y}')(\mathbf{x}')^*) \\
&= \|(\mathbf{x} - \mathbf{x}')\mathbf{y}^*\|_F^2 + \|\mathbf{x}'(\mathbf{y} - \mathbf{y}')^*\|_F^2 + 2\langle \mathbf{x} - \mathbf{x}', \mathbf{x}' \rangle \langle \mathbf{y} - \mathbf{y}', \mathbf{y} \rangle \\
&= \|\mathbf{x} - \mathbf{x}'\|_2^2 + \|\mathbf{y} - \mathbf{y}'\|_2^2 + 2\left(\langle \mathbf{x}, \mathbf{x}' \rangle - \|\mathbf{x}'\|_2^2\right)\left(\|\mathbf{y}\|_2^2 - \langle \mathbf{y}, \mathbf{y}' \rangle\right) \\
&\leq \|\mathbf{x} - \mathbf{x}'\|_2^2 + \|\mathbf{y} - \mathbf{y}'\|_2^2 \ .
\end{aligned}$$

The inequality in the last step follows from $\langle \mathbf{x}, \mathbf{x}' \rangle \leq 1 = \|\mathbf{x}'\|_2^2$ and $\langle \mathbf{y}, \mathbf{y}' \rangle \leq 1 = \|\mathbf{y}\|_2^2$ using the Cauchy-Schwarz inequality. Therefore, we have shown that

$$\mathbb{E}|X_{\mathbf{x},\mathbf{y}} - X_{\mathbf{x}',\mathbf{y}'}|^2 \leq \mathbb{E}|Y_{\mathbf{x},\mathbf{y}} - Y_{\mathbf{x}',\mathbf{y}'}|^2 \ . \tag{9.27}$$

It follows from Gordon's Lemma 8.28 and Remark 8.29 that

$$\begin{aligned}
\mathbb{E} \inf_{\mathbf{x} \in T} \|\mathbf{A}\mathbf{x}\|_2 &= \mathbb{E} \inf_{\mathbf{x} \in T} \max_{\mathbf{y} \in S^{m-1}} X_{\mathbf{x},\mathbf{y}} \geq \mathbb{E} \inf_{\mathbf{x} \in T} \max_{\mathbf{y} \in S^{m-1}} Y_{\mathbf{x},\mathbf{y}} \\
&= \mathbb{E} \inf_{\mathbf{x} \in T} \max_{\mathbf{y} \in S^{m-1}} \{\langle \mathbf{g}, \mathbf{x} \rangle + \langle \mathbf{h}, \mathbf{y} \rangle\} = \mathbb{E} \inf_{\mathbf{x} \in T} \{\langle \mathbf{g}, \mathbf{x} \rangle + \|\mathbf{h}\|_2\} \\
&= \mathbb{E}\|\mathbf{h}\|_2 - \mathbb{E} \sup_{\mathbf{x} \in T} \langle \mathbf{g}, \mathbf{x} \rangle = E_m - \ell(T) \ ,
\end{aligned}$$

where we have once applied the symmetry of a standard Gaussian vector.

Similarly to the proof of Proposition A.17 we argue that the function $F(\mathbf{A}) := \inf_{\mathbf{x} \in T} \|\mathbf{A}\mathbf{x}\|_2$ is Lipschitz with respect to the Frobenius norm. Indeed, for two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times N}$,

$$\inf_{\mathbf{x}\in T} \|\mathbf{A}\mathbf{x}\|_2 \le \inf_{\mathbf{x}\in T} (\|\mathbf{B}\mathbf{x}\|_2 + \|(\mathbf{A}-\mathbf{B})\mathbf{x}\|_2) \le \inf_{\mathbf{x}\in T} (\|\mathbf{B}\mathbf{x}\|_2 + \|\mathbf{A}-\mathbf{B}\|_{2\to 2}) \le \inf_{\mathbf{x}\in T} \|\mathbf{B}\mathbf{x}\|_2 + \|\mathbf{A}-\mathbf{B}\|_F \, .$$

Hereby, we have used that $T \subset S^{N-1}$ and that the operator norm is bounded by the Frobenius norm, see (A.16). Replacing the role of $\mathbf{A}$ and $\mathbf{B}$ we conclude that $|F(\mathbf{A}) - F(\mathbf{B})| \le \|\mathbf{A}-\mathbf{B}\|_F$. It follows from concentration of measure, Theorem 8.38, that

$$\mathbb{P}\big(\inf_{\mathbf{x}\in T} \|\mathbf{A}\mathbf{x}\|_2 \le \mathbb{E} \inf_{\mathbf{x}\in T} \|\mathbf{A}\mathbf{x}\|_2 - t\big) \le e^{-t^2/2} \, .$$

A combination with the estimate on the expectation $\mathbb{E}\inf_{\mathbf{x}\in T} \|\mathbf{A}\mathbf{x}\|_2$ derived above concludes the proof. □

Clearly, the estimate in the above theorem is only non-trivial if $\ell(T) < E_m$. Considering the recovery condition of Theorem 4.34 we are led to bounding the Gaussian widths of $T := T(\mathbf{x}) \cap S^{N-1}$, where

$$T(\mathbf{x}) = \operatorname{cone}\{\mathbf{z} - \mathbf{x} : \mathbf{z} \in \mathbb{R}^N, \|\mathbf{z}\|_1 \le \|\mathbf{x}\|_1\} \, .$$

Indeed, if

$$\inf_{\mathbf{x}\in T\cap S^{N-1}} \|\mathbf{A}\mathbf{x}\|_2 > 0$$

then $T \cap \ker \mathbf{A} = \emptyset$ and $T(\mathbf{x}) \cap \ker \mathbf{A} = \{\mathbf{0}\}$, so that Theorem 4.34 implies exact recovery of $\mathbf{x}$ from $\mathbf{A}\mathbf{x}$ via ℓ_1-minimization.

Recall the notion of polar cone in (B.3). The polar of $T(\mathbf{x})$ is the normal cone of the ℓ_1-norm at $\mathbf{x}$,

$$\begin{aligned} \mathcal{N}(\mathbf{x}) &= T(\mathbf{x})^\circ \\ &= \big\{\mathbf{z} \in \mathbb{R}^N : \langle \mathbf{z}, \mathbf{w} - \mathbf{x}\rangle \le 0 \text{ for all } \mathbf{w} \text{ such that } \|\mathbf{w}\|_1 \le \|\mathbf{x}\|_1\big\} \, . \end{aligned} \tag{9.28}$$

The next result bounds the Gaussian widths of T in terms of an expression in the normal cone $\mathcal{N}(\mathbf{x})$.

Proposition 9.20. *Let $\mathbf{g} \in \mathbb{R}^N$ be a standard Gaussian random vector. Then*

$$\ell(T(\mathbf{x}) \cap S^{N-1}) \le \mathbb{E} \min_{\mathbf{z}\in\mathcal{N}(\mathbf{x})} \|\mathbf{g} - \mathbf{z}\|_2 \, . \tag{9.29}$$

Proof. It follows from (B.39) that

$$\begin{aligned} \ell(T(\mathbf{x}) \cap S^{N-1}) &= \mathbb{E} \max_{\mathbf{z}\in T(\mathbf{x}), \|\mathbf{z}\|_2 = 1} \langle \mathbf{g}, \mathbf{z}\rangle \le \mathbb{E} \max_{\mathbf{z}\in T(\mathbf{x}), \|\mathbf{z}\|_2 \le 1} \langle \mathbf{g}, \mathbf{z}\rangle \\ &\le \mathbb{E} \min_{\mathbf{z}\in T(\mathbf{x})^\circ} \|\mathbf{g} - \mathbf{z}\|_2 \, . \end{aligned}$$

By definition of the normal cone, this establishes the claim. □

The previous result suggests to compute the normal cone of the ℓ_1-norm at a sparse vector.

Lemma 9.21. *Let $\mathbf{x} \in \mathbb{R}^N$ with $\operatorname{supp} \mathbf{x} = S \subset [N]$. Then*

$$\mathcal{N}(\mathbf{x}) = \left\{ \mathbf{z} \in \mathbb{R}^N, z_\ell = t \operatorname{sgn}(x_\ell) \text{ for } \ell \in S,\ |z_\ell| \le t \text{ for } \ell \in \overline{S},\ \text{for some } t \ge 0 \right\} . \tag{9.30}$$

Proof. If $\mathbf{z}$ is contained in the right hand side of (9.30) then, for $\mathbf{w}$ such that $\|\mathbf{w}\|_1 \le \|\mathbf{x}\|_1$,

$$\langle \mathbf{z}, \mathbf{w}-\mathbf{x} \rangle = \langle \mathbf{z}, \mathbf{w} \rangle - \langle \mathbf{z}, \mathbf{x} \rangle \le \|\mathbf{z}\|_\infty \|\mathbf{w}\|_1 - \|\mathbf{z}\|_\infty \|\mathbf{x}\|_1 = \|\mathbf{z}\|_\infty (\|\mathbf{w}\|_1 - \|\mathbf{x}\|) \le 0 ,$$

hence, $\mathbf{z} \in \mathcal{N}(\mathbf{x})$.

Now assume that $\mathbf{z} \in \mathcal{N}(\mathbf{x})$. If there would exist $\ell \in S$ such that $z_\ell \neq \|\mathbf{z}\|_\infty \operatorname{sgn}(x_\ell)$, then $\langle \mathbf{z}, \mathbf{x} \rangle \le \kappa \|\mathbf{z}\|_\infty \|\mathbf{x}\|_1$ for some $\kappa < 1$ and we can find a vector $\mathbf{w}$ with $\|\mathbf{w}\|_1 = \|\mathbf{x}\|_1$ such that $w_j = t \operatorname{sgn}(z_j)$ for an appropriate $t > 0$ for those $j \in [N]$ such that $|z_j| = \|\mathbf{z}\|_\infty$ and $w_j = 0$ for the remaining $j \in [N]$. Then

$$\langle \mathbf{z}, \mathbf{w} - \mathbf{x} \rangle \ge \|\mathbf{z}\|_\infty \|\mathbf{w}\|_1 - \kappa \|\mathbf{z}\|_\infty \|\mathbf{x}\|_1 = \|\mathbf{z}\|_\infty \|\mathbf{x}\|_1 (1 - \kappa) > 0$$

gives a contradiction to $\mathbf{z} \in \mathcal{N}(\mathbf{x})$, so that necessarily $z_\ell = \|\mathbf{z}\|_\infty \operatorname{sgn}(x_\ell)$ for all $\ell \in S$. Obviously, $|z_\ell| \le \|\mathbf{z}\|_\infty$ for all $\ell \in [N]$. Setting $t = \|\mathbf{z}\|_\infty$, we see that $\mathbf{z}$ is contained in the right hand side of (9.30). □

Now we are equipped to estimate the desired Gaussian widths.

Proposition 9.22. *Let $\mathbf{x} \in \mathbb{R}^N$ be s-sparse. Then*

$$\left(\ell(T(\mathbf{x}) \cap S^{N-1}) \right)^2 \le 2s \ln(2.34N/s) . \tag{9.31}$$

Proof. It follows from Proposition 9.20 and Hölder's inequality that

$$(\ell(T(\mathbf{x}) \cap S^{N-1}))^2 \le \left(\mathbb{E} \min_{\mathbf{z} \in \mathcal{N}(\mathbf{x})} \|\mathbf{g} - \mathbf{z}\|_2 \right)^2 \le \mathbb{E} \min_{\mathbf{z} \in \mathcal{N}(\mathbf{x})} \|\mathbf{g} - \mathbf{z}\|_2^2 . \tag{9.32}$$

Let $S = \operatorname{supp} \mathbf{x}$. Then $\operatorname{card}(S) = s$ and the normal cone $\mathcal{N}(\mathbf{x})$ is given by (9.30). We have

$$\min_{\mathbf{z} \in \mathcal{N}(\mathbf{x})} \|\mathbf{g} - \mathbf{z}\|_2^2 = \min_{\substack{t \ge 0 \\ |z_\ell| \le t, \ell \in \overline{S}}} \sum_{\ell \in S} (g_\ell - t \operatorname{sgn}(\mathbf{x}_\ell))^2 + \sum_{\ell \in \overline{S}} (g_\ell - z_\ell)^2 . \tag{9.33}$$

A straightforward computation shows that (see also Exercise 15.1)

$$\min_{|z_\ell| \le t} (g_\ell - z_\ell)^2 = S_t(g_\ell)^2 ,$$

where S_t is the soft-thresholding operator (B.17),

$$S_t(u) = \begin{cases} u - t & \text{if } u \geq t \,, \\ 0 & \text{if } |u| \leq t \,, \\ u + t & \text{if } u \leq -t \,. \end{cases}$$

Hence, for fixed $t > 0$ independent of $\mathbf{g}$,

$$\begin{aligned} \min_{\mathbf{z} \in \mathcal{N}(\mathbf{x})} \|\mathbf{g} - \mathbf{z}\|_2^2 &\leq \mathbb{E}\left[\sum_{\ell \in S} (g_\ell - t\mathrm{sgn}(\mathbf{x}_\ell))^2\right] + \mathbb{E}\left[\sum_{\ell \in \overline{S}} S_t(g_\ell)^2\right] \\ &= s\mathbb{E}(g+t)^2 + \sum_{\ell \in \overline{S}} \mathbb{E} S_t(g_\ell)^2 = s(1+t^2) + (N-s)\mathbb{E} S_t(g)^2 \,, \end{aligned}$$

where g is a standard (univariate) normal distributed random variable. It remains to estimate $\mathbb{E}S_t(g)^2$. Applying symmetry of g and S_t as well as integration by parts we get

$$\begin{aligned} &\mathbb{E} S_t(g)^2 = \frac{2}{\sqrt{2\pi}} \int_0^\infty S_t(u)^2 e^{-u^2/2} du = \sqrt{\frac{2}{\pi}} \int_t^\infty (u-t)^2 e^{-u^2/2} du \\ &= \sqrt{\frac{2}{\pi}} \left(\int_t^\infty (u-t) u e^{-u^2/2} du - t \int_t^\infty (u-t) e^{-u^2/2} du \right) \\ &= \sqrt{\frac{2}{\pi}} \left((u-t) e^{-u^2/2} \Big|_{u=t}^{u=\infty} + \int_t^\infty e^{-u^2/2} du + t^2 \int_t^\infty e^{-u^2/2} du - t e^{-t^2/2} \right) \\ &= \sqrt{\frac{2}{\pi}} \left((1+t^2) \int_t^\infty e^{-u^2/2} du - t e^{-t^2/2} \right) \,. \end{aligned}$$

We apply Lemma C.8 to reach

$$\mathbb{E} S_t(g)^2 \leq \sqrt{\frac{2}{\pi}} \left(\frac{1+t^2}{t} e^{-t^2/2} - t e^{-t^2/2} \right) = \sqrt{\frac{2}{\pi}} t^{-1} e^{-t^2/2} \,. \tag{9.34}$$

Now we choose $t = \sqrt{2\ln(N/s)}$. This choice gives

$$\begin{aligned} \min_{\mathbf{z} \in \mathcal{N}(\mathbf{x})} \|\mathbf{g} - \mathbf{z}\|_2^2 &\leq s(1 + 2\ln(N/s)) + (N-s)\sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{2\ln(N/s)}} \frac{s}{N} \\ &= s \left(2\ln(N/s) + \sqrt{\frac{\pi}{2}} (1 - s/N) \frac{1}{\sqrt{2\ln(N/s)}} + 1 \right) \,. \end{aligned} \tag{9.35}$$

This is already a slightly better (but more complicated) bound than claimed.

Let $\beta \in (0,1)$ be a parameter to be determined later. If $s \leq \beta N$ then the second term in (9.35) is bounded by

$$\sqrt{\frac{\pi}{2}} (1 - s/N) \frac{1}{\sqrt{2\ln(N/s)}} \leq \sqrt{\frac{\pi}{4\ln(\beta^{-1})}} =: c_1(\beta) \,.$$

If $\beta N \le s \le N$ we set $\alpha := N/s \in [1, \beta^{-1}]$. Then the term on the right hand side takes the form $\sqrt{\pi/2}(1-\alpha^{-1})/\sqrt{2\ln(\alpha)}$. By concavity of the logarithm, setting $\alpha = t + (1-t)\beta^{-1}$ for $t \in [0,1]$, we have

$$\ln(\alpha) \ge t\ln(1) + (1-t)\ln(\beta^{-1}) = \frac{\alpha - 1}{\beta^{-1} - 1}\ln(\beta^{-1}) \, .$$

This implies

$$\frac{1-\alpha^{-1}}{\sqrt{\ln(\alpha)}} \le \frac{1-\alpha^{-1}}{\sqrt{\frac{\alpha-1}{\beta^{-1}-1}\ln(\beta^{-1})}} = \alpha^{-1}\sqrt{\alpha-1}\sqrt{\frac{\beta^{-1}-1}{\ln(\beta^{-1})}} \le \frac{1}{2}\sqrt{\frac{\beta^{-1}-1}{\ln(\beta^{-1})}} \, ,$$

where we have used that $\alpha^{-1}\sqrt{\alpha-1} \le 1/2$ for all $\alpha \ge 1$. We obtain, for $s \ge \beta N$,

$$\sqrt{\frac{\pi}{2}}(1-s/N)\frac{1}{\sqrt{2\ln(N/s)}} \le \frac{\sqrt{\pi}}{4}\sqrt{\frac{\beta^{-1}-1}{\ln(\beta^{-1})}} =: c_2(\beta) \, .$$

Choosing $\beta = 1/5$ gives $c_1(\beta) = c_2(\beta) = \sqrt{\pi/(4\ln(5))} \approx 0.6986$. Therefore, with $c_3 = 1 + c_1(\beta) = 1 + \sqrt{\pi/(4\ln(5))} \approx 1.6986$ we obtain

$$\min_{\mathbf{z}\in\mathcal{N}(\mathbf{x})} \|\mathbf{g}-\mathbf{z}\|_2^2 \le 2s(\ln(N/s) + c_3/2) = 2s\ln(cN/s)$$

with $c = \exp(1/2 + \sqrt{\pi/(4\ln(5))}/2) \approx 2.3380 < 2.34$. We arrived at the desired estimate. □

Proof (of Theorem 9.16). Set $t = \sqrt{2\ln(\varepsilon^{-1})}$. By Proposition 9.22 and since $E_m \ge m/\sqrt{m+1}$, see Proposition 8.1(b), the conditions in Theorem 9.16 ensure that

$$E_m - \ell(T(\mathbf{x}) \cap S^{N-1}) - t \ge 0 \, .$$

It follows from Theorem 9.19 that

$$\begin{aligned} &\mathbb{P}\left(\min_{T(\mathbf{x})\cap S^{N-1}} \|\mathbf{A}\mathbf{x}\|_2 > 0\right) \\ &\ge \mathbb{P}\left(\min_{T(\mathbf{x})\cap S^{N-1}} \|\mathbf{A}\mathbf{x}\|_2 > E_m - \ell(T(\mathbf{x})\cap S^{N-1}) - t\right) \\ &\ge 1 - e^{-t^2/2} = 1 - \varepsilon \, . \end{aligned} \tag{9.36}$$

This implies that $T(\mathbf{x}) \cap \ker \mathbf{A} = \{\mathbf{0}\}$ with probability at least $1-\varepsilon$. An application of Theorem 4.34 concludes the proof. □

Proof (of Theorem 9.18). With the same notation as in the previous proof, the assumptions of Theorem 9.18 imply

$$E_m - \ell(T(\mathbf{x}) \cap S^{N-1}) - \tau - t \ge 0 \, .$$

As in (9.36) we conclude that

$$\mathbb{P}\left(\min_{\mathbf{z}\in T(\mathbf{x})\cap S^{N-1}} \|\mathbf{A}\mathbf{z}\|_2 \geq \tau\right) \leq 1-\varepsilon\,.$$

The claim follows then from Theorem 4.36. □

Remark 9.23. The alternative choice $t = \sqrt{2\ln((N-s)/s)-1}$ — valid if $s < (1+e)^{-1}N$ — in the proof of Proposition 9.22 allows to deduce the slightly more precise estimate

$$\ell(T(\mathbf{x})\cap S^{N-1}) \leq 2s\left(\ln\left(\frac{N-s}{s}\right) + \frac{2e}{\sqrt{2\pi}\sqrt{\ln((N-s)/s)-1}}\right)\,.$$

Therefore, the recovery condition (9.23) can in this case be refined to

$$\frac{m^2}{m+1} \geq 2s\left(\sqrt{\ln\left(\frac{N-s}{s}\right) + \frac{2e}{\sqrt{2\pi}\sqrt{\ln((N-s)/s)-1}}} + \sqrt{\frac{\ln(\varepsilon^{-1})}{s}}\right)^2\,,$$

to ensure nonuniform recovery via ℓ_1-minimization with probability at least $1-\varepsilon$. Roughly speaking for large N, mildly large s and large ratio N/s we therefore get the "asymptotic" recovery condition

$$m \geq 2s\ln(N/s)\,. \tag{9.37}$$

This is the general rule of thumb for compressive sensing, and reflects well empirical tests for sparse recovery using Gaussian matrices, but also different random matrices. However, our proof of this result is restricted to the Gaussian case.

9.3 Gaussian Random Matrices

We return now to uniform recovery and specialize to Gaussian matrices, where we can provide explicit constants. We treat again the restricted isometry property, but also give a direct estimate for the null space property of Gaussian matrices. For the latter, the constants turn out to be very reasonable.

Restricted Isometry Property

In this section we give an alternative proof of the restricted isometry property for Gaussian matrices, that provides explicit constants (which are better than the ones the previous analysis would provide when specializing to Gaussian matrices). The approach of this section is based on concentration of measure, Theorem 8.38, and on the Slepian-Gordon lemmas, and therefore does not generalize to subgaussian matrices.

We start with estimates for the extremal singular values of a Gaussian random matrix.

Theorem 9.24. *Let* $\mathbf{A}$ *be an* $m \times s$ *Gaussian matrix with* $m > s$*, and let* $\sigma_{\min}$*,* $\sigma_{\max}$ *be the smallest resp. largest singular value of the renormalized matrix* $\frac{1}{\sqrt{m}}\mathbf{A}$*. Then, for* $t > 0$*,*

$$\mathbb{P}(\sigma_{\max} \geq 1 + \sqrt{s/m} + t) \leq e^{-mt^2/2} , \tag{9.38}$$

$$\mathbb{P}(\sigma_{\min} \leq 1 - \sqrt{s/m} - t) \leq e^{-mt^2/2} . \tag{9.39}$$

Proof. By Proposition A.17 the extremal singular values are 1-Lipschitz functions with respect to the Frobenius norm (which corresponds to the ℓ_2-norm by identifying $\mathbb{R}^{m\times s}$ with $\mathbb{R}^{ms}$). Therefore, it follows from concentration of measure for Gaussian vectors, Theorem 8.38, that in particular, the largest singular value of the non-normalized matrix $\mathbf{A}$ satisfy

$$\mathbb{P}(\sigma_{\max}(\mathbf{A}) \geq \mathbb{E}[\sigma_{\max}(\mathbf{A})] + r) \leq e^{-r^2/2} . \tag{9.40}$$

Let us estimate the expectation above. For this task we will use the Slepian Lemma 8.26.

Let $S^{s-1} = \{\mathbf{x} \in \mathbb{R}^s, \|\mathbf{x}\|_2 = 1\}$ denote the sphere in $\mathbb{R}^s$. Observe that

$$\sigma_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in S^{s-1}} \sup_{\mathbf{y}\in S^{m-1}} \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle . \tag{9.41}$$

As in the proof of Theorem 9.19 we introduce two Gaussian processes by

$$\begin{aligned} X_{\mathbf{x},\mathbf{y}} &:= \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle = \operatorname{tr}(\mathbf{A}\mathbf{x}\mathbf{y}^*) , \\ Y_{\mathbf{x},\mathbf{y}} &:= \sum_{j=1}^{N} g_j x_j + \sum_{k=1}^{m} h_j y_j = \langle \mathbf{x}, \mathbf{g}\rangle + \langle \mathbf{y}, \mathbf{h}\rangle , \end{aligned} \tag{9.42}$$

where $\mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^s$ are two independent standard Gaussian vectors. Then $\sigma_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in S^{s-1}} \sup_{\mathbf{y}\in S^{m-1}} X_{\mathbf{x},\mathbf{y}}$. By (9.27) we have

$$\mathbb{E}|X_{\mathbf{x},\mathbf{y}} - X_{\mathbf{x}',\mathbf{y}'}|^2 \leq \mathbb{E}|Y_{\mathbf{x},\mathbf{y}} - Y_{\mathbf{x}',\mathbf{y}'}|^2 . \tag{9.43}$$

Slepian's Lemma 8.26 (see also Remark 8.29) implies that

$$\begin{aligned} \mathbb{E}\,\sigma_{\max}(\mathbf{A}) &= \mathbb{E} \sup_{\mathbf{x}\in S^{s-1},\mathbf{y}\in S^{m-1}} X_{\mathbf{x},\mathbf{y}} \leq \mathbb{E} \sup_{\mathbf{x}\in S^{s-1},\mathbf{y}\in S^{m-1}} Y_{\mathbf{x},\mathbf{y}} \\ &= \mathbb{E} \sup_{\mathbf{x}\in S^{s-1}} \langle \mathbf{g}, \mathbf{x}\rangle + \mathbb{E} \sup_{\mathbf{y}\in S^{m-1}} \langle \mathbf{h}, \mathbf{y}\rangle = \mathbb{E}\|\mathbf{g}\|_2 + \mathbb{E}\|\mathbf{h}\|_2 \\ &\leq \sqrt{\mathbb{E}\|\mathbf{g}\|_2^2} + \sqrt{\mathbb{E}\|\mathbf{h}\|_2^2} = \sqrt{s} + \sqrt{m} . \end{aligned}$$

The inequality on the third line is Cauchy-Schwarz and the last equality follows from Proposition 8.1(b). Plugging this estimate into (9.40) shows that

$$\mathbb{P}(\sigma_{\max}(\mathbf{A}) \geq \sqrt{m} + \sqrt{s} + r) \leq e^{-r^2/2} .$$

Rescaling by $\frac{1}{\sqrt{m}}$ shows the estimate (9.38) for the largest singular value of $\frac{1}{\sqrt{m}}\mathbf{A}$.

The smallest singular value $\sigma_{\min}(\mathbf{A}) = \inf_{\mathbf{x}\in S^{s-1}} \|\mathbf{A}\mathbf{x}\|_2$ can be estimated with the help of Theorem 9.19 (which used concentration of measure for Lipschitz functions in its proof as well). The required Gaussian width of $T = S^{s-1}$ is given, for a standard Gaussian vector $\mathbf{g}$ in $\mathbb{R}^s$, by

$$\ell(S^{s-1}) = \mathbb{E} \sup_{\mathbf{x}\in S^{s-1}} \langle \mathbf{x}, \mathbf{g}\rangle = \mathbb{E}\|\mathbf{g}\|_2 = E_s \, .$$

By Proposition 8.1(c) and Lemma C.4 we further obtain

$$E_m - \ell(S^{s-1}) = \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)} - \sqrt{2}\frac{\Gamma((s+1)/2)}{\Gamma(s/2)} \geq \sqrt{m} - \sqrt{s} \, .$$

Together with Theorem 9.19 this concludes the proof. □

With this tool at hand we can easily show the restricted isometry property of Gaussian matrices.

Theorem 9.25. *Let* $\mathbf{A}$ *be an* $m \times N$ *Gaussian matrix with* $m < N$*. For* $\eta, \varepsilon \in (0,1)$ *assume that*

$$m \geq 2\eta^{-2}\left(s\ln(eN/s) + \ln(2\varepsilon^{-1})\right). \tag{9.44}$$

Then with probability at least $1-\varepsilon$ *the restricted isometry constant* δ_s *of* $\frac{1}{\sqrt{m}}\mathbf{A}$ *satisfies*

$$\delta_s \leq 2\left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)\eta + \left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)^2 \eta^2. \tag{9.45}$$

Remark 9.26. Note that (9.45) implies the simpler inequality $\delta_s \leq C\eta$ with $C = 2(1+\sqrt{1/2}) + (1+\sqrt{1/2})^2 \approx 6.3284$. In other words, the condition

$$m \geq \tilde{C}\delta^{-2}\left(s\ln(eN/s) + \ln(2\varepsilon^{-1})\right)$$

with $\tilde{C} = 2C^2 \approx 80.1$ implies $\delta_s \leq \delta$. In most situations, that is, if $s \ll N$, the statement of the theorem provides better constants. For instance, if $2\ln(eN/s) \geq 8$, that is, $N/s \geq e^3 \approx 20.08$ and $\eta = 0.16$ then $\delta_s \leq 0.48 < 0.4931$ (compare Theorem 6.11 concerning ℓ_1-minimization) provided

$$m \geq 78.13\left(s\ln(eN/s) + \ln(2\varepsilon^{-1})\right).$$

Further, in the limit, as $N/s \to \infty$ we get $\delta_s \leq C_1\eta + C_2\eta^2$ with $C_1 = 2$ and $C_2 = 1$. Then the choice $\eta = 0.22$ yields $\delta_s \leq 0.4884$ under the condition $m \geq 41.32\left(s\ln(eN/s) + \ln(2\varepsilon^{-1})\right)$ in this asymptotic regime.

Proof (of Theorem 9.25). We proceed similarly as in Theorem 9.10. Let $S \subset [N]$ be of cardinality s. Clearly, $\mathbf{A}_S$ is an $m \times s$ Gaussian matrix and the eigenvalues of $\frac{1}{m}\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}$ are contained in $[\sigma_{\min}^2 - 1, \sigma_{\max}^2 - 1]$ where $\sigma_{\min}$, $\sigma_{\max}$ are the extremal singular values of $\frac{1}{\sqrt{m}}\mathbf{A}_S$. Denote $\widetilde{\mathbf{A}}_S = \frac{1}{\sqrt{m}}\mathbf{A}$. Theorem 9.24 implies that

$$\begin{aligned}\|\widetilde{\mathbf{A}}_S^*\widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} &\le \max\left\{(1+\sqrt{s/m}+\eta)^2 - 1, 1 - (1-(\sqrt{s/m}+\eta))^2\right\} \\ &= 2(\sqrt{s/m}+\eta) + (\sqrt{s/m}+\eta)^2 \,.\end{aligned}$$

with probability at least $1 - 2\exp(-m\eta^2/2)$. Taking the union bound over all $\binom{N}{s}$ and in view of the definition of the restricted isometry constant, $\delta_s = \max_{S\subset[N],\mathrm{card}(S)} \|\widetilde{\mathbf{A}}_S^*\widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2}$ we obtain

$$\begin{aligned}\mathbb{P}\left(\delta_s > 2(\sqrt{s/m}+\eta) + (\sqrt{s/m}+\eta)^2\right) &\le 2\binom{N}{s} e^{-m\eta^2/2} \\ &\le 2\left(\frac{eN}{s}\right)^s e^{-m\eta^2/2} \,.\end{aligned}$$

In the second inequality we have applied Lemma C.5. The last term is dominated by ε due to condition (9.44), which also implies $\sqrt{s/m} \le \frac{\eta}{\sqrt{2\ln(eN/s)}}$. The conclusion of the theorem follows. □

Null Space Property

Our next theorem states stable uniform recovery with Gaussian random matrices via ℓ_1-minimization. It is established by directly showing the stable null space property in Definition 4.10 rather than by relying on the restricted isometry property.

Theorem 9.27. *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a random draw of a Gaussian matrix. Assume that*

$$\frac{m^2}{m+1} \ge 2s\ln(eN/s)\left(\rho^{-1} + D(s/N) + \sqrt{\frac{\ln(\varepsilon^{-1})}{s\ln(eN/s)}}\right)^2 \,. \tag{9.46}$$

where, for $\alpha \in (0,1)$,

$$D(\alpha) := \inf_{\kappa>0}\left\{\sqrt{(1+\kappa) + \frac{(1+\kappa)\ln(1+\kappa^{-1})}{2\ln(e\alpha^{-1})}} + \left(\frac{2(1-\alpha)}{\pi e^2 \ln^3(e\alpha^{-1})}\right)^{1/4}\right\} \,. \tag{9.47}$$

Then with probability at least $1-\varepsilon$ the following holds for every vector $\mathbf{x} \in \mathbb{R}^N$. Let $\mathbf{x}^\sharp$ be the minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Ax} = \mathbf{Az}$. Then

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \le \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x})_1 \, .$$

The function D satisfies $D(\alpha) \le 2.05$ for all $\alpha \in (0,1)$ and

$$\lim_{\alpha \to 0} D(\alpha) = 1 \, .$$

Remark 9.28. (a) Roughly speaking, for large N, mildly large s and small quotient s/N (which is the situation of most interest in compressive sensing) then Condition (9.46) turns into

$$m \ge 2(1+\rho^{-1})^2 s \ln(eN/s) \, .$$

(b) The proof proceeds by establishing the null space property of order s with constant ρ. Letting $\rho = 1$ yields therefore uniform exact recovery of all s-sparse vectors under roughly the condition

$$m \ge 8 s \ln(eN/s) \, . \tag{9.48}$$

(c) The claims on D above can be seen as follows. The choise $\kappa = 0.35$, bounding the term $(1-\alpha)$ by 1, and then setting $\alpha = 1$ gives the bound $D(\alpha) \le 2.05$. When $\alpha \to 0$ we may choose for instance $\kappa = \left(\exp(\sqrt{\ln(e\alpha^{-1})}) - 1\right)^{-1}$ in the definition of D so that also $\kappa \to 0$, and we conclude that $\lim_{\alpha \to 0} D(\alpha) = 1$.

The proof proceeds with a similar strategy as in the previous section. In particular, we use Gordon's escape through the mesh Theorem 9.19. For $\rho \in (0,1]$ we introduce the set

$$T_{\rho,s} := \left\{ \mathbf{w} \in \mathbb{R}^N : \|\mathbf{w}_S\|_1 \ge \rho \|\mathbf{w}_{\overline{S}}\|_1 \text{ for some } S \subset [N], \operatorname{card}(S) = s \right\} \, .$$

If

$$\min\{\|\mathbf{A}\mathbf{w}\|_2 : \mathbf{w} \in T_{\rho,s} \cap S^{N-1}\} > 0 \tag{9.49}$$

then

$$\|\mathbf{v}_S\|_1 < \rho \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{for all } \mathbf{v} \in \ker \mathbf{A} \setminus \{\mathbf{0}\}, S \subset [N], \operatorname{card}(S) = s \, ,$$

so that the stable null space property holds. This implies that we have stable recovery of all (approximately) s-sparse vectors by Theorem 4.11. Following Theorem 9.19 we are led to study the Gaussian widths of the set $T_{\rho,s} \cap S^{N-1}$. As a first step we relate this problem to the following simpler set

$$K_{\rho,s} := \left\{ \mathbf{u} \in \mathbb{R}^N : u_\ell \ge 0 \text{ for all } \ell \in [N], \sum_{\ell=1}^{s} u_\ell \ge \rho \sum_{\ell=s+1}^{N} u_\ell \right\} \, , \tag{9.50}$$

which is a convex cone. Our next result is similar to Proposition 9.20. We recall that the nonincreasing rearrangement $\mathbf{g}^*$ of a vector $\mathbf{g}$ has entries $g_j^* = |g_{\ell_j}|$ with a permutation $j \mapsto \ell_j$ of $[N]$ such that $g_1^* \ge g_2^* \ge \cdots \ge g_N^* \ge 0$, see Definition 2.4.

Proposition 9.29. *Let* $\mathbf{g} \in \mathbb{R}^N$ *be a standard Gaussian vector and* $\mathbf{g}^*$ *its non-increasing rearrangement. Then*

$$\ell(T_{\rho,s} \cap S^{N-1}) \le \mathbb{E} \min_{\mathbf{z} \in K^*_{\rho,s}} \|\mathbf{g}^* + \mathbf{z}\|_2 \,,$$

where $K^*_{\rho,s}$ *is the dual cone of* $K_{\rho,s}$*, see* (B.2)*.*

Proof. Consider the maximization problem $\max_{\mathbf{w} \in T_{\rho,s} \cap S^{N-1}} \langle \mathbf{g}, \mathbf{w} \rangle$ appearing in the definition of the Gaussian widths (9.25) (since $T_{\rho,s} \cap S^{N-1}$ is compact the maximum is attained). Changing the sign of any entry of a vector $\mathbf{w} \in T_{\rho,s}$ preserves membership in this set, as well as any permutation of the entries of $\mathbf{w}$. It follows that

$$\max_{\mathbf{w} \in T_{\rho,s} \cap S^{N-1}} \langle \mathbf{g}, \mathbf{w} \rangle = \max_{\mathbf{w} \in T_{\rho,s} \cap S^{N-1}} \langle \mathbf{g}^*, \mathbf{w}^* \rangle \,.$$

Now if $\mathbf{w}$ ranges through all vectors in $T_{\rho,s}$ then $\mathbf{w}^*$ ranges through $K_{\rho,s}$. Therefore,

$$\max_{\mathbf{w} \in T_{\rho,s} \cap S^{N-1}} \langle \mathbf{g}, \mathbf{w} \rangle = \max_{\mathbf{u} \in K_{\rho,s} \cap S^{N-1}} \langle \mathbf{g}^*, \mathbf{u} \rangle \le \min_{\mathbf{z} \in K^*_{\rho,s}} \|\mathbf{g}^* + \mathbf{z}\|_2 \,,$$

where the inequality follows from (B.38). By definition of the Gaussian widths (9.25) the claim follows. □

The previous results suggests to compute the dual cone $K^*_{\rho,s}$.

Lemma 9.30. *The dual cone of* $K_{\rho,s}$ *defined in* (9.50) *is given by*

$$K^*_{\rho,s} = \big\{ \mathbf{z} \in \mathbb{R}^N : z_\ell = t, \ell \in [s],\, z_\ell \ge -\rho t, \ell = s+1, \dots, N,\, t \ge 0 \big\} \,.$$

Proof. Take a vector $\mathbf{z}$ in the right hand set. Then, for any $\mathbf{u} \in K_{\rho,s}$,

$$\langle \mathbf{z}, \mathbf{u} \rangle = \sum_{\ell=1}^{s} z_\ell u_\ell + \sum_{\ell=s+1}^{N} z_\ell u_\ell \ge t \sum_{\ell=1}^{s} u_\ell - t\rho \sum_{\ell=s+1}^{N} u_\ell \ge 0 \,. \tag{9.51}$$

Therefore, $\mathbf{z} \in K^*_{\rho,s}$. The converse inclusion is shown in a similar way as in the proof of Lemma 9.21. □

With this preparation we estimate the Gaussian widths of $T_{\rho,s} \cap S^{N-1}$.

Proposition 9.31. *It holds*

$$\ell(T_{\rho,s} \cap S^{N-1}) \le \sqrt{2s \ln(eN/s)} \left(\rho^{-1} + D(s/N) \right) \,,$$

where D *is the function in* (9.47)*.*

Proof. By Proposition (9.29) it remains to estimate

$$E := \mathbb{E} \min_{\mathbf{z}\in K^*_{\rho,s}} \|\mathbf{g}^* + \mathbf{z}\|_2 \le \mathbb{E} \min_{\mathbf{z}\in K^*_{\rho,s}} \|\mathbf{g}^* + \mathbf{z}\|_2$$
$$= \mathbb{E} \min_{\substack{t\ge 0 \\ z_\ell \ge -\rho t, \ell=s+1,\dots,N}} \sqrt{\sum_{\ell=1}^{s} (g_\ell^* + t)^2} + \sqrt{\sum_{\ell=s+1}^{N} (g_\ell^* + z_\ell)^2} \,.$$

Consider a fixed $t \ge 0$. Then

$$\mathbb{E} \min_{\mathbf{z}\in K^*_{\rho,s}} \|\mathbf{g}^* + \mathbf{z}\|_2 \le \mathbb{E} \left[\sum_{\ell=1}^{s} (|g_\ell^*| + t)^2\right]^{1/2} + \mathbb{E} \left[\min_{z_\ell \ge -\rho t} \sum_{\ell=s+1}^{N} (|g_\ell| + z_\ell)^2\right]^{1/2}$$
$$\le \mathbb{E} \sqrt{\sum_{\ell=1}^{s} (g_\ell^*)^2} + t\sqrt{s} + \mathbb{E} \left[\sum_{\ell=s+1}^{N} S_{\rho t}(g_\ell)^2\right]^{1/2} , \qquad (9.52)$$

where g is a (univariate) standard Gaussian and $S_{\rho t}$ is the soft-thresholding operator (B.17). It follows from Hölder's inequality and (9.34) that the last term above can be estimated by

$$\mathbb{E} \left[\sum_{\ell=s+1}^{N} S_{\rho t}(g_\ell)^2\right]^{1/2} \le \left[\mathbb{E} \sum_{\ell=s+1}^{N} S_{\rho t}(g_\ell)^2\right]^{1/2} = \sqrt{(N-s)\mathbb{E} S_{\rho t}(g)^2}$$
$$\le \sqrt{(N-s)\sqrt{\frac{2}{\pi}} \frac{e^{-(\rho t)^2/2}}{\rho t}} \,.$$

It remains to estimate the first term in (9.52). By Hölder's inequality and Proposition 8.2, for any $\kappa > 0$,

$$\mathbb{E} \sqrt{\sum_{\ell=1}^{s} (g_\ell^*)^2} = \mathbb{E} \max_{S\subset[N],\mathrm{card}(S)=s} \|\mathbf{g}_S\|_2 \le \sqrt{\mathbb{E} \max_{S\subset[N],\mathrm{card}(S)=s} \|\mathbf{g}_S\|_2^2} \,.$$
$$\le \sqrt{(2+2\kappa) \ln \binom{N}{s} + (1+\kappa)\ln(1+\kappa^{-1})s}$$
$$\le \sqrt{(2+2\kappa)s\ln(eN/s) + (1+\kappa)\ln(1+\kappa^{-1})s} \,. \qquad (9.53)$$

Altogether we have estimated

$$E \le \sqrt{(2+2\kappa)s\ln(eN/s) + (1+\kappa)\ln(1+\kappa^{-1})s} + t\sqrt{s}$$
$$+ \sqrt{(N-s)\sqrt{\frac{2}{\pi}} \frac{e^{-(\rho t)^2/2}}{\rho t}} \,. \qquad (9.54)$$

We choose $t = \rho^{-1}\sqrt{2\ln(eN/s)}$ to obtain

$$
\begin{aligned}
E &\le \sqrt{(2+2\kappa)s\ln(eN/s) + (1+\kappa)\ln(1+\kappa^{-1})s} + \rho^{-1}\sqrt{2s\ln(eN/s)} \\
&\quad + \sqrt{s\frac{N-s}{N}\sqrt{\frac{2}{\pi}}\frac{1}{e\sqrt{2\ln(eN/s)}}} \\
&= \sqrt{2s\ln(eN/s)} \\
&\quad \times \left(\rho^{-1} + \sqrt{(1+\kappa) + \frac{(1+\kappa)\ln(1+\kappa^{-1})}{2\ln(eN/s)}} + \left(\frac{2(1-s/N)}{\pi e^2 \ln^3(eN/s)}\right)^{1/4}\right) .
\end{aligned}
$$

Taking the infimum over $\kappa > 0$ shows that

$$
E \le \sqrt{2s\ln(eN/s)}\left(\rho^{-1} + D(s/N)\right) .
$$

This completes the proof. □

In view of Theorem 4.11, the uniform recovery result of Theorem 9.27 is now an immediate consequence of the following statement.

Corollary 9.32. *Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a random draw of a Gaussian matrix. Let $s < N, \rho \in (0,1], \varepsilon \in (0,1)$ such that*

$$
\frac{m^2}{m+1} \ge 2s\ln(eN/s)\left(\rho^{-1} + D(s/N) + \sqrt{\frac{\ln(\varepsilon^{-1})}{s\ln(eN/s)}}\right)^2 .
$$

Then with probability at least $1-\varepsilon$ the matrix $\mathbf{A}$ satisfies the stable null space property of order s with constant ρ.

Proof. Taking into account the preceding results, the proof is a variation of the one of Theorem 9.16, see also Exercise 9.8. □

9.4 Relation to Johnson-Lindenstrauss Embeddings

The Johnson-Lindenstrauss Lemma is not a statement connected with sparsity per se, but it is closely related to the concentration inequality (9.6) for subgaussian matrices leading to the restricted isometry property. Assume that we are given a finite set $\{\mathbf{x}_1, \dots, \mathbf{x}_M\} \subset \mathbb{R}^N$ of points. If N is large then it is usually computationally expensive to process these points. Therefore, it is of interest to project these points into a lower dimensional space while preserving essential geometrical properties such as mutual distances. The Johnson-Lindenstrauss lemma states that such lower dimensional embeddings exist. For simplicity we state our results for the real case, but note that it has immediate extensions to $\mathbb{C}^N$ (for instance, simply by identifying $\mathbb{C}^N$ with $\mathbb{R}^{2N}$).

Lemma 9.33. *Let $\mathbf{x}_1, \dots, \mathbf{x}_M \in \mathbb{R}^N$ be an arbitrary set of points and $\eta > 0$. If $m > C\eta^{-2}\ln(M)$, then there exists a matrix $\mathbf{B} \in \mathbb{R}^{m\times N}$ such that*

$$
(1-\eta)\|\mathbf{x}_j - \mathbf{x}_\ell\|_2^2 \le \|\mathbf{B}(\mathbf{x}_j - \mathbf{x}_\ell)\|_2^2 \le (1+\eta)\|\mathbf{x}_j - \mathbf{x}_\ell\|_2^2
$$

for all $j, \ell \in [M]$. The constant $C > 0$ is universal.

Proof. Considering the set

$$E = \{\mathbf{x}_j - \mathbf{x}_\ell : \ 1 \le j < \ell \le M\}$$

of cardinality $\mathrm{card}(E) \le M(M-1)/2$, it is enough to show that

$$(1-\eta)\|\mathbf{x}\|_2^2 \le \|\mathbf{B}\mathbf{x}\|_2^2 \le (1+\eta)\|\mathbf{x}\|_2^2 \quad \text{for all } \mathbf{x} \in E\,. \tag{9.55}$$

We take $\mathbf{B} = \frac{1}{\sqrt{m}}\mathbf{A} \in \mathbb{R}^{m\times N}$, where $\mathbf{A}$ is a random draw of a subgaussian matrix. Then (9.6) implies that for any fixed $\mathbf{x} \in E$ and an appropriate constant $\tilde{c}$

$$\mathbb{P}\big(\,\big|\|\mathbf{B}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \ge \eta\|\mathbf{x}\|_2^2\big) \le 2\exp(-\tilde{c}m\eta^2)\,.$$

By the union bound (9.55) holds simultaneously for all $\mathbf{x} \in E$ with probability at least

$$1 - M^2 e^{-\tilde{c}m\eta^2}\,.$$

Setting $\varepsilon = M^2\exp(-\tilde{c}m\eta^2)$ so that $m = \tilde{c}^{-1}\eta^{-2}\ln(M^2/\varepsilon)$ inequality (9.55) holds with probability at least $1-\varepsilon$, and existence of a map with the desired property is established when $\varepsilon < 1$. This gives the claim with $C = 2\tilde{c}^{-1}$. □

This proof shows that the concentration inequality (9.6) is closely related to the Johnson-Lindenstrauss lemma. As (9.6) implies the restricted isometry property by Theorem 9.10, one may even say that in this sense the Johnson-Lindenstrauss lemma implies the restricted isometry property. We will show next that in some sense also the converse holds: Given a matrix $\mathbf{A}$ satisfying the restricted isometry, randomization of the column signs of $\mathbf{A}$ provides a Johnson-Lindenstrauss embedding. For a Rademacher sequence $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)$ we denote $\mathbf{D}_{\boldsymbol{\epsilon}} = \mathrm{diag}(\boldsymbol{\epsilon})$ the diagonal matrix with $\boldsymbol{\epsilon}$ on the diagonal.

Theorem 9.34. *Let $E \subset \mathbb{R}^N$ be a finite point set of cardinality $\mathrm{card}(E) = M$. Fix $\eta, \varepsilon \in (0,1)$. Let $\mathbf{A} \in \mathbb{R}^{m\times N}$, and assume that its restricted isometry constant satisfies $\delta_{2s} \le \eta/4$ for some $s \ge 16\ln(4M/\varepsilon)$. Then with probability exceeding $1-\varepsilon$*

$$(1-\eta)\|\mathbf{x}\|_2^2 \le \|\mathbf{A}\mathbf{D}_{\boldsymbol{\epsilon}}\mathbf{x}\|_2^2 \le (1+\eta)\|\mathbf{x}\|_2^2 \quad \textit{for all } \mathbf{x} \in E\,.$$

Remark 9.35. (a) Without randomization of the column signs, the theorem is false. Indeed, there is no assumption on the point set E. Therefore, if we choose the points of E to be in the kernel of the matrix $\mathbf{A}$ (which is not assumed random here), there is no chance that the lower bound can hold. Randomization of the column signs ensures that the probability, that E falls in the kernel of $\mathbf{A}\mathbf{D}_{\boldsymbol{\epsilon}}$ (or close to it) is very small.

(b) There is no direct condition on the embedding dimension m in the previous theorem, but of course, the requirement $\delta_{2s} \le \eta/4$ for $\mathbf{A} \in \mathbb{R}^{m\times n}$ poses an indirect condition on m. For "good" matrices one expects that this requires $m \ge C\eta^{-2}s\ln^\alpha(N)$, say, so that the condition on s in the previous

result turns into $m \geq C\eta^{-2}\ln^{\alpha}(N)\ln(M/\varepsilon)$. In comparison to the original Johnson-Lindenstrauss Lemma 9.33 we only observe an additional factor of $\ln^{\alpha}(N)$.

(c) The theorem allows to derive Johnson-Lindenstrauss embeddings also for other types of matrices rather than just subgaussian random matrices. In Chapter 12 we will see indeed different types of matrices $\mathbf{A}$ satisfying the restricted isometry property, so that $\mathbf{AD}_{\boldsymbol{\epsilon}}$ will provide a Johnson-Lindenstrauss embedding; for instance partial random Fourier matrices. It seems presently not known how to show the Johnson-Lindenstrauss embedding directly for such type of matrices.

Proof. Without loss of generality we may assume that all $\mathbf{x} \in E$ are normalized, $\|\mathbf{x}\|_2 = 1$. Consider a fixed $\mathbf{x} \in E$. Similarly as in the proof of Theorem 6.8 we partition $\mathbf{x}$ into blocks of size s according to its non-increasing rearrangement. More precisely, $S_1 \subset [N]$ is an index set of s largest absolute entries of the vector $\mathbf{x}$, $S_2 \subset [N]\setminus S_1$ is an index set of s largest absolute entries of $\mathbf{x}$ in $[N] \setminus S_1$, and so on. Note that as usual, $\mathbf{x}_S$ (and similar expressions below) can both have the meaning of restricting the vector $\mathbf{x}$ to the indices in S as well as being the vector whose entries are set to zero out S.

We write

$$\begin{aligned}\|\mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}\|_2^2 &= \|\mathbf{AD}_{\boldsymbol{\epsilon}}\sum_j \mathbf{x}_{S_j}\|_2^2 \\ &= \sum_j \|\mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_j}\|_2^2 + 2\langle \mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_1}, \mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{\overline{S_1}}\rangle + \sum_{\substack{j,\ell\geq 2\\ j\neq\ell}} \langle \mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_j}, \mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_\ell}\rangle\ .\end{aligned} \tag{9.56}$$

As $\mathbf{A}$ possesses the restricted isometry property, $\delta_s \leq \eta/4$, and since $\|\mathbf{D}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_j}\|_2 = \|\mathbf{x}_{S_j}\|_2$ the first term satisfies

$$(1-\eta/4)\|\mathbf{x}\|_2^2 = (1-\eta/4)\sum_j \|\mathbf{x}_{S_j}\|_2^2 \leq \sum_j \|\mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_j}\|_2^2 \leq (1+\eta/4)\|\mathbf{x}\|_2^2\ .$$

To estimate the second term in (9.56) we consider

$$X := \langle \mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_1}, \mathbf{AD}_{\boldsymbol{\epsilon}}\mathbf{x}_{\overline{S_1}}\rangle = \langle \mathbf{v}, \boldsymbol{\epsilon}_{\overline{S_1}}\rangle = \sum_{\ell\notin S_1} \epsilon_\ell v_\ell$$

with $\mathbf{v} \in \mathbb{R}^{\overline{S_1}}$ given by

$$\mathbf{v} = \mathbf{D}_{\mathbf{x}_{\overline{S_1}}}\mathbf{A}_{\overline{S_1}}^*\mathbf{A}_{S_1}\mathbf{D}_{\mathbf{x}_{S_1}}\boldsymbol{\epsilon}_{S_1}\ .$$

Hereby, we exploited that $D_{\boldsymbol{\epsilon}}\mathbf{x} = D_{\mathbf{x}}\boldsymbol{\epsilon}$. Observe that $\mathbf{v}$ and $\boldsymbol{\epsilon}_{\overline{S_1}}$ are stochastically independent. We aim at applying Hoeffding's inequality, Corollary 7.21, which requires to estimate the 2-norm of the vector $\mathbf{v}$,

$$\begin{aligned}\|\mathbf{v}\|_2 &= \sup_{\|\mathbf{z}\|_2\le 1} \langle \mathbf{z}, \mathbf{v}\rangle = \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j\ge 2} \langle \mathbf{z}_{S_j}, \mathbf{D}_{\mathbf{x}_{S_j}} \mathbf{A}_{S_j}^* \mathbf{A}_{S_1} \mathbf{D}_{\boldsymbol{\epsilon}_{S_1}} \mathbf{x}_{S_1}\rangle \\ &\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j\ge 2} \|\mathbf{z}_{S_j}\|_2 \|\mathbf{D}_{\mathbf{x}_{S_j}} \mathbf{A}_{S_j}^* \mathbf{A}_{S_1} \mathbf{D}_{\boldsymbol{\epsilon}_{S_1}}\|_{2\to 2} \|\mathbf{x}_{S_1}\|_2 \\ &\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j\ge 2} \|\mathbf{A}_{S_j}^* \mathbf{A}_{S_1}\|_{2\to 2} \|\mathbf{z}_{S_j}\|_2 \|\mathbf{x}_{S_j}\|_\infty \|\mathbf{x}_{S_1}\|_2 \ ,\end{aligned}$$

where we have used that $\|\mathbf{D}_{\mathbf{x}}\|_{2\to 2} = \|\mathbf{x}\|_\infty$ and $\|\boldsymbol{\epsilon}\|_\infty = 1$. It follows from Lemma 6.9 and by construction of the partitioning $S_1, S_2, \dots$ that $\|\mathbf{x}_{S_j}\|_\infty \le s^{-1/2}\|\mathbf{x}_{S_{j-1}}\|_2$. Moreover, $\|\mathbf{A}_{S_j}^* \mathbf{A}_{S_1}\|_{2\to 2} \le \delta_{2s}$ for $j \ge 2$ by Proposition 6.3, and $\|\mathbf{x}_{S_1}\|_2 \le \|\mathbf{x}\|_2 \le 1$. We continue our estimation with

$$\begin{aligned}\|\mathbf{v}\|_2 &\le \frac{\delta_{2s}}{\sqrt{s}} \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j\ge 2} \|\mathbf{z}_{S_j}\|_2 \|\mathbf{x}_{S_{j-1}}\|_2 \\ &\le \frac{\delta_{2s}}{\sqrt{s}} \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j\ge 2} \frac{1}{2}\left(\|\mathbf{z}_{S_j}\|_2^2 + \|\mathbf{x}_{S_{j-1}}\|_2^2\right) \le \frac{\delta_{2s}}{\sqrt{s}} \ ,\end{aligned}$$

where we have used that $\sum_j \|\mathbf{x}_{S_j}\|_2^2 = \|\mathbf{x}\|_2^2 = 1$. By Hoeffding's inequality (7.30) and independence of $\mathbf{v}$ and $\boldsymbol{\epsilon}_{\overline{S_1}}$ we have, for $t > 0$,

$$\mathbb{P}(|X| \ge t) \le 2\exp\left(-\frac{t^2 s}{2\delta_{2s}^2}\right) \le 2\exp\left(-\frac{8\, s t^2}{\eta^2}\right) \ . \tag{9.57}$$

Next we consider the third term in (9.56), which can be written as

$$Y := \sum_{\substack{j,\ell\ge 2\\ j\ne\ell}} \langle \mathbf{A}\mathbf{D}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_j}, \mathbf{A}\mathbf{D}_{\boldsymbol{\epsilon}}\mathbf{x}_{S_\ell}\rangle = \sum_{j,\ell\in[N]} \epsilon_j \epsilon_\ell B_{j,\ell} = \boldsymbol{\epsilon}^* \mathbf{B} \boldsymbol{\epsilon} \ ,$$

where $\mathbf{B} \in \mathbb{R}^{N\times N}$ is a symmetric matrix with zero diagonal given entrywise by

$$B_{i,\ell} = \begin{cases} x_i \mathbf{a}_i^* \mathbf{a}_\ell x_\ell & \text{if } i,\ell \in \overline{S_1} \text{ and } i,\ell \text{ are contained in different blocks } S_k \ , \\ 0 & \text{otherwise .} \end{cases}$$

Here, the $\mathbf{a}_j$, $j \in [N]$, denote the columns of $\mathbf{A}$ as usual. We are thus lead to estimating the tail of a Rademacher chaos, which by Proposition 8.13 requires to bound the spectral and Frobenius norm of $\mathbf{B}$. By symmetry the spectral norm can be estimated similarly as above by

$$\begin{aligned}
\|\mathbf{B}\|_{2\to 2} &= \sup_{\|\mathbf{z}\|_2\le 1} \langle \mathbf{B}\mathbf{z}, \mathbf{z}\rangle = \sup_{\|\mathbf{z}\|_2\le 1} \sum_{\substack{j,\ell\ge 2\\ j\neq \ell}} \langle \mathbf{z}_{S_j}, \mathbf{D}_{\mathbf{x}_{S_j}} \mathbf{A}_{S_j}^* \mathbf{A}_{S_\ell} \mathbf{D}_{\mathbf{x}_{S_\ell}} \mathbf{z}_{S_\ell}\rangle \\
&\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{\substack{j,\ell\ge 2\\ j\neq \ell}} \|\mathbf{z}_{S_j}\|_2 \|\mathbf{z}_{S_\ell}\|_2 \|\mathbf{x}_{S_j}\|_\infty \|\mathbf{x}_{S_\ell}\|_\infty \|\mathbf{A}_{S_j}^* \mathbf{A}_{S_\ell}\|_{2\to 2} \\
&\le \delta_{2s} \sup_{\|\mathbf{z}\|_2\le 1} \sum_{\substack{j,\ell\ge 2\\ j\neq \ell}} \|\mathbf{z}_{S_j}\|_2 \|\mathbf{z}_{S_\ell}\|_2 s^{-1/2} \|\mathbf{x}_{S_{j-1}}\|_2 s^{-1/2} \|\mathbf{x}_{S_{\ell-1}}\|_2 \\
&\le \frac{\delta_{2s}}{4s} \sup_{\|\mathbf{z}\|_2\le 1} \sum_{\substack{j,\ell\ge 2\\ j\neq \ell}} \left(\|\mathbf{x}_{S_{j-1}}\|_2^2 + \|\mathbf{z}_{S_j}\|_2^2\right) \left(\|\mathbf{x}_{S_{\ell-1}}\|_2^2 + \|\mathbf{z}_{S_\ell}\|_2^2\right) \\
&\le \delta_{2s}/s \ .
\end{aligned}$$

The Frobenius norm obeys the bound

$$\begin{aligned}
\|\mathbf{B}\|_F^2 &= \sum_{\substack{j,k\ge 2\\ j\neq k}} \sum_{i\in S_j} \sum_{\ell\in S_k} (x_i \mathbf{a}_i^* \mathbf{a}_\ell x_\ell)^2 = \sum_{\substack{j,k\ge 2\\ j\neq k}} \sum_{i\in S_j} x_i^2 \mathbf{a}_i^* \sum_{\ell\in S_k} \mathbf{a}_\ell x_\ell^2 \mathbf{a}_\ell^* \mathbf{a}_i \\
&= \sum_{\substack{j,k\ge 2\\ j\neq k}} \sum_{i\in S_j} x_i^2 \|\mathbf{D}_{\mathbf{x}_{S_k}} \mathbf{A}_{S_k}^* \mathbf{a}_i\|_2^2 \le \sum_{\substack{j,k\ge 2\\ j\neq k}} \sum_{i\in S_j} x_i^2 \|\mathbf{x}_{S_k}\|_\infty^2 \|\mathbf{A}_{S_k}^* \mathbf{a}_i\|_2^2 \\
&\le \delta_{2s}^2 \sum_{\substack{j,k\ge 2\\ j\neq k}} \|\mathbf{x}_{S_j}\|_2^2 s^{-1} \|\mathbf{x}_{S_k}\|_2^2 \le \frac{\delta_{2s}^2}{s} \ .
\end{aligned}$$

Hereby, we have used that $\|\mathbf{A}_{S_k}^* \mathbf{a}_i\|_2 = \|\mathbf{A}_{S_k}^* \mathbf{a}_i\|_{2\to 2} \le \delta_{s+1} \le \delta_{2s}$ by Proposition 6.5. It follows from Proposition 8.13 that the tail of the third term (9.56) can be estimated by

$$\begin{aligned}
\mathbb{P}(|Y| \ge r) &\le 2\exp\left(-\min\left\{\frac{3\,r^2}{128\,\|\mathbf{B}\|_F^2}, \frac{r}{32\,\|\mathbf{B}\|_{2\to 2}}\right\}\right) \\
&\le 2\exp\left(-\min\left\{\frac{3\,sr^2}{128\,\delta_{2s}^2}, \frac{sr}{32\,\delta_{2s}}\right\}\right) \ . \\
&\le 2\exp\left(-s\min\left\{\frac{3\,r^2}{8\,\eta^2}, \frac{t}{8\,\eta}\right\}\right) \ .
\end{aligned}$$

Now we choose $t = \eta/8$ and $r = \eta/2$. Then plugging into the previous estimate and into (9.57), and combining with (9.56) shows that

$$(1-\eta)\|\mathbf{x}\|_2^2 \le \|\mathbf{A}\mathbf{D}_{\boldsymbol{\epsilon}}\mathbf{x}\|_2^2 \le (1+\eta)\|\mathbf{x}\|_2^2 \tag{9.58}$$

for a single $\mathbf{x} \in E$ with probability at least

$$1 - 2\exp(-s/8) - 2\exp(-s\min\{3/32, 1/16\}) \ge 1 - 4\exp(-s/16) \ .$$

Taking the union bound over all $\mathbf{x} \in E$ shows that (9.58) holds for all $\mathbf{x} \in E$ simultaneously with probability at least

$$1 - 4M \exp(-s/16) \geq 1 - \varepsilon \,,$$

under the condition $s \geq 16 \ln(4M/\varepsilon)$. This concludes the proof. □

Notes

Section 9.1 follows the general idea of the paper [302] by S. Mendelson, A. Pajor and N. Tomczak-Jaegermann, and independently developed in [24] by R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. There, however, the restricted isometry property is considered without squares on the ℓ_2-norms. As a result, the proof given here is slightly different. Similar techniques were also used in extensions [358], including the D-RIP [71] covered in Exercise 9.11, and the corresponding notion of the restricted isometry property in low-rank matrix recovery [75, 362], see Exercise 9.12. E. Candés and T. Tao have been the first to show the restricted isometry property for Gaussian matrices in [82]. They essentially followed the approach given in Section 9.3. They relied on the condition number estimate for Gaussian random matrices of Theorem 9.24. The proof method of the latter based on Slepian's and Gordon's lemma as well as on concentration of measure follows [118].

The nonuniform recovery result of Theorem 9.14 has been shown in [17], see also [77] for a very similar approach. The accurate estimate of the required number of samples in the Gaussian case, Theorem 9.16, appeared in slightly different form in [92], where also far reaching extension to other situations such as low rank matrix recovery are treated. The estimate of the null space property for Gaussian random matrices, Theorem 9.27, has not appeared elsewhere in this form. Similar ideas, however, were used in [374, 393].

The escape through the mesh theorem 9.19 is essentially due to Gordon [202], where it appeared with slightly worse constants. It was used first in compressed sensing by M. Rudelson and R. Vershynin in [374], see also [393, 92].

The Johnson-Lindenstrauss-Lemma appeared in [256] for the first time. A different proof was given in [113]. Theorem 9.34 on the relation of the restricted isometry property to the Johnson-Lindenstrauss lemma was shown by F. Krahmer and R. Ward in [268].

Random matrices were initially introduced in the context of mathematical physics by E. Wigner. There is a large body of literature on the asymptotic analysis of the spectrum of random matrices when the matrix dimension tends to infinity. A well-known result in this context states that the empirical distribution of Wigner random matrices (Hermitian random matrices with independent entries up to symmetries) converges to the famous semi-circle law. We refer to the monographs [10, 21] for further information on asymptotic random matrix theory.

The methods employed in this chapter fall into the area of nonasymptotic random matrix theory [438, 376], which considers spectral properties of random matrices in fixed (but usually large) dimension. M. Rudelson and R. Vershynin [375] exploited methods developed in compressive sensing (among other techniques) and established an open conjecture on the smallest singular value of square Bernoulli random matrices. By distinguishing the action of the matrix on compressible and incompressible vectors they were able to achieve their breakthrough. The action on compressible vectors is handled in the same way as the restricted isometry property is shown for rectangular random matrices in Section 9.1.

Sparse recovery with Gaussian matrices via polytope geometry. D. Donoho and J. Tanner [144, 143, 132, 145] approach the analysis of sparse recovery via ℓ_1-minimization using Gaussian random matrices via the geometric characterization of Corollary 4.39. They consider an asymptotic scenario where the dimension N tends to infinity, and $m = m_N$ and $s = s_N$ are such that

$$\lim_{N\to\infty} \frac{m_N}{N} = \delta \qquad \text{and} \qquad \lim_{N\to\infty} \frac{s_N}{m_N} = \rho$$

for some $\delta, \rho \in [0,1]$. They show that there exist thresholds that separate regions in the plane $[0,1]^2$ of parameters (δ, ρ), where recovery succeeds and recovery fails with probability tending to 1 as $N \to \infty$. In other words, a phase transition phenomenon is happening for high dimensions N. They distinguish a strong threshold $\rho_S = \rho_S(\delta)$, and a weak threshold $\rho_W = \rho_W(\delta)$.

In our terminology, the strong threshold corresponds to uniform recovery via ℓ_1-minimization. In the limit as $N \to \infty$, if $\rho < \rho_S(\delta)$ with $\delta = m/N$ then $s < \rho m$ implies recovery of all s-sparse vectors with high probability. Moreover, if $\rho > \rho_S(\delta)$ and $s < \rho m$ then recovery of all s-sparse recovery fails with high probability.

The weak threshold corresponds to nonuniform recovery. (The formulation in [144, 143, 132, 145] is slightly different than our notion of nonuniform recovery, but for Gaussian random matrices both notions are equivalent.) In the limit as $N \to \infty$, if $\rho < \rho_W(\delta)$ with $\delta = m/N$ then $s < \rho m$ implies that a fixed s-sparse vector is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization with high probability using a draw of an $m \times N$ Gaussian random matrix $\mathbf{A}$. Conversely, if $\rho > \rho_W(\delta)$ and $s > \rho m$ then ℓ_1-minimization fails to recover a given s-sparse vector from $\mathbf{y} = \mathbf{A}\mathbf{x}$ with high probability.

Unfortunately, no closed forms for the functions ρ_W and ρ_S are available. Nevertheless, D. Donoho and J. Tanner [132, 145] provide complicated implicit expressions and compute these functions numerically, see Figure 9.1. Moreover, they derive the asymptotic behavior of $\rho_W(\delta)$, $\rho_S(\delta)$ when $\delta \to 0$, that is, in the relevant scenario when m is significantly smaller than N:

$$\rho_S(\delta) \sim \frac{1}{2e\ln((\sqrt{\pi}\delta)^{-1})} \qquad \text{and} \qquad \rho_W(\delta) \sim \frac{1}{2\ln(\delta^{-1})} \qquad \delta \to 0\,.$$

As consequence we roughly obtain the following statements for large N:

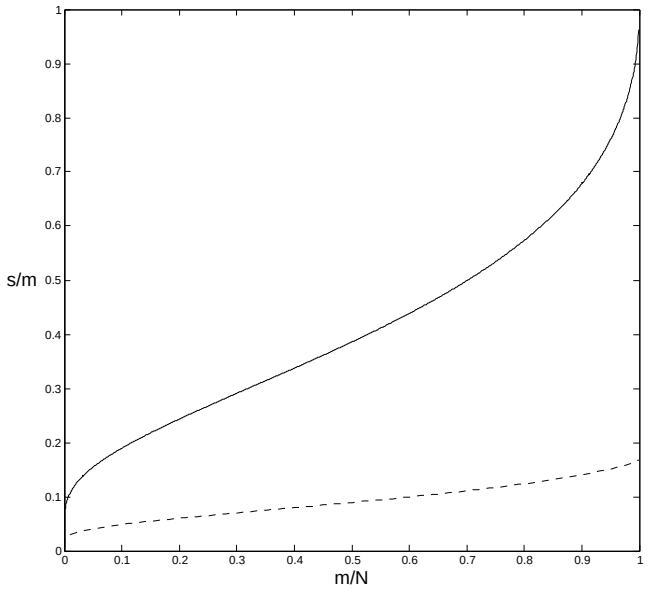

Fig. 9.1. Strong threshold $\rho_S = \rho_S(\delta)$ (dashed curve - -), weak threshold $\rho_W(\delta)$ (solid curve –), $\delta = m/N$, $\rho = s/m$.

- **Uniform recovery.** If

$$m > 2es \ln(N/(\sqrt{\pi}m))$$

  then with high probability on the draw of a Gaussian random matrix, every s-sparse vector $\mathbf{x}$ is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization. Conversely, if $m < 2es \ln(N/(\sqrt{\pi}m))$ then recovery of all s-sparse vectors $\mathbf{x}$ fails with high probability.
- **Nonuniform recovery.** If

$$m > 2s \ln(N/m)$$

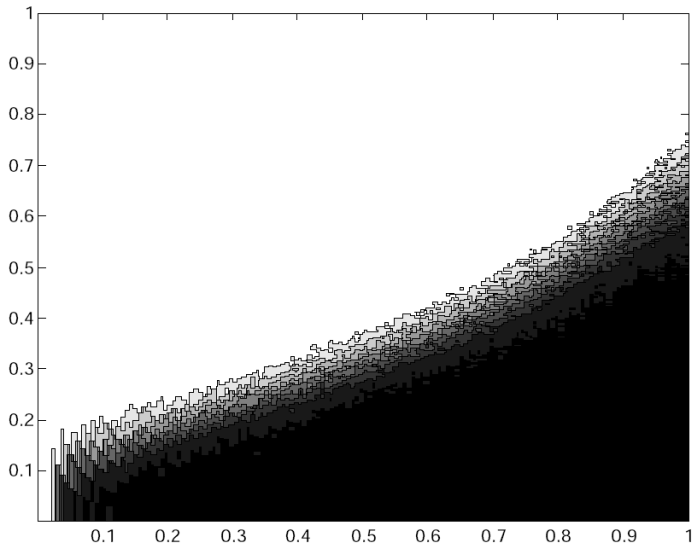

Fig. 9.2. Weak threshold observed empirically.

then a fixed s-sparse vector $\mathbf{x}$ is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization with high probability on the draw of a Gaussian random matrix $\mathbf{A}$. Conversely, if $m < 2s\ln(N/m)$ then with high probability ℓ_1-minimization fails to recover a fixed s-sparse vector $\mathbf{x}$ using a random draw of a Gaussian matrix.

The rather involved analysis of D. Donoho and J. Tanner builds on the characterization of sparse recovery in Corollary 4.39. Stated in slightly different notation, s-sparse recovery is equivalent to s-neighborliness of the projected polytope $\mathbf{A}B_1^N$: every set of s vertices of $\mathbf{A}B_1^N$ (not containing antipodal points) spans an $s-1$-face of $\mathbf{A}B_1^N$, see also Exercise 4.15. This property is investigated directly using work by F. Affentranger and R. Schneider [3], and by A. Vershik and P. Sporyshev [435] on random polytopes. Additionally, Donoho and Tanner provide thresholds for the case that it is known a priori that the sparse vector has only nonnegative entries [145, 143]. This information can be used as an additional constraint in the ℓ_1-minimization problem and in this case one has to analyze the projected simplex $\mathbf{A}S^N$, where S^N is the standard simplex, that is, the convex hull of the canonical unit vectors and the zero vector.

It is presently unclear whether this approach can be extended to other types of random matrices than Gaussian, for instance, Bernoulli matrices, although the same weak threshold is observed empirically for a variety of random matrices [146]. For illustration, an empirical phase diagram is shown

in Figure 9.2. The polytope approach does not seem to cover stability and robustness of reconstruction.

The fact that this analysis also provides precise statements about the failure of recovery via ℓ_1-minimization, allows to deduce that the constant 2 in our nonuniform recovery analysis for Gaussian random matrices in Section 9.2 is optimal, see (9.23) and (9.37). Moreover, the constant 8 appearing in our analysis of the null space property in Theorem 9.27, see also (9.48), is not optimal but at least not too far from the optimal value $2e$. In contrast to the polytope approach however, Theorem 9.27 covers also the stability of reconstruction.

A similar precise phase transition analysis of the restricted isometry constants of Gaussian random matrices has been performed in [20, 43].

Message passing algorithms [128] in connection with Gaussian random matrices also allow a precise asymptotic analysis.

Exercises

9.1. Let $q \in (0,1)$ and let $\mathcal{S}$ be the unit sphere of $\mathbb{R}^n$ with respect to the ℓ_q-quasinorm. Prove that, for each $\rho > 0$, there exists a subset $\mathcal{U}$ of $\mathcal{S}$ such that

$$\min_{\mathbf{u}\in\mathcal{U}} \|\mathbf{z}-\mathbf{u}\|_q^q \le \rho \qquad \text{for all } \mathbf{z} \in \mathcal{S}$$

and

$$\operatorname{card}(\mathcal{U}) \le \left(1+\frac{2}{\rho}\right)^{n/q}.$$

9.2. Coherence of a Bernoulli random matrix.
Let $\mathbf{A} = (\mathbf{a}_1|\mathbf{a}_2|\cdots|\mathbf{a}_N)$ be an $m \times N$ Bernoulli matrix. Let μ be the coherence of $m^{-1/2}\mathbf{A}$, i.e., $\mu = m^{-1}\max_{j\neq k}|\langle \mathbf{a}_j, \mathbf{a}_k\rangle|$. Show that

$$\mu \le 2\sqrt{\frac{\ln(N/\varepsilon)}{m}}$$

with probability at least $1-\varepsilon^2$.

9.3. Concentration inequality for Gaussian matrices. Let $\mathbf{A}$ be an $m \times N$ standard Gaussian random matrix. Show that, for $\mathbf{x} \in \mathbb{R}^N$ and $t \in (0,1)$,

$$\mathbb{P}\big(\,\big|m^{-1}\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \ge t\|\mathbf{x}\|_2^2\big) \le 2\exp\big(-m(t^2/4 - t^3/6)\big)\,. \tag{9.59}$$

Show that this concentration inequality holds as well for a Bernoulli random matrix.

9.4. Smallest singular value of a subgaussian matrix.
Let $\mathbf{B}$ be an $m \times s$ subgaussian random matrix, and let $\sigma_{\min}$ be the smallest singular value of $\frac{1}{\sqrt{m}}\mathbf{B}$. Show that, for $t \in (0,1)$,

$$\mathbb{P}\left(\sigma_{\min} \le 1 - c_1\sqrt{\frac{s}{m}} - t\right) \le 2\exp(-c_2 m t^2)\ .$$

Provide values for the constants $c_1, c_2 > 0$, possibly in terms of $\tilde{c}$ in (9.6).

9.5. Extremal singular values of complex Gaussian matrices.

9.6. Nonuniform recovery for Gaussian matrices.
Let $\mathbf{x} \in \mathbb{C}^N$ be an s-sparse vector. Let $\mathbf{A}$ be an $m \times N$ Gaussian random matrix. Show that if, for some $\varepsilon \in (0,1)$

$$m \ge s\left[\sqrt{2\ln(2(N-s)/\varepsilon)} + 1 + \sqrt{2\ln(2/\varepsilon)/s}\right]^2$$

then with probability at least $1-\varepsilon$ the vector $\mathbf{x}$ is the unique solution to the ℓ_1-minimization problem $\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$.

9.7. Suppose that $\mathbf{A}$ is an $m \times N$ Gaussian random matrix. For $0 < \delta < 1$, prove that the mixed-norm restricted isometry property

$$(1-\delta)\sqrt{m}\|\mathbf{x}\|_2 \le \|\mathbf{Ax}\|_1 \le (1+\delta)\sqrt{m}\|\mathbf{x}\|_2 \qquad \text{for all } s\text{-sparse } \mathbf{x} \in \mathbb{R}^N$$

is fulfilled with high probability, provided that

$$m \ge \mathrm{c}(\delta)\, s\, \ln(eN/s).$$

9.8. Verify Corollary 9.32 in detail.

9.9. Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a random matrix satisfying the concentration inequality (9.7). Given $\delta > 0$, prove that the matrix $\mathbf{A}$ satisfies the *homogeneous restricted isometry property*

$$\left(1-\sqrt{\frac{r}{s}}\delta\right)\|\mathbf{x}\|_2^2 \le \|\mathbf{Ax}\|_2^2 \le \left(1+\sqrt{\frac{r}{s}}\delta\right)\|\mathbf{x}\|_2^2 \quad \text{for all } r\text{-sparse } \mathbf{x} \in \mathbb{C}^N,\ r \le s,$$

with probability at least $1 - N^{-c_1}$ provided $m \ge c_2\delta^{-2}s\ln(N)$.

9.10. Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a random matrix, for which all columns are independent and uniformly distributed on the sphere S^{m-1}. Show that its restricted isometry constant satisfies $\delta_s \le \delta$ with probability at least $1-\varepsilon$ provided

$$m \ge C\delta^{-2}\left(s\ln(eN/s) + \ln(2\varepsilon^{-1})\right)\ ,$$

where $C > 0$ is an appropriate universal constant.

9.11. D-RIP
Let $\mathbf{D} \in \mathbb{R}^{N\times M}$ (the dictionary) with $M \geq N$ and $\mathbf{A} \in \mathbb{R}^{m\times N}$ (the measurement matrix). The restricted isometry constants δ_s adapted to $\mathbf{D}$ are defined to be the smallest constants such that

$$(1-\delta_s)\|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1+\delta_s)\|\mathbf{z}\|_2^2$$

for all $\mathbf{z} \in \mathbb{R}^N$ of the form $\mathbf{z} = \mathbf{D}\mathbf{x}$ for some s-sparse $\mathbf{x} \in \mathbb{R}^M$. (This notion appears in the recovery of vectors that are sparse with respect to an overcomplete $\mathbf{D}$.)

Let $\mathbf{A}$ be an $m \times N$ subgaussian random matrix. Show that the restricted isometry constants adapted to $\mathbf{D}$ of $m^{-1/2}\mathbf{A}$ satisfy $\delta_s \leq \delta$ with probability at least $1-\varepsilon$ provided that

$$m \geq C\delta^{-2}\left(s\ln(M/s) + \ln(2\varepsilon^{-1})\right) .$$

9.12. Rank-RIP for subgaussian measurement maps.
For a measurement map $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ the rank-restricted isometry constant δ_s is defined as the smallest number such that

$$(1-\delta_s)\|\mathbf{X}\|_F^2 \leq \|\mathcal{A}(\mathbf{X})\|_2^2 \leq (1+\delta_s)\|\mathbf{X}\|_F^2 \quad \text{for all } \mathbf{X} \text{ of rank at most } s .$$

A measurement map $\mathcal{A}$ is called subgaussian if all the entries $\mathcal{A}_{jk\ell}$ in the representation

$$\mathbf{A}(\mathbf{X})_j = \sum_{k,\ell} \mathcal{A}_{jk\ell} X_{k\ell}$$

are independent mean-zero subgaussian random variables of variance 1 with the same subgaussian parameter c. Show that the restricted isometry constants of $1/\sqrt{m}\mathcal{A}$ satisfy $\delta_s \leq \delta$ with probability at least $1-\varepsilon$ provided that

$$m \geq C\delta^{-2}\left(s(n_1+n_2) + \ln(2\varepsilon^{-1})\right) .$$

(Why is this bound optimal?)

As a first step show that the covering numbers of the set $D_s = \{\mathbf{X} \in \mathbb{R}^{n_1\times n_2} : \|\mathbf{X}\|_F \leq 1, \mathrm{rank}(\mathbf{X}) \leq s\}$ satisfy

$$\mathcal{N}(D_s, \|\cdot\|_F, \rho) \leq (1+6/\rho)^{(n_1+n_2+1)s} .$$

Hint: Use the (reduced) singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^*$, where $\mathbf{U} \in \mathbb{R}^{n_1\times s}, \mathbf{V} \in \mathbb{R}^{n_2\times s}$ have orthonormal columns and $\mathbf{D} \in \mathbb{R}^{s\times s}$ is diagonal. Cover the sets of the three components $\mathbf{U}, \mathbf{V}, \mathbf{D}$ separately with respect to suitable norms.

9.13. Largest singular value via Dudley's inequality.
Let $\mathbf{A}$ be an $m \times s$ (unnormalized) subgaussian random matrix. Use Dudley's inequality, Theorem 8.23, to show that

$$\mathbb{E}\|\mathbf{A}\|_{2\to 2} \leq C(\sqrt{m} + \sqrt{s}) .$$

10

Gelfand Widths of ℓ_1-Balls

In this chapter, we make a detour via the geometry of ℓ_1^N in order to underline the optimality of random sensing in terms of the number of measurements. In Section 10.1, we introduce several notions of widths, and show that Gelfand widths are closely to the worst case reconstruction error of compressive sensing methods over classes of vectors. In Section 10.2, we establish upper and lower bounds for the Gelfand widths of ℓ_1-balls. In fact, methods from compressive sensing turn out to be appropriate tools to tackle this venerable problem originating from pure mathematics. We give further instances of methods from compressive sensing being used successfully in Banach space geometry in Section 10.3, where we establish lower and upper bounds of certain Kolmogorov widths as well as Kashin's decomposition theorem. Although this is not mandatory, we only consider vector spaces over the field of real numbers in this chapter.

10.1 Definitions and Relation to Compressive Sensing

We introduce in this section several notions of widths. We start with the classical notion of Gelfand widths.

Definition 10.1. *The Gelfand m-width of a subset K of a normed space X is defined as*

$$d^m(K,X) := \inf\left\{ \sup_{\mathbf{x}\in K\cap L^m} \|\mathbf{x}\|, \;\; L^m \text{ subspace of } X \text{ with } \operatorname{codim}(L^m) \leq m \right\}.$$

Since a subspace L^m of X is of codimension at most m if and only if there exists linear functionals $\lambda_1, \ldots, \lambda_m \in X^*$ such that

$$L^m = \{\mathbf{x} \in X : \lambda_i(\mathbf{x}) = 0 \text{ for all } i \in [m]\} = \ker \mathbf{A},$$

where $\mathbf{A} : X \to \mathbb{R}^m$, $\mathbf{x} \mapsto [\lambda_1(\mathbf{x}), \ldots, \lambda_m(\mathbf{x})]^\top$, we also have the representation

$$d^m(K,X) = \inf\left\{ \sup_{\mathbf{x}\in K\cap \ker \mathbf{A}} \|\mathbf{x}\|, \;\; \mathbf{A} : X \to \mathbb{R}^m \text{ linear} \right\}.$$

We readily observe that the sequence $(d^m(K,X))_{m\geq 0}$ is nonincreasing. Its first term is $d^0(K,X) = \sup_{\mathbf{x}\in K} \|\mathbf{x}\|$. If $N := \dim(X)$ is finite, then $d^m(K,X) = 0$ for all $m \geq N$, provided that C contains the zero vector. If otherwise $\dim(X)$ is infinite, then $\lim_{m\to\infty} d^m(K,X) = 0$ as soon as the set C is compact — see Exercise 10.2.

We now highlight the pivotal role of Gelfand widths in compressive sensing. To do so, we show that they are comparable to quantities that measure the worst-case reconstruction errors of optimal measurement/reconstruction schemes. We call the first of these quantities the (nonadaptive) compressive widths. Here comes their precise definition.

Definition 10.2. *The compressive m-width of a subset K of a normed space X is defined as*

$$E^m(K,X) := \inf\left\{ \sup_{\mathbf{x}\in K} \|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|, \;\; \mathbf{A} : X \to \mathbb{R}^m \text{ linear}, \Delta : \mathbb{R}^m \to X \right\}.$$

In this definition, the measurement scheme associated to the linear map $\mathbf{A}$ is *nonadaptive*, in the sense that the m linear functionals $\lambda_1, \ldots, \lambda_m \in X^*$ given by $\mathbf{A}\mathbf{x} = [\lambda_1(\mathbf{x}), \ldots, \lambda_m(\mathbf{x})]^\top$ are chosen once and for all. In contrast, we may also consider the *adaptive* setting, where the choice of a measurement depends on the result of previous measurements according to a specific rule. In this way, the measurement scheme is represented by the *adaptive* map $F : X \to \mathbb{R}^m$ defined by

$$F(\mathbf{x}) = \begin{bmatrix} \lambda_1(\mathbf{x}) \\ \lambda_{2;\lambda_1(\mathbf{x})}(\mathbf{x}) \\ \vdots \\ \lambda_{m;\lambda_1(\mathbf{x}),\ldots,\lambda_{m-1}(\mathbf{x})}(\mathbf{x}) \end{bmatrix}, \tag{10.1}$$

where the functionals $\lambda_1, \lambda_{2;\lambda_1(\mathbf{x})}, \ldots, \lambda_{m;\lambda_1(\mathbf{x}),\ldots,\lambda_{m-1}(\mathbf{x})}$ are all linear. This leads to the introduction of the adaptive compressive width.

Definition 10.3. *The adaptive compressive m-width of a subset K of a normed space X is defined as*

$$E^m_{\mathrm{ada}}(K,X) := \inf\left\{ \sup_{\mathbf{x}\in K} \|\mathbf{x} - \Delta(F(\mathbf{x}))\|, \;\; F : X \to \mathbb{R}^m \text{ adaptive}, \Delta : \mathbb{R}^m \to X \right\}.$$

The intuitive expectation that adaptivity improves the performance of the measurement/reconstruction scheme is invalid, at least when considering worst cases over K. The following theorem indeed shows that, under some mild conditions, the nonadaptive and the adaptive compressive sensing widths are comparable, and that they are both comparable to the Gelfand width.

Theorem 10.4. *If K is a subset of a normed space X, then*

$$E^m_{\mathrm{ada}}(K,X) \le E^m(K,X).$$

If the subset K satisfies $-K = K$, then

$$d^m(K,X) \le E^m_{\mathrm{ada}}(K,X).$$

If the set K further satisfies $K + K \subseteq a\,K$ for some positive constant a, then

$$E^m(K,X) \le a\,d^m(K,X).$$

Proof. The first inequality is straightforward, because any linear measurement map $\mathbf{A} : X \to \mathbb{R}^m$ can be considered adaptive.

Let us now assume that the set K satisfies $-K = K$. We consider an adaptive map $F : X \to \mathbb{R}^m$ of the form (10.1) and a reconstruction map $\Delta : \mathbb{R}^m \to X$. We define the linear map $\mathbf{A} : X \to \mathbb{R}^m$ by $\mathbf{A}(\mathbf{x}) = [\lambda_1(\mathbf{x}), \lambda_{2;0}(\mathbf{x}), \dots, \lambda_{m;0,\dots,0}(\mathbf{x})]^\top$ and we set $L^m := \ker \mathbf{A}$. Since this is a subspace of X satisfting $\operatorname{codim}(L^m) \le m$, the definition of Gelfand width implies

$$d^m(K,X) \le \sup_{\mathbf{v} \in K \cap \ker \mathbf{A}} \|\mathbf{v}\|. \tag{10.2}$$

We notice that, for $\mathbf{v} \in \ker \mathbf{A}$, we have $\lambda_1(\mathbf{v}) = 0$, then $\lambda_{2;\lambda_1(\mathbf{v})}(\mathbf{v}) = \lambda_{2;0}(\mathbf{v})$, and so on until $\lambda_{m;\lambda_1(\mathbf{v}),\dots,\lambda_{m-1}(\mathbf{v})}(\mathbf{v}) = \lambda_{m;0,\dots,0}(\mathbf{v}) = 0$, so that $F(\mathbf{v}) = 0$. Thus, for any $\mathbf{v} \in K \cap \ker \mathbf{A}$, we observe that

$$\|\mathbf{v} - \Delta(0)\| = \|\mathbf{v} - \Delta(F(\mathbf{v}))\| \le \sup_{\mathbf{x} \in K} \|\mathbf{x} - \Delta(F(\mathbf{x}))\|,$$

and likewise, since $-\mathbf{v} \in K \cap \ker \mathbf{A}$, that

$$\|-\mathbf{v} - \Delta(0)\| = \|-\mathbf{v} - \Delta(F(-\mathbf{v}))\| \le \sup_{\mathbf{x} \in K} \|\mathbf{x} - \Delta(F(\mathbf{x}))\|.$$

We derive that, for any $\mathbf{v} \in K \cap \ker \mathbf{A}$,

$$\begin{aligned} \|\mathbf{v}\| &= \Big\| \frac{1}{2}(\mathbf{v} - \Delta(0)) - \frac{1}{2}(-\mathbf{v} - \Delta(0)) \Big\| \le \frac{1}{2}\|\mathbf{v} - \Delta(0)\| + \frac{1}{2}\|-\mathbf{v} - \Delta(0)\| \\ &\le \sup_{\mathbf{x} \in K} \|\mathbf{x} - \Delta(F(\mathbf{x}))\|. \end{aligned} \tag{10.3}$$

According to (10.2) and (10.3), we have

$$d^m(K,X) \le \sup_{\mathbf{x} \in K} \|\mathbf{x} - \Delta(F(\mathbf{x}))\|.$$

The inequality $d^m(K,X) \le E^m_{\mathrm{ada}}(K,X)$ follows by taking the infimum over all possible F and Δ.

Let us finally also assume that $K + K \subseteq a\,K$ for some positive constant a. We consider a subspace L^m of the space X with $\operatorname{codim}(L^m) \le m$. We choose

a linear map $\mathbf{A} : X \to \mathbb{R}^m$ such that $\ker \mathbf{A} = L^m$, and we define a map $\Delta : \mathbb{R}^m \to X$ in such a way that

$$\Delta(\mathbf{y}) \in K \cap \mathbf{A}^{-1}(\mathbf{y}) \qquad \text{for all } \mathbf{y} \in \mathbf{A}(K).$$

We then deduce that

$$E^m(K,X) \le \sup_{\mathbf{x}\in K} \|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\| \le \sup_{\mathbf{x}\in K} \Big[\sup_{\mathbf{z}\in K\cap \mathbf{A}^{-1}(\mathbf{A}\mathbf{x})} \|\mathbf{x}-\mathbf{z}\| \Big].$$

For $\mathbf{x} \in K$ and $\mathbf{z} \in K \cap \mathbf{A}^{-1}(\mathbf{A}\mathbf{x})$, we observe that the vector $\mathbf{x} - \mathbf{z}$ belongs to $K + (-K) \subseteq a\,K$ and to $\ker \mathbf{A} = L^m$ as well. Therefore, we obtain

$$E^m(K,X) \le \sup_{\mathbf{u}\in aK\cap L^m} \|\mathbf{u}\| = a \sup_{\mathbf{v}\in K\cap L^m} \|\mathbf{v}\|.$$

Taking the infimum over L^m, we conclude that $E^m(K,X) \le a\, d^m(K,X)$. □

In the next section, we give matching upper and lower bounds for the Gelfand width $d^m(B_1^N, \ell_p^N)$ of ℓ_1-balls in ℓ_p^N when $1 < p \le 2$, see Propositions 10.9 and 10.10. They provide the following result.

Theorem 10.5. *For $1 < p \le 2$ and $m < N$ there exist absolute constants c_1, c_2 depending only on p such that*

$$c_1 \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p} \le d^m(B_1^N, \ell_p^N) \le c_2 \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

We immediately obtain corresponding estimates for the compressive widths, where we recall that $A \asymp B$ means that there exist absolute constants c_1, c_2 such that $c_1 A \le B \le c_2 A$.

Corollary 10.6. *For $1 < p \le 2$ and $m < N$, the nonadaptive and adaptive compressive widths satisfy*

$$E_{\mathrm{ada}}^m(B_1^N, \ell_p^N) \asymp E^m(B_1^N, \ell_p^N) \asymp \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

Proof. Since $-B_1^N = B_1^N$ and $B_1^N + B_1^N \subseteq 2B_1^N$, Theorem 10.4 implies

$$d^m(B_1^N, \ell_p^N) \le E_{\mathrm{ada}}^m(B_1^N, \ell_p^N) \le E^m(B_1^N, \ell_p^N) \le 2\, d^m(B_1^N, \ell_p^N).$$

Theorem 10.5 therefore concludes the proof. □

The lower estimate is of particular significance in compressive sensing. Indeed, under the condition

$$m \ge c\, s \ln\left(\frac{eN}{s}\right), \tag{10.4}$$

we have seen that there are matrices $\mathbf{A} \in \mathbb{R}^{m \times N}$ with small restricted isometry constants and reconstruction maps providing the stability estimate

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_p \leq \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1 \qquad \text{for all } \mathbf{x} \in \mathbb{R}^N.$$

Such reconstruction maps include, for instance, basis pursuit, iterative hard thresholding, or orthogonal matching pursuit, see Chapter 6. Conversely, we can now show that the existence of Δ and $\mathbf{A}$ — or Δ and an adaptive F — providing such a stability estimate forces the number of measurements to be bounded from below as in (10.4).

Proposition 10.7. *For $1 < p \leq 2$, suppose that there exist $\mathbf{A} \in \mathbb{R}^{N \times m}$ and a map $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ such that, for all $\mathbf{x} \in \mathbb{R}^N$,*

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_p \leq \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1. \tag{10.5}$$

Then, for some constant $c_1, c_2 > 0$ depending only on C,

$$m \geq c_1 \, s \ln\left(\frac{eN}{s}\right),$$

provided $s > c_2$.
The same statement holds true for an adaptive map $F : \mathbb{R}^N \to \mathbb{R}^m$ in place of a linear map $\mathbf{A}$.

Proof. It is enough to prove the statement for an adaptive map $F : \mathbb{R}^N \to \mathbb{R}^m$. We notice that (10.5) implies

$$E^m_{\mathrm{ada}}(B_1^N, \ell_p^N) \leq \frac{C}{s^{1-1/p}} \sup_{\mathbf{x} \in B_1^N} \sigma_s(\mathbf{x})_1 \leq \frac{C}{s^{1-1/p}}.$$

But, in view of Theorem 10.6, there is a constant $c > 0$ such that

$$c \, \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p} \leq E^m_{\mathrm{ada}}(B_1^N, \ell_p^N).$$

Thus, for some constant $c' > 0$,

$$c' \, \min\left\{1, \frac{\ln(eN/m)}{m}\right\} \leq \frac{1}{s}.$$

We derive either $s \leq 1/c'$ or $m \geq c' s \ln(eN/m)$. The hypothesis $s > c_2 := 1/c'$ allows to discard the first alternative. Calling upon Lemma C.6, the second alternative gives $m \geq c_1 s \ln(eN/m)$ with $c_1 = c'e/(1+e)$. This is the desired result. □

The restrictions $s > c_2$ and $p > 1$ will be removed in the nonadaptive setting by Theorem 11.7. Accepting that this theorem is true for now, we can state the following result on the minimal number of measurement needed to enforce the restricted isometry property.

Corollary 10.8. *A matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $2s$th restricted isometry constant $\delta_{2s} < 1/3$, say, must have a number of rows bounded below by*

$$m \geq c\, s \ln\left(\frac{eN}{s}\right)$$

for some constant $c > 0$ depending only on δ_{2s}.

Proof. If $\delta_{2s} < 1/3$ and if Δ is the ℓ_1-minimization reconstruction map, we know from Theorem 6.11 that (10.5) with $p = 2$ holds for some constant C depending only on δ_{2s}. The previous argument yields the result. □

10.2 Estimate for the Gelfand Widths of ℓ_1-Balls

In this section, we establish the two-sided estimate of Theorem 10.6 for the Gelfand widths of the unit ℓ_1-balls in ℓ_p^N when $1 \leq p \leq 2$. We separate the lower and upper estimates.

Upper Bound

With the results of compressive sensing that we have already established it is rather simple to bound the Gelfand widths from above. For instance, recovery theorems such as Theorems 6.11, 6.20, and 6.27, applied to matrices with the restricted isometry property imply that

$$E^m(B_1^N, \ell_p^N) \leq \frac{C}{s^{1-1/p}} \sup_{\mathbf{x} \in B_1^N} \sigma_s(\mathbf{x})_1 \leq \frac{C}{s^{1-1/p}}$$

when m is of the order of $s \ln(eN/s)$, or equivalently (see Lemma C.6), of the order of $m/\ln(eN/m)$. Then, using Theorem 10.4, we get

$$d^m(B_1^N, \ell_p^N) \leq E^m(B_1^N, \ell_p^N) \leq C' \left\{\frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

A more rigorous and self-contained argument (not relying on any recovery theorems) is given below. It is strongly inspired by the ideas of compressive sensing.

Proposition 10.9. *There is a constant $C > 0$ such that, for $1 < p \leq 2$ and $m < N$,*

$$d^m(B_1^N, \ell_p^N) \leq C \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

Proof. Using the inequality $\|\mathbf{x}\|_p \le \|\mathbf{x}\|_1$, $\mathbf{x} \in \mathbb{R}^N$, in the definition of Gelfand width immediately gives

$$d^m(B_1^N, \ell_p^N) \le 1.$$

As a result, if $m \le c \ln(eN/m)$ with $c := 144(1 + e^{-1})$, then

$$d^m(B_1^N, \ell_p^N) \le \min\left\{1, \frac{c \ln(eN/m)}{m}\right\}^{1-1/p}. \tag{10.6}$$

On the other hand, if $m > c \ln(eN/m)$, we define $s \ge 1$ to be the largest integer smaller than $m/(c \ln(eN/m))$, so that

$$\frac{m}{2c \ln(eN/m)} \le s < \frac{m}{c \ln(eN/m)}.$$

Note that $m > c\, s \ln(eN/m)$ yields $m > c' s \ln(eN/s)$ with $c' = 144$, see Lemma C.6. Then Theorem 9.25 with $\eta = 1/6$ and $\epsilon = 2\exp(-m/144)$ guarantees the existence of a measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ with restricted isometry constant

$$\delta_s(\mathbf{A}) \le \delta := 4\eta + 4\eta^2 = 1/9,$$

since $m \ge 72(s \ln(eN/s) - m/144)$, i.e., $m \ge 144 s \ln(eN/s)$. (Instead of Theorem 9.25, we could alternatively use the easier Theorem 9.2 or Theorem 9.10 on the restricted isometry property of subgaussian random matrices, which however does not specify the constants.)

Partitioning the index set $[N]$ as the union $S_0 \cup S_1 \cup S_2 \cup \dots$ of index sets of size s in such a way that $|x_i| \ge |x_j|$ whenever $i \in S_{k-1}$, $j \in S_k$, and $k \ge 1$, we recall from Lemma 6.9 that $\|\mathbf{x}_{S_k}\|_2 \le \|\mathbf{x}_{S_{k-1}}\|_1/\sqrt{s}$ for all $k \ge 1$. Therefore, for $\mathbf{x} \in L^m := \ker \mathbf{A}$, we have

$$\begin{aligned}
\|\mathbf{x}\|_p &\le \sum_{k \ge 0} \|\mathbf{x}_{S_k}\|_p \le \sum_{k \ge 0} s^{1/p-1/2} \|\mathbf{x}_{S_k}\|_2 \le \sum_{k \ge 0} \frac{s^{1/p-1/2}}{\sqrt{1-\delta}} \|\mathbf{A}(\mathbf{x}_{S_k})\|_2 \\
&= \frac{s^{1/p-1/2}}{\sqrt{1-\delta}} \left[\|\mathbf{A}(-\sum_{k \ge 1} \mathbf{x}_{S_k})\|_2 + \sum_{k \ge 1} \|\mathbf{A}(\mathbf{x}_{S_k})\|_2 \right] \\
&\le \frac{s^{1/p-1/2}}{\sqrt{1-\delta}} \left[2 \sum_{k \ge 1} \|\mathbf{A}(\mathbf{x}_{S_k})\|_2 \right] \le 2\sqrt{\frac{1+\delta}{1-\delta}}\, s^{1/p-1/2} \sum_{k \ge 1} \|\mathbf{x}_{S_k}\|_2 \\
&\le 2\sqrt{\frac{1+\delta}{1-\delta}}\, s^{1/p-1/2} \sum_{k \ge 1} \|\mathbf{x}_{S_{k-1}}\|_1/\sqrt{s} = 2\sqrt{\frac{1+\delta}{1-\delta}}\, \frac{1}{s^{1-1/p}} \sum_{k \ge 1} \|\mathbf{x}_{S_{k-1}}\|_1 \\
&\le 2\sqrt{\frac{1+\delta}{1-\delta}} \left(\frac{2c \ln(eN/m)}{m} \right)^{1-1/p} \|\mathbf{x}\|_1.
\end{aligned}$$

Using $\delta = 7/9$ and $2^{1-1/p} \le 2$, it follows that, for all $\mathbf{x} \in B_1^N \cap L^m$,

$$\|\mathbf{x}\|_p \le 8\sqrt{2} \left\{ \frac{c \ln(eN/m)}{m} \right\}^{1-1/p}.$$

This shows that, if $m > c\ln(eN/m)$, then

$$d^m(B_1^N, \ell_p^N) \le 8 \min\left\{1, \frac{c\ln(eN/m)}{m}\right\}^{1-1/p}. \tag{10.7}$$

Combining (10.6) and (10.7), we conclude

$$d^m(B_1^N, \ell_p^N) \le C \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}$$

with $C = 8\sqrt{2}c = 1152\sqrt{2}(1+e^{-1})$, which is the desired upper bound. □

Lower Bound

We now establish the lower bound for the Gelfand width of ℓ_1-balls in ℓ_p^N for $1 < p \le \infty$. This bound matches the previous upper bound up to a multiplicative constant. We point out that a lower bound where the minimum does not appear would be invalid, since the width $d^m(B_1^N, \ell_p^N)$ is bounded above by one, hence cannot exceed $c\ln(eN/m)/m$ for large N.

Proposition 10.10. *There is a constant $c > 0$ such that, for $1 < p \le \infty$ and $m < N$,*

$$d^m(B_1^N, \ell_p^N) \ge c \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

The proof of this proposition relies again on the methods of compressive sensing. In particular, it requires the important result stated next.

Theorem 10.11. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times N}$, if every $2s$-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is a minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{Az} = \mathbf{Ax}$, then*

$$m \ge c_1 s \ln\left(\frac{N}{c_2 s}\right),$$

where $c_1 = 1/\ln 9$ and $c_2 = 4$.

This is based on the key combinatorial lemma that follows.

Lemma 10.12. *Given integers $s < N$, there exist*

$$n \ge \left(\frac{N}{4s}\right)^{s/2} \tag{10.8}$$

subsets $S_1, \dots, S_n$ of $[N]$, such that each S_j has cardinality s and

$$\operatorname{card}(S_i \cap S_j) < \frac{s}{2} \qquad \text{whenever } i \ne j. \tag{10.9}$$

Proof. We may assume that $s \le N/4$, for otherwise it suffices to take $n = 1$ subset of $[N]$. Let $\mathcal{B}$ denote the family of subsets of $[N]$ having cardinality s. We draw an element $S_1 \in \mathcal{B}$ and we collect in a family $\mathcal{A}_1$ all the sets $S \in \mathcal{B}$ such that $\mathrm{card}(S_1 \cap S) \ge s/2$. We have

$$\mathrm{card}(\mathcal{A}_1) = \sum_{k=\lceil s/2 \rceil}^{s} \binom{s}{k}\binom{N-s}{s-k} \le 2^s \max_{\lceil s/2 \rceil \le k \le s} \binom{N-s}{s-k} = 2^s \binom{N-s}{\lfloor s/2 \rfloor},$$

where the last equality holds because $\lfloor s/2 \rfloor \le (N-s)/2$ when $s \le N/2$. We observe that any set $S \in \mathcal{B} \setminus \mathcal{A}_1$ satisfies $\mathrm{card}(S_1 \cap S) < s/2$. Next, we draw an element $S_2 \in \mathcal{B} \setminus \mathcal{A}_1$, provided that the latter is nonempty. As before, we collect in a family $\mathcal{A}_2$ all the sets $S \in \mathcal{B} \setminus \mathcal{A}_1$ such that $\mathrm{card}(S_2 \cap S) \ge s/2$, we remark that

$$\mathrm{card}(\mathcal{A}_2) \le 2^s \binom{N-s}{\lfloor s/2 \rfloor},$$

and we observe that any set $S \in \mathcal{B} \setminus (\mathcal{A}_1 \cup \mathcal{A}_2)$ satisfies $\mathrm{card}(S_1 \cap S) < s/2$ and $\mathrm{card}(S_2 \cap S) < s/2$. We repeat the procedure of selecting sets $S_1, \ldots, S_n$ until $\mathcal{B} \setminus (\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n)$ is empty. In this way, (10.9) is automaticall fulfilled. Moreover,

$$\begin{aligned} n &\ge \frac{\mathrm{card}(\mathcal{B})}{\max_{1 \le i \le n} \mathrm{card}(\mathcal{A}_i)} \ge \frac{\binom{N}{s}}{2^s \binom{N-s}{\lfloor s/2 \rfloor}} \\ &= \frac{1}{2^s} \frac{N(N-1)\cdots(N-s+1)}{(N-s)(N-s-1)\cdots(N-s-\lfloor s/2 \rfloor + 1)} \frac{1}{s(s-1)\cdots(\lfloor s/2 \rfloor + 1)} \\ &\ge \frac{1}{2^s} \frac{N(N-1)\cdots(N-\lceil s/2 \rceil + 1)}{s(s-1)\cdots(s-\lceil s/2 \rceil + 1)} \ge \frac{1}{2^s} \left(\frac{N}{s}\right)^{\lceil s/2 \rceil} \ge \left(\frac{N}{4s}\right)^{s/2}. \end{aligned}$$

This shows that (10.8) is fulfilled, too, and concludes the proof. □

With this lemma at hand, we can turn to the proof of the theorem.

Proof (of Theorem 10.11). Let us consider the quotient space

$$X := \ell_1^N / \ker \mathbf{A} = \big\{[\mathbf{x}] := \mathbf{x} + \ker \mathbf{A}, \mathbf{x} \in \mathbb{R}^N\big\},$$

which is normed with

$$\|[\mathbf{x}]\| := \inf_{\mathbf{v} \in \ker \mathbf{A}} \|\mathbf{x} - \mathbf{v}\|_1, \qquad \mathbf{x} \in \mathbb{R}^N.$$

Given a $2s$-sparse vector $\mathbf{x} \in \mathbb{R}^N$, we notice that every vector $\mathbf{z} = \mathbf{x} - \mathbf{v}$ with $\mathbf{v} \in \ker \mathbf{A}$ satisfies $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$. Thus, our assumption gives $\|[\mathbf{x}]\| = \|\mathbf{x}\|_1$. Let $S_1, \ldots, S_n$ be the sets introduced in Lemma 10.12, and let us define s-sparse vectors $\mathbf{x}^1, \ldots, \mathbf{x}^n \in \mathbb{R}^N$ with unit ℓ_1-norms by

$$x_k^i = \begin{cases} 1/s & \text{if } k \in S_i, \\ 0 & \text{if } k \notin S_i. \end{cases} \tag{10.10}$$

For $1 \le i \ne j \le n$, we have $\|[\mathbf{x}^i] - [\mathbf{x}^j]\| = \|[\mathbf{x}^i - \mathbf{x}^j]\| = \|\mathbf{x}^i - \mathbf{x}^j\|_1$, since the vector $\mathbf{x}^i - \mathbf{x}^j$ is $2s$-sparse. We also have $\|\mathbf{x}^i - \mathbf{x}^j\|_1 > 1$, since $|x_k^i - x_k^j|$ equals $1/s$ if $k \in S_i \Delta S_j$ and vanishes otherwise and since $\mathrm{card}(S_i \Delta S_j) > s$. We conclude that

$$\|[\mathbf{x}^i] - [\mathbf{x}^j]\| > 1 \qquad \text{for all } 1 \le i \ne j \le n.$$

This shows that $\{[\mathbf{x}^1], \dots, [\mathbf{x}^n]\}$ is a 1-separating subset of the unit sphere of X, which has dimension $r := \mathrm{rank}(\mathbf{A}) \le m$. According to Proposition C.3, this implies that $n \le 3^r \le 3^m$. In view (10.8), we obtain

$$\left(\frac{N}{4s}\right)^{s/2} \le 3^m.$$

Taking the logarithm on both sides gives the desired result. □

We are now ready to prove the main result of this section.

Proof (of Proposition 10.10). With $c' := 2/(1 + 4 \ln 9)$, we are going to show that

$$d^m(B_1^N, \ell_p^N) \ge \frac{\mu^{1-1/p}}{2^{2-1/p}}, \qquad \text{where } \mu := \min\left\{1, \frac{c' \ln(eN/m)}{m}\right\}.$$

The result will then follow with $c = \min\{1, c'\}^{1-1/p}/2^{2-1/p} \ge \min\{1, c'\}/4$. By way of contradiction, we assume that $d^m(B_1^N, \ell_p^N) < \mu^{1-1/p}/2^{2-1/p}$. This implies the existence of a subspace L^m of $\mathbb{R}^N$ with $\mathrm{codim}(L^m) \le m$ such that, for all $\mathbf{v} \in L^m \setminus \{0\}$,

$$\|\mathbf{v}\|_p < \frac{\mu^{1-1/p}}{2^{2-1/p}} \|\mathbf{v}\|_1.$$

Let us consider a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ such that $\ker \mathbf{A} = L^m$. Let us also define an integer $s \ge 1$ by $s := \lfloor 1/\mu \rfloor$, so that

$$\frac{1}{2\mu} < s \le \frac{1}{\mu}.$$

We have in this way, for all $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$,

$$\|\mathbf{v}\|_p < \frac{1}{2}\left(\frac{1}{2s}\right)^{1-1/p} \|\mathbf{v}\|_1.$$

The inequality $\|\mathbf{v}\|_1 \le N^{1-1/p} \|\mathbf{v}\|_p$ ensures that $1 < (N/2s)^{1-1/p}/2$, hence that $2s < N$. Then, for $S \subseteq [N]$ with $\mathrm{card}(S) \le 2s$ and for $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$, we have

$$\|\mathbf{v}_S\|_1 \le (2s)^{1-1/p} \|\mathbf{v}_S\|_p \le (2s)^{1-1/p} \|\mathbf{v}\|_p < \frac{1}{2} \|\mathbf{v}\|_1.$$

This is the null space property (4.2) of order $2s$. Thus, according to Theorem 4.5, every $2s$-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is uniquely recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ by ℓ_1-minimization. Theorem 10.11 now implies that

$$m \geq c_1 s \ln\Big(\frac{N}{c_2 s}\Big), \qquad c_1 = \frac{1}{\ln 9}, \quad c_2 = 4.$$

Theorem 2.13 also implies that $m \geq 2(2s) = c_2 s$. It follows that

$$m \geq c_1 s \ln\Big(\frac{N}{m}\Big) = c_1 s \ln\Big(\frac{eN}{m}\Big) - c_1 s > \frac{c_1}{2\mu} \ln\Big(\frac{eN}{m}\Big) - \frac{c_1}{4} m.$$

After rearrangement, we deduce

$$m > \frac{2c_1}{4 + c_1} \frac{\ln(eN/m)}{\min\big\{1, c' \ln(eN/m)/m\big\}} \geq \frac{2c_1}{4 + c_1} \frac{\ln(eN/m)}{c' \ln(eN/m)/m} = m.$$

This is the desired contradiction. □

10.3 Applications to the Geometry of Banach Spaces

Let us now make a slight detour and highlight two applications of the previous results and their proofs in Banach space geometry. By relating the Gelfand widths to their duals, the Kolmogorov widths, we obtain also lower and upper bounds for those. Moreover, we show that $\mathbb{R}^{2m}$ can be splitted into two orthogonal subspaces on which the ℓ_1-norm and the ℓ_2-norm are essentially equivalent. This is called a Kashin splitting.

Kolmogorov widths

Let us start with the definition.

Definition 10.13. *The Kolmogorov m-width of a subset K of a normed space X is defined as*

$$d_m(K, X) := \inf \left\{ \sup_{\mathbf{x} \in K} \inf_{\mathbf{z} \in X_m} \|\mathbf{x} - \mathbf{z}\|, \ X_m \text{ subspace of } X \text{ with } \dim(X_m) \leq m \right\}.$$

The Kolmogorov widths of ℓ_p-balls in ℓ_q are closely related to certain Gelfand widths as shown by the following duality result.

Theorem 10.14. *Let $1 \leq p, q \leq \infty$ and p^*, q^* such that $1/p^* + 1/p = 1$ and $1/q^* + 1/q = 1$. Then*

$$d_m(B_p^N, \ell_q^N) = d^m(B_{q^*}^N, \ell_{p^*}^N).$$

The proof uses a classical observation about best approximation.

Lemma 10.15. *Let Y be a finite-dimensional subspace of a normed space X. Given $\mathbf{x} \in X \setminus Y$ and $\mathbf{y}^\star \in Y$, the following properties are equivalent:*

(a) $\mathbf{y}^\star$ is a best approximation to $\mathbf{x}$ from Y,
(b) $\|\mathbf{x} - \mathbf{y}^\star\| = \lambda(\mathbf{x})$ for some linear functional $\lambda \in B_{X^*}$ vanishing on Y.

Proof. Let us first assume that (b) holds. To derive (a), we simply observe that $\lambda(\mathbf{y}) = 0$ for all $\mathbf{y} \in Y$, so that

$$\|\mathbf{x} - \mathbf{y}^\star\| = \lambda(\mathbf{x}) = \lambda(\mathbf{x} - \mathbf{y}) \le \|\lambda\| \, \|\mathbf{x} - \mathbf{y}\| \le \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{y} \in Y.$$

Conversely, let us assume that (a) holds. We define a linear functional $\widetilde{\lambda}$ on the space $[Y \oplus \operatorname{span}(\mathbf{x})]$ by

$$\widetilde{\lambda}(\mathbf{y} + t\mathbf{x}) = t \, \|\mathbf{x} - \mathbf{y}^\star\| \quad \text{for all } \mathbf{y} \in Y \text{ and } t \in \mathbb{R}.$$

It is readily seen that $\widetilde{\lambda}$ vanishes on Y. Besides, for $\mathbf{y} \in Y$ and $t \neq 0$, we have

$$|\widetilde{\lambda}(\mathbf{y} + t\mathbf{x})| = |t| \, \|\mathbf{x} - \mathbf{y}^\star\| \le |t| \, \|\mathbf{x} - (-\mathbf{y}/t)\| = \|\mathbf{y} + t\mathbf{x}\|.$$

This inequality — which remains valid for $t = 0$ — allows to derive $\|\widetilde{\lambda}\| \le 1$. The linear functional λ required in (b) is the Hahn-Banach extension of the linear functional $\widetilde{\lambda}$ to the whole space X. □

Proof (of Theorem 10.14). Given a subspace X_m of ℓ_q^N with $\dim(X_m) \le m$ and a vector $\mathbf{x} \in B_p^N$, Lemma 10.15 shows that

$$\inf_{\mathbf{z} \in X_m} \|\mathbf{x} - \mathbf{z}\|_q =: \|\mathbf{x} - \mathbf{z}^\star\|_q \le \sup_{\mathbf{u} \in B_{q^*}^N \cap X_m^\perp} \langle \mathbf{u}, \mathbf{x} \rangle.$$

Moreover, for all $\mathbf{u} \in B_{q^*}^N \cap X_m^\perp$, we have

$$\langle \mathbf{u}, \mathbf{x} \rangle = \langle \mathbf{u}, \mathbf{x} - \mathbf{z}^\star \rangle = \|\mathbf{u}\|_{q^*} \|\mathbf{x} - \mathbf{z}^\star\|_q \le \|\mathbf{x} - \mathbf{z}^\star\|_q.$$

We deduce the equality

$$\inf_{\mathbf{z} \in X_m} \|\mathbf{x} - \mathbf{z}\|_q = \sup_{\mathbf{u} \in B_{q^*}^N \cap X_m^\perp} \langle \mathbf{u}, \mathbf{x} \rangle.$$

It follows that

$$\begin{aligned} \sup_{\mathbf{x} \in B_p^N} \inf_{\mathbf{z} \in X_m} \|\mathbf{x} - \mathbf{z}\|_q &= \sup_{\mathbf{x} \in B_p^N} \sup_{\mathbf{u} \in B_{q^*}^N \cap X_m^\perp} \langle \mathbf{u}, \mathbf{x} \rangle = \sup_{\mathbf{u} \in B_{q^*}^N \cap X_m^\perp} \sup_{\mathbf{x} \in B_p^N} \langle \mathbf{u}, \mathbf{x} \rangle \\ &= \sup_{\mathbf{u} \in B_{q^*}^N \cap X_m^\perp} \|\mathbf{u}\|_{p^*}. \end{aligned}$$

Taking the infimum over all subspaces X_m with $\dim(X_m) \le m$ and noticing the one-to-one correspondence between the subspaces $X_m^\perp$ and the subspaces L^m with $\operatorname{codim}(L^m) \le m$, we conclude

$$d_m(B_p^N, \ell_q^N) = d^m(B_{q^*}^N, \ell_{p^*}^N).$$

This is the desired identity. □

Our estimate on the Gelfand widths in Theorem 10.5 immediately implies now the following estimate of the Kolmogorov widths of ℓ_p^N-balls in ℓ_∞^N for $p \in [2, \infty)$.

Theorem 10.16. *Let $2 \le p < \infty$ and $m < N$. Then there exist constants $c_1, c_2 > 0$ depending only on p such that*

$$c_1 \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1/p} \le d_m(B_p^N, \ell_\infty^N) \le c_2 \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1/p}.$$

Kashin's Decomposition Theorem

If we specify the upper estimate of the Gelfand width of the unit ℓ_1-ball in ℓ_2^N to the case $N = 2m$, we obtain $d^m(B_1^{2m}, \ell_2^{2m}) \le C/\sqrt{m}$, which says that there is a subspace E of $\mathbb{R}^{2m}$ such that

$$\|\mathbf{x}\|_2 \le \frac{C}{\sqrt{m}} \|\mathbf{x}\|_1 \qquad \text{for all } \mathbf{x} \in E.$$

Together with $\|\mathbf{x}\|_1 \le \sqrt{2m}\,\|\mathbf{x}\|_2$, which is valid for any $\mathbf{x} \in \mathbb{R}^{2m}$, this says that the norms $\|\cdot\|_1/\sqrt{m}$ and $\|\cdot\|_2$ are comparable on E. In other words, as a subspace of ℓ_1^{2m}, the m-dimensional space E is almost Euclidean. Kashin's decomposition theorem states something more, namely that one can find an m-dimensional space E such that both E and its orthogonal complement $E^\perp$, as subspaces of ℓ_1^{2m}, are almost Euclidean.

Theorem 10.17. *There exist universal constants $\alpha, \beta > 0$ such that, for any $m \ge 1$, the space $\mathbb{R}^{2m}$ contains two orthogonal subspaces E and $E^\perp$ of dimension m satisfying*

$$\alpha \sqrt{m}\,\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \beta \sqrt{m}\,\|\mathbf{x}\|_2 \tag{10.11}$$

for all $\mathbf{x} \in E$ and all $\mathbf{x} \in E^\perp$.

Proof. The first inequality in (10.11) holds with $\beta := \sqrt{2}$ regardless of the subspace E of $\mathbb{R}^{2m}$ considered, so we focus on the second inequality. Let $\mathbf{G}$ be an $m \times m$ matrix whose entries are independent Gaussian random variables with mean zero and variance $1/m$. We define two full-rank $m \times (2m)$ matrices by

$$\mathbf{A} := \left[\, \mathbf{Id} \;\middle|\; \mathbf{G} \,\right], \qquad \mathbf{B} := \left[\, \mathbf{G}^* \;\middle|\; -\mathbf{Id} \,\right],$$

and we consider the m-dimensional space $E := \ker \mathbf{A}$. In view of $\mathbf{B}\mathbf{A}^* = 0$, we have $E^\perp = \operatorname{im} \mathbf{A}^* \subseteq \ker \mathbf{B}$, and $E^\perp = \ker \mathbf{B}$ follows from dimension arguments. We are going to show that, given any $t \in (0,1)$ and any $\mathbf{x} \in \mathbb{R}^{2m}$, the matrices $\mathbf{M} = \mathbf{A}$ and $\mathbf{M} = \mathbf{B}$ satisfy the concentration inequality

$$\mathbb{P}\big(\big|\|\mathbf{M}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \ge t\|\mathbf{x}\|_2^2\big) \le 2\exp\big(-\tilde{c}t^2 m\big) \tag{10.12}$$

for some constant $\tilde{c} > 0$. Fixing $0 < \delta < 1$, say $\delta := \sqrt{2/3}$, Theorem 9.10 with $\epsilon = 2\exp(-\tilde{c}m/4)$ implies that $\delta_s(\mathbf{A}) \le \delta$ and $\delta_s(\mathbf{B}) \le \delta$ with probability at least $1 - 4\exp(-\tilde{c}m/4)$ provided

$$m \ge \frac{2}{3\tilde{c}\delta^2}[s(9 + 2\ln(2m/s)) + \tilde{c}m/2], \text{ i.e., } \tilde{c}m \ge 2s(9 + 2\ln(2m/s)). \quad (10.13)$$

We take $m > 8\ln(2)/\tilde{c}$ to make the above probability positive, and we also take $m > 1/(2\gamma)$ for a constant γ small enough to have $4\gamma(9 + 2\ln(2/\gamma)) \le \tilde{c}$. In this way, the integer $s := \lfloor 2\gamma m \rfloor \ge 1$ satisfies $\gamma m \le s \le 2\gamma m$, and (10.13) is therefore fulfilled. Let now $\mathbf{x} \in E \cup E^\perp$, i.e., $\mathbf{x} \in \ker \mathbf{M}$ for $\mathbf{M} = \mathbf{A}$ or $\mathbf{M} = \mathbf{B}$. Reproducing the argument in the proof of Proposition 10.9, starting with the partition $[N] = S_0 \cup S_1 \cup S_2 \cup \cdots$, we arrive at

$$\|\mathbf{x}\|_2 \le 2\sqrt{\frac{1+\delta}{1-\delta}}\frac{\|\mathbf{x}\|_1}{\sqrt{s}} \le \frac{2(\sqrt{2}+\sqrt{3})}{\sqrt{\gamma m}}\|\mathbf{x}\|_1.$$

This is the desired inequality with $\sqrt{\gamma}/(2(\sqrt{2}+\sqrt{3}))$ taking the role of α when $m > m_* := \max\{8\ln(2)/\tilde{c}, 1/(2\gamma)\}$. When $m \le m_*$, the desired inequality simply follows from $\|\mathbf{x}_1\| \ge \|\mathbf{x}\|_2 \ge \sqrt{m}\|\mathbf{x}\|_2/\sqrt{m_*}$. The result is therefore acquired with $\alpha := \min\{\sqrt{\gamma}/(2(\sqrt{2}+\sqrt{3})), 1/\sqrt{m_*}\}$. It remains to establish the concentration inequality (10.12). In the case $\mathbf{M} = \mathbf{A}$ — the case $\mathbf{M} = \mathbf{B}$ being similar — we notice that, with $\mathbf{x} = [\mathbf{u}, \mathbf{v}]^\top$

$$\begin{aligned}\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| = \big|\|\mathbf{u} + \mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{u}\|_2^2 - \|\mathbf{v}\|_2^2\big| &= \big|2\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle + \|\mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big| \\ &\le 2\big|\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle\big| + \big|\|\mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big|.\end{aligned}$$

Thus, if $\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \ge t\|\mathbf{x}\|_2^2$, at least one of the following two alternatives holds:

$$2\big|\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle\big| \ge \frac{t}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2), \quad \text{in which case } \big|\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle\big| \ge \frac{t}{2}\|\mathbf{u}\|_2\|\mathbf{v}\|_2,$$
$$\big|\|\mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big| \ge \frac{t}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2), \quad \text{in which case } \big|\|\mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big| \ge \frac{t}{2}\|\mathbf{v}\|_2^2.$$

In terms of probability, we have

$$\begin{aligned}\mathbb{P}\big(\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \ge t\|\mathbf{x}\|_2^2\big) \le{}& \mathbb{P}\big(\big|\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle\big| \ge t\|\mathbf{u}\|_2\|\mathbf{v}\|_2/2\big) \\ &+ \mathbb{P}\big(\big|\|\mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big| \ge t\|\mathbf{v}\|_2^2/2\big).\end{aligned}$$

For the first of these probabilities, we observe that

$$\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle = \sum_{i=1}^m u_i \sum_{j=1}^m g_{i,j} v_j \ \sim\ \sum_{i=1}^m u_i \mathcal{N}\big(0, \|\mathbf{v}\|_2^2/m\big) \ \sim\ \mathcal{N}\big(0, \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2/m\big),$$

so that the standard tail estimate of Proposition 7.5 gives

$$\mathbb{P}\big(\big|\langle \mathbf{u}, \mathbf{G}\mathbf{v}\rangle\big| \ge t\|\mathbf{u}\|_2\|\mathbf{v}\|_2/2\big) = \mathbb{P}\big(|g| \ge t\sqrt{m}/2\big) \le \exp\big(-t^2 m/8\big).$$

For the second probability, we recall from Exercise 9.3 (see also Lemma 9.7, where the constant is not specified) that

$$\mathbb{P}\big(\big|\|\mathbf{G}\mathbf{v}\|_2^2 - \|\mathbf{v}\|_2^2\big| \ge t\|\mathbf{v}\|_2^2/2\big) \le 2\exp\big(-(t^2/16 - t^3/48)m\big) \le 2\exp\big(-t^2 m/24\big),$$

while this probability is also less than one. As a consequence of the previous estimates, we obtain

$$\mathbb{P}\big(\big|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \ge t\|\mathbf{x}\|_2^2\big) \le \exp\big(-t^2 m/8\big) + \min\big\{1, 2\exp\big(-t^2 m/24\big)\big\}.$$

To complete the proof, it remains to notice that the latter is smaller than $2\exp(-\tilde{c}t^2 m)$ for the properly chosen constant $\tilde{c} = \ln(4/3)/\ln(2^{12})$. □

Notes

The definition of Gelfand widths sometimes appear with $\operatorname{codim}(L^m) = m$ instead of $\operatorname{codim}(L^m) \le m$, see for instance A. Pinkus' book [334]. This is of course equivalent to the definition we have used.

We have coined the terms nonadaptive and adaptive compressive widths for the quantity $E^m(C, X)$ and $E^m_{\mathrm{ada}}(C, X)$. In the compressive sensing literature, the nonadaptive compressive width appeared, along with the corresponding part of Theorem 10.4, in [102], see also [130, 319, 185]. The other part of Theorem 10.4 is an instance of general results from Information-Based Complexity showing that 'adaptivity does not help', see [320].

The lower estimate for Gelfand widths of ℓ_1-balls given in Proposition 10.10 was obtained by A. Garnaev and E. Gluskin in [189]. Their original proof, which is reproduced in Exercise 10.9, dealt with the dual Kolmogorov width. The proof relying only on compressive sensing techniques presented here was proposed in [185], where the case of ℓ_p-balls, $0 < p \le 1$, was treated in a similar way. The key combinatorial lemma, namely Lemma 10.12, follows [185, 301], but it had also been used in other areas before, see e.g. [317, 64, 204].

For $1 < q < p \le \infty$, the order of the Gelfand widths of ℓ_q-balls in ℓ_p^N is known; see [288, pages 481-482] and the references therein for the dual statement about Kolmogorov widths. Precisely, for $1 \le m < N$, we have

- if $1 < q < p \le 2$,

$$d^m(B_q^N, \ell_p^N) \asymp \min\left\{1, \frac{N^{1-1/q}}{m^{1/2}}\right\}^{\frac{1/q-1/p}{1/q-1/2}},$$

- if $1 < q \le 2 < p \le \infty$,

$$d^m(B_q^N, \ell_p^N) \asymp \max\left\{\frac{1}{N^{1/q-1/p}}, \left(1 - \frac{m}{N}\right)^{1/2} \min\left(1, \frac{N^{1-1/q}}{m^{1/2}}\right)\right\},$$

- if $2 \le q < p \le \infty$,

$$d^m(B_q^N, \ell_p^N) \asymp \max\left\{ \frac{1}{N^{1/q-1/p}}\,,\ \left(1 - \frac{m}{N}\right)^{\frac{1/q-1/p}{1-2/p}} \right\}.$$

Theorem 10.17 was first established by B. Kashin in [259]. S. Szarek then gave a shorter proof in [396]. The argument presented here is close to a proof given by G. Schechtman in [382], which implicitly contained a few ideas now familiar in compressive sensing.

Exercises

10.1. Determine the Gelfand widths $d^m(B_1^N, \ell_2^N)$, $1 \le m < N$, of the unit ℓ_1-ball in the Euclidean space ℓ_2^N when $N = 2$ and $N = 3$.

10.2. For a compact subset K of an infinite-dimensional normed space K, prove that $\lim_{m\to\infty} d^m(C, X) = 0$.

10.3. Let K be the subset of $L_2(\mathbb{T})$ defined by

$$K := \left\{ g \in \mathcal{C}^1(\mathbb{T}) : \|g'\|_2 \le 1 \right\}.$$

Prove that

$$d_0(K, L_2(\mathbb{T})) = \infty, \qquad d_{2n-1}(K, L_2(\mathbb{T})) = d_{2n}(K, L_2(\mathbb{T})) = \frac{1}{n} \quad \text{for } n \ge 1.$$

Evaluate first the quantity

$$\sup_{f \in L_2(\mathbb{T})} \inf_{g \in \mathcal{T}_{n-1}} \|f - g\|_2,$$

where

$$\mathcal{T}_{n-1} := \operatorname{span}[1, \sin(x), \cos(x), \dots, \sin((n-1)x), \cos((n-1)x)]$$

is the space of trigonometric polynomials of degree at most $n - 1$.

10.4. Prove that

$$d^m(B_{X_n}, X) = 1, \quad X_n \text{ an } n\text{-dimensional subspace of } X, \quad m < n. \quad (10.14)$$

Prove also that

$$d_m(B_{X_n}, X) = 1, \quad X_n \text{ an } n\text{-dimensional subspace of } X, \quad m < n. \quad (10.15)$$

For (10.15), use the so-called *theorem of deviation of subspaces*:
If U and V be two finite-dimensional subspaces of a normed space X with

$\dim(V) > \dim(U)$, *then there exists a nonzero vector* $\mathbf{v} \in V$ *to which zero is a best approximation from* U*, i.e.,*

$$\|\mathbf{v}\| \le \|\mathbf{v} - \mathbf{u}\| \qquad \text{for all} \quad \mathbf{u} \in U.$$

You should derive this theorem from *Borsuk–Ulam theorem*:
If a continuous map F *from the sphere* $\mathbb{S}^n$ *— relative to an arbitrary norm — of* $\mathbb{R}^{n+1}$ *into* $\mathbb{R}^n$ *is antipodal, i.e.,*

$$F(-\mathbf{x}) = -F(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathbb{S}^n,$$

then it vanishes at least once, i.e.,

$$F(\mathbf{x}) = 0 \qquad \text{for some } \mathbf{x} \in \mathbb{S}^n.$$

10.5. Let K be a subset of a normed space X with $0 \in K$. Prove that

$$d^m(K, X) \le 2E^m_{\mathrm{ada}}(K, X).$$

10.6. Let $B^N_{1,+}$ be the subset of the unit ball B^N_1 consisting of all nonnegative vectors, i.e.,

$$B^N_{1,+} = \{\mathbf{x} \in B^N_1 : x_j \ge 0 \text{ for all } j \in [N]\}.$$

Prove that

$$d^m(B^N_1, \ell^N_2) \le 2E^m(B^N_{1,+}, \ell^N_2),$$

and deduce that

$$E^m(B^N_{1,+}, \ell^N_2) \asymp \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1/2}.$$

10.7. For $1 \le p < q \le \infty$ and $m < N$, prove that

$$d^m(B^N_p, \ell^N_q) \ge \frac{1}{(m+1)^{1/p-1/q}}.$$

10.8. For $\mathbf{A} \in \mathbb{R}^{m \times N}$ and $s \ge 2$, show that if every s-sparse vector $\mathbf{x} \in \mathbb{R}^N$ is a minimizer of $\|\mathbf{z}\|_1$ subject to $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, then $m \ge c\, s \ln(eN/s)$ for some constant $c > 0$, but that this does not hold for $s = 1$.

10.9. Original proof of the lower bound
This problem aims at establishing the lower bound of Proposition 10.10 by way of the Kolmogorov width $d_m(B^N_p, \ell^N_\infty)$, $1 \le p \le \infty$.
(a) Given a subset C of the normed space X, for $\varepsilon > 2d_m(C, X)$ and $t > 0$, prove that the maximal number of points in $C \cap tB_X$ with mutual distance in X exceeding ε satisfies

$$P(\varepsilon, C \cap tB_X, X) \le \left(1 + 2\frac{t + d_m(C, X)}{\varepsilon - 2d_m(C, X)}\right)^m.$$

(b) For $1 \le k \le N$ and $0 < \varepsilon < k^{-1/p}$, prove that

$$P(\varepsilon, B_p^N \cap k^{-1/p} B_\infty^N, \ell_\infty^N) \ge 2^k \binom{N}{k}.$$

(c) Conclude that, for $1 \le m < N$,

$$d_m(B_p^N, \ell_\infty^N) \ge \frac{1}{3} \min\left\{1, \frac{\ln(3N/m)}{6m}\right\}^{1/p}.$$

10.10. Observe that Kashin's decomposition theorem also applies to ℓ_p^{2m} with $1 < p \le 2$ instead of ℓ_1^{2m}, i.e., observe that there are orthogonal subspaces E and $E^\perp$ of dimension m such that

$$\alpha\, m^{1/p-1/2} \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_p \le \beta\, m^{1/p-1/2} \|\mathbf{x}\|_2$$

for all $\mathbf{x} \in E$ and all $\mathbf{x} \in E^\perp$, where $\alpha, \beta > 0$ are absolute constants.

11

Instance Optimality and Quotient Property

This chapter investigates further properties of ℓ_1-minimization as a reconstruction map. In Section 11.1, which deals with general reconstruction maps, the concept of instance optimality is introduced. The minimal number of measurements to achieve ℓ_1-instance optimality is determined, complementing some results from Chapter 10. It is also revealed that ℓ_2-instance optimality is not a good concept for the range of parameters typical to compressive sensing— this explains, in retrospect, the appearance of $\sigma_s(\mathbf{x})_1$ instead of $\sigma_s(\mathbf{x})_2$ in estimates for the reconstruction error. It is nonetheless established in Section 11.4 that the ℓ_1-minimization allows for a weaker form of the ℓ_2-instance optimality. The tools needed for the analysis of this nonuniform instance optimality are developed in Sections 11.2 and 11.3. There, the equality-constrained ℓ_1-minimization is investigated in the presence of nonzero measurement error. In Section 11.2, the concept of quotient property is introduced, and it is proved to imply stability and robustness estimates for the equality-constrained ℓ_1-minimization. It is then shown in Section 11.3 that different versions of the quotient property hold with high probability for Gaussian matrices and for subgaussian matrices.

11.1 Uniform Instance Optimality

When a measurement–reconstruction scheme is assessed for s-sparse recovery, it is natural to compare the reconstruction error for a vector $\mathbf{x} \in \mathbb{C}^N$ to the error of best s-term approximation

$$\sigma_s(\mathbf{x})_p = \inf \big\{ \|\mathbf{x} - \mathbf{z}\|_p, \ \mathbf{z} \in \mathbb{C}^N \text{ is } s\text{-sparse} \big\}.$$

This motivates the introduction of the *instance optimality* concept.

Definition 11.1. *Given $p \geq 1$, a pair of measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ is called* ℓ_p-instance optimal *of order s with constant $C > 0$ if*

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_p \le C\,\sigma_s(\mathbf{x})_p \qquad \textit{for all } \mathbf{x} \in \mathbb{C}^N.$$

In Theorems 6.11, 6.20, and 6.24, we have seen examples of ℓ_1-instance optimal pairs, i.e., a matrix $\mathbf{A}$ with small restricted isometry constants δ_{2s}, δ_{6s}, or δ_{20s}, together with a reconstruction map Δ corresponding to basis pursuit, iterative hard thresholding, or orthogonal matching pursuit, respectively. In fact, more general statements have been established where the reconstruction error was measured in ℓ_q for $q \ge 1$. With the following terminology, the previous pairs $(\mathbf{A}, \Delta)$ are mixed (ℓ_q, ℓ_1)-instance optimal.

Definition 11.2. *Given $q \ge p \ge 1$, a pair of measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ and reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ is called mixed* (ℓ_q, ℓ_p)-instance optimal *of order s with constant $C > 0$ if*

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_q \le \frac{C}{s^{1/p-1/q}}\,\sigma_s(\mathbf{x})_p \qquad \textit{for all } \mathbf{x} \in \mathbb{C}^N.$$

Remark 11.3. The term $s^{1/p-1/q}$ in this definition is not only motivated by the results previously mentioned. Indeed, since we are mainly interested in the reconstruction of compressible vectors, we want to compare the reconstruction error $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_q$ to the error of best approximation $\sigma_s(\mathbf{x})_q$ for vectors $\mathbf{x} \in \mathbb{C}^N$ belonging to balls B_r^N or $B_{r,\infty}^N$ with $r < 1$. By considering the nonincreasing rearrangements of such vectors, we can easily observe that

$$\sup_N \sup_{\mathbf{x}\in B_{r,\infty}^N} \sigma_s(\mathbf{x})_q \asymp \frac{1}{s^{1/r-1/q}} \asymp \frac{1}{s^{1/p-1/q}} \sup_N \sup_{\mathbf{x}\in B_{r,\infty}^N} \sigma_s(\mathbf{x})_p. \tag{11.1}$$

This justifies that we should compare $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_q$ to $\sigma_s(\mathbf{x})_p / s^{1/p-1/q}$.

Our goal is to determine conditions on the number of measurements under which instance optimality can be achieved for some pair of measurement matrix and reconstruction map. We start with a useful characterization for the existence of instance optimal pairs. We stress that the condition (11.2) below reduces, when $q = p = 1$, to

$$\|\mathbf{v}\|_1 \le C\,\sigma_{2s}(\mathbf{v})_1 \qquad \text{for all } \mathbf{v} \in \ker \mathbf{A}.$$

This is reminiscent of the null space property of order $2s$ for recovery via ℓ_1-minimization as formulated in (4.3), namely

$$\|\mathbf{v}\|_1 < 2\,\sigma_{2s}(\mathbf{v})_1 \qquad \text{for all } \mathbf{v} \in \ker \mathbf{A}.$$

The link between arbitrary instance optimal pairs $(\mathbf{A}, \Delta)$ and the pair $(\mathbf{A}, \Delta_1)$, where Δ_1 denotes the ℓ_1-minimization reconstruction map, will be further investigated in Exercise 11.5.

Theorem 11.4. *Let $q \ge p \ge 1$ and a measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ be given. If there exits a reconstruction map Δ making the pair (A, Δ) mixed* (ℓ_q, ℓ_p)*-instance optimal of order s with constant C, then*

$$\|\mathbf{v}\|_q \le \frac{C}{s^{1/p-1/q}} \sigma_{2s}(\mathbf{v})_p \qquad \text{for all } \mathbf{v} \in \ker \mathbf{A}. \tag{11.2}$$

Conversely, if (11.2) *is fulfilled, then there exists a reconstruction map* Δ *making the pair* $(\mathbf{A}, \Delta)$ *mixed* (ℓ_q, ℓ_p)*-instance optimal of order* s *with constant* $2C$*.*

Proof. Let us first assume that (A, Δ) is a mixed (ℓ_q, ℓ_p)-instance optimal pair of order s with constant C. Given $\mathbf{v} \in \ker \mathbf{A}$, let S be an index set of s largest entries of $\mathbf{v}$ in modulus. The instance optimality implies $-\mathbf{v}_S = \Delta(\mathbf{A}(-\mathbf{v}_S))$. Since $\mathbf{A}(-\mathbf{v}_S) = \mathbf{A}(\mathbf{v}_{\overline{S}})$, we have $-\mathbf{v}_S = \Delta(\mathbf{A}(\mathbf{v}_{\overline{S}}))$. We now derive (11.2) from

$$\begin{aligned} \|\mathbf{v}\|_q &= \|\mathbf{v}_{\overline{S}} + \mathbf{v}_S\|_q = \|\mathbf{v}_{\overline{S}} - \Delta(\mathbf{A}(\mathbf{v}_{\overline{S}}))\|_q \\ &\le \frac{C}{s^{1/p-1/q}}\, \sigma_s(\mathbf{v}_{\overline{S}})_p = \frac{C}{s^{1/p-1/q}}\, \sigma_{2s}(\mathbf{v})_p. \end{aligned}$$

Conversely, let us assume that (11.2) holds for some measurement matrix $\mathbf{A}$. We define a reconstruction map by

$$\Delta(\mathbf{y}) := \operatorname{argmin}\{\sigma_s(\mathbf{z})_p \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{y}\}.$$

For $\mathbf{x} \in \mathbb{C}^N$, applying (11.2) to $\mathbf{v} := \mathbf{x} - \Delta(\mathbf{A}\mathbf{x}) \in \ker \mathbf{A}$ yields

$$\begin{aligned} \|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_q &\le \frac{C}{s^{1/p-1/q}} \sigma_{2s}(\mathbf{x} - \Delta(\mathbf{A}\mathbf{x}))_p \\ &\le \frac{C}{s^{1/p-1/q}} \big(\sigma_s(\mathbf{x})_p + \sigma_s(\Delta(\mathbf{A}\mathbf{x}))_p\big) \le \frac{2C}{s^{1/p-1/q}} \sigma_s(\mathbf{x})_p, \end{aligned}$$

where we have used the triangle inequality $\sigma_{2s}(\mathbf{u}+\mathbf{v})_p \le \sigma_s(\mathbf{u})_p + \sigma_s(\mathbf{v})_p$ and the definition of $\Delta(\mathbf{A}\mathbf{x})$. This proves that (A, Δ) is a mixed (ℓ_q, ℓ_p)-instance optimal pair of order s with constant $2C$. □

Theorem 11.4 allows to prove that ℓ_2-instance optimality is not a pertinent concept in compressive sensing, since ℓ_2-instance optimal pairs — even of order 1 — can only exist if the number m of measurements is comparable to the dimension N. Note that this assertion will be moderated in Theorems 11.21 and 11.23, where we switch from a uniform point of view to a nonuniform point of view.

Theorem 11.5. *If a pair of measurement matrix and reconstruction map is* ℓ_2*-instance optimal of order* $s \ge 1$ *with constant* C*, then*

$$m \ge c\, N, \tag{11.3}$$

for some constant c *depending only on* C*.*

Proof. According to Theorem 11.4, the measurement matrix $\mathbf{A}$ in the instance optimal pair satisfies

$$\|\mathbf{v}\|_2 \le C\,\sigma_s(\mathbf{v})_2 \qquad \text{for all } \mathbf{v} \in \ker \mathbf{A}.$$

In particular, specifying this condition to $s = 1$ yields $\|\mathbf{v}\|_2^2 \le C^2(\|\mathbf{v}\|_2^2 - |v_j|^2)$ for all $\mathbf{v} \in \ker \mathbf{A}$ and all $j \in [N]$, i.e., $C^2|v_j|^2 \le (C^2-1)\|\mathbf{v}\|_2^2$. If $(\mathbf{e}_1, \dots, \mathbf{e}_N)$ denotes the canonical basis of $\mathbb{C}^N$, this means that $|\langle \mathbf{v}, \mathbf{e}_j\rangle| \le C'\,\|\mathbf{v}\|_2$ for all $\mathbf{v} \in \ker \mathbf{A}$ and all $j \in [N]$, where $C' := \sqrt{(C^2-1)/C^2}$. Thus, if P represents the orthogonal projector onto $\ker \mathbf{A}$, we have

$$N - m \le \dim(\ker \mathbf{A}) = \operatorname{tr}(P) = \sum_{j=1}^{N} \langle P\mathbf{e}_j, \mathbf{e}_j\rangle \le \sum_{j=1}^{N} C'\,\|P\mathbf{e}_j\|_2 \le N\,C'.$$

This immediately implies the desired result with $c = 1 - \sqrt{(C^2-1)/C^2}$. □

We now turn our attention to ℓ_1-instance optimality and (ℓ_q, ℓ_1)-instance optimality for $q \ge 1$. As already recalled, we have established in Chapter 6 that several reconstruction algorithms give rise to mixed (ℓ_q, ℓ_1)-instance optimal pairs $(\mathbf{A}, \Delta)$, provided the measurement matrix $\mathbf{A}$ has small restricted isometry constants. Moreover, Theorem 9.11 guarantees that this occurs with high probability for subgaussian random matrices $\mathbf{A}$ provided $m \ge c\,s\ln(eN/s)$ for some constant $c > 0$. Theorems 11.6 and 11.7 below show that a smaller number m of measurements is impossible. For $1 < q \le 2$, this was already derived using Gelfand width estimates in Proposition 10.7 with the proviso that s is large. This proviso will be lifted shortly. In the case $q = 1$, Gelfand width estimates can no longer be used, but the tools developed in Chapter 10 are still appropriate to deal with this more delicate case.

Theorem 11.6. *If a pair of measurement matrix and reconstruction map is ℓ_1-instance optimal of order s with constant C, then*

$$m \ge c\,s\,\ln(eN/s) \tag{11.4}$$

for some constant c depending only on C.

Proof. We call upon Lemma 10.12 to construct $n \ge (N/4s)^{s/2}$ index sets $S_1, \dots, S_n$ of size s satisfying $\operatorname{card}(S_i \cap S_j) < s/2$ for all $1 \le i \ne j \le n$. We consider the s-sparse vectors $\mathbf{x}^1, \dots, \mathbf{x}^n$ already defined in 10.10 by

$$x_k^i = \begin{cases} 1/s & \text{if } k \in S_i, \\ 0 & \text{if } k \notin S_i. \end{cases}$$

We notice that $\|\mathbf{x}^i\|_1 = 1$ and that $\|\mathbf{x}^i - \mathbf{x}^j\|_1 > 1$ for all $1 \le i \ne j \le n$. Let $(\mathbf{A}, \Delta)$ denote the ℓ_1-instance optimal pair of order s with constant C. Setting $\rho := 1/(2(C+1))$, we claim that $\{\mathbf{A}(\mathbf{x}^i + \rho B_1^N), i \in [n]\}$ is a disjoint collection of subsets of $\mathbf{A}(\mathbb{C}^N)$, which has dimension $d \le m$. Indeed, if there

existed indices $i \neq j$ and vectors $\mathbf{z}, \mathbf{z}' \in \rho B_1^N$ such that $\mathbf{A}(\mathbf{x}^i+\mathbf{z}) = \mathbf{A}(\mathbf{x}^j+\mathbf{z}')$, then a contradiction would follow from

$$\begin{aligned}\|\mathbf{x}^i - \mathbf{x}^j\|_1 &= \|\big(\mathbf{x}^i + \mathbf{z} - \Delta(\mathbf{A}(\mathbf{x}^i + \mathbf{z}))\big) - \big(\mathbf{x}^j + \mathbf{z}' - \Delta(\mathbf{A}(\mathbf{x}^j + \mathbf{z}'))\big) - \mathbf{z} + \mathbf{z}'\|_1 \\ &\leq \|\mathbf{x}^i + \mathbf{z} - \Delta(\mathbf{A}(\mathbf{x}^i + \mathbf{z}))\|_1 + \|\mathbf{x}^j + \mathbf{z}' - \Delta(\mathbf{A}(\mathbf{x}^j + \mathbf{z}'))\|_1 + \|\mathbf{z}\|_1 + \|\mathbf{z}'\|_1 \\ &\leq C\,\sigma_s(\mathbf{x}^i + \mathbf{z})_1 + C\,\sigma_s(\mathbf{x}^j + \mathbf{z}')_1 + \|\mathbf{z}\|_1 + \|\mathbf{z}'\|_1 \\ &\leq C\,\|\mathbf{z}\|_1 + C\,\|\mathbf{z}'\|_1 + \|\mathbf{z}\|_1 + \|\mathbf{z}'\|_1 \leq 2\,(C+1)\,\rho = 1.\end{aligned}$$

Next, we readily observe that the collection $\{\mathbf{A}(\mathbf{x}^i+\rho B_1^N), i \in [n]\}$ is contained in $(1+\rho)\mathbf{A}(B_1^N)$. Therefore, considering the volume of this collection, we deduce

$$\sum_{i\in[n]} \operatorname{vol}\big(\mathbf{A}(\mathbf{x}^i + \rho B_1^N)\big) \leq \operatorname{vol}\big((1+\rho)\mathbf{A}(B_1^N)\big)\,.$$

Using homogeneity and translation invariance of the volume, we derive

$$n\,\rho^d\,\operatorname{vol}\big(\mathbf{A}(B_1^N)\big) \leq (1+\rho)^d\,\operatorname{vol}\big(\mathbf{A}(B_1^N)\big)\,.$$

This yields

$$\left(\frac{N}{4s}\right)^{s/2} \leq n \leq \left(1+\frac{1}{\rho}\right)^d = (2C+3)^d \leq (2C+3)^m. \tag{11.5}$$

Taking the logarithms in (11.5) on the one hand, and on the other hand, remarking that the pair $(\mathbf{A}, \Delta)$ allows exact recovery of s-sparse vectors, we obtain

$$\frac{m}{s} \geq \frac{\ln(N/4s)}{2\ln(2C+3)}, \qquad \frac{m}{s} \geq 2.$$

Combining these two inequalities leads to

$$\Big(2\ln(2C+3)+2\Big)\,\frac{m}{s} \geq \ln(N/4s) + \ln(e^4) = \ln(e^4 N/4s) \geq \ln(eN/s).$$

This is the desired result where $c = 1/(2(\ln(2C+3)+1))$. □

With the help of Theorem 11.6, we can prove that the requirement (11.4) on the number of measurements is also imposed by mixed (ℓ_q, ℓ_1)-instance optimality when $q > 1$. This is formally stated in the following theorem.

Theorem 11.7. *Given $q > 1$, if a pair of measurement matrix and reconstruction map is mixed (ℓ_q, ℓ_1)-instance optimal of order s with constant C, then*

$$m \geq c\,s\,\ln(eN/s)$$

for some constant c depending only on C.

The proof is omitted, since it is a simple consequence of Theorem 11.6 and of the following lemma, which roughly says that mixed (ℓ_q, ℓ_1)-instance optimality is preserved when decreasing q.

Lemma 11.8. *Given $q \geq q' \geq p \geq 1$, if a pair $(\mathbf{A}, \Delta)$ is mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C, then there is a reconstruction map Δ' making the pair $(\mathbf{A}, \Delta')$ mixed $(\ell_{q'}, \ell_p)$-instance optimal of order s with constant C' depending only on C.*

Proof. Let us consider a vector $\mathbf{v} \in \ker \mathbf{A}$. Since the pair $(\mathbf{A}, \Delta)$ is mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C, Theorem 11.4 yields

$$\|\mathbf{v}\|_q \leq \frac{C}{s^{1/p-1/q}} \sigma_{2s}(\mathbf{v})_p.$$

Let S denote an index set of $3s$ largest entries of $\mathbf{v}$ in modulus. We have

$$\begin{aligned} \|\mathbf{v}_S\|_{q'} &\leq (3s)^{1/q'-1/q} \|\mathbf{v}_S\|_q \leq (3s)^{1/q'-1/q} \|\mathbf{v}\|_q \\ &\leq (3s)^{1/q'-1/q} \frac{C}{s^{1/p-1/q}} \sigma_{2s}(\mathbf{v})_p = \frac{3^{1/q'-1/q}\, C}{s^{1/p-1/q'}} \sigma_{2s}(\mathbf{v})_p \\ &\leq \frac{3\, C}{s^{1/p-1/q'}} \sigma_{2s}(\mathbf{v})_p. \end{aligned}$$

Moreover, we derive from Proposition 2.3 that

$$\|\mathbf{v}_{\overline{S}}\|_{q'} \leq \frac{1}{s^{1/p-1/q'}} \sigma_{2s}(\mathbf{v})_p.$$

Thus, we obtain

$$\|\mathbf{v}\|_{q'} \leq \|\mathbf{v}_S\|_{q'} + \|\mathbf{v}_{\overline{S}}\|_{q'} \leq \frac{3\, C + 1}{s^{1/p-1/q'}} \sigma_{2s}(\mathbf{v})_p.$$

In view of the converse part of Theorem 11.4, the desired result holds with $C' = 2(3\, C + 1)$. □

In parallel with Lemma 11.8, it can be proved that mixed (ℓ_q, ℓ_p)-instance optimality is also preserved when decreasing p instead of q, see Exercise 11.2.

11.2 Robustness and Quotient Property

Many reconstruction algorithms introduced in Chapter 3 have been proved to be stable — instance optimal, to be using the terminology of the previous section. In fact, Theorems 6.20, 6.24, and 6.27 showed that algorithms such as iterative hard thresholding, orthogonal matching pursuit, and compressive sampling matching pursuit are in addition robust, in the sense that estimates of the type

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x} + \mathbf{e})\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(\mathbf{x})_1 + D\|\mathbf{e}\|_2 \tag{11.6}$$

are valid for all $\mathbf{x} \in \mathbb{C}^N$ and all $\mathbf{e} \in \mathbb{C}^m$. The ℓ_2-norm $\|\mathbf{e}\|_2$ of the error between the ideal measurement $\mathbf{A}\mathbf{x}$ and the inaccurate measurement $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$

came into play in these results, but we will also investigate robustness estimates (11.6) where other norms $\|\mathbf{e}\|$ are used. We stress that running iterative hard thresholding or compressive sampling matching pursuit presents the drawback that an estimation of the targeted sparsity s is required. Running the inequality-constrained ℓ_1-minimization, on the other hand, necessitates an estimation not of the targeted sparsity but of the measurement error. This can also appear as a significant drawback. Moreover, it does not lead to robustness estimates exactly similar to (11.6). Precisely, let $\mathbf{A}$ be the realization of a renormalized $m \times N$ subgaussian matrix with $m \geq c\, s \ln(eN/m)$, and let

$$\Delta_{1,\eta}(\mathbf{y}) := \operatorname{argmin}\{\|\mathbf{z}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta\} \tag{11.7}$$

be the output of the inequality-constrained ℓ_1-minimization. Then, with high probability, for any $1 \leq p \leq 2$, the robust estimate takes the form

$$\|\mathbf{x} - \Delta_{1,\eta}(\mathbf{A}\mathbf{x} + \mathbf{e})\|_p \leq \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1 + D s^{1/p-1/2} \eta, \tag{11.8}$$

valid for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$ with $\|\mathbf{e}\|_2 \leq \eta$. Thus, setting

$$s_* := s_*(m, N) := \frac{m}{\ln(eN/m)},$$

we derive that, for any $1 \leq p \leq 2$, if $s \leq s_*/c$, then

$$\|\mathbf{x} - \Delta_{1,\eta}(\mathbf{A}\mathbf{x} + \mathbf{e})\|_p \leq \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1 + D' s_*^{1/p-1/2} \eta$$

holds for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$ with $\|\mathbf{e}\|_2 \leq \eta$. The purpose of this section and the next one is to show that such robustness results are also achieved by using the equality-constrained ℓ_1-minimization given by

$$\Delta_1(\mathbf{y}) := \operatorname{argmin}\{\|\mathbf{z}\|_1 \text{ subject to } \mathbf{A}\mathbf{z} = \mathbf{y}\}, \tag{11.9}$$

without the need to quantify the ℓ_2-norm η of the measurement error beforehand. The measurement process involves Gaussian and subgaussian matrices. These matrices, introduced in Definition 9.1, are required to have entries with variance 1. Here, the measurement matrices are renormalized to have entries with variance $1/m$. The first main result pertains to the Gaussian case.

Theorem 11.9. *There exist absolute constants $c_1, c_2, c_3, C, D > 0$ such that, for any $1 \leq p \leq 2$, if $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ where $\mathbf{A}$ is an $m \times N$ Gaussian matrix, then with probability at least $1 - 3\exp(-c_1 m)$, the ℓ_p-error estimates*

$$\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_p \leq \frac{C}{s^{1-1/p}} \sigma_s(\mathbf{x})_1 + D s_*^{1/p-1/2} \|\mathbf{e}\|_2 \tag{11.10}$$

are valid for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, provided

$$N \geq c_2 m \qquad \text{and} \qquad s \leq c_3 s_* = \frac{c_3 m}{\ln(eN/m)}.$$

The second main result concerns the more general subgaussian matrices. In this case, the ℓ_2-norm on the measurement error has to be slightly adjusted to

$$\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)} := \max\left\{\|\mathbf{e}\|_2, \sqrt{\ln(eN/m)}\|\mathbf{e}\|_\infty\right\}.$$

Theorem 11.10. *For any $1 \le p \le 2$, if $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ where $\mathbf{A}$ is an $m \times N$ subgaussian matrix with symmetric entries, then there exist constants $c_1, c_2, c_3, C, D > 0$ depending only on the subgaussian distributions such that, with probability at least $1 - 5\exp(-c_1 m)$, the ℓ_p-error estimates*

$$\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_p \le \frac{C}{s^{1-1/p}}\sigma_s(\mathbf{x})_1 + D s_*^{1/p-1/2}\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)} \quad (11.11)$$

are valid for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, provided

$$N \ge c_2 m \qquad \text{and} \qquad s \le c_3 s_* = \frac{c_3 m}{\ln(eN/m)}.$$

The fundamental tool for establishing these theorems is a new property of the measurement matrix called the *quotient property*. In this section, we show that the estimates (11.10) and (11.11) are implied by the quotient property, and in the next section we establish the quotient property for random matrices.

Definition 11.11. *Given $q \ge 1$, a measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ is said to have the* ℓ_q-quotient property *with constant d relative to a norm $\|\cdot\|$ on $\mathbb{C}^m$ if, for all $\mathbf{e} \in \mathbb{C}^m$, there exits $\mathbf{u} \in \mathbb{C}^N$ with*

$$\mathbf{A}\mathbf{u} = \mathbf{e} \quad \text{and} \quad \|\mathbf{u}\|_q \le d\, s_*^{1/q-1/2}\|\mathbf{e}\|,$$

where $s_* := m/\ln(eN/m)$.

We point out that the quotient property is a natural assumption to make, since it is implied by the prospective robustness estimate

$$\|\mathbf{x} - \Delta_1(\mathbf{A}\mathbf{x} + \mathbf{e})\|_q \le \frac{C}{s^{1-1/q}}\sigma_s(\mathbf{x})_1 + D s_*^{1/q-1/2}\|\mathbf{e}\|. \quad (11.12)$$

Indeed, setting $\mathbf{x} = 0$ in (11.12) gives $\|\Delta_1(\mathbf{e})\|_q \le D s_*^{1/q-1/2}\|\mathbf{e}\|$. This implies — if $q = 1$, it is equivalent to — the ℓ_q-quotient property by taking $\mathbf{u} = \Delta_1(\mathbf{e})$. The ℓ_1-quotient property asserts that the image under $\mathbf{A}$ of the ℓ_1-ball of radius $d\sqrt{s_*}$ covers the unit ball relative to $\|\cdot\|$. The terminology *quotient property* is explained by a reformulation involving the quotient norm of the set $[\mathbf{e}] = \mathbf{u} + \ker\mathbf{A}$ of preimages of a vector $\mathbf{e} = \mathbf{A}\mathbf{u} \in \mathbb{C}^m$, i.e.,

$$\|[\mathbf{e}]\|_{\ell_q/\ker\mathbf{A}} := \inf\left\{\|\mathbf{u} + \mathbf{v}\|_q,\, \mathbf{v} \in \ker\mathbf{A}\right\} = \inf\left\{\|\mathbf{z}\|_q,\, \mathbf{A}\mathbf{z} = \mathbf{e}\right\}.$$

Thus, the ℓ_q-quotient property is equivalent to

$$\|[\mathbf{e}]\|_{\ell_q/\ker\mathbf{A}} \le d s_*^{1/q-1/2}\|\mathbf{e}\| \quad \text{for all } \mathbf{e} \in \mathbb{C}^m.$$

Another reformulation, used in Section 11.3 to establish the quotient property for random matrices, involves the dual norm of the norm $\|\cdot\|$, but is not needed at this point. The rest of this section is of a deterministic nature, and Theorems 11.9 and 11.10 will become simple consequences of Theorem 11.12 below as soon as we verify that its two hypotheses hold with high probability for random matrices. Note that the first hypothesis — the robust null space property— is already acquired. Incidentally, we point out that the robust null space property is also a natural assumption to make, since it is necessary for prospective estimate (11.12), as we can see by setting $\mathbf{x} = \mathbf{v} \in \mathbb{C}^N$ and $\mathbf{e} = -\mathbf{A}\mathbf{v} \in \mathbb{C}^m$ — see also Remark 4.24.

Theorem 11.12. *Given $s_* := m/\ln(eN/m)$, if a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies*

- *the ℓ_2-robust null space property of order $c\,s_*$ with constants $0 < \rho < 1$ and $\tau > 0$ relative to a norm $\|\cdot\|$,*
- *the ℓ_1-quotient property with constant d relative to the norm $\|\cdot\|$,*

then, for all $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$,

$$\|\mathbf{x} - \Delta_1(\mathbf{A}\mathbf{x} + \mathbf{e})\|_q \le \frac{C}{s^{1-1/q}}\sigma_s(\mathbf{x})_1 + D s_*^{1/q-1/2}\|\mathbf{e}\|, \qquad 1 \le q \le 2,$$

whenever $s \le c\,s_*$. The constants C and D depend only on ρ, τ, c, and d.

The next two lemmas account for Theorem 11.12. The first lemma asserts that the mixed instance optimality and the simultaneous quotient property— to be defined below — together yield the desired robustness estimates. The second lemma asserts that robust null space property and ℓ_1-quotient property together the yield simultaneous quotient property. Let us now introduce the notion of simultaneous quotient property.

Definition 11.13. *Given $q \ge 1$, a matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ is said to have the* simultaneous (ℓ_q, ℓ_1)-quotient property *with constants d and d' relative to a norm $\|\cdot\|$ on $\mathbb{C}^m$ if, for all $\mathbf{e} \in \mathbb{C}^m$, there exits $\mathbf{u} \in \mathbb{C}^N$ with*

$$\mathbf{A}\mathbf{u} = \mathbf{e} \quad \text{and} \quad \begin{cases} \|\mathbf{u}\|_q \le d\, s_*^{1/q-1/2}\|\mathbf{e}\|, \\ \|\mathbf{u}\|_1 \le \quad d'\, s_*^{1/2}\|\mathbf{e}\|. \end{cases}$$

The two lemmas mentioned above formally read as follows.

Lemma 11.14. *Given $q \ge 1$, if a measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ and a reconstruction map Δ are such that*

- *$(\mathbf{A}, \Delta)$ is a mixed (ℓ_q, ℓ_1)-instance optimal pair of order $s \le c\,s_*$ with constant C,*
- *$\mathbf{A}$ has the simultaneous (ℓ_q, ℓ_1)-quotient property with constants d and d' relative to a norm $\|\cdot\|$,*

then, for all $\mathbf{x} \in \mathbb{C}^N$ *and* $\mathbf{e} \in \mathbb{C}^m$,

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x} + \mathbf{e})\|_q \le \frac{C}{s^{1-1/q}} \sigma_s(\mathbf{x})_1 + D s_*^{1/q-1/2} \|\mathbf{e}\|, \qquad D := Cd' + d.$$

Proof. For $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$, the simultaneous (ℓ_q, ℓ_1)-quotient property ensures the existence of $\mathbf{u} \in \mathbb{C}^N$ satisfying

$$\mathbf{A}\mathbf{u} = \mathbf{e} \quad \text{and} \quad \begin{cases} \|\mathbf{u}\|_q \le d\, s_*^{1/q-1/2} \|\mathbf{e}\|, \\ \|\mathbf{u}\|_1 \le \quad d'\, s_*^{1/2} \|\mathbf{e}\|. \end{cases} \tag{11.13}$$

Using the instance optimality, we then derive

$$\begin{aligned} \|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x} + \mathbf{e})\|_q = \|\mathbf{x} - \Delta(\mathbf{A}(\mathbf{x} + \mathbf{u}))\|_q &\le \|\mathbf{x} + \mathbf{u} - \Delta(\mathbf{A}(\mathbf{x} + \mathbf{u}))\|_q + \|\mathbf{u}\|_q \\ &\le \frac{C}{s^{1-1/q}} \sigma_s(\mathbf{x} + \mathbf{u})_1 + \|\mathbf{u}\|_q \\ &\le \frac{C}{s^{1-1/q}} (\sigma_s(\mathbf{x})_1 + \|\mathbf{u}\|_1) + \|\mathbf{u}\|_q. \end{aligned}$$

Substituting the inequalities of (11.13) into the latter yields the result. □

Lemma 11.15. *Given* $q \ge 1$ *and a norm* $\|\cdot\|$ *on* $\mathbb{C}^m$*, if a measurement matrix* $\mathbf{A} \in \mathbb{C}^{m \times N}$ *satisfies*

- *the* ℓ_q*-robust null space property of order* $c\, s_*$ *with constants* $\rho > 0$ *and* $\tau > 0$ *relative to* $s_*^{1/q-1/2} \|\cdot\|$,
- *the* ℓ_1*-quotient property with constant* d *relative to* $\|\cdot\|$,

then the matrix $\mathbf{A}$ *also satisfies the simultaneous* (ℓ_q, ℓ_1)*-quotient property relative to* $\|\cdot\|$ *with constants* $D := (1+\rho)d/c^{1-1/q} + \tau$ *and* $D' := d$.

Proof. Let us consider a vector $\mathbf{e} \in \mathbb{C}^m$. By the ℓ_1-quotient property, there exists $\mathbf{u} \in \mathbb{C}^N$ such that $\mathbf{A}\mathbf{u} = \mathbf{e}$ and $\|\mathbf{u}\|_1 \le d s_*^{1/2} \|\mathbf{e}\|$. Next, we establish the estimate $\|\mathbf{u}\|_q \le D s_*^{1/q-1/2} \|\mathbf{e}\|$ for some constant D. For an index set S of $c\, s_*$ largest entries of $\mathbf{u}$ in modulus, we first use Proposition 2.3 to derive

$$\|\mathbf{u}_{\overline{S}}\|_q \le \frac{1}{(c\, s_*)^{1-1/q}} \|\mathbf{u}\|_1.$$

We then use the ℓ_q-robust null space property of order $c\, s_*$ to write

$$\|\mathbf{u}_S\|_q \le \frac{\rho}{(c\, s_*)^{1-1/q}} \|\mathbf{u}_{\overline{S}}\|_1 + \tau s_*^{1/q-1/2} \|\mathbf{A}\mathbf{u}\| \le \frac{\rho}{(c\, s_*)^{1-1/q}} \|\mathbf{u}\|_1 + \tau s_*^{1/q-1/2} \|\mathbf{e}\|.$$

It follows that

$$\|\mathbf{u}\|_q = \|\mathbf{u}_{\overline{S}} + \mathbf{u}_S\|_q \le \|\mathbf{u}_{\overline{S}}\|_q + \|\mathbf{u}_S\|_q \le \frac{1+\rho}{(c\, s_*)^{1-1/q}} \|\mathbf{u}\|_1 + \tau s_*^{1/q-1/2} \|\mathbf{e}\|.$$

The estimate $\|\mathbf{u}\|_1 \le d s_*^{1/2} \|\mathbf{e}\|$ yields the desired result. □

Now that Lemmas 11.14 and 11.15 have been established, Theorem 11.12 can be derived with the help of results from Chapter 4.

Proof (of Theorem 11.12). We assume that $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the ℓ_2-robust null space property of order $c\, s_*$ with constant $0 < \rho < 1$ and $\tau > 0$ relative to $\|\cdot\|$, as well as the ℓ_1-quotient property with constant d relative to $\|\cdot\|$. Then, for any $1 \le q \le 2$, Definition 4.20 and the considerations afterwards ensure that $\mathbf{A}$ satisfies the ℓ_q-robust null space property of order $c\, s_*$ with constant $0 < \rho < 1$ and $\tau\, c^{1/q-1/2} > 0$ relative to $s_*^{1/q-1/2}\|\cdot\|$. Lemma 11.15 now implies that $\mathbf{A}$ satisfies the simultaneous (ℓ_q, ℓ_1)-quotient property with constants $D = (1+\rho)d/c^{1-1/q} + \tau\, c^{1/q-1/2} \le (1+\rho)d/\min\{1,c\}^{1/2} + \tau\, \max\{1,c\}^{1/2}$ and $D' = d$. Next, for any $1 \le q \le 2$, Theorem 4.23 ensures that the pair $(\mathbf{A}, \Delta_1)$ is mixed (ℓ_q, ℓ_1)-instance optimal of any order $s \le c\, s_*$ with constant $C = (1+\rho)^2/(1+\rho)$. Lemma 11.14 finally yields the desired estimate with constants depending only on ρ, τ, c, and d. □

11.3 Quotient Property for Random Matrices

In this section, we prove the ℓ_1-quotient property for certain random matrices. First, we focus on Gaussian matrices, where the ℓ_1-quotient property holds relative to the ℓ_2-norm. Second, we analyze general subgaussian random matrices, where the ℓ_1-quotient property holds relative to a slight alteration of the ℓ_2-norm. The basis of both arguments is a convenient reformulation of the quotient property involving the dual norm of a norm $\|\cdot\|$ (see Definition A.4), i.e.,

$$\|\mathbf{e}\|_* := \sup_{\|\mathbf{y}\|=1} |\langle \mathbf{y}, \mathbf{e}\rangle|, \qquad \mathbf{e} \in \mathbb{C}^m.$$

Lemma 11.16. *For $q \ge 1$, a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ has the ℓ_q-quotient property with constant d relative to a norm $\|\cdot\|$ if and only if*

$$\|\mathbf{e}\|_* \le d\, s_*^{1/q-1/2} \|\mathbf{A}^*\mathbf{e}\|_{q^*} \qquad \text{for all } \mathbf{e} \in \mathbb{C}^m, \tag{11.14}$$

where $s_* := \dfrac{m}{\ln(eN/m)}$ and where $q^* := \dfrac{q}{q-1}$ is the conjugate exponent of q.

Proof. Let us assume that $\mathbf{A}$ has the ℓ_q-quotient property. For $\mathbf{e} \in \mathbb{C}^m$, we have $\|\mathbf{e}\|_* = \langle \mathbf{y}, \mathbf{e}\rangle$ for some $\mathbf{y} \in \mathbb{C}^m$ with $\|\mathbf{y}\| = 1$. The vector $\mathbf{y}$ can be written as $\mathbf{y} = \mathbf{A}\mathbf{u}$ for some $\mathbf{u} \in \mathbb{C}^N$ with $\|\mathbf{u}\|_q \le d s_*^{1/q-1/2}$. We deduce (11.14) from

$$\|\mathbf{e}\|_* = \langle \mathbf{A}\mathbf{u}, \mathbf{e}\rangle = \langle \mathbf{u}, \mathbf{A}^*\mathbf{e}\rangle \le \|\mathbf{u}\|_q \|\mathbf{A}^*\mathbf{e}\|_{q^*} \le d s_*^{1/q-1/2} \|\mathbf{A}^*\mathbf{e}\|_{q^*}.$$

Conversely, let us assume that (11.14) holds. We consider the case $q > 1$ first. For $\mathbf{e} \in \mathbb{C}^m \backslash \{0\}$ — the case $\mathbf{e} = 0$ is clear — we let $\mathbf{u} \in \mathbb{C}^N \backslash \{0\}$ be a minimizer

of $\|\mathbf{z}\|_q$ subject to $\mathbf{A}\mathbf{z} = \mathbf{e}$. Our goal is to show that $\|\mathbf{u}\|_q \le d s_*^{1/q-1/2}\|\mathbf{e}\|$. Let us fix a vector $\mathbf{v} \in \ker \mathbf{A}$. Given $\tau = te^{i\theta}$ with $t > 0$ small enough to have $\mathbf{u} + \tau\mathbf{v} \neq 0$, we consider the vector $\mathbf{w}^\tau \in \mathbb{C}^N$ whose entries are given by

$$w_j^\tau := \frac{\operatorname{sgn}(u_j + \tau v_j)\,|u_j + \tau v_j|^{q-1}}{\|\mathbf{u} + \tau\mathbf{v}\|_q^{q-1}}, \qquad j \in [N].$$

We notice that $\langle \mathbf{w}^\tau, \mathbf{u} + \tau\mathbf{v}\rangle = \|\mathbf{u} + \tau\mathbf{v}\|_q$ with $\|\mathbf{w}^\tau\|_{q^*} = 1$. We also notice that the vector $\mathbf{w} := \lim_{\tau\to 0} \mathbf{w}^\tau$ is well-defined and independent of $\mathbf{v}$, thanks to the assumption $q > 1$. It satisfies $\langle \mathbf{w}, \mathbf{u}\rangle = \|\mathbf{u}\|_q$ with $\|\mathbf{w}\|_{q^*} = 1$. Then the definition of $\mathbf{u}$ yields

$$\operatorname{Re}\langle \mathbf{w}^\tau, \mathbf{u}\rangle \le \|\mathbf{u}\|_q \le \|\mathbf{u} + \tau\mathbf{v}\|_q = \operatorname{Re}\langle \mathbf{w}^\tau, \mathbf{u} + \tau\mathbf{v}\rangle,$$

so that $\operatorname{Re}\langle \mathbf{w}^\tau, e^{i\theta}\mathbf{v}\rangle \ge 0$. Taking the limit as t tends to zero, we obtain $\operatorname{Re}\langle \mathbf{w}, e^{i\theta}\mathbf{v}\rangle \ge 0$ independently of θ, hence $\langle \mathbf{w}, \mathbf{v}\rangle = 0$. Since this is true for all $\mathbf{v} \in \ker \mathbf{A}$, we have $\mathbf{w} \in (\ker \mathbf{A})^\perp = \operatorname{ran} \mathbf{A}^*$. Therefore, we can write $\mathbf{w} = \mathbf{A}^*\mathbf{y}$ for some $\mathbf{y} \in \mathbb{C}^m$. According to (11.14), we have $\|\mathbf{y}\|_* \le d s_*^{1/q-1/2}$. It now follows that

$$\|\mathbf{u}\|_q = \langle \mathbf{w}, \mathbf{u}\rangle = \langle \mathbf{A}^*\mathbf{y}, \mathbf{u}\rangle = \langle \mathbf{y}, \mathbf{A}\mathbf{u}\rangle = \langle \mathbf{y}, \mathbf{e}\rangle \le \|\mathbf{y}\|_*\|\mathbf{e}\| \le d s_*^{1/q-1/2}\|\mathbf{e}\|.$$

This establishes the ℓ_q-quotient property in the case $q > 1$. We use some limiting arguments for the case $q = 1$. Precisely, let us consider a sequence of numbers $q_n > 1$ converging to 1. For each n, in view of $\|\mathbf{A}^*\mathbf{e}\|_\infty \le \|\mathbf{A}^*\mathbf{e}\|_{q_n^*}$, the property (11.14) for $q = 1$ implies a similar property for $q = q_n$ provided d is changed to $d s_*^{1/q_n^*}$. Given $\mathbf{e} \in \mathbb{C}^m$, the preceding argument yields a vector $\mathbf{u}^n \in \mathbb{C}^N$ with $\mathbf{A}\mathbf{u}^n = \mathbf{e}$ and $\|\mathbf{u}^n\|_{q_n} \le d s_*^{1/q_n^*} s_*^{1/q_n - 1/2}\|\mathbf{e}\| = d s_*^{1/2}\|\mathbf{e}\|$. Since the sequence $(\mathbf{u}^n)$ is bounded in ℓ_∞-norm, it has a convergent subsequence. Denoting by $\mathbf{u} \in \mathbb{C}^N$ its limit, we obtain $\mathbf{A}\mathbf{u} = \mathbf{e}$ and $\|\mathbf{u}\|_1 \le d s_*^{1/2}\|\mathbf{e}\|$ by letting n tend to infinity. This settles the case $q = 1$. □

Remark 11.17. In the case of a real matrix $\mathbf{A}$, we can also consider a real version of the quotient property, i.e., for all $\mathbf{e} \in \mathbb{R}^m$, there exits $\mathbf{u} \in \mathbb{R}^N$ with

$$\mathbf{A}\mathbf{u} = \mathbf{e} \quad \text{and} \quad \|\mathbf{u}\|_q \le d\, s_*^{1/q-1/2}\|\mathbf{e}\|.$$

The real and complex versions are in fact equivalent, up to a possible change of the constant d. A real version of Lemma 11.16 also holds. Exercise 11.6 asks for a detailed verification of these statements. When we establish the quotient property for random matrices, we actually prove the real analog of (11.14), i.e., $\|\mathbf{e}\|_* \le d s_*^{1/q-1/2}\|\mathbf{A}^*\mathbf{e}\|_{q^*}$ for all $\mathbf{e} \in \mathbb{R}^m$.

Gaussian Matrices

We are now in the position to prove the ℓ_1-quotient property for Gaussian matrices, and then to deduce Theorem 11.9. We point out that the numerical constants in the following theorems have not been optimized, they have simply been chosen for convenience.

Theorem 11.18. *For $N \geq 2m$, if $\mathbf{A}$ is an $m \times N$ Gaussian matrix, then the matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ has the ℓ_1-quotient property with constant $D = 34$ relative to the ℓ_2-norm with probability at least*

$$1 - \exp(-m/100).$$

Proof. According to Lemma 11.16 and Remark 11.17, we need to prove that

$$\mathbb{P}\big(\|\mathbf{e}\|_2 \leq D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\|_\infty \text{ for all } \mathbf{e} \in \mathbb{R}^m\big) \geq 1 - \exp(-m/100). \tag{11.15}$$

To this end, we separate two cases: $2m \leq N < Cm$ and $N \geq Cm$, where $C = 165^6$ for reason that will become apparent later. In the first case, by considering the renormalized matrix $\mathbf{B} := \sqrt{m/N}\widetilde{\mathbf{A}}^* = \mathbf{A}^*/\sqrt{N} \in \mathbb{R}^{N\times m}$, we notice that the existence of $\mathbf{e} \in \mathbb{R}^m$ such that $\|\mathbf{e}\|_2 > D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\|_\infty$ implies

$$\|\mathbf{e}\|_2 > D\sqrt{\frac{s_*N}{m}}\|\mathbf{B}\mathbf{e}\|_\infty \geq D\sqrt{\frac{s_*}{m}}\|\mathbf{B}\mathbf{e}\|_2 \geq \frac{D}{\sqrt{\ln(eN/m)}}\sigma_{\min}(\mathbf{B})\|\mathbf{e}\|_2.$$

In view of $N < Cm$, we derive

$$\sigma_{\min}(\mathbf{B}) < \frac{\sqrt{\ln(eC)}}{D} = 1 - \sqrt{\frac{m}{N}} - t,$$

If $D \geq 6\sqrt{\ln(eC)}$ (which is satisfied for $D = 34$ and $C = 165^6$), we have

$$t := 1 - \sqrt{\frac{m}{N}} - \frac{\sqrt{\ln(eC)}}{D} \geq 1 - \sqrt{\frac{1}{2}} - \frac{1}{6} \geq \frac{1}{10}.$$

Calling upon Theorem 9.24, we obtain

$$\begin{aligned}\mathbb{P}\big(\|\mathbf{e}\|_2 > 34\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\|_\infty \text{ for some } \mathbf{e} \in \mathbb{R}^m\big) &\leq \mathbb{P}\bigg(\sigma_{\min}(\mathbf{B}) < 1 - \sqrt{\frac{m}{N}} - t\bigg)\\ &\leq \exp\bigg(-\frac{t^2N}{2}\bigg) = \exp\bigg(-\frac{N}{200}\bigg) \leq \exp\bigg(-\frac{m}{100}\bigg).\end{aligned}$$

This establishes (11.15) in the case $2m \leq N < Cm$. The case $N \geq Cm$ is more delicate. Here, with $D = 8$, we will prove the stronger statement

$$\mathbb{P}\big(\|\mathbf{e}\|_2 > D\sqrt{s_*}|\!|\!|\widetilde{\mathbf{A}}^*\mathbf{e}|\!|\!| \text{ for some } \mathbf{e} \in \mathbb{R}^m\big) \leq \exp\big(-m/3\big). \tag{11.16}$$

The norm appearing in this statement is defined by

$$|\!|\!|\mathbf{z}|\!|\!| := \frac{1}{2h}\sum_{\ell=1}^{2h}\|\mathbf{z}_{T_\ell}\|_\infty, \qquad \mathbf{z} \in \mathbb{R}^N, \tag{11.17}$$

for some integer $1 \leq h \leq N/2$ and some fixed partition $T_1, \ldots, T_{2h}$ of $[N]$. Each set T_ℓ can be chosen to have size $\lfloor N/h \rfloor$ or $\lfloor N/h \rfloor + 1$. The straightforward

inequality $|||\mathbf{z}||| \le \|\mathbf{z}\|_\infty$ explains why (11.16) implies (11.15). Another key property of the norm defined in (11.17) is the existence, for any $\mathbf{z} \in \mathbb{R}^N$, of a subset L of $[2h]$ of size h such that

$$\|\mathbf{z}_{T_\ell}\|_\infty \le 2|||\mathbf{z}||| \quad \text{for all } \ell \in L.$$

Indeed, the inequality

$$|||\mathbf{z}||| \ge \frac{1}{2h} \sum_{\ell:\|\mathbf{z}_{T_\ell}\|_\infty > 2|||\mathbf{z}|||} \|\mathbf{z}_{T_\ell}\|_\infty \ge \frac{1}{h} \operatorname{card}(\{\ell : \|\mathbf{z}_{T_\ell}\|_\infty > 2|||\mathbf{z}|||\})\, |||\mathbf{z}|||$$

implies $\operatorname{card}(\{\ell : \|\mathbf{z}_{T_\ell}\|_\infty > 2|||\mathbf{z}|||\}) \le h$, i.e., $\operatorname{card}(\{\ell : \|\mathbf{z}_{T_\ell}\|_\infty \le 2|||\mathbf{z}|||\}) \ge h$. Therefore, for a fixed $\mathbf{e} \in \mathbb{R}^m$ and with $d := D/2$, we have

$$\begin{aligned}
&\mathbb{P}\big(\|\mathbf{e}\|_2 > d\sqrt{s_*}|||\widetilde{\mathbf{A}}^*\mathbf{e}|||\big) \\
&\quad \le \mathbb{P}\bigg(\|(\widetilde{\mathbf{A}}^*\mathbf{e})_{T_\ell}\|_\infty < \frac{2\|\mathbf{e}\|_2}{d\sqrt{s_*}} \text{ for all } \ell \text{ in some } L \subseteq [2h], \operatorname{card}(L) = h\bigg) \\
&\quad \le \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \mathbb{P}\bigg(\max_{j\in T_\ell} \Big|(\widetilde{\mathbf{A}}^*\mathbf{e})_j\Big| < \frac{2\|\mathbf{e}\|_2}{d\sqrt{s_*}} \text{ for all } \ell \in L\bigg) \\
&\quad = \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \mathbb{P}\bigg(\Big|(\widetilde{\mathbf{A}}^*\mathbf{e})_j\Big| < \frac{2\|\mathbf{e}\|_2}{d\sqrt{s_*}} \text{ for all } j \in \cup_{\ell\in L} T_\ell\bigg) \\
&\quad = \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \prod_{j\in\cup_{\ell\in L}T_\ell} \mathbb{P}\bigg(\Big|(\widetilde{\mathbf{A}}^*\mathbf{e})_j\Big| < \frac{2\|\mathbf{e}\|_2}{d\sqrt{s_*}}\bigg).
\end{aligned}$$

For each $j \in \cup_{\ell\in L} T_\ell$, we notice that $(\widetilde{\mathbf{A}}^*\mathbf{e})_j = \sum_{i=1}^m a_{i,j}e_i/\sqrt{m}$ is a zero-mean Gaussian random variable with variance $\|\mathbf{e}\|_2^2/m$. Therefore, if g represents a standard normal random variable, we obtain

$$\begin{aligned}
\mathbb{P}\big(\|\mathbf{e}\|_2 > d\sqrt{s_*}|||\widetilde{\mathbf{A}}^*\mathbf{e}|||\big) &\le \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \prod_{j\in\cup_{\ell\in L}T_\ell} \mathbb{P}\bigg(|g| < \frac{2\sqrt{m/s_*}}{d}\bigg) \\
&= \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \bigg(1 - \mathbb{P}\bigg(|g| \ge \frac{2\sqrt{m/s_*}}{d}\bigg)\bigg)^{\operatorname{card}(\cup_{\ell\in L}T_\ell)} \\
&\le \binom{2h}{h}\bigg(1 - \mathbb{P}\bigg(|g| \ge \frac{2\sqrt{m/s_*}}{d}\bigg)\bigg)^{N/2}. \qquad (11.18)
\end{aligned}$$

At this point, we bound from below the tail of a standard normal variable as

$$\mathbb{P}\Big(|g| \ge \frac{2\sqrt{m/s_*}}{d}\Big) = \sqrt{\frac{2}{\pi}} \int_{2\sqrt{m/s_*}/d}^{\infty} \exp(-t^2/2)dt$$
$$\ge \sqrt{\frac{2}{\pi}} \int_{2\sqrt{m/s_*}/d}^{4\sqrt{m/s_*}/d} \exp(-t^2/2)dt \ge \sqrt{\frac{2}{\pi}} \frac{2\sqrt{m/s_*}}{d} \exp\Big(-\frac{8m/s_*}{d^2}\Big)$$
$$\ge \frac{\sqrt{8/\pi}}{d} \exp\Big(-\frac{8}{d^2} \ln\Big(\frac{eN}{m}\Big)\Big) = \frac{\sqrt{8/\pi}}{d} \Big(\frac{m}{eN}\Big)^{8/d^2}. \tag{11.19}$$

Substituting (11.19) into (11.18), while using the inequalities $\binom{n}{k} \le \big(\frac{en}{k}\big)^k$ (see Lemma C.5) and $1 - x \le \exp(-x)$, we derive

$$\mathbb{P}\big(\|\mathbf{e}\|_2 > d\sqrt{s_*}|||\widetilde{\mathbf{A}}^*\mathbf{e}|||\big) \le (2e)^h \exp\Big(-\frac{\sqrt{8/\pi}}{d}\Big(\frac{m}{eN}\Big)^{8/d^2}\Big)^{N/2}$$
$$= \exp\Big(\ln(2e)h - \frac{\sqrt{2/\pi}}{de^{8/d^2}}\, m^{8/d^2} N^{1-8/d^2}\Big). \tag{11.20}$$

We now use covering arguments to deduce a probability estimate applied to all $\mathbf{e} \in \mathbb{R}^m$ simultaneously. According to Proposition C.3, with $0 < \delta < 1$ to be chosen later, we can find a δ-covering $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ of the unit sphere of ℓ_2^m with cardinality $n \le (1 + 2/\delta)^m$. Let us suppose that there exists $\mathbf{e} \in \mathbb{R}^m$ with $\|\mathbf{e}\|_2 > D\sqrt{s_*}|||\widetilde{\mathbf{A}}^*\mathbf{e}|||$. Without loss of generality, we may assume that $\|\mathbf{e}\|_2 = 1$, hence $\|\mathbf{e} - \mathbf{e}_i\|_2 \le \delta$ for some $i \in [n]$. It follows that

$$D\sqrt{s_*}|||\widetilde{\mathbf{A}}^*\mathbf{e}_i||| \le D\sqrt{s_*}|||\widetilde{\mathbf{A}}^*\mathbf{e}||| + D\sqrt{s_*}|||\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i)|||$$
$$< 1 + D\frac{\sqrt{s_*}}{2h} \sum_{\ell=1}^{2h} \|(\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i))_{T_\ell}\|_\infty$$
$$\le 1 + D\frac{\sqrt{s_*}}{2h} \sum_{\ell=1}^{2h} \|(\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i))_{T_\ell}\|_2$$
$$\le 1 + D\sqrt{\frac{s_*}{2h}} \|\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i)\|_2.$$

Applying Theorem 9.24 to the renormalized matrix $\mathbf{B} = \mathbf{A}^*/\sqrt{N}$, we obtain

$$\mathbb{P}\Big(\sigma_{\max}(\mathbf{B}) > 1 + 2\sqrt{\frac{m}{N}}\Big) \le \exp\Big(-\frac{m}{2}\Big). \tag{11.21}$$

Thus, in the likely case $\sigma_{\max}(\mathbf{B}) \le 1 + 2\sqrt{m/N}$, whence $\sigma_{\max}(\mathbf{B}) \le \sqrt{2}$ provided $C \ge 12 + 8\sqrt{2}$ (which is satisfied for $C = 165$[6]), we have

$$\|\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i)\|_2 \le \sigma_{\max}(\widetilde{\mathbf{A}}^*)\|\mathbf{e} - \mathbf{e}_i\|_2 = \sqrt{\frac{N}{m}}\, \sigma_{\max}(\mathbf{B})\|\mathbf{e} - \mathbf{e}_i\|_2 \le \sqrt{\frac{2N}{m}}\delta.$$

In turn, we deduce

$$d\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| = \frac{1}{2}\Big(D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\|\Big) \le \frac{1}{2}\Big(1 + D\sqrt{\frac{s_* N}{h\,m}}\,\delta\Big) \le \|\mathbf{e}_i\|_2$$

where the last equality holds because of the choice

$$\delta := \frac{1}{D}\sqrt{\frac{h}{N}}$$

and the facts that $s_* \le m$ and that $\|\mathbf{e}_i\|_2 = 1$. Summarizing the previous considerations yields

$$\begin{aligned}
&\mathbb{P}\big(\|\mathbf{e}\|_2 > D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\| \text{ for some } \mathbf{e}\in\mathbb{R}^m\big)\\
&\quad= \mathbb{P}\bigg(\|\mathbf{e}\|_2 > D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\| \text{ for some } \mathbf{e}\in\mathbb{R}^m \text{ and } \sigma_{\max}(B) > 1+2\sqrt{\frac{m}{N}}\bigg)\\
&\quad+ \mathbb{P}\bigg(\|\mathbf{e}\|_2 > D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\| \text{ for some } \mathbf{e}\in\mathbb{R}^m \text{ and } \sigma_{\max}(B) \le 1+2\sqrt{\frac{m}{N}}\bigg)\\
&\quad\le \mathbb{P}\bigg(\sigma_{\max}(B) > 1+2\sqrt{\frac{m}{N}}\bigg) + \mathbb{P}\big(\|\mathbf{e}_i\|_2 > d\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| \text{ for some } i\in[n]\big).
\end{aligned}$$

By (11.21), the first term on the right-hand side is bounded by $\exp(-N/2)$. Moreover, a union bound, the inequality $n \le (1+2/\delta)^m \le \exp(2m/\delta)$, and the probability estimate (11.20) applied to the fixed $\mathbf{e}_i \in \mathbb{R}^m$ show that the second term is bounded by

$$\exp\bigg(2D\sqrt{\frac{N}{h}}\,m + \ln(2e)h - \frac{\sqrt{2/\pi}}{de^{8/d^2}}\,m^{8/d^2}N^{1-8/d^2}\bigg).$$

We substitute the values $d = 4$ (corresponding to $D = 8$) and we make the choice $h = \lceil m^{2/3}N^{1/3}\rceil$ (so that $1 \le h \le N/2$ for the constant $C = 165^6$) to bound the second term by

$$\begin{aligned}
&\exp\bigg(16\,m^{2/3}N^{1/3} + 2\ln(2e)\,m^{2/3}N^{1/3} - \frac{1}{\sqrt{8\pi e}}m^{1/2}N^{1/2}\bigg)\\
&\quad= \exp\bigg(-\bigg[\frac{1}{\sqrt{8\pi e}} - \frac{2\ln(2e^9)}{(N/m)^{1/6}}\bigg]m^{1/2}N^{1/2}\bigg)\\
&\quad\le \exp\bigg(-\bigg[\frac{1}{\sqrt{8\pi e}} - \frac{2\ln(2e^9)}{C^{1/6}}\bigg]m^{1/2}N^{1/2}\bigg) \le \exp\bigg(-\frac{m^{1/2}N^{1/2}}{300}\bigg).
\end{aligned}$$

The choice $C = 165^6$ accounts for the last inequality. Putting the two bounds together, we obtain

$$\begin{aligned}
&\mathbb{P}\big(\|\mathbf{e}\|_2 > 8\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\| \text{ for some } \mathbf{e}\in\mathbb{R}^m\big)\\
&\qquad\le \exp\Big(-\frac{m}{2}\Big) + \exp\Big(-\frac{m^{1/2}N^{1/2}}{300}\Big) \le \exp\Big(-\frac{m}{3}\Big).
\end{aligned}$$

This establishes (11.16) in the case $N \ge Cm$, and concludes the proof. □

We now prove the main robustness estimate for Gaussian matrices.

Proof (of Theorem 11.9). Under the assumption $N \ge c_2 m$ with $c_2 := 2$, Theorem 11.18 guarantees that the matrix $\widetilde{\mathbf{A}}$ has the ℓ_1-quotient property relative to the ℓ_2-norm with probability at least $1 - \exp(-m/100)$. Moreover, the assumption $s \le c_3 s_*$ with $c_3 := 1/1400$ reads $m \ge 1400 s \ln(eN/m)$. Lemma C.6, in view of $1400/\ln(1400e) \ge 160$ and $\ln(eN/s) \ge \ln(eN/(2s))$, implies the inequality $m \ge 80(2s)\ln(eN/(2s))$. This is equivalent to

$$\frac{5m}{4} \ge 80(2s)\ln\Big(\frac{eN}{2s}\Big) + \frac{m}{4}, \qquad \text{i.e.,} \quad m \ge \frac{2}{\eta^2}(2s)\ln\Big(\frac{eN}{2s}\Big) + \frac{2}{\eta^2}\ln\Big(\frac{2}{\epsilon}\Big),$$

where $\eta := 1/\sqrt{32}$ and $\epsilon := 2\exp(-m/320)$. Theorem 9.25 implies that, with probability at least $1 - 2\exp(-m/320)$, the restricted isometry constant of the matrix $\widetilde{\mathbf{A}}$ satisfies

$$\begin{aligned}\delta_{2s} &\le 2\bigg(1 + \frac{1}{\sqrt{2\ln(eN/(2s))}}\bigg)\eta + \bigg(1 + \frac{1}{\sqrt{2\ln(eN/(2s))}}\bigg)^2 \eta^2 \\ &\le 2\bigg(1 + \frac{1}{\sqrt{2\ln(1400e)}}\bigg)\frac{1}{\sqrt{32}} + \bigg(1 + \frac{1}{\sqrt{2\ln(1400e)}}\bigg)^2 \frac{1}{32} \approx 0.489.\end{aligned}$$

In this case, Theorem 6.12 ensures that the matrix $\widetilde{\mathbf{A}}$ has the ℓ_2-robust null space property of order s. Thus, with probability at least

$$1 - \exp(-m/100) - 2\exp(m/320) \ge 1 - 3\exp(-c_1 m), \qquad c_1 := 1/320,$$

the matrix $\widetilde{\mathbf{A}}$ satisfies both the ℓ_1-quotient property relative to the ℓ_2-norm and the ℓ_2-robust null space property of order $s \le c_3 s_*$. The conclusion now follows from Theorem 11.12. □

Subgaussian Matrices

For renormalized Bernoulli matrices, the ℓ_1-quotient property relative to the ℓ_2-norm, namely

$$\text{for all } \mathbf{e} \in \mathbb{C}^m, \text{ there exits } \mathbf{u} \in \mathbb{C}^N \text{ with } \widetilde{\mathbf{A}}\mathbf{u} = \mathbf{e} \text{ and } \|\mathbf{u}\|_1 \le d\sqrt{s_*}\|\mathbf{e}\|_2,$$

cannot be true. Indeed, since such a matrix $\widetilde{\mathbf{A}}$ has entries $\tilde{a}_{i,j} = \pm 1/\sqrt{m}$, the ℓ_1-quotient property applied to the vectors $\mathbf{e}_i = [0, \dots, 0, 1, 0, \dots, 0]^\top \in \mathbb{C}^m$ would give rise to vectors $\mathbf{u} \in \mathbb{C}^N$ for which

$$1 = (\widetilde{\mathbf{A}}\mathbf{u})_i = \sum_{j=1}^N \tilde{a}_{i,j} u_j \le \frac{\|\mathbf{u}\|_1}{\sqrt{m}} \le \frac{d\sqrt{s_*}\|\mathbf{e}\|_2}{\sqrt{m}} = \frac{d}{\sqrt{\ln(eN/m)}}.$$

Thus, to obtain robustness estimates, the strategy consists in eliminating these troublesome vectors $\mathbf{e}_i$ by clipping the ℓ_2-ball around them. This explains the introduction of the norm defined, for $\alpha \ge 1$, by

$$\|\mathbf{y}\|^{(\alpha)} := \max\{\|\mathbf{y}\|_2, \alpha\|\mathbf{y}\|_\infty\}, \qquad \mathbf{y} \in \mathbb{C}^m. \tag{11.22}$$

Then the ℓ_1-quotient property relative to this norm applied to the vectors $\mathbf{e}_i = [0, \dots, 0, 1, 0, \dots, 0]^\top \in \mathbb{C}^m$ yields

$$1 \le \frac{d\sqrt{s_*}\|\mathbf{e}\|^{(\alpha)}}{\sqrt{m}} = \frac{d\alpha}{\sqrt{\ln(eN/m)}}.$$

This dictates the choice $\alpha = \sqrt{\ln(eN/m)} \ge 1$ for Bernoulli matrices. Here is a precise statement about the ℓ_1-quotient property for Bernoulli matrices, among others.

Theorem 11.19. *If $\mathbf{A}$ is an $m \times N$ matrix whose entries are independent symmetric random variables with variance 1 and fourth moment bounded by some $\mu^4 \ge 1$, and if the concentration inequality*

$$\mathbb{P}\big(\big|N^{-1}\|\mathbf{A}^*\mathbf{y}\|_2^2 - \|\mathbf{y}\|_2^2\big| > t\|\mathbf{y}\|_2^2\big) \le 2\exp(-\tilde{c}t^2N) \tag{11.23}$$

holds for all $\mathbf{y} \in \mathbb{R}^m$ and $t \in (0,1)$, then there exist constants $C, D > 0$ depending only on μ and $\tilde{c}$ such that, with probability at least

$$1 - 3\exp(-m),$$

the matrix $\widetilde{\mathbf{A}} := \frac{1}{\sqrt{m}}\mathbf{A}$ has the ℓ_1-quotient property with constant D relative to the norm $\|\cdot\|^{(\alpha)}$, $\alpha := \sqrt{\ln(eN/m)}$, provided $N \ge C\,m$.

The arguments follow the same lines as the Gaussian case. In particular, estimates from below for tail probabilities involving the dual norm of $\|\cdot\|^{(\alpha)}$ are needed. We start by comparing this dual norm to a more tractable norm.

Lemma 11.20. *For an integer $k \ge 1$, the dual norm of $\|\cdot\|^{(\sqrt{k})}$ is comparable with the norm $|\cdot|_k$ defined by*

$$|\mathbf{y}|_k := \max\left\{\sum_{\ell=1}^k \|\mathbf{y}_{B_\ell}\|_2,\ B_1, \dots, B_k \text{ form a partition of } [m]\right\},$$

in the sense that

$$\sqrt{\frac{1}{k}}\,|\mathbf{y}|_k \le \|\mathbf{y}\|_*^{(\sqrt{k})} \le \sqrt{\frac{2}{k}}\,|\mathbf{y}|_k, \qquad \mathbf{y} \in \mathbb{C}^m. \tag{11.24}$$

Proof. We define a norm on $\mathbb{C}^m \times \mathbb{C}^m$ by

$$\|(\mathbf{u}, \mathbf{v})\| := \max\big(\|\mathbf{u}\|_2, \sqrt{k}\|\mathbf{v}\|_\infty\big).$$

This makes the linear map $T : \mathbf{z} \in (\mathbb{C}^m, \|\cdot\|^{(\sqrt{k})}) \mapsto (\mathbf{z}, \mathbf{z}) \in (\mathbb{C}^m \times \mathbb{C}^m, \|(\cdot,\cdot)\|)$ an isometry from $\mathbb{C}^m$ onto $X := T(\mathbb{C}^m)$. Let us now fix a vector $\mathbf{y} \in \mathbb{C}^m$. We have

$$\|\mathbf{y}\|_*^{(\sqrt{k})} = \max_{\|\mathbf{u}\|^{(\sqrt{k})}=1} |\langle \mathbf{u}, \mathbf{y}\rangle| = \max_{\|(\mathbf{u},\mathbf{u})\|=1} |\langle T^{-1}((\mathbf{u},\mathbf{u})), \mathbf{y}\rangle| = \|\lambda\|_{X^*},$$

where we have defined the linear functional λ on X by $\lambda(\mathbf{x}) := \langle T^{-1}(\mathbf{x}), \mathbf{y}\rangle$. The Hahn–Banach extension theorem now ensures the existence of a linear functional $\widetilde{\lambda}$ defined on $\mathbb{C}^m \times \mathbb{C}^m$ such that $\widetilde{\lambda}(\mathbf{x}) = \lambda(\mathbf{x})$ for all $\mathbf{x} \in X$ and $\|\widetilde{\lambda}\|_* = \|\lambda\|_{X^*}$. This functional can be written, for some $(\mathbf{y}', \mathbf{y}'') \in \mathbb{C}^m \times \mathbb{C}^m$, as $\widetilde{\lambda}(\mathbf{u}, \mathbf{v}) = \langle (\mathbf{u}, \mathbf{v}), (\mathbf{y}', \mathbf{y}'')\rangle = \langle \mathbf{u}, \mathbf{y}'\rangle + \langle \mathbf{v}, \mathbf{y}''\rangle$ for all $(\mathbf{u}, \mathbf{v}) \in \mathbb{C}^m \times \mathbb{C}^m$. The identity $\widetilde{\lambda}(T(\mathbf{z})) = \lambda(T(\mathbf{z}))$, i.e., $\langle \mathbf{z}, \mathbf{y}' + \mathbf{y}''\rangle = \langle \mathbf{z}, \mathbf{y}\rangle$, for all $\mathbf{z} \in \mathbb{C}^m$ yields $\mathbf{y}' + \mathbf{y}'' = \mathbf{y}$. Moreover, the equality $\|\widetilde{\lambda}\|_* = \|\mathbf{y}'\|_2 + \|\mathbf{y}''\|_1/\sqrt{k}$ yields $\|\mathbf{y}\|_*^{(\sqrt{k})} = \|\mathbf{y}'\|_2 + \|\mathbf{y}''\|_1/\sqrt{k}$. Now, choosing optimal partitions $B_1', \ldots, B_k'$ and $B_1'', \ldots, B_k''$ of $[m]$, we observe that

$$|\mathbf{y}'|_k = \sum_{\ell=1}^{k} \|\mathbf{y}'_{B_\ell'}\|_2 \le \sqrt{k} \sqrt{\sum_{\ell=1}^{k} \|\mathbf{y}'_{B_\ell'}\|_2^2} = \sqrt{k}\|\mathbf{y}'\|_2,$$

$$|\mathbf{y}''|_k = \sum_{\ell=1}^{k} \|\mathbf{y}''_{B_\ell''}\|_2 \le \sum_{\ell=1}^{k} \|\mathbf{y}''_{B_\ell''}\|_1 = \|\mathbf{y}''\|_1.$$

It follows that

$$|\mathbf{y}|_k = |\mathbf{y}' + \mathbf{y}''|_k \le |\mathbf{y}'|_k + |\mathbf{y}''|_k \le \sqrt{k}\left(\|\mathbf{y}'\|_2 + \|\mathbf{y}''\|_1/\sqrt{k}\right) = \sqrt{k}\|\mathbf{y}\|_*^{(\sqrt{k})}.$$

This proves the leftmost inequality of (11.24).

For the second inequality, given a fixed vector $\mathbf{y} \in \mathbb{C}^m$, we consider a vector $\mathbf{u} \in \mathbb{C}^m$ with $\|\mathbf{u}\|^{(\sqrt{k})} = 1$ such that $\|\mathbf{y}\|_*^{(\sqrt{k})} = \langle \mathbf{u}, \mathbf{y}\rangle$. The definition of $\|\mathbf{u}\|^{(\sqrt{k})}$ implies that $\|\mathbf{u}\|_2 \le 1$ and that $\|\mathbf{u}\|_\infty \le 1/\sqrt{k}$. Now we define — if possible — the integer $m_1 > 1$ as the smallest integer $\le m$ such that

$$\sum_{i=1}^{m_1} |u_i|^2 > \frac{1}{k}, \quad \text{so that} \quad \sum_{i=1}^{m_1-1} |u_i|^2 \le \frac{1}{k}, \quad \text{and} \quad \sum_{i=1}^{m_1} |u_i|^2 \le \frac{2}{k}.$$

Likewise, we define — if possible — the integer $m_2 > m_1 + 1$ as the smallest integer $\le m$ such that

$$\sum_{i=m_1+1}^{m_2} |u_i|^2 > \frac{1}{k}, \quad \text{so that} \quad \sum_{i=m_1+1}^{m_2-1} |u_i|^2 \le \frac{1}{k}, \quad \text{and} \quad \sum_{i=m_1+1}^{m_2} |u_i|^2 \le \frac{2}{k},$$

the integer $m_3 > m_2 + 1$ as the smallest integer $\le m$ such that

$$\sum_{i=m_2+1}^{m_3} |u_i|^2 > \frac{1}{k}, \quad \text{so that} \quad \sum_{i=m_2+1}^{m_3-1} |u_i|^2 \le \frac{1}{k}, \quad \text{and} \quad \sum_{i=m_2+1}^{m_3} |u_i|^2 \le \frac{2}{k},$$

and so on. We notice that the last m_h defined in this way has index $h < k$. Indeed, if m_k was defined, with $m_0 := 0$, we would obtain a contradiction from

$$\|\mathbf{u}\|_2^2 \ge \sum_{\ell=1}^{k} \sum_{i=m_{\ell-1}+1}^{m_\ell} |u_i|^2 > \sum_{\ell=1}^{k} \frac{1}{k} = 1.$$

We also notice that, because m_{h+1} is undefined, we have

$$\sum_{i=m_h+1}^{m} |u_i|^2 \le \frac{1}{k}.$$

We now set $B_\ell = \{m_{\ell-1}+1, \dots, m_\ell\}$ for $1 \le \ell \le h$, $B_{h+1} := \{m_h+1, \dots, m\}$, and $B_\ell = \emptyset$ for $h+2 \le \ell \le k$. In view of $\|\mathbf{u}_{B_\ell}\|_2 \le \sqrt{2/k}$ for all $1 \le \ell \le k$, we derive

$$\begin{aligned}\|\mathbf{y}\|_*^{(\sqrt{k})} = \langle \mathbf{u}, \mathbf{y}\rangle &= \sum_{\ell=1}^{k} \langle \mathbf{u}_{B_\ell}, \mathbf{y}_{B_\ell}\rangle \le \sum_{\ell=1}^{k} \|\mathbf{u}_{B_\ell}\|_2 \|\mathbf{y}_{B_\ell}\|_2 \le \sqrt{\frac{2}{k}} \sum_{\ell=1}^{k} \|\mathbf{y}_{B_\ell}\|_2 \\ &\le \sqrt{\frac{2}{k}}\, |\mathbf{y}|_k.\end{aligned}$$

This proves the rightmost inequality of (11.24). □

We are now ready to carry on with the proof of Theorem 11.19.

Proof (of Theorem 11.19). Let us suppose that $N \ge C\,m$, where the constant $C \ge 1$ has to meet three requirements determined below. We set

$$\beta := \sqrt{\frac{\ln(eN/m)}{\ln(eC)}} \ge 1.$$

Since $\beta \le \alpha$, hence $\|\mathbf{e}\|^{(\beta)} \le \|\mathbf{e}\|^{(\alpha)}$, the ℓ_1-quotient property relative to $\|\cdot\|^{(\beta)}$ implies the ℓ_1-quotient property relative to $\|\cdot\|^{(\alpha)}$, so we concentrate on the ℓ_1-quotient property relative to the norm $\|\cdot\|^{(\beta)}$. Precisely, according to Lemma 11.16 and Remark 11.17, we need to prove that

$$\mathbb{P}\big(\|\mathbf{e}\|_*^{(\beta)} \le D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\|_\infty \text{ for all } \mathbf{e} \in \mathbb{R}^m\big) \ge 1 - 3\exp\big(-m\big).$$

As in the proof of the Gaussian case, we prove the stronger statement

$$\mathbb{P}\big(\|\mathbf{e}\|_*^{(\beta)} \le D\sqrt{s_*}|\!|\!|\widetilde{\mathbf{A}}^*\mathbf{e}|\!|\!| \text{ for all } \mathbf{e} \in \mathbb{R}^m\big) \ge 1 - 3\exp\big(-m\big) \qquad (11.25)$$

with $D := 16\sqrt{\ln(eC)}$. The norm $|\!|\!|\cdot|\!|\!| \le \|\cdot\|_\infty$ is the norm defined in (11.17). We therefore assume that there exists $\mathbf{e} \in \mathbb{R}^m$ such that

$$\|\mathbf{e}\|_*^{(\beta)} > D\sqrt{s_*}|\!|\!|\widetilde{\mathbf{A}}^*\mathbf{e}|\!|\!|.$$

Introducing the integer $k := \lfloor \beta^2 \rfloor \ge 1$, for which

$$k \le \beta^2 < 2k,$$

we have $\|\mathbf{y}\|^{(\sqrt{k})} \le \|\mathbf{y}\|^{(\beta)}$ for all $\mathbf{y} \in \mathbb{R}^m$, and in turn $\|\mathbf{y}\|_*^{(\sqrt{k})} \ge \|\mathbf{y}\|_*^{(\beta)}$ for all $\mathbf{y} \in \mathbb{R}^m$. Assuming without loss of generality that $\|\mathbf{e}\|_*^{(\sqrt{k})} = 1$, we obtain $D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\| < 1$. Moreover, for $0 < \delta < 1$ to be chosen later, we consider a δ-covering $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ of the unit sphere of $(\mathbb{R}^m, \|\cdot\|_*^{(\sqrt{k})})$ with cardinality $n \le (1 + 2/\delta)^m$. Selecting an integer $i \in [n]$ such that $\|\mathbf{e} - \mathbf{e}_i\|_*^{(\sqrt{k})} \le \delta$, it follows that

$$\begin{aligned} D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| &\le D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\| + D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i)\| \\ &< 1 + D\frac{\sqrt{s_*}}{2h}\sum_{\ell=1}^{2h}\|(\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i))_{T_\ell}\|_\infty \le 1 + D\frac{\sqrt{s_*}}{2h}\sum_{\ell=1}^{2h}\|(\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i))_{T_\ell}\|_2 \\ &\le 1 + D\frac{\sqrt{s_*}}{\sqrt{2h}}\|\widetilde{\mathbf{A}}^*(\mathbf{e} - \mathbf{e}_i)\|_2 \le 1 + D\sqrt{\frac{s_*N}{2hm}}\sigma_{\max}(\mathbf{B})\|\mathbf{e} - \mathbf{e}_i\|_2, \end{aligned} \tag{11.26}$$

where $\mathbf{B} \in \mathbb{R}^{N\times m}$ is the renormalized matrix $\mathbf{B} := \sqrt{m/N}\,\widetilde{\mathbf{A}}^* = \mathbf{A}^*/\sqrt{N}$. Let us observe that, if $B_1, \dots, B_k$ is an optimal partition for $|\mathbf{e} - \mathbf{e}_i|_k$, then

$$\|\mathbf{e} - \mathbf{e}_i\|_2 \le \sum_{\ell=1}^{k}\|(\mathbf{e} - \mathbf{e}_i)_{B_\ell}\|_2 = |\mathbf{e} - \mathbf{e}_i|_k \le \sqrt{k}\,\delta, \tag{11.27}$$

where Lemma 11.20 was used in the last inequality. Thus, under the assumption that $\sigma_{\max}(\mathbf{B}) \le \sqrt{2}$, (11.26) and (11.27) yield

$$D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| < 1 + D\sqrt{\frac{ks_*}{m}}\sqrt{\frac{N}{h}}\,\delta \le 1 + \frac{D}{\sqrt{\ln(eC)}}\sqrt{\frac{N}{m}}\,\delta \le \frac{4}{\sqrt{k}}|\mathbf{e}_i|_k,$$

where the last inequality holds because of the choice

$$\delta := \frac{\sqrt{\ln(eC)}}{D}\sqrt{\frac{h}{N}}$$

and of the fact that $1 = \|\mathbf{e}_i\|_*^{(\sqrt{k})} \le 2|\mathbf{e}_i|/\sqrt{k}$. Summarizing the previous considerations gives, with $d := D/4$,

$$\begin{aligned} \mathbb{P}\big(\|\mathbf{e}\|_*^{(\beta)} > D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\|_\infty \text{ for some } \mathbf{e} \in \mathbb{R}^m\big) &\le \mathbb{P}\big(\sigma_{\max}(B) > \sqrt{2}\big) \\ &\quad + \mathbb{P}\big(|\mathbf{e}_i|_k > d\sqrt{ks_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| \text{ for some } i \in [n]\big). \end{aligned} \tag{11.28}$$

For the first term on the right-hand side of (11.28), we call upon Theorem 9.8 to obtain

$$\begin{aligned} \mathbb{P}\big(\sigma_{\max}(B) > \sqrt{2}\big) = \mathbb{P}\big(\sigma_{\max}^2(B) > 2\big) &\le \mathbb{P}\big(\|\mathbf{B}^*\mathbf{B} - \mathbf{Id}\|_{2\to 2} > 1\big) \\ &\le 2\exp\bigg(-\frac{\tilde{c}N}{2}\bigg) \end{aligned}$$

provided

$$N \geq \frac{2}{3\,\tilde{c}}\Big(7m + \tilde{c}N\Big), \qquad \text{i.e.,} \qquad N \geq \frac{14}{\tilde{c}}m.$$

The first requirement imposed on C is therefore $C \geq 14/\tilde{c}$. In this case, we have the bound

$$\mathbb{P}\big(\sigma_{\max}(B) > \sqrt{2}\big) \leq 2\exp\big(-7m\big). \tag{11.29}$$

For the first term on the right-hand side of (11.28), we begin by bounding the probability $\mathbb{P}\big(|\mathbf{e}|_k > d\sqrt{ks_*}\,|\!|\!|\widetilde{\mathbf{A}}^*\mathbf{e}|\!|\!|\big)$ for fixed vectors $\mathbf{e} \in \mathbb{R}^m$. As in the proof of Theorem 11.18, using the existence of a subset L of $[N]$ of size h such that $\|\mathbf{z}_{T_\ell}\|_\infty \leq 2|\!|\!|\mathbf{z}|\!|\!|$ for any $\mathbf{z} \in \mathbb{R}^N$, we observe that

$$\begin{aligned}
&\mathbb{P}\big(|\mathbf{e}|_k > d\sqrt{ks_*}\,|\!|\!|\widetilde{\mathbf{A}}^*\mathbf{e}|\!|\!|\big) \\
&\quad\leq \mathbb{P}\bigg(\|(\widetilde{\mathbf{A}}^*\mathbf{e})_{T_\ell}\|_\infty < \frac{2|\mathbf{e}|_k}{d\sqrt{ks_*}} \text{ for all } \ell \text{ in some } L \subseteq [2h], \operatorname{card}(L) = h\bigg) \\
&\quad\leq \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \mathbb{P}\bigg(\max_{j\in T_\ell} |(\widetilde{\mathbf{A}}^*\mathbf{e})_j| < \frac{2|\mathbf{e}|_k}{d\sqrt{ks_*}} \text{ for all } \ell \in L\bigg) \\
&\quad= \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \prod_{j\in\cup_{\ell\in L}T_\ell} \mathbb{P}\bigg(|(\widetilde{\mathbf{A}}^*\mathbf{e})_j| < \frac{2|\mathbf{e}|_k}{d\sqrt{ks_*}}\bigg) \\
&\quad= \sum_{L\subseteq[2h],\operatorname{card}(L)=h} \prod_{j\in\cup_{\ell\in L}T_\ell} \bigg(1 - 2\,\mathbb{P}\bigg((\widetilde{\mathbf{A}}^*\mathbf{e})_j \geq \frac{2|\mathbf{e}|_k}{d\sqrt{ks_*}}\bigg)\bigg),
\end{aligned} \tag{11.30}$$

where the symmetry of the random variables $a_{i,j}$ was used in the last step. Let now $B_1, \ldots, B_k$ denote a partition of $[m]$ such that $|\mathbf{e}|_k = \sum_{\ell=1}^k \|\mathbf{e}_{B_\ell}\|_2$. For each $j \in \cup_{\ell\in L}T_\ell$, we have

$$\begin{aligned}
\mathbb{P}\bigg((\widetilde{\mathbf{A}}^*\mathbf{e})_j \geq \frac{2|\mathbf{e}|_k}{d\sqrt{ks_*}}\bigg) &= \mathbb{P}\bigg(\sum_{\ell=1}^k \sum_{i\in B_\ell} \frac{a_{i,j}}{\sqrt{m}} e_i \geq \sum_{\ell=1}^k \frac{2\|\mathbf{e}_{B_\ell}\|_2}{d\sqrt{ks_*}}\bigg) \\
&\geq \mathbb{P}\bigg(\sum_{i\in B_\ell} a_{i,j}e_i \geq \frac{2}{d}\sqrt{\frac{m}{ks_*}}\|\mathbf{e}_{B_\ell}\|_2 \text{ for all } \ell \in [k]\bigg) \\
&= \prod_{\ell\in[k]} \mathbb{P}\bigg(\sum_{i\in B_\ell} a_{i,j}e_i \geq \frac{2}{d}\sqrt{\frac{m}{ks_*}}\|\mathbf{e}_{B_\ell}\|_2\bigg) \\
&= \prod_{\ell\in[k]} \frac{1}{2}\,\mathbb{P}\bigg(\Big|\sum_{i\in B_\ell} a_{i,j}e_i\Big| \geq \frac{2}{d}\sqrt{\frac{m}{ks_*}}\|\mathbf{e}_{B_\ell}\|_2\bigg),
\end{aligned}$$

where the last step follows again from the symmetry of the random variables. For each $\ell \in [k]$, we use Lemma 7.17 to obtain

$$\mathbb{P}\Big(\Big|\sum_{i\in B_\ell} a_{i,j}e_i\Big| \ge \frac{2}{d}\sqrt{\frac{m}{ks_*}}\|\mathbf{e}_{B_\ell}\|_2\Big) \ge \frac{1}{\mu^4}\Big(1-\frac{4m}{d^2ks_*}\Big)^2 \ge \frac{1}{\mu^4}\Big(1-\frac{8m}{d^2\beta^2 s_*}\Big)^2$$
$$= \frac{1}{\mu^4}\Big(1-\frac{8\ln(eC)}{d^2}\Big)^2 = \frac{1}{4\mu^4},$$

where we made use of the value $d = D/4 = 4\sqrt{\ln(eC)}$. It follows that

$$\mathbb{P}\Big((\widetilde{\mathbf{A}}^*\mathbf{e})_j \ge \frac{2|\mathbf{e}|_k}{d\sqrt{ks_*}}\Big) \ge \Big(\frac{1}{8\mu^4}\Big)^k \ge \Big(\frac{1}{8\mu^4}\Big)^{\beta^2} = \exp\big(-\beta^2\ln(8\mu^4)\big)$$
$$= \exp\bigg(-\ln\Big(\frac{eN}{m}\Big)\frac{\ln(8\mu^4)}{\ln(eC)}\bigg) \ge \Big(\frac{m}{eN}\Big)^{1/2}. \quad (11.31)$$

The last inequality holds by virtue of a second requirement on C, namely $C \ge 64\mu^8/e$. Substituting (11.31) into (11.30), while using $1-x \le \exp(-x)$, we obtain

$$\mathbb{P}\big(|\mathbf{e}|_k > d\sqrt{ks_*}\,\|\widetilde{\mathbf{A}}^*\mathbf{e}\|\big) \le \sum_{L\subseteq[2h],\mathrm{card}(L)=h} \exp\bigg(-2\Big(\frac{m}{eN}\Big)^{1/2}\bigg)^{\mathrm{card}(\cup_{\ell\in L}T_\ell)}$$
$$\le \binom{2h}{h}\exp\bigg(-2\Big(\frac{m}{eN}\Big)^{1/2}\bigg)^{N/2} \le \exp\bigg(\ln(2e)h - \frac{1}{e^{1/2}}m^{1/2}N^{1/2}\bigg).$$

Thus, in view of $n \le (1+2/\delta)^m \le \exp(2m/\delta)$, we derive

$$\mathbb{P}\big(|\mathbf{e}_i|_k > d\sqrt{ks_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| \text{ for some } i\in[n]\big) \le n\,\mathbb{P}\big(|\mathbf{e}|_k > d\sqrt{ks_*}\,\|\widetilde{\mathbf{A}}^*\mathbf{e}\|\big)$$
$$\le \exp\bigg(\frac{2}{\delta}m + \ln(2e)h - \frac{1}{e^{1/2}}m^{1/2}N^{1/2}\bigg).$$

We now choose $h := \lceil m^{2/3}N^{2/3}\rceil$ (so that $1 \le h \le N/2$ when $C \ge 64\mu^8/e$). We then have $h \le 2\,m^{2/3}N^{2/3}$ and $2/\delta = 32(N/h)^{1/2} \le 32(N/m)^{1/3}$. It follows that

$$\mathbb{P}\big(|\mathbf{e}_i|_k > d\sqrt{ks_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| \text{ for some } i\in[n]\big)$$
$$\le \exp\bigg(32m^{2/3}N^{1/3} + 2\ln(2e)m^{2/3}N^{1/3} - \frac{1}{e^{1/2}}m^{1/2}N^{1/2}\bigg)$$
$$\le \exp\bigg(-\bigg[\frac{1}{e^{1/2}} - \frac{2\ln(2e^{17})}{(N/m)^{1/6}}\bigg]m^{1/2}N^{1/2}\bigg)$$
$$\le \exp\bigg(-\bigg[\frac{1}{e^{1/2}} - \frac{2\ln(2e^{17})}{C^{1/6}}\bigg]m^{1/2}N^{1/2}\bigg).$$

A third requirement on C, namely $C^{1/6} \ge 4e^{1/2}\ln(2e^{17})$, implies that

$$\mathbb{P}\big(|\mathbf{e}_i|_k > d\sqrt{ks_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}_i\| \text{ for some } i\in[n]\big)$$
$$\le \exp\bigg(-\frac{m^{1/2}N^{1/2}}{2e^{1/2}}\bigg) \le \exp\bigg(-\frac{4}{e}m\bigg). \quad (11.32)$$

In the last step, we simply used the second requirement $C \geq 64\mu^8/e$. Finally, substituting (11.32) and (11.29) into (11.28), we conclude that

$$\begin{aligned}\mathbb{P}\big(\|\mathbf{e}\|_*^{(\beta)} > D\sqrt{s_*}\|\widetilde{\mathbf{A}}^*\mathbf{e}\|_\infty \text{ for some } \mathbf{e} \in \mathbb{R}^m\big) \\ \leq 2\exp\big(-7m\big) + \exp\big(-4m/e\big) \leq 3\exp\big(-m\big).\end{aligned}$$

We have proved the desired estimate (11.25). □

We now prove the main robustness estimate for subgaussian matrices.

Proof (of Theorem 11.10). According to the definition of subgaussian matrices and to the bound on moments in terms of tail probabilities, i.e., Definition 9.1 and Proposition 7.13, the symmetric entries of the subgaussian matrix $\mathbf{A}$ have fourth moments bounded by some $\mu^4 \geq 1$. Moreover, according to Lemma 9.7, the concentration inequality (11.23) is satisfied. Thus, by choosing c_2 properly, Theorem 11.19 guarantees that, with probability at least $1 - 3\exp(-m)$, the matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ has the ℓ_1-quotient property relative to the norm $\|\cdot\|^{(\alpha)}$, $\alpha := \sqrt{\ln(eN/m)}$. Furthermore, according to Theorem 9.10, there is a constant $\tilde{c} > 0$ such that $\delta_{2s}(\widetilde{\mathbf{A}}) < 1/3$ with probability at least $1-2\exp(-\tilde{c}m/15)$ provided

$$m \geq \frac{6}{\tilde{c}}\Big[s\Big(18 + 4\ln\Big(\frac{N}{2s}\Big)\Big) + \frac{2\tilde{c}}{15}m\Big], \quad \text{i.e.,} \quad m \geq \frac{60}{\tilde{c}}s\Big(9 + 2\ln\Big(\frac{N}{2s}\Big)\Big).$$

Since $9 + 2\ln(N/2s) \leq 9\ln(eN/s)$, this follows from $m \geq (540/\tilde{c})s\ln(eN/s)$. Using Lemma C.6, we observe that this condition is implied by the condition $s \leq c_3 s_*$ — which is equivalent to $m \geq (1/c_3)s\ln(eN/m)$ — provided c_3 is chosen small enough to have $c_3\ln(e/c_3) \leq \tilde{c}/540$. Theorem 6.12 now ensures that the matrix $\widetilde{\mathbf{A}}$ satisfies the ℓ_2-robust null space property of order s relative to $\|\cdot\|_2$. Since $\|\cdot\|_2 \leq \|\cdot\|^{(\alpha)}$, it also satisfies the ℓ_2-robust null space property of order s relative to $\|\cdot\|^{(\alpha)}$. Thus, with probability at least

$$1 - 3\exp(-m) - 2\exp(-\tilde{c}m/15) \geq 1 - 5\exp(-c_1 m), \quad c_1 := \min\{1, \tilde{c}/15\},$$

the matrix $\widetilde{\mathbf{A}}$ satisfies both the ℓ_1-quotient property and the ℓ_2-robust null space property of order $s \leq s_*/c_3$ relative to the norm $\|\cdot\|^{(\alpha)}$. The conclusion now follows from Theorem 11.12. □

11.4 Nonuniform Instance Optimality

In Section 11.1, we have established that the uniform ℓ_2-instance optimality—the property that $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_2 \leq C\sigma_s(\mathbf{x})_2$ for all $\mathbf{x} \in \mathbb{C}^N$ — was only possible in the case $m \geq cN$ which is irrelevant in compressive sensing. In this section, we change the point of view, as we fix $\mathbf{x} \in \mathbb{C}^N$ at the start. We are going to prove, for the ℓ_1-minimization map, that the *nonuniform ℓ_2-instance*

optimality — the property that $\|\mathbf{x} - \Delta_1(\mathbf{A}\mathbf{x})\|_2 \le C\sigma_s(\mathbf{x})_2$ for this fixed $\mathbf{x} \in \mathbb{C}^N$ — occurs with high probability on the draw of an $m \times N$ random matrix $\mathbf{A}$, provided $m \ge cs\ln(eN/s)$. We notice that such estimates hold for other algorithms such as iterative hard thresholding, hard thresholding pursuit, orthogonal matching pursuit, and compressive sampling matching pursuit: indeed, under some restricted isometry conditions, Theorems 6.20, 6.24, and 6.27 yield $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_2 \le C\|\mathbf{A}\mathbf{x}_{\overline{S}}\|_2$, where S denotes an index set of s largest absolute entries of $\mathbf{x}$, and the desired estimate follows from the the concentration inequality $\|\mathbf{A}\mathbf{x}_{\overline{S}}\|_2 \le 2\|\mathbf{x}_{\overline{S}}\|_2 = 2\sigma_s(\mathbf{x})_2$. However, these algorithms (except perhaps orthogonal matching pursuit) necessitate s as an input. Advantageously, the ℓ_1-minimization does not. For the ℓ_1-minimization, the key to proving the nonuniform ℓ_2-instance optimality lies in the stable and robust estimates of Theorems 11.9 and 11.10. We begin with the easier case of Gaussian matrices. The result also incorporates measurement error.

Theorem 11.21. *There exist absolute constants $c_1, c_2, c_3, C, D > 0$ such that, if $\mathbf{x} \in \mathbb{C}^N$ is a fixed vector and if $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ where $\mathbf{A}$ is an $m \times N$ Gaussian matrix, then, with probability at least $1 - 5\exp(-c_1 m)$, the ℓ_2-error estimates*

$$\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_2 \le C\sigma_s(\mathbf{x})_2 + D\|\mathbf{e}\|_2 \tag{11.33}$$

are valid for all $\mathbf{e} \in \mathbb{C}^m$, provided

$$N \ge c_2 m, \qquad s \le c_3 s_* = \frac{c_3 m}{\ln(eN/m)}.$$

Proof. Let S denote a set of s largest absolute entries of $\mathbf{x}$. We have

$$\begin{aligned}\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_2 &\le \|\mathbf{x}_{\overline{S}}\|_2 + \|\mathbf{x}_S - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_2 \\ &= \sigma_s(\mathbf{x})_2 + \|\mathbf{x}_S - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x}_S + \mathbf{e}')\|_2, \end{aligned} \tag{11.34}$$

where $\mathbf{e}' := \widetilde{\mathbf{A}}\mathbf{x}_{\overline{S}} + \mathbf{e}$. Taking the conditions $N \ge c_2 m$ and $s \le c_3 s_*$ into account, Theorem 11.9 applied to $\mathbf{x}_S \in \mathbb{C}^N$ and $\mathbf{e}' \in \mathbb{C}^m$ yields

$$\|\mathbf{x}_S - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x}_S + \mathbf{e}')\|_2 \le D\|\mathbf{e}'\|_2 \le D\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S}}\|_2 + D\|\mathbf{e}\|_2 \tag{11.35}$$

with probability at least $1 - 3\exp(-c_1' m)$ for some constant $c_1' > 0$. Next, the concentration inequality for Gaussian matrices (see Exercise 9.3) ensures that

$$\|\mathbf{A}\mathbf{x}_{\overline{S}}\|_2 \le 2\|\mathbf{x}_{\overline{S}}\|_2 = 2\sigma_s(\mathbf{x})_2 \tag{11.36}$$

with probability at least $1 - 2\exp(-m/12)$. We finally derive (11.33) by combining (11.34), (11.35), and (11.36). The desired probability is at least $1 - 3\exp(-c_1' m) - 2\exp(-m/12) \ge 1 - 5\exp(-c_1 m)$, $c_1 := \min\{c_1', 1/12\}$. □

In the same spirit, a nonuniform mixed (ℓ_q, ℓ_p)-instance optimality result for Gaussian matrices can be proved for any $1 \le p \le q \le 2$. It reads as follows.

Theorem 11.22. *There exist absolute constants $c_1, c_2, c_3, C, D > 0$ such that, for $1 \le p \le q \le 2$, if $\mathbf{x} \in \mathbb{C}^N$ is a fixed vector and if $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$, where $\mathbf{A}$ is an $m \times N$ Gaussian matrix, then, with probability at least $1 - 5\exp(-c_1 m)$, the error estimates*

$$\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q \le \frac{C}{s^{1/p-1/q}}\sigma_s(\mathbf{x})_p + D s_*^{1/q-1/2}\|\mathbf{e}\|_2$$

are valid for all $\mathbf{e} \in \mathbb{C}^m$, provided

$$N \ge c_2 m, \qquad s \le c_3 s_* = \frac{c_3 m}{\ln(eN/m)}.$$

Proof. If $c_1', c_2', c_3', C', D' > 0$ are the constants of Theorem 11.9, we define $c_3 := c_3'/3$. Then, for $s \le c_3 s_*$, we consider an index set of S largest absolute entries of $\mathbf{x}$, and an index set T of $t := \lceil c_3 s_* \rceil \ge s$ next largest absolute entries of $\mathbf{x}$. We have

$$\begin{aligned}\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q &\le \|\mathbf{x}_{\overline{S\cup T}}\|_q + \|\mathbf{x}_{S\cup T} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q \\ &\le \frac{1}{t^{1/p-1/q}}\|\mathbf{x}_{\overline{S}}\|_p + \|\mathbf{x}_{S\cup T} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x}_{S\cup T} + \mathbf{e}')\|_q, \end{aligned} \tag{11.37}$$

where we have used Proposition 2.3 and set $\mathbf{e}' := \widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}} + \mathbf{e}$ in the last inequality. Taking $c_2 = c_2'$ and noticing that $s + t \le c_3' s_*$, Theorem 11.9 applied to $\mathbf{x}_{S\cup T} \in \mathbb{C}^N$ and $\mathbf{e}' \in \mathbb{C}^m$ yields

$$\begin{aligned}\|\mathbf{x}_{S\cup T} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x}_{S\cup T} + \mathbf{e}')\|_q &\le D s_*^{1/q-1/2}\|\mathbf{e}'\|_2 \\ &\le D s_*^{1/q-1/2}\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_2 + D s_*^{1/q-1/2}\|\mathbf{e}\|_2 \\ &\le D\frac{t^{1/q-1/2}}{c_3^{1/q-1/2}}\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_2 + D s_*^{1/q-1/2}\|\mathbf{e}\|_2 \end{aligned} \tag{11.38}$$

with probability at least $1 - 3\exp(-c_1' m)$. The concentration inequality for Gaussian matrices (see Exercise 9.3), in conjunction with Proposition 2.3, gives

$$\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_2 \le 2\|\mathbf{x}_{\overline{S\cup T}}\|_2 \le \frac{2}{t^{1/p-1/2}}\|\mathbf{x}_{\overline{S}}\|_p \tag{11.39}$$

with probability at least $1 - 2\exp(-m/12)$. Combining (11.37), (11.38), and (11.39), we deduce

$$\begin{aligned}\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q &\le \frac{1 + 2D c_3^{1/2-1/q}}{t^{1/p-1/2}}\|\mathbf{x}_{\overline{S}}\|_p + D s_*^{1/q-1/2}\|\mathbf{e}\|_2 \\ &\le \frac{1 + 2D c_3^{-1/2}}{s^{1/p-1/2}}\sigma_s(\mathbf{x})_p + D s_*^{1/q-1/2}\|\mathbf{e}\|_2. \end{aligned}$$

The desired probability is $1 - 3\exp(-c_1' m) - 2\exp(-m/12) \ge 1 - 5\exp(-c_1 m)$, $c_1 := \min\{c_1', 1/12\}$. □

The previous results extend to subgaussian matrices. We do not isolate the ℓ_2-instance optimality here, as we state the nonuniform mixed instance optimality directly. As in Section 11.3, the ℓ_2-norm on the measurement error is replaced by

$$\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)} := \max\left\{\|\mathbf{e}\|_2, \sqrt{\ln(eN/m)}\|\mathbf{e}\|_\infty\right\}.$$

Theorem 11.23. *For any $1 \le p \le q \le 2$, if $\mathbf{x} \in \mathbb{C}^N$ is a fixed vector and if $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ where $\mathbf{A}$ is an $m \times N$ subgaussian matrix with symmetric entries, then there exist constants $c_1, c_2, c_3, c_4, C, D > 0$ depending only on the subgaussian distributions such that, with probability at least $1 - 9\exp(-c_1\sqrt{m})$, the error estimates*

$$\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q \le \frac{C}{s^{1/p-1/q}}\sigma_s(\mathbf{x})_p + Ds_*^{1/q-1/2}\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)}$$

are valid for all $\mathbf{e} \in \mathbb{C}^m$, provided

$$c_2 m \le N \le \frac{m}{e}\exp(c_3\sqrt{m}), \qquad s \le c_4 s_* = \frac{c_4 m}{\ln(eN/m)}.$$

Proof. The argument is similar to the one used in the proof of Theorem 11.22, with the addition of step (11.42). If $c_1', c_2', c_3', C', D' > 0$ are the constants of Theorem 11.10, we define $c_4 := c_3'/3$. Then, for $s \le c_4 s_*$, we consider an index set of S largest absolute entries of $\mathbf{x}$, and an index set T of $t := \lceil c_4 s_* \rceil \ge s$ next largest absolute entries of $\mathbf{x}$. We have

$$\begin{aligned}\|\mathbf{x} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q &\le \|\mathbf{x}_{\overline{S\cup T}}\|_q + \|\mathbf{x}_{S\cup T} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x} + \mathbf{e})\|_q \\ &\le \frac{1}{t^{1/p-1/q}}\|\mathbf{x}_{\overline{S}}\|_p + \|\mathbf{x}_{S\cup T} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x}_{S\cup T} + \mathbf{e}')\|_q. \end{aligned} \tag{11.40}$$

Taking $c_2 = c_2'$ and noticing that $s + t \le c_3' s_*$, Theorem 11.10 applied to $\mathbf{x}_{S\cup T} \in \mathbb{C}^N$ and $\mathbf{e}' \in \mathbb{C}^m$ yields

$$\begin{aligned}&\|\mathbf{x}_{S\cup T} - \Delta_1(\widetilde{\mathbf{A}}\mathbf{x}_{S\cup T} + \mathbf{e}')\|_q \le Ds_*^{1/q-1/2}\|\mathbf{e}'\|^{\left(\sqrt{\ln(eN/m)}\right)} \\ &\le Ds_*^{1/q-1/2}\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|^{\left(\sqrt{\ln(eN/m)}\right)} + Ds_*^{1/q-1/2}\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)} \\ &\le D\frac{t^{1/q-1/2}}{c_3^{1/q-1/2}}\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|^{\left(\sqrt{\ln(eN/m)}\right)} + Ds_*^{1/q-1/2}\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)} \end{aligned} \tag{11.41}$$

with probability at least $1 - 5\exp(-c_1' m)$. By the concentration inequality of Lemma 9.7, we have

$$\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_2 \le 2\|\mathbf{x}_{\overline{S\cup T}}\|_2$$

with probability at least $1 - 2\exp(-\tilde{c}m)$ for a constant $\tilde{c}$ depending only on the subgaussian distributions. Moreover, for each $i \in [m]$, Theorem 7.27 guarantees that the inequality

$$|(\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}})_i| \le \frac{2}{\sqrt{\ln(eN/m)}}\|\mathbf{x}_{\overline{S\cup T}}\|_2$$

holds with probability at least $1-2\exp\big(-cm/\ln(eN/m)\big)$ for a constant c depending only on the subgaussian distributions. It follows that

$$\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_\infty \le \frac{2}{\sqrt{\ln(eN/m)}}\|\mathbf{x}_{\overline{S\cup T}}\|_2 \tag{11.42}$$

with probability at least $1-2m\exp\big(-cm/\ln(eN/m)\big)$. We note that this probability is at least $1-2\exp(-\sqrt{m})$ when $N \le m\exp(c_3\sqrt{m})/e$, $c_3 := c/2$, since

$$m\exp\Big(\frac{-cm}{\ln(eN/m)}\Big) \le m\exp\Big(\frac{-cm}{c_3\sqrt{m}}\Big) = \exp\big(\ln(m)-2\sqrt{m}\big) \le \exp\big(-\sqrt{m}\big).$$

We have obtained

$$\begin{aligned}\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|^{\left(\sqrt{\ln(eN/m)}\right)} &= \max\{\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_2, \sqrt{\ln(eN/m)}\|\widetilde{\mathbf{A}}\mathbf{x}_{\overline{S\cup T}}\|_\infty\}\\ &\le 2\|\mathbf{x}_{\overline{S\cup T}}\|_2 \le \frac{2}{t^{1/p-1/2}}\|\mathbf{x}_{\overline{S}}\|_p,\end{aligned}$$

with probability at least $1-2\exp(-\tilde{c}m)-2\exp(-\sqrt{m}) \ge 1-4\exp(-c'\sqrt{m})$, $c' := \min\{\tilde{c},1\}$. Combining (11.37), (11.40), and (11.41), we deduce

$$\begin{aligned}\|\mathbf{x}-\Delta_1(\widetilde{\mathbf{A}}\mathbf{x}+\mathbf{e})\|_q &\le \frac{1+2Dc_3^{1/2-1/q}}{t^{1/p-1/2}}\|\mathbf{x}_{\overline{S}}\|_p + Ds_*^{1/q-1/2}\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)}\\ &\le \frac{1+2Dc_3^{-1/2}}{s^{1/p-1/2}}\sigma_s(\mathbf{x})_p + Ds_*^{1/q-1/2}\|\mathbf{e}\|^{\left(\sqrt{\ln(eN/m)}\right)}.\end{aligned}$$

The desired probability is $1-5\exp(-c_1'm)-4\exp(-c'\sqrt{m})$, which is at least $1-9\exp(-c_1\sqrt{m})$, $c_1 := \min\{c_1', c'\}$. □

Notes

The notions of instance optimality and mixed instance optimality were introduced by A. Cohen, W. Dahmen, and R. DeVore in [102]. Theorems 11.4 and 11.5 are taken from this article. The other major theorem of Section 11.1, namely Theorem 11.6 on the minimal number of measurements for ℓ_1-instance optimality, is taken from [185].

The ℓ_1-quotient property was introduced in the context of Compressive Sensing by P. Wojtaszczyk in [446]. The content of Section 11.2 essentially follows the ideas of this article, except that we replaced the restricted isometry property by the weaker notion of robust null space property, and that we gave error estimates in ℓ_q-norm for all $1 \le q \le 2$. The ℓ_1-quotient property

for Gaussian matrices was proved in [446], too, save for the extra requirement that $N \geq cm \ln^{\xi}(m)$ for some $\xi > 0$ — this issue was resolved here with the use of the norm defined in (11.17). As a matter of fact, the ℓ_1-quotient property for Gaussian matrices had been established earlier in a different context by E. Gluskin in [197], where a certain optimality of the probability estimate was also proved [CHECK THIS!].

Gaussian matrices are not the only random matrices that satisfy the ℓ_1-quotient property relative to the ℓ_2-norm, and in turn the estimates of Theorem 11.9. It was established in [182] that Weibull matrices also do. For matrices satisfying the restricted isometry property, P. Wojtaszczyk also showed in [447] that the estimates of Theorem 11.9 can be obtained with a modified ℓ_1-minimization in which one artificially adds columns to the matrix $\mathbf{A}$.

The ℓ_1-quotient property relative to the norm $\max\{\|\cdot\|_2, \alpha\|\cdot\|_\infty\}$ was introduced in the context of Compressive Sensing by R. DeVore, G. Petrova, and P. Wojtaszczyk in [125], where it was proved for Bernoulli random matrices. As for the Gaussian case, it had been established earlier in a different context by A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann in [284]. We followed the proof of [284], because of a slight flaw in the proof of [125], namely that the vectors in their δ-covering depend on the random matrix, hence the concentration inequality cannot be applied directly to them. The key Lemma 11.20 was proved by S. Montgomery-Smith in [307].

The results given in Section 11.4 on the nonuniform ℓ_2-instance optimality appeared (under a different terminology) in [446] and [125].

Exercises

11.1. Verify in details the observation made in (11.1).

11.2. For $q \geq p \geq p' \geq 1$, prove that if a pair $(\mathbf{A}, \Delta)$ is mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C, then it is also mixed $(\ell_q, \ell_{p'})$-instance optimal of order $\lceil s/2 \rceil$ with constant C' depending only on C. Combine this result with Theorem 11.7 to derive that mixed (ℓ_q, ℓ_p)-instance optimal pairs $(\mathbf{A}, \Delta)$ of order s, where $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $\Delta : \mathbb{C}^m \to \mathbb{C}^N$, can only exist if $m \geq c\, s \ln(eN/s)$. For $q > p > 1$, improve this bound using the estimate for the Gelfand width $d^m(B_p^N, \ell_q^N)$ given on page 299.

11.3. Prove that if the coherence of a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ with ℓ_2-normalized columns satisfies $\mu(\mathbf{A}) < 1/4$, then the operator norm $\|\mathbf{A}\|_{2 \to 2}$ cannot be bounded by an absolute constant $C > 0$ unless $m \geq c\, N$ for some constant $c > 0$ depending on C.

11.4. Let a measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ be given and let $0 < p < 1$. Prove that if there is a reconstruction map Δ such that $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_p \leq C \sigma_{2s}(\mathbf{x})_p$ for all $\mathbf{x} \in \mathbb{C}^N$, then $\|\mathbf{v}\|_p \leq C \sigma_{2s}(\mathbf{v})_p$ for all $\mathbf{v} \in \ker \mathbf{A}$. Prove conversely that if $\|\mathbf{v}\|_p \leq C \sigma_{2s}(\mathbf{v})_p$ for all $\mathbf{v} \in \ker \mathbf{A}$, then there is a reconstruction map Δ such that $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_p \leq 2^{1/p} C \sigma_{2s}(\mathbf{x})_p$ for all $\mathbf{x} \in \mathbb{C}^N$.

11.5. Let a measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ be given. Suppose that, for some integer $s \geq 1$ and some constant $C \geq 1$, there exists a reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ such that

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_2 \leq \frac{C}{\sqrt{s}} \|\mathbf{x}\|_1 \qquad \text{for all } \mathbf{x} \in \mathbb{C}^N.$$

Prove that the pair $(\mathbf{A}, \Delta_1)$ is mixed (ℓ_2, ℓ_1)-instance optimal of order t with constant $(2+\rho)/(1-\rho)$ provided $\rho := 2C\sqrt{t/s} < 1$. Deduce that the existence of a pair (A, Δ) which is mixed (ℓ_2, ℓ_1)-instance optimal of order $\lceil 9C^2 t \rceil$ with constant C implies that the pair $(\mathbf{A}, \Delta_1)$ is mixed (ℓ_2, ℓ_1)-instance optimal of order t with constant 8.

11.6. Let $\mathbf{A} \in \mathbb{R}^{m \times N}$ and let $\|\cdot\|$ be a norm on $\mathbb{C}^m$ invariant by complex conjugation, i.e., satisfying $\|\overline{\mathbf{y}}\| = \|\mathbf{y}\|$ for all $\mathbf{y} \in \mathbb{C}^m$. For $q \geq 1$, prove that the real and complex versions of the ℓ_q-quotient property, namely

$$\forall \mathbf{e} \in \mathbb{R}^m, \exists \mathbf{u} \in \mathbb{R}^N : \mathbf{A}\mathbf{u} = \mathbf{e}, \ \|\mathbf{u}\|_q \leq d\, s_*^{1/q-1/2} \|\mathbf{e}\|, \tag{11.43}$$

$$\forall \mathbf{e} \in \mathbb{C}^m, \exists \mathbf{u} \in \mathbb{C}^N : \mathbf{A}\mathbf{u} = \mathbf{e}, \ \|\mathbf{u}\|_q \leq d\, s_*^{1/q-1/2} \|\mathbf{e}\|, \tag{11.44}$$

are equivalent, in the sense that (11.44) implies (11.43) with the same constant d and (11.43) implies (11.44) with the constant d replaced by $2d$.

11.7. Prove Lemma 11.16 in the case $q = 1$ without using limiting arguments.

11.8. Prove that the dual norm of the norm $\|\cdot\|^{(\alpha)}$ introduced in (11.22) can be expressed as

$$\|\mathbf{y}\|_*^{(\alpha)} = \inf \left\{ \|\mathbf{y}'\|_2 + \frac{1}{\alpha} \|\mathbf{y}''\|_1, \mathbf{y}' + \mathbf{y}'' = \mathbf{y} \right\}, \qquad \mathbf{y} \in \mathbb{C}^m.$$

11.9. Let $q \geq 1$ and let $\|\cdot\|$ be a norm on $\mathbb{C}^m$. Given a matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$, suppose that there exist $D > 0$ and $0 < \rho < 1$ such that, for each $\mathbf{e} \in \mathbb{C}^m$, one can find $\mathbf{u} \in \mathbb{C}^N$ with $\|\mathbf{A}\mathbf{u} - \mathbf{e}\| \leq \rho \|\mathbf{e}\|$ and $\|\mathbf{u}\|_q \leq D s_*^{1/q-1/2} \|\mathbf{e}\|$. Prove that the matrix $\mathbf{A}$ satisfies the ℓ_q-quotient property with constant $D/(1-\rho)$ relative to the norm $\|\cdot\|$.

11.10. Let $q \geq 1$ and let $\|\cdot\|$ be a norm on $\mathbb{C}^m$. Suppose that a pair of measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ and reconstruction map $\Delta : \mathbb{C}^N \to \mathbb{C}^m$ is mixed (ℓ_q, ℓ_1)-instance optimal of order $s \leq c s_*$ and that $\mathbf{A}$ has the simultaneous (ℓ_q, ℓ_1)-quotient property relative to $\|\cdot\|$. Prove that there is a constant $D > 0$ such that

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_q \leq \|\mathbf{x}_{\overline{S}}\|_q + D s_*^{1/q-1/2} \|\mathbf{A}\mathbf{x}_{\overline{S}}\|$$

for any $\mathbf{x} \in \mathbb{C}^N$ and any index set $S \subseteq [N]$ of size s.

12

Random Sampling in Bounded Orthonormal Systems

We have seen in the previous chapters that subgaussian random matrices provide optimal measurement matrices for compressive sensing. While this is a very important insight for the theory, the use of such type of "completely random" matrices, where all entries are independent, is limited for practical purposes. Indeed, subgaussian random matrices do not possess any structure. However, structure is important for several reasons:

- Applications may impose certain structure on the measurement matrix due to physical or other constraints.
- Structure of the measurement matrix often allows to have fast matrix-vector multiplication algorithms — exploiting for instance the fast Fourier transform (FFT) — for both the matrix itself and its adjoint. This is crucial for speed-ups in any recovery algorithm (including ℓ_1-minimization), and only in this situation can large scale problems be treated with compressive sensing techniques.
- For large unstructured matrices difficulties in storing the matrix entries arise, while a structured matrix is usually generated by a number of parameters much smaller than the number of matrix entries, so that it is much easier to store.

From this point of view, it is important to investigate whether certain structured random matrices may provide similar recovery guarantees as the ones for subgaussian random matrices. By a structured random matrix, we mean a structured matrix that is generated by a random choice of parameters.

An important setup at the core and the origin of the field, that we will study exclusively below, arises from random sampling of functions whose expansion into a bounded orthonormal system (see the precise definition below) is sparse or compressible. Special cases consist in sampling of sparse trigonometric polynomials and in recovery of sparse vectors from random samples of its Fourier transform. The associated random sampling matrix is then a random partial Fourier matrix, and it has a fast matrix vector multiplication routine using the FFT. The analysis of the resulting random measurement ma-

trices becomes more involved than the one for subgaussian random matrices because the entries are not independent anymore. In this context nonuniform recovery results are simpler to derive than uniform recovery results based on the restricted isometry property. We will proceed by increasing difficulty of the proofs.

Other types of structured random matrices, including partial random circulant matrices, will be discussed briefly in the Notes section.

12.1 Bounded Orthonormal Systems

An important class of structured random matrices is connected with random sampling of functions in certain finite-dimensional function spaces. We require an orthonormal basis of functions that are uniformly bounded in the L^∞-norm. The most prominent example consists of the trigonometric system. In a discrete setup, the resulting matrix is a random partial Fourier matrix, which was the first structured random matrix investigated in compressive sensing.

Let $\mathcal{D} \subset \mathbb{R}^d$ be endowed with a probability measure ν. Further, let $\Phi = \{\phi_1, \dots, \phi_N\}$ be an orthonormal system of complex-valued functions on $\mathcal{D}$, that is, for $j, k \in [N]$,

$$\int_{\mathcal{D}} \phi_j(\mathbf{t})\overline{\phi_k(\mathbf{t})} d\nu(\mathbf{t}) = \delta_{j,k} = \begin{cases} 0 & \text{if } j \neq k\,, \\ 1 & \text{if } j = k\,. \end{cases} \tag{12.1}$$

Definition 12.1. *We call $\Phi = \{\phi_1, \dots, \phi_N\}$ a bounded orthonormal system (BOS) with constant K if it satisfies* (12.1) *and if*

$$\|\phi_j\|_\infty := \sup_{\mathbf{t}\in\mathcal{D}} |\phi_j(\mathbf{t})| \leq K \quad \text{for all } j \in [N]\,. \tag{12.2}$$

The smallest value that the constant K can take is $K = 1$. Indeed,

$$1 = \int_{\mathcal{D}} |\phi_j(\mathbf{t})|^2 d\nu(\mathbf{t}) \leq \sup_{\mathbf{t}\in\mathcal{D}} |\phi_j(\mathbf{t})|^2 \int_{\mathcal{D}} d\nu(\mathbf{t}) \leq K^2\,.$$

In the extreme case $K = 1$ we necessarily have $|\phi_j(\mathbf{t})| = 1$ for ν-almost all $\mathbf{t} \in \mathcal{D}$ as revealed by the same chain of inequalities.

Note that *some* bound K can be found for most reasonable sets of functions $\{\phi_j, j \in [N]\}$. The crucial point of the boundedness condition (12.2) is that K should ideally be independent of N. Intuitively, such a condition excludes for instance that the functions ϕ_j are very localized in small regions of $\mathcal{D}$.

We consider functions of the form

$$f(\mathbf{t}) = \sum_{k=1}^{N} x_k \phi_k(\mathbf{t})\,, \quad \mathbf{t} \in \mathcal{D}. \tag{12.3}$$

Let $\mathbf{t}_1, \dots, \mathbf{t}_m \in \mathcal{D}$ be some sampling points and suppose we have given the sample values

$$y_\ell = f(\mathbf{t}_\ell) = \sum_{k=1}^{N} x_k \phi_k(\mathbf{t}_\ell) \;, \quad \ell \in [m] \;.$$

Introducing the *sampling matrix* $\mathbf{A} \in \mathbb{C}^{m \times N}$ with entries

$$A_{\ell,k} = \phi_k(\mathbf{t}_\ell) \;, \qquad \ell \in [m], k \in [N] \;, \tag{12.4}$$

the vector $\mathbf{y} = [y_1, \ldots, y_m]^\top$ of sample values (measurements) can be written in the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} \;, \tag{12.5}$$

where $\mathbf{x} = [x_1, \ldots, x_N]^\top$ is the vector of coefficients in (12.3).

Our task is to reconstruct the function f, or equivalently its vector $\mathbf{x}$ of coefficients, from the vector of samples $\mathbf{y}$. We wish to perform this task with as few samples as possible. Without further assumptions, this is impossible if $m < N$. As common in this book we therefore introduce sparsity.

A function f of the form (12.3) is called s-sparse with respect to $(\phi_1, \ldots, \phi_N)$ if its coefficient vector $\mathbf{x}$ is s-sparse. . The problem of recovering an s-sparse function from m sample values, reduces to the compressive sensing problem with measurement matrix given by the sampling matrix $\mathbf{A}$ in (12.4).

Since it is to date open to derive good compressive sensing results for deterministic matrices, we now introduce randomness. We assume that the sampling points $\mathbf{t}_1, \ldots, \mathbf{t}_m$ are selected independently at random according to the probability measure ν. This means that $\mathbb{P}(\mathbf{t}_\ell \in B) = \nu(B)$, $\ell \in [m]$, for a measurable subset $B \subset \mathcal{D}$. We call the associated matrix (12.4) then the *random sampling matrix* associated to a BOS with constant $K \geq 1$. Note that this matrix has stochastically independent rows, but the entries within each row are not independent. Indeed, for fixed ℓ the entries $A_{\ell,k}$, $k \in [N]$, all depend on the single random sampling point $\mathbf{t}_\ell$.

Before continuing with the general theory, we give some important examples of bounded orthonormal systems.

1. **Trigonometric Polynomials.** Let $\mathcal{D} = [0, 1]$ and for $k \in \mathbb{Z}$ set

   $$\phi_k(t) = e^{2\pi i k t} \;.$$

   The probability measure ν is the Lebesgue measure on $[0, 1]$. Then for all $j, k \in \mathbb{Z}$,

   $$\int_0^1 \phi_k(t)\overline{\phi_j(t)}dt = \delta_{j,k} \;. \tag{12.6}$$

   The constant in (12.2) is $K = 1$. For a subset $\Gamma \subset \mathbb{Z}$ of size N we then consider the trigonometric polynomials of the form

   $$f(t) = \sum_{k \in \Gamma} x_k \phi_k(t) = \sum_{k \in \Gamma} x_k e^{2\pi i k t} \;.$$

The common choice $\Gamma = \{-q, -q+1, \ldots, q-1, q\}$ results in trigonometric polynomials of degree at most q (then $N = 2q+1$). We emphasize, however, that an arbitrary choice of $\Gamma \subset \mathbb{Z}$ of size $\mathrm{card}(\Gamma) = N$ is possible. Introducing sparsity on the coefficient vector $\mathbf{x} \in \mathbb{C}^N$ then leads to the notion of s-sparse trigonometric polynomials.

The sampling points $t_1, \ldots, t_m$ will be chosen independently and uniformly at random from $[0,1]$. The entries of the associated structured random matrix $\mathbf{A}$ are given by

$$A_{\ell,k} = e^{2\pi i k t_\ell} , \quad \ell \in [m] , \quad k \in \Gamma . \tag{12.7}$$

Such a matrix $\mathbf{A}$ is a Fourier type matrix, sometimes also called nonequispaced Fourier matrix.

This example extends to multivariate trigonometric polynomials on $[0,1]^d$, $d \in \mathbb{N}$. Indeed, the monomials $\phi_{\mathbf{k}}(\mathbf{t}) = e^{2\pi i \langle \mathbf{k}, \mathbf{t} \rangle}$, $\mathbf{k} \in \mathbb{Z}^d$, $\mathbf{t} \in [0,1]^d$, form an orthonormal system on $[0,1]^d$. For readers familiar with abstract harmonic analysis we mention that this example can be further generalized to characters of a compact commutative group. The corresponding measure will be the Haar measure of the group.

2. **Real Trigonometric Polynomials.** Instead of the complex exponentials above, we may also take the real functions

$$\begin{aligned} \phi_{2k}(t) &= \sqrt{2}\cos(2\pi k t) , \quad k \in \mathbb{N} , \quad \phi_0(t) = 1 , \\ \phi_{2k-1}(t) &= \sqrt{2}\sin(2\pi k t) , \quad k \in \mathbb{N} . \end{aligned} \tag{12.8}$$

They also form an orthonormal system on $\mathcal{D} = [0,1]$ with respect to the Lebesgue measure and the constant in (12.2) is $K = \sqrt{2}$. The sampling points $t_1, \ldots, t_m$ are chosen again according to the uniform distribution on $[0,1]$.

3. **Discrete Orthonormal Systems.** Let $\mathbf{U} \in \mathbb{C}^{N \times N}$ be a unitary matrix. The normalized columns $\sqrt{N}\mathbf{u}_k \in \mathbb{C}^N$, $k \in [N]$, form an orthonormal system with respect to the discrete uniform measure on $[N]$, $\nu(B) = \mathrm{card}(B)/N$ for $B \subset [N]$, i.e.,

$$\frac{1}{N} \sum_{t=1}^{N} \sqrt{N}\mathbf{u}_k(t) \overline{\sqrt{N}\mathbf{u}_\ell(t)} = \langle \mathbf{u}_k, \mathbf{u}_\ell \rangle = \delta_{k,\ell} , \quad k, \ell \in [N] .$$

Here, $\mathbf{u}_k(t) := U_{t,k}$ denotes the tth entry of the kth column of $\mathbf{U}$. The boundedness condition (12.2) requires that the normalized entries of $\mathbf{U}$ are bounded, i.e.,

$$\sqrt{N} \max_{k,t \in [N]} |U_{tk}| = \max_{k,t \in [N]} |\sqrt{N}\mathbf{u}_k(t)| \le K . \tag{12.9}$$

Choosing the points $t_1, \ldots, t_m$ independently and uniformly at random from $[N]$ corresponds then to creating the random matrix $\mathbf{A}$ by selecting

its rows independently and uniformly at random from the rows of $\sqrt{N}U$, that is,

$$\mathbf{A} = \sqrt{N}\mathbf{R}_T\mathbf{U} ,$$

where $T = \{t_1, \dots, t_m\}$ and $\mathbf{R}_T : \mathbb{C}^N \to \mathbb{C}^m$ denotes the random subsampling operator

$$(\mathbf{R}_T\mathbf{z})_\ell = z_{t_\ell} , \quad \ell \in [m] . \tag{12.10}$$

Compressive sensing in this context corresponds to the situation where only the entries of $\tilde{\mathbf{y}} = \sqrt{N}\mathbf{U}\mathbf{x} \in \mathbb{C}^N$ on T are observed for an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$. In other words, $\mathbf{y} = \mathbf{R}_T\tilde{\mathbf{y}} \in \mathbb{C}^m$, and we wish to recover $\mathbf{x}$ from $\mathbf{y}$.

Note that it may happen with non-zero probability that a row of $\sqrt{N}\mathbf{U}$ is selected more than once because the probability measure is discrete in this example. Hence, $\mathbf{A}$ is allowed to have repeated rows. One can avoid this effect by passing to a different probability model where the subset $\{t_1, \dots, t_m\} \subset [N]$ is selected uniformly at random among all subsets of $[N]$ of cardinality m. This probability model requires a slightly different analysis than the model described above, and we will discuss such issues at the end of this section. However, the difference between the two models is very slight and the final recovery results are almost the same. We refer to Section 12.6 for details.

4. **Partial Discrete Fourier Transform.** An important example of the setup just described is the partial discrete Fourier matrix $\mathbf{F} \in \mathbb{C}^{N\times N}$ with entries

$$F_{\ell,k} = \frac{1}{\sqrt{N}} e^{2\pi i \ell k/N}, \quad \ell, k \in [N] . \tag{12.11}$$

The Fourier matrix $\mathbf{F}$ is unitary, see Exercise 12.1. The constant in the boundedness condition (12.9) is clearly $K = 1$. The result $\hat{\mathbf{x}} = \mathbf{F}\mathbf{x}$ is called the Fourier transform of $\mathbf{x}$. Applying the setup of the previous example to this situation, results in the problem of reconstructing a sparse vector $\mathbf{x}$ from m random entries of its Fourier transform $\hat{\mathbf{x}}$, that are independent and uniformly distributed on $\mathbb{Z}_N := \{1, \dots, N\}$. The resulting matrix $\mathbf{A}$ is called *random partial Fourier matrix*. Such a matrix can also be seen as a special case of the nonequispaced Fourier type matrix in (12.7) with the points t_ℓ being chosen from the grid $\mathbb{Z}_N$ instead of the whole interval $[0, 1]$. Note that the discrete Fourier matrix in (12.11) can also be extended to higher dimensions, i.e., to grids $\mathbb{Z}_N^d$ for $d \in \mathbb{N}$.

A crucial point for applications is that the Fourier transform can be computed quickly using the FFT. It computes the Fourier transform of a vector $\mathbf{x} \in \mathbb{C}^N$ in complexity $\mathcal{O}(N \ln N)$. Then also a partial Fourier matrix $\mathbf{A} = \mathbf{R}_T\mathbf{F}$ has a fast matrix vector multiplication. Simply compute $\mathbf{F}\mathbf{x}$ via the FFT and then omit all entries outside T. Similarly, the application of the adjoint, $\mathbf{A}^*\mathbf{y}$, can be evaluated fast by extending the vector $\mathbf{y}$ with zeros outside T and then applying $\mathbf{F}^*$, which can also be computed via the FFT.

5. **Hadamard Transform.** The Hadamard transform $\mathbf{H} = \mathbf{H}_d \in \mathbb{R}^{2^d \times 2^d}$ can be seen as a Fourier transform on $\mathbb{Z}_2^d = \{0,1\}^d$. Writing out indices $j, \ell \in [2^d]$ into a binary expansion,

$$j = \sum_{k=1}^{d} j_k 2^{k-1} + 1 \quad \text{and} \quad \ell = \sum_{k=1}^{d} \ell_l 2^{\ell-1} + 1$$

with $j_k, \ell_k \in \{0,1\}$, an entry $H_{j,\ell}$ of the Hadamard matrix $\mathbf{H}_d$ is given by

$$H_{j,\ell} = \frac{1}{2^{d/2}} (-1)^{\sum_{k=1}^{d} j_k \ell_k} .$$

The Hadamard matrix is orthogonal and self-adjoint, that is, $\mathbf{H}_d = \mathbf{H}_d^* = \mathbf{H}_d^{-1}$. The constant in (12.2) or (12.9) is once more $K = 1$. The Hadamard transform also has a fast matrix-vector multiplication algorithm, which operates in complexity $\mathcal{O}(N \ln N)$, where $N = 2^d$. The algorithm uses recursively the identity

$$\mathbf{H}_d = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{H}_{d-1} & \mathbf{H}_{d-1} \\ \mathbf{H}_{d-1} & -\mathbf{H}_{d-1} \end{pmatrix} , \quad \mathbf{H}_0 = 1 ,$$

which can be taken as an alternative recursive definition for the Hadamard matrix. A slightly different description of the Hadamard transform will be discussed in the next section.

6. **Incoherent Bases.** Let $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{N \times N}$ be two unitary matrices. Their columns $(\mathbf{v}_\ell)_{\ell=1}^N$ and $(\mathbf{w}_\ell)_{\ell=1}^N$ form two orthonormal bases of $\mathbb{C}^N$. Assume that a vector $\mathbf{z} \in \mathbb{C}^N$ is sparse with respect to the basis $(\mathbf{v}_\ell)$ rather than the canonical basis, that is, $\mathbf{z} = \mathbf{V}\mathbf{x}$ for a sparse vector $\mathbf{x}$. Further, assume that $\mathbf{z}$ is sampled with respect to the basis $(\mathbf{w}_\ell)$, i.e., we obtain measurements

$$y_k = \langle \mathbf{z}, \mathbf{w}_{t_k} \rangle , \quad k \in [m]$$

with $T := \{t_1, \ldots, t_m\} \subset [N]$. In matrix vector form this can be written

$$\mathbf{y} = \mathbf{R}_T \mathbf{W}^* \mathbf{z} = \mathbf{R}_T \mathbf{W}^* \mathbf{V} \mathbf{x} ,$$

where $\mathbf{R}_T$ is again the sampling operator (12.10). Defining the unitary matrix $\mathbf{U} := \mathbf{W}^* \mathbf{V} \in \mathbb{C}^{N \times N}$ we are back to the situation of the third example. The condition (12.9) now reads

$$\sqrt{N} \max_{\ell,k \in [N]} |\langle \mathbf{v}_\ell, \mathbf{w}_k \rangle| \leq K . \tag{12.12}$$

The bases $(\mathbf{v}_\ell), (\mathbf{w}_\ell)$ are called incoherent if K can be chosen small. The two previous examples fall into this setting by choosing one of the bases as the canonical basis, $\mathbf{W} = \mathbf{Id} \in \mathbb{C}^{N \in N}$. The Fourier basis and the canonical basis are actually maximally incoherent, since $K = 1$ in this case.

Further examples, namely Haar wavelets in connection with noiselets as well as Legendre polynomials will be mentioned in the Notes section.

We recall that Figure 1.2 in Chapter 1 shows an example of exact recovery of a 10-sparse vector in dimension 300 from 30 Fourier samples (example (iv) above) using ℓ_1-minimization. For comparison the reconstruction via ℓ_2-minimization is also shown.

12.2 Uncertainty Principles and Lower Bounds

In this section we concentrate essentially on the Fourier system of Example 4 and on the Hadamard matrix of Example 5 in the previous section in order to illustrate some basic facts and bounds that arise in random sampling of bounded orthonormal systems. In particular, we provide lower bounds on the minimal number of measurements, see (12.29), which are slightly stronger than the ones obtained in Chapter 10 in the general setup using Gelfand widths.

We recall that $\mathbf{F} \in \mathbb{C}^{N \times N}$ is the Fourier transform matrix with entries

$$F_{\ell,k} = \frac{1}{\sqrt{N}} e^{2\pi i \ell k/N}, \quad \ell, k \in [N] \, .$$

(Note that in order to be consistent with the general notation in this book, we use the index set $[N] = \{1, \dots, N\}$, although in the literature one often finds the index set $\{0, 1, \dots, N-1\}$ in connection with the Fourier matrix.) With the stated normalization $\mathbf{F}$ is unitary. For a vector $\mathbf{x} \in \mathbb{C}^N$, its Fourier transform is denoted

$$\widehat{\mathbf{x}} = \mathbf{F}\mathbf{x}.$$

Uncertainty principles state that a vector cannot be simultaneously localized both in time and frequency. In other words, it is impossible that both $\mathbf{x}$ and $\widehat{\mathbf{x}}$ are concentrated in a small portion of $[N]$. Various versions of the uncertainty principle make the notion of localization precise.

We present a general discrete version for incoherent bases (see Example 6 above). Let $\mathbf{V} = (\mathbf{v}_1 | \cdots | \mathbf{v}_N)$, $\mathbf{W} = (\mathbf{w}_1 | \cdots | \mathbf{w}_N) \in \mathbb{C}^{N \times N}$ be two unitary matrices that are mutually incoherent, that is,

$$\sqrt{N} \max_{\ell,k \in [N]} |\langle \mathbf{v}_\ell, \mathbf{w}_k \rangle| \leq K \tag{12.13}$$

for some small $K \geq 1$. Taking the pairs of Fourier and identity matrix, $\mathbf{V} = \mathbf{F}, \mathbf{W} = \mathbf{Id}$, we get the optimal constant $K = 1$.

Theorem 12.2. *Let $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{N \times N}$ be two mutually incoherent unitary matrices with parameter K in* (12.13). *Let $\mathbf{y} \in \mathbb{C}^N \setminus \{0\}$ and $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ be the representation coefficients in $\mathbf{y} = \mathbf{V}\mathbf{x} = \mathbf{W}\mathbf{z}$. Then*

$$\|\mathbf{x}\|_0 + \|\mathbf{z}\|_0 \geq \frac{2\sqrt{N}}{K} \, . \tag{12.14}$$

Proof. Since $\mathbf{V}$ is unitary, left multiplication of the identity $\mathbf{Vx} = \mathbf{Wz}$ by $\mathbf{V}^*$ yields $\mathbf{x} = \mathbf{V}^*\mathbf{Wz}$. An entry of $\mathbf{x}$ satisfies

$$|x_k| = |(\mathbf{V}^*\mathbf{Wz})_k| = |\sum_\ell (\mathbf{V}^*\mathbf{W})_{k,\ell} z_\ell| \le \sum_\ell |\langle \mathbf{w}_\ell, \mathbf{v}_k\rangle||z_\ell|$$
$$\le \max_{\ell,k} |\langle \mathbf{w}_\ell, \mathbf{v}_k\rangle| \, \|\mathbf{z}\|_1 \le \frac{K}{\sqrt{N}} \|\mathbf{z}\|_1 \, .$$

Summation over $k \in \operatorname{supp} \mathbf{x}$ yields

$$\|\mathbf{x}\|_1 \le \|\mathbf{x}\|_0 \frac{K}{\sqrt{N}} \|\mathbf{z}\|_1 \, .$$

Left-multiplication by $\mathbf{W}^*$ of $\mathbf{Vx} = \mathbf{Wz}$ similarly yields

$$\|\mathbf{z}\|_1 \le \|\mathbf{z}\|_0 \frac{K}{\sqrt{N}} \|\mathbf{x}\|_1 \, .$$

Multiplication of both inequalities and division by $\|\mathbf{x}\|_1 \|\mathbf{z}\|_1$ implies the inequality $1 \le \|\mathbf{z}\|_0 \|\mathbf{x}\|_0 K^2/N$, or expressed differently

$$\sqrt{\|\mathbf{z}\|_0 \|\mathbf{x}\|_0} \ge \frac{\sqrt{N}}{K} \, .$$

Using that the arithmetic mean dominates the geometric mean we obtain that

$$\frac{\|\mathbf{z}\|_0 + \|\mathbf{x}\|_0}{2} \ge \sqrt{\|\mathbf{z}\|_0 \|\mathbf{x}\|_0} \ge \frac{\sqrt{N}}{K} \, .$$

This completes the proof. □

Specializing to the pair of identity matrix and Fourier matrix, for which $K = 1$, we arrive at the next consequence.

Corollary 12.3. *Let $\mathbf{x} \in \mathbb{C}^N \setminus \{0\}$. Then*

$$\|\mathbf{x}\|_0 + \|\widehat{\mathbf{x}}\|_0 \ge 2\sqrt{N} \, , \tag{12.15}$$

where $\widehat{\mathbf{x}} = \mathbf{Fx}$ is the discrete Fourier transform of $\mathbf{x}$.

This uncertainty principle has consequences for signal separation (Exercise 12.2) and it implies a weak result concerning recovery from undersampled measurements (Exercise 12.3). Our motivation for the above statements is rather that they have converses that motivate to consider random sets of samples. Indeed, the bound (12.15) cannot be improved in general, since the next proposition shows that it is sharp for so-called *delta trains*.

Proposition 12.4. *Let $N = n^2$ be a square. Set $\mathbf{x} \in \mathbb{C}^N$ to be the vector with entries*

$$x_\ell = \begin{cases} 1 & \text{if } \ell = 0 \mod n, \\ 0 & \text{otherwise} \, . \end{cases} \tag{12.16}$$

Then $\widehat{\mathbf{x}} = \mathbf{x}$ and $\|\mathbf{x}\|_0 = \|\widehat{\mathbf{x}}\|_0 = \sqrt{N}$.

Proof. By definition of the Fourier transform we have, for $j \in [n^2]$,

$$\widehat{x}_j = \frac{1}{n} \sum_{\ell=1}^{n^2} x_\ell e^{2\pi i \ell j / n^2} = \frac{1}{n} \sum_{k=1}^{n} e^{2\pi i k j / n} = \begin{cases} 1 & \text{if } j = 0 \mod n \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

This shows that $\widehat{\mathbf{x}} = \mathbf{x}$. □

Using delta trains we can illustrate why one cannot work with arbitrary sampling sets $T \subset [N]$ for sparse recovery from Fourier measurements. Suppose that $N = n^2$ is a square, and let $\mathbf{x}$ be defined as in (12.16). We consider the set of sampling points $T := [n^2] \setminus \{n, 2n, \ldots, n^2\}$. Then by the previous proposition, the restriction of $\widehat{\mathbf{x}}$ to T is the zero vector, that is,

$$\mathbf{y} = \mathbf{R}_T \mathbf{F} \mathbf{x} = 0 \,.$$

Any reasonable algorithm will output $\mathbf{x}^\sharp = 0$ from $\mathbf{y} = 0$. In other words, this sampling scheme cannot distinguish $\mathbf{x}$ from the zero vector. Observe that $s = \|\mathbf{x}\|_0 = n$, but the number of samples satisfies

$$m = \operatorname{card}(T) = n^2 - n \,.$$

In conclusion, for this choice of sampling set not even $m = s^2 - s$ samples are sufficient. This example gives an indication why we move to random choices of sampling sets T. Indeed, the sampling set of the example is very structured, and this is essentially the reason why it allows counterexamples. Good sampling sets rather possess only very little additive structure, and the simplest way to construct an unstructured set of numbers is to choose it at random.

Next we investigate a general lower bound on the number m of samples for s-sparse recovery in dimension N. We have seen in Chapter 10 that for a general stable sparse recovery problem we have the lower bound

$$m \geq Cs \ln(N/s) \,.$$

We will construct an example that shows that the term $\ln(N/s)$ has to be replaced by $\ln N$ in the context of random sampling in bounded orthonormal systems. To this end, we use the Hadamard transform $\mathbf{H}$ introduced in Example 5.

The Hadamard transform is related to Fourier analysis on the additive group $\mathbb{Z}_2^n = (\{0,1\}^n, +)$, which has cardinality $N = 2^n$. Addition is understood modulo 2. We give here a slightly different description of the Hadamard matrix than in the previous section. The constant function $\chi_0 = 1$ on $\mathbb{Z}_2$ and the function

$$\chi_1(t) = \begin{cases} 1 & \text{if } t = 0 \,, \\ -1 & \text{if } t = 1 \,. \end{cases}$$

are the characters on $\mathbb{Z}_2$, that is, $\chi_j(t+r) = \chi_j(t)\chi_j(r)$ for $j, t, r \in \{0,1\}$. We also observe that $\chi_{j+k}(t) = \chi_j(t)\chi_k(t)$. One easily checks that the characters

are orthonormal with respect to the normalized counting measure on $\mathbb{Z}_2$, that is,

$$\langle \chi_j, \chi_k \rangle := \frac{1}{2} \sum_{t \in \{0,1\}} \chi_j(t)\chi_k(t) = \delta_{jk} \ .$$

For $\mathbf{j}, \mathbf{t} \in \mathbb{Z}_2^n$ we define a character of $\mathbb{Z}_2^n$ as the tensor product

$$\chi_{\mathbf{j}}(\mathbf{t}) = \prod_{\ell=1}^{n} \chi_{j_\ell}(t_\ell) \ .$$

By the corresponding properties on $\mathbb{Z}_2$ we have

$$\chi_{\mathbf{j}}(\mathbf{t}+\mathbf{r}) = \chi_{\mathbf{j}}(\mathbf{t})\chi_{\mathbf{j}}(\mathbf{r}) \quad \text{and} \quad \chi_{\mathbf{j}+\mathbf{k}}(\mathbf{t}) = \chi_{\mathbf{j}}(\mathbf{t})\chi_{\mathbf{k}}(\mathbf{t}) \ . \tag{12.17}$$

It follows from the orthonormality of the χ_j that these functions are orthonormal with respect to the counting measure on $\mathbb{Z}_2^n$, that is,

$$\langle \chi_{\mathbf{j}}, \chi_{\mathbf{k}} \rangle = 2^{-n} \sum_{\mathbf{t} \in \mathbb{Z}_2^n} \chi_{\mathbf{j}}(\mathbf{t})\chi_{\mathbf{k}}(\mathbf{t}) = \delta_{\mathbf{j},\mathbf{k}} \ . \tag{12.18}$$

The uniform bound of these functions is $K = 1$. The (unnormalized) Hadamard transform (Fourier transform on $\mathbb{Z}_2^n$) of a vector $\mathbf{x}$ indexed by $\mathbb{Z}_2^n$ is then defined entry-wise as

$$z_{\mathbf{j}} = (\mathbf{H}\mathbf{x})_{\mathbf{j}} = 2^{-n/2} \sum_{\mathbf{t} \in \mathbb{Z}_2^n} x_{\mathbf{t}} \chi_{\mathbf{j}}(\mathbf{t}) \ .$$

Key to our lower estimate is the fact that an arbitrary subset of $\mathbb{Z}_2^n$ contains (the translate of) a large subgroup of $\mathbb{Z}_2^n$.

Lemma 12.5. *For any subset Λ of $\mathbb{Z}_2^n$, if $N := \mathrm{card}(\mathbb{Z}_2^n) = 2^n$ and if $\kappa := \mathrm{card}(\Lambda)/N$ satisfies $\log_2(\kappa^{-1}) \geq 10\, N^{-3/4}$, then there exist an element $\mathbf{b} \in \mathbb{Z}_2^n$ and a subgroup Γ of $\mathbb{Z}^n$ such that*

$$\mathbf{b} + \Gamma \subset \Lambda \quad \textit{and} \quad \mathrm{card}(\Gamma) \geq \frac{n}{8 \log_2(\kappa^{-1})}. \tag{12.19}$$

Proof. We iteratively construct elements $\gamma_0, \gamma_1, \ldots, \gamma_p \in \mathbb{Z}_2^n$ and subsets $\Lambda_0, \Lambda_1, \ldots, \Lambda_p$ of $\mathbb{Z}_2^n$ as follows: we set $\gamma_0 = 0$ and $\Lambda_0 := \Lambda$, and, for $j \geq 1$, with $G(\gamma_0, \ldots, \gamma_{j-1})$ denoting the group generated by $\gamma_0, \ldots, \gamma_{j-1}$, we define

$$\gamma_j := \mathrm{argmax}\ \mathrm{card}((\gamma + \Lambda_{j-1}) \cap \Lambda_{j-1}), \quad \gamma \notin G(\gamma_0, \ldots, \gamma_{j-1}), \tag{12.20}$$

$$\Lambda_j := (\gamma_j + \Lambda_{j-1}) \cap \Lambda_{j-1}. \tag{12.21}$$

The condition $\gamma \notin G(\gamma_0, \ldots, \gamma_{j-1})$ guarantees that $G(\gamma_0, \ldots, \gamma_j)$ is twice as large as $G(\gamma_0, \ldots, \gamma_{j-1})$, so that $\mathrm{card}(G(\gamma_0, \ldots, \gamma_j)) = 2^j$ follows by induction. Therefore, the construction of $\gamma_1, \ldots, \gamma_p$ via (12.20) is possible as long as

$2^{p-1} < N$, and in particular for p chosen as in (12.23) below. Let us now show that property (12.21) implies, for $j \geq 1$,

$$\Lambda_j + G(\gamma_0, \ldots, \gamma_j) \subseteq \Lambda_{j-1} + G(\gamma_0, \ldots, \gamma_{j-1}). \tag{12.22}$$

Indeed, for $\mathbf{g} \in \Lambda_j + G(\gamma_0, \ldots, \gamma_j)$, we write $\mathbf{g} = \lambda_j + \sum_{\ell=1}^{j} \delta_\ell \gamma_\ell$ for some $\lambda_j \in \Lambda_j$ and some $\delta_1, \ldots, \delta_j \in \{0,1\}$. In view of $\Lambda_j = (\gamma_j + \Lambda_{j-1}) \cap \Lambda_{j-1}$, we can always write $\lambda_j = \lambda_{j-1} + \delta_j \gamma_j$ for some $\lambda_{j-1} \in \Lambda_{j-1}$ — if $\delta_j = 0$, we use $\lambda_j \in \Lambda_{j-1}$, and if $\delta_j = 1$, we use $\lambda_j \in \gamma_j + \Lambda_{j-1}$. It follows that

$$\mathbf{g} = \lambda_{j-1} + \delta_j \gamma_j + \sum_{\ell=1}^{j} \delta_\ell \gamma_\ell = \lambda_{j-1} + \sum_{\ell=1}^{j-1} \delta_\ell \gamma_\ell \in \Lambda_{j-1} + G(\gamma_0, \ldots, \gamma_{j-1}).$$

This establishes (12.22). We derive that $\Lambda_p + G(\gamma_0, \ldots, \gamma_p) \subseteq \Lambda_0 + G(\gamma_0) = \Lambda$ by immediate induction. Thus, choosing $\Gamma = G(\gamma_0, \ldots, \gamma_p)$ and picking any $\mathbf{b} \in \Lambda_p$, we have $\mathbf{b} + \Gamma \subset \Lambda$. It remains to prove that the size of Γ is large, and that an element $\mathbf{b} \in \Lambda_p$ does exist. By considering $p \geq 0$ such that

$$2^{p-1} < \frac{n}{8 \log_2(\kappa^{-1})} \leq 2^p, \tag{12.23}$$

we immediately obtain the second part of (12.19). To show that $\mathrm{card}(\Lambda_p) > 0$, we use property (12.21). For $j \geq 1$, the observation that the maximum is larger than the average leads to

$$\begin{aligned}
\mathrm{card}(\Lambda_j) &\geq \frac{1}{N - 2^{j-1}} \sum_{\gamma \in \mathbb{Z}_2^n \setminus G(\gamma_0, \ldots, \gamma_{j-1})} \mathrm{card}((\gamma + \Lambda_{j-1}) \cap \Lambda_{j-1}) \\
&= \frac{1}{N - 2^{j-1}} \Bigg[\sum_{\gamma \in \mathbb{Z}_2^n} \mathrm{card}((\gamma + \Lambda_{j-1}) \cap \Lambda_{j-1}) \\
&\qquad - \sum_{\gamma \in G(\gamma_0, \ldots, \gamma_{j-1})} \mathrm{card}((\gamma + \Lambda_{j-1}) \cap \Lambda_{j-1}) \Bigg]
\end{aligned}$$

On the one hand, we have

$$\sum_{\gamma \in G(\gamma_0, \ldots, \gamma_{j-1})} \mathrm{card}((\gamma + \Lambda_{j-1}) \cap \Lambda_{j-1}) \leq \sum_{\gamma \in G(\gamma_0, \ldots, \gamma_{j-1})} \mathrm{card}(\Lambda_{j-1}) \leq 2^{j-1} \mathrm{card}(\Lambda_{j-1}).$$

On the other hand, with $\mathbb{1}_A$ denoting the characteristic function of a set A, we have

$$\begin{aligned}
\sum_{\gamma \in \mathbb{Z}_2^n} \mathrm{card}((\gamma + \Lambda_{j-1}) \cap \Lambda_{j-1}) &= \sum_{\gamma \in \mathbb{Z}_2^n} \sum_{\mathbf{h} \in \Lambda_{j-1}} \mathbb{1}_{\gamma + \Lambda_{j-1}}(\mathbf{h}) \\
&= \sum_{\mathbf{h} \in \Lambda_{j-1}} \sum_{\gamma \in \mathbb{Z}_2^n} \mathbb{1}_{\mathbf{h} - \Lambda_{j-1}}(\gamma) = \sum_{\mathbf{h} \in \Lambda_{j-1}} \mathrm{card}(\Lambda_{j-1}) = \mathrm{card}(\Lambda_{j-1})^2.
\end{aligned}$$

As a result, we obtain

$$\mathrm{card}(\Lambda_j) \geq \frac{\mathrm{card}(\Lambda_{j-1})}{N - 2^{j-1}} \Big[\mathrm{card}(\Lambda_{j-1}) - 2^{j-1}. \Big]$$

By induction, this implies the estimate

$$\mathrm{card}(\Lambda_j) \geq \kappa^{2^j} N \Big(1 - \frac{2^{j-1}}{N} \sum_{\ell=0}^{j-1} \kappa^{-2^\ell} \Big) . \tag{12.24}$$

Indeed, this holds for $j = 0$, and if it holds for $j-1$, then

$$\begin{aligned}
&\mathrm{card}(\Lambda_j) \\
&\geq \kappa^{2^{j-1}} N \left(1 - \frac{2^{j-2}}{N} \sum_{\ell=0}^{j-2} \kappa^{-2^\ell} \right) \frac{\kappa^{2^{j-1}} N}{N - 2^{j-1}} \left(1 - \frac{2^{j-2}}{N} \sum_{\ell=0}^{j-2} \kappa^{-2^\ell} - \frac{2^{j-1}}{\kappa^{2^{j-1}} N} \right) \\
&\geq \kappa^{2^j} N \left(1 - \frac{2^{j-1}}{N} \sum_{\ell=0}^{j-2} \frac{1}{2\kappa^{2^\ell}} \right) \left(1 - \frac{2^{j-1}}{N} \left(\sum_{\ell=0}^{j-2} \frac{1}{2\kappa^{2^\ell}} + \frac{1}{\kappa^{2^{j-1}}} \right) \right) \\
&\geq \kappa^{2^j} N \left(1 - \frac{2^{j-1}}{N} \sum_{\ell=0}^{j-1} \frac{1}{\kappa^{2^\ell}} \right) .
\end{aligned}$$

This finishes the inductive justification of (12.24). Since $\sum_{\ell=0}^{p-1} \kappa^{-2^\ell} \leq p \kappa^{-2^{p-1}}$, we derive in particular

$$\mathrm{card}(\Lambda_p) \geq \kappa^{2^p} N \Big(1 - \frac{2^{p-1}}{N} \, p \, \kappa^{-2^{p-1}} \Big) = \kappa^{2^{p-1}} \Big(\kappa^{2^{p-1}} N - 2^{p-1} p \Big) .$$

Using the leftmost inequality in (12.23), as well as $p \leq n$ and the assumption $\log_2(\kappa^{-1}) \geq 10\, N^{-3/4}$, we obtain

$$\begin{aligned}
\mathrm{card}(\Lambda_p) &\geq \kappa^{2^{p-1}} \left(\kappa^{n/(8 \log_2(\kappa^{-1}))} 2^n - \frac{n^2}{8 \log_2(\kappa^{-1})} \right) \\
&= \kappa^{2^{p-1}} \left(2^{n(1-1/8)} - \frac{n^2\, 2^{3n/4}}{80} \right) = \kappa^{2^{p-1}} 2^{3n/4} \left(2^{n/8} - \frac{n^2}{80} \right) > 0.
\end{aligned}$$

The proof is now complete. □

Remark 12.6. The condition $\log_2(\kappa^{-1}) \geq 10\, N^{-3/4}$ in the previous lemma can be replaced by any condition of the type $\log_2(\kappa^{-1}) \geq c_\beta N^{-\beta}$, $0 < \beta < 1$. This only requires to adjust the constants.

The next result is analogous to Proposition 12.4 and indicates the reason for the importance of having large subgroups.

Proposition 12.7. *Given a subgroup G of $\mathbb{Z}_2^n$, the set*

$$G^\perp := \{\boldsymbol{\lambda} \in \mathbb{Z}_2^n : \sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{g}) \neq 0\} \tag{12.25}$$

forms another subgroup of $\mathbb{Z}_n$. Furthermore, the Hadamard transform of the vector $\mathbf{x} \in \mathbb{C}^{\mathbb{Z}_2^n}$ with entries

$$x_{\mathbf{j}} = \begin{cases} 1 & \text{if } \mathbf{j} \in G\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

is given by

$$\widehat{z}_{\mathbf{k}} = \begin{cases} \operatorname{card}(G) & \text{if } \mathbf{k} \in G^\perp\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

In particular, $\|\mathbf{z}\|_0 \cdot \|\widehat{\mathbf{z}}\|_0 = \operatorname{card}(G) \cdot \operatorname{card}(G^\perp) = 2^n$.

Proof. First, we observe that $\mathbf{0} \in G^\perp$ because $\chi_{\mathbf{0}} = 1$ is the constant function and that $-\boldsymbol{\lambda} \in G^\perp$ whenever $\boldsymbol{\lambda} \in G^\perp$ because any element of $\mathbb{Z}_2^n$ is its own inverse. Then, using the fact that G is a group, we obtain, for all $\mathbf{g}, \mathbf{h} \in G$ and $\boldsymbol{\lambda} \in G^\perp$,

$$\sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{g}) = \sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{h} + \mathbf{g}) = \chi_{\boldsymbol{\lambda}}(\mathbf{h}) \sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{g})\,.$$

In view of $\sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{g}) \neq 0$, we deduce

$$\chi_{\boldsymbol{\lambda}}(\mathbf{h}) = 1 \qquad \text{for all } \mathbf{h} \in G \text{ and } \boldsymbol{\lambda} \in G^\perp\,. \tag{12.26}$$

In particular, given $\boldsymbol{\lambda}, \boldsymbol{\rho} \in G^\perp$, we derive

$$\sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}+\boldsymbol{\rho}}(\mathbf{g}) = \sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{g}) \chi_{\boldsymbol{\rho}}(\mathbf{g}) = \sum_{\mathbf{g} \in G} 1 = \operatorname{card}(G) \neq 0\,,$$

which shows that $\boldsymbol{\lambda} + \boldsymbol{\rho} \in G^\perp$. We have established that $G^\perp$ is a group. The special case $\boldsymbol{\rho} = 0$ of the previous identity reads

$$\sum_{\mathbf{g} \in G} \chi_{\boldsymbol{\lambda}}(\mathbf{g}) = \operatorname{card}(G) \qquad \text{for all } \boldsymbol{\lambda} \in G^\perp\,. \tag{12.27}$$

The definition of the unnormalized Hadamard transform then yields

$$\widehat{z}_{\mathbf{k}} = \sum_{\mathbf{g} \in G} \chi_{\mathbf{k}}(\mathbf{g}) = \begin{cases} \operatorname{card}(G) & \text{if } \mathbf{k} \in G^\perp\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Finally, summing (12.27) over all $\boldsymbol{\lambda} \in G^\perp$ and using (12.25) as well as the orthogonality relation (12.18), we obtain

$$\begin{aligned}\operatorname{card}(G)\cdot\operatorname{card}(G^\perp) &= \sum_{\boldsymbol{\lambda}\in G^\perp}\sum_{\mathbf{g}\in G}\chi_{\boldsymbol{\lambda}}(\mathbf{g}) = \sum_{\boldsymbol{\lambda}\in\mathbb{Z}_2^n}\sum_{\mathbf{g}\in\mathbb{Z}_2^n}\chi_{\boldsymbol{\lambda}}(\mathbf{g}) \\ &= \sum_{\mathbf{g}\in G}\sum_{\boldsymbol{\lambda}\in\mathbb{Z}_2^n}\chi_{\mathbf{g}}(\boldsymbol{\lambda}) = 2^n\sum_{\mathbf{g}\in G}\langle\chi_{\mathbf{g}},\chi_{\mathbf{0}}\rangle = 2^n\,.\end{aligned}$$

This completes the proof. □

Now we are in the position to provide a lower bound on the number m of measurements for recovery of s-sparse vectors in $\mathbb{C}^N$, $N = 2^n$ from samples of the Hadamard transform. The bound applies to an arbitrary (nonrandom) set of m samples.

Theorem 12.8. *Let T be an arbitrary subset of $\mathbb{Z}_2^n$ of size m. If $m \le N/2$ and $m \ge c\,N^{1/4}$ where $N = 2^n$ and $c = 10\ln 2 \approx 6.93$, then there exists a nonzero vector $\mathbf{x} \in \mathbb{C}^N$ whose Hadamard transform vanishes on T and whose sparsity obeys*

$$\|\mathbf{x}\|_0 \le \frac{16\,m}{\log_2(N)}\,. \tag{12.28}$$

Proof. We consider the set $\Lambda := \mathbb{Z}_2^n \setminus T$. With $\kappa := \operatorname{card}(\Lambda)/N = 1 - \frac{m}{N}$, the concavity of the logarithm, as well as the assumption on m, yields

$$\log_2(\kappa^{-1}) = -\log_2\left(1-\frac{m}{N}\right)\begin{cases}\ge \dfrac{m}{\ln(2)\,N} \ge 10\,N^{-3/4}\,,\\ \le \dfrac{2\,m}{N}\,.\end{cases}$$

Thus, Lemma 12.5 guarantees the existence of an element $\mathbf{b} \in \mathbb{Z}_2^n$ and a subgroup Γ of $\mathbb{Z}_2^n$ such that $\mathbf{b} + \Gamma \subset \Lambda$ and $\operatorname{card}(\Gamma) \ge n/(8\log_2(\kappa^{-1}))$. The vector $\mathbf{z} \in \mathbb{C}^{\mathbb{Z}_2^n}$ introduced in Proposition (12.7) with $G := \Gamma^\perp$ satisfies

$$\|\mathbf{z}\|_0 = \operatorname{card}(\Gamma^\perp) = \frac{N}{\operatorname{card}(\Gamma)} \le \frac{8\log_2(\kappa^{-1})\,N}{n} \le \frac{16\,m}{N}\,,$$

and consequently so does the vector $\mathbf{x} \in \mathbb{C}^{\mathbb{Z}_2^n}$ defined by $x_{\mathbf{k}} = \chi_{\mathbf{b}}(\mathbf{k})z_{\mathbf{k}}$. It remains to verify that the Hadamard transform of $\mathbf{x}$ vanishes on T. For this purpose, we notice that, for any $\mathbf{j} \in \mathbb{Z}_2^n$,

$$\widehat{x}_{\mathbf{j}} = 2^{-n/2}\sum_{\mathbf{t}\in\mathbb{Z}_2^n} x_{\mathbf{t}}\chi_{\mathbf{j}}(\mathbf{t}) = 2^{-n/2}\sum_{\mathbf{t}\in\mathbb{Z}_2^n} h_{\mathbf{t}}\chi_{\mathbf{j}+\mathbf{b}}(\mathbf{t}) = \widehat{\mathbf{h}}_{\mathbf{j}+\mathbf{b}}\,.$$

Hence, according to Proposition (12.7), we have $\widehat{x}_{\mathbf{j}} = 0$ if $\mathbf{j} + \mathbf{b} \notin G^\perp$, i.e., $\mathbf{j} \notin \mathbf{b} + \Gamma$, which does occur when $\mathbf{j} \in T$. This concludes the proof by noting that $(G^\perp)^\perp = G$. □

The result below shows that, for random sampling in bounded orthonormal systems, a factor $\ln(N)$ must appear in the in the number of measurements. This is in contrast to other measurement matrices, where the logarithmic factor can be lowered to $\ln(N/s)$, see Chapters 9 and 10.

Corollary 12.9. *Let T be an arbitrary subset of $\mathbb{Z}_2^n$ with size $m \leq N/2$. The existence of a method to recover every s-sparse vector from the samples indexed by T of its Hadamard transform imposes*

$$m > C\, s \ln(N)\,, \qquad C = \frac{1}{8\ln(2)} \approx 0.1803, \tag{12.29}$$

provided $m \geq cN^{1/4}$, $c = 10 \ln 2$.
Without restrictions on m, the existence of a stable method to recover every s-sparse vector from the samples indexed by T of its Hadamard transform imposes

$$m > Cs \ln(N)$$

for some constant C depending on the stability requirement.

Remark 12.10. Recall that by a stable recovery method we mean a mapping $\Delta : \mathbb{C}^N \to \mathbb{C}^m$ such that for given $\mathbf{A} \in \mathbb{C}^{m\times N}$ and all $\mathbf{x} \in \mathbb{C}^N$ we have the stability estimate

$$\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_1 \leq \widehat{C}\sigma_s(\mathbf{x})_1\,.$$

Proof. Suppose that $m \geq cN^{1/4}$ and that a method to recover every s-sparse vector from the samples indexed by T of its Hadamard transform exists. Let us decompose the nonzero vector $\mathbf{x} \in \mathbb{C}^{\mathbb{Z}_2^n}$ of Theorem 12.8 as $\mathbf{x} = \mathbf{u} - \mathbf{v}$ for two distinct vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathbb{Z}_2^n}$ of sparsity at most $(\|\mathbf{x}\|_0 + 1)/2$. Since the Hadamard transforms of $\mathbf{u}$ and $\mathbf{v}$ are identical on T, we must have $(\|\mathbf{x}\|_0 + 1)/2 > s$, i.e.,

$$2s \leq \|\mathbf{x}\|_0 \leq \frac{16\, m}{\log_2(N)},$$

and (12.29) follows. In the case $m \leq cN^{1/4}$, we know from Theorem 11.6 that if a stable method to recover every s-sparse vector from the samples indexed by T of its Hadamard transform exists, then there is a constant c' such that

$$m \geq c'\, s \ln(N/m) \geq c'\, s \ln(N^{3/4}/c) \geq C\, s\, \ln(N)$$

for some appropriate constant C. This concludes the proof. □

12.3 Nonuniform Recovery – Random Sign Patterns

We start with nonuniform recovery guarantees for random sampling in bounded orthonormal systems. In order to simplify the argument, we assume in this section that the signs of the nonzero coefficients of the vector to be recovered are random. Recall that the recovery condition in Theorem 4.25 depends only on the signs of $\mathbf{x}$ on its support, so that the magnitudes of the entries of $\mathbf{x}$ do not play any role. This is the reason why we impose randomness only on the signs of the entries. In this way, $\mathbf{x}$ certainly becomes random as well. But in

contrast to Chapter 13, where we focus on recovery of random signals using deterministic matrices $\mathbf{A}$, the support of $\mathbf{x}$ is still kept arbitrary here. Due to the deterministic support, the randomness in $\mathbf{x}$ can be considered mild and we will indeed remove the assumption on the randomness of the signs in the next section at the cost of a more complicated approach.

Recall that we consider the random sampling matrix $\mathbf{A}$ associated to a BOS with constant $K \geq 1$ introduced in (12.4). The sampling points $\mathbf{t}_1, \ldots, \mathbf{t}_m$ are chosen independently at random according to the probability measure ν.

Theorem 12.11. *Let $\mathbf{x} \in \mathbb{C}^N$ be s-sparse with support S, $\operatorname{card}(S) = s$, and such that its sign sequence $\operatorname{sgn}(\mathbf{x}_S)$ forms a Rademacher or Steinhaus sequence. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be the random sampling matrix associated to a BOS with constant $K \geq 1$. Assume that*

$$m \geq CK^2 s \ln^2(6N/\varepsilon) \, . \tag{12.30}$$

Then with probability at least $1 - \varepsilon$ basis pursuit recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$. The constant C is no larger than $88/3 \approx 29.33$.

In Theorem 12.18 below we will improve this result by replacing the exponent 2 by 1 at the log-factor in (12.30).

The proof of Theorem 12.11 requires some preparatory results to be provided next. As a crucial tool we use the recovery condition for individual vectors of Corollary 4.27. This requires to investigate the conditioning of the submatrix $\mathbf{A}_S$ associated to the support S of the vector to be recovered. The proof of the corresponding result stated next is based on the noncommutative Bernstein inequality of Theorem 8.14.

Theorem 12.12. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be the random sampling matrix associated to a BOS with constant $K \geq 1$. Let $S \subset [N]$ be of cardinality $\operatorname{card}(S) = s$. Then, for $\delta \in (0, 1)$, the normalized matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}} \mathbf{A}$ satisfies*

$$\| \widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id} \|_{2 \to 2} \leq \delta \, .$$

with probability at least

$$1 - 2s \exp\left(- \frac{3m\delta^2}{8K^2 s} \right) \tag{12.31}$$

Remark 12.13. Expressed differently, $\| \mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id} \|_{2 \to 2} \leq \delta$ with probability at least $1 - \varepsilon$ provided $m \geq (8/3) K^2 \delta^{-2} s \ln(2s/\varepsilon)$.

Proof. Denote $\mathbf{Y}_\ell = (\overline{\phi_j(\mathbf{t}_\ell)})_{j \in S} \in \mathbb{C}^s$ a column vector of $\mathbf{A}_S^*$. By independence of the $\mathbf{t}_\ell$, these are independent random vectors. Their ℓ_2-norm is bounded by

$$\|\mathbf{Y}_\ell\|_2 = \sqrt{\sum_{j\in S} |\phi_j(\mathbf{t}_\ell)|^2} \le K\sqrt{s}\,. \tag{12.32}$$

Furthermore, for $j,k \in S$,

$$\mathbb{E}\left(\mathbf{Y}_\ell \mathbf{Y}_\ell^*\right)_{j,k} = \mathbb{E}\left[\phi_j(\mathbf{t}_\ell)\overline{\phi_k(\mathbf{t}_\ell)}\right] = \int_{\mathcal{D}} \phi_j(\mathbf{t})\overline{\phi_k(\mathbf{t})} d\nu(\mathbf{t}) = \delta_{j,k}\,,$$

or in other words, $\mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^* = \mathbf{Id}$. Observe that

$$\widetilde{\mathbf{A}}_S^*\widetilde{\mathbf{A}}_S - \mathbf{Id} = \frac{1}{m}\sum_{\ell=1}^m (\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^*)\,.$$

The matrices $\mathbf{X}_\ell = \mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^* \in \mathbb{C}^{s\times s}$ have mean zero. Moreover,

$$\|\mathbf{X}_\ell\|_{2\to 2} = \max_{\|\mathbf{x}\|_2=1} |\langle \mathbf{Y}_\ell\mathbf{Y}_\ell^*\mathbf{x},\mathbf{x}\rangle - \|\mathbf{x}\|_2^2| = |\|\mathbf{Y}_\ell\|_2^2 - 1| \le K^2 s\,,$$

and since $\mathbf{Y}_\ell\mathbf{Y}_\ell^*\mathbf{Y}_\ell\mathbf{Y}_\ell^* = \|\mathbf{Y}_\ell\|_2^2\mathbf{Y}_\ell\mathbf{Y}_\ell^*$ we have

$$\begin{aligned}\mathbb{E}\mathbf{X}_\ell^2 &= \mathbb{E}\left(\mathbf{Y}_\ell\mathbf{Y}_\ell^*\mathbf{Y}_\ell\mathbf{Y}_\ell^* - 2\mathbf{Y}_\ell\mathbf{Y}_\ell^* + \mathbf{Id}\right) = \mathbb{E}\left((\|\mathbf{Y}_\ell\|_2^2 - 2)\mathbf{Y}_\ell\mathbf{Y}_\ell^*\right) + \mathbf{Id} \\ &\preccurlyeq (K^2 s - 2)\mathbb{E}[\mathbf{Y}_\ell\mathbf{Y}_\ell^*] + \mathbf{Id} \preccurlyeq K^2 s\,\mathbf{Id}\,. \end{aligned} \tag{12.33}$$

The variance parameter in (8.28) can therefore be estimated by

$$\sigma^2 := \left\|\sum_{\ell=1}^m \mathbb{E}(\mathbf{X}_\ell^2)\right\|_{2\to 2} \le mK^2 s\|\mathbf{Id}\|_{2\to 2} = K^2 sm\,.$$

The noncommutative Bernstein inequality (8.30) yields, for $\delta \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left(\left\|\widetilde{\mathbf{A}}_S^*\widetilde{\mathbf{A}}_S - \mathbf{Id}\right\|_{2\to 2} > \delta\right) &= \mathbb{P}\left(\|\sum_{\ell=1}^m \mathbf{X}_\ell\|_{2\to 2} > \delta m\right) \\ &\le 2s\exp\left(-\frac{\delta^2 m^2/2}{K^2 sm + K^2 s\delta m/3}\right) \le 2s\exp\left(-\frac{3}{8}\frac{\delta^2 m}{K^2 s}\right)\,.\end{aligned}$$

The proof is completed. □

The above result implies also the following coherence bound. Note that we do not require normalization of the columns (in contrast to Chapter 5). Furthermore, coherence estimates can also be shown with simpler techniques as pursued below, which do not require bounds on condition numbers, see for instance Exercise 12.5.

Corollary 12.14. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be the random sampling matrix (12.4) associated to a BOS with constant $K \ge 1$, and μ the coherence of $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$. Then*

$$\mu \le \sqrt{\frac{16K^2\ln(2N^2/\varepsilon)}{3m}}$$

with probability at least $1-\varepsilon$.

Proof. We denote the columns of $\widetilde{A}$ by $\widetilde{\mathbf{a}}_j$, $j \in [N]$. Let $S = \{j,k\}$ be a two element set. Then the matrix $\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}$ contains $\langle \widetilde{\mathbf{a}}_j, \widetilde{\mathbf{a}}_k \rangle$ as a matrix entry. Since the absolute value of any entry of a matrix is bounded by the operator norm (Lemma A.10) we have

$$|\langle \widetilde{\mathbf{a}}_j, \widetilde{\mathbf{a}}_k \rangle| \le \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} \, .$$

By Theorem 12.12 applied with $s = 2$ the probability that the operator norm on the right is not bounded by δ is at most

$$2 \times 2 \exp\left(-\frac{3m\delta^2}{8K^2 \times 2} \right) \, .$$

Taking the union bound over all $N(N-1)/2 \le N^2/2$ two element sets $S \subset [N]$ shows that

$$\mathbb{P}(\mu \ge \delta) \le 2N^2 \exp\left(-\frac{3m\delta^2}{16K^2} \right) \, .$$

Requiring that the right-hand side is at most ε leads to the desired conclusion. □

Proposition 12.15. *Let $S \subset [N]$ and let $\mathbf{x} \in \mathbb{C}^N$ be a vector support on S whose sign vector $\operatorname{sgn}(\mathbf{x}_S)$ is a Rademacher or Steinhaus sequence. If $\mathbf{A} \in \mathbb{C}^{m \times N}$ is such that $\mathbf{A}_S$ is injective and*

$$\|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 \le \alpha < 1 \quad \text{for all } \ell \notin S \, . \tag{12.34}$$

then, with probability at least

$$1 - 2N \exp(-\alpha^{-2}/2) \, ,$$

the vector $\mathbf{x}$ is the unique solution to the ℓ_1-minimization problem (P_1) *with $\mathbf{y} = \mathbf{A}\mathbf{x}$.*

Note that we need $\alpha < \big(2\ln(2N)\big)^{-1/2}$ to obtain a nontrivial statement.

Proof. In the Rademacher case, the union bound and Hoeffding's inequality (see Corollary 7.21 for the real case, and Corollary 8.8 for the general complex case) yield

$$\begin{aligned} \mathbb{P}(\max_{\ell \notin S} |\langle \mathbf{A}_S^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S) \rangle| \ge 1) &\le \sum_{\ell \notin S} \mathbb{P}\left(|\langle \mathbf{A}_S^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S) \rangle| \ge \|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 \alpha^{-1} \right) \\ &\le N 2 \exp(-\alpha^{-2}/2) \, . \end{aligned}$$

In the Steinhaus case we even obtain a better estimate from Corollary 8.10. An application of Corollary 4.27 finishes the proof. □

Remark 12.16. In Chapter 13 we will actually choose the matrix $\mathbf{A}$ deterministic and the support set S at random. Of course, also in this situation Proposition 12.15 remains applicable.

Next, we provide two conditions which ensure that $\|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2$ is small. The first condition requires that $\mathbf{A}_S$ is well-conditioned and that the coherence of $\mathbf{A}$, defined in (5.1), is small. (In contrast to Chapter 5 we do not impose $\mathbf{A}$ to have normalized columns here, although this will be satisfied in most of the later examples anyway.)

Proposition 12.17. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ with coherence μ and let $S \subset [N]$ of size s. Assume that $\|\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \le \delta$ for some $\delta \in (0,1)$. Then*

$$\|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 \le \frac{\sqrt{s}\mu}{1-\delta} \qquad \text{for all } \ell \notin S \,.$$

Proof. Since $\|\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \le \delta < 1$ the matrix $\mathbf{A}_S$ is injective and by Lemma A.13,

$$\|(\mathbf{A}_S^* \mathbf{A}_S)^{-1}\|_{2\to 2} \le \frac{1}{1-\delta} \,.$$

By definition of the operator norm

$$\begin{aligned} \|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 &= \|(\mathbf{A}_S^* \mathbf{A}_S)^{-1} \mathbf{A}_S^* \mathbf{a}_\ell\|_2 \le \|(\mathbf{A}_S^* \mathbf{A}_S)^{-1}\|_{2\to 2} \|\mathbf{A}_S^* \mathbf{a}_\ell\|_2 \\ &\le (1-\delta)^{-1} \|\mathbf{A}_S^* \mathbf{a}_\ell\|_2 \,. \end{aligned} \tag{12.35}$$

The second term in (12.35) can be estimated using the coherence, namely,

$$\|\mathbf{A}_S^* \mathbf{a}_\ell\|_2 = \sqrt{\sum_{j\in S} |\langle \mathbf{a}_\ell, \mathbf{a}_j\rangle|^2} \le \sqrt{s}\mu \,.$$

Combining the two estimates completes the proof. □

In Exercise 12.4 an alternative way of bounding the term $\|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2$ is provided. Both bounds only require that one column-submatrix of $\mathbf{A}$, or at least only a small number of them, is well-conditioned, while the restricted isometry property requires that all column-submatrices of a certain size are well-conditioned simultaneously. Indeed, it is significantly simpler to prove well-conditionedness for a single column-submatrix of a structured random matrix.

Now we are in the position to prove the nonuniform recovery result stated in Theorem 12.11.

Proof (of Theorem 12.11). Set $\alpha = \sqrt{s}u/(1-\delta)$ for some $\delta, u \in (0,1)$ to be chosen later. Let μ be the coherence of $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$. By Proposition 12.15 the probability that recovery by basis pursuit fails is upper bounded by

$$P = 2Ne^{-\alpha^{-2}/2} + \mathbb{P}\left(\|\widetilde{\mathbf{A}}_S^\dagger \widetilde{\mathbf{a}}_\ell\|_2 \geq \alpha \quad \text{for some } \ell \in [N] \setminus S\right)$$
$$\leq 2Ne^{-\alpha^{-2}/2} + \mathbb{P}(\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} > \delta) + \mathbb{P}(\mu > u)\,. \tag{12.36}$$

Here, we also used Proposition 12.17. Theorem 12.12 yields $\mathbb{P}(\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} > \delta) \leq \varepsilon/3$ under the condition

$$m \geq \frac{8K^2}{3\delta^2} s \ln(6s/\varepsilon)\,. \tag{12.37}$$

Corollary 12.14 asserts that $\mathbb{P}(\mu > u) \leq \varepsilon/3$ provided

$$m \geq \frac{16K^2}{3u^2} \ln(6N^2/\varepsilon)\,,$$

which (since $\ln(6N^2/\varepsilon) \leq 2\ln(6N/\varepsilon)$) is implied by

$$m \geq \frac{32K^2}{3t^2} \ln(6N/\varepsilon)\,. \tag{12.38}$$

Set $u = 2\delta/\sqrt{s}$. Then (12.38) implies (12.37), and $\alpha = 2\delta/(1-\delta)$. Next we set $\delta^{-2} = 11\ln(6N/\varepsilon)$. Then the first term in (12.36) is bounded by

$$2N\exp(-\alpha^{-2}/2) \leq 2N\exp\left(-\frac{(1-\delta)^2}{8\delta^2}\right)$$
$$= 2N\exp\left(-(1-(11\ln(6N/\varepsilon))^{-1/2})^2 \cdot 11\ln(6N/\varepsilon)/8\right)$$
$$\leq 2N\exp(-C\ln(6N/\varepsilon)) \leq \varepsilon/3\,,$$

where $C = 11(1-(11\ln(72))^{-1/2})^2/8 \approx 1.003 \geq 1$. Hereby, we tacitly assumed $N \geq 12$ because otherwise the statement is not interesting. (Even if $s = 1$ then the smallest possible m required by Theorem 12.11 is larger than 12, in particular it would be larger than N if $N < 12$.) Plugging the value of δ into the definition of u, that is, $u = (cs\ln(6N/\varepsilon))^{-1/2}$ with $c = 11/4$, and then into (12.38) we find that recovery by basis pursuit fails with probability at most ε provided

$$m \geq \frac{32 \cdot 11}{3 \cdot 4} K^2 s \ln^2(6N/\varepsilon)\,.$$

This completes the proof. □

Unfortunately, the exponent 2 at the log-term in (12.30) is not optimal. The next statement improves on this exponent. Unlike the previous result its proof does not use the coherence, but rather a sophisticated way of bounding the term $\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2$ using Corollary 8.42.

Theorem 12.18. *Let $\mathbf{x} \in \mathbb{C}^N$ be an s-sparse vector with support S, $\mathrm{card}(S) = s$, and such that its sign sequence $\mathrm{sgn}(\mathbf{x}_S)$ forms a Rademacher or Steinhaus sequence. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be the random sampling matrix (12.4) associated to a BOS with constant $K \geq 1$. Assume that $s \geq 36 \ln(6N/\varepsilon)$ and*

$$m \geq 18 K^2 s \ln(6N/\varepsilon) \, . \tag{12.39}$$

Then with probability at least $1 - \varepsilon$ basis pursuit recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$.

We note that the condition that $s \geq 36 \ln(6N/\varepsilon)$ is not severe. If s is smaller — meaning that it is really tiny — then we can even use the coherence bound of Corollary 12.14 together with the simple recovery condition $(2s-1)\mu$ of Theorem 5.15. If $s \leq c \ln(6N/\varepsilon)$ then $s^2 \leq cs \ln(6N/\varepsilon)$ and we obtain (uniform) recovery under the condition $m \geq C K^2 s \ln^2(6N/\varepsilon)$ for an appropriate constant C.

We start with a technical lemma.

Lemma 12.19. *With the notation of Theorem 12.18 let $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}} \mathbf{A}$. Then, for $t > 0$,*

$$\mathbb{P}\left(\max_{j \in \overline{S}} \| \widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j \|_2 \geq \sqrt{\frac{K^2 s}{m}} + t \right) \leq N \exp\left(- \frac{m t^2}{K^2 \sqrt{s}} \frac{1}{\frac{2}{\sqrt{s}} + 4 \sqrt{\frac{K^2 s}{m}} + 2t/3} \right) .$$

Proof. Fix $j \in \overline{S}$. We introduce the vectors $\mathbf{X}_\ell = (\phi_k(\mathbf{t}_\ell))_{k \in S} \in \mathbb{C}^S$ and $\mathbf{Y}_\ell = \Big(\phi_j(\mathbf{t}_\ell) \overline{\phi_k(\mathbf{t}_\ell)} \Big)_{k \in S} = \phi_j(\mathbf{t}_\ell) \mathbf{X}_\ell \in \mathbb{C}^S$. Then

$$\| \widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j \|_2 = \frac{1}{m} \left\| \sum_{\ell=1}^m \mathbf{Y}_\ell \right\|_2 \, .$$

Our aim is to apply the ℓ_2-Bernstein inequality of Corollary 8.42. The $\mathbf{Y}_\ell$ are independent copies of a single random vector $\mathbf{Y}$ that satisfies $\mathbb{E}\mathbf{Y} = 0$ by orthonormality of the ϕ_j and because $j \notin S$. It can be bounded by

$$\| \mathbf{Y} \|_2 = \| \mathbf{X} \|_2 |\langle \mathbf{e}_j, \mathbf{X} \rangle| \leq \sqrt{s} K^2 \, ,$$

by the boundedness condition (12.2) and since $\mathrm{card}(S) \leq s$. Furthermore,

$$\mathbb{E}\| \mathbf{Y} \|_2^2 = \mathbb{E}[|\langle \mathbf{X}_\ell, \mathbf{e}_j \rangle|^2 \| \mathbf{X} \|_2^2] \leq K^2 \mathbb{E} \| \mathbf{X} \|_2^2 = K^2 \sum_{k \in S} \mathbb{E} |\phi_k(\mathbf{t})|^2 = K^2 s \, .$$

For an estimate of the weak variance we observe that for a vector $\mathbf{z} \in \mathbb{C}^S$ with $\| \mathbf{z} \|_2 \leq 1$,

$$\mathbb{E} |\langle \mathbf{Y}, \mathbf{z} \rangle|^2 = \mathbb{E}[|\langle \mathbf{X}, \mathbf{e}_j \rangle|^2 |\langle \mathbf{X}, \mathbf{z} \rangle|^2] \leq K^2 \mathbb{E}[\mathbf{z}^* \mathbf{X} \mathbf{X}^* \mathbf{z}] = K^2 \| \mathbf{z} \|_2^2 \leq K^2$$

again by the orthonormality condition (12.1). Hence,

$$\sigma^2 = \sup_{\|\mathbf{z}\|_2 \le 1} \mathbb{E}|\langle \mathbf{z}, \mathbf{Y}\rangle|^2 \le K^2 \, .$$

The ℓ_2-Bernstein inequality (8.87) yields

$$\mathbb{P}(\|\sum_{\ell=1}^m \mathbf{Y}_\ell\|_2 \ge \sqrt{msK^2}+t) \le \exp\left(-\frac{t^2/2}{mK^2 + 2\sqrt{s}K^2\sqrt{msK^2} + t\sqrt{s}K^2/3}\right) \, .$$

Rescaling by $1/m$ and taking the union bound over all $j \in \overline{S}$ yields the claimed probability estimate. □

Proof (of Theorem 12.18). As suggested by Proposition 12.15, we investigate

$$\|\mathbf{A}_S^\dagger \mathbf{a}_j\|_2 \le \|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S)^{-1}\|_{2\to 2} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 \, , \quad j \notin S \, ,$$

where $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ and the $\widetilde{\mathbf{a}}_j$ denote its columns. The operator norm satisfies $\|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S)^{-1}\|_{2\to 2} \le (1-\delta)^{-1}$ provided $\|\widetilde{\mathbf{A}}_S^* \mathbf{A}_S - \mathbf{Id}\|_{2\to 2} \le \delta$, the latter being treated by Theorem 12.12. For the remaining term we set $t = \alpha K\sqrt{s/m}$ in Lemma 12.19 to obtain

$$\begin{aligned} &\mathbb{P}\left(\max_{j\in\overline{S}} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 \ge (1+\alpha)K\sqrt{\frac{s}{m}}\right) \\ &\le N \exp\left(-\frac{\alpha^2\sqrt{s}}{2/\sqrt{s} + (4+2\alpha/3)K\sqrt{s/m}}\right) \, . \end{aligned} \tag{12.40}$$

Set $v = (1+\alpha)K\sqrt{\frac{s}{m}}$. If $\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} \le \delta$ and $\max_{j\notin S} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 \le v$ then $\max_{j\notin S} \|\widetilde{\mathbf{A}}_S^\dagger \widetilde{\mathbf{a}}_j\|_2 \le v/(1-\delta)$. Therefore, by Proposition 12.15 and Theorem 12.12 the probability that basis pursuit fails to recover $\mathbf{x}$ is bounded by

$$\begin{aligned} &\mathbb{P}(\max_{j\notin S} |\langle \widetilde{\mathbf{A}}_S^\dagger \widetilde{\mathbf{a}}_j, \operatorname{sgn}(\mathbf{x}_S)\rangle| \ge 1) \\ &\le \mathbb{P}\left(\max_{j\notin S} |\langle \widetilde{\mathbf{A}}_S^\dagger \widetilde{\mathbf{a}}_j, \operatorname{sgn}(\mathbf{x}_S)\rangle| \ge 1 \,\middle|\, \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} \le \delta \right. \\ &\qquad\qquad \left. \& \max_{j\notin S} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 \le v\right) \\ &+ \mathbb{P}(\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} \ge \delta) + \mathbb{P}(\max_{j\notin S} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 \ge v) \\ &\le 2N \exp\left(-\frac{(1-\delta)^2}{2v^2}\right) + 2s \exp\left(-\frac{3m\delta^2}{8K^2 s}\right) \qquad (12.41) \\ &+ N \exp\left(-\frac{\alpha^2\sqrt{s}}{2/\sqrt{s} + (4+2\alpha/3)K\sqrt{s/m}}\right) \, . \qquad (12.42) \end{aligned}$$

Let us choose $\delta = \alpha = 1/2$. Then the second term in (12.41) is bounded by $\varepsilon/3$ provided

$$m \geq \frac{32}{3} s K^2 \ln(6s/\varepsilon) \, . \tag{12.43}$$

The first term in (12.41) does not exceed $\varepsilon/3$ provided

$$v^{-2} \geq 2(1-\delta)^{-2} \ln(6N/\varepsilon) \, ,$$

which, by definition of v, is equivalent to

$$m \geq \frac{2(1+\alpha)^2}{(1-\delta)^2} K^2 s \ln(6N/\varepsilon) = 18 K^2 s \ln(6N/\varepsilon) \, . \tag{12.44}$$

Suppose that this condition holds. Using the assumption $s \geq 36 \ln(6N/\varepsilon)$ we bound the term in (12.42) by

$$\begin{aligned} & N \exp\left(- \frac{\sqrt{s}/4}{2/\sqrt{s} + \frac{13}{3} K \sqrt{s/m}} \right) \leq N \exp\left(- \frac{\sqrt{s}/4}{\frac{2}{6\sqrt{\ln(6N/\varepsilon)}} + \frac{13}{3\sqrt{18 \ln(6N/\varepsilon)}}} \right) \\ & = N \exp\left(- \frac{\sqrt{s \ln(6N/\varepsilon)}}{4(1/3 + 13/(3\sqrt{18}))} \right) \leq N \exp\left(- \frac{6 \ln(6N/\varepsilon)}{4(1/3 + 13/(3\sqrt{18}))} \right) \\ & \leq \varepsilon/6 \, . \end{aligned}$$

because $6/(4(1/3 + 13/(3\sqrt{18}))) \approx 1.1072 > 1$.

Since (12.43) is implied by (12.44), we have shown that the probability that basis pursuit fails to recover $\mathbf{x}$ is at most ε provided that condition (12.43) together with $s \geq 36 \ln(6N/\varepsilon)$ holds. □

By slightly tuning the constants α, δ in the above proof, one may still improve a little on the constants in Theorem 12.18.

12.4 Nonuniform Recovery – Deterministic Sign Patterns

As already announced, we remove in this section the assumption that the sign pattern of the nonzero coefficients are required to be random. This means the coefficient vector is completely arbitrary (but fixed). Only the sampling matrix is randomly chosen. The main result of this section reads as follows.

Theorem 12.20. *Let $\mathbf{x} \in \mathbb{C}^N$ be s-sparse. Choose $\mathbf{A} \in \mathbb{C}^{m \times N}$ to be the random sampling matrix* (12.4) *associated to a BOS with constant $K \geq 1$. Assume that*

$$m \geq C K^2 s \ln(N) \ln(\varepsilon^{-1}) \, , \tag{12.45}$$

where $C > 0$ is a universal constant. Then with probability at least $1 - \varepsilon$ basis pursuit recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Remark 12.21. The proof reveals the more precise condition

$$m \geq 2cK^2 s \ln(4N) \ln(8\epsilon^{-1}) + \ln(4) cK^2 s \left(2\ln(s) + 8\ln(4\epsilon^{-1}) + 16\right)$$

with $c \approx 70.43$.

The previous result can be made stable under noise and sparsity defect. (Note that the error bound holding under the restricted isometry property shown in the next section is stronger, but requires slightly more samples.)

Theorem 12.22. *Let $\mathbf{x} \in \mathbb{C}^N$ and choose $\mathbf{A} \in \mathbb{C}^{m \times N}$ to be the random sampling matrix (12.4) associated to a BOS with constant $K \geq 1$. Let $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta\sqrt{m}$ for some $\eta \geq 0$ and let $\mathbf{x}^\sharp$ be a solution to*

$$\min_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta\sqrt{m} \,. \tag{12.46}$$

If

$$m \geq CK^2 s \ln(N) \ln(\varepsilon^{-1}) \tag{12.47}$$

then with probability at least $1 - \varepsilon - N^{-c}$ the reconstruction error satisfies

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \leq C_1 \sigma_s(\mathbf{x})_1 + C_2 \sqrt{s}\eta \,.$$

The constants $C, C_1, C_2, c > 0$ are universal.

Remark 12.23. The assumption $\|\mathbf{e}\|_2 \leq \eta\sqrt{m}$ on the noise is natural. If $f(\mathbf{t}) = \sum_{\ell \in [N]} x_\ell \phi_\ell(\mathbf{t})$ is the function associated with $\mathbf{x}$ then it is satisfied under the pointwise error estimate $|f(\mathbf{t}_\ell) - y_\ell| \leq \eta$ for $\ell \in [m]$.

In contrast to the approach of the previous section, the proof of these results relies on the recovery condition via an inexact dual in Theorem 4.31 and its extension to stable recovery in Theorem 4.32. As before, we introduce the rescaled matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$, where $\mathbf{A}$ is the sampling matrix in (12.4). The term $\|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S)^{-1}\|_{2\to 2}$ in (4.25) will be treated with Theorem 12.12 by noticing that $\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} \leq \delta$ implies $\|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S)^{-1}\|_{2\to 2} \leq (1-\delta)^{-1}$ (Lemma A.13). The other terms in Theorem 4.31 will be estimated based on the following lemmas together with some estimates from the previous section. All the following results refer to the rescaling sampling matrix $\widetilde{\mathbf{A}}$ as just introduced.

Lemma 12.24. *Let $\mathbf{v} \in \mathbb{C}^N$ with $\operatorname{supp} \mathbf{v} = S$, $\operatorname{card}(S) = s$. Then, for $t > 0$,*

$$\mathbb{P}(\|\widetilde{\mathbf{A}}_{\overline{S}}^* \widetilde{\mathbf{A}}\mathbf{v}\|_\infty \geq t\|\mathbf{v}\|_2) \leq 4N \exp\left(-\frac{m}{4K^2} \frac{t^2}{1 + \sqrt{s/18}\, t}\right) \,. \tag{12.48}$$

Proof. Note that

$$\|\widetilde{\mathbf{A}}_{\overline{S}}^{*}\widetilde{\mathbf{A}}\mathbf{v}\|_\infty = \max_{k\in\overline{S}} |\langle \mathbf{e}_k, \widetilde{\mathbf{A}}^{*}\widetilde{\mathbf{A}}\mathbf{v}\rangle| \,,$$

where $\mathbf{e}_k$ denotes the kth canonical vector. Without loss of generality, we may assume that $\|\mathbf{v}\|_2 = 1$. Denote

$$\mathbf{X}_\ell = (\overline{\phi_j(\mathbf{t}_\ell)})_{j\in[N]} \in \mathbb{C}^N \,. \tag{12.49}$$

Let $k \in \overline{S}$ and write

$$\langle \mathbf{e}_k, \widetilde{\mathbf{A}}^{*}\widetilde{\mathbf{A}}\mathbf{v}\rangle = \frac{1}{m}\sum_{\ell=1}^{m} \langle \mathbf{e}_k, \mathbf{X}_\ell \mathbf{X}_\ell^{*}\mathbf{v}\rangle = \frac{1}{m}\sum_{\ell=1}^{m} \mathbf{Y}_\ell$$

with $\mathbf{Y}_\ell = \langle \mathbf{e}_k, \mathbf{X}_\ell \mathbf{X}_\ell^{*}\mathbf{v}\rangle$. We aim to apply Bernstein's inequality in Corollary (7.31). To this end we note that the $\mathbf{Y}_\ell$ are independent and satisfy $\mathbb{E}\mathbf{Y}_\ell = \langle \mathbf{e}_k, \mathbb{E}[\mathbf{X}_\ell \mathbf{X}_\ell^{*}]\mathbf{v}\rangle = \langle \mathbf{e}_k, \mathbf{v}\rangle = 0$ since $k \notin S = \operatorname{supp}\mathbf{v}$. Next it follows from the Cauchy-Schwarz inequality that

$$\begin{aligned} |\mathbf{Y}_\ell| &= |\langle \mathbf{e}_k, \mathbf{X}_\ell \mathbf{X}_\ell^{*}\mathbf{v}\rangle| = |\langle \mathbf{e}_k, \mathbf{X}_\ell\rangle\langle \mathbf{X}_\ell, \mathbf{v}\rangle| = |\langle \mathbf{e}_k, \mathbf{X}_\ell\rangle||\langle (\mathbf{X}_\ell)_S, \mathbf{v}_S\rangle| \\ &\le |\phi_k(\mathbf{t}_\ell)|\|(\mathbf{X}_\ell)_S\|_2\|\mathbf{v}\|_2 \le K^2\sqrt{s} \,. \end{aligned}$$

Hereby, we used that $|\phi_k(\mathbf{t}_\ell)| \le K$ by the boundedness condition (12.2), and that $\|(\mathbf{X}_\ell)_S\|_2 \le K\sqrt{s}$, compare (12.32). The variance of $\mathbf{Y}_\ell$ can be estimated as

$$\begin{aligned} \mathbb{E}|\mathbf{Y}_\ell|^2 &= \mathbb{E}\left[\langle \mathbf{e}_k, \mathbf{X}_\ell \mathbf{X}_\ell^{*}\mathbf{v}\rangle\langle \mathbf{X}_\ell \mathbf{X}_\ell^{*}\mathbf{v}, \mathbf{e}_k\rangle\right] = \mathbb{E}\left[|\langle \mathbf{e}_k, \mathbf{X}_\ell\rangle|^2 \mathbf{v}^{*}\mathbf{X}_\ell \mathbf{X}_\ell^{*}\mathbf{v}\right] \\ &\le K^2 \mathbf{v}^{*}\mathbb{E}[\mathbf{X}_\ell \mathbf{X}_\ell^{*}]\mathbf{v} = K^2\|\mathbf{v}\|_2^2 = K^2 \end{aligned}$$

by the orthonormality relation (12.1), i.e., $\mathbb{E}[\mathbf{X}_\ell^{*}\mathbf{X}_\ell] = \mathbf{Id}$. Clearly, $\mathrm{Re}(\mathbf{Y}_\ell)$ and $\mathrm{Im}(\mathbf{Y}_\ell)$ satisfy the same bounds as $\mathbf{Y}_\ell$ itself. The union bound, the fact that $|z|^2 = \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2$ for any complex number z, and Bernstein's inequality (7.41) yield, for $t > 0$,

$$\begin{aligned} &\mathbb{P}(|\langle \mathbf{e}_k, \widetilde{\mathbf{A}}^{*}\widetilde{\mathbf{A}}\mathbf{v}\rangle| \ge t) \\ &\le \mathbb{P}\left(\left|\frac{1}{m}\sum_{\ell=1}^{m}\mathrm{Re}(\mathbf{Y}_\ell)\right| \ge t/\sqrt{2}\right) + \mathbb{P}\left(\left|\frac{1}{m}\sum_{\ell=1}^{m}\mathrm{Im}(\mathbf{Y}_\ell)\right| \ge t/\sqrt{2}\right) \\ &\le 4\exp\left(-\frac{(mt)^2/4}{mK^2 + K^2\sqrt{s}tm/(3\sqrt{2})}\right) = 4\exp\left(-\frac{m}{4K^2}\frac{t^2}{1+\sqrt{s/18}\,t}\right) \,. \end{aligned}$$

Taking the union bound over all $k \in \overline{S}$ completes the proof. □

Note that in the real-valued case (that is, the functions ϕ_j as well as the vector $\mathbf{v}$ are real-valued) the constant 4 in the probability estimate (12.48) above can be replaced by 2 in both instances.

Lemma 12.25. *Let $S \subset [N]$ with $\mathrm{card}(S) \le s$ and $\mathbf{v} \in \mathbb{C}^S$ with $\|\mathbf{v}\|_2 = 1$. Then, for $t > 0$,*

$$\mathbb{P}\left(\|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id})\mathbf{v}\|_2 \ge \sqrt{\frac{K^2 s}{m}} + t \right) \le \exp\left(-\frac{mt^2}{2K^2 s} \frac{1}{1 + 2\sqrt{\frac{K^2 s}{m}} + t/3} \right) .$$

Proof. Again we may assume without loss of generality that $\|\mathbf{v}\|_2 = 1$. Similarly to the previous proof we introduce vectors $\mathbf{X}_\ell = (\overline{\phi_j(\mathbf{t}_\ell)})_{j \in S} \in \mathbb{C}^S$. Note that

$$(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id})\mathbf{v} = \frac{1}{m} \sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})\mathbf{v} = \frac{1}{m} \sum_{\ell=1}^m \mathbf{Y}_\ell$$

with vectors $\mathbf{Y}_\ell = (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})\mathbf{v} \in \mathbb{C}^S$. Our aim is to apply the vector-valued Bernstein inequality of Corollary 8.42. Observe to this end that the $\mathbf{Y}_\ell$ are independent copies of a single random vector $\mathbf{Y}$ because the $\mathbf{X}_\ell$ are independent copies of a random vector $\mathbf{X}$, and they satisfy $\mathbb{E}\mathbf{Y}_\ell = \mathbb{E}\mathbf{Y} = \mathbb{E}(\mathbf{X}\mathbf{X}^* - \mathbf{Id})\mathbf{v} = 0$. Furthermore,

$$\mathbb{E}\|\mathbf{Y}_\ell\|_2^2 = \mathbb{E}\|(\mathbf{X}\mathbf{X}^* - \mathbf{Id})\mathbf{v}\|_2^2 = \mathbb{E}\left[|\langle \mathbf{X}, \mathbf{v} \rangle|^2 \|\mathbf{X}\|_2^2 \right] - 2\mathbb{E}|\langle \mathbf{X}, \mathbf{v} \rangle|^2 + 1 .$$

Observe that

$$|\langle \mathbf{X}, \mathbf{v} \rangle| = |\sum_{j \in S} \overline{\mathbf{v}_j} \phi_j(\mathbf{t})| \le \|\mathbf{v}\|_2 \sqrt{s} K = \sqrt{s} K .$$

by the Cauchy Schwarz inequality and the boundedness condition (12.2), which also implies $\|\mathbf{X}\|_2^2 = \sum_{j \in S} |\phi_j(\mathbf{t})|^2 \le sK^2$. Furthermore,

$$\mathbb{E}|\langle \mathbf{X}, \mathbf{v} \rangle|^2 = \sum_{j,k \in S} \mathbf{v}_j \overline{\mathbf{v}_k} \mathbb{E}[\phi_k(\mathbf{t}) \overline{\phi_j}(\mathbf{t})] = \|\mathbf{v}\|_2^2 = 1$$

by orthogonality (12.1). Hence,

$$\begin{aligned} \mathbb{E}\|\mathbf{Y}\|_2^2 &\le \mathbb{E}\left[|\langle \mathbf{X}, \mathbf{v} \rangle|^2 \|\mathbf{X}\|_2^2 \right] - 2\mathbb{E}|\langle \mathbf{X}, \mathbf{v} \rangle|^2 + 1 \le (sK^2 - 2)\mathbb{E}|\langle \mathbf{X}, \mathbf{v} \rangle|^2 + 1 \\ &= sK^2 - 1 \le sK^2 . \end{aligned}$$

For the uniform bound, observe that

$$\begin{aligned} \|\mathbf{Y}\|_2^2 &= \|(\mathbf{X}\mathbf{X}^* - \mathbf{Id})\mathbf{v}\|_2^2 = |\langle \mathbf{X}, \mathbf{v} \rangle|^2 \|\mathbf{X}\|_2^2 - 2|\langle \mathbf{X}, \mathbf{v} \rangle|^2 + 1 \\ &= |\langle \mathbf{X}, \mathbf{v} \rangle|^2 (\|\mathbf{X}\|_2^2 - 2) + 1 \le sK^2(sK^2 - 2) + 1 \le s^2 K^4 , \end{aligned}$$

so that $\|\mathbf{Y}\|_2 \le sK^2$ for all realizations of $\mathbf{Y}$. Further, we simply bound the weak variance by the strong variance,

$$\sigma^2 = \sup_{\|\mathbf{z}\|_2 \le 1} \mathbb{E}|\langle \mathbf{z}, \mathbf{Y} \rangle|^2 \le \mathbb{E}\|\mathbf{Y}\|_2^2 \le sK^2 .$$

Then the ℓ_2-valued Bernstein inequality (8.87) yields

$$\mathbb{P}(\|\sum_{\ell=1}^{m} \mathbf{Y}_\ell\|_2 \geq \sqrt{msK^2} + t) \leq \exp\left(-\frac{t^2/2}{msK^2 + 2sK^2\sqrt{msK^2} + tsK^2/3}\right) ,$$

so that with t replaced by mt we obtain

$$\mathbb{P}\left(\|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id})\mathbf{v}\|_2 \geq \sqrt{\frac{K^2 s}{m}} + t\right) \leq \exp\left(-\frac{mt^2}{2K^2 s} \frac{1}{1 + 2\sqrt{\frac{K^2 s}{m}} + t/3}\right) .$$

This completes the proof. □

Next we provide a variant of Lemma 12.19, which is more convenient here.

Lemma 12.26. *For $0 < t \leq 2\sqrt{s}$,*

$$\mathbb{P}\left(\max_{j \in \overline{S}} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 \geq t\right) \leq 2(s+1)N \exp\left(-\frac{3}{10}\frac{mt^2}{K^2 s}\right) . \tag{12.50}$$

Proof. Fix $j \in \overline{S}$. Similarly as before, we introduce the vectors $\mathbf{X}_\ell = (\phi_k(\mathbf{t}_\ell))_{k \in S} \in \mathbb{C}^S$ and $\mathbf{Y}_\ell = \left(\phi_j(\mathbf{t}_\ell)\overline{\phi_k(\mathbf{t}_\ell)}\right)_{k \in S} = \phi_j(\mathbf{t}_\ell)\mathbf{X}_\ell \in \mathbb{C}^S$. Then

$$\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j\|_2 = \frac{1}{m}\left\|\sum_{\ell=1}^{m} \mathbf{Y}_\ell\right\|_2 .$$

Our goal is to apply the noncommutative Bernstein inequality in Theorem 8.14 and its extension in Exercise 8.7 by treating the $\mathbf{Y}_\ell$ as matrices and noting that the operator norm of $\mathbf{Y}_\ell$ equals then its ℓ_2-norm. The $\mathbf{Y}_\ell$ are independent and satisfy $\mathbb{E}\mathbf{Y}_\ell = 0$ by orthonormality of the ϕ_j and because $j \notin S$. They can be bounded by

$$\|\mathbf{Y}_\ell\|_{2\to 2} = \|\mathbf{Y}_\ell\|_2 = \|\mathbf{X}_\ell\|_2 |\langle \mathbf{e}_j, \mathbf{X}_\ell\rangle| \leq \sqrt{s}K^2 ,$$

by the boundedness condition (12.2) and since $\mathrm{card}(S) \leq s$. Furthermore,

$$\begin{aligned}\mathbb{E}[\mathbf{Y}_\ell^* \mathbf{Y}_\ell] = \mathbb{E}\|\mathbf{Y}_\ell\|_2^2 = \mathbb{E}[|\langle \mathbf{X}_\ell, \mathbf{e}_j\rangle|^2 \|\mathbf{X}_\ell\|_2^2] &\leq K^2 \mathbb{E}\|\mathbf{X}_\ell\|_2^2 = K^2 \sum_{k \in S} \mathbb{E}|\phi_k(\mathbf{t})|^2 \\ &= K^2 s .\end{aligned}$$

Moreover

$$\mathbb{E}[\mathbf{Y}_\ell \mathbf{Y}_\ell^*] = \mathbb{E}[|\phi_j(\mathbf{t}_\ell)|^2 \mathbf{X}_\ell \mathbf{X}_\ell^*] \preccurlyeq K^2 \mathbb{E}[\mathbf{X}_\ell \mathbf{X}_\ell^*] = K^2 \mathbf{Id} .$$

Therefore, the variance parameter σ^2 in (8.115) satisfies

$$\sigma^2 = \max\left\{ \| \sum_{\ell=1}^{m} \mathbb{E}[\mathbf{Y}_\ell \mathbf{Y}_\ell^*] \|_{2\to 2}, \| \sum_{\ell=1}^{m} \mathbb{E}[\mathbf{Y}_\ell^* \mathbf{Y}_\ell] \|_{2\to 2} \right\} \leq K^2 m s \, .$$

The version of the noncommutative Bernstein inequality for rectangular random matrices (8.116) yields

$$\mathbb{P}\left(\| \sum_{\ell=1}^{m} \mathbf{Y}_\ell \|_2 \geq u \right) \leq 2(s+1) \exp\left(- \frac{u^2/2}{K^2 m s + u\sqrt{s} K^2/3} \right) \, .$$

Setting $u = mt$, taking the union bound over $j \in [N]$, and using that $0 < t \leq 2\sqrt{s}$ yields

$$\begin{aligned} &\mathbb{P}\left(\max_{j \in \overline{S}} \| \widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_j \|_2 \geq t \right) = \mathbb{P}\left(\max_{j \in \overline{S}} \| \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{Y}_\ell \|_2 \geq t \right) \\ &\leq 2N(s+1) \exp\left(- \frac{mt^2/2}{K^2 s + t\sqrt{s} K^2/3} \right) \leq 2N(s+1) \exp\left(- \frac{3}{10} \frac{mt^2}{K^2 s} \right) \, . \end{aligned}$$

This completes the proof. □

Before passing to the proof of Theorem 12.20 we provide a slightly weaker result, which we strengthen afterwards.

Proposition 12.27. *Let $\mathbf{x} \in \mathbb{C}^N$ be s-sparse. Choose $\mathbf{A} \in \mathbb{C}^{m \times N}$ to be the random sampling matrix* (12.4) *associated to a BOS with constant $K \geq 1$. Assume that*

$$m \geq c K^2 s \left[2 \ln(4N) \ln(12\varepsilon^{-1}) + \ln(s) \ln(12\varepsilon^{-1} \ln(s)) \right] \, ,$$

with $c = 8e^2(1 + (1/\sqrt{8} + 1/6)/e) \approx 70.43$. Then basis pursuit recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$ with probability at least $1 - \varepsilon$.

Remark 12.28. If $\ln(s) \ln(\ln s) \leq c \ln(N)$, then the above result already implies Theorem 12.20.

Proof. The proof relies on the so-called golfing scheme, and an application of the recovery result in Theorem 4.31 for ℓ_1-minimization based on an inexact dual vector. We partition the m independent samples into L disjoint blocks of sizes $m_1, \ldots, m_L$ to be specified later; in particular, $m = \sum_{j=1}^{L} m_j$. These blocks correspond to row submatrices of $\mathbf{A}$, which we denote by $\mathbf{A}^{(1)} \in \mathbb{C}^{m_1 \times N}, \ldots, \mathbf{A}^{(L)} \in \mathbb{C}^{m_L \times N}$. It will be crucial below that these submatrices are stochastically independent. As before, we also introduce the rescaled matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}} \mathbf{A}$.

Let $S = \operatorname{supp} \mathbf{x}$. We set $\mathbf{u}^{(0)} = 0 \in \mathbb{C}^N$ and define recursively

$$\mathbf{u}^{(n)} = \frac{1}{m_n} (\mathbf{A}^{(n)})^* \mathbf{A}_S^{(n)} (\operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n-1)}) + \mathbf{u}^{(n-1)} \, , \tag{12.51}$$

for $n = 1, \ldots, L$. The vector $\mathbf{u} = \mathbf{u}^{(L)}$ will then serve as a candidate for the inexact dual of Theorem 4.31. By construction of $\mathbf{u}$ there exists indeed a vector $\mathbf{h} \in \mathbb{C}^m$ such that $\mathbf{u} = \mathbf{A}^*\mathbf{h}$, and by rescaling also $\mathbf{u} = \widetilde{\mathbf{A}}^*\widetilde{\mathbf{h}}$, for some $\widetilde{\mathbf{h}}$. For the sake of simpler notation we introduce $\mathbf{w}^{(n)} = \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n)}$. Observe that

$$\begin{aligned} \mathbf{w}^{(n)} &= \left(\mathbf{Id} - \frac{1}{m_n}(\mathbf{A}_S^{(n)})^*\mathbf{A}_S^{(n)}\right)\mathbf{w}^{(n-1)} \\ &= \prod_{k=1}^{n}\left(\mathbf{Id} - \frac{1}{m_k}(\mathbf{A}_S^{(k)})^*\mathbf{A}_S^{(k)}\right)\operatorname{sgn}(\mathbf{x}_S)\ , \end{aligned} \tag{12.52}$$

and

$$\mathbf{u} = \sum_{n=1}^{L}\frac{1}{m_n}(\mathbf{A}^{(n)})^*\mathbf{A}_S^{(n)}\mathbf{w}^{(n-1)}\ . \tag{12.53}$$

We will now verify the conditions of Theorem 4.31. For this task we will use the lemmas proven above. First we require the following inequalities,

$$\|\mathbf{w}^{(n)}\|_2 \le \left(\sqrt{\frac{K^2 s}{m_n}} + r_n\right)\|\mathbf{w}^{(n-1)}\|_2\ , \quad n \in [L]\ , \tag{12.54}$$

$$\left\|\frac{1}{m_n}(\mathbf{A}_{\overline{S}}^{(n)})^*\mathbf{A}_S^{(n)}\mathbf{w}^{(n-1)}\right\|_\infty \le t_n\|\mathbf{w}^{(n-1)}\|_2\ , \quad n \in [L]\ , \tag{12.55}$$

where the parameters r_n, t_n will be specified below. The probability $p_1(n)$ that (12.54) does not hold can be be bounded using Lemma 12.25,

$$p_1(n) \le \exp\left(-\frac{m_n r_n^2}{2K^2 s}\frac{1}{1 + 2\sqrt{\frac{K^2 s}{m_n}} + r_n/3}\right)\ .$$

Due to Lemma 12.24 the probability $p_2(n)$ that (12.55) does not hold is bounded by

$$p_2(n) \le 4N\exp\left(-\frac{m_n}{4K^2}\frac{t_n^2}{1 + \sqrt{s/18}\,t_n}\right)\ . \tag{12.56}$$

Let $r_n' := \sqrt{K^2 s/m_n} + r_n$. Then the definition of $\mathbf{w}^{(n)}$ yields

$$\|\operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 = \|\mathbf{w}^{(L)}\|_2 \le \|\operatorname{sgn}(\mathbf{x}_S)\|_2\prod_{n=1}^{L} r_n' \le \sqrt{s}\prod_{n=1}^{L} r_n'\ .$$

Furthermore, (12.53) yields

$$\begin{aligned} \|\mathbf{u}_{\overline{S}}\|_\infty &\le \sum_{n=1}^{L}\|\frac{1}{m_n}(\mathbf{A}_{\overline{S}}^{(n)})^*\mathbf{A}_S^{(n)}\mathbf{w}^{(n-1)}\|_\infty \le \sum_{n=1}^{L} t_n\|\mathbf{w}^{(n-1)}\|_2 \\ &\le \sqrt{s}\sum_{n=1}^{L} t_n\prod_{j=1}^{n-1} r_j' \end{aligned}$$

with the understanding that $\prod_{j=1}^{n-1} r'_j = 1$ if $n = 1$. Next we need to set the parameters $L, m_1, \ldots, m_L, r_1, \ldots, r_L, t_1, \ldots, t_L$ such that $\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \le \gamma$ and $\|\mathbf{u}_{\overline{S}}\|_\infty \le \theta$ for some appropriate values of $\theta, \gamma < 1$, see also Theorem 4.31. We choose

$$
\begin{aligned}
&L = \lceil \ln(s)/2 \rceil + 2 \,, \\
&m_1, m_2 \ge cK^2 s \ln(4N) \ln(2\varepsilon^{-1}) \,, \quad \text{and} \quad m_n \ge cK^2 s \ln(2L\varepsilon^{-1}) \,, \; n = 3, \ldots, L \,, \\
&r_1 = r_2 = \frac{1}{2e\sqrt{\ln(4N)}} \,, \quad \text{and} \quad r_n = (2e)^{-1} \,, \quad n = 3, \ldots, L \,, \\
&t_1 = t_2 = \frac{1}{e\sqrt{s}} \,, \quad \text{and} \quad t_n = \frac{\ln(4N)}{e\sqrt{s}} \,, \quad n = 3, \ldots, L \,,
\end{aligned}
$$

where $c = 8e^2(1 + e^{-1}(1/\sqrt{8} + 1/6)) \approx 70.43$. Then $r'_1, r'_2 \le 1/(e\sqrt{\ln(4N)})$ and $r'_n \le e^{-1}$, $n = 3, \ldots, L$. Furthermore,

$$
\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \le \sqrt{s} \prod_{n=1}^{L} r'_n \le \sqrt{s} e^{-\ln(s)/2 - 2} = e^{-2} \,,
$$

and

$$
\|\mathbf{u}_{\overline{S}}\|_\infty \le e^{-1} \left(1 + \frac{1}{e\sqrt{\ln(4N)}} + \sum_{n=2}^{L-1} e^{-n} \right) \le \frac{e^{-1}}{1 - e^{-1}} = \frac{1}{e-1} \,.
$$

The probabilities $p_1(n)$ can be estimated as

$$
\begin{aligned}
p_1(1), p_1(2) &\le \exp\left(-\frac{m_1 r_1^2}{2K^2 s} \frac{1}{1 + \sqrt{K^2 s / m_1} + r_1/3} \right) \\
&\le \exp\left(-\frac{c \ln(4N) \ln(2\varepsilon^{-1})}{8e^2 \ln(4N)} \frac{1}{1 + (c \ln(4N) \ln(2\varepsilon^{-1}))^{-1/2} + 1/(6e\sqrt{\ln(4N)})} \right) \\
&\le \varepsilon/2 \,, \qquad (12.57)
\end{aligned}
$$

by definition of c and similarly

$$
p_1(n) \le \varepsilon/(2L) \,, \quad n = 3, \ldots, L \,.
$$

This yields $\sum_{n=1}^{L} p_1(n) \le 2\varepsilon$. By (12.56), the definitions of the parameters and of the constant c, we obtain

$$
\begin{aligned}
p_2(1), p_2(2) &\le 4N \exp\left(-\frac{cK^2 s \ln(4N) \ln(2\varepsilon^{-1})}{4K \left(e^2 s (1 + \sqrt{s/18}/(e\sqrt{s})) \right)} \right) \\
&= 4N \exp\left(-\frac{c}{4e^2(1 + 1/(e\sqrt{18}))} \ln(4N) \ln(2\varepsilon) \right) \\
&\le 4N \exp\left(-\ln(4N) - \ln(2\varepsilon^{-1}) \right) = \varepsilon/2 \,,
\end{aligned}
$$

where we have used that $2ab \geq a + b$ for $a, b \geq 1$. A similar estimate gives $p_2(n) \leq \varepsilon/(2L)$ for $n \geq 3$, so that again $\sum_{n=1}^{L} p_2(n) \leq 2\varepsilon$.

The overall number of samples obeys

$$\begin{aligned} m &= \sum_{n=1}^{L} m_n = m_1 + m_2 + \sum_{n=3}^{L} m_n \\ &\geq 2cK^2 s \ln(4N) \ln(2\varepsilon^{-1}) + cK^2 \lceil \ln(s)/2 \rceil s \ln(2\lceil \ln(s)/2 \rceil \varepsilon^{-1}) \, . \end{aligned}$$

Hence, the proposed choices of the m_n are possible if

$$m \geq cK^2 s \left[2 \ln(4N) \ln(2\varepsilon^{-1}) + \ln(s) \ln(2\varepsilon^{-1} \ln(s)) \right] \, . \tag{12.58}$$

By Theorem 12.12 we have $\|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\| \leq 1/2$ with probability at least $1 - 2s \exp\left(-\frac{3m}{32K^2 s}\right)$. Hence, the first part of condition (4.25) of Theorem 4.31 holds with $\alpha = 2$, that is, $\|(\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S)^{-1}\|_{2\to 2} \leq 2$ provided

$$m \geq \frac{32}{3} K^2 s \ln(2s\varepsilon^{-1}). \tag{12.59}$$

In the notation of Theorem 4.31 we have so far chosen parameters $\alpha = 2$, $\gamma = e^{-2}$ and $\theta = (e-1)^{-1}$. The condition $\theta + \alpha\beta\gamma < 1$ together with the second part of (4.25) translates into

$$\max_{\ell \in \overline{S}} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_\ell\|_2 \leq \beta$$

with $\beta < \gamma^{-1}\alpha^{-1}(1-\theta) = e^2(e-2)/(2(e-1)) \approx 1.544$. Let us choose $\beta = 3/2$, say. Lemma 12.26 together with $(s+1) \leq N$ implies that

$$\mathbb{P}\left(\max_{\ell \in \overline{S}} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{a}}_\ell\|_2 \geq \beta \right) \leq 2N^2 \exp\left(-\frac{3}{8} \frac{m\beta^2}{K^2 s} \right) \, .$$

This term is bounded by ε provided $m \geq (10/3)\beta^{-2} K^2 s \ln(2N^2/\varepsilon)$ which is implied by

$$m \geq CK^2 s \ln(2N/\varepsilon) \tag{12.60}$$

with $C = 20\beta^{-2}/3 \approx 2.96$.

Altogether we have shown that the conditions (4.25) and (4.26) of Theorem 4.31 hold simultaneously with probability at least $1 - 6\varepsilon$ provided conditions (12.58), (12.59), (12.60) hold. Replacing ε by $\varepsilon/6$, and noting that (12.58) is stronger than (12.59) and (12.60) concludes the proof of Proposition 12.27. □

Remark 12.29. The name golfing scheme of the method of this proof comes from the strategy that with each iteration n, the vector $\mathbf{u}^{(n)}$ gets closer to the desired inexact dual vector, like in golf where the ball (ideally) comes closer to the hole with each hit.

Now we modify the previous proof by a nice trick to obtain the main result of this section.

Proof (of Theorem 12.20). We use the basic structure of the previous proof. The strengthening of the result is based on the idea that we can sample slightly more row blocks $\mathbf{A}^{(n)}$ of the matrix $\mathbf{A}$ than in the previous proof. Then we use only a part of them such that (12.54) and (12.55) are satisfied. The probability that these inequalities hold only for a fraction of the samples is much higher than the probability that they hold simultaneously for all sampled blocks. The fact that we have to choose slightly more blocks will not deteriorate the overall number m of samples - in contrast, it actually decreases m because the size m_n of each block can be chosen smaller.

To be more precise, we choose a number $L' > L$ of row submatrices to be determined below. As in the previous proof we set $\mathbf{u}^{(0)} = 0$ and define recursively $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$ (for $n = 1, 2$ we do not allow replacements) via (12.51). Next we continue with the recursive definition of $\mathbf{u}^{(n)}$, but always check whether the associated $\mathbf{w}^{(n)} = \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n)}$ satisfies (12.54) and (12.55). If these conditions are not satisfied we "discard" this particular n in the sense that we replace $\mathbf{A}^{(n)}$ by $\mathbf{A}^{(n+1)}$ (and also all subsequent $\mathbf{A}^{(\ell)}$ by $\mathbf{A}^{(\ell+1)}$, $\ell > n$). Then we redefine $\mathbf{u}^{(n)}$ and $\mathbf{w}^{(n)}$ using the modified $\mathbf{A}^{(n)}$. We continue in this way by always discarding an n when (12.54) and (12.55) are not satisfied, until we arrive at $n = L$ (below we estimate the probability that this actually happens). Since the $\mathbf{A}^{(n)}$ are independent, the events that (12.54) and (12.55) hold for a given $n \in [L']$ are independent.

With respect to the previous proof, we use a slightly different definition of m_n, $n \geq 3$,

$$m_n \geq cK^2 s \ln(2\rho^{-1}),$$

for some $\rho \in (0, 1)$ to be defined below. The remaining quantities L, m_1, m_2, r_n, t_n are defined in the same way as before. Again the probabilities $p_1(1), p_1(2), p_2(1), p_2(2) \leq \varepsilon/2$. We need to determine the probability that (12.54) and (12.55) hold for at least $L - 2$ choices of $n \in \{3, 4, \ldots, L'\}$. By (12.57) and the modified definition of m_n we have $p_1(n) \leq \rho/2$ and $p_2(n) \leq \rho/2$, $n \geq 3$, so that the event B_n that both (12.54) and (12.55) hold for a given $n \geq 3$ occurs with probability at least $1 - \rho$. The event that B_n occurs for at least $L - 2$ choices of n has probability larger than the event that

$$\sum_{n=3}^{L'} X_n \geq L - 2,$$

where the X_n are independent random variables that take the value 1 with probability $1 - \rho$ and the value 0 with probability ρ. Clearly, $\mathbb{E}X_n = 1 - \rho$ and $X_n - \mathbb{E}X_n \leq 1$ for all n. Set $J := L' - 2$. Hoeffding's inequality, Theorem 7.20, shows that

$$\mathbb{P}\left(\sum_{n=3}^{L'} X_n < (1-\rho)J - \sqrt{J}t\right) = \mathbb{P}\left(\sum_{n=3}^{L'}(X_n - \mathbb{E}X_n) < -\sqrt{J}t\right) \le e^{-t^2/2}.$$

Setting $L = (1-\rho)J - \sqrt{J}t$ and solving for t yields

$$\mathbb{P}\left(\sum_{n=3}^{L'} X_n < L\right) \le \exp\left(-\frac{((1-\rho)J - L)^2}{2J}\right).$$

The choice

$$J = \left\lceil \frac{2}{1-\rho}L + \frac{2}{(1-\rho)^2}\ln(\widetilde{\varepsilon}^{-1}) \right\rceil \tag{12.61}$$

implies that the event B_n occurs at least L times with probability at least $1-\widetilde{\varepsilon}$. The overall number of samples satisfies

$$\begin{aligned} m &= m_1 + m_2 + \sum_{n=3}^{L'} m_n \ge 2cK^2 s \ln(4N)\ln(2\epsilon^{-1}) + JcK^2 s \ln(2\rho^{-1}) \\ &= 2cK^2 s \ln(4N)\ln(2\epsilon^{-1}) + \left\lceil \frac{2}{1-\rho}L + \frac{2}{(1-\rho)^2}\ln(\widetilde{\varepsilon}^{-1}) \right\rceil cK^2 s \ln(2\rho^{-1}) \\ &= 2cK^2 s \ln(4N)\ln(2\epsilon^{-1}) \\ &\quad + \left\lceil \frac{2}{1-\rho}\lceil \ln(s)/2 \rceil + 2) + \frac{2}{(1-\rho)^2}\ln(\widetilde{\varepsilon}^{-1}) \right\rceil cK^2 s \ln(2\rho^{-1}) \,. \end{aligned} \tag{12.62}$$

Choosing $\rho = 1/2$, this condition is implied by

$$m \ge 2cK^2 s \ln(4N)\ln(2\varepsilon^{-1}) + \ln(4)cK^2 s(2\ln(s) + 8\ln(\widetilde{\varepsilon}^{-1}) + 16). \tag{12.63}$$

Note that with $\widetilde{\varepsilon} = \varepsilon$ this condition is stronger than (12.59) and (12.60). A ltogether we showed that ℓ_1-minimization recovers $\mathbf{x}$ with probability at least $1 - 4\varepsilon$. Replacing ε by $\varepsilon/4$ and realizing that (12.63) is implied by

$$m \ge CK^2 s \ln(N)\ln(\varepsilon^{-1})$$

with an appropriate constant C concludes the proof. □

Let us finally consider stable recovery.

Proof (of Theorem 12.22). The proof is based on the inexact dual condition of Theorem 4.32. We use the golfing scheme of the previous proof, and in particular, we make the same choices of the parameters L, L', J, r_n, r_n', t_n as before. We only slightly change the conditions on the m_n as follows,

$$\begin{aligned} m_1, m_2 &\ge cK^2 s \ln(4N)\ln(2\varepsilon^{-1}) \,, \\ m_n &\ge cK^2 s \ln(2\rho^{-1})\ln(2\varepsilon^{-1}), \quad n = 3, \dots, L' \,, \end{aligned}$$

where $\rho = 1/2$. We impose the additional constraint that

$$\frac{m}{m_n} \leq C'(r'_j)^2 \ln(4N) \,, \tag{12.64}$$

for an appropriate constant $C' > 0$. This is possible by the condition on m and by definition of the r_j. Moreover, we now choose $\widetilde{\varepsilon} = N^{-c}$ in the definition of J in (12.61).

Let $S \subset [N]$ with $\mathrm{card}(S)$ be an index set of s largest coefficients of $\mathbf{x}$. Conditions (4.29), (4.30), (4.31), (4.32) of Theorem 4.32 with $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ in place of $\mathbf{A}$ hold with probability at least $1 - \varepsilon - N^{-c}$ with appropriate values of the constants $\delta, \beta, \gamma, \theta$. This follows from the arguments above. (The additional factor of $\ln(2\varepsilon^{-1})$ in the definition of m_n, $n \geq 3$, does not change the analysis. The reader is invited to check all details in Exercise 12.6.) The resulting number of samples in (12.63) satisfies, with the modified value of $\widetilde{\varepsilon} = N^{-c}$,

$$m = \sum_{n=1}^{L'} m_n \geq 2cK^2 s \ln(4N) \ln(2\varepsilon^{-1}) + \ln(4) cK^2 s (2\ln(s) + 8c\ln(N) + 16)$$

so that the choices of m_n are possible as proposed when the constant C in (12.47) is set appropriately.

It remains to verify (4.33) for the vector $\mathbf{h} \in \mathbb{C}^m$ constructed in the previous proof such that $\mathbf{u} = \widetilde{\mathbf{A}}^* \mathbf{h}$. For notational simplicity, we assume that in the setting of the previous proof the first L values of n are taken for the construction of the inexact dual, that is, $\mathbf{u}$ is given by (12.51) and using the rescaled matrices $\widetilde{\mathbf{A}}^{(n)} = \frac{1}{\sqrt{m}}\mathbf{A}^{(n)}$ gives

$$\mathbf{u} = \sum_{n=1}^{L} \frac{1}{m_n} (\mathbf{A}^{(n)})^* \mathbf{A}_S^n \mathbf{w}^{(n-1)} = \sum_{n=1}^{L} \frac{m}{m_n} (\widetilde{\mathbf{A}}^{(n)})^* \widetilde{\mathbf{A}}_S^n \mathbf{w}^{(n-1)} \,.$$

Hence, $\mathbf{u} = \widetilde{\mathbf{A}}^* \mathbf{h}$ with $\mathbf{h}^* = ((\mathbf{h}^{(1)})^*, \ldots, (\mathbf{h}^{(L)})^*, 0, \ldots, 0)$ and

$$\mathbf{h}^{(n)} = \frac{m}{m_n} \widetilde{\mathbf{A}}^{(n)} \mathbf{w}^{(n-1)} \in \mathbb{C}^{m_n} \,, \quad n = 1, \ldots, L \,.$$

Then

$$\begin{aligned} \|\mathbf{h}\|_2^2 &= \sum_{n=1}^{L} \|\mathbf{h}^{(n)}\|_2^2 = \sum_{n=1}^{L} \frac{m}{m_n} \left\| \sqrt{\frac{m}{m_n}} \widetilde{\mathbf{A}}_S^{(n)} \mathbf{w}^{(n-1)} \right\|_2^2 \\ &= \sum_{n=1}^{L} \frac{m}{m_n} \left\| \sqrt{\frac{1}{m_n}} \mathbf{A}_S^{(n)} \mathbf{w}^{(n-1)} \right\|_2^2 \,. \end{aligned}$$

We also recall the relation (12.52) of the vectors $\mathbf{w}^n$. This gives, for $n \geq 1$,

$$\left\| \sqrt{\frac{1}{m_n}} \mathbf{A}_S^{(n)} \mathbf{w}^{(n-1)} \right\|_2^2 = \left\langle \frac{1}{m_n} (\mathbf{A}_S^{(n)})^* \mathbf{A}_S^{(n)} \mathbf{w}^{(n-1)}, \mathbf{w}^{n-1} \right\rangle$$
$$= \left\langle \frac{1}{m_n} (\mathbf{A}_S^{(n)})^* \mathbf{A}_S^{(n)} - \mathbf{Id}) \mathbf{w}^{(n-1)}, \mathbf{w}^{n-1} \right\rangle + \|\mathbf{w}^{(n-1)}\|_2^2$$
$$= \left\langle \mathbf{w}^{(n)}, \mathbf{w}^{(n-1)} \right\rangle + \|\mathbf{w}^{(n-1)}\|_2^2 \leq \|\mathbf{w}^{(n)}\|_2 \|\mathbf{w}^{(n-1)}\|_2 + \|\mathbf{w}^{(n-1)}\|_2^2 \,.$$

Recall from (12.54) that $\|\mathbf{w}^{(n)}\|_2 \leq r'_n \|\mathbf{w}^{(n-1)}\|_2 \leq \|\mathbf{w}^{(n-1)}\|_2$ (except for an event of probability at most ϵ). This gives

$$\left\| \sqrt{\frac{1}{m_n}} \mathbf{A}_S^{(n)} \mathbf{w}^{(n-1)} \right\|_2^2 \leq 2\|\mathbf{w}^{(n-1)}\|_2^2 \leq 2\|\mathbf{w}^{(0)}\|_2^2 \prod_{j=1}^{n} (r'_j)^2$$
$$= 2\|\mathrm{sgn}(\mathbf{x})_S\|_2^2 \prod_{j=1}^{n} (r'_j)^2 = 2s \prod_{j=1}^{n-1} (r'_j)^2 \,.$$

The definition of the constants r_n and the additional constraint (12.64) therefore yield

$$\|\mathbf{h}\|_2^2 \leq 2s \sum_{n=1}^{L} \frac{m}{m_n} \prod_{j=1}^{n-1} (r'_j)^2 \leq C' s \ln(4N) \sum_{n=1}^{L} (r'_n)^2 \prod_{j=1}^{n} (r'_j)^2$$
$$\leq C'(2e)^{-2} s \sum_{n=1}^{L} \prod_{j=2}^{n} (r'_j)^2 \leq C'' s \,,$$

where we used the convention that $\prod_{j=2}^{1} (r'_n)^2 = 1$ and that $\prod_{j=2}^{n} (r'_j)^2 \leq (2e)^{-2(n-1)}$ for $n \geq 2$. Therefore, all conditions of Theorem 4.32 are satisfied for $\mathbf{x}$ and $\widetilde{\mathbf{A}}$ with probability at least $1 - \varepsilon - N^{-c}$. Noting that the optimization problem

$$\min_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \left\| \widetilde{A}\mathbf{z} - \frac{1}{\sqrt{m}} \mathbf{y} \right\|_2 \leq \eta$$

is equivalent to (12.46) completes the proof. □

12.5 Restricted Isometry Property

In this section we derive an estimate for the restricted isometry constants of the random matrix $\mathbf{A}$ in (12.4) associated to random sampling in a bounded orthonormal system. This will lead to a stable and uniform recovery result for ℓ_1-minimization.

The main result of this section reads as follows.

Theorem 12.30. *Let* $\mathbf{A} \in \mathbb{C}^{m \times N}$ *be the random matrix* (12.4) *associated to random sampling in a bounded orthogonal system obeying* (12.2) *with some constant* $K \geq 1$. *Let* $\delta \in (0,1)$. *If*

$$m \geq CK^2 \delta^{-2} s \ln^3(s) \ln(N) \tag{12.65}$$

then with probability at least $1 - N^{-\gamma \ln^3(s)}$ *the restricted isometry constant* δ_s *of* $\frac{1}{\sqrt{m}}\mathbf{A}$ *satisfies* $\delta_s \leq \delta$. *The constants* $C, \gamma > 0$ *are universal.*

Remark 12.31. Since $s \leq N$, the condition (12.65) is implied by the simpler condition

$$m \geq CK^2 \delta^{-2} s \ln^4(N) \, .$$

The probability of success may then be strengthened to $1 - N^{-\gamma \ln^3(N)}$.

The above theorem follows from the more precise result stated next.

Theorem 12.32. *Let* $\mathbf{A} \in \mathbb{C}^{m \times N}$ *be the random sampling matrix* (12.4) *associated to random sampling in a bounded orthogonal system obeying* (12.2) *with some constant* $K \geq 1$. *Let* $\varepsilon, \eta_1, \eta_2 \in (0,1)$. *If*

$$\frac{m}{\ln(9m)} \geq \overline{C} \eta_1^{-2} K^2 s \ln^2(9s) \ln(8N) \, , \tag{12.66}$$

$$m \geq \widetilde{C} \eta_2^{-2} K^2 s \ln(\varepsilon^{-1}) \, . \tag{12.67}$$

then with probability at least $1 - \varepsilon$ *the restricted isometry constant* δ_s *of* $\frac{1}{\sqrt{m}}\mathbf{A}$ *satisfies* $\delta_s \leq \eta_1 + \eta_1^2 + \eta_2$. *The constants may be chosen* $\widetilde{C} = 32/3 \approx 10.66$ *and* $\overline{C} = c_0 C^2 = 23\,328$ *where* $c_0 = 162$ *and* $C = 12$ *is the constant from Dudley's inequality in Theorem 8.23.*

Remark 12.33. The constants in the previous result are definitely not nice, and certainly not optimal. However, an improvement is probably cumbersome, and it is questionable whether this provides more insight.

Before proceeding we briefly show how Theorem 12.32 implies Theorem 12.30.

Proof (of Theorem 12.30). It follows from Lemma C.7 that (12.65) with an appropriate constant $C > 0$ implies (12.66) with $\eta_1 = \delta$. Furthermore, if $\varepsilon = N^{-\gamma \ln^3(s)}$ for an appropriate $\gamma > 0$, then (12.65) implies as well (12.67) with $\eta_2 = \delta$. Therefore, Theorem 12.32 implies $\delta_s \leq 3\delta$ with probability at least $1 - N^{-\gamma \ln^3(s)}$, which is the claim after rescaling constants. □

Using the results of Chapter 6 we obtain the following result concerning recovery of sparse polynomials with respect to the orthonormal system $\{\phi_j : j \in [N]\}$ from random samples.

Corollary 12.34. *Suppose that*

$$m \geq CK^2 s \ln^3(s) \ln(N) \, .$$

Then

(a) with probability at least $1 - N^{-\gamma \ln^3(s)}$ *every* s*-sparse vector* $\mathbf{x} \in \mathbb{C}^N$ *is recovered from the samples* $\mathbf{y} = \mathbf{A}\mathbf{x} = \left(\sum_{j=1}^N x_j \phi_j(\mathbf{t}_\ell) \right)_{\ell=1}^m$ *by basis pursuit.*

(b) More generally, with probability at least $1 - N^{-\gamma \ln^3(s)}$ *the following statement holds for every* $\mathbf{x} \in \mathbb{C}^N$*. Let noisy samples* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ *be given with*

$$\|\mathbf{e}\|_2 = \sqrt{\sum_{\ell=1}^m |e_\ell|^2} \leq \eta \sqrt{m}$$

and let $\mathbf{x}^\sharp$ *be the solution of the* ℓ_1*-minimization problem*

$$\text{minimize}_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta \sqrt{m} \, . \tag{12.68}$$

Then

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_2 \leq C_1 \frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + C_2 \eta \, .$$

All constants $C, C_1, C_2, \gamma > 0$ *are universal.*

Proof. Combine Theorem 12.32 with Theorem 6.11 for the normalized matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}} \mathbf{A}$. □

Remark 12.35. (a) The assumption $\|\mathbf{e}\|_2 \leq \eta \sqrt{m}$ on the noise is satisfied if each sample is taken with accuracy η, that is, $|y_\ell - (\mathbf{A}\mathbf{x})_\ell| = |y_\ell - f(\mathbf{t}_\ell)| \leq \eta$.

(b) Of course, the above result applies verbatim to the other algorithms as well for which recovery under conditions on the restricted isometry constants have been shown in Chapter 6. This includes Iterative Hard Thresholding, Hard Tresholding Pursuit, Orthogonal Matching Pursuit and Compressive Sampling Matching Pursuits.

Compared to the recovery condition of Theorems 12.18, 12.20, and 12.22 we pay some $\ln(s)$-factors, but we gain uniform recovery and we improve on the stability estimate. Compared to the condition of Theorem 9.10 for sub-Gaussian random matrices ensuring small restricted isometry constants (and, hence, uniform recovery by basis pursuit), which involves a factor of $\ln(N/s)$, we also obtain more log factors.

In the remainder of this section we develop the proof of Theorem 12.32. We first note that – unlike in the case of Gaussian (or sub-Gaussian) random matrices – the strategy of taking the probabilistic bound (12.31) for the condition number of a single column submatrix and then applying the union bound over all collections of s-element subsets of the N columns of $\mathbf{A}$ only leads to a rather poor estimate of the required samples m that allow recovery,

see Exercise 12.7. Indeed, the estimate (12.92) of m scales quadratically in s, while the desired estimate (12.66) obeys a linear scaling up to some log-factors. Below we pursue a different strategy that uses Dudley's inequality, Theorem 8.23, as a main tool.

Proof (of Theorem 12.32). We use the characterization of the restricted isometry constants in (6.2),

$$\delta_s = \max_{S \subset N, \mathrm{card}(S) \le s} \|\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2 \to 2} \, .$$

Let us introduce the set

$$D_{s,N} := \{\mathbf{z} \in \mathbb{C}^N : \|\mathbf{z}\|_2 \le 1, \|\mathbf{z}\|_0 \le s\} = \bigcup_{S \subset [N], \mathrm{card}(S) = s} B_S \, , \qquad (12.69)$$

where B_S denotes the unit sphere in $\mathbb{C}^S$ with respect to the ℓ_2-norm. The quantity

$$|\!|\!| \mathbf{B} |\!|\!|_s := \sup_{\mathbf{z} \in D_{s,N}} |\langle \mathbf{B}\mathbf{z}, \mathbf{z} \rangle|$$

defines a norm on self-adjoint matrices $\mathbf{B} = \mathbf{B}^* \in \mathbb{C}^{N \times N}$ (a semi-norm on all of $\mathbb{C}^{N \times N}$). Since $\widetilde{\mathbf{A}}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}$ is indeed selfadjoint, we have

$$\delta_s = |\!|\!| \widetilde{\mathbf{A}}^* \widetilde{\mathbf{A}} - \mathbf{Id} |\!|\!|_s \, .$$

Let $\mathbf{X}_\ell = \left(\overline{\phi_j(\mathbf{t}_\ell)}\right)_{j=1}^N \in \mathbb{C}^N$ be the random column vector of $\mathbf{A}^*$ associated to the sampling point $\mathbf{t}_\ell$, $\ell \in [m]$. Then $\mathbf{X}_\ell^*$ is a row of $\mathbf{A}$. Observe that $\mathbb{E}\mathbf{X}_\ell \mathbf{X}_\ell^* = \mathbf{Id}$ by the orthogonality relation (12.1). We can express the restricted isometry constant of $\widetilde{\mathbf{A}}$ as

$$\delta_s = |\!|\!| \frac{1}{m} \sum_{\ell=1}^m \mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id} |\!|\!|_s = \frac{1}{m} |\!|\!| \sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbb{E}\mathbf{X}_\ell \mathbf{X}_\ell^*) |\!|\!|_s \, . \qquad (12.70)$$

Let us first consider the expectation of δ_s. Using symmetrization (Lemma 8.4) we estimate

$$\mathbb{E} |\!|\!| \sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbb{E}\mathbf{X}_\ell \mathbf{X}_\ell^*) |\!|\!|_s \le 2\mathbb{E} |\!|\!| \sum_{\ell=1}^m \epsilon_\ell \mathbf{X}_\ell \mathbf{X}_\ell^* |\!|\!|_s \, . \qquad (12.71)$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$ is a Rademacher sequence, which is independent from the random sampling points $\mathbf{t}_\ell$, $\ell \in [m]$. The following lemma, which heavily relies on Dudley's inequality, is key to the estimate of the expectation above.

Lemma 12.36. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be vectors in $\mathbb{C}^N$ with $\|\mathbf{x}_\ell\|_\infty \le K$ for all $\ell \in [m]$. Then, for $s \le m$,*

$$\mathbb{E}\|\sum_{\ell=1}^{m}\epsilon_\ell \mathbf{x}_\ell\mathbf{x}_\ell^*\|_s \le C_1 K\sqrt{s}\ln^2(9s)\sqrt{\ln(8N)\ln(9m)}\sqrt{\|\sum_{\ell=1}^{m}\mathbf{x}_\ell\mathbf{x}_\ell^*\|_s}\,, \quad (12.72)$$

where $C_1 = \sqrt{2}C_0 C = 12\sqrt{2}C_0 \approx 54$. *Here,* $C = 12$ *is the constant in Dudley's inequality and* $C_0 = 3.1821$.

Proof. Observe that

$$E := \mathbb{E}\|\sum_{\ell=1}^{m}\epsilon_\ell \mathbf{x}_\ell\mathbf{x}_\ell^*\|_s = \mathbb{E}\sup_{\mathbf{u}\in D_{s,N}}\left|\sum_{\ell=1}^{m}\epsilon_\ell|\langle \mathbf{x}_\ell,\mathbf{u}\rangle|^2\right|\,.$$

This is the supremum of a Rademacher process, $X_\mathbf{u} = \sum_{\ell=1}^m \epsilon_\ell|\langle \mathbf{x}_\ell,\mathbf{u}\rangle|^2$, which has associated pseudo-metric

$$d(\mathbf{u},\mathbf{v}) = \left(\mathbb{E}|X_\mathbf{u} - X_\mathbf{v}|^2\right)^{1/2} = \sqrt{\sum_{\ell=1}^{m}\left(|\langle \mathbf{x}_\ell,\mathbf{u}\rangle|^2 - |\langle \mathbf{x}_\ell,\mathbf{v}\rangle|^2\right)^2}\,,$$

see also (8.44). Then, for $\mathbf{u},\mathbf{v}\in D_{s,N}$, the triangle inequality gives

$$\begin{aligned} d(\mathbf{u},\mathbf{v}) &= \left(\sum_{\ell=1}^{m}\left(|\langle \mathbf{x}_\ell,\mathbf{u}\rangle| - |\langle \mathbf{x}_\ell,\mathbf{v}\rangle|\right)^2\left(|\langle \mathbf{x}_\ell,\mathbf{u}\rangle| + |\langle \mathbf{x}_\ell,\mathbf{v}\rangle|\right)^2\right)^{1/2} \\ &\le \max_{\ell\in[m]}\big||\langle \mathbf{x}_\ell,\mathbf{u}\rangle| - |\langle \mathbf{x}_\ell,\mathbf{v}\rangle|\big| \sup_{\mathbf{u},\mathbf{v}\in D_{s,N}}\sqrt{\sum_{\ell=1}^{m}(|\langle \mathbf{x}_\ell,\mathbf{u}\rangle| + |\langle \mathbf{x}_\ell,\mathbf{v}\rangle|)^2} \\ &\le 2R\max_{\ell\in[m]}|\langle \mathbf{x}_\ell,\mathbf{u}-\mathbf{v}\rangle|\,, \end{aligned}$$

where

$$R = \sup_{\mathbf{u}\in D_{s,N}}\sqrt{\sum_{\ell=1}^{m}|\langle \mathbf{x}_\ell,\mathbf{u}\rangle|^2} = \sqrt{\|\sum_{\ell=1}^{m}\mathbf{x}_\ell\mathbf{x}_\ell^*\|_s}\,.$$

We further introduce the auxiliary seminorm

$$\|\mathbf{u}\|_X := \max_{\ell\in[m]}|\langle \mathbf{x}_\ell,\mathbf{u}\rangle|\,, \quad \mathbf{u}\in\mathbb{C}^N\,.$$

We derived that the rescaled process $X_\mathbf{u}/(2R)$ satisfies

$$\left(\mathbb{E}|X_\mathbf{u}/(2R) - X_\mathbf{v}/(2R)|^2\right)^{1/2} \le \|\mathbf{u}-\mathbf{v}\|_X.$$

Observe that the vector $\mathbf{u} = 0$ is contained in $D_{s,N}$. It follows from Dudley's inequality (8.48) with $t_0 = 0$ that

$$E \le 2CR\int_0^{\Delta(D_{s,N},\|\cdot\|_X)/2}\sqrt{\ln(\sqrt{2}\mathcal{N}(D_{s,N},\|\cdot\|_X,t))}dt\,, \quad (12.73)$$

with $C = 12$. By the Cauchy-Schwarz inequality, for $\mathbf{u} \in D_{s,N}$

$$\|\mathbf{u}\|_X = \max_{\ell \in [m]} |\langle \mathbf{x}_\ell, \mathbf{u} \rangle| \leq \|\mathbf{u}\|_1 \max_{\ell \in [m]} \|\mathbf{x}_\ell\|_\infty \leq K\sqrt{s}\|\mathbf{u}\|_2 \leq K\sqrt{s} \, . \quad (12.74)$$

Therefore, the diameter $\Delta(D_{s,N}, \|\cdot\|_X) = \sup_{\mathbf{u},\mathbf{v} \in D_{s,N}} \|\mathbf{u} - \mathbf{v}\|_X$ satisfies

$$\Delta(D_{s,N}, \|\cdot\|_X) \leq 2K\sqrt{s} \, .$$

Our next task is to estimate the covering numbers $\mathcal{N}(D_{s,N}, \|\cdot\|_X, t)$. We will do this in two different ways. One estimate will be good for small values of t and the other one for large values of t. For large values, we introduce the norm

$$\|\mathbf{z}\|_1^* := \sum_{j=1}^N (|\operatorname{Re}(z_j)| + |\operatorname{Im}(z_j)|) \, , \mathbf{z} \in \mathbb{C}^N \, ,$$

which is the usual ℓ_1-norm after identification of $\mathbb{C}^N$ with $\mathbb{R}^{2N}$. Then by the Cauchy-Schwarz inequality we have the embedding

$$D_{s,N} \subset \sqrt{2s} B^N_{\|\cdot\|_1^*} = \{\mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_1^* \leq \sqrt{2s}\} \, .$$

The next lemma provides an estimate of the covering numbers of an arbitrary subset of $B^N_{\|\cdot\|_1^*}$.

Lemma 12.37. *Let U be a subset of $B^N_{\|\cdot\|_1^*}$ and $0 < t < \sqrt{2}K$. Then*

$$\sqrt{\ln(\sqrt{2}\mathcal{N}(U, \|\cdot\|_X, t))} \leq 6K\sqrt{\ln(9m)\ln(8N)}t^{-1} \, .$$

Proof. Fix $\mathbf{x} \in U$. The idea is to approximate $\mathbf{x}$ by a finite set of very sparse vectors. In order to find a vector $\mathbf{z}$ from this finite set that is close to $\mathbf{x}$ we use the so called empirical method of Maurey. To this end we define a random vector $\mathbf{Z}$ that takes the value $\operatorname{sgn}(\operatorname{Re}(x_j))\mathbf{e}_j$ with probability $|\operatorname{Re}(x_j)|$, the value $i \operatorname{sgn}(\operatorname{Im}(x_j))\mathbf{e}_j$ with probability $|\operatorname{Im}(x_j)|$ for $j = 1, \ldots, N$, and the zero vector 0 with probability $1 - \|\mathbf{x}\|_1^*$. Here, $\mathbf{e}_j$ denotes the jth canonical unit vector, $(\mathbf{e}_j)_k = \delta_{j,k}$. Since $\|\mathbf{x}\|_1^* \leq 1$ this is a valid probability distribution. Note that

$$\mathbb{E}\mathbf{Z} = \sum_{j=1}^N \operatorname{sgn}(\operatorname{Re}(x_j))|\operatorname{Re}(x_j)|\mathbf{e}_j + i\sum_{j=1}^N \operatorname{sgn}(\operatorname{Im}(x_j))|\operatorname{Im}(x_j)|\mathbf{e}_j = \mathbf{x} \, .$$

Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$ be independent copies of $\mathbf{Z}$, where M is a number to be determined later. We attempt to approximate $\mathbf{x}$ with the M-sparse vector

$$\mathbf{z} = \frac{1}{M}\sum_{k=1}^M \mathbf{Z}_k \, .$$

We estimate the expected distance of $\mathbf{z}$ to $\mathbf{x}$ in $\|\cdot\|_X$ by first using symmetrization (Lemma 8.4),

$$\mathbb{E}\|\mathbf{z}-\mathbf{x}\|_X = \mathbb{E}\|\frac{1}{M}\sum_{k=1}^{M}(\mathbf{Z}_k - \mathbb{E}\mathbf{Z}_k)\|_X \le \frac{2}{M}\mathbb{E}\|\sum_{k=1}^{M}\epsilon_k\mathbf{Z}_k\|_X$$
$$= \frac{2}{M}\mathbb{E}\max_{\ell\in[m]}\left|\sum_{k=1}^{M}\epsilon_k\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle\right| ,$$

where $\boldsymbol{\epsilon}$ is a Rademacher sequence, which is independent of $(\mathbf{Z}_1,\dots,\mathbf{Z}_M)$. Now we fix a realization of $(\mathbf{Z}_1,\dots,\mathbf{Z}_M)$ and consider only expectation and probability with respect to $\boldsymbol{\epsilon}$ for the moment (that is, conditional on $(\mathbf{Z}_1,\dots,\mathbf{Z}_M)$). Since $\|\mathbf{x}_\ell\|_\infty \le K$ and $\mathbf{Z}_k$ has at most one non-zero component of magnitude 1, we have $|\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle| \le K$. It follows that

$$\|(\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle)_{k=1}^M\|_2 \le \sqrt{M}K, \quad \ell\in[m] .$$

It follows from Theorem 8.8 that the random variable $Y_\ell := \sum_{k=1}^M \epsilon_k\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle$ satisfies (conditional on the $\mathbf{Z}_k$),

$$\mathbb{P}_{\boldsymbol{\epsilon}}(|\mathbf{Y}_\ell| \ge \sqrt{M}Kt) \le 2e^{-t^2/2} , \quad t>0 .$$

Therefore, by the union bound

$$\mathbb{P}_{\boldsymbol{\epsilon}}(\max_{\ell\in[m]}|\mathbf{Y}_\ell| \ge \sqrt{M}Kt) \le 2m\, e^{-t^2/2} .$$

Proposition 7.14 yields then

$$\mathbb{E}_{\boldsymbol{\epsilon}}\max_{\ell\in[m]}\left|\sum_{k=1}^{M}\epsilon_k\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle\right| \le C\sqrt{M}K\sqrt{\ln(8m)} .$$

with $C=3/2$. (In the real case we may also use Proposition 7.29 which implies the slight better estimate $\mathbb{E}_{\boldsymbol{\epsilon}}\max_{\ell\in[m]}\left|\sum_{k=1}^M\epsilon_k\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle\right| \le \sqrt{2MK^2\ln(2m)}$.) By Fubini's theorem we finally obtain

$$\mathbb{E}\|\mathbf{z}-\mathbf{x}\|_X \le \frac{2}{M}\mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\boldsymbol{\epsilon}}\max_{\ell\in[m]}\left|\sum_{k=1}^{M}\epsilon_k\langle\mathbf{x}_\ell,\mathbf{Z}_k\rangle\right| \le \frac{3K}{\sqrt{M}}\sqrt{\ln(8m)} .$$

This implies that there exists a vector of the form

$$\mathbf{z} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{z}_k, \tag{12.75}$$

where each $\mathbf{z}_k$ is one of the vectors in $\{\pm\mathbf{e}_j, \pm i\mathbf{e}_j, 0 : j\in[N]\}$, such that

$$\|\mathbf{z}-\mathbf{x}\|_X \le \frac{3K}{\sqrt{M}}\sqrt{\ln(8m)}. \tag{12.76}$$

(Note that $\mathbf{z}$ has sparsity at most M.) In particular,

$$\|\mathbf{z}-\mathbf{x}\|_X \le t/2 \tag{12.77}$$

provided

$$\frac{3K}{\sqrt{M}}\sqrt{\ln(8m)} \le t/2. \tag{12.78}$$

Each $\mathbf{z}_k$ can take $4N+1$ values, so that $\mathbf{z}$ can take at most $(4N+1)^M \le (5N)^M$ values. For each $\mathbf{x} \in U$ we can therefore find a vector $\mathbf{z}$ of the form (12.75) such that $\|\mathbf{x}-\mathbf{z}\|_X \le t/2$. The definition of the covering numbers requires that each point of the covering belongs to U as well, but we only know that the points $\mathbf{z}$ are contained in $B^N_{\|\cdot\|_1^*}$. We can correct for this by replacing each point $\mathbf{z}$ by a point $\mathbf{z}' \in U$ with $\|\mathbf{z}-\mathbf{z}'\|_X \le t/2$ provided such a point exists. If such a point $\mathbf{z}'$ does not exist then we simply discard $\mathbf{z}$ as it will not be needed for the covering of U. Then for every $\mathbf{x} \in U$ we can find a point $\mathbf{z}' \in U$ from the new covering such that $\|\mathbf{x}-\mathbf{z}'\|_X \le \|\mathbf{x}-\mathbf{z}\|_X + \|\mathbf{z}-\mathbf{z}'\|_X \le t$. Again the number of points $\mathbf{z}'$ of the covering is bounded by $(5N)^M$.

The choice

$$M = \left\lfloor \frac{36K^2}{t^2}\ln(9m) \right\rfloor$$

satisfies (12.78). Indeed, then

$$\begin{aligned} M &\ge \frac{36K^2}{t^2}\ln(9m) - 1 \ge \frac{36K^2}{t^2}\ln(8m) + \frac{36K^2\ln(9/8)}{t^2} - 1 \\ &\ge \frac{36K^2}{t^2}\ln(8m) + \frac{36\ln(9/8)}{2} - 1 \ge \frac{36K^2}{t^2}\ln(8m) \end{aligned}$$

since $t \le \sqrt{2}K$ and $\frac{36\ln(9/8)}{2} \approx 2.12 > 1$. Therefore, (12.78) is satisfied. We deduce that the covering numbers can be estimated by

$$\begin{aligned} &\sqrt{\ln(\sqrt{2}N(U,\|\cdot\|_X,t))} \le \sqrt{\ln(\sqrt{2}(5N)^M)} \le \sqrt{\left\lfloor \frac{36K^2}{t^2}\ln(9m)\right\rfloor \ln(\sqrt{2}\cdot 5N)} \\ &\le 6K\sqrt{\ln(9m)\ln(8N)}t^{-1}, \end{aligned}$$

This completes the proof of the lemma. □

The estimate of the previous lemma will be good for large values of t. For small values of t we use a volumetric argument, that is, Proposition C.3. Note that $\|\mathbf{x}\|_X \le K\sqrt{s}\|\mathbf{x}\|_2$ for $\mathbf{x} \in D_{s,N}$ by (12.74). Using subadditivity (C.4) of the covering numbers, we obtain

$$\begin{aligned}\mathcal{N}(D_{s,N}, \|\cdot\|_X, t) &\le \sum_{S\subset[N],\mathrm{card}(S)=s} \mathcal{N}(B_S, K\sqrt{s}\|\cdot\|_2, t) \\ &= \sum_{S\subset[N],\mathrm{card}(S)=s} \mathcal{N}\left(B_S, \|\cdot\|_2, \frac{t}{K\sqrt{s}}\right) \le \binom{N}{s}\left(1+\frac{2K\sqrt{s}}{t}\right)^{2s} \\ &\le \left(\frac{eN}{s}\right)^s \left(1+\frac{2K\sqrt{s}}{t}\right)^{2s}.\end{aligned}$$

Hereby, we have also used the covering number estimate of Lemma C.3 (noting that we treat the s-dimensional complex unit ball, which is isometric to the real $2s$-dimensional unit ball), and the bound of the binomial coefficient in Lemma C.5. Together with Lemma 12.37 we get the two bounds

$$\begin{aligned}\sqrt{\ln(\sqrt{2}\mathcal{N}(D_{s,N}, \|\cdot\|_X, t))} &\le 6K\sqrt{2s}\sqrt{\ln(9m)\ln(8N)}t^{-1}, \quad 0 < t \le 2K\sqrt{s}, \\ \sqrt{\ln(\sqrt{2}\mathcal{N}(D_{s,N}, \|\cdot\|_X, t))} &\le \sqrt{2s}\sqrt{\ln(2^{1/4}eN/s) + \ln(1+2K\sqrt{s}/t)} \\ &\le \sqrt{2s}\left(\sqrt{\ln(cN/s)} + \sqrt{\ln(1+2K\sqrt{s}/t)}\right), \ t > 0\end{aligned}$$

with $c = 2^{1/4}e$. Next we combine these inequalities to estimate the "Dudley integral" in (12.73). We obtain, for arbitrary $\kappa \in (0, \Delta(D_{s,N})/2)$,

$$\begin{aligned}I &:= \int_0^{\Delta(D_{s,N})/2} \sqrt{\ln(\sqrt{2}\mathcal{N}(D_{s,N}, \|\cdot\|_X, t)}dt \\ &\le \sqrt{2s}\int_0^{\kappa}\left(\sqrt{\ln(cN/s)} + \sqrt{\ln\left(1+2K\sqrt{s}/t\right)}\right)dt \\ &\quad + 6K\sqrt{2s\ln(9m)\ln(8N)}\int_{\kappa}^{K\sqrt{s}} t^{-1}dt \\ &\le \sqrt{2s}\left(\kappa\sqrt{\ln(cN/s)} + \kappa\sqrt{\ln(e(1+2K\sqrt{s}/\kappa))}\right. \\ &\qquad \left. +6K\sqrt{\ln(9m)\ln(8N)}\ln(K\sqrt{s}/\kappa)\right).\end{aligned}$$

Hereby, we have applied Lemma C.10. The choice $\kappa = K/3$ yields

$$\begin{aligned}I &\le \sqrt{2s}K\left(\frac{1}{3}\sqrt{\ln(cN/s)} + \frac{1}{3}\sqrt{\ln(e(1+6\sqrt{s}))}\right. \\ &\qquad \left. +6\sqrt{\ln(9m)\ln(8N)}\ln(\sqrt{9s})\right) \\ &\le \sqrt{2s}KC_0\sqrt{\ln(9m)\ln(8N)}\ln(9s),\end{aligned}$$

where

$$C_0 := \frac{1}{3\sqrt{\ln(9)}\ln(9)} + \frac{1}{3}\sqrt{\frac{\ln(7e/3)}{\ln(9)^2} + \frac{1}{2\ln(9)}}\frac{1}{\sqrt{\ln(9)\ln(24)}} + 3 \approx 3.1821.$$

Hereby, we tacitly assumed $N \geq 3$ (otherwise the estimate is not interesting). Combining the above estimates with (12.73) completes the proof of Lemma 12.36 with $C_1 = \sqrt{2}C_0C = 12\sqrt{2}C_0 \approx 54$. □

Proof (of Theorem 12.32, continued). Let us now complete the proof of Theorem 12.32.

Estimate of Expectation. Recall from (12.70) that

$$E := \mathbb{E}\delta_s = m^{-1}\mathbb{E}\Big\|\sum_{\ell=1}^{m}(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathbf{Id})\Big\|_s$$

Set $G(K,s,m,N) = K\sqrt{s}\ln(9s)\sqrt{\ln(8N)\ln(9m)}$. Then Fubini's theorem, (12.71) and Lemma 12.36 implies that

$$\begin{aligned} E = m^{-1}\mathbb{E}\Big\|\sum_{\ell=1}^{m}(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathbf{Id})\Big\|_s &\leq \frac{2}{m}\mathbb{E}_{\mathbf{X}}\mathbb{E}_\epsilon\Big\|\sum_{\ell=1}^{m}\epsilon_\ell\mathbf{X}_\ell\mathbf{X}_\ell^*\Big\|_s \\ &\leq \frac{2C_1G(K,s,m,N)}{\sqrt{m}}\mathbb{E}_X\sqrt{\Big\|m^{-1}\sum_{\ell=1}^{m}\mathbf{X}_\ell\mathbf{X}_\ell^*\Big\|_s}\,. \end{aligned}$$

Inserting the identity $\mathbf{Id}$, applying the triangle inequality, $\|\mathbf{Id}\|_s = 1$ and using the Cauchy-Schwarz inequality for expectations we obtain

$$\begin{aligned} E &\leq 2C_1\frac{G(K,s,m,N)}{\sqrt{m}}\mathbb{E}\sqrt{m^{-1}\Big\|\sum_{\ell=1}^{m}(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathbf{Id})\Big\|_s + 1} \\ &\leq 2C_1\frac{G(K,s,m,N)}{\sqrt{m}}\sqrt{E+1}\,. \end{aligned}$$

Setting $D := 2C_1\frac{G(K,s,m,N)}{\sqrt{m}}$, we get $E \leq D\sqrt{E+1}$. Squaring this inequality and completing the squares yields $(E - D^2/2)^2 \leq D^2 + D^4/4$, which gives

$$E \leq \sqrt{D^2 + D^4/4} + D^2/2 \leq D + D^2\,. \tag{12.79}$$

If

$$D = \frac{2C_1K\sqrt{2s}\ln(9s)\sqrt{\ln(9m)\ln(8N)}}{\sqrt{m}} \leq \eta_1 \tag{12.80}$$

for some $\eta_1 \in (0,1)$ then

$$E = \mathbb{E}\delta_s \leq \eta_1 + \eta_1^2\,.$$

Probability estimate. It remains to show that δ_s does not deviate much from its expectation with high probability. To this end we use the deviation inequality of Theorem 8.39. By definition of the norm $\|\cdot\|_s$ we can write

$$
\begin{aligned}
m\delta_s &= |\!|\!|\sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})|\!|\!|_s = \sup_{S\subset[N],\mathrm{card}(S)\le s} \|\sum_{\ell=1}^m ((\mathbf{X}_\ell)_S (\mathbf{X}_\ell)_S^* - \mathbf{Id}_S)\|_{2\to 2} \\
&= \sup_{(z,w)\in Q_{s,N}^2} \mathrm{Re}\left(\left\langle \sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})\mathbf{z}, \mathbf{w}\right\rangle\right) \\
&= \sup_{(z,w)\in Q_{s,N}^{2,*}} \sum_{\ell=1}^m \mathrm{Re}\left(\langle (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})\mathbf{z}, \mathbf{w}\rangle\right) ,
\end{aligned}
$$

where $(\mathbf{X}_\ell)_S$ denotes the vector $\mathbf{X}_\ell$ restricted to the entries in S, and $Q_{s,N}^2 = \bigcup_{S\subset[N],\mathrm{card}(S)\le s} Q_{S,N}$ where

$$
Q_{S,N} = \{(\mathbf{z},\mathbf{w}) \,:\, \mathbf{z},\mathbf{w}\in\mathbb{C}^N \,,\ \|\mathbf{z}\|_2 = \|\mathbf{w}\|_2 = 1 \,,\ \mathrm{supp}\,\mathbf{z}, \mathrm{supp}\,\mathbf{w} \subset S\} \,.
$$

Further, $Q_{s,N}^{2,*}$ denotes a dense countable subset of $Q_{s,N}^2$. Introducing $f_{\mathbf{z},\mathbf{w}}(\mathbf{X}) = \mathrm{Re}(\langle(\mathbf{X}\mathbf{X}^* - \mathbf{Id})\mathbf{z},\mathbf{w}\rangle)$ we therefore have

$$
m^{-1}\delta_s = \sup_{(\mathbf{z},\mathbf{w})\in Q_{s,N}^{2,*}} \sum_{\ell=1}^m f_{\mathbf{z},\mathbf{w}}(\mathbf{X}_\ell) \,.
$$

Let us check the boundedness of $f_{\mathbf{z},\mathbf{w}}$ for $(\mathbf{z},\mathbf{w})\in Q_{S,N}$ with $\mathrm{card}(S)\le s$,

$$
\begin{aligned}
|f_{\mathbf{z},\mathbf{w}}(\mathbf{X})| &\le |\langle(\mathbf{X}\mathbf{X}^* - \mathbf{Id})\mathbf{z},\mathbf{w}\rangle| \le \|\mathbf{z}\|_2\|\mathbf{w}\|_2 \|\mathbf{X}^S\mathbf{X}_S^* - \mathbf{Id}_S\|_{2\to 2} \\
&\le \|\mathbf{X}_S(\mathbf{X}_S)^* - \mathbf{Id}_S\|_{1\to 1} = \max_{j\in S} \sum_{k\in S} |\phi_j(\mathbf{t})\overline{\phi_k(\mathbf{t})} - \delta_{j,k}| \\
&\le sK^2
\end{aligned}
$$

by the boundedness condition (12.2). Hereby, we used that the operator norm on ℓ_2 is bounded by the one on ℓ_1 for self-adjoint matrices (Lemma A.9), as well as the explicit expression (A.9) for $\|\cdot\|_{1\to 1}$. For the variance term σ^2 we estimate

$$
\begin{aligned}
\mathbb{E}|f_{\mathbf{z},\mathbf{w}}(\mathbf{X}_\ell)|^2 &\le \mathbb{E}|\langle(\mathbf{X}\mathbf{X}^* - \mathbf{Id})\mathbf{z},\mathbf{w}\rangle|^2 \\
&= \mathbb{E}\mathbf{w}^*(\mathbf{X}_S\mathbf{X}_S^* - \mathbf{Id})\mathbf{z}((\mathbf{X}_S\mathbf{X}_S^* - \mathbf{Id})\mathbf{z})^*\mathbf{w} \\
&\le \|\mathbf{w}\|_2^2 \mathbb{E}\|(\mathbf{X}_S\mathbf{X}_S^* - \mathbf{Id})\mathbf{z}((\mathbf{X}_S\mathbf{X}_S^* - \mathbf{Id})\mathbf{z})^*\|_{2\to 2} \\
&= \mathbb{E}\|(\mathbf{X}_S\mathbf{X}_S^* - \mathbf{Id})\mathbf{z}\|_2^2 = \mathbb{E}\|\mathbf{X}_S\|_2^2|\langle\mathbf{X},\mathbf{z}\rangle|^2 - 2\mathbb{E}|\langle\mathbf{X},\mathbf{z}\rangle|^2 + 1 \,.
\end{aligned}
$$

Hereby we used that $\|\mathbf{u}\mathbf{u}^*\|_{2\to 2} = \|\mathbf{u}\|_2^2$, see (A.13). Observe that

$$
\|\mathbf{X}_S\|_2^2 = \sum_{\ell\in S} |\phi_\ell(\mathbf{t})|^2 \le sK^2
$$

by the boundedness condition (12.2). Furthermore,

$$
\mathbb{E}|\langle\mathbf{X},z\rangle|^2 = \sum_{j,k\in S} z_j\overline{z_k}\mathbb{E}[\phi_k(\mathbf{t})\overline{\phi_j}(\mathbf{t})] = \|\mathbf{z}\|_2^2 = 1
$$

by orthogonality (12.1). Hence,

$$\begin{aligned}\mathbb{E}|f_{\mathbf{z},\mathbf{w}}(\mathbf{X}_\ell)|^2 &\le \mathbb{E}\|\mathbf{X}_S\|_2^2|\langle \mathbf{X},\mathbf{z}\rangle|^2 - 2\mathbb{E}|\langle \mathbf{X},\mathbf{z}\rangle|^2 + 1 \le (sK^2-2)\mathbb{E}|\langle \mathbf{X},\mathbf{z}\rangle|^2 + 1 \\ &= sK^2 - 1 < sK^2 \,.\end{aligned}$$

Now we are prepared to apply Theorem 8.39. Under the condition (12.80) this gives

$$\begin{aligned}&\mathbb{P}(\delta_s \ge \eta_1 + \eta_1^2 + \eta_2) \le \mathbb{P}(\delta_s \ge \mathbb{E}\delta_s + \eta_2) \\ &= \mathbb{P}(\|\!|\sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})|\!\|_s \ge \mathbb{E}\|\!|\sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbf{Id})|\!\|_s + \eta_2 m) \\ &\le \exp\left(-\frac{(\eta_2 m)^2}{2msK^2 + 4(\eta_1+\eta_1^2)msK^2 + 2\eta_2 msK^2/3}\right) \\ &= \exp\left(-\frac{m\eta_2^2}{sK^2}\frac{1}{2+4(\eta_1+\eta_1^2)+2\eta_2/3}\right) \le \exp\left(-c(\eta_1)\frac{m\eta_2^2}{sK^2}\right),\end{aligned}$$

with $c(\eta_1) = (2+4(\eta_1+\eta_1^2)+2/3)^{-1} \le (2+8+2/3)^{-1} = \frac{3}{32}$. The left hand term is less than ε provided

$$m \ge \widetilde{C}\eta_2^{-2} K^2 s \ln(\varepsilon^{-1})$$

with $\widetilde{C} = 32/3 \approx 10.66$.

Taking also (12.80) into account, we proved that $\delta_s \le \eta_1 + \eta_1^2 + \eta_2$ with probability at least $1-\varepsilon$ provided that m satisfies the two conditions

$$\begin{aligned}\frac{m}{\ln(9m)} &\ge \overline{C}\eta_1^{-2} K^2 s \ln^2(9s)\ln(8N)\,, \\ m &\ge \widetilde{C}\eta_2^{-2} K^2 s \ln(\varepsilon^{-1})\,.\end{aligned}$$

with $\overline{C} = 8C_1^2 = 16C_0^2C^2 = 16\cdot 12^2 \cdot C_0^2 \approx 23\,328$. Here, $C = 12$ is the constant of Dudley's inequality, Theorem 8.23. This finally completes the proof of Theorem 12.32. □

12.6 Discrete Bounded Orthonormal Systems

The two previous sections developed general bounds for sparse recovery of randomly sampled functions that have a sparse expansion in terms of a bounded orthonormal system. Several examples mentioned in Section 12.1 were actually discrete, i.e., the functions ϕ_k are actually the columns (or rows) of a unitary matrix $\mathbf{U} \in \mathbb{C}^{N\times N}$, $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{Id}$, with bounded entries,

$$\sqrt{N}\max_{k,t\in[N]} |U_{tk}| \le K\,, \tag{12.81}$$

see also (12.81). Among the mentioned examples were the Fourier matrix F and the matrix $\mathbf{U} = \mathbf{W}^*\mathbf{V}$ resulting from two incoherent orthonormal bases V, W.

Randomly sampling of entries corresponds to selecting the rows of the measurement matrix $\mathbf{A}$ uniformly at random from the rows of U. As already mentioned above, the probability model of taking the samples independently and uniformly at random has the slight disadvantage that some rows may be selected more than once with non-zero probability. In order to avoid this drawback, we discuss the following probability model. Let $\mathbf{u}_j^* \in \mathbb{C}^N$, $j \in [N]$, be the rows of $\mathbf{U} \in \mathbb{C}^{N \times N}$.

- **Selecting subsets uniformly at random.** In this probability model we choose the set $\Omega \subset [N]$ of rows uniformly at random among all subsets of $[N]$ of size m. This means that each subset is selected with equal probabilty. Since the number $\binom{N}{m}$ of such subsets is finite this is a valid probability model. The matrix $\mathbf{A}$ consists then of the rows $\mathbf{u}_j^*$, $j \in \Omega$. Clearly, $\mathbf{A}$ has exactly m rows in this probability model.

A matrix $\mathbf{A}$ resulting from selecting a subset of rows of $\mathbf{U}$ in the above way will be called a random partial unitary matrix. If $\mathbf{U} = \mathbf{F} \in \mathbb{C}^N$ is the Fourier matrix then we call $\mathbf{A}$ a random partial Fourier matrix.

The difficulty in analyzing the above probability model above consists in the fact that the events that $\mathbf{u}_j^*$, $j \in [N]$, has been selected as one of the rows, are not independent. We resolve this problem by simply relating results for this probability model to the results in the previous sections derived for the model of selecting rows (that is, the sampling points) independently at random. We only state the analogue of the uniform recovery result in Corollary (12.34)(a). Analogues of other statements in the previous sections can be derived as well.

Corollary 12.38. *Let $\mathbf{U} \in \mathbb{C}^{N \times N}$ be a unitary matrix with constant K in (12.81). Suppose that m, s, N are such that*

$$m \geq CK^2 s \ln^3(s) \ln(N). \tag{12.82}$$

Choose $\mathbf{A} \in \mathbb{C}^{m \times N}$ to be the matrix derived from U via selecting m rows uniformly at random from all m-element subsets of $[N]$. Then with probability at least $1 - N^{-\gamma \ln^3(s)}$ every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is recovered from the samples $\mathbf{y} = \mathbf{A}\mathbf{x}$ via ℓ_1-minimization.

Proof. Let $T' = \{t'_1, \ldots, t'_m\}$, where the $t'_\ell \in [N]$ are selected independently and uniformly at random from $[N]$. The size of T' is then random as well, since some of the t_ℓ may coincide. Further, for $k \leq m$ let $T_k \subset [N]$ be a subset of $[N]$ that is chosen uniformly at random among all subsets of cardinality k. For some subset $T \subset [N]$ let $F(T)$ be the event that ℓ_1-minimization fails to recover every s-sparse $\mathbf{x}$ from the samples on T, that is, from $\mathbf{y} = \mathbf{R}_T\mathbf{U}\mathbf{x}$. Next we note that it follows from Theorem (4.5) together with Remark (4.6)

that for $T \subset \widehat{T} \subset [N]$ that $F(\widehat{T}) \subset F(T)$. In other words, adding samples decreases the probability of failure. In particular, $\mathbb{P}(F(T_m)) \leq \mathbb{P}(F(T_k))$ for all $k \leq m$. Furthermore, conditionally on the event that $\operatorname{card}(T') = k$ for $k \leq m$, T' has the same distribution as T_k. We obtain

$$\begin{aligned}\mathbb{P}(F(T')) &= \sum_{k=1}^{m} \mathbb{P}(F(T')|\operatorname{card}(T') = k)\mathbb{P}(\operatorname{card}(T') = k) \\ &= \sum_{k=1}^{m} \mathbb{P}(F(T_k))\mathbb{P}(\operatorname{card}(T') = k) \geq \mathbb{P}(F(T_m)) \sum_{k=1}^{m} \mathbb{P}(\operatorname{card}(T') = k) \\ &= \mathbb{P}(F(T_m)).\end{aligned}$$

So the probability of failure in the model of selecting rows uniformly at random among all subsets of size m is bounded by the failure probability in the model of Corollary (12.34). This yields the claim. □

Another discrete probability model of interest uses Bernoulli selectors, see also Exercise (12.8).

12.7 Relation to the Λ_1-Problem

In this section we consider a discrete bounded orthonormal system, that is, the setup of Example 3. Let $\mathbf{U} \in \mathbb{C}^{N \times N}$ be a unitary matrix and set K as in (12.9),

$$K = \sqrt{N} \max_{k,t \in [N]} |U_{tk}|$$

We will compare the ℓ_1-norm and ℓ_2-norm of expansions in terms of subsets of this discrete bounded orthonormal system. To be more concrete, one may think of the Fourier matrix $\mathbf{U} = \mathbf{F}$ with entries $F_{jk} = e^{2\pi ijk/N}$ and constant $K = 1$.

Let $\Lambda \subset [N]$ and denote the (orthonormal) *rows* of $\mathbf{U}$ by $\mathbf{v}_k \in \mathbb{C}^N$, that is, $\mathbf{A}^\top = (\mathbf{v}_1|\ldots|\mathbf{v}_N)$. (In the Fourier case $(\mathbf{v}_k)_j = e^{2\pi ijk/N}$.) The trivial relation between the ℓ_1-norm and ℓ_2-norm (Hölder's inequality) implies that for all coefficient sequences $(b_k)_{k\in\Lambda} \in \mathbb{C}^\Lambda$,

$$\frac{1}{\sqrt{N}} \| \sum_{k\in\Lambda} b_k \mathbf{v}_k \|_1 \leq \| \sum_{k\in\Lambda} b_k \mathbf{v}_k \|_2 \, .$$

A valid converse of the above inequality is given by the trivial estimate $\|\cdot\|_2 \leq \|\cdot\|_1$. The Λ_1-problem consists in finding a large subset $\Lambda \subset [N]$ such that the much better estimate

$$\| \sum_{k\in\Lambda} b_k \mathbf{v}_k \|_2 \leq \frac{D(N)}{\sqrt{N}} \| \sum_{k\in\Lambda} b_k \mathbf{v}_k \|_1 \tag{12.83}$$

holds for all $(b_k)_{k\in\Lambda} \in \mathbb{C}^N$ and a "small" constant $D(N)$, say, $D(N) = C\log^\alpha(N)$. Such a Λ will be called a Λ_1-set. Then the ℓ_2-norm and the ℓ_1-norm (scaled by the factor $N^{-1/2}$) of corresponding orthogonal expansions on Λ will be almost equivalent.

Any singleton $\Lambda = \{\ell\}$, $\ell \in [N]$, is a Λ_1-set because by orthonormality and uniform boundedness

$$1 = \|\mathbf{v}_\ell\|_2^2 = \sum_{j=1}^N |(\mathbf{v}_\ell)_j|^2 \le \frac{K}{\sqrt{N}} \sum_{j=1}^N |(\mathbf{v}_\ell)_j| = \frac{K}{\sqrt{N}} \|\mathbf{v}_\ell\|_1 \,,$$

that is, $\|\mathbf{v}_\ell\|_1 \ge K^{-1}\sqrt{N}$, so that, for any $b_\ell \in \mathbb{C}$,

$$\|b_\ell \mathbf{v}_\ell\|_2 = |b_\ell| \le \frac{K}{\sqrt{N}} \|b_\ell \mathbf{v}_\ell\|_1$$

and (12.83) holds with $D(N) = K$ for $\Lambda = \{\ell\}$. However, singleton sets are of limited interest, and we would like to have large sets Λ, that is, $\mathrm{card}(\Lambda) \ge cN$.

It turns out that the null space property (and therefore the RIP) is quite related to the Λ_1-problem as stated next.

Proposition 12.39. *Let $\mathbf{U} \in \mathbb{C}^{N\times N}$ be a unitary matrix with rows $\mathbf{v}_\ell \in \mathbb{C}^N$, and let $\Omega \subset [N]$. Assume that the matrix $\mathbf{A} = \mathbf{R}_\Omega \mathbf{U}$, that is, the restriction of $\mathbf{U}$ to the columns indexed by Ω, satisfies the ℓ_2-robust null space property of order s with constants ρ and $\tau > 0$, see Definition 4.20. Then the complement $\overline{\Omega} = [N] \setminus \Omega$ is a Λ_1-set in the sense that*

$$\|\sum_{j\in\overline{\Omega}} b_\ell \mathbf{v}_\ell\|_2 \le \frac{1+\rho}{\sqrt{s}} \|\sum_{j\in\overline{\Omega}} b_\ell \mathbf{v}_\ell\|_1$$

for all $(b_j)_{j\in\overline{\Omega}} \in \mathbb{C}^{\overline{\Omega}}$.

Proof. Inequality (4.20) specialized to $p = q = 2$ and $\mathbf{u} = \mathbf{z} - \mathbf{x} \in \ker \mathbf{A}$ implies

$$\|\mathbf{u}\|_2 \le \frac{1+\rho}{\sqrt{s}} \|\mathbf{u}\|_1 \quad \text{for all } \mathbf{u} \in \ker \mathbf{A} \,. \tag{12.84}$$

Since $\mathbf{A}$ is the row submatrix of a unitary matrix, its kernel is spanned by the rows left out in $\mathbf{A}$, that is, by the ones indexed by $\overline{\Omega}$. Therefore, any $\mathbf{u} \in \ker \mathbf{A}$ takes the form

$$\mathbf{u} = \sum_{\ell\in\overline{\Omega}} b_\ell \mathbf{v}_\ell \,.$$

Combining these facts concludes the proof. □

Since the restricted isometry property implies the ℓ_2-robust null space property (Theorem 6.12) we can combine the above proposition with the RIP estimate for bounded orthonormal system to arrive at the following theorem on the Λ_1-problem.

Theorem 12.40. *Let $c \in (0,1)$. Then there exists a set $\Lambda \subset [N]$ with $\mathrm{card}(\Lambda) \geq cN$ such that*

$$\|\sum_{j\in\Lambda} b_\ell \mathbf{v}_\ell\|_2 \leq \frac{CK\log^2(N)}{\sqrt{N}} \|\sum_{j\in\Lambda} b_\ell \mathbf{v}_\ell\|_1 \tag{12.85}$$

for all $(b_\ell)_{\ell\in\Lambda} \in \mathbb{C}^\Lambda$. The constant C depends only on c, more precisely $C = C'(1-c)^{-1/2}$ for some universal constant C'.

Note that a slightly better estimate in terms of the log-factors is available, see the Notes section below. It is, however, a consequence of Lemma 12.5, that the term $\log^2(N)$ cannot be improved to something better than $\sqrt{\log N}$ in general, see Exercise 12.9.

Proof. Let $m = \lfloor (1-c)N \rfloor$. Then Theorem 12.30 (see also Remark 12.31) implies the existence of a set $\Omega \subset [N]$ such that the restricted isometry constant of the matrix $\mathbf{A} = \frac{1}{\sqrt{m}}\mathbf{R}_\Omega \mathbf{U}$ satisfies $\delta_{2s} \leq \delta^* := 0.4$ for the choice

$$s = \lceil C_0 \frac{m}{K^2 \log^4(N)} \rceil,$$

where C_0 is a universal constant. Then it follows from Theorem 6.12 that $\mathbf{A}$ satisfies the ℓ_2-robust null space property with constants ρ, τ depending only on δ^*. Clearly, the kernel of $\mathbf{A}$ does not depend on the scaling of $\mathbf{A}$, so that (12.84) holds also for $\mathbf{R}_\Omega \mathbf{U}$ and Proposition 12.39 applies to $\Lambda = \overline{\Omega}$ which has cardinality $\mathrm{card}(\Lambda) \geq cN$. We conclude that

$$\|\sum_{j\in\Lambda} b_\ell \mathbf{v}_\ell\|_2 \leq \frac{1+\rho}{\sqrt{s}} \|\sum_{j\in\Lambda} b_\ell \mathbf{v}_\ell\|_1 \ .$$

Taking into account our choices of s and m we arrive at

$$\|\sum_{j\in\Lambda} b_\ell \mathbf{v}_\ell\|_2 \leq \frac{1+\rho}{\sqrt{C_0(1-c)}} \frac{K\log^2(N)}{\sqrt{N}} \|\sum_{j\in\Lambda} b_\ell \mathbf{v}_\ell\|_1 \ .$$

This completes the proof. □

Notes

Background on Fourier analysis (Examples 1, 4, 5) can be found, for instance, in [172, 203, 337, 391, 442]. The complex exponentials of Examples 1 can be generalized to characters of commutative groups, see for instance [173, 378]. The sampling matrix (12.7) arising from continuously sampling trigonometric expansions has an (approximate) fast matrix vector multiplication called the nonequispaced fast Fourier transform [345]. Like the FFT, see Appendix C.1, it has complexity $\mathcal{O}(N \log N)$.

The uncertainty principle for the discrete Fourier transform in Corollary 12.3 was shown by Donoho and Stark in [142], where they also realized that the uncertainty principle is not only a negative statement, but can as well be used to derive positive conclusions about signal separation and recovery, see also [136]. Later in [158], Elad and Bruckstein derived the discrete uncertainty principle for general pairs of bases, Theorem 12.2. Kuppinger, Durisi and Bölcskei extended this further to an uncertainty principe for pairs of possibly redundant systems in [272]. An overview on uncertainty principles in general, including the classical uncertainty principles of Heisenberg and the one of Hardy, is provided in [174].

Lemma 12.5 concerning the existence of translates of large subgroups in arbitrary subsets of $\mathbb{Z}_2^n$ that then leads to the lower bound (12.29) of the necessary number of samples in undersampled Hadamard transforms involving a $\log N$ factor goes back to the work of Bourgain and Talagrand on the Λ_1-problem [402], but was published much later in [214].

The nonuniform recovery result Theorem 12.11 with random sign pattern seems to have first appeared in [355], while its improvement, Theorem 12.18, s was shown by E. Candès and J. Romberg in [79]. The idea of using random signs in order to derive recovery bounds for ℓ_1-minimization appeared first in [419]. The nonuniform recovery result of Theorem 12.20, in which the randomness in the signs of the coefficient vectors is removed, was shown by E. Candès and Y. Plan in [74]. The key technique in their proof, that is, the golfing scheme, was developed by Gross in [211] in the context of matrix completion and more general low rank matrix recovery problems, see also [361]. Instead of the deviation result for sums of random vectors in ℓ_2, Corollary 8.42 and the noncommutative Bernstein inequality (8.26), which were used in Section 12.4 to derive Lemmas 12.24, 12.25, 12.26, they use a slightly weaker version of the vector Bernstein inequality by D. Gross [211, Theorem 11], which also allows to remove the factor $(s+1)$ in (12.50). (This factor, however, is not important as it only enters in a term $\ln(2N(s+1)) \leq \ln(2N^2) \leq 2\ln(2N)$.) Moreover, E. Candès and Y. Plan also showed stronger stability estimates than the one of Theorem 12.22, in which the factor $\sqrt{s}$ can essentially be replaced by $\ln(s)^{3/2}$, while still keeping the bound (12.47) on the number of required samples (in contrast to the bound on the restricted isometry constants which involves more log-factors). To do so they introduced weak restricted isometry constants, and estimated these. This requires additional steps compared to the proof of the restricted isometry property in Section 12.5, see [74] for details.

The special case of partial random Fourier matrices (Example 4 in Section 12.1) was treated already in the first contribution of E. Candès, J. Romberg and T. Tao to compressive sensing [72]. They provided a nonuniform recovery result for deterministic sign patterns (in the noiseless case), where the number m of samples scales as

$$m \geq Cs\ln(N/\varepsilon) \tag{12.86}$$

in order to achieve recovery via ℓ_1-minimization with probability at least $1-\varepsilon$. This estimate was extended to random sampling of sparse trigonometric polynomials (as described in Example 1 in Section 12.1) by Rauhut in [352]. It is remarkable that this bound is still slightly better with regard to the dependence in ε than the result for general bounded orthonormal system, Theorem 12.20, where one encounters the term $\ln(N)\ln(\varepsilon^{-1})$ in contrast to $\ln(N/\varepsilon) = \ln(N) + \ln(\varepsilon^{-1})$ above. (For instance with $\varepsilon = N^{-\gamma}$ the first term results in $\gamma \ln^2(N)$, while the second only yields $(\gamma+1)\ln(N)$.) It is presently not clear how to arrive at a bound of the form (12.86) for general systems. The rather long proof of the sufficient condition (12.86) in [72, 352] heavily uses the algebraic structure of the Fourier system, and proceeds via involved combinatorial estimates. It does not seem possible to extend this approach to general bounded orthonormal systems.

The restricted isometry property for partial random Fourier matrices (Example 4) was first analyzed by E. Candès and T. Tao in [82], where they obtained the bound $m \geq C_\delta s \ln^5(N)\ln(\varepsilon^{-1})$ for the number of required samples, to achieve the restricted isometry property with sparsity s with probability at least $1-\varepsilon$. This estimate was then improved by M. Rudelson and R. Vershynin in [374] to $m \geq C_\delta s \ln^3(s)\ln(N)\ln(\varepsilon^{-1})$. (The proofs in both papers [82, 374] actually apply to more general discrete orthonormal systems as described in Example 3.) H. Rauhut [353, 355] generalized to possibly continuous bounded orthonormal systems and improved the probability estimate to the one stated in Theorem 12.30 by using Bernstein's inequality for suprema of empirical processes, Theorem 8.39. We followed Rudelson and Vershynin's approach in Section 12.5 to estimate the expected restricted isometry constants. With similar techniques it is also possible to directly establish the null space property for random sampling matrices arising from bounded orthonormal systems. We refer to [87] for details on this and for many other facts relating compressive sensing, random matrices and Banach space geometry.

Applications to recovery of functions in high dimensions are given in [103].

Further examples of bounded orthonormal systems. We discuss two other examples to which the developed theory applies. Since detailed proofs would lead too far from the scope of this book, we only mention the basic facts and refer to further literature for the details.

Haar wavelets and noiselets. This example is a special case of Example 6, which is potentially useful for image processing applications. It is convenient to start with a continuous description of Haar-wavelets and noiselets [104], and then pass to the discrete setup via sampling. The Haar scaling function on $\mathbb{R}$ is defined as the characteristic function of the interval $[0,1)$,

$$\phi(x) = \chi_{[0,1)}(x) = \begin{cases} 1 & \text{if } x \in [0,1), \\ 0 & \text{otherwise.} \end{cases} \tag{12.87}$$

The Haar wavelet is then defined as

$$\psi(x) = \phi(2x) - \phi(2x-1) = \begin{cases} 1 & \text{if } x \in [0, 1/2), \\ -1 & \text{if } x \in [1/2, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{12.88}$$

Further, denote

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad \phi_k(x) = \phi(x-k), \quad x \in \mathbb{R}, j \in \mathbb{Z}, k \in \mathbb{Z}. \tag{12.89}$$

It is straightforward to verify [445] that, for $n \in \mathbb{N}$, the Haar-wavelet system

$$\Psi_n := \{\phi_k, k \in \mathbb{Z}\} \cup \{\psi_{j,k},\, k = 0, \ldots, 2^j - 1, j = 0, \ldots, n-1\} \tag{12.90}$$

forms an orthonormal basis of

$$V_n = \{f \in L^2([0,1]) : f \text{ is constant on } [k2^{-n}, (k+1)2^{-n}), k = 0, \ldots, 2^n - 1\}.$$

Now let $N = 2^n$ for some $n \in \mathbb{N}$. Since the functions $\psi_{j,k}$, $j \leq n-1$, are constant on intervals of the form $[2^{-n}k, 2^{-n}(k+1))$ we conclude that the vectors $\widetilde{\phi}, \widetilde{\psi}^{(j,k)} \in \mathbb{C}^N$, $j = 0, \ldots, n-1, k = 0, \ldots, 2^j - 1$, with entries

$$\widetilde{\phi}_t = 2^{-n/2}\phi(t/N), \quad t = 0, \ldots, N-1$$
$$\widetilde{\psi}_t^{(j,k)} = 2^{-n/2}\psi_{j,k}(t/N), \quad t = 0, \ldots, N-1$$

form an orthonormal basis of $\mathbb{C}^N$. We collect these vectors as the columns of a unitary matrix $\Psi \in \mathbb{C}^{N \times N}$.

Next we introduce the noiselet system on $[0,1]$. Let $g_1 = \phi = \chi_{[0,1)}$ be the Haar scaling function and define, for $r \geq 1$, recursively the complex-valued functions

$$g_{2r}(x) = (1-i)g_r(2x) + (1+i)g_r(2x-1),$$
$$g_{2r+1}(x) = (1+i)g_r(2x) + (1-i)g_r(2x-1).$$

It is shown in [104] that the functions $\{2^{-n/2}g_r, r = 2^n, \ldots, 2^{n+1} - 1\}$ form an orthonormal basis of V_n. The key property for us consists in the fact that they are maximally incoherent with respect to the Haar basis. Indeed, Lemma 10 in [104] states that

$$\left| \int_0^1 g_r(x)\psi_{j,k}(x)dx \right| = 1 \quad \text{provided } r \geq 2^j - 1, \quad 0 \leq k \leq 2^j - 1. \tag{12.91}$$

For the discrete noiselet basis on $\mathbb{C}^N$, $N = 2^n$, we take the vectors

$$\widetilde{g}_t^{(r)} = 2^{-n}g_{N+r}(t/N), \quad r = 0, \ldots, N-1, \quad t = 0, \ldots, N-1.$$

Again, since the functions $g_{N+r}, r = 0, \ldots, N-1$, are constant on intervals of the form $[2^{-n}k, 2^{-n}(k+1))$ it follows that the vectors $\widetilde{g}^{(r)}$, $r = 0, \ldots, N-1$, form an orthonormal basis of $\mathbb{C}^N$. We collect these as columns into a unitary

matrix $G \in \mathbb{C}^{N \times N}$. Due to (12.91) the unitary matrix $U = G^* \Psi \in \mathbb{C}^{N \times N}$ satisfies (12.9) with $K = 1$ – or in other words, the incoherence condition (12.12) for the Haar basis and the noiselet basis holds with the optimal constant $K = 1$.

Due to the their recursive definition, both the Haar wavelet transform and the noiselet transform, that is, the application of Ψ and G and their adjoints, come with a fast algorithm that computes a matrix vector multiply in $\mathcal{O}(N \log(N))$ time.

As a simple signal model, images or other types of signals are sparse in the Haar wavelet basis. The setup of this chapter corresponds to randomly sampling such functions with respect to noiselets. For more information on wavelets we refer to [101, 114, 293, 445].

Legendre polynomials and more general orthogonal polynomial systems. The Legendre polynomials P_j, $j = 0, 1, 2, \ldots$, form a system of orthogonal polynomials, where L_j is a polynomial of precise degree j, and orthonormality is with respect to the normalized Lebesgue measure $dx/2$ on $[-1, 1]$, that is,

$$\frac{1}{2} \int_{-1}^{1} L_j(x) L_k(x) dx = \delta_{j,k} \, .$$

We refer to [13, 96, 397] for details on orthogonal polynomials, and in, particular on Legendre polynomials. The supremum norm of Legendre polynomials is given by [397]

$$\|L_j\|_\infty = \sup_{t \in [-1,1]} |L_j(t)| = \sqrt{2j+1} \, ,$$

so considering the polynomials $L_j, j = 0, \ldots, N-1$, yields the constant $K = \sqrt{2N-1}$ in (12.2). Unfortunately, K grows therefore rather quickly with N. Plugging this value of K for instance in the estimate (12.65) for the sufficient number of samples ensuring the RIP estimate $\delta_s \leq \delta$ yields

$$m \geq C \delta^{-2} N s \ln^3(s) \ln(N) \, .$$

This estimate is useless for compressive sensing because the number of measurements is required to be larger than the signal length N.

Of course, the question arises whether better estimates are possible, and indeed, the described problem can be circumvented with a trick [360]. The crucial point is that L_2-normalized Legendre polynomials P_j only grow unboundedly with j near the endpoint points ± 1 of the interval $[-1, 1]$. Define the function

$$v(t) = (\pi/2)^{1/2} (1 - t^2)^{1/4} \, .$$

Then Theorem 7.3.3 in [397] states that, for all $j \geq 1$,

$$\sup_{t \in [-1,1]} v(t) |L_j(t)| \leq \sqrt{2 + 1/j} \leq \sqrt{3} \, .$$

We define the auxiliary function system $Q_j(t) = v(t)L_j(t)$. Orthogonality is then with respect to the Chebyshev measure (arcsine distribution)

$$d\nu(t) = \pi^{-1}(1-t^2)^{-1/2}dt \ ,$$

where the normalization is such that ν is a probability measure on $[-1,1]$. Indeed,

$$\begin{aligned}\int_{-1}^{1} Q_j(t)Q_k(t)d\nu(t) &= \frac{1}{2}\int_{-1}^{1} L_j(t)L_k(t)v(t)^2(1-t^2)^{-1}dt \\ &= \int_{-1}^{1} L_j(t)L_k(t)dt = \delta_{j,k} \ .\end{aligned}$$

Therefore, the system $\{Q_j\}_{j=0}^{N-1}$ forms a bounded orthonormal system with constant $K = \sqrt{3}$ with respect to the Chebyshev measure. Clearly, the results derived in this chapter are valid therefore for the random sampling matrix $\mathbf{B} \in \mathbb{R}^{m\times N}$ having entries

$$B_{\ell,j} = Q_j(t_\ell),$$

where the t_ℓ are sampled independently according to the Chebyshev measure ν. (This causes that more sample points lie near the endpoints $[-1,1]$ compared to sampling from the uniform measure.) For instance the restricted isometry constant δ_s of $\frac{1}{\sqrt{m}}\mathbf{B}$ satisfies $\delta_s \le \delta$ provided $m \ge C\delta^{-2}s\ln(s)^3\ln(N)$. Multiplying with the function $v(t)$ can be interpreted as preconditioning of the Legendre sampling matrix. Defining $\mathbf{A} \in \mathbb{R}^{m\in N}$, $\mathbf{D} \in \mathbb{R}^{m\times m}$ via

$$A_{\ell,j} = L_j(t_\ell), \quad \text{and} \quad \mathbf{D} = \operatorname{diag}(v(t_\ell), \ell \in [m])$$

we realize that $\mathbf{B} = \mathbf{D}\mathbf{A}$. Since $\mathbf{D}$ is invertible with probability 1, the matrices $\mathbf{A}$ and $\mathbf{B}$ have the same null space almost surely. Now if $\frac{1}{\sqrt{m}}\mathbf{B}$ satisfies the restricted isometry property, say $\delta_{2s} < 0.4931$, then by Theorem 6.12 it satisfies the ℓ_2-robust null space property, and in particular, the stable null space property. The latter depends only on the null space of $\mathbf{B}$ which coincides with the one of $\mathbf{A}$, so that also $\mathbf{A}$ satisfies then the stable null space property. By Theorem 4.11 this in turn ensures stable sparse recovery via ℓ_1-minimization using the matrix $\mathbf{A}$. Altogether, choosing m independent random sampling points according to the Chebyshev measure ν with $m \ge C's\ln(s)^3\ln(N)$, the sampling matrix $\mathbf{A}$ satisfies the stable null space property of order s, and we have stable s-sparse recovery via ℓ_1-minimization. This setting is also interesting because it provides an example of a matrix $\mathbf{A}$ that does not satisfy the restricted isometry property itself, but it does possess the null space property. Also sampling points have to be sampled according to the Chebyshev measure although the Legendre polynomials are orthogonal according to the uniform measure.

Another view on the above example is that the diagonal matrix $\mathbf{D}$ serves as a preconditioner for $\mathbf{A}$, so that $\mathbf{B} = \mathbf{D}\mathbf{A}$ satisfies the restricted isometry.

Given Legendre type measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, we may multiply afterwards by the diagonal matrix $\mathbf{D}$, $\mathbf{y}' = \mathbf{D}\mathbf{y} = \mathbf{D}\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}$, and work directly with $\mathbf{B} = \mathbf{D}\mathbf{A}$ and the transformed measurements $\mathbf{y}'$ in any recovery algorithm. In this way, also Iterative Hard Thresholding (IHT), Iterative Thresholding Pursuit and CoSaMP can be used in the setup of random sampling of sparse Legendre polynomial expansions.

It is important to note that the Legendre transform matrix $\mathbf{A}$ has fast matrix vector multiplication algorithms, see [253, 230, 112, 344, 428], which may speed up recovery algorithms.

Extensions to other orthogonal polynomial expansions on $[-1, 1]$ are possible, where orthogonality is with respect to a weight function that satisfies a mild continuity condition. This includes for instance all Jacobi polynomials $P_k^{\alpha,\beta}$ with $\alpha, \beta \geq -1/2$ [397]. It is quite interesting that for all these families of orthogonal polynomials random sampling is with respect to the Chebyshev measure ν. We refer to [360] for details.

Spherical harmonics. Extensions of the previous example to the system of spherical harmonics [13] (an orthonormal system for $L^2(S^2)$, where S^2 is the 2-sphere in $\mathbb{R}^3$) are given in [67, 359]. Unfortunately, even the preconditioning trick above so far only yields the restricted isometry property provided $m \geq Cs \ln^3(s) N^{1/6} \ln(N)$ in [67] after an earlier bound in [359], where $N^{1/4}$ appeared instead of $N^{1/6}$. The result of [67] was established in a more general context by developing involved weighted L^∞ bounds for eigenfunctions of the Laplace operator on certain manifolds including the 2-sphere and thereby improving on estimates for associated Legendre polynomials in [269]. The key ingredient consists in identifying the right sampling measure. We refer to [67] for details.

An application of sparse spherical harmonic expansions for the inpainting problem of the cosmic microwave background are contained in [1]. Fast matrix vector multiplication algorithms for sampling matrices involving spherical harmonics are provided for instance in [230].

The **$\mathbf{\Lambda_1}$-problem** was investigated by Bourgain and Talagrand [402], who treated the case of general (not necessarily discrete) bounded orthonormal systems ϕ_j, $j \in [N]$, where orthonormality is with respect to a probability measure ν. The main result in [402] states the existence of a subset $\Lambda \subset [N]$ with $\mathrm{card}(\Lambda) \geq cN$ such that

$$\| \sum_{\ell \in \Lambda} b_\ell \phi_\ell \|_{L^2(\nu)} \leq CK \sqrt{\ln N \ln\ln N} \| \sum_{\ell \in \Lambda} b_\ell \phi_\ell \|_{L^1(\nu)} \, .$$

(Note that the factor $1/\sqrt{N}$ has to be introduced in the discrete setting of Section 12.7 because the usual ℓ_1 and ℓ_2-norms are not taken with respect to a probability measure, in contrast to the spaces $L^2(\nu)$ and $L^1(\nu)$ above.) It follows from Lemma 12.5 that a factor of $\ln N$ has to be present in this estimate, see Exercise 12.9. It is conjectured, however, that the term $\ln\ln N$

can be removed, but this conjecture remains open until today. Taking this fact into account together with the relation of the RIP with the Λ_1-problems (see the proof of Theorem 12.40), it seems to be a very hard problem to remove all log-factors expect one factor of $\ln N$ from the RIP estimate (12.65), as this would imply a positive solution to this conjecture (at least in the discrete case). Further results on the Λ_1-problem are contained in the paper [214], which also treats Kashin type decompositions for bounded orthonormal systems. The Λ_p-problem, for $p > 2$, was solved by Bourgain in [51], see also [52] for more information on this topic.

Signal separation: Similar mathematics as developed in this Chapter has been used in the problem of separating a signal that is a decomposition of two components, see p. 15 for a description of signal separation problems in general. One component is assumed to be sparse, and the other one sparse in the Fourier domain [142, 78, 419]. Assuming that the support set is random in at least one of the components then one can show the separation is possible via ℓ_1-minimization provided that the sparsity s in both components does not exceed $N/\sqrt{\ln N}$, where N is the signal length [78]. The proof methods are similar to the ones used for the nonuniform recovery guarantees of this Chapter.

Further types of structured random matrices. There are further types of structured random matrices, which are of interest for certain applications of compressive sensing. At the time of writing these type of random matrices and their interplay with compressive sensing were not yet completely understood. Therefore, we decided not to cover their analysis in detail. We mention below what is known about these.

Partial random circulant matrices. For a vector $\mathbf{b} = (b_0, b_1, \dots, b_{N-1}) \in \mathbb{C}^N$ the associated circulant matrix $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{b}) \in \mathbb{C}^{N \times N}$ is defined entry-wise by

$$\Phi_{k,j} = b_{j-k \mod N}, \quad k, j = 1, \dots, N.$$

The application of Φ to a vector is the discrete circular convolution,

$$(\mathbf{\Phi x})_j = (\mathbf{x} * \widetilde{\mathbf{b}})_j = \sum_{\ell=1}^{N} x_\ell \widetilde{b}_{j-\ell \mod N},$$

where $\widetilde{b}_j = b_{N-j}$. Let $\Theta \subset [N]$ be an arbitrary (deterministic) subset of cardinality $m < N$. Then we define the partial circulant matrix $\mathbf{\Phi}^\Theta = \mathbf{\Phi}^\Theta(\mathbf{b}) = \mathbf{R}_\Theta \mathbf{\Phi}(\mathbf{b}) \in \mathbb{C}^{m \times N}$ as the submatrix of $\mathbf{\Phi}$ consisting of the rows indexed by Θ. The application of a partial circulant matrix is clearly convolution with $\mathbf{b}$ followed by subsampling on Θ. It is important from a computational viewpoint that circulant matrices can be diagonalized using the discrete Fourier transform, see e.g. [198]. Therefore, there is a fast matrix vector multiplication algorithm for partial circulant matrices of complexity $\mathcal{O}(N \log(N))$ that uses the FFT.

Choosing the generator $\mathbf{b} = \boldsymbol{\epsilon}$ to be a Rademacher sequence makes the matrix $\mathbf{\Phi}^{\Theta} = \mathbf{\Phi}^{\Theta}(\boldsymbol{\epsilon})$ a structured random matrix, which is called partial random circulant matrix. It is then of interest to study recovery guarantees for ℓ_1-minimization and the restricted isometry property of the resulting matrix.

Of particular relevance is the case $N = mL$ with $L \in \mathbb{N}$ and $\Theta = \{L, 2L, \ldots, mL\}$. Then the application of $\mathbf{\Phi}^{\Theta}(\mathbf{b})$ corresponds to convolution with the sequence $\mathbf{b}$ followed by a downsampling by a factor of L. This setting was studied numerically in [427] by Tropp et al. (using orthogonal matching pursuit). Also of interest is the case $\Theta = [m]$ which was studied in [22, 228, 229].

Nonuniform recovery guarantees in the spirit of Theorem (12.11) for partial random circulant matrices in connection with ℓ_1-minimization were derived in [354, 355]. A sufficient condition on the number of samples is $m \geq Cs\log^2(N/\varepsilon)$ for recovery with probability at least $1 - \varepsilon$. After first non-optimal bounds in [22, 228, 229, 357], the so far best estimate on the restricted isometry constants of $\mathbf{\Phi}^{\Theta}(\boldsymbol{\varepsilon})$ developed by F. Krahmer, S. Mendelson and H. Rauhut in [266] states that $\delta_s \leq \delta$ with high probability provided

$$m \geq C\delta^{-2} s \log^2(s) \log^2(N) \, .$$

The proof uses chaining methods, and the analysis of the corresponding covering numbers uses some of the results developed in Section 12.5.

Time-Frequency structured random matrices. Introduce the translation and modulation (frequency shift) operators on $\mathbb{C}^m$ by

$$(\mathbf{T}^k \mathbf{g})_j = h_{j \ominus k} \quad \text{and} \quad (\mathbf{M}^\ell \mathbf{g})_j = e^{2\pi i \ell j/n} g_j \, ,$$

where $\ominus$ is subtraction modulo m. The operators $\boldsymbol{\pi}(\lambda) = \mathbf{M}^\ell \mathbf{T}^k$, $\lambda = (k, \ell)$, are called time-frequency shifts and the system $\{\boldsymbol{\pi}(\lambda) : \lambda \in [m]\times[m]\}$ of all time-frequency shifts forms a basis of the matrix space $\mathbb{C}^{m\times m}$ [275, 267]. Given a vector $\mathbf{g} \in \mathbb{C}^n$, the system of all possible time-frequency shifts of $\mathbf{g}$,

$$\{\boldsymbol{\pi}(\lambda)\mathbf{g}, \lambda \in [m] \times [m]\}$$

is called a full Gabor system with window $\mathbf{g}$ [210]. The matrix $\mathbf{A} = \mathbf{A}_{\mathbf{g}} \in \mathbb{C}^{m\times m^2}$ whose columns list the vectors $\boldsymbol{\pi}(\lambda)\mathbf{g}$, $\lambda \in [n]\times[n]$, of the Gabor system is referred to as Gabor synthesis matrix [356, 275, 98]. Note that $\mathbf{A}_{\mathbf{g}}$ allows for fast matrix vector multiplication algorithms based on the FFT, see for instance [166, 167]. Note that the matrix constructed in the proof of Proposition 5.13 is actually a Gabor synthesis matrix with window $g_j = \frac{1}{\sqrt{m}} e^{2\pi i j^3/m}$, which has small coherence $\mu = 1/\sqrt{m}$ (in the case that $m \geq 5$ is prime).

Let us choose the vector $\mathbf{g}$ at random,

$$\mathbf{g} = \frac{1}{\sqrt{m}} \boldsymbol{\epsilon} \, ,$$

where $\boldsymbol{\epsilon} \in \mathbb{C}^m$ is a Steinhaus sequence, that is, its entries are independent and uniformly distributed on the torus $\{z \in \mathbb{C}, |z| = 1\}$. Then the matrix $\mathbf{A} = \mathbf{A_g}$ becomes a structured random matrix, and we are interested in its performance for compressive sensing. A nonuniform recovery result is shown in [356], where the s sparse vector $\mathbf{x}$ is fixed (with deterministic sign pattern), then $\mathbf{A_g}$ is chosen at random and $\mathbf{y} = \mathbf{A}_g\mathbf{x}$ is observed. Exact recovery via ℓ_1-minimization occurs with high probability provided

$$m \leq cs \ln(m) \ .$$

(Note that in this setup $N = m^2$, so that $\ln(N) = \ln(m^2) = 2\ln(m)$.) After a first non-optimal estimate in [332], it was shown in [266] that the restricted isometry constants of $\mathbf{A_g}$ satisfies $\delta_s \leq \delta$ with high probabiliy provided

$$m \leq c\delta^{-2} s \log^2(s) \log^2(m) \ .$$

Sparse recovery with time-frequency structured random matrices has potential applications for the channel identification problem [333] in wireless communications and sonar [392, 303], as well as in radar [232]. Note that the results in [333] and [232] were derived based on coherence estimates and an analysis for random signals [419], similarly to the one outlined in Chapter 13.

More background on time-frequency analysis can be found in Gröchenig's excellent book [210].

Random Demodulator. For some engineering applications it is hard to realize sampling at random time-locations in hardware, especially, when the sampling rate is very high. In order to overcome this technological problem, one may instead multiply with random sign flips at a very high rate, integrate the signal over some time period and then sample equidistantly at a relatively low sampling rate [426]. The advantage is that all these components can be realized in hardware relatively easy. In particular, performing a sign flip at a very high rate is much simpler to realize than sampling at this high rate with high accuracy. In mathematical terms, the sampling matrix modelling this sensing scenario can be descriped as follows. Let $\mathbf{F} \in \mathbb{C}^{N \times N}$ be the N-dimensional discrete Fourier matrix. Further, let $\mathbf{D}_{\boldsymbol{\epsilon}} \in \mathbb{R}^{N \times N}$ be a random diagonal matrix having a Rademacher sequence $\boldsymbol{\epsilon}$ on its diagonal, and let finally $\mathbf{H} \in \mathbb{R}^{m \times N}$ modelling the integration process, where we assume for simplicity that m divides N. The jth row of $\mathbf{H}$ has N/m ones starting in column jN/m and is zeros elsewhere. An example for $m = 3$ and $N = 12$ is

$$\mathbf{H} = \begin{pmatrix} 1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1 \end{pmatrix} \ .$$

The measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is then the structured random matrix

$$\mathbf{A} = \mathbf{H}\mathbf{D}_{\boldsymbol{\epsilon}}\mathbf{F} \ ,$$

where the randomness comes from the Rademacher vector on the diagonal of $\mathbf{D}$. It has been shown in [426] that the restricted isometry constants of a suitably rescaled version of $\mathbf{A}$ satisfy $\delta_s \leq \delta$ with high probability provided

$$m \geq C_\delta s \ln^6(N) \, .$$

Therefore, the above described sampling mechanism can efficiently reconstruct signals that are s-sparse in the Fourier domain from m measurements using various algorithms including ℓ_1-minimization. The proof of the restricted isometry property uses parts of the analysis developed in Section 12.5. We refer to [426] for details.

Fast John-Lindenstrauss mappings. The combination of the bound of the restricted isometry property of Theorem 12.32 for random sampling matrices $\mathbf{A}$ including the random partial Fourier matrix together with Theorem 9.34 provides a Johnson-Lindenstrauss embedding for the mapping $\mathbf{AD}_\varepsilon$, where $\boldsymbol{\varepsilon}$ is a Rademacher vector, see Exercise 12.10. The important feature of $\mathbf{AD}_\varepsilon$ in contrast to a subgaussian random matrix is that comes with a fast matrix multiplication routine when $\mathbf{A}$ is for instance the partial random Fourier matrix [268, 6, 7]. A. Hinrichs and J. Vybiral investigated a similar scenario when $\mathbf{A}$ is a partial random circulant matrix [236, 439], see also [268, 357].

Sublinear Fourier Algorithms. It was notes even before the area of compressive sensing began to evolve that it is possible to design of algorithms for computing Fourier transforms of vectors that are sparse in the Fourier domain, which have sublinear runtime in the signal length N [193, 455]. (Since one needs to report only the locations and values of the s non-zero entries there is no a-priori contradiction in having a sublinear time algorithm.) Such algorithms are based on random samples in the time domain. In contrast to the setup of this chapter, however, the samples are not all independent in order to have enough algebraic structure that allows for fast computation. Although these algorithms were initially designed for fast computations, one can separate the sampling and the computation process so that they apply also in compressive sensing setups. A very appealing construction making use of prime numbers and the Chinese remainder theorem was presented by M. Iwen in [248, 249]. He provides a deterministic version of the algorithm, which uses $m \geq Cs^2 \log^4(N)$ samples and has runtime $\mathcal{O}(s^2 \log^4(N))$, and a randomized variant, which requires $m \geq Cs \log^4 N$ samples in runs in time $\mathcal{O}(s \log^4(N))$. A numerical evaluation of sublinear Fourier algorithms is presented in [250, 386]. In Chapter 14 we will see a sublinear sparse recovery algorithms in the different context of lossless expanders.

Exercises

12.1. Show that the Fourier matrix defined in (12.11) is unitary.

12.2. Let $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ with $\|\mathbf{x}\|_0 + \|\mathbf{z}\|_0 < 2\sqrt{N}$. Set $\mathbf{y} = \mathbf{x} + F\mathbf{z} \in \mathbb{C}^N$. Show that $(\mathbf{x}|\mathbf{z})$ is the unique solution to $\mathbf{y} = \mathbf{x}' + F\mathbf{z}'$ among all $\mathbf{x}', \mathbf{z}'$ with $\|\mathbf{x}'\|_0 + \|\mathbf{z}'\|_0 \leq 2\sqrt{N}$. In particular, the signal $\mathbf{y}$ can be separated uniquely into the components $\mathbf{x}$ and $\mathbb{K}\mathbf{z}$ under such sparsity assumption.

12.3. Let $T \subset [N]$ be an arbitrary subset of cardinality m. Show that every s-sparse $x \in \mathbb{C}^N$ with $s \leq \sqrt{N}$ can be recovered from its samples of the Fourier transform on T, i.e., from $y = R_T F x$ provided

$$m \geq N - \sqrt{N}.$$

12.4. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ and $S \subset [N]$. Assume that

$$\|\mathbf{A}_{S \cup \{\ell\}}^* \mathbf{A}_{S \cup \{\ell\}} - \mathbf{Id}\| \leq \delta \quad \text{for all } \ell \in [N] \setminus S \, .$$

Show that $\|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 \leq \frac{\delta}{1-\delta}$.

12.5. Let $\Gamma \subset \mathbb{Z}$ with $\mathrm{card}(\Gamma) = N$. Consider the non-equispaced random Fourier matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ from Example 1 in Section12.1. Improve Corollary 12.14 for this case using Corollary 8.10 (with $\lambda = 4/5$): Let μ be the coherence of the normalized matrix $\widetilde{\mathbf{A}} = \frac{1}{\sqrt{m}} \mathbf{A}$. Show that

$$\mu \leq \sqrt{\frac{5 \ln(5N^2/(2\varepsilon))}{4m}}$$

with probability at least $1 - \varepsilon$.

12.6. Check all details in the proof of Theorem 12.22.

12.7. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be the random matrix in (12.4) associated to sampling in bounded orthogonal systems. Use the probabilistic estimate (12.31) and the union bound to show that the restricted isometric constant δ_s of $\frac{1}{\sqrt{m}} \mathbf{A}$ satisfies $\delta_s \leq \delta$ with probability at least $1 - \varepsilon$ provided

$$m \geq \frac{8K^2}{3\delta^2} s^2 \left(\ln(eN/s) + \ln(2s/\varepsilon)/s \right) \, . \tag{12.92}$$

(In other words, the union bound is not strong enough to provide good estimates of δ_s, in particular, the union bound does not provide linear scaling of m in s.)

12.8. Bernoulli selectors.
Let $\mathbf{U} \in \mathbb{C}^{N \times N}$ be a unitary matrix with constant K in (12.81). Let δ_j, $j \in [N]$, be independent Bernoulli selectors, that is, random variables that take the value 1 with probability m/N and 0 with probability $1 - m/N$. Define the random sampling set $T = \{j, \delta_j = 1\}$, and let $\mathbf{A}$ be the random submatrix of $\mathbf{U}$ defined by $\mathbf{A} = \mathbf{R}_T \mathbf{U}$.

(a) In this context $\mathrm{card}(T)$ is random. Show that $\mathbb{E}\,\mathrm{card}(T) = m$ and derive an upper bound on $\mathbb{P}(|m - \mathrm{card}(T)| \geq t)$ for $t > 0$.

(b) Let $S \subset [N]$ with $\mathrm{card}(S) = s$. Set $\widetilde{\mathbf{A}} = \sqrt{N/m}\,\mathbf{A} = \sqrt{N/m} R_T U$. Then $\widetilde{\mathbf{A}}^* \widetilde{\mathbf{A}} = \frac{N}{m} \sum_{j=1}^N \delta_j \mathbf{X}_j \mathbf{X}_j^*$ where $(X_j)_t = \overline{U_{tj}}$, $t \in N$. Use the matrix Bernstein inequality to derive an upper bound on $\mathbb{P}(\|\widetilde{A}_S^* \widetilde{\mathbf{A}}_S - \mathbf{Id}\|_{2\to 2} \geq t)$ for $t > 0$.

12.9. Lower bound for the Λ_1-problem.
Let $H \in \mathbb{C}^{N\times N}$, $N = 2^n$, be the Hadamard matrix, as described in Example 5 and Section 12.2. Denote by $\mathbf{v}_\ell \in \mathbb{C}^N$ the columns of H. Let $\Lambda \subset [N]$ be an arbitrary subset of cardinality $\mathrm{card}(\Lambda) = cN$ for some $c \in (0,1)$. Show that there exists a vector $\mathbf{a} \in \mathbb{C}^\Lambda \setminus \{0\}$ such that

$$\|\sum_{j\in\Lambda} a_j \mathbf{v}_\ell\|_2 \geq c' \sqrt{\frac{\ln(N)}{N}} \|\sum_{j\in\Lambda} a_j \mathbf{v}_\ell\|_1 \ ,$$

where c' is a constant that only depends on c.
Consequently, the factor $\ln(N)^2$ in (12.85), cannot be improved to a better term than $\sqrt{\ln(N)}$ in general.

12.10. Fast Johnson-Lindenstrauss mappings.
Let $\mathbf{x}_1, \ldots, \mathbf{x}_M \in \mathbb{C}^N$ be an arbitrary set of points. Let $\mathbf{A}$ be the $m \times N$ random sampling matrix (12.4) associated to a bounded orthonormal system with constant $K \geq 1$ and $\mathbf{D}_\varepsilon \in \mathbb{R}^{N\times N}$ a diagonal matrix with a Rademacher vector $\boldsymbol{\epsilon}$ on the diagonal. Show that if $m \geq C\eta^{-2} \ln(M) \ln^4(N)$ then with high probability the matrix $\Phi = \mathbf{A}\mathbf{D}_{\boldsymbol{\varepsilon}} \in \mathbb{C}^{m\times N}$ provides a Johnson-Lindenstrauss embedding in the sense that

$$(1-\eta)\|\mathbf{x}_j - \mathbf{x}_k\|_2^2 \leq \|\Phi\mathbf{x}_j - \Phi\mathbf{x}_k\|_2^2 \leq (1+\eta)\|\mathbf{x}_j - \mathbf{x}_k\|_2^2 \quad \text{for all } j,k \in [M] \ .$$

13

Recovery of Random Signals

In this chapter we slightly change the point of view and work with deterministic measurement matrices but treat the sparse signal to be recovered as random. In particular, the support set of the sparse vector (and additionally the signs of the nonzero coefficients) will be chosen at random. In this scenario only mild conditions on the coherence of the measurement matrix are needed, indeed, much weaker than the ones outlined in Chapter 5. Recall that those conditions ensuring recovery of all s-sparse vectors together with the lower bound on the coherence, Theorem 5.7, lead to the quadratic bottleneck stating that the best possible bound on the required number of measurements m that can be derived with the coherence takes the form $m \geq Cs^2$. In contrast, we will see that it is possible to recover a random s-sparse vector using ℓ_1-minimization with $m \geq Cs \ln(N)$ measurements with high probability, provided the coherence satisfies $\mu \leq c(\ln N)^{-1}$. The latter condition is satisfied for many deterministic constructions of measurements and is indeed much milder than the optimal achievable bound $\mu \leq cm^{-1/2}$. Moreover, the coherence has the advantage that it is easy to evaluate for an explicitly given matrix.

Clearly, the results in this chapter are weaker than the ones of Chapter 5 in the sense that they apply only to *most* signals instead of to *all* signals, but they show that the deterministic bounds using coherence may be somewhat pessimistic even if no bounds on the restricted isometry constants are available. Moreover, the analysis in this chapter shows that one has to be careful in drawing conclusions from numerical experiments where often signals are generated at random; the result of testing measurement matrices on random signals does not tell much about recovery of *all* signals, or about the restricted isometry constants of the matrix. Indeed, the results of this chapter apply also to the counterexamples outlined in Section 12.2, where there are s-sparse signals which cannot be recovered from fewer than cs^2 measurements. Nevertheless *most* s-sparse signals can be recovered from far fewer samples.

The results of this chapter are especially important in the context of sparse approximation, where the matrix $\mathbf{A}$ takes the role of a redundant dictionary,

and $\mathbf{y} \in \mathbb{C}^m$ is a signal of interest that has a sparse representation in terms of $\mathbf{A}$, that is, $\mathbf{y} = \mathbf{A}\mathbf{x}$. In such context, one cannot design $\mathbf{A}$, but rather one uses a fixed deterministic matrix. Since it is very hard, and so far open, to verify the restricted isometry property for deterministic matrices in the optimal range of parameters, results for random signals that overcome the limits of deterministic results are important. Nevertheless, such types of bounds are also important in the context of compressive sensing when $\mathbf{A}$ takes the role of a measurement matrix that can be designed – especially in situations where good bounds for the restricted isometry property are not (yet) available.

We first derive bounds on the conditioning of a random column submatrix of a a given matrix. The methods draw on moment bounds, decoupling, and matrix deviation inequalities as developed in Chapter 8. In Section 13.2 we then develop recovery guarantees for ℓ_1-minimization based on Corollary 4.27 on recovery of individual sparse vectors.

13.1 Conditioning of Random Submatrices

Throughout this chapter we assume that the measurement matrix $\mathbf{A} = [\mathbf{a}_1|\dots|\mathbf{a}_N] \in \mathbb{C}^{m\times N}$ has ℓ_2-normalized columns, $\|\mathbf{a}_j\|_2 = 1$, and coherence

$$\mu = \max_{k\neq\ell} |\langle \mathbf{a}_k, \mathbf{a}_\ell\rangle| \, .$$

We will use two probability models for selecting a random support set $S \subset [N]$.

- **Uniform Model**. S is selected uniformly at random among all subsets of $[N]$ of cardinality $s \leq N$.
- **Bernoulli Model**. Choose $\delta = s/N$, and introduce independent Bernoulli selectors δ_j, $j \in [N]$, that take the value 1 with probability δ and the value 0 with probability $1-\delta$. Then define the random set

$$S = \{j \in [N], \delta_j = 1\} \, .$$

  The cardinality of S in this probability is random as well but its expectation satisfies $\mathbb{E}\,\mathrm{card}(S) = s$ according to choice $\delta = s/N$. By Hoeffding's inequality, Theorem 7.20, the size of S concentrates around s,

$$\mathbb{P}(|\,\mathrm{card}(S) - s| \geq t\sqrt{s}) = \mathbb{P}(|\sum_{j=1}^{N}(\delta_j - \delta)| \geq t\sqrt{s}) \leq 2e^{-t^2/2}.$$

  To see that Hoeffding's inequality applies, note that $|\delta_j - \delta| \leq 1$ and $\mathbb{E}(\delta_j - \delta) = 0$.

The first probability model may be more intuitive because the cardinality of S is always m, but the second probability model is easier to analyze because of the independence of the Bernoulli selectors δ_j. In any case, both probability models are closely related as we will see below.

We are interested in the conditioning of $\mathbf{A}_S$, that is, in the operator norm

$$\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \, .$$

We have the following probabilistic bound on this norm.

Theorem 13.1. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$, $m \le N$, with ℓ_2-normalized columns and coherence μ, and let $s \in [N]$. Select S at random according to the uniform model ($\operatorname{card}(S) = s$) or to the Bernoulli model ($\mathbb{E}\operatorname{card}(S) = s$). Assume that, for $\eta, \varepsilon \in (0,1)$,*

$$\mu \le c\frac{\eta}{\ln(N/\varepsilon)} \, , \tag{13.1}$$

$$\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2 \le c\frac{\eta^2}{\ln(N/\varepsilon)} \tag{13.2}$$

for an appropriate constant $c > 0$. Then

$$\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \le \eta$$

with probability at least $1 - \varepsilon$.

Remark 13.2. (a) The proof reveals the more precise estimate

$$\begin{aligned}&\mathbb{P}(\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \ge c_1\mu u + c_2\sqrt{\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2 u} + 2e\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2)\\ &\le c_3 N^4 \exp(-u)\end{aligned}$$

with $c_1 \approx 4.8078$, $c_2 \approx 11.21$ and $c_3 \approx 70.15$.

(b) Also a bound on the expectation can be shown. In case of the Bernoulli model,

$$\mathbb{E}\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \le 16\ln(2N)\mu + \sqrt{128\ln(2N)\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2} + 2\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2 \, . \tag{13.3}$$

In order for this result to have any value, of course the quantity $\frac{s}{N}\|\mathbf{A}\|^2$ should be small. Let us comment on this. First note that $\operatorname{tr}(\mathbf{A}^*\mathbf{A}) \le m\|\mathbf{A}^*\mathbf{A}\|_{2\to 2} = m\|\mathbf{A}\|_{2\to 2}^2$ because $\mathbf{A}^*\mathbf{A}$ has rank at most m so that

$$\|\mathbf{A}\|_{2\to 2}^2 \ge \frac{\operatorname{tr}(\mathbf{A}^*\mathbf{A})}{m} = \frac{N}{m} \, . \tag{13.4}$$

Equality is achieved for a unit norm tight frame. Indeed, recall from Definition 5.6 that a tight frame satisfies $\mathbf{A}\mathbf{A}^* = \lambda\mathbf{Id}_m$ so that it remains to verify that $\lambda = N/m$ when $\mathbf{A}$ has columns with unit ℓ_2-norm. In this case

$$\lambda m = \operatorname{tr}(\lambda\mathbf{Id}_m) = \operatorname{tr}(\mathbf{A}\mathbf{A}^*) = \operatorname{tr}(\mathbf{A}^*\mathbf{A}) = N \, ,$$

which yields the claimed relation $\|\mathbf{A}\|_{2\to 2}^2 = N/m$.

Unit norm tight frames are important examples in the context of sparse approximation, and they appear very frequently. An important special case of a unit norm tight frame arises when the columns of $\mathbf{A}$ form the union of several orthonormal bases. In this important case of a unit norm tight frame we therefore have

$$\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2 = \frac{s}{m} \, . \tag{13.5}$$

Choosing the probability $\varepsilon = N^{-2}$, say, condition (13.2) becomes then the familiar one

$$m \geq c\eta^{-2} s \ln(N) \, ,$$

while (13.1) is only a very mild condition on the coherence of $\mathbf{A}$,

$$\mu \leq c\eta \ln^{-1}(N) \, .$$

We develop the proof of Theorem (13.1) in several steps. Let us start with some notation. We introduce the hollow Gram matrix

$$\mathbf{H} = \mathbf{A}^* \mathbf{A} - \mathbf{Id} \, .$$

The matrix $\mathbf{H}$ has zero diagonal because $\mathbf{A}$ has ℓ_2-normalized columns by assumption. Let $\mathbf{P}_S$ be the projection operator onto S, that is, for $\mathbf{x} \in \mathbb{C}^N$,

$$(\mathbf{P}_S \mathbf{x})_\ell = \begin{cases} \mathbf{x}_\ell & \text{if } \ell \in S \, , \\ 0 & \text{if } \ell \notin S \, . \end{cases}$$

With this notation we realize that

$$\|\mathbf{A}_S^* \mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} = \|\mathbf{P}_S \mathbf{H} \mathbf{P}_S\|_{2\to 2} \, .$$

We will analyze the Bernoulli model and later reduce the uniform model to the Bernoulli model. We simply write $\mathbf{P} = \mathbf{P}_S$ and realize that P is the random diagonal matrix

$$\mathbf{P} = \operatorname{diag}(\delta_j, j \in [N]) \, .$$

We will bound the moments of $\|\mathbf{PHP}\|_{2\to 2}$. The fact that $\mathbf{P}$ appears twice makes a direct estimate difficult. We use decoupling to replace one instance with an independent copy of $\mathbf{P}$. Theorem 8.12 implies that, for $p \geq 1$,

$$(\mathbb{E}\|\mathbf{PHP}\|_{2\to 2}^p)^{1/p} \leq 2(\mathbb{E}\|\mathbf{P'HP}\|_{2\to 2}^p)^{1/p},$$

where $\mathbf{P'}$ is an independent copy of $\mathbf{P}$. Then the matrix $\mathbf{B} = \mathbf{P'H}$ is independent of $\mathbf{P}$. We first derive a moment estimate for $\|\mathbf{BP}\|_{2\to 2}$ with general $\mathbf{B}$.

Theorem 13.3. *Let $\mathbf{B} \in \mathbb{C}^{N\times N}$ and $\mathbf{P} = \operatorname{diag}\{\delta_j, j \in [N]\}$ be a random diagonal matrix of Bernoulli variables with mean $\delta \in [0,1]$. Let $p \geq 2$. Then*

$$(\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^p)^{1/p} \leq C(C_2 N)^{2/p}\sqrt{p}(\mathbb{E}\|\mathbf{BP}\|_{1\to 2}^p)^{1/p} + \sqrt{\delta}\|\mathbf{B}\|_{2\to 2} \, .$$

The constants satisfy $C \leq e^{1/(2e)}\sqrt{2/e} \approx 1.0310$ and $C_2 \leq 4.1878$.

Proof. Writing $\mathbf{B} = (\mathbf{b}_1|\cdots|\mathbf{b}_N)$ with column vectors $\mathbf{b}_j \in \mathbb{C}^N$ we observe that

$$\|\mathbf{BP}\|_{2\to 2}^2 = \|\mathbf{BPB}^*\|_{2\to 2} = \|\sum_{j=1}^N \delta_j \mathbf{b}_j \mathbf{b}_j^*\|_{2\to 2} ,$$

since $\mathbf{P} = \mathbf{P}^2$. Plugging in the mean, followed by the triangle inequality and symmetrization, Lemma 8.4, yields, for $r \geq 1$,

$$\begin{aligned}(\mathbb{E}\|\mathbf{BPB}^*\|_{2\to 2}^r)^{1/r} &\leq (\mathbb{E}\|\sum_{j=1}^N (\delta_j - \delta)\mathbf{b}_j\mathbf{b}_j^*\|_{2\to 2}^r)^{1/r} + \delta\|\sum_{j=1}^N \mathbf{b}_j\mathbf{b}_j^*\|_{2\to 2} \\ &\leq 2(\mathbb{E}\|\sum_{j=1}^N \epsilon_j\delta_j\mathbf{b}_j\mathbf{b}_j^*\|_{2\to 2}^r)^{1/r} + \delta\|\mathbf{BB}^*\|_{2\to 2} ,\end{aligned}$$

where $\boldsymbol{\epsilon}$ is a Rademacher sequence. The tail inequality for matrix Rademacher sums, Proposition 8.20, states that conditionally on $\boldsymbol{\delta}$,

$$\mathbb{P}_\varepsilon(\|\sum_{j=1}^N \epsilon_j\delta_j\mathbf{b}_j\mathbf{b}_j^*\|_{2\to 2} \geq t\sigma) \leq 2Ne^{-t^2/2} , \quad t > 0 ,$$

where

$$\begin{aligned}\sigma &= \big\|\sum_{j=1}^N (\delta_j\mathbf{b}_j\mathbf{b}_j^*)^2\big\|_{2\to 2}^{1/2} = \big\|\sum_{j=1}^N \delta_j^2\|\mathbf{b}_j\|_2^2\mathbf{b}_j\mathbf{b}_j^*\big\|_{2\to 2}^{1/2} \\ &\leq \max_{j\in[N]}\{\delta_j\|\mathbf{b}_j\|_2\}\big\|\sum_{j=1}^N \delta_j\mathbf{b}_j\mathbf{b}_j^*\big\|_{2\to 2}^{1/2} = \|\mathbf{BP}\|_{1\to 2}\|\mathbf{BP}\|_{2\to 2} ,\end{aligned}$$

where we have applied the explicit expression (A.10) of the norm $\|\cdot\|_{1\to 2}$. It follows from Proposition 7.13 that, for $r \geq 1$,

$$\left(\mathbb{E}_\varepsilon\|\sum_{j=1}^N \epsilon_j\delta_j\mathbf{b}_j\mathbf{b}_j^*\|_{2\to 2}^r\right)^{1/r} \leq C(C_2N)^{1/r}\sqrt{r}\|\mathbf{BP}\|_{1\to 2}\|\mathbf{BP}\|_{2\to 2}$$

with $C = e^{1/(2e)}e^{-1/2} \approx 0.729$ and $C_2 = 2C_{2,2} \approx 4.1878$. Taking expectation also with respect to $\boldsymbol{\delta}$ and applying the Cauchy-Schwarz inequality yields

$$\mathbb{E}\|\sum_{j=1}^N \epsilon_j\delta_j\mathbf{b}_j\mathbf{b}_j^*\|_{2\to 2}^r \leq C_2N\cdot C^r r^{r/2}(\mathbb{E}\|\mathbf{BP}\|_{1\to 2}^{2r})^{1/2}(\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^{2r})^{1/2} .$$

By combining the above estimates and choosing $r = p/2$ we arrive at

$$
\begin{aligned}
&(\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^p)^{2/p} = (\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^{2r})^{1/r} = (\mathbb{E}\|\mathbf{BPB}\|_{2\to 2}^r)^{1/r} \\
&\le 2(\mathbb{E}\|\sum_{j=1}^N \epsilon_j \delta_j \mathbf{b}_j \mathbf{b}_j^*\|_{2\to 2}^r)^{1/r} + \delta\|\mathbf{B}\|_{2\to 2}^2 \\
&\le 2(C_2 N)^{1/r} C_1 \sqrt{r} (\mathbb{E}\|\mathbf{BP}\|_{1\to 2}^{2r})^{1/(2r)} (\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^{2r})^{1/(2r)} + \delta\|\mathbf{B}\|_{2\to 2}^2 \\
&= 2(C_2 N)^{2/p} C_1 \sqrt{p/2} (\mathbb{E}\|\mathbf{BP}\|_{1\to 2}^p)^{1/p} (\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^p)^{1/p} + \delta\|\mathbf{B}\|_{2\to 2}^2 \,.
\end{aligned}
$$

Setting $E := (\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^p)^{1/p}$, this inequality takes the form $E^2 \le \alpha E + \beta$. Completing square gives $(E - \alpha/2)^2 \le \alpha^2/4 + \beta$ so that

$$
E \le \alpha/2 + \sqrt{\alpha^2/4 + \beta} \le \alpha + \sqrt{\beta} \,. \tag{13.6}
$$

We conclude that

$$
(\mathbb{E}\|\mathbf{BP}\|_{2\to 2}^p)^{1/p} \le (C_2 N)^{2/p} \sqrt{2} C_1 \sqrt{p} (\mathbb{E}\|\mathbf{BP}\|_{1\to 2}^p)^{1/p} + \delta^{1/2}\|\mathbf{B}\|_{2\to 2} \,.
$$

This finishes the proof. □

Remark 13.4. It follows from (8.114) that

$$
\mathbb{E}_\varepsilon \|\sum_{j=1}^N \delta_j \mathbf{b}_j \mathbf{b}_j^*\|_{2\to 2} \le \sqrt{2\ln(2N)} \|\mathbf{BP}\|_{1\to 2} \|\mathbf{BP}\|_{2\to 2} \,.
$$

Proceeding in the same way as in the previous proof shows that

$$
\mathbb{E}\|\mathbf{BP}\|_{2\to 2} \le \mathbb{E}(\|\mathbf{BP}\|_{2\to 2}^2)^{1/2} \le \sqrt{8\ln(2N)} (\mathbb{E}\|\mathbf{BP}\|_{1\to 2}^2)^{1/2} + \sqrt{\delta}\|\mathbf{B}\|_{2\to 2} \,. \tag{13.7}
$$

The above lemma requires a moment bound for $\|\mathbf{BP}\|_{1\to 2}$. Noting that we will later use $\mathbf{B} = \mathbf{P}'\mathbf{H}$, we actually need to estimate $\|\mathbf{P}'\mathbf{HP}\|_{1\to 2} = \|\mathbf{P}'\widetilde{\mathbf{B}}\|_{1\to 2}$ with $\widetilde{\mathbf{B}} = \mathbf{HP}$. The next lemma requires the norm

$$
\|\mathbf{B}\|_{\max} := \max_{j,k} |B_{j,k}| \,,
$$

that is, the ℓ_∞-norm over all matrix entries.

Lemma 13.5. *Let $\mathbf{B} \in \mathbb{C}^{N\times N}$ and $\mathbf{P} = \operatorname{diag}\{\delta_j, j \in [N]\}$ be a random diagonal matrix of Bernoulli variables with mean $\delta \in [0,1]$. Let $p \ge 2$. Then*

$$
(\mathbb{E}\|\mathbf{PB}\|_{1\to 2}^p)^{1/p} \le C_3 (2N)^{2/p} \sqrt{p} (\mathbb{E}\|\mathbf{PB}\|_{\max}^p)^{1/p} + \sqrt{\delta}\|\mathbf{B}\|_{1\to 2} \tag{13.8}
$$

with $C_3 = 2(2e)^{-1/2} \approx 0.8578$, and, for $u > 0$,

$$
\mathbb{P}(\|\mathbf{PB}\|_{1\to 2} \ge \sqrt{2\delta}\|\mathbf{B}\|_{1\to 2} + 2\|\mathbf{B}\|_{\max} u) \le 4N^2 e^{-u^2} \,. \tag{13.9}
$$

Proof. Similarly as in the previous proof we set $E := (\mathbb{E}\|\mathbf{PB}\|_{1\to 2}^p)^{1/p}$ and $r = p/2$. Symmetrization, Lemma 8.4, and the explicit expression for $\|\cdot\|_{1\to 2}$ yields

$$\begin{aligned} E^2 &= \left(\mathbb{E}\big(\max_{k\in[N]} \sum_{j=1}^N \delta_j |B_{jk}|^2\big)^r\right)^{1/r} \\ &\le 2\left(\mathbb{E}_{\boldsymbol{\delta}}\mathbb{E}_\varepsilon \max_{k\in[N]} \Big|\sum_{j=1}^N \epsilon_j\delta_j |B_{jk}|^2\Big|^r\right)^{1/r} + \delta\|\mathbf{B}\|_{1\to 2}^2 \,. \end{aligned} \tag{13.10}$$

Estimating the maximum by a sum and using Khintchine's inequality (8.9) we get

$$\begin{aligned} &\left(\mathbb{E}_\varepsilon \max_{k\in[N]} \Big|\sum_{j=1}^N \epsilon_j\delta_j |B_{jk}|^2\Big|^r\right)^{1/r} \le \left(\sum_{k=1}^N \mathbb{E}_\varepsilon \Big|\sum_{j=1}^N \epsilon_j\delta_j |B_{jk}|^2\Big|^r\right)^{1/r} \\ &\le 2^{1/r} e^{-1/2}\sqrt{r}\left(\sum_{k=1}^N \big(\sum_{j=1}^N \delta_j |B_{j,k}|^4\big)^{r/2}\right)^{1/r} \\ &\le 2^{1/r)} e^{-1/2}\sqrt{r} N^{1/r} \max_{k\in[N]} \sqrt{\Big(\max_{j\in[N]} \delta_j |B_{j,k}|^2\Big)\sum_{j=1}^N \delta_j |B_{j,k}|^2} \\ &= (2N)^{1/r} e^{-1/2}\sqrt{r}\|\mathbf{PB}\|_{\max}\|\mathbf{PB}\|_{1\to 2} \,. \end{aligned}$$

By the Cauchy-Schwarz inequality

$$\mathbb{E}_{\boldsymbol{\delta}}\mathbb{E}_\varepsilon \max_{k\in[N]} \Big|\sum_{j=1}^N \epsilon_j\delta_j |B_{jk}|^2\Big|^r \le 2Ne^{-r/2} r^{r/2} \left(\mathbb{E}\|\mathbf{PB}\|_{\max}^{2r}\right)^{1/2} \left(\mathbb{E}\|\mathbf{PB}\|_{1\to 2}^{2r}\right)^{1/2} \,.$$

Altogether

$$\begin{aligned} E^2 &\le 2e^{-1/2}(2N)^{1/r}\sqrt{r}(\mathbb{E}\|\mathbf{PB}\|_{\max}^{2r})^{1/(2r)}(\mathbb{E}\|\mathbf{PB}\|_{1\to 2}^{2r})^{1/(2r)} + \delta\|\mathbf{B}\|_{1\to 2}^2 \\ &\le 2e^{-1/2}(2N)^{2/p}\sqrt{p/2}(\mathbb{E}\|\mathbf{PB}\|_{\max}^p)^{1/p} E + \delta\|\mathbf{B}\|_{1\to 2}^2 \,. \end{aligned}$$

As above, since solutions to $E^2 \le \alpha E + \beta$ satisfy (13.6) we reach

$$E \le 2(2e)^{-1/2}(2N)^{2/p}\sqrt{p}(\mathbb{E}\|\mathbf{PB}\|_{\max}^p)^{1/p} + \delta^{1/2}\|\mathbf{B}\|_{1\to 2} \,.$$

While (13.9) with slightly worse constants can be deduced from (13.8) we find it instructive to derive the probability bound (13.9) via moment generating functions. For $\theta > 0$ we obtain by using symmetrization, Lemma 8.4, with the convex nondecreasing function $F(u) = \exp(\theta u)$,

$$\mathbb{E}\exp(\theta(\|\mathbf{PB}\|_{1\to 2}^2 - \delta\|\mathbf{B}\|_{1\to 2}^2)) \le \mathbb{E}\exp\left(2\theta \max_{k\in[N]} \Big|\sum_{j=1}^N \epsilon_j\delta_j|B_{jk}|^2\Big|\right)$$
$$\le \sum_{k=1}^N \mathbb{E}\exp\left(2\theta\Big|\sum_{j=1}^N \epsilon_j\delta_j|B_{jk}|^2\Big|\right) \le 2N\mathbb{E}_{\boldsymbol{\delta}}\exp(2\theta^2\|\mathbf{B}\|_{\max}^2\|\mathbf{PB}\|_{1\to 2}^2)\,,$$

where in the last step we have used the fact that $\sum_{j=1}^N \epsilon_j|B_{j,k}|^2$ is subgaussian by Theorem 7.27. Assuming that $2\theta\|\mathbf{B}\|_{\max}^2 \le 1/2$, Hölder's (or Jensen's) inequality gives

$$\exp(-\theta\delta\|\mathbf{B}\|_{1\to 2}^2)\mathbb{E}[\exp(\theta(\|\mathbf{PB}\|_{1\to 2}^2))] \le 2N\mathbb{E}[\exp(\theta\|\mathbf{PB}\|_{1\to 2}^2/2)]$$
$$\le 2N\left(\mathbb{E}[\exp(\theta\|\mathbf{PB}\|_{1\to 2}^2)]\right)^{1/2}\,.$$

Rearranging this inequality results in

$$\mathbb{E}\big[\exp(\theta(\|\mathbf{PB}\|_{1\to 2}^2 - 2\delta\|\mathbf{B}\|_{1\to 2}^2))\big] \le 4N^2 \quad \text{for all } 0<\theta\le \frac{1}{4\|B\|_{\max}^2}\,.$$

Markov's inequality together with the choice $\theta = 1/(4\|\mathbf{B}\|_{\max}^2)$ yields

$$\mathbb{P}(\|\mathbf{PB}\|_{1\to 2}^2 - 2\delta\|\mathbf{B}\|_{1\to 2}^2 \ge t) \le 4N^2e^{-\theta t} = 4N^2e^{-t/(4\|\mathbf{B}\|_{\max}^2)}\,.$$

Taking square roots inside the probability above and substituting $u = \sqrt{t}/(2\|\mathbf{B}\|_{\max})$ implies

$$\mathbb{P}(\|\mathbf{PB}\|_{1\to 2} \ge \sqrt{2\delta}\|\mathbf{B}\|_{1\to 2} + 2\|\mathbf{B}\|_{\max}u) \le 4N^2e^{-u^2}\,.$$

This point completes the proof. □

Remark 13.6. Using the fact that the random variable $\sum_{j=1}^N \epsilon_j\delta_j|B_{j,k}|^2$ is subgaussian conditional on δ by Theorem 7.27, one may invoke Theorem 7.29 to deduce

$$\mathbb{E}_\varepsilon \max_{k\in[N]} \Big|\sum_{j=1}^N \epsilon_j\delta_j|B_{jk}|^2\Big| \le \sqrt{2\ln(2N)}\|\mathbf{PB}\|_{\max}\|\mathbf{PB}\|_{1\to 2}\,.$$

Proceeding further as in the previous proof, one reaches

$$\mathbb{E}\|\mathbf{PB}\|_{1\to 2} \le (\mathbb{E}\|\mathbf{PB}\|_{1\to 2}^2)^{1/2}$$
$$\le \sqrt{8\ln(2N)}(\mathbb{E}\|\mathbf{PB}\|_{\max}^2)^{1/2} + \sqrt{\delta}\|\mathbf{B}\|_{1\to 2}\,. \tag{13.11}$$

Proof (of Theorem 13.1). We first derive a moment estimate for $\|\mathbf{PHP}\|_{2\to 2}$. Using the decoupling inequality (8.18) (noticing that $\mathbf{H}$ has zero diagonal) and applying Theorem 13.3 twice, we get, for $p \ge 2$,

$$
\begin{aligned}
&(\mathbb{E}\|\mathbf{PHP}\|_{2\to 2}^p)^{1/p} \le 2(\mathbb{E}\|\mathbf{PHP'}\|_{2\to 2}^p)^{1/p} \\
&\le 2\left(\mathbb{E}(C(C_2N)^{2/p}\sqrt{p}(\mathbb{E}_{P'}\|\mathbf{PHP'}\|_{1\to 2}^p)^{1/p} + \sqrt{\delta}\|\mathbf{PH}\|_{2\to 2})^p\right)^{1/p} \\
&\le 2C(C_2N)^{2/p}\sqrt{p}(\mathbb{E}\|\mathbf{PHP'}\|_{1\to 2}^p)^{1/p} + 2\sqrt{\delta}(\mathbb{E}\|\mathbf{HP}\|_{2\to 2}^p)^{1/p} \\
&\le 2C(C_2N)^{2/p}\sqrt{p}(\mathbb{E}\|\mathbf{PHP'}\|_{1\to 2}^p)^{1/p} + 2\sqrt{\delta}\cdot C(C_2N)^{2/p}\sqrt{p}(\mathbb{E}\|\mathbf{HP}\|_{1\to 2}^p)^{1/p} \\
&\quad + 2\delta\|\mathbf{H}\|_{2\to 2}^2 \,.
\end{aligned}
$$

Hereby, we have also used that $\|\mathbf{HP}\|_{2\to 2} = \|(\mathbf{HP})^*\|_{2\to 2} = \|\mathbf{PH}\|_{2\to 2}$ since $\mathbf{H}$ and $\mathbf{P}$ are self-adjoint. An application of Lemma 13.5 leads to

$$
\begin{aligned}
&(\mathbb{E}\|\mathbf{PHP}\|_{2\to 2}^p)^{1/p} \\
&\le 2C(C_2N)^{2/p}\sqrt{p}\left(C_3(2N)^{2/p}\sqrt{p}(\mathbb{E}\|\mathbf{PHP'}\|_{\max}^p)^{1/p} + \sqrt{\delta}(\mathbb{E}\|\mathbf{HP'}\|_{1\to 2}^p)^{1/p}\right) \\
&\quad + 2C(C_2N)^{2/p}\sqrt{p\delta}(\mathbb{E}\|\mathbf{HP}\|_{1\to 2}^p)^{1/p} + 2\delta\|\mathbf{H}\|_{2\to 2} \\
&= C_4(2C_2N^2)^{2/p}p(\mathbb{E}\|\mathbf{PHP'}\|_{\max}^p)^{1/p} + C_5(C_2N)^{2/p}\sqrt{p\delta}(\mathbb{E}\|\mathbf{HP}\|_{1\to 2}^p)^{1/p} \\
&\quad + 2\delta\|\mathbf{H}\|_{2\to 2} \,,
\end{aligned}
$$

with $C_4 = 2CC_3 = 2e^{1/(2e)}\sqrt{2/e}\cdot 2(2e)^{-1/2} = 4e^{1/(2e)-1} \approx 1.7687$ and $C_5 = 4C = 4e^{1/(2e)}\sqrt{2/e} \approx 4.1239$. Here we also used that $\mathbf{P'}$ is an independent copy of $\mathbf{P}$.

Next we exploit the properties of $\mathbf{H}$. Clearly, $\mu = \|\mathbf{H}\|_{\max}$ so that $\|\mathbf{PHP'}\|_{\max} \le \mu$ for any realization of $\mathbf{P}$ and $\mathbf{P'}$. Moreover,

$$
\|\mathbf{H}\|_{1\to 2} = \|\mathbf{A}^*\mathbf{A} - \mathbf{Id}\|_{1\to 2} \le \|\mathbf{A}^*\mathbf{A}\|_{1\to 2} = \max_{k\in[N]} \|\mathbf{A}^*\mathbf{a}_k\|_2 \le \|\mathbf{A}\|_{2\to 2}
$$

because the columns $\mathbf{a}_k$ are ℓ_2-normalized. It follows that

$$
\|\mathbf{HP}\|_{1\to 2} \le \|\mathbf{H}\|_{1\to 2} \le \|\mathbf{A}\|_{2\to 2} \tag{13.12}
$$

for any realization of $\mathbf{P}$. Moreover,

$$
\|\mathbf{H}\|_{2\to 2} = \|\mathbf{A}^*\mathbf{A} - \mathbf{Id}\|_{2\to 2} = \max\{1, \|\mathbf{A}\|_{2\to 2}^2 - 1\} \le \|\mathbf{A}\|_{2\to 2}^2 \,,
$$

because $\|\mathbf{A}\|_{2\to 2}^2 \ge N/m$ by (13.4). Therefore, we get

$$
\begin{aligned}
&(\mathbb{E}\|\mathbf{PHP}\|_{2\to 2}^p)^{1/p} \\
&\le (2C_2N^2)^{2/p}\left(C_4 p\,\mu + C_5\sqrt{p\delta}\|\mathbf{A}\|_{2\to 2}\right) + 2\delta\|\mathbf{A}\|_{2\to 2}^2 \,.
\end{aligned}
$$

It follows from Proposition 7.15 that, for $u \ge 2$,

$$
\mathbb{P}(\|\mathbf{PHP}\|_{2\to 2} \ge 2e\delta\|\mathbf{A}\|_{2\to 2}^2 + eC_4\mu u + eC_5\sqrt{\delta}\|\mathbf{A}\|_{2\to 2}\sqrt{u}) \le C_6N^4\exp(-u)
$$

with $C_6 = (2C_2)^2 \approx 70.15$. This implies that

$$
\|\mathbf{PHP}\|_{2\to 2} \le \eta
$$

with probability at least $1-\varepsilon$ provided

$$eC_4\mu\ln(C_6N^4/\epsilon) \le \eta/6\ , \quad eC_5\sqrt{\delta}\|\mathbf{A}\|_{2\to 2}\sqrt{\ln(C_6N^4/\epsilon)} \le 4\eta/5\ ,$$
$$\text{and} \quad 2e\delta\|\mathbf{A}\|_{2\to 2}^2 \le \eta/30\ .$$

The first two relations are equivalent to

$$\mu \le \frac{\eta}{C_7\ln(C_6N^4/\varepsilon)}\ ,$$
$$\delta\|\mathbf{A}\|_{2\to 2}^2 \le \frac{\eta^2}{C_8\ln(C_6N^4/\varepsilon)}\ , \tag{13.13}$$

with $C_7 = 6eC_4 \approx 28.85$, $C_8 = 25e^2C_5^2/16 \approx 196.35$. Then the second of these inequalities also implies $2e\delta\|\mathbf{A}\|_{2\to 2}^2 \le \eta/30$. Noting that $\delta = s/N$ finishes the proof for the Bernoulli model.

For the uniform model we proceed similarly as in the proof of Corollary 12.38 to bound the probability by the one for the Bernoulli model. Let $\mathbb{P}_B$ denote the probability in the Bernoulli model and $\mathbb{P}_{U,r}$ the one in the uniform model, where S is selected uniformly at random among all subsets of cardinality r. Then, for $t > 0$,

$$\begin{aligned}
&\mathbb{P}_B(\|\mathbf{PHP}\|_{2\to 2} \ge t)\\
&= \sum_{r=0}^{N} \mathbb{P}_B(\|\mathbf{P}_S\mathbf{H}\mathbf{P}_S\|_{2\to 2} \ge t|\,\mathrm{card}(S)=r)\,\mathbb{P}_B(\mathrm{card}(S)=r)\\
&\ge \sum_{r=s}^{N} \mathbb{P}_B(\|\mathbf{P}_S\mathbf{H}\mathbf{P}_S\|_{2\to 2} \ge t|\,\mathrm{card}(S)=r)\,\mathbb{P}_B(\mathrm{card}(S)=r)\\
&= \sum_{r=s}^{N} \mathbb{P}_{U,r}(\|\mathbf{P}_S\mathbf{H}\mathbf{P}_S\|_{2\to 2} \ge t)\,\mathbb{P}_B(\mathrm{card}(S)=r)\ .
\end{aligned} \tag{13.14}$$

Since the norm of a submatrix does not exceed the norm of the full matrix, we have for subsets $S \subset S' \subset [N]$

$$\|\mathbf{P}_S\mathbf{H}\mathbf{P}_S\|_{2\to 2} \le \|\mathbf{P}_{S'}\mathbf{H}\mathbf{P}_{S'}\|_{2\to 2}\ ,$$

which implies that

$$\mathbb{P}_{U,r+1}(\|\mathbf{P}_S\mathbf{H}\mathbf{P}_S\|_{2\to 2} \ge t) \le \mathbb{P}_{U,r}(\|\mathbf{P}_S\mathbf{H}\mathbf{P}_S\|_{2\to 2} \ge t)\ .$$

Moreover, since s is an integer, it is the median of the binomial distribution, see (7.6), so that

$$\sum_{r=s}^{N} \mathbb{P}_B(\mathrm{card}(S)=r) = \mathbb{P}_B(\mathrm{card}(S)\ge s) \le 1/2\ .$$

It follows that

$$\begin{aligned}\mathbb{P}_B(\|\mathbf{PHP}\|_{2\to 2} \geq t) &\geq \mathbb{P}_{U,s}(\|\mathbf{P}_S\mathbf{HP}_S\|_{2\to 2} \geq t)\mathbb{P}_B(\operatorname{card}(S) \geq s) \\ &\geq \frac{1}{2}\mathbb{P}_{U,s}(\|\mathbf{P}_S\mathbf{HP}_S\|_{2\to 2} \geq t) \, .\end{aligned}$$

This shows the claim also for the uniform model. □

13.2 Sparse Recovery via ℓ_1-Minimization

Based on the previous result on the conditioning of random submatrices we derive a sparse recovery result for random sparse signals via ℓ_1-minimization. Here we choose both the support of the nonzero coefficients as well as the signs of the nonzeros at random.

Theorem 13.7. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$, $m \leq N$, with ℓ_2-normalized columns and coherence μ, let $s \in [N]$. Select S at random according to the uniform model ($\operatorname{card}(S) = s$) or to the Bernoulli model ($\mathbb{E}\operatorname{card}(S) = s$). Choose a sparse vector $\mathbf{x} \in \mathbb{C}^N$ with $\operatorname{supp}\mathbf{x} = S$ and such that the signs of the nonzeros $\operatorname{sgn}(\mathbf{x}_S)$ form either a Steinhaus or a Rademacher sequence, which is independent of S. If*

$$\mu \leq \frac{c}{\ln(N/\varepsilon)} \, , \tag{13.15}$$

$$\frac{s}{N}\|\mathbf{A}\|_{2\to 2}^2 \leq \frac{c}{\ln(N/\varepsilon)} \tag{13.16}$$

for an appropriate constant $c > 0$, then ℓ_1-minimization applied to $\mathbf{y} = \mathbf{Ax}$ recovers $\mathbf{x}$ exactly with probability at least $1 - \varepsilon$.

Explicit constants can be found in the proof, see (13.18). We recall from (13.5) that for a unit norm tight frame relation (13.16) is satisfied under the familiar condition

$$m \geq Cs\ln(N/\varepsilon) \, ,$$

and only the mild condition (13.15) is imposed on the coherence.

Proof. The proof relies on the recovery result for vectors with random signs in Proposition 12.15, which in turn builds on the recovery conditions for individual vectors, Corollary 4.27. We are hence led to bounding the term

$$\max_{\ell\notin S} \|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 = \max_{\ell\notin S} \|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\mathbf{A}_S^*\mathbf{a}_\ell\|_2$$

for a random choice of S. If $\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \leq \eta$, as analyzed in Theorem 13.1, then $\|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\|_{2\to 2} \leq (1-\eta)^{-1}$ and we get the bound

$$\max_{\ell\notin S} \|\mathbf{A}_S^\dagger \mathbf{a}_\ell\|_2 \leq (1-\eta)^{-1} \max_{\ell\notin S} \|\mathbf{A}_S^*\mathbf{a}_\ell\|_2 \, .$$

Using $\mathbf{H} = \mathbf{A}^*\mathbf{A} - \mathbf{Id}$ and the projection $\mathbf{P} = \mathbf{P}_S$ as in the previous section, we realize that

$$\max_{\ell \notin S} \|\mathbf{A}_S^* \mathbf{a}_\ell\|_2 = \|\mathbf{PH}(\mathbf{Id} - \mathbf{P})\|_{1\to 2} \le \|\mathbf{PH}\|_{1\to 2} \, .$$

Assuming the Bernoulli model with $\delta = s/N$ for now, Lemma 13.5 implies that

$$\mathbb{P}(\|\mathbf{PH}\|_{1\to 2} \ge \sqrt{2\delta}\|\mathbf{H}\|_{1\to 2} + 2\|\mathbf{H}\|_{\max} u) \le 4N^2 e^{-u^2}$$

Since $\|\mathbf{H}\|_{\max} = \mu$ and $\|\mathbf{H}\|_{1\to 2} \le \|\mathbf{A}\|_{2\to 2}$, see (13.12), we therefore get

$$\mathbb{P}(\|\mathbf{PH}\|_{1\to 2} \ge \sqrt{2\delta}\|\mathbf{A}\|_{2\to 2} + 2\mu u) \le 4N^2 e^{-u^2} \, . \tag{13.17}$$

It follows from Proposition 12.15 that, for any $\alpha \in (0,1)$, the probability of failure of reconstruction via ℓ_1-minimization can be bounded by

$$\begin{aligned} P &:= \mathbb{P}(\max_{\ell \notin S} \|\mathbf{A}_S^\dagger \mathbf{a}_\ell\| \ge \alpha) + 2(N-s)e^{-\alpha^{-2}/2} \\ &\le 2Ne^{-\alpha^{-2}/2} + \mathbb{P}(\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \ge 3/4) + \mathbb{P}(\|\mathbf{PH}\|_{1\to 2} \ge \alpha/4) \, , \end{aligned}$$

where we have set $\eta = 3/4$ in the inequalities in the beginning of this proof. Let us choose $\alpha = 1/\sqrt{2\ln(C_6 N^4/\varepsilon)}$. Then the first term is bounded by $\varepsilon/8$. Assume that with the constants $C_6 \approx 70.15, C_7 \approx 28.85, C_8 \approx 196.35$ from the proof of Theorem 13.1

$$\mu \le \frac{3/4}{C_7 \ln(C_6 N^4/\varepsilon)} \, , \quad \text{and} \quad \sqrt{\delta}\|\mathbf{A}\|_{2\to 2} \le \frac{3/4}{\sqrt{C_8 \ln(C_6 N^4/\varepsilon)}} \, . \tag{13.18}$$

Then, by Theorem 13.1,

$$\mathbb{P}(\|\mathbf{A}_S^*\mathbf{A}_S - \mathbf{Id}_S\|_{2\to 2} \ge 3/4) \le \varepsilon \, .$$

The second inequality in (13.18) also implies

$$\sqrt{2\delta}\|\mathbf{A}\|_{2\to 2} \le c_1 \alpha$$

with $c_1 = 3/(2C_8^{1/2}) \approx 0.107$. Further, with $c_2 = 0.14$ and $u = C_9\sqrt{\ln(C_6 N^4/\varepsilon)}$ for $C_9 = \frac{c_1 \cdot 4}{\sqrt{2}\cdot 3} C_7 \approx 3.808$ we have $\mu u \le c_2 \alpha$ so that

$$\mu u + \sqrt{2\delta}\|\mathbf{A}\|_{2\to 2} \le (c_1 + c_2)\alpha \le \alpha/4 \, .$$

Therefore, by (13.17) we get

$$\begin{aligned} \mathbb{P}(\|\mathbf{PH}\|_{1\to 2} \ge \alpha/4) &\le \mathbb{P}(\|\mathbf{PH}\|_{1\to 2} \ge \sqrt{\delta}\|\mathbf{A}\|_{2\to 2} + 2\mu u) \le 4N^2 \exp(-u^2) \\ &= 4N^2 \exp(-C_9^2 \ln(C_6 N^4/\varepsilon)) \le 4N^2 \frac{\varepsilon}{C_6 N^4} \le c_3 \varepsilon \, , \end{aligned}$$

with $c_3 = 4/C_6 \approx 0.0570$. Altogether the failure probability is bounded by $\varepsilon + \varepsilon/8 + c_3\epsilon$ and replacing ε by $\varepsilon/(1 + 0.125 + c_3)$ completes the proof for the Bernoulli model.

For the uniform model we proceed similarly as in the proof of Theorem 13.1 to show that

$$\mathbb{P}_{U,s}(\|\mathbf{PH}\|_{1\to 2} \geq t) \leq 2\mathbb{P}_B(\|\mathbf{PH}\|_{1\to 2} \geq t) \,, \tag{13.19}$$

where again $\mathbb{P}_{U,s}$ denotes the probability under the uniform model, where subsets S are selected uniformly at random among all subsets of cardinality s, while $\mathbb{P}_B$ denotes the probability under the Bernoulli model. With this point the proof is concluded in the same way as above. □

Notes

Theorem 13.7 explains why one can expect recovery of s-sparse signals from $m \geq Cs\log(N)$ measurements under much milder conditions on the coherence as in Chapter 5 and even in situations when estimates on the restricted isometry constants are unavailable or even known to fail. In particular, usual numerical performance evaluations take the support set of the signal and the non-zero coefficients at random, so that the results of this chapter explain the high success rate of these experiments. However, one should be careful when drawing conclusion for the recovery of "real-world" signals from such numerical experiments. Certainly, Theorem 13.7 still indicates that recovery is possible under mild conditions, but it is often hard to argue rigorously that the support set of a "natural" signal is random. For instance, the wavelet coefficients of a natural image follow the edges of an image, so that the nonzero (large) coefficients are rather organized in trees. Such tree structure is certainly not random – at least the support set does not follow a *uniform* distribution. Therefore, the results of the preceding chapters holding for *all* sparse signals remain very important. Moreover, we did not derive results on the stability of reconstruction. Although it is possible to show stability under *random* noise [73], such results are weaker in nature than the ones based on the restricted isometry property.

Conditioning of random submatrices (subdictionaries) based on coherence was first studied by J. Tropp in [419], where he derived slightly weaker estimates. Indeed, the bounds in [419] require in addtion to (13.2) that $\mu^2 s\ln(s) \leq c$, which is harder to satisfy than (13.1) (unless s is tiny, in which case the "quadratic" bounds of Chapter 5 would also be fine). J. Tropp refined his estimates later in [418] to the ones presented in this chapter. Using more sophisticated decoupling techniques together with the matrix Chernoff inequality [424], S. Chrétien and S. Darses [97] obtain slightly better constants than the ones stated in Theorem 13.1 on the conditioning of random submatrices. Candès and Plan applied Tropp's result in the context of statistical

sparse estimation using the Dantzig selector, where they also allowed noise on the measurements [73]. Tropp's paper [419] also contains refined results for the case where the matrix $\mathbf{A}$ is the concatenation of two orthonormal bases. Candès and Romberg's paper [78] treats the special case of the concatenation of the identity and the Fourier basis, see also [420].

Tropp's original methods in [419, 418] use noncommutative Khintchine inequalities (8.112) instead of the tail inequality for matrix-valued Rademacher sums, (8.36). The "random compression bound" of Theorem 13.3 goes back to Rudelson and Vershynin [373], see also [421, Proposition 12].

Analyses of sparse recovery algorithms for random choices of signals have been carried out as well in the context of multichannel sparse recovery or multiple measurement vectors [160, 208], where the measurement matrix $\mathbf{A}$ is applied to a collection of sparse signals $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)} \in \mathbb{C}^N$ with common support, that is,

$$(\mathbf{y}^{(1)} | \cdots | \mathbf{y}^{(L)}) = \mathbf{A}(\mathbf{x}^{(1)} | \cdots | \mathbf{x}^{(L)})$$

and $\operatorname{supp} \mathbf{x}^{(\ell)} = S$ for all $\ell \in [L]$. In this context a nonzero coefficient is actually a vector $\mathbf{x}_k = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) \in \mathbb{C}^L$, which will be chosen at random (for instance, according to a multivariate Gaussian distribution or the uniform distribution on the sphere). The results in [160, 208] apply to multichannel variants of ℓ_1-minimization and greedy algorithms, and predict that the probability failure decreases exponentially in L provided that a very mild condition on the number of samples hold. The estimates outlined in this chapter are partly used in these contributions.

The bound on the conditioning of random matrices, Theorem 13.1 is somewhat related to the Bourgain-Tzafriri restricted invertibility theorem [55, 56]. We state a strengthened version due to Spielman and Srivastava [389], and Casazza and Pfander [85] (who provided an upper bound for the first time).

Theorem 13.8. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ with ℓ_2-normalized columns and $\alpha \in (0, 1)$ be a prescribed parameter. There exists a subset $S \subset [N]$ with*

$$\operatorname{card}(S) \geq \frac{\alpha^2 N}{\|\mathbf{A}\|_{2\to 2}^2}$$

such that

$$(1-\alpha)^2 \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}_S \mathbf{x}\|_2^2 \leq (1-\alpha)^{-2} \|x\|_2^2 \,.$$

for all $\mathbf{x} \in \mathbb{C}^S$.

The assumptions in this theorem are certainly weaker than the one of Theorem 13.1, in particular, no reference to the coherence or a similar quantity is made. But the statement is only about *existence* of a submatrix with controlled smallest singular and not about properties of *most* (that is, random) submatrices. Indeed, one cannot expect Theorem 13.1 to hold without any assumption on the coherence because a random submatrix of a matrix which consists of a duplicated orthonormal basis (hence, $\mu = 1$) will contain a duplicated column with high probability, so that the singular value will be zero.

Nevertheless, well-conditioned submatrices certainly exist in this case such as one copy of the orthonormal basis. Further information on the restricted invertibility theorem can be found, for instance, in [86, 422, 436].

There is also a relation of Theorem 13.1 with another theorem of Bourgain and Tzafriri [56]:

Theorem 13.9. *Let* $\mathbf{H} \in \mathbb{C}^{N \times N}$ *with* $\|\mathbf{H}\|_{2\to 2} \le 1$ *whose entries satisfy*

$$|H_{j,k}| \le \frac{1}{\ln^2 N} \, .$$

Select a subset $S \subset [N]$ *of size* cN *uniformly at random. Then* $\|\mathbf{P}_S \mathbf{H} \mathbf{P}_S\|_{2\to 2} \le 1/2$ *with probability at least* $1 - N^{-c}$.

This theorem clearly applies not only to matrices of the form $\mathbf{H} = \mathbf{A}^* \mathbf{A} - \mathbf{Id}$ (that is, with zero diagonal), but has a slightly stronger condition on the size of the matrix entries than (13.1).

Exercises

13.1. Verify (13.7) and (13.11) in detail.

13.2. Verify (13.3) in detail.

13.3. Verify (13.19) in detail.

14

Lossless Expanders in Compressive Sensing

In this chapter, we introduce another type of matrices that can be used when reconstructing a sparse vector from a limited number of measurements. They are adjacency matrices of certain bipartite graphs called lossless expanders. These objects are defined in Section 14.1, where some of their useful properties are established. In Section 14.2, we resort to combinatorial arguments to show that lossless expanders do indeed exist. Then, in Section 14.3, we prove that using their adjacency matrices as measurement matrices allow for a stable and robust reconstruction of sparse vectors via ℓ_1-minimization. One of the nice features of this approach is that the robust null space property can be proved directly in the ℓ_1-setting, without resorting to auxiliary tools such as restricted isometry properties. Section 14.4 shows the stability and robustness of a thresholding-based algorithm and, finally, Section 14.5 presents a simple sublinear-time algorithm.

14.1 Definitions and Basic Properties

Throughout this chapter, we consider *bipartite* graphs, i.e., graphs $G = (L, R, E)$ where each edge $e := \overline{j\,i} \in E$ connects a left vertex $j \in L$ with a right vertex $i \in R$. Removing vertices if necessary, we assume that every vertex is attached to an edge. The sets L and R are identified with $[N]$ and $[m]$, respectively, where $N := \mathrm{card}(L)$ and $m := \mathrm{card}(R)$. The *degree* of a left vertex is the number of right vertices it connects with. A bipartite graph is called *left regular* with degree d if all left vertices have the same degree d. For such *left d-regular* bipartite graphs, given a set $J \subseteq [N]$ of left vertices, the cardinality of the set

$$E(J) := \{\overline{j\,i} \in E \text{ with } j \in J\}$$

of all edges emanating from J is exactly

$$\mathrm{card}(E(J)) = d\ \mathrm{card}(J).$$

The set

$$R(J) = \{i \in R : \overline{j\,i} \in E \text{ with } j \in J\}$$

of right vertices connected to J satisfies

$$\operatorname{card}(R(J)) \le d \operatorname{card}(J).$$

Equality occurs if and only if no two edges emanating from J share a common right vertex. In the typical situation where the number N of left vertices is much larger than the number m of right vertices, such an equality cannot be met for large sets J. However, we shall see that an almost-equality can be met for small sets J. This almost-equality constitutes the expansion property, and left regular bipartite graphs with this property are called lossless expanders. The precise definition is given below. We stress the difference between this concept and the better-known concept of expanders which involves classical (unipartite) graphs — see the Notes section.

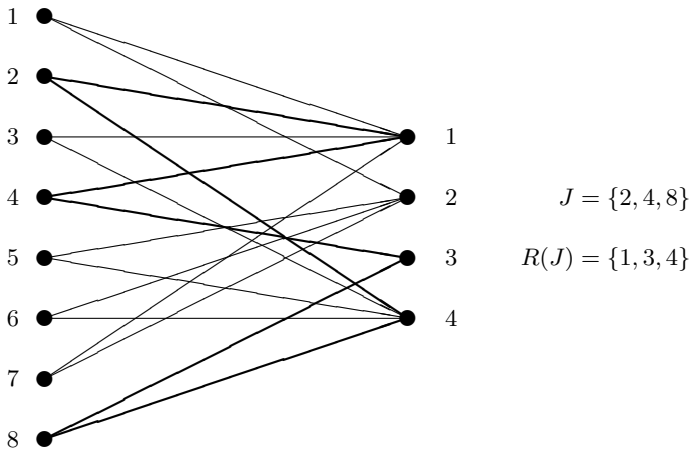

Fig. 14.1. A left regular bipartite graph with left degree two

Definition 14.1. *A left regular bipartite graph with left degree d is called a* (s, d, θ)-lossless expander *if it satisfies the expansion property*

$$\operatorname{card}(R(J)) \ge (1 - \theta)\, d \operatorname{card}(J) \tag{14.1}$$

for all sets J of left vertices such that $\operatorname{card}(J) \le s$. *The smallest $\theta \ge 0$ for which the expansion property holds is denoted by θ_s.*

It is readily seen that

$$0 = \theta_1 \le \theta_2 \le \cdots \le \theta_s \le \theta_{s+1} \le \cdots \le \theta_N.$$

It is also possible to compare the constants θ_t of higher order in terms of the constants θ_s of lower order, similarly to Proposition 6.6 for restricted isometry constants.

Proposition 14.2. *For integers $k, s \geq 1$,*

$$\theta_{ks} \leq (k-1)\theta_{2s} + \theta_s.$$

Proof. Let T be a set of left vertices satisfying $t := \mathrm{card}(T) \leq ks$. We partition T as $T = S_1 \cup \cdots \cup S_k$, where each S_ℓ satisfies $s_\ell := \mathrm{card}(S_\ell) \leq s$. We have

$$\begin{aligned}\mathrm{card}(R(T)) &= \mathrm{card}\Big(\bigcup_{1\leq \ell \leq k} R(S_\ell)\Big) \\ &\geq \sum_{1\leq \ell \leq k} \mathrm{card}(R(S_\ell)) - \sum_{1\leq \ell_1 < \ell_2 \leq k} \mathrm{card}(R(S_{\ell_1}) \cap R(S_{\ell_2})).\end{aligned}$$

In view of $\mathrm{card}(R(S_\ell)) \geq (1-\theta_s) d s_\ell$ and of

$$\begin{aligned}\mathrm{card}(R(S_{\ell_1}) \cap R(S_{\ell_2})) &= \mathrm{card}(R(S_{\ell_1})) + \mathrm{card}(R(S_{\ell_2})) - \mathrm{card}(R(S_{\ell_1}) \cup R(S_{\ell_2})) \\ &\leq d s_{\ell_1} + d s_{\ell_2} - (1-\theta_{2s}) d (s_{\ell_1} + s_{\ell_2}) = \theta_{2s} d (s_{\ell_1} + s_{\ell_2}),\end{aligned}$$

we then obtain

$$\begin{aligned}\mathrm{card}(R(T)) &\geq \sum_{1\leq \ell \leq k} (1-\theta_s) d s_\ell - \sum_{1\leq \ell_1 < \ell_2 \leq k} \theta_{2s} d (s_{\ell_1} + s_{\ell_2}) \\ &= (1-\theta_s) dt - \frac{\theta_{2s} d}{2} \Big(\sum_{1\leq \ell_1, \ell_2 \leq k} (s_{\ell_1} + s_{\ell_2}) - \sum_{1\leq \ell_1 \leq k} (s_{\ell_1} + s_{\ell_1}) \Big) \\ &= (1-\theta_s) dt - \frac{\theta_{2s} d}{2} \Big(\sum_{1\leq \ell_1 \leq k} (k s_{\ell_1} + t) - 2t \Big) \\ &= (1-\theta_s) dt - \frac{\theta_{2s} d}{2} (2kt - 2t) = (1 - \theta_s - (k-1)\theta_{2s}) dt.\end{aligned}$$

This shows that $\theta_{ks} \leq \theta_s + (k-1)\theta_{2s}$, as announced. □

We now formulate two lemmas and a corollary to be used in Sections 14.3 and 14.4. They all formalize the intuition that collisions at right vertices are rare in a lossless expander.

Lemma 14.3. *Given a left d-regular bipartite graph, if disjoint sets J and K of left vertices satisfy $\mathrm{card}(J) + \mathrm{card}(K) \leq s$, then the set*

$$E(K; J) := \{\overline{j\,i} \in E(K) \text{ with } i \in R(J)\}$$

is small in the sense that

$$\mathrm{card}(E(K; J)) \leq \theta_s \, d \, s.$$

Proof. We separate the set E_0 of edges emanating from $J \cup K$ into three distinct subsets:

- the set E_1 of edges emanating from J,
- the set E_2 of edges emanating from K and whose right vertices are not connected to any left vertex in J,
- the set E_3 of edges emanating from K and whose right vertices are also connected to left vertices in J.

We need to bound the cardinality of the set $E(K;J) = E_3$. In view of $\operatorname{card}(E_0) = d\operatorname{card}(J \cup K) = d(\operatorname{card}(J) + \operatorname{card}(K))$ and $\operatorname{card}(E_1) = d\operatorname{card}(J)$, we have

$$\operatorname{card}(E_3) = \operatorname{card}(E_0) - \operatorname{card}(E_1) - \operatorname{card}(E_2) = d\operatorname{card}(K) - \operatorname{card}(E_2). \quad (14.2)$$

We now observe that each right vertex $i \in R(K) \setminus R(J)$ gives rise to at least one edge emanating from K whose right vertex is not connected to any left vertex in J, so that

$$\operatorname{card}(E_2) \geq \operatorname{card}(R(K) \setminus R(J)) = \operatorname{card}(R(J \cup K)) - \operatorname{card}(R(J)).$$

We now take

$$\begin{aligned} \operatorname{card}(R(J)) &\leq d\operatorname{card}(J) \\ \operatorname{card}(R(J \cup K)) &\geq (1-\theta)d\operatorname{card}(J \cup K) = (1-\theta)\, d\, (\operatorname{card}(J) + \operatorname{card}(K)) \end{aligned}$$

into account to derive the inequality

$$\operatorname{card}(E_2) \geq (1-\theta)d\operatorname{card}(K) - \theta d\operatorname{card}(J). \quad (14.3)$$

Substituting (14.3) into (14.2), we conclude that

$$\operatorname{card}(E_3) \leq \theta\, d\, (\operatorname{card}(K) + \operatorname{card}(J)),$$

which is the desired result. □

Lemma 14.4. *For each right vertex i of a left d-regular bipartite graph, let $\ell(i)$ denote a fixed left vertex connected to i. If S is a set of size s, then*

$$E'(S) := \{\overrightarrow{j\,i} \in E(S) : j \neq \ell(i)\}$$

is small in the sense that

$$\operatorname{card}(E'(S)) \leq \theta_s\, d\, s.$$

Proof. The set $E(S)$ of edges emanating from S is partitioned as $E(S) = E'(S) \cup E''(S)$, where $E''(S) := \{\overrightarrow{\ell(i)\,i}, i \in R(S)\}$. Since $\operatorname{card}(E(S)) = d\, s$ and $\operatorname{card}(E''(S)) = \operatorname{card}(R(S)) \geq (1-\theta_s)\, d\, s$, we conclude that $\operatorname{card}(E'(S)) = \operatorname{card}(E(S)) - \operatorname{card}(E''(S)) \leq \theta_s\, d\, s$. □

Corollary 14.5. *Given a left d-regular bipartite graph, if S is a set of s left indices, then the set*

$$R_1(S) := \big\{ i \in R(S) : \text{there is a unique } j \in S \text{ with } \overline{j\,i} \in E \big\}$$

of right vertices connected to exactly one left vertex in S is large in the sense that

$$\operatorname{card}(R_1(S)) \geq (1 - 2\theta_s)\, d\, s.$$

Proof. Fixing a left vertex $\ell(i)$ for each right vertex i as in Lemma 14.4, any $i \in R_{\geq 2}(S) := R(S) \setminus R_1(S)$ gives rise to at least one edge in $E'(S)$. Thus, $\operatorname{card}(R_{\geq 2}(S) \leq \operatorname{card}(E'(S)) \leq \theta_s\ ds$, and consequently $\operatorname{card}(R_1(S)) = \operatorname{card}(R(S)) - \operatorname{card}(R_{\geq 2}(S)) \geq (1 - 2\theta_s)\, d\, s$. □

14.2 Existence of Lossless Expanders

In this section, we prove that lossless expanders with parameters relevant to compressive sensing do exist. As a matter of fact, we prove that most left regular bipartite graphs are lossless expanders, i.e., that random left regular bipartite graphs are, with high probability, lossless expanders.

Theorem 14.6. *For $0 < \epsilon < 1/2$, the proportion of (s, d, θ)-lossless expanders among all left d-regular bipartite graphs with N left vertices and m right vertices exceeds $1 - \epsilon$ provided that*

$$d = \left\lceil \frac{1}{\theta} \ln\Big(\frac{eN}{\epsilon s}\Big) \right\rceil,$$

$$m \geq c_\theta\, s \ln\Big(\frac{eN}{\epsilon s}\Big), \qquad\qquad c_\theta := \frac{2e^{2/\theta}}{\theta}.$$

Proof. Since each of the left vertices $j \in [N]$ connects to a set $R(j) \subseteq [m]$ of d right vertices, the total number of left d-regular bipartite graphs is

$$\binom{m}{d}^N.$$

Among these graphs, a graph fails to be an (s, d, θ)-lossless expander if there exists a set $J \subseteq [N]$ with $2 \leq j := \operatorname{card}(J) \leq s$ such that $\operatorname{card}(R(J)) < (1 - \theta)dj$, i.e.,

$$R(J) \subseteq I \quad \text{for some set } I \subseteq [m] \text{ with } \operatorname{card}(I) = r_j := \lceil (1 - \theta)dj \rceil - 1.$$

For fixed sets I and J, the number of left d-regular bipartite graphs satisfying the latter is

$$\binom{r_j}{d}^j \binom{m}{d}^{N-j}.$$

Taking the union over all possible sets I and J, we see that the number of left d-regular bipartite graphs that are not (s,d,θ)-lossless expanders is at most

$$\sum_{j=2}^{s} \binom{N}{j} \binom{m}{r_j} \binom{r_j}{d}^j \binom{m}{d}^{N-j}.$$

Therefore, the proportion of graphs that are not (s,d,θ)-lossless expanders among the left d-regular bipartite graphs is at most

$$p := \sum_{j=2}^{s} p_j, \qquad \text{where} \quad p_j := \binom{N}{j} \binom{m}{r_j} \left(\frac{\binom{r_j}{d}}{\binom{m}{d}} \right)^j.$$

Using the simple inequalities of Lemma C.5, namely

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k,$$

we obtain, for each $2 \le j \le s$,

$$p_j \le \left(\frac{eN}{j}\right)^j \left(\frac{em}{r_j}\right)^{r_j} \left(\frac{\left(\frac{er_j}{d}\right)^d}{\left(\frac{m}{d}\right)^d} \right)^j = \left(\frac{eN}{j}\right)^j e^{r_j + dj} \left(\frac{r_j}{m}\right)^{dj - r_j}.$$

We now observe that

$$r_j \le (1-\theta)\, d\, j \le d\, j, \qquad\qquad m \ge e^{2/\theta} \frac{2}{\theta} \ln\left(\frac{eN}{\epsilon s}\right) s \ge e^{2/\theta} d\, s.$$

Taking $j \le s$ into account, we derive

$$\begin{aligned} p_j &\le \left(\frac{eN}{j}\right)^j e^{(2-\theta)dj} \left(\frac{d\, j}{e^{2/\theta} d\, s}\right)^{dj - r_j} \le \left(\frac{eN}{j}\right)^j e^{(2-\theta)dj} \left(\frac{j}{e^{2/\theta} s}\right)^{\theta dj} \\ &= \left(\frac{eN}{j}\, e^{-\theta d} \left(\frac{j}{s}\right)^{\theta d} \right)^j \le \left(\frac{eN}{j}\, \frac{\epsilon s}{eN} \left(\frac{j}{s}\right)^{\theta d} \right)^j = \left(\epsilon \left(\frac{j}{s}\right)^{\theta d - 1} \right)^j \le \epsilon^j. \end{aligned}$$

It follows that

$$p = \sum_{j=2}^{s} p_j \le \sum_{j=2}^{s} \epsilon^j \le \sum_{j=2}^{\infty} \epsilon^j = \frac{\epsilon^2}{1-\epsilon} < \epsilon,$$

which is the desired result. □

To obtain a result where the targeted probability does not enter the number of measurements, one can simply make a specific choice for ϵ, e.g. $\epsilon = s/(eN)$.

Corollary 14.7. *A bipartite graph with N left vertices and m right vertices drawn at random among all left d-regular graphs, $d := \lceil 2\ln(eN/s)/\theta \rceil$, satisfies $\theta_s \le \theta$ with probability at least*

$$1 - \frac{s}{eN}$$

provided

$$m \ge \frac{4e^{2/\theta}}{\theta} \ln\Big(\frac{eN}{s}\Big).$$

Discarding the dependence on θ, Corollary 14.7 is optimal in the sense that the existence of a lossless expander forces the number m of right vertices to satisfy $m \ge c\, s \ln(eN/s)$ for some $c > 0$, as we shall see in Corollary 14.13.

14.3 Sparse Recovery via Basis Pursuit

In this section, we prove that lossless expanders provide suitable measurement matrices for basis pursuit. These matrices are the adjacency matrices of the bipartite graph, defined as follows.

Definition 14.8. *The* adjacency matrix *of a bipartite graph $G = ([N], [m], E)$ is the $m \times N$ matrix $\mathbf{A}$ with entries*

$$A_{i,j} = \begin{cases} 1 & \text{if } \overline{j\,i} \in E, \\ 0 & \text{if } \overline{j\,i} \notin E. \end{cases}$$

It is completely equivalent, and sometimes more appropriate, to think of a (s, d, θ)-lossless as a matrix $\mathbf{A}$ populated with zeros and ones, with d ones per column, and such that there are at least $(1-\theta)dk$ nonzero rows in any submatrix of $\mathbf{A}$ composed of $k \le s$ columns. Because of their zero-one structure, such matrices present some advantages over subgaussian random matrices, notably they require less storage space and they allow for faster computations. They also allow for stable and robust sparse recovery, as established below. As usual, perfect recovery is obtained in the particular case where the vector $\mathbf{x}$ is exactly s-sparse and the measurement error η equals zero.

Theorem 14.9. *Suppose that $\mathbf{A} \in \{0,1\}^{m \times N}$ is the adjacency matrix of a left d-regular bipartite graph satisfying*

$$\theta_{2s} < \frac{1}{6}.$$

For $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{e} \in \mathbb{C}^m$ with $\|\mathbf{e}\|_1 \le \eta$, if $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, then a solution $\mathbf{x}^\sharp$ of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_1 \le \eta,$$

approximates the vector $\mathbf{x}$ *with* ℓ_1*-error*

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \le \frac{2(1-2\theta)}{(1-6\theta)}\,\sigma_s(\mathbf{x})_1 + \frac{4}{(1-6\theta)d}\,\eta.$$

According to Theorem 4.18, this is a corollary of the following result.

Theorem 14.10. *The adjacency matrix* $\mathbf{A} \in \{0,1\}^{m\times N}$ *of a left* d*-regular bipartite graph satisfies the* ℓ_1*-robust null space property of order* s *provided* $\theta_{2s} < 1/6$*, precisely*

$$\|\mathbf{v}_S\|_1 \le \frac{2\theta_{2s}}{1-4\theta_{2s}}\,\|\mathbf{v}_{\overline{S}}\|_1 + \frac{1}{(1-4\theta_{2s})d}\,\|\mathbf{A}\mathbf{v}\|_1 \tag{14.4}$$

for all $\mathbf{v} \in \mathbb{C}^N$ *and all* $S \subseteq [N]$ *with* $\mathrm{card}(S) = s$*.*

We isolate the following two lemmas for the proof of Theorem 14.10.

Lemma 14.11. *Let* $\mathbf{A} \in \{0,1\}^{m\times N}$ *be the adjacency matrix of a left* d*-regular bipartite graph. If* S *and* T *are two disjoint subsets of* $[N]$ *and if* $\mathbf{x} \in \mathbb{C}^N$*, then*

$$\|(\mathbf{A}\mathbf{x}_S)_{R(T)}\|_1 \le \theta_{s+t}\,d\,(s+t)\,\|\mathbf{x}_S\|_\infty,$$

where $s = \mathrm{card}(S)$ *and* $t = \mathrm{card}(T)$*.*

Proof. We estimate the term $\|(\mathbf{A}\mathbf{x}_S)_{R(T)}\|_1$ as

$$\begin{aligned}
\|(\mathbf{A}\mathbf{x}_S)_{R(T)}\|_1 &= \sum_{i\in R(T)} |(\mathbf{A}\mathbf{x}_S)_i| = \sum_{i=1}^m \mathbf{1}_{\{i\in R(T)\}} \Big| \sum_{j\in S} A_{i,j} x_j \Big| \\
&\le \sum_{i=1}^m \mathbf{1}_{\{i\in R(T)\}} \sum_{j\in S} \mathbf{1}_{\{\overline{j\,i}\in E\}} |x_j| \\
&= \sum_{j\in S}\sum_{i=1}^m \mathbf{1}_{\{i\in R(T) \text{ and } \overline{j\,i}\in E\}} |x_j| = \sum_{\overline{j\,i}\in E(S;T)} |x_j| \\
&\le \mathrm{card}(E(S;T))\,\|\mathbf{x}_S\|_\infty.
\end{aligned}$$

The conclusion follows from the bound on $\mathrm{card}(E(S;T))$ of Lemma 14.3. □

Lemma 14.12. *Let* $\mathbf{A} \in \{0,1\}^{m\times N}$ *be the adjacency matrix of a left* d*-regular bipartite graph. Given an* s*-sparse vector* $\mathbf{w} \in \mathbb{C}^N$*, let* $\mathbf{w}' \in \mathbb{C}^m$ *be defined by* $w'_i := w_{\ell(i)}$*,* $i \in [m]$*, where*

$$\ell(i) := \mathrm{argmax}\{|w_j|, \overline{j\,i} \in E\}.$$

Then

$$\|\mathbf{A}\mathbf{w} - \mathbf{w}'\|_1 \le \theta_s\,d\,\|\mathbf{w}\|_1.$$

Proof. We may and do assume that the left vertices are ordered so that

$$|w_1| \geq |w_2| \geq \cdots \geq |w_s| \geq |w_{s+1}| = \cdots = |w_N| = 0.$$

In this way, the edge $\overline{\ell(i)\, i}$ can be thought of as the first edge arriving at the right vertex i. Since the vector $\mathbf{w} \in \mathbb{C}^N$ is supported on $S := [s]$, and since $\ell(i) \in S$ whenever $i \in R(S)$, we have

$$(\mathbf{A}\mathbf{w} - \mathbf{w}')_i = \sum_{j=1}^{N} A_{i,j} w_j - w_{\ell(i)} = \sum_{j \in S} \mathbf{1}_{\{\overline{j\, i} \in E \text{ and } j \neq \ell(i)\}} w_j.$$

Thus, we obtain

$$\begin{aligned} \|\mathbf{A}\mathbf{w} - \mathbf{w}'\|_1 &= \sum_{i=1}^{m} \left| \sum_{j \in S} \mathbf{1}_{\{\overline{j\, i} \in E \text{ and } j \neq \ell(i)\}} w_j \right| \leq \sum_{i=1}^{m} \sum_{j \in S} \mathbf{1}_{\{\overline{j\, i} \in E \text{ and } j \neq \ell(i)\}} |w_j| \\ &\leq \sum_{j \in S} \left(\sum_{i=1}^{m} \mathbf{1}_{\{\overline{j\, i} \in E \text{ and } j \neq \ell(i)\}} \right) |w_j| = \sum_{j=1}^{s} c_j |w_j|, \end{aligned}$$

where $c_j := \sum_{i=1}^{m} \mathbf{1}_{\{\overline{j\, i} \in E \text{ and } j \neq \ell(i)\}}$. For all $k \in [s]$, we observe that

$$\begin{aligned} C_k := \sum_{j=1}^{k} c_j &= \sum_{j=1}^{k} \sum_{i=1}^{m} \mathbf{1}_{\{\overline{j\, i} \in E \text{ and } j \neq \ell(i)\}} = \operatorname{card}(\{\overline{j\, i} \in E([k]), j \neq \ell(i)\}) \\ &\leq \theta_s\, d\, k, \end{aligned} \tag{14.5}$$

where the last inequality was derived from Lemma 14.4. Setting $C_0 = 0$ and performing a *summation by parts*, we have

$$\begin{aligned} \sum_{j=1}^{s} c_j\, |w_j| &= \sum_{j=1}^{s} (C_j - C_{j-1}) |w_j| = \sum_{j=1}^{s} C_j |w_j| - \sum_{j=1}^{s} C_{j-1} |w_j| \\ &= \sum_{j=1}^{s} C_j |w_j| - \sum_{j=0}^{s-1} C_j |w_{j+1}| = \sum_{j=1}^{s-1} C_j (|w_j| - |w_{j+1}|) + C_s |w_s|. \end{aligned}$$

Since $|w_j| - |w_{j+1}| \geq 0$, the bound (14.5) yields

$$\sum_{j=1}^{s} c_j\, |w_j| \leq \sum_{j=0}^{s-1} \theta_s\, d\, j (|w_j| - |w_{j+1}|) + \theta_s\, d\, s\, |w_s| = \sum_{j=1}^{s} \theta_s\, d\, |w_j|, \tag{14.6}$$

where the last equality was derived by reversing the summation by parts process after replacing c_j by $\theta_s d$. The result is proved. □

We are know ready to prove the key result of this section.

Proof (of Theorem 14.10). Let $\mathbf{v} \in \mathbb{C}^N$ be a fixed vector, and let S_0 be an index set of s largest absolute entries of $\mathbf{v}$, S_1 an index set of next s largest absolute entries, etc. It is enough to establish (14.4) for $S = S_0$. We start by writing

$$\begin{aligned} d\,\|\mathbf{v}_{S_0}\|_1 = d \sum_{j \in S_0} |v_j| = \sum_{\overline{j\,i} \in E(S_0)} |v_j| = \sum_{i \in R(S_0)} \sum_{\substack{j \in S_0 \\ \overline{j\,i} \in E}} |v_j| \\ = \sum_{i \in R(S_0)} |v_{\ell(i)}| + \sum_{i \in R(S_0)} \sum_{\substack{j \in S_0 \setminus \{\ell(i)\} \\ \overline{j\,i} \in E}} |v_j|, \end{aligned} \tag{14.7}$$

where the notation of Lemma 14.12 has been used. We now observe that, for $i \in R(S_0)$,

$$(A\mathbf{v})_i = \sum_{j \in [N]} A_{i,j} v_j = \sum_{\substack{j \in [N] \\ \overline{j\,i} \in E}} v_j = \sum_{k \ge 0} \sum_{\substack{j \in S_k \\ \overline{j\,i} \in E}} v_j = v_{\ell(i)} + \sum_{\substack{j \in S_0 \setminus \{\ell(i)\} \\ \overline{j\,i} \in E}} v_j + \sum_{k \ge 1} \sum_{\substack{j \in S_k \\ \overline{j\,i} \in E}} v_j.$$

It follows that

$$|v_{\ell(i)}| \le \sum_{\substack{j \in S_0 \setminus \{\ell(i)\} \\ \overline{j\,i} \in E}} |v_j| + \sum_{k \ge 1} \sum_{\substack{j \in S_k \\ \overline{j\,i} \in E}} |v_j| + |(A\mathbf{v})_i|.$$

Summing over all $i \in R(S_0)$ and substituting into (14.7), we obtain

$$d\,\|\mathbf{v}_{S_0}\|_1 \le 2 \sum_{i \in R(S_0)} \sum_{\substack{j \in S_0 \setminus \{\ell(i)\} \\ \overline{j\,i} \in E}} |v_j| + \sum_{k \ge 1} \sum_{i \in R(S_0)} \sum_{\substack{j \in S_k \\ \overline{j\,i} \in E}} |v_j| + \|\mathbf{A}\mathbf{v}\|_1. \tag{14.8}$$

For the first term in the right-hand side of (14.8), we apply Lemma 14.12 to $\mathbf{w} = |\mathbf{v}_{S_0}|$ (i.e., $w_j = v_j$ if $j \in S_0$ and $w_j = 0$ otherwise) to obtain

$$\sum_{i \in R(S_0)} \sum_{\substack{j \in S_0 \setminus \{\ell(i)\} \\ \overline{j\,i} \in E}} |v_j| = \|\mathbf{A}\mathbf{w} - \mathbf{w}'\|_1 \le \theta_s\, d\,\|\mathbf{w}\|_1 = \theta_s\, d\,\|\mathbf{v}_{S_0}\|_1. \tag{14.9}$$

For the second term in the right-hand side of (14.8), we apply Lemma 14.11 to obtain

$$\begin{aligned} \sum_{k \ge 1} \sum_{i \in R(S_0)} \sum_{\substack{j \in S_k \\ \overline{j\,i} \in E}} |v_j| = \sum_{k \ge 1} \|(\mathbf{A}\mathbf{v}_{S_k})_{R(S_0)}\|_1 \le \sum_{k \ge 1} \theta_{2s}\, d\, 2s\, \|\mathbf{v}_{S_k}\|_\infty \\ \le 2\theta_{2s}\, d \sum_{k \ge 1} \|\mathbf{v}_{S_{k-1}}\|_1 \le 2\theta_{2s}\, d\,\|\mathbf{v}\|_1. \end{aligned} \tag{14.10}$$

Finally, substituting (14.9) and (14.10) into (14.8), we deduce

$$\begin{aligned} d\,\|\mathbf{v}_{S_0}\|_1 &\le 2\,\theta_s\,d\,\|\mathbf{v}_{S_0}\|_1 + 2\,\theta_{2s}\,d\,\|\mathbf{v}\|_1 + \|\mathbf{A}\mathbf{v}\|_1 \\ &= 4\,\theta_{2s}\,d\,\|\mathbf{v}_{S_0}\|_1 + 2\,\theta_{2s}\,d\,\|\mathbf{v}_{\overline{S_0}}\|_1 + \|\mathbf{A}\mathbf{v}\|_1. \end{aligned}$$

Rearranging the latter leads to the desired inequality (14.4). □

To close this section, we highlight that the exact s-sparse recovery by basis pursuit using lossless expanders provides, in retrospect, a lower bound for the number of right vertices in a lossless expander.

Corollary 14.13. *For $s \ge 2$ and $\theta < 1/25$, an (s, d, θ)-lossless expander with N left vertices must have a number m of right vertices bounded below by*

$$m \ge \frac{c_1}{\theta}\, s\, \ln\Big(\frac{c_2 \theta N}{s}\Big)$$

for some absolute constants $c_1, c_2 > 0$.

Proof. Let us consider $k := \lfloor 1/(25\theta) \rfloor \ge 1$ and $s' := \lfloor s/2 \rfloor \ge 1$. According to Proposition 14.2, we have

$$\theta_{4ks'} \le 4k\,\theta_{2s'} \le 4k\,\theta \le \frac{4}{25} < \frac{1}{6}.$$

Therefore, Theorem 14.9 implies that every $2ks'$ sparse vector $\mathbf{x} \in \mathbb{R}^N$ is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$ via ℓ_1-minimization. Theorem 10.11 then implies that, with $c = 1/\ln 9$,

$$m \ge c\,k\,s' \ln\Big(\frac{N}{4ks'}\Big).$$

In view of $1/(50\theta) \le k \le 1/(25\theta)$ and of $2s/3 \le s' \le s/2$, we conclude that

$$m \ge \frac{c}{75\,\theta}\, s \ln\Big(\frac{25\theta N}{2s}\Big),$$

which is the desired result with $c_1 = c/75$ and $c_2 = 25/2$. □

14.4 Sparse Recovery via an Iterative Thresholding Algorithm

In this section, we prove that lossless expanders provide suitable measurement matrices for other algorithms besides basis pursuit. First, in the real setting, we consider a variation of the iterative hard thresholding algorithm. Precisely, starting with an initial s-sparse vector $\mathbf{x}^0 \in \mathbb{R}^N$, typically $\mathbf{x}^0 = 0$, we iterate the scheme

$$\mathbf{x}^{n+1} = H_s(\mathbf{x}^n + \mathcal{M}(\mathbf{y} - \mathbf{A}\mathbf{x}^n)). \tag{14.11}$$

The nonlinear operator $\mathcal{M} = \mathcal{M}_{\mathbf{A}}$ is the *median operator*, which is defined componentwise by

$$(\mathcal{M}(\mathbf{z}))_j := \text{median}[z_i, i \in R(j)] \quad \text{for } \mathbf{z} \in \mathbb{C}^m \text{ and } j \in [N].$$

Here, $R(j) = R(\{j\})$ denotes the set of right vertices connected to j, and the median of the d numbers z_i, $i \in R(j)$, is defined to be the $\lceil d/2 \rceil$th largest of these numbers. The properties of the algorithm (14.11) are very similar to the properties established in Section 6.3 for the iterative hard thresholding algorithm.

Theorem 14.14. *Suppose that the adjacency matrix $\mathbf{A} \in \{0,1\}^{m \times N}$ of a left d-regular bipartite graph satisfies*

$$\theta_{3s} < \frac{1}{12}.$$

Then, for $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{e} \in \mathbb{R}^m$, and $S \subseteq [N]$ with $\text{card}(S) = s$, the sequence $(\mathbf{x}^n)$ defined by (14.11) with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ satisfies, for any $n \geq 0$,

$$\|\mathbf{x}^n - \mathbf{x}_S\|_1 \leq \rho^n \|\mathbf{x}^0 - \mathbf{x}_S\|_1 + \frac{\tau}{d} \|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_1, \tag{14.12}$$

where $\rho < 1$ and τ depend only on θ_{3s}. In particular, if the sequence $(\mathbf{x}^n)$ clusters around some $\mathbf{x}^\sharp \in \mathbb{R}^N$, then

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \leq C\sigma_s(\mathbf{x})_1 + \frac{D}{d}\|\mathbf{e}\|_1$$

for some constants $C, D > 0$ depending only on θ_{3s}.

The proof relies on the fact that the median operator approximately inverts the action of $\mathbf{A}$ on sparse vectors. We state this as a lemma involving the slightly more general *quantile operators* $\mathcal{Q}_k$ in place of $\mathcal{M} = \mathcal{Q}_{\lceil d/2 \rceil}$. It is defined componentwise by

$$(\mathcal{Q}_k(\mathbf{z}))_j := q_k[z_i, i \in R(j)] \quad \text{for } \mathbf{z} \in \mathbb{C}^m \text{ and } j \in [N],$$

where the quantile q_k denotes the kth largest element, i.e.,

$$q_k[a_1, \ldots, a_d] = a_{\pi(k)}$$

if $\pi : [d] \to [d]$ is a permutation for which $a_{\pi(1)} \geq a_{\pi(2)} \geq \cdots \geq a_{\pi(d)}$. We will use the following observations, to be established in Exercise 14.9,

$$|q_k[a_1, \ldots, a_d]| \leq q_k[|a_1|, \ldots, |a_d|], \qquad \text{if } 2k \leq d+1, \tag{14.13}$$

$$q_k[b_1, \ldots, b_d] \leq \frac{b_1 + \cdots + b_d}{k} \qquad \text{if } b_j \geq 0 \text{ for all } j. \tag{14.14}$$

Lemma 14.15. *Let $\mathbf{A} \in \{0,1\}^{m \times N}$ be the adjacency matrix of a left d-regular bipartite graph and let k be an integer satisfying $2\theta_s d < k \leq (d+1)/2$. If S is a subset of $[N]$ with size s, then*

$$\|\big(\mathcal{Q}_k(\mathbf{A}\mathbf{x}_S + \mathbf{e}) - \mathbf{x}\big)_S\|_1 \leq \frac{2\theta_s d}{k - 2\theta_s d}\|\mathbf{x}_S\|_1 + \frac{1}{k - 2\theta_s d}\|\mathbf{e}_{R(S)}\|_1 \tag{14.15}$$

for all $\mathbf{x} \in \mathbb{R}^N$ and all $\mathbf{e} \in \mathbb{R}^m$.

Proof. According to the definition of $\mathcal{Q}_k$ and to (14.13), we have

$$\begin{aligned}\|\big(\mathcal{Q}_k(\mathbf{A}\mathbf{x}_S+\mathbf{e})-\mathbf{x}\big)_S\|_1 &= \sum_{j\in S}\big|q_k\big[(\mathbf{A}\mathbf{x}_S+\mathbf{e})_i, i\in R(j)\big]-x_j\big| \\ &= \sum_{j\in S}\big|q_k\big[(\mathbf{A}\mathbf{x}_S)_i+e_i-x_j, i\in R(j)\big]\big| \\ &\le \sum_{j\in S} q_k\big[|(\mathbf{A}\mathbf{x}_S)_i-x_j+e_i|, i\in R(j)\big] \\ &= \sum_{j\in S} q_k\Big[\Big|\sum_{\substack{\ell\in S\setminus\{j\}\\ \overline{\ell\, i}\in E}} x_\ell+e_i\Big|, i\in R(j)\Big].\end{aligned}$$

We now proceed by induction on $s=\mathrm{card}(S)$ to show that the latter is bounded above by the right-hand side of (14.15). If $s=1$, i.e., if $S=\{j\}$ for some $j\in S$ so that there is no $\ell\in S\setminus\{j\}$, we have the stronger estimate

$$q_k\Big[\Big|\sum_{\substack{\ell\in S\setminus\{j\}\\ \overline{\ell\, i}\in E}} x_\ell+e_i\Big|, i\in R(j)\Big] = q_k[|e_i|, i\in R(j)] \le \frac{1}{k}\|\mathbf{e}_{R(j)}\|_1,$$

where we have used (14.14). Let us now assume that the induction hypothesis holds up to $s-1$ for some $s\ge 2$, and let us show that it holds for s, too. For $S\subseteq[N]$ with $\mathrm{card}(S)=s$ and for $j\in S$, we introduce the set

$$R_1(j,S) := R(j)\setminus \bigcup_{\ell\in S\setminus\{j\}} R(\ell)$$

of right vertices connected only to j in S. We recall from Corollary 14.5 that

$$\sum_{j\in S}\mathrm{card}(R_1(j,S)) = \mathrm{card}(R_1(S)) \ge (1-2\theta_s)d\,s. \tag{14.16}$$

Thus, there exists $j^*\in S$ such that $r:=\mathrm{card}(R_1(j^*,S))\ge(1-2\theta_s)d$. This means that there are at most $d-r\le 2\theta_s d$ right vertices in $R(j^*)\setminus R_1(j^*,S)$. By definition of q_k, there exist k distinct $i_1,\dots,i_k\in R(j^*)$ such that, for all $h\in[k]$,

$$q_k\Big[\Big|\sum_{\substack{\ell\in S\setminus\{j^*\}\\ \overline{\ell\, i}\in E}} x_\ell+e_i\Big|, i\in R(j^*)\Big] \le \Big|\sum_{\substack{\ell\in S\setminus\{j^*\}\\ \overline{\ell\, i_h}\in E}} x_\ell+e_{i_h}\Big|. \tag{14.17}$$

At least $k':=k-(d-r)\ge k-2\theta_s d$ elements among $i_1,\dots,i_k$ are in $R_1(j^*,S)$. Averaging (14.17) over these elements i_h, keeping in mind that there are no $\ell\in S\setminus\{j^*\}$ with $\overline{\ell\, i_h}\in E$ in this case, we obtain

$$q_k\Big[\Big|\sum_{\substack{\ell\in S\setminus\{j^*\}\\ \overline{\ell\, i}\in E}} x_\ell+e_i\Big|, i\in R(j^*)\Big] \le \frac{1}{k-2\theta_s d}\|\mathbf{e}_{R_1(j^*,S)}\|_1. \tag{14.18}$$

On the other hand, if $T := S \setminus \{j^*\}$ and if $j \in T$, we have

$$\Big| \sum_{\substack{\ell \in S \setminus \{j\} \\ \overline{\ell\, i} \in E}} x_\ell + e_i \Big| = \Big| \sum_{\substack{\ell \in T \setminus \{j\} \\ \overline{\ell\, i} \in E}} x_\ell + \mathbf{1}_{\{\overline{j^*\, i} \in E\}} x_{j^*} + e_i \Big|.$$

Applying the induction hypothesis with S replaced by T and e_i replaced by $e_i' = \mathbf{1}_{\{\overline{j^*\, i} \in E\}} x_{j^*} + e_i$ gives, in view of $\theta_{s-1} \le \theta_s$,

$$\sum_{j \in T} q_k \Big[\Big| \sum_{\substack{\ell \in S \setminus \{j\} \\ \overline{\ell\, i} \in E}} x_\ell + e_i \Big|, i \in R(j) \Big] \le \frac{2\theta_s d}{k - 2\theta_s d} \|\mathbf{x}_T\|_1 + \frac{1}{k - 2\theta_s d} \|\mathbf{e}'_{R(T)}\|_1. \tag{14.19}$$

In order to bound $\|\mathbf{e}'_{R(T)}\|_1$, we observe that

$$\begin{aligned} \sum_{i \in R(T)} \mathbf{1}_{\{\overline{j^*\, i} \in E\}} &= \sum_{i=1}^{m} \mathbf{1}_{\{\overline{j^*\, i} \in E \text{ and } \overline{j\, i} \in E \text{ for some } j \in T\}} \\ &= \sum_{i \in R(j^*)} \mathbf{1}_{\{\overline{j\, i} \in E \text{ for some } j \in T\}} = \operatorname{card}(R(j^*) \setminus R_1(j^*, S)) \le 2\theta_s d, \end{aligned}$$

which allows to derive

$$\|\mathbf{e}'_{R(T)}\|_1 \le \sum_{i \in R(T)} \mathbf{1}_{\{\overline{j^*\, i} \in E\}} |x_{j^*}| + \|\mathbf{e}_{R(T)}\|_1 \le 2\theta_s d\, |x_{j^*}| + \|\mathbf{e}_{R(T)}\|_1.$$

Taking this bound into account in (14.19) and summing with (14.18) gives,

$$\begin{aligned} &\sum_{j \in S} q_k \Big[\Big| \sum_{\substack{\ell \in S \setminus \{j\} \\ \overline{\ell\, i} \in E}} x_\ell + e_i \Big|, i \in R(j) \Big] \\ &= \sum_{j \in T} q_k \Big[\Big| \sum_{\substack{\ell \in S \setminus \{j\} \\ \overline{\ell\, i} \in E}} x_\ell + e_i \Big|, i \in R(j) \Big] + q_k \Big[\Big| \sum_{\substack{\ell \in S \setminus \{j^*\} \\ \overline{\ell\, i} \in E}} x_\ell + e_i \Big|, i \in R(j^*) \Big] \\ &\le \frac{2\theta_s d}{k - 2\theta_s d} \|\mathbf{x}_T\|_1 + \frac{1}{k - 2\theta_s d} \big(2\theta_s d\, |x_{j^*}| + \|\mathbf{e}_{R(T)}\|_1 \big) + \frac{1}{k - 2\theta_s d} \|\mathbf{e}_{R_1(j^*, S)}\|_1 \\ &\le \frac{2\theta_s d}{k - 2\theta_s d} \|\mathbf{x}_S\|_1 + \frac{1}{k - 2\theta_s d} \|\mathbf{e}_{R(S)}\|_1, \end{aligned}$$

where the fact that $R_1(j^*, S)$ and $R(T)$ are disjoint subsets of $R(S)$ was used in the last inequality. This concludes the inductive proof. □

Proof (of Theorem 14.14). We are going to prove that, for any $n \ge 0$,

$$\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_1 \le \rho \|\mathbf{x}^n - \mathbf{x}_S\|_1 + \frac{(1-\rho)\tau}{d} \|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_1. \tag{14.20}$$

We use the triangle inequality and the fact that $\mathbf{x}^{n+1}$ is a better s-term approximation than $\mathbf{x}_S$ to $\mathbf{u}^{n+1} := (\mathbf{x}^n + \mathcal{M}(\mathbf{y} - \mathbf{A}\mathbf{x}^n))_{T^{n+1}}$, where $T^{n+1} := S \cup \mathrm{supp}(\mathbf{x}^n) \cup \mathrm{supp}(\mathbf{x}^{n+1})$, to derive that

$$\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_1 \le \|\mathbf{x}^{n+1} - \mathbf{u}^{n+1}\|_1 + \|\mathbf{x}_S - \mathbf{u}^{n+1}\|_1 \le 2\|\mathbf{x}_S - \mathbf{u}^{n+1}\|_1.$$

Since $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{A}\mathbf{x}_S + \mathbf{e}'$ with $\mathbf{e}' := \mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}$, Lemma 14.15 implies that

$$\begin{aligned}\|\mathbf{x}^{n+1} - \mathbf{x}_S\|_1 &\le 2\|(\mathbf{x}_S - \mathbf{x}^n - \mathcal{M}(\mathbf{A}(\mathbf{x}_S - \mathbf{x}^n) + \mathbf{e}'))_{T^{n+1}}\|_1 \\ &\le \frac{4\theta_{3s}d}{\lceil d/2\rceil - 2\theta_{3s}d}\|\mathbf{x}_S - \mathbf{x}^n\|_1 + \frac{2}{\lceil d/2\rceil - 2\theta_{3s}d}\|\mathbf{e}'\|_1 \\ &\le \frac{8\theta_{3s}}{1-4\theta_{3s}}\|\mathbf{x}_S - \mathbf{x}^n\|_1 + \frac{4}{(1-4\theta_{3s})d}\|\mathbf{e}'\|_1.\end{aligned}$$

This is the desired inequality (14.20) with $\rho := 8\theta_{3s}/(1-4\theta_{3s}) < 1$ and $\tau := 4/(1-12\theta_{3s})$. The estimate (14.12) follows by immediate induction. Next, if $\mathbf{x}^\sharp$ is a cluster point of the sequence $(\mathbf{x}^n)_{n\ge 0}$, we deduce

$$\|\mathbf{x}^\sharp - \mathbf{x}_S\|_1 \le \frac{\tau}{d}\|\mathbf{A}\mathbf{x}_{\overline{S}} + \mathbf{e}\|_1 \le \frac{\tau}{d}\|\mathbf{A}\mathbf{x}_{\overline{S}}\|_1 + \frac{\tau}{d}\|\mathbf{e}\|_1,$$

where we choose S as an index set of s largest absolute entries of $\mathbf{x}$. In view of the inequality

$$\|\mathbf{A}\mathbf{v}\|_1 = \sum_{i=1}^m \left|\sum_{j=1}^N a_{i,j}v_j\right| \le \sum_{j=1}^N \sum_{i=1}^m a_{i,j}|v_j| = \sum_{j=1}^N d|v_j| = d\|\mathbf{v}\|_1$$

applied to $\mathbf{v} = \mathbf{x}_{\overline{S}}$, it follows that

$$\|\mathbf{x}^\sharp - \mathbf{x}\|_1 \le \|\mathbf{x}_{\overline{S}}\|_1 + \|\mathbf{x}^\sharp - \mathbf{x}_S\|_1 \le (1+\tau)\sigma_s(\mathbf{x})_1 + \frac{\tau}{d}\|\mathbf{e}\|_1.$$

This is the desired estimate with $C = 1 + \tau$ and $D = \tau$. □

14.5 Sparse Recovery via a Simple Sublinear Algorithm

The relative simplicity of the algorithm of Section 14.4 is counterbalanced by the nonoptimality of its runtime. Indeed, the dimension N enters at least linearly when forming $\mathbf{x}^n + \mathcal{M}(\mathbf{y} - \mathbf{A}\mathbf{x}^n)$. One aims, however, at exploiting some features of the measurement matrix in order to devise algorithms with a smaller runtime than linear in N, for instance polylogarithmic in N and polynomial in s. Such algorithms are called *sublinear-time algorithms*. This section illustrates that sublinear-time algorithms are indeed possible, although the most sophisticated ones are not presented. As a first indication of this possibility, we consider the special case of 1-sparse vectors. We introduce the *bit-tester* matrix $\mathbf{B} \in \{0,1\}^{\ell \times N}$, $\ell := \lceil \log_2(N)\rceil$, defined by

$$\mathbf{B} = \begin{bmatrix} b_1(1) & \cdots & b_1(N) \\ \vdots & \ddots & \vdots \\ b_\ell(1) & \cdots & b_\ell(N) \end{bmatrix},$$

where $b_i(j) \in \{0,1\}$ denotes the ith digit in the binary expansion of $j-1$, i.e.,

$$j - 1 = b_\ell(j)2^{\ell-1} + b_{\ell-1}(j)2^{\ell-2} + \cdots + b_2(j)2 + b_1(j). \tag{14.21}$$

If the support of $\mathbf{x} \in \mathbb{C}^N$ is a singleton $\{j\}$, then the value of j is deduced from the measurement $\mathbf{Bx} = [b_1(j), \ldots, b_\ell(j)]^\top$ via (14.21). Moreover, if we append a row of ones after the last row of $\mathbf{B}$ to form the augmented bit-tester matrix

$$\mathbf{B}' = \begin{bmatrix} b_1(1) & \cdots & b_1(N) \\ \vdots & \ddots & \vdots \\ b_\ell(1) & \cdots & b_\ell(N) \\ 1 & \cdots & 1 \end{bmatrix},$$

then the measurement vector $\mathbf{B}'\mathbf{x} = [b_1(j), \ldots, b_\ell(j), x_j]^\top$ allows to determine both j and x_j using only a number of algebraic operations roughly proportional to $\log_2(N)$. This simple strategy can be extended from 1-sparse vectors to s-sparse vectors with $s > 1$ using lossless expanders. Precisely, given a matrix $\mathbf{A} \in \{0,1\}^{m \times N}$ with d ones per columns, we first construct a matrix $\mathbf{A}' \in \{0,1\}^{m' \times N}$ whose $m' = m(\ell+1)$ rows are all the pointwise products of rows of $\mathbf{A}$ with rows of $\mathbf{B}'$, precisely

$$A'_{(i-1)(\ell+1)+k,j} = B'_{k,j} A_{i,j}, \qquad i \in [m],\ k \in [\ell+1],\ j \in [N]. \tag{14.22}$$

Next, given $\mathbf{y} \in \mathbb{C}^m$, we construct a sequence of vectors $(\mathbf{x}^n)$ starting with $\mathbf{x}^0 = 0$ and iterating the instructions

- for all $i \in [m]$ satisfying $v_i := (\mathbf{y} - \mathbf{A}'\mathbf{x}^n)_{i(\ell+1)} \neq 0$, compute the integer

$$j_i := 1 + \frac{1}{v_i} \sum_{k=1}^{\ell} (\mathbf{y} - \mathbf{A}'\mathbf{x}^n)_{(i-1)(\ell+1)+k}\, 2^{k-1},$$

- if there are $r \geq d/2$ distinct right vertices $i_1, \cdots, i_r \in [m]$ such that $(j_{i_1}, v_{i_1}) = \cdots = (j_{i_r}, v_{i_r}) =: (j, v)$, set

$$x_j^{n+1} = x_j^n + v.$$

The procedure stops when $v_i = 0$ for all $i \in [m]$, i.e., when $\mathbf{A}'\mathbf{x}^n = \mathbf{y}$. If $\mathbf{A}$ is the adjacency matrix of a lossless expander and, neglecting stability and robustness issues, if $\mathbf{y} = \mathbf{A}'\mathbf{x}$ for some exactly s-sparse $\mathbf{x} \in \mathbb{C}^N$, then each j_i is an integer, each v_i is accurate, and the algorithm subsequently recovers the vector $\mathbf{x}$ in a finite number of iterations. The number of measurements approaches the optimal value $c\, s \log_2(N/s)$ up to the logarithmic factor $\log_2(N)$.

Theorem 14.16. *If $m' \geq c\, s \log_2(N/s) \log_2(N)$, then there is a measurement matrix $\mathbf{A}' \in \{0,1\}^{m' \times N}$ such that the procedure described above reconstructs every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}'\mathbf{x}$ with a number of algebraic operations at most proportional to $s^2 \log_2(N) \log_2(N/s) \log_2(s)$.*

Proof. Let $\mathbf{A} \in \{0,1\}^{m \times N}$ be the adjacency matrix of a left regular bipartite graph satisfying $\theta_s < 1/16$, and let d denotes its left degree. According to Theorem 14.6, such a matrix exists provided $m \asymp s \log_2(N/s)$. Let $\mathbf{A}' \in \{0,1\}^{m \times N}$ be the matrix constructed in (14.22). Its number of rows satisfies $m' = m(\ell+1) \asymp s \log_2(N/s) \log_2(N)$. We claim that, if $(\mathbf{x}^n)$ is the sequence produced by the algorithm described above and if $S^n := \operatorname{supp}(\mathbf{x} - \mathbf{x}^n)$, then we have $\operatorname{card}(S^{n+1}) < \operatorname{card}(S^n)/2$, so that $\mathbf{x}^{\bar{n}} = \mathbf{x}$ when $\bar{n} = \lceil \log_2(s) \rceil$. To justify this claim, we observe that elements $i \notin R(S^n)$ do not produce any change from $\mathbf{x}^n$ to $\mathbf{x}^{n+1}$, since

$$v_i = (\mathbf{A}'(\mathbf{x} - \mathbf{x}^n))_{i(\ell+1)} = \sum_{j \in S^n} A'_{i(l+1),j} (\mathbf{x} - \mathbf{x}^n)_j = \sum_{j \in S^n} A_{i,j} (\mathbf{x} - \mathbf{x}^n)_j = 0.$$

Next, we prove that elements $i \notin R_1(S^n) = \cup_{j \in S^n} R_1(j, S^n)$ create many zero entries in $\mathbf{x} - \mathbf{x}^{n+1}$. Indeed, let $i \in R_1(j^*, S^n)$ for some $j^* \in S^n$, i.e., the right vertex i is connected only to the left vertex j^* in S^n. We have, for any $k \in [\ell+1]$,

$$\begin{aligned}(\mathbf{y} - \mathbf{A}'\mathbf{x}^n)_{(i-1)(\ell+1)+k} &= (\mathbf{A}'(\mathbf{x} - \mathbf{x}^n))_{(i-1)(\ell+1)+k} \\ &= \sum_{j \in S^n} A'_{(i-1)(\ell+1)+k,j} (\mathbf{x} - \mathbf{x}^n)_j = \sum_{j \in S^n} B_{k,j} A_{i,j} (\mathbf{x} - \mathbf{x}^n)_j \\ &= B_{k,j^*} (\mathbf{x} - \mathbf{x}^n)_{j^*}.\end{aligned}$$

In particular, since $B_{\ell+1,j^*} = 1$, setting $k = \ell + 1$ yields

$$(\mathbf{y} - \mathbf{A}'\mathbf{x}^n)_{i(\ell+1)} = (\mathbf{x} - \mathbf{x}^n)_{j^*} \neq 0.$$

Furthermore, since $B_{k,j^*} = b_k(j^*)$ for $k \in [\ell]$, we obtain

$$\begin{aligned}\sum_{k=0}^{\ell} (\mathbf{y} - \mathbf{A}'\mathbf{x}^n)_{(i-1)(\ell+1)+k}\, 2^{k-1} &= \sum_{k=0}^{\ell} b_k(j^*)\, 2^{k-1} (\mathbf{x} - \mathbf{x}^n)_{j^*} \\ &= (j^* - 1)(\mathbf{x} - \mathbf{x}^n)_{j^*}.\end{aligned}$$

This means that $v_i = (\mathbf{x} - \mathbf{x}^n)_{j^*}$ and that $j_i = j^*$. Thus, it follows that $(\mathbf{x} - \mathbf{x}^{n+1})_{j^*} = x_{j^*} - (x^n_{j^*} + v_i) = 0$ provided $\operatorname{card}(R_1(j^*, S^n)) \geq d/2$. If t denotes the number of such j^*, Corollary 14.5 implies that

$$\begin{aligned}(1 - 2\theta_s) d \operatorname{card}(S^n) \leq \operatorname{card}(R_1(S^n)) &= \sum_{j \in S^n} \operatorname{card}(R_1(j, S^n)) \\ &\leq t\, d + (\operatorname{card}(S^n) - t)\, d/2,\end{aligned}$$

which yields $t \geq (1-4\theta_s)\operatorname{card}(S^n)$. Therefore, at least $(1-4\theta_s)\operatorname{card}(S^n)$ zeros entries of $\mathbf{x}-\mathbf{x}^{n+1}$ are created by elements $i \in R_1(S^n)$. Finally, we take into account that elements $i \in R(S^n) \setminus R_1(S^n)$ may potentially corrupt zero entries of $\mathbf{x}-\mathbf{x}^n$ to nonzero entries of $\mathbf{x}-\mathbf{x}^{n+1}$. For a corruption to occur, we need a group of at least $d/2$ elements in $R(S^n) \setminus R_1(S^n)$, which has size at most $2\theta_s d \operatorname{card}(S^n)$, hence the number of corruptions is at most $4\theta_s \operatorname{card}(S^n)$. Putting everything together, we deduce the desired claim from

$$\begin{aligned}\operatorname{card}(S^{n+1}) &\leq \operatorname{card}(S^n) - (1-4\theta_s)\operatorname{card}(S^n) + 4\theta_s \operatorname{card}(S^n) = 8\theta_s \operatorname{card}(S^n)\\ &< \frac{\operatorname{card}(S^n)}{2}.\end{aligned}$$

It now remains to count the number of algebraic operations the procedure requires. At each iteration, we notice that the first step requires $\mathcal{O}(m(s+\ell s)) = \mathcal{O}(sm\ell)$ operations, since the sparsity of $\mathbf{x}^n$ ensures that each component of $\mathbf{A}'\mathbf{x}^n$ can be computed in $\mathcal{O}(s)$ operations, and that the second step requires $\mathcal{O}(s)$ operations, since the previous argument ensures that at most $\mathcal{O}(s)$ entries change from $\mathbf{x}^n$ to $\mathbf{x}^{n+1}$. Overall, the total number of algebraic operations is then $\mathcal{O}(\bar{n}sm\ell) = \mathcal{O}(\log_2(s)s^2\log_2(N/s)\log_2(N))$. □

Notes

Some authors use the terms unbalanced expander or left regular bipartite expander instead of lossless expander. We opted for the terminology of [241]. As already pointed out, a lossless expander is different from an expander. We present here two equivalent definitions of the latter. They both concern an undirected graph $G=(V,E)$, with set of vertices V and set of edges E, which is d-regular in the sense that the number d of edges attached to a vertex is the same for all vertices. For $0<\mu<1$, the combinatorial property defining a *μ-edge expander* is $\operatorname{card}(E(S,\overline{S})) \geq \mu d \operatorname{card}(S)$ for all $S \subseteq V$ with $\operatorname{card}(S) \leq \operatorname{card}(V)/2$, where $E(S,\overline{S})$ denotes the set of edges between S and its complement $\overline{S}$. For $0<\lambda<1$, the algebraic property defining a λ-expander uses its adjacency matrix $\mathbf{A}$ defined by $A_{i,j}=1$ if there is an edge connecting i and j, $A_{i,j}=0$ if there is none. Note the usual identification of V to $[n]$ with $n := \operatorname{card}(V)$. Since the matrix $\mathbf{A}/d$ is symmetric and *stochastic*, i.e it has nonnegative entries summing to one along each row and along each column, it has n real eigenvalues $\lambda_1 = 1, \lambda_2, \ldots, \lambda_n$ ordered as $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. The graph G is then called a *λ-expander* if $|\lambda_2| \leq \lambda$, or equivalently if its *spectral gap* $1-|\lambda_2|$ is at least $1-\lambda$. The combinatorial and algebraic definitions are equivalent, since a λ-expander is a $(1-\lambda)/2$-edge expander and a μ-edge expander is a $1-\mu^2/2$-expander. We refer the reader to [15, Chapter 21] for more details on the subject.

There are no deterministic construction of lossless expanders with optimal parameters available to date, but there exist explicit constructions with $d \asymp (\log(N)\log(s))^{1+1/\alpha}$ and $m \asymp d^2 s^{1+\alpha}$ for any $\alpha > 0$, see [217].

The stable null space property for adjacency matrices of lossless expanders was established by R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss in [32]. We mainly followed their arguments to prove the robust null space property in Theorem 14.10, except that we did not call upon the ℓ_1-restricted isometry property that they established first — see Exercise 14.4.

The algorithm (14.11) is a modification of the sparse matching pursuit algorithm proposed by R. Berinde, P. Indik, and M. Ružić in [33]. The analysis of the latter is also based on Lemma 14.15, see Exercise 14.10. The way we proved Lemma 14.15 differs from the original proof of [33]. There are other iterative algorithms yielding stable and robust reconstruction using adjacency matrices of lossless expanders, see the survey [246] by P. Indik and A. Gilbert. For instance, the expander matching pursuit algorithm of [247] precedes the sparse matching pursuit algorithm and runs in linear time, while the HHS (heavy hitters on steroids) pursuit of [195] runs in sublinear time. The sublinear-time algorithm of Theorem 14.16 is taken from [32], and the one of Exercise 14.11 from [252], but they were not designed with stability in mind. There are also sublinear-time algorithms for matrices other than adjacency matrices of lossless expanders, for instance [248, 227, 226] deals with partial Fourier matrices.

Exercises

14.1. Show that the expansion property (14.1) for $\mathrm{card}(S) = s$ does not necessarily imply the expansion property for $\mathrm{card}(S) < s$.

14.2. Prove that a left d-regular bipartite graph is a $(s, d, (d-1)/d)$-lossless expander if and only if, for any set S of left vertices with $\mathrm{card}(S) \le s$, one can find for each $j \in J$ an edge $\overline{j\, i_j}$ in such a way that the right vertices i_j, $j \in S$, are all distinct. You may use *Hall's theorem*: for finite subsets $X_1, X_2, \ldots, X_n$ of a set X, one can find distinct points $x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$ if and only $\mathrm{card}(\cup_{k\in K} X_k) \ge \mathrm{card}(K)$ for all $K \subseteq [n]$.

14.3. Let $R_{\ge k}(S)$ be the set of right vertices connected to at least k left vertices of a set S in a left d-regular bipartite graph. Prove that the graph is an (s, d, θ)-lossless expander if and only if $\sum_{k\ge 2} \mathrm{card}(R_{\ge k}(S)) \le \theta d\, \mathrm{card}(S)$ for any set S of at most s left vertices. Deduce that $\mathrm{card}(R_{\ge 2}(S)) \le \theta d\, \mathrm{card}(S)$ for any set S of at most s left vertices if the graph is an (s, d, θ)-lossless expander.

14.4. Prove that the $m \times N$ adjacency matrix $\mathbf{A}$ of a (s, d, θ)-lossless expander satisfies the property that

$$d(1-2\theta)\|\mathbf{z}\|_1 \le \|\mathbf{A}\mathbf{z}\|_1 \le d\|\mathbf{z}\|_1 \quad \text{for all } s\text{-sparse } \mathbf{z} \in \mathbb{C}^N,$$

which can be interpreted as a scaled restricted isometry property in ℓ_1.

14.5. For a fixed $\delta > 0$, suppose that a measurement matrix $\mathbf{A} \in \{0,1\}^{m \times N}$ satisfies $\delta_s(\gamma \mathbf{A}) \leq \delta$ for some $\gamma > 0$. Let c and r denote the minimal number of ones per columns of $\mathbf{A}$ and the maximal number of ones per row of $\mathbf{A}$. Show that

$$c \leq \frac{r\,m}{N}.$$

Observe also that $c \geq (1-\delta)/\gamma^2$ by considering a suitable 1-sparse vector. Then, by considering any vector in $\{0,1\}^N$ with exactly s ones, deduce that

$$c \leq \frac{1+\delta}{1-\delta}\frac{m}{s}.$$

Next, by considering a suitable vector in $\{0,1\}^N$ with exactly $t := \min\{r,s\}$ ones, observe that

$$t \leq \frac{1+\delta}{\gamma^2} \leq \frac{1+\delta}{1-\delta}\,c.$$

Separating the cases $r \geq s$ and $r < s$, conclude that

$$m \geq \min\Big\{\frac{1-\delta}{1+\delta}\,N, \Big(\frac{1-\delta}{1+\delta}\Big)^2 s^2\Big\},$$

so that matrices populated with zeros and ones do not satisfy the classical restricted isometry property in the parameter range relevant to compressive sensing.

14.6. Let $\mathbf{A} \in \{0,1\}^{m \times N}$ be adjacency matrix of a left regular bipartite graph, and let $S \subseteq [N]$ be a fixed index set. Suppose that every nonnegative vector supported on S is uniquely recovered via ℓ_1-minimization with measurement matrix $\mathbf{A}$. Prove that every nonnegative vector $\mathbf{x}$ supported on S is in fact the unique vector in the set $\{\mathbf{z} \in \mathbb{R}^N : \mathbf{z} \geq 0, \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}\}$.

14.7. Extend Theorem 14.9 to the case of a measurement error considered in ℓ_p-norms, $p \geq 1$. Precisely, given the adjacency matrix $\mathbf{A}$ of a left d-regular bipartite graph with $\theta_{2s} < 1/6$, prove that a solution $\mathbf{x}^\sharp$ of

$$\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} \ \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_p \leq \eta,$$

where $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_p \leq \eta$, satisfies

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \leq \frac{2(1-2\theta)}{(1-6\theta)}\,\sigma_s(\mathbf{x})_1 + \frac{4}{(1-6\theta)d}\,\frac{s^{1-1/p}}{d^{1/p}}\,\eta.$$

14.8. For the adjacency matrix $\mathbf{A} \in \{0,1\}^{m \times N}$ of a left regular bipartite graph, and let $\mathbf{A}' \in \{-1,1\}^{m \times N}$ be the matrix obtained from $\mathbf{A}$ by replacing the zeros by negative ones. Given $\mathbf{x} \in \mathbb{C}^N$, prove that the solutions of the two problems

$$\begin{aligned} &\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}, \\ &\text{minimize } \|\mathbf{z}\|_1 \quad \text{subject to } \mathbf{A}'\mathbf{z} = \mathbf{A}'\mathbf{x}, \end{aligned}$$

are the identical.

14.9. For the quantiles q_k, prove the inequalities (14.13), (14.14), as well as

$$q_k[a_1,\dots,a_d] \le q_k[b_1,\dots,b_d] \qquad \text{if } a_j \le b_j \text{ for all } j,$$
$$q_{2k}[a_1+b_1,\dots,a_d+b_d] \le q_k[a_1,\dots,a_d] + q_k[b_1,\dots,b_d] \quad \text{if } a_j, b_j \ge 0 \text{ for all } j.$$

14.10. Establish an analog of Theorem 14.14 when $\theta_{4s} < 1/20$ for the *sparse matching pursuit* algorithm consisting in the scheme

$$\mathbf{u}^{n+1} := H_{2s}(\mathcal{M}(\mathbf{y} - \mathbf{A}\mathbf{x}^n)), \quad \mathbf{x}^{n+1} := H_s(\mathbf{x}^n + \mathbf{u}^{n+1}).$$

14.11. Let $\mathbf{A} \in \{0,1\}^{m\times N}$ be the adjacency matrix of an (s,d,θ)-lossless expander. If θ is small enough, prove that every s-sparse vector is recovered from $\mathbf{y} = \mathbf{A}\mathbf{x}$ in a finite number of iterations of the algorithm

- for each $i \in [m]$, compute
$$v_i := (\mathbf{y} - \mathbf{A}\mathbf{x}^n)_i,$$
- if there are $i_1,\dots,i_r \in R(j)$ with $r \ge d/2$ and $v_{i_1} = \cdots = v_{i_r} =: v \ne 0$,
$$x_j^{n+1} = x_j^n + v.$$

15

Algorithms for ℓ_1-Minimization

Throughout this book ℓ_1-minimization plays a central role as recovery method for compressive sensing. So far, however, no *algorithm* for this minimization problem was introduced. In Chapter 3 we mentioned that ℓ_1-minimization can be recast as a linear program in the real case, see (P'_1), and as a second order cone program in the complex case, see $(\mathrm{P}'_{1,\eta})$. For linear and second order cone programs standard software is available, which is based on interior point methods or the older simplex method for linear programs. While such standard software works reliably and is straightforward to use, it is designed for general linear and second order cone problems. It turns out that algorithms which are developed specifically for ℓ_1-minimization may be significantly faster than general purpose methods. This chapter introduces and analyzes several of these algorithms. The homotopy method, which is restricted to the real case, is somewhat similar to orthogonal matching pursuit, but is guaranteed to always provide the ℓ_1-minimizer. In Section 15.2 we introduce an algorithm due to A. Chambolle and T. Pock. It applies actually to a whole class of optimization problems which are similar to ℓ_1-minimization. Our third algorithm, iteratively reweighted least squares, is only a proxy for ℓ_1-minimization. But its formulation is motivated by ℓ_1-minimization and in certain cases it indeed provides the ℓ_1-minimizer. Under the stable null space property (equivalent to exact and approximate sparse recovery via ℓ_1-minimization) we will show the same error guarantees as valid for ℓ_1-minimization. But in general, its output maybe different from the ℓ_1-minimizer.

15.1 The Homotopy Method

The homotopy method solves the ℓ_1-minimization problem

$$\min \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} \tag{15.1}$$

in the real case, that is, for $\mathbf{A} \in \mathbb{R}^{m \times N}$ and $\mathbf{y} \in \mathbb{R}^m$. Moreover, a slight variant solves the quadratically constrained ℓ_1-minimization problem

$$\min \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta \,. \tag{15.2}$$

For $\lambda > 0$, we consider the ℓ_1-regularized least squares functional

$$F_\lambda(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1, \quad \mathbf{x} \in \mathbb{R}^N \,, \tag{15.3}$$

and $\mathbf{x}_\lambda$ to be a minimizer of it. When $\lambda = \widehat{\lambda}$ is large enough then $\mathbf{x}_{\widehat{\lambda}} = \mathbf{0}$. Furthermore, essentially have $\lim_{\lambda\to 0} \mathbf{x}_\lambda = \mathbf{x}^\sharp$, where $\mathbf{x}^\sharp$ is a minimizer of (15.1). A precise statement is contained in the next result.

Proposition 15.1. *Assume that $\mathbf{A}\mathbf{x} = \mathbf{y}$ has a solution. If the minimizer $\mathbf{x}^\sharp$ of (15.1) is unique then*

$$\lim_{\lambda\to 0} \mathbf{x}_\lambda = \mathbf{x}^\sharp \,.$$

More generally, if the minimizer of (15.1) is not unique then the $\mathbf{x}_\lambda$ are bounded and every accumulation point of $\mathbf{x}_\lambda$ is a minimizer of (15.1).

Proof. For the boundedness, observe that $F_\lambda(\mathbf{x}_\lambda) \leq F_\lambda(\mathbf{0}) = \|\mathbf{y}\|_2^2/2$. Pick an arbitrary sequence $(\lambda_n)_{n\in N} \subset (0,\infty)$ that converges monotonically to 0. For simplicity we write $\mathbf{x}^n = \mathbf{x}_{\lambda_n}$. Let $\mathbf{x}^\sharp$ be a minimizer of the ℓ_1-minimization problem (15.1). Then

$$\|\mathbf{x}^n\|_1 \leq \frac{1}{\lambda_n} F_{\lambda_n}(\mathbf{x}^n) \leq \frac{1}{\lambda_n} F_{\lambda_n}(\mathbf{x}^\sharp) = \|\mathbf{x}^\sharp\|_1 \,,$$

since $\mathbf{A}\mathbf{x}^\sharp = \mathbf{y}$ by (15.1). Therefore, we may restrict our considerations to the compact set

$$K := \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_1 \leq \|\mathbf{x}^\sharp\|_1\} \,.$$

Introduce the function $F : K \to \mathbb{R}$, $F(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2/2$. By compactness of K, F is trivially coercive, see Definition C.13. Moreover, denoting by F_n the functions F_{λ_n} restricted to K, we observe that the sequence F_n is monotonically decreasing in n and converges to F. By Proposition C.14, any accumulation point of the sequence $(\mathbf{x}^n)_n$ is a minimizer $\mathbf{x}'$ of F so that it satisfies $\mathbf{A}\mathbf{x}' = \mathbf{y}$. Therefore, $\|\mathbf{x}'\| \geq \|\mathbf{x}^\sharp\|_1$ On the other hand, by definition of the set K we also have $\|\mathbf{x}'\|_1 = \lim_{n\to\infty} \|\mathbf{x}^n\|_1 \leq \|\mathbf{x}^\sharp\|_1$. It follows that

$$\|\mathbf{x}'\|_1 = \|\mathbf{x}^\sharp\|_1 \,,$$

so that every accumulation point of $(\mathbf{x}^n)_n$ is a minimizer of (15.1). If the minimizer is unique, then this argument shows that any subsequence of $(\mathbf{x}^n)_n$ converges to $\mathbf{x}^\sharp$, so that the full sequence converges to $\mathbf{x}^\sharp$. □

The basic idea of the homotopy method is to follow the solution $\mathbf{x}_\lambda$ from $\mathbf{x}_{\widehat{\lambda}} = 0$ to $\mathbf{x}^\sharp$. As we will show below, the solution path $\lambda \mapsto \mathbf{x}_\lambda$ is piecewise linear, and it is enough to trace the endpoints of the linear pieces.

By Theorem B.21 the minimizer of (15.3) can be characterized using the subdifferential defined in (B.11). The subdifferential of F_λ is given by

$$\partial F_\lambda(\mathbf{x}) = \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) + \lambda \partial \|\mathbf{x}\|_1 \ ,$$

where the subdifferential of the ℓ_1-norm is given by

$$\partial \|\mathbf{x}\|_1 = \{\mathbf{v} \in \mathbb{R}^N : \ v_\ell \in \partial |x_\ell|, \ell \in [N]\} \ .$$

Hereby, the subdifferential of the absolute value is given by

$$\partial |z| = \begin{cases} \{\mathrm{sgn}(z)\}, & \text{if } z \neq 0 \ , \\ [-1, 1] & \text{if } z = 0 \ . \end{cases}$$

A vector $\mathbf{x}$ is the minimizer of F_λ if and only if $\mathbf{0} \in \partial F_\lambda(\mathbf{x})$, see Theorem B.21. By the above this is equivalent to

$$(\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y}))_\ell = -\lambda \,\mathrm{sgn}(x_\ell) \qquad \text{if } x_\ell \neq 0 \ , \tag{15.4}$$

$$|(\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y}))_\ell| \leq \lambda \qquad \text{if } x_\ell = 0 \ , \tag{15.5}$$

for $\ell \in [N]$.

The homotopy method starts with $\mathbf{x}^{(0)} = \mathbf{x}_\lambda = \mathbf{0}$. By condition (15.5) the corresponding λ is chosen as $\lambda = \lambda^{(0)} = \|\mathbf{A}^*\mathbf{y}\|_\infty$.

In the further steps $j = 1, 2, \ldots$ the algorithm varies λ, computes corresponding minimizers $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$, and maintains an active (support) set T_j. Denote by

$$\mathbf{c}^{(j)} = \mathbf{A}^*(\mathbf{A}\mathbf{x}^{(j-1)} - \mathbf{y})$$

the current residual vector.

Step 1: Let

$$\ell^{(1)} := \arg\max_{\ell \in [N]} |(\mathbf{A}^*\mathbf{y})_\ell| = \arg\max_{\ell \in [N]} |c_\ell^{(1)}| \ . \tag{15.6}$$

One assumes here and also in the further steps that the maximum is attained at only one index ℓ. The case that the maximum is attained simultaneously at two or more indexes ℓ (which almost never happens) requires more complications that will not be covered here (but see the Notes section).

Now set $T_1 = \{\ell^{(1)}\}$. We introduce the vector $\mathbf{d} \in \mathbb{R}^N$ describing the direction of the solution (homotopy) path with entries

$$d_{\ell^{(1)}}^{(1)} = \|\mathbf{a}_{\ell^{(1)}}\|_2^{-2} \mathrm{sgn}((\mathbf{A}^*\mathbf{y})_{\ell^{(1)}}) \quad \text{ and } \quad d_\ell^{(1)} = 0 \ , \quad \ell \neq \ell^{(1)} \ .$$

The first linear piece of the solution path then takes the form

$$\mathbf{x} = \mathbf{x}(\gamma) = \mathbf{x}^{(0)} + \gamma \mathbf{d}^{(1)} = \gamma \mathbf{d}^{(1)} \ , \quad \gamma \in [0, \gamma^{(1)}]$$

with some $\gamma^{(1)}$ to be determined below. One verifies with the definition of $\mathbf{d}^{(1)}$ that (15.4) is always satisfied for $\mathbf{x} = \mathbf{x}(\gamma)$ and $\lambda = \lambda(\gamma) = \lambda^{(0)} - \gamma$, $\gamma \in [0, \lambda^{(0)}]$. The next breakpoint is found by determining the maximal $\gamma = \gamma^{(1)} > 0$ for which (15.5) is still satisfied. Using the notation $(t)_+ = \max\{t, 0\}$, this gives

$$\gamma^{(1)} = \min_{\ell \neq \ell^{(1)}} \left\{ \left(\frac{\lambda^{(0)} + c_\ell^{(1)}}{1 - (\mathbf{A}^* \mathbf{A} \mathbf{d}^{(1)})_\ell} \right)_+ , \left(\frac{\lambda^{(0)} - c_\ell^{(1)}}{1 + (\mathbf{A}^* \mathbf{A} \mathbf{d}^{(1)})_\ell} \right)_+ \right\} . \tag{15.7}$$

Then $\mathbf{x}^{(1)} = \mathbf{x}(\gamma^{(1)}) = \gamma^{(1)} \mathbf{d}^{(1)}$ is the next minimizer of F_λ for $\lambda = \lambda^{(1)} := \lambda^{(0)} - \gamma^{(1)}$. This $\lambda^{(1)}$ satisfies $\lambda^{(1)} = \|\mathbf{c}^{(2)}\|_\infty$. Let $\ell^{(2)}$ be the index where the minimum in (15.7) is attained (where we again assume that the minimum is attained only at one index) and put $T_2 = \{\ell^{(1)}, \ell^{(2)}\}$.

Step j: The new direction $\mathbf{d}^{(j)}$ of the homotopy path is determined by

$$\mathbf{A}_{T_j}^* \mathbf{A}_{T_j} \mathbf{d}_{T_j}^{(j)} = \operatorname{sgn}(\mathbf{c}_{T_j}^{(j)}) . \tag{15.8}$$

This amounts to solving a linear system of equations of size $|T_j| \times |T_j|$, where $|T_j| \leq j$. Outside the components in T_j we set $d_\ell^{(j)} = 0$, $\ell \notin T_j$. The next linear piece of the path is then given by

$$\mathbf{x}(\gamma) = \mathbf{x}^{(j-1)} + \gamma \mathbf{d}^{(j)}, \quad \gamma \in [0, \gamma^{(j)}] .$$

The maximal γ such that $\mathbf{x}(\gamma)$ satisfies (15.5) is

$$\gamma_+^{(j)} = \min_{\ell \notin T_j} \left\{ \left(\frac{\lambda^{(j-1)} + c_\ell^{(j)}}{1 - (\mathbf{A}^* \mathbf{A} \mathbf{d}^{(j)})_\ell} \right)_+ , \left(\frac{\lambda^{(j-1)} - c_\ell^{(j)}}{1 + (\mathbf{A}^* \mathbf{A} \mathbf{d}^{(j)})_\ell} \right)_+ \right\} . \tag{15.9}$$

The maximal γ such that $\mathbf{x}(\gamma)$ satisfies (15.4) is given by

$$\gamma_-^{(j)} = \min_{\ell \in T_j} \left\{ \left(-\mathbf{x}_\ell^{(j-1)} / d_\ell^{(j)} \right)_+ \right\} . \tag{15.10}$$

The next breakpoint is given by $\mathbf{x}^{(j)} = \mathbf{x}(\gamma^{(j)})$ with $\gamma^{(j)} = \min\{\gamma_+^{(j)}, \gamma_-^{(j)}\}$. If $\gamma_+^{(j)}$ determines the minimum, then the index $\ell_+^{(j)} \notin T_j$ providing the minimum in (15.9) is added to the active set, $T_{j+1} = T_j \cup \{\ell_+^{(j)}\}$. If $\gamma^{(j)} = \gamma_-^{(j)}$, then the index $\ell_-^{(j)} \in T_j$ is removed from the active set, $T_{j+1} = T_j \setminus \{\ell_-^{(j)}\}$. We update $\lambda^{(j)} = \lambda^{(j-1)} - \gamma^{(j)}$. Then we have by construction that $\lambda^{(j)} = \|\mathbf{c}^{(j+1)}\|_\infty$.

The algorithm stops when $\lambda^{(j)} = \|\mathbf{c}^{(j+1)}\|_\infty = 0$, i.e., when the residual vanishes, and outputs $\mathbf{x}^\sharp = \mathbf{x}^{(j)}$.

Theorem 15.2. *Assume that the minimizer $\mathbf{x}^\sharp$ of ℓ_1-minimization problem* (15.1) *is unique. If in each step the minimum in* (15.9) *and* (15.10) *is attained in only one index ℓ, then the homotopy algorithm as described outputs $\mathbf{x}^\sharp$.*

Proof. Following the description of the algorithm above, it only remains to show that the algorithm eventually stops. To this end, we note that the sign patterns $\operatorname{sgn}(\mathbf{x}_{\lambda^{(i)}})$ are different for each i. Indeed, if they would be the same for two parameters $\lambda^{(i)}$ and $\lambda^{(j)}$, $j > i$, then (15.4) would imply that, for all ℓ such that $\operatorname{sgn}(\mathbf{x}_{\lambda^{(i)}})_\ell = \operatorname{sgn}(\mathbf{x}_{\lambda^{(j)}})_\ell =: \sigma_\ell \neq 0$,

$$(\mathbf{A}^*\mathbf{A}(\mathbf{x}_{\lambda^{(i)}} - \mathbf{x}_{\lambda^{(j)}}))_\ell = (\lambda^{(i)} - \lambda^{(j)})\sigma_\ell \,,$$

which in turn would mean that $\mathbf{x}_{\lambda^{(j)}}$ would be on the interior of the linear piece of the homotopy path starting from $\mathbf{x}_{\lambda^{(i)}}$. However, by construction, the points $\mathbf{x}_{\lambda^{(j)}}$ are always endpoints of these linear pieces. Since there exist only a finite number of possible sign patterns the algorithm eventually has to stop. □

Remark 15.3. The theorem still holds if $\mathbf{Ax} = \mathbf{y}$ has a solution but the ℓ_1-minimizer is not unique. Then the homotopy method computes a minimizer of (15.1). The case that the minimum in (15.9) and (15.10) is attained in more than one index ℓ is very unlikely. The algorithm may be modified in this case, see the Notes section.

If the algorithm is stopped earlier at some iteration j then obviously it yields the minimizer of $F_\lambda = F_{\lambda^{(j)}}$. In particular, obvious stopping rules may also be used to solve the problems

$$\min \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{Ax} - \mathbf{y}\|_2 \le \eta \tag{15.11}$$

$$\text{or} \quad \min \|\mathbf{Ax} - \mathbf{y}\|_2 \quad \text{subject to } \|\mathbf{x}\|_1 \le \delta \,. \tag{15.12}$$

The first of these appears in (15.2), and the second is called the LASSO (least absolute shrinkage and selection operator), see Chapter 3.

The LARS (least angle regression) algorithm is a simple modification of the homotopy method, which only adds elements to the active set in each step. So $\gamma_-^{(j)}$ in (15.10) is not considered. (Sometimes the homotopy method is therefore also called modified LARS.) Clearly, LARS is not guaranteed any more to yield the solution of (15.1). However, it is observed empirically that often in sparse recovery problems, the homotopy method never removes elements from the active set, so that in this case LARS and homotopy perform the same steps. If the solution of (15.1) is s-sparse and the homotopy method never removes elements then the solution is obtained after precisely s-steps. Furthermore, the most demanding computational part at step j is then the solution of the $j \times j$ linear system of equations (15.8).

In conclusion, the homotopy and LARS methods are very efficient for sparse recovery problems - provided the solution is very sparse. For only mildly sparse solutions the methods in the next sections may be better suited.

15.2 Chambolle and Pock's Primal Dual Algorithm

This section covers an iterative primal dual algorithm for the numerical solution of general optimization problems including the various ℓ_1-minimization problems appearing in this book. We require some knowledge of convex analysis and optimization as covered in Appendix B.

Remark 15.4. We formulate everything below in the complex setting of $\mathbb{C}^N$, although the material in Appendix B is treated only for the real case. As noted there, everything carries over to the complex case by identifying $\mathbb{C}^N$ with $\mathbb{R}^{2N}$. The only formal difference when making this identification concrete is that complex inner products have to be replaced by real inner products $\mathrm{Re}(\langle \cdot, \cdot \rangle)$. Reversely, everything below holds also if $\mathbb{C}^N$ is replaced by $\mathbb{R}^N$, of course.

We consider a general optimization problem of the form

$$\min_{\mathbf{x} \in \mathbb{C}^N} F(\mathbf{A}\mathbf{x}) + G(\mathbf{x}) \;, \tag{15.13}$$

with $\mathbf{A} \in \mathbb{C}^{m \times N}$ and extended real-valued lower semicontinuous convex functions $F : \mathbb{C}^m \to (-\infty, \infty]$, $G : \mathbb{C}^N \to (-\infty, \infty]$, see Definition B.13 for the notion of lower semi-continuity. (Note that the function value ∞ is allowed, so that the requirement of continuity would be too strong). We will explain in detail below, how ℓ_1-minimization fits into this framework.

The dual problem of (15.13) is given by

$$\max_{\boldsymbol{\xi} \in \mathbb{C}^m} -F^*(\boldsymbol{\xi}) - G^*(-\mathbf{A}^*\boldsymbol{\xi}) \;, \tag{15.14}$$

see (B.48). Here F^* and G^* are the convex conjugate functions of F and G (Definition B.17).

Theorem B.30 states that strong duality holds for the primal dual pair (15.13) and (15.14) under mild assumptions on F and G, which are always met in the special cases of our interest. Furthermore, the joint primal dual optimization of (15.13) and (15.14) is equivalent to solving the saddle point problem

$$\min_{\mathbf{x} \in \mathbb{C}^N} \max_{\boldsymbol{\xi} \in \mathbb{C}^m} \mathrm{Re}(\langle \mathbf{A}\mathbf{x}, \boldsymbol{\xi} \rangle) + G(\mathbf{x}) - F^*(\boldsymbol{\xi}) \;. \tag{15.15}$$

The algorithm we will describe below uses the proximal mappings (B.13) of F^* and G. It will be convenient to introduce another parameter $\tau > 0$ into these mappings by setting

$$P_G(\tau; \mathbf{z}) := P_{\tau G}(\mathbf{z}) = \arg\min_{\mathbf{x} \in \mathbb{C}^N} \left\{ \tau G(\mathbf{x}) + \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 \right\} \;, \quad \mathbf{z} \in \mathbb{C}^N \;, \tag{15.16}$$

and $P_{F^*}(\tau; \mathbf{z})$ is defined in the same way.

We assume that $P_{F^*}(\tau; \mathbf{z})$ and $P_G(\tau; \mathbf{z})$ are easy to evaluate. Note that by Moreau's identity (B.15) the proximal mapping associated with F^* is easy to compute once the one associated with F is. Although the algorithm can be formulated for arbitrary convex F and G, it will only be efficient under this assumption because it relies on a repeated application of the proximal mappings.

Primal Dual Algorithm

Parameters: $\theta \in [0,1]$, $\tau, \sigma > 0$ such that $\tau\sigma\|\mathbf{A}\|_{2\to 2} < 1$.
Initialization: $\mathbf{x}^0 \in \mathbb{C}^N, \boldsymbol{\xi}^0 \in \mathbb{C}^m$, $\bar{\mathbf{x}}^0 = \mathbf{x}^0$.
Iteration: repeat until stopping criterion is met at $n = \bar{n}$

$$\boldsymbol{\xi}^{n+1} := P_{F^*}(\sigma; \boldsymbol{\xi}^n + \sigma\mathbf{A}\bar{\mathbf{x}}^n) \,, \tag{PD$_1$}$$
$$\mathbf{x}^{n+1} := P_G(\tau; \mathbf{x}^n - \tau\mathbf{A}^*\boldsymbol{\xi}^{n+1}) \,, \tag{PD$_2$}$$
$$\bar{\mathbf{x}}^{n+1} := \mathbf{x}^{n+1} + \theta(\mathbf{x}^{n+1} - \mathbf{x}^n) \,. \tag{PD$_3$}$$

Output: Approximation $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$ to solution of primal problem (15.13),
Approximation $\boldsymbol{\xi}^\sharp = \boldsymbol{\xi}^{\bar{n}}$ to solution of dual problem (15.14).

We will analyze this algorithm for the parameter choice $\theta = 1$. In the case that F^* or G are uniformly convex, an acceleration can be achieved by varying the parameters θ, τ, σ during the iterations, see the Notes section.

A possible stopping criterion may be based on the primal dual gap (B.29), which in our case reads

$$E(\mathbf{x}, \boldsymbol{\xi}) = F(\mathbf{A}\mathbf{x}) + G(\mathbf{x}) + F^*(\boldsymbol{\xi}) + G^*(-\mathbf{A}^*\boldsymbol{\xi}) \geq 0 \,.$$

For the primal dual optimum $(\mathbf{x}^*, \boldsymbol{\xi}^*)$ we have $E(\mathbf{x}^*, \boldsymbol{\xi}^*) = 0$ and $E(\mathbf{x}^n, \boldsymbol{\xi}^n) \leq \eta$ for some prescribed tolerance $\eta > 0$ can be taken as a criterion to stop the iterations at n.

Remark 15.5. In two of the examples below, F can take the value ∞, so that E may also be infinite during the iterations and gives only limited information about the quality of the approximation of the iterates to the optimal solution. In this case, one may modify the primal dual gap so that the value ∞ does not occur anymore. Empirically, a modified primal dual gap still provides a good stopping criterion.

Note that if $\mathbf{A}$ and $\mathbf{A}^*$ allow fast matrix multiplication routines then the primal dual algorithm can easily exploit this fact for speed up.

A variant of the algorithm is obtained by interchanging the updates for $\boldsymbol{\xi}^{n+1}$ and $\mathbf{x}^{n+1}$ and carrying along an auxiliary variable $\bar{\boldsymbol{\xi}}^n$, that is,

$$\mathbf{x}^{n+1} = P_G(\tau; \mathbf{x}^n - \tau\mathbf{A}^*\bar{\boldsymbol{\xi}}^n) \,,$$
$$\boldsymbol{\xi}^{n+1} = P_{F^*}(\sigma; \boldsymbol{\xi}^n + \sigma\mathbf{A}\mathbf{x}^{n+1}) \,,$$
$$\bar{\boldsymbol{\xi}}^{n+1} = \boldsymbol{\xi}^{n+1} + \theta(\boldsymbol{\xi}^{n+1} - \boldsymbol{\xi}^n) \,.$$

The algorithm can be interpreted as a fixed point iteration:

Proposition 15.6. *A point* $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ *is a fixed point of the iterations* (PD$_1$), (PD$_2$), (PD$_3$) *(for any choice of* θ*) if and only if* $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ *is a saddle point*

of (15.15)*, that is, a primal-dual optimal point for the problems* (15.13) *and* (15.14)*.*

Proof. It follows from the characterization of the proximal mapping in Proposition B.23 that a fixed point $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ satisfies

$$\begin{aligned} \boldsymbol{\xi}^\sharp + \sigma \mathbf{A}\mathbf{x}^\sharp &\in \boldsymbol{\xi}^\sharp + \sigma \partial F^*(\boldsymbol{\xi}^\sharp) \,, \\ \mathbf{x}^\sharp - \tau \mathbf{A}^* \boldsymbol{\xi}^\sharp &\in \mathbf{x}^\sharp + \tau \partial G(\mathbf{x}^\sharp) \,, \end{aligned}$$

where ∂F^* and ∂G are the subdifferentials of F^* and G, see Definition B.20. Equivalently,

$$\mathbf{0} \in -\mathbf{A}\mathbf{x}^\sharp + \partial F^*(\boldsymbol{\xi}^\sharp) \quad \text{and} \quad \mathbf{0} \in \mathbf{A}^* \boldsymbol{\xi}^\sharp + \partial G(\mathbf{x}^\sharp) \,.$$

By Theorem B.21 these relations are equivalent to $\mathbf{x}^\sharp$ being the minimum of the function $\mathbf{x} \mapsto \mathrm{Re}(\langle \mathbf{x}, \mathbf{A}^* \boldsymbol{\xi}^\sharp \rangle) + G(\mathbf{x}) - F^*(\boldsymbol{\xi}^\sharp)$ and $\boldsymbol{\xi}^\sharp$ being the maximum of the function $\boldsymbol{\xi} \mapsto \mathrm{Re}(\langle \mathbf{A}\mathbf{x}^\sharp, \boldsymbol{\xi} \rangle) + G(\mathbf{x}^\sharp) - F^*(\boldsymbol{\xi})$. This is equivalent to $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ being a saddle point of (15.15).

These arguments show as well the converse that a saddle point of (15.15) is a fixed point of the primal dual algorithm. □

Before continuing with the analysis of this algorithm let us illustrate the setup for various ℓ_1-minimization problems.

Example 15.7. (a) The ℓ_1-minimization problem

$$\min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} \tag{15.17}$$

is equivalent to (15.13) with $G(\mathbf{x}) = \|\mathbf{x}\|_1$ and

$$F(\mathbf{z}) = \chi_{\{\mathbf{y}\}}(\mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z} = \mathbf{y} \,, \\ \infty & \text{if } \mathbf{z} \neq \mathbf{y} \,, \end{cases}$$

the characteristic function of the singleton $\{\mathbf{y}\}$. Note that F is trivially lower semicontinuous. By Example (B.19) the convex conjugates are given by

$$\begin{aligned} F^*(\boldsymbol{\xi}) &= \mathrm{Re}(\langle \boldsymbol{\xi}, \mathbf{y} \rangle) \,, \\ G^*(\boldsymbol{\zeta}) &= \chi_{B_{\|\cdot\|_\infty}}(\boldsymbol{\zeta}) = \begin{cases} 0 & \text{if } \|\boldsymbol{\zeta}\|_\infty \leq 1 \,, \\ \infty & \text{otherwise} \quad . \end{cases} \end{aligned} \tag{15.18}$$

Since points where the target function takes the value $-\infty$ can be discarded when maximizing, we can make such constraint explicit so that the dual program (15.14) becomes

$$\max_{\boldsymbol{\xi} \in \mathbb{C}^m} -\mathrm{Re}(\langle \mathbf{y}, \boldsymbol{\xi} \rangle) \quad \text{subject to } \|\mathbf{A}^* \boldsymbol{\xi}\|_\infty \leq 1 \,.$$

(Note that in Appendix B.5 the dual of the ℓ_1-minimization problem is derived in a slightly different way, see (B.31) and (B.32) and the preceding notes.) The saddle point problem (15.15) reads

$$\min_{\mathbf{x}\in\mathbb{C}^N} \max_{\boldsymbol{\xi}\in\mathbb{C}^m} \mathrm{Re}(\langle \mathbf{A}\mathbf{x}-\mathbf{y}, \boldsymbol{\xi}\rangle) + \|\mathbf{x}\|_1 \ . \tag{15.19}$$

The proximal mapping of F is the projection onto $\{\mathbf{y}\}$, that is, the constant map

$$P_F(\sigma;\boldsymbol{\xi}) = \mathbf{y}, \quad \text{for all } \boldsymbol{\xi}\in\mathbb{R}^m \ .$$

By Moreau's identity (B.15) (or by a straightforward direct computation) the proximal mapping of F^* is therefore

$$P_{F^*}(\sigma;\boldsymbol{\xi}) = \boldsymbol{\xi} - \sigma\mathbf{y} \ .$$

For the proximal mapping of $G(\mathbf{x}) = \|\mathbf{x}\|_1$ we first observe that by a straightforward computation the proximal mapping of the complex absolute value function satisfies, for $z\in\mathbb{C}$.

$$\begin{aligned} P_{|\cdot|}(\tau;z) &= \arg\min_{x\in\mathbb{C}} \left\{ \frac{1}{2}|x-z|^2 + \tau|x| \right\} = \begin{cases} \mathrm{sgn}(z)(|z|-\tau) & \text{if } |z|\geq\tau \ , \\ 0 & \text{otherwise} \end{cases} \\ &=: S_\tau(z) \ , \end{aligned} \tag{15.20}$$

where the sign function is given by $\mathrm{sgn}(z) = z/|z|$ for $z\neq 0$, as usual. The function $S_\tau(z)$ is called (complex) soft thresholding operator. (Note that in the real case it is computed in (B.17).) Since the optimization problem defining the proximal mapping of $\|\cdot\|_1$ decouples, $P_G(\tau;\mathbf{z}) =: \mathcal{S}_\tau(\mathbf{z})$ is given component-wise by

$$P_G(\tau;\mathbf{z})_\ell = S_\tau(z_\ell), \quad \ell\in[N] \ . \tag{15.21}$$

The primal dual algorithm for the ℓ_1-minimization problem (15.17) reads then

$$\begin{aligned} \boldsymbol{\xi}^{n+1} &= \boldsymbol{\xi}^n + \sigma(\mathbf{A}\bar{\mathbf{x}}^n - \mathbf{y}) \ , \\ \mathbf{x}^{n+1} &= \mathcal{S}_\tau(\mathbf{x}^n - \tau\mathbf{A}^*\boldsymbol{\xi}^{n+1}) \ , \\ \bar{\mathbf{x}}^{n+1} &= \mathbf{x}^{n+1} + \theta(\mathbf{x}^{n+1} - \mathbf{x}^n) \ . \end{aligned}$$

(b) The quadratically constraint ℓ_1-minimization problem

$$\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2 \leq \eta \tag{15.22}$$

takes the form (15.13) with $G(\mathbf{x}) = \|\mathbf{x}\|_1$ and

$$F(\mathbf{z}) = \chi_{B(\mathbf{y},\eta)}(\mathbf{z}) = \begin{cases} 0 & \text{if } \|\mathbf{z}-\mathbf{y}\|_2 \leq \eta \ , \\ \infty & \text{otherwise} \ . \end{cases}$$

The function F is lower semicontinuous because the set $B(\mathbf{y}, \eta)$ is closed. Example B.19(d) shows that its convex conjugate is given by

$$F^*(\boldsymbol{\xi}) = \sup_{\mathbf{z}:\|\mathbf{z}-\mathbf{y}\|_2 \le \eta} \mathrm{Re}(\langle \mathbf{z}, \boldsymbol{\xi} \rangle) = \mathrm{Re}(\langle \mathbf{y}, \boldsymbol{\xi} \rangle) + \eta \|\boldsymbol{\xi}\|_2 \, .$$

The convex conjugate of G is given by (15.18). The dual problem to (15.22) is therefore

$$\max_{\xi \in \mathbb{C}^m} - \mathrm{Re}(\langle \mathbf{y}, \boldsymbol{\xi} \rangle) - \eta \|\boldsymbol{\xi}\|_2 \quad \text{subject to } \|\mathbf{A}^* \boldsymbol{\xi}\|_\infty \le 1 \, ,$$

while the associated saddle point problem is given by

$$\min_{\mathbf{x} \in \mathbb{C}^N} \max_{\boldsymbol{\xi} \in \mathbb{C}^m} \mathrm{Re}(\langle \mathbf{A}\mathbf{x} - \mathbf{y}, \boldsymbol{\xi} \rangle) - \eta \|\boldsymbol{\xi}\|_2 + \|\mathbf{x}\|_1 \, . \tag{15.23}$$

The proximal mapping of F is the orthogonal projection onto the ball $B(\mathbf{y}, \eta)$,

$$\begin{aligned} P_F(\sigma; \boldsymbol{\xi}) &= \arg \min_{\boldsymbol{\zeta} \in \mathbb{C}^m : \|\boldsymbol{\zeta} - \mathbf{y}\|_2 \le \eta} \|\boldsymbol{\zeta} - \boldsymbol{\xi}\|_2 \\ &= \begin{cases} \boldsymbol{\xi} & \text{if } \|\boldsymbol{\xi} - \mathbf{y}\|_2 \le \eta \, , \\ \mathbf{y} + \dfrac{\eta}{\|\boldsymbol{\xi} - \mathbf{y}\|_2} (\boldsymbol{\xi} - \mathbf{y}) & \text{otherwise} \, . \end{cases} \end{aligned}$$

By Moreau's identity (B.15) the proximal mapping of F^* is given by

$$P_{F^*}(\sigma; \boldsymbol{\xi}) = \begin{cases} \mathbf{0} & \text{if } \|\boldsymbol{\xi} - \sigma \mathbf{y}\|_2 \le \eta\sigma \, , \\ \left(1 - \dfrac{\eta\sigma}{\|\boldsymbol{\xi} - \sigma \mathbf{y}\|_2} \right) (\boldsymbol{\xi} - \sigma \mathbf{y}) & \text{otherwise} \, . \end{cases}$$

After these computations our primal dual algorithm reads

$$\begin{aligned} \boldsymbol{\xi}^{n+1} &= P_{F^*}(\sigma; \boldsymbol{\xi}^n + \sigma \mathbf{A} \bar{\mathbf{x}}^n) \\ &= \begin{cases} \mathbf{0} & \text{if } \|\sigma^{-1} \boldsymbol{\xi}^n + \mathbf{A} \bar{\mathbf{x}}^n - \mathbf{y}\|_2 \le \eta \, , \\ \left(1 - \dfrac{\eta\sigma}{\|\boldsymbol{\xi}^n + \sigma(\mathbf{A} \bar{\mathbf{x}}^n - \mathbf{y})\|_2} \right) (\boldsymbol{\xi}^n + \sigma(\mathbf{A} \bar{\mathbf{x}}^n - \mathbf{y})) & \text{otherwise} \, , \end{cases} \\ \mathbf{x}^{n+1} &= \mathcal{S}_\tau(\mathbf{x}^n - \tau \mathbf{A}^* \boldsymbol{\xi}^{n+1}) \, , \\ \bar{\mathbf{x}}^{n+1} &= \mathbf{x}^{n+1} + \theta(\mathbf{x}^{n+1} - \mathbf{x}^n) \, . \end{aligned}$$

(c) Consider the ℓ_1-regularized least squares problem

$$\min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{x}\|_1 + \frac{\gamma}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \, , \tag{15.24}$$

with some regularization parameter $\gamma > 0$. This problem is equivalent to (15.3) after the parameter change $\lambda = \gamma^{-1}$. It can be written in the form (15.13) with $G(\mathbf{x}) = \|\mathbf{x}\|_1$ and

$$F(\mathbf{x}) = \frac{\gamma}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \,.$$

Clearly, F is even continuous in this case. It follows either from a direct computation or from Proposition B.18(d) and (e) together with Example B.19(a) that

$$F^*(\boldsymbol{\xi}) = \mathrm{Re}(\langle \mathbf{y}, \boldsymbol{\xi}\rangle) + \frac{1}{2\gamma}\|\boldsymbol{\xi}\|_2^2 \,.$$

The dual to (15.24) is the optimization problem

$$\max_{\boldsymbol{\xi}\in\mathbb{C}^m} -\mathrm{Re}(\langle \mathbf{y}, \boldsymbol{\xi}\rangle) - \frac{1}{2\gamma}\|\boldsymbol{\xi}\|_2^2 \quad \text{subject to } \|\mathbf{A}^*\boldsymbol{\xi}\|_\infty \le 1 \,,$$

and the associated saddle point problem reads

$$\min_{\mathbf{x}\in\mathbb{C}^N} \max_{\boldsymbol{\xi}\in\mathbb{C}^m} \mathrm{Re}(\langle \mathbf{A}\mathbf{x} - \mathbf{y}, \boldsymbol{\xi}\rangle) - \frac{1}{2\gamma}\|\boldsymbol{\xi}\|_2^2 + \|\mathbf{x}\|_1 \,.$$

A straightforward calculation gives

$$P_F(\sigma;\boldsymbol{\xi}) = \frac{\sigma\gamma}{1+\sigma\gamma}\mathbf{y} + \frac{1}{\sigma\gamma+1}\boldsymbol{\xi} \,.$$

By Moreau's identity (B.15)

$$P_{F^*}(\sigma;\boldsymbol{\xi}) = \left(1 - \frac{\gamma}{\gamma+\sigma}\right)\boldsymbol{\xi} - \frac{\sigma^2}{\gamma+\sigma}\mathbf{y} \,.$$

With these relations our primal dual algorithm for the numerical solution of (15.24) is given by

$$\begin{aligned}
\boldsymbol{\xi}^{n+1} &= \left(1 - \frac{\gamma}{\gamma+\sigma}\right)\boldsymbol{\xi}^n + \frac{\sigma^2}{\gamma+\sigma}(\mathbf{A}\bar{\mathbf{x}}^n - \mathbf{y}) \,, \\
\mathbf{x}^{n+1} &= \mathcal{S}_\tau(\mathbf{x}^n - \tau\mathbf{A}^*\boldsymbol{\xi}^{n+1}) \,, \\
\bar{\mathbf{x}}^{n+1} &= \mathbf{x}^{n+1} + \theta(\mathbf{x}^{n+1} - \mathbf{x}^n) \,.
\end{aligned}$$

Let us turn to the analysis of the primal dual algorithm in the general situation. For this purpose we introduce the Lagrangian

$$L(\mathbf{x},\boldsymbol{\xi}) := \mathrm{Re}(\langle \mathbf{A}\mathbf{x}, \boldsymbol{\xi}\rangle) + G(\mathbf{x}) - F^*(\boldsymbol{\xi}) \,,$$

and the partial primal-dual gap, which for two sets $B_1 \subset \mathbb{C}^N$ and $B_2 \subset \mathbb{C}^m$ is defined as

$$\mathcal{G}_{B_1,B_2}(\mathbf{x},\boldsymbol{\xi}) := \sup_{\boldsymbol{\xi}'\in B_2} L(\mathbf{x},\boldsymbol{\xi}') - \inf_{\mathbf{x}'\in B_1} L(\mathbf{x}',\boldsymbol{\xi}) \,. \tag{15.25}$$

This is a variant of the primal dual gap defined in (B.29), which is more convenient in our context, and obviously motivated by the saddle point

problem (15.15). As soon as $B_1 \times B_2$ contains a saddle point $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\xi}})$, then $\mathcal{G}_{B_1,B_2}(\mathbf{x}, \boldsymbol{\xi}) \geq 0$ for all $\mathbf{x}, \boldsymbol{\xi}$, and $\mathcal{G}(\mathbf{x}, \boldsymbol{\xi}) = 0$ if and only if $(\mathbf{x}, \boldsymbol{\xi})$ is a saddle point of $\mathcal{G}_{B_1,B_2}$. Therefore, we can take $\mathcal{G}_{B_1,B_2}(\mathbf{x}, \boldsymbol{\xi})$ as a measure of how far the pair $(\mathbf{x}, \boldsymbol{\xi})$ is away from the optimum of the saddle point problem (15.15). But note that there is no general upper bound of the ℓ_2-distance of $(\mathbf{x}, \boldsymbol{\xi})$ to the optimum $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ by the primal dual gap.

The convergence of the primal dual algorithm is settled by the following theorem.

Theorem 15.8. *Assume that the problem* (15.15) *has a saddle point. Choose* $\theta = 1$ *and* $\sigma, \tau > 0$ *such that* $\tau\sigma\|\mathbf{A}\|_{2\to 2}^2 < 1$. *Let* $(\mathbf{x}^n, \bar{\mathbf{x}}^n, \boldsymbol{\xi}^n)$, $n \geq 0$ *be the sequence generated by* (PD$_1$), (PD$_2$), (PD$_3$).

(a) The sequence $(\mathbf{x}^n, \boldsymbol{\xi}^n)$ *converges to a saddle point* $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ *of* (15.15). *In particular,* $\mathbf{x}^n$ *converges to a minimizer of* (15.13).

(b) Define $\mathbf{x}_M := M^{-1} \sum_{n=1}^M \mathbf{x}^n$ *and* $\boldsymbol{\xi}_M := M^{-1} \sum_{n=1}^M \boldsymbol{\xi}^n$. *Let* B_1, B_2 *be bounded sets such that* $B_1 \times B_2$ *contains a saddle point* $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$. *Then*

$$\mathcal{G}_{B_1,B_2}(\mathbf{x}_M, \boldsymbol{\xi}_M) \leq \frac{D(B_1, B_2)}{M}\,, \tag{15.26}$$

where $D(B_1, B_2) := (2\tau)^{-1} \sup_{\mathbf{x}\in B_1} \|\mathbf{x} - \mathbf{x}^0\|_2^2 + (2\sigma)^{-1} \sup_{\boldsymbol{\xi}\in B_2} \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_2^2$.

Remark 15.9. Due to (15.26) one says that our algorithm converges at rate $\mathcal{O}(M^{-1})$. Inequality (15.26) also holds for sets B_1, B_2 not necessarily containing a saddle point, but in this case $\mathcal{G}_{B_1,B_2}$ has limited interpretation because it may get negative. Note that (15.26) does neither imply a rate of convergence of $\|\mathbf{x}^n - \mathbf{x}^\sharp\|_2$, nor of $\|\boldsymbol{\xi}^n - \boldsymbol{\xi}^\sharp\|_2$, but in practice the algorithm converges reasonably fast also in this sense.

We develop the proof in several steps. In order to simplify notation we introduce, for a sequence $(\mathbf{u}^n)_{n\in\mathbb{N}_0}$ (of scalars or vectors), the divided difference

$$\Delta_\tau \mathbf{u}^n := \frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{\tau}, \quad n \in \mathbb{N}\,.$$

This term can be interpreted as a discrete derivative with step size τ. Slightly abusing notation we also use Δ_τ to write related expressions such as

$$\Delta_\tau \|\mathbf{u}^{n+1}\|_2^2 = \frac{\|\mathbf{u}^{n+1}\|_2^2 - \|\mathbf{u}^n\|_2^2}{\tau}\,.$$

We have the following identities which closely resemble corresponding relations for the usual (continuous) derivative.

Lemma 15.10. *Let* $\mathbf{u}, \mathbf{u}^n \in \mathbb{C}^N$, $n \in \mathbb{N}_0$. *Then*

$$2\,\mathrm{Re}(\langle \Delta_\tau \mathbf{u}^n, \mathbf{u}^n - \mathbf{u}\rangle) = \Delta_\tau \|\mathbf{u} - \mathbf{u}^n\|_2^2 + \tau \|\Delta_\tau \mathbf{u}^n\|_2^2\,. \tag{15.27}$$

Moreover, if $\mathbf{v}^n$, $n \in \mathbb{N}_0$, *is another sequence of vectors, then the following discrete integration by parts formula holds for* $M \in \mathbb{N}$,

$$\tau \sum_{n=1}^{M} \big(\langle \Delta_\tau \mathbf{u}^n, \mathbf{v}^n \rangle + \langle \mathbf{u}^{n-1}, \Delta_\tau \mathbf{v}^n \rangle \big) = \langle \mathbf{u}^M, \mathbf{v}^M \rangle - \langle \mathbf{u}^0, \mathbf{v}^0 \rangle \, . \tag{15.28}$$

Proof. Set $\widetilde{\mathbf{u}}^n = \mathbf{u}^n - \mathbf{u}$. Then $\Delta_\tau \widetilde{\mathbf{u}}^n = \Delta_\tau \mathbf{u}^n$ and

$$\begin{aligned} 2\tau \langle \Delta_\tau \mathbf{u}^n, \mathbf{u}^n - \mathbf{u} \rangle &= 2\tau \langle \Delta_\tau \widetilde{\mathbf{u}}^n, \widetilde{\mathbf{u}}^n \rangle = 2 \langle \widetilde{\mathbf{u}}^n - \widetilde{\mathbf{u}}^{n-1}, \widetilde{\mathbf{u}}^n \rangle \\ &= \langle \widetilde{\mathbf{u}}^n - \widetilde{\mathbf{u}}^{n-1}, \widetilde{\mathbf{u}}^n - \widetilde{\mathbf{u}}^{n-1} \rangle + \langle \widetilde{\mathbf{u}}^n - \widetilde{\mathbf{u}}^{n-1}, \widetilde{\mathbf{u}}^n + \widetilde{\mathbf{u}}^{n-1} \rangle \, . \end{aligned}$$

Noting that

$$\mathrm{Re}(\langle \widetilde{\mathbf{u}}^n - \widetilde{\mathbf{u}}^{n-1}, \widetilde{\mathbf{u}}^n + \widetilde{\mathbf{u}}^{n-1} \rangle) = \|\widetilde{\mathbf{u}}^n\|_2^2 - \|\widetilde{\mathbf{u}}^{n-1}\|_2^2 = \tau \Delta_\tau \|\widetilde{\mathbf{u}}^n\|_2^2$$

completes the proof of the first statement. Next observe that

$$\Delta_\tau \langle \mathbf{u}^n, \mathbf{v}^n \rangle = \frac{\langle \mathbf{u}^n, \mathbf{v}^n \rangle - \langle \mathbf{u}^{n-1}, \mathbf{v}^{n-1} \rangle}{\tau} = \langle \Delta_\tau \mathbf{u}^n, \mathbf{v}^n \rangle + \langle \mathbf{u}^{n-1}, \Delta_\tau \mathbf{v}^n \rangle \, .$$

Summing this identity over $n = 1, \ldots, M$ and using the telescoping identity $\tau \sum_{n=1}^{M} \Delta_\tau \langle \mathbf{u}^n, \mathbf{v}^n \rangle = \langle \mathbf{u}^M, \mathbf{v}^M \rangle - \langle \mathbf{u}^0, \mathbf{v}^0 \rangle$ gives the second statement. □

Lemma 15.11. *Let* $(\mathbf{x}^n, \bar{\mathbf{x}}^n, \boldsymbol{\xi}^n)$, $n \geq 0$, *be the sequence generated by* (PD$_1$), (PD$_2$), (PD$_3$), *and let* $\mathbf{x} \in \mathbb{C}^N$, $\boldsymbol{\xi} \in \mathbb{C}^m$ *be arbitrary. Then, for* $n \in \mathbb{N}$,

$$\begin{aligned} &\frac{1}{2} \Delta_\sigma \|\boldsymbol{\xi} - \boldsymbol{\xi}^n\|_2^2 + \frac{1}{2} \Delta_\tau \|\mathbf{x} - \mathbf{x}^n\|_2^2 + \frac{\sigma}{2} \|\Delta_\sigma \boldsymbol{\xi}^n\|_2^2 + \frac{\tau}{2} \|\Delta_\tau \mathbf{x}^n\|_2^2 \\ &\leq L(\mathbf{x}, \boldsymbol{\xi}^n) - L(\mathbf{x}^n, \boldsymbol{\xi}) + \mathrm{Re}(\langle \mathbf{A}(\mathbf{x}^n - \bar{\mathbf{x}}^{n-1}), \boldsymbol{\xi} - \boldsymbol{\xi}^n \rangle) \, . \end{aligned} \tag{15.29}$$

Proof. It follows from the characterization of the proximal mapping in Proposition B.23 that the iterates satisfy the relations (replacing $n+1$ by n)

$$\begin{aligned} \boldsymbol{\xi}^{n-1} + \sigma \mathbf{A} \bar{\mathbf{x}}^{n-1} &\in \boldsymbol{\xi}^n + \sigma \partial F^*(\boldsymbol{\xi}^n) \, , \\ \mathbf{x}^{n-1} - \tau \mathbf{A}^* \boldsymbol{\xi}^n &\in \mathbf{x}^n + \tau \partial G(\mathbf{x}^n) \, , \end{aligned}$$

where ∂F^* and ∂G are the subdifferentials of F^* and G. By Definition (B.11) of the subdifferential (and recalling that inner products have to be replaced by $\mathrm{Re}(\langle \cdot, \cdot \rangle)$ when passing from the real to the complex case) this implies

$$\begin{aligned} \mathrm{Re}(\langle -\boldsymbol{\xi}^n + \boldsymbol{\xi}^{n-1} + \sigma \mathbf{A} \bar{\mathbf{x}}^{n-1}, \boldsymbol{\xi} - \boldsymbol{\xi}^n \rangle) &\leq \sigma F^*(\boldsymbol{\xi}) - \sigma F^*(\boldsymbol{\xi}^n) \, , \\ \mathrm{Re}(\langle -\mathbf{x}^n + \mathbf{x}^{n-1} - \tau \mathbf{A}^* \boldsymbol{\xi}^n, \mathbf{x} - \mathbf{x}^n \rangle) &\leq \tau G(\mathbf{x}) - \tau G(\mathbf{x}^n) \, , \end{aligned}$$

or, with our definition of the divided difference,

$$\begin{aligned} \mathrm{Re}(\langle \Delta_\sigma \boldsymbol{\xi}^n, \boldsymbol{\xi}^n - \boldsymbol{\xi} \rangle) + \mathrm{Re}(\langle \mathbf{A} \bar{\mathbf{x}}^{n-1}, \boldsymbol{\xi} - \boldsymbol{\xi}^n \rangle) &\leq F^*(\boldsymbol{\xi}) - F^*(\boldsymbol{\xi}^n) \, , \\ \mathrm{Re}(\langle \Delta_\tau \mathbf{x}^n, \mathbf{x}^n - \mathbf{x} \rangle) - \mathrm{Re}(\langle \mathbf{A}(\mathbf{x} - \mathbf{x}^n), \boldsymbol{\xi}^n \rangle) &\leq G(\mathbf{x}) - G(\mathbf{x}^n) \, . \end{aligned}$$

Summing both inequalities and exploiting (15.27) yields

$$\begin{aligned}
&\frac{1}{2}\Delta_\sigma\|\boldsymbol{\xi}-\boldsymbol{\xi}^n\|_2^2+\frac{1}{2}\Delta_\tau\|\mathbf{x}-\mathbf{x}^n\|_2^2+\frac{\sigma}{2}\|\Delta_\sigma\boldsymbol{\xi}^n\|_2^2+\frac{\tau}{2}\|\Delta_\tau\mathbf{x}^n\|_2^2\\
&\le F^*(\boldsymbol{\xi})-F^*(\boldsymbol{\xi}^n)+G(\mathbf{x})-G(\mathbf{x}^n)\\
&+\operatorname{Re}(\langle\mathbf{A}(\mathbf{x}-\mathbf{x}^n),\boldsymbol{\xi}^n\rangle)-\operatorname{Re}(\langle\mathbf{A}\bar{\mathbf{x}}^{n-1},\boldsymbol{\xi}-\boldsymbol{\xi}^n\rangle)\\
&=(\operatorname{Re}(\langle\mathbf{A}\mathbf{x},\boldsymbol{\xi}^n\rangle)+G(\mathbf{x})-F^*(\boldsymbol{\xi}^n))-(\operatorname{Re}(\langle\mathbf{A}\mathbf{x}^n,\boldsymbol{\xi}\rangle)+G(\mathbf{x}^n)-F^*(\boldsymbol{\xi}))\\
&+\operatorname{Re}(\langle\mathbf{A}(\mathbf{x}^n-\bar{\mathbf{x}}^{n-1}),\boldsymbol{\xi}-\boldsymbol{\xi}^n\rangle)\,.
\end{aligned}$$

This finishes the proof. □

Remark 15.12. Inequality (15.29) suggests that one would ideally set $\bar{\mathbf{x}}^{n-1}=\mathbf{x}^n$. However, this would lead to an implicit scheme, where the equations defining the iterations become as hard to solve as the original problem.

Lemma 15.13. *Let $(\mathbf{x}^n,\bar{\mathbf{x}}^n,\boldsymbol{\xi}^n)$, $n\ge 0$ be the sequence generated by* (PD$_1$), (PD$_2$), (PD$_3$) *with the parameter choice $\theta=1$, and let $\mathbf{x}\in\mathbb{C}^N$, $\boldsymbol{\xi}\in\mathbb{C}^m$ be arbitrary. Then, for $M\in\mathbb{N}$,*

$$\begin{aligned}
&\sum_{n=1}^{M}(L(\mathbf{x}^n,\boldsymbol{\xi})-L(\mathbf{x},\boldsymbol{\xi}^n))+\frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}^M\|_2^2+\frac{1-\sigma\tau\|\mathbf{A}\|_{2\to2}^2}{2\sigma}\|\boldsymbol{\xi}-\boldsymbol{\xi}^M\|_2^2\\
&+\frac{1-\sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to2}}{2\tau}\sum_{n=1}^{M-1}\|\mathbf{x}^n-\mathbf{x}^{n-1}\|_2^2+\frac{1-\sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to2}}{2\sigma}\sum_{n=1}^{M}\|\boldsymbol{\xi}^n-\boldsymbol{\xi}^{n-1}\|_2^2\\
&\le\frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}^0\|_2^2+\frac{1}{2\sigma}\|\boldsymbol{\xi}-\boldsymbol{\xi}^0\|_2^2\,. \qquad (15.30)
\end{aligned}$$

Proof. First note that $\mathbf{x}^n-\bar{\mathbf{x}}^{n-1}=\mathbf{x}^n-\mathbf{x}^{n-1}-(\mathbf{x}^{n-1}-\mathbf{x}^{n-2})=\tau^2\Delta_\tau\Delta_\tau\mathbf{x}^n=:\tau^2\Delta_\tau^2\mathbf{x}^n$ for $n\ge 2$, and the formula extends to $n=1$ when setting $\mathbf{x}^{-1}=\mathbf{x}^0$ because by definition $\bar{\mathbf{x}}^0=\mathbf{x}^0$. In particular $\Delta_\tau\mathbf{x}^0=0$. Summing inequality (15.29) from $n=1$ to $n=M$ gives

$$\begin{aligned}
&\frac{1}{2\sigma}(\|\boldsymbol{\xi}-\boldsymbol{\xi}^M\|_2^2-\|\boldsymbol{\xi}-\boldsymbol{\xi}^0\|_2^2)+\frac{1}{2\tau}(\|\mathbf{x}-\mathbf{x}^M\|_2^2-\|\mathbf{x}-\mathbf{x}^0\|_2^2)\\
&+\frac{1}{2\sigma}\sum_{n=1}^{M}\|\boldsymbol{\xi}^n-\boldsymbol{\xi}^{n-1}\|_2^2+\frac{1}{2\tau}\sum_{n=1}^{M}\|\mathbf{x}^n-\mathbf{x}^{n-1}\|_2^2\\
&\le\sum_{n=1}^{M}(L(\mathbf{x},\boldsymbol{\xi}^n)-L(\mathbf{x}^n,\boldsymbol{\xi}))+\tau^2\sum_{n=1}^{M}\operatorname{Re}(\langle\mathbf{A}\Delta_\tau^2\mathbf{x}^n,\boldsymbol{\xi}-\boldsymbol{\xi}^n\rangle)\,. \qquad (15.31)
\end{aligned}$$

Next we exploit the discrete integration by parts formula (15.28) and $\Delta_\tau\mathbf{x}^0=\mathbf{0}$ to reach

$$
\begin{aligned}
&\tau^2 \sum_{n=1}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau^2 \mathbf{x}^n, \boldsymbol{\xi} - \boldsymbol{\xi}^n \rangle) \\
&= \tau^2 \sum_{n=1}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^{n-1}, \Delta_\tau \boldsymbol{\xi}^n \rangle) + \tau \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^M, \boldsymbol{\xi} - \boldsymbol{\xi}^M \rangle) \\
&= \sigma\tau \sum_{n=1}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^{n-1}, \Delta_\sigma \boldsymbol{\xi}^n \rangle) + \tau \operatorname{Re}(\langle \Delta_\tau \mathbf{x}^M, \mathbf{A}^*(\boldsymbol{\xi} - \boldsymbol{\xi}^M) \rangle) \, .
\end{aligned}
$$

Since $2ab \leq \alpha a^2 + b^2/\alpha$ for positive a, b, α we have

$$
\begin{aligned}
&\tau\sigma \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^{n-1}, \Delta_\sigma \boldsymbol{\xi}^n \rangle) \leq \tau\sigma \|\mathbf{A}\|_{2\to 2} \|\Delta_\tau \mathbf{x}^{n-1}\|_2 \|\Delta_\sigma \boldsymbol{\xi}^n\|_2 \\
&\leq \frac{\tau\sigma\|\mathbf{A}\|_{2\to 2}}{2} \left(\alpha \|\Delta_\tau \mathbf{x}^{n-1}\|_2^2 + \alpha^{-1} \|\Delta_\sigma \boldsymbol{\xi}^n\|_2^2 \right) \\
&\leq \frac{\sigma\alpha\|\mathbf{A}\|_{2\to 2}}{2\tau} \|\mathbf{x}^{n-1} - \mathbf{x}^{n-2}\|_2^2 + \frac{\tau\|\mathbf{A}\|_{2\to 2}}{2\alpha\sigma} \|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2^2 \, .
\end{aligned}
$$

We choose $\alpha = \sqrt{\tau/\sigma}$ to get

$$
\begin{aligned}
&\tau\sigma \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^{n-1}, \Delta_\sigma \boldsymbol{\xi}^n \rangle) \\
&\leq \frac{\sqrt{\tau\sigma}\|\mathbf{A}\|_{2\to 2}}{2\tau} \|\mathbf{x}^{n-1} - \mathbf{x}^{n-2}\|_2^2 + \frac{\sqrt{\tau\sigma}\|\mathbf{A}\|_{2\to 2}}{2\sigma} \|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2^2 \, .
\end{aligned} \tag{15.32}
$$

Similarly, we have

$$
\begin{aligned}
&\tau \operatorname{Re}(\langle \Delta_\tau \mathbf{x}^M, \mathbf{A}^*(\boldsymbol{\xi} - \boldsymbol{\xi}^M) \rangle) \leq \frac{\tau}{2} \left(\|\Delta_\tau \mathbf{x}^M\|_2^2 + \|\mathbf{A}\|_{2\to 2}^2 \|\boldsymbol{\xi} - \boldsymbol{\xi}^M\|_2^2 \right) \\
&= \frac{1}{2\tau} \|\mathbf{x}^M - \mathbf{x}^{M-1}\|_2^2 + \frac{\tau\sigma\|\mathbf{A}\|_{2\to 2}^2}{2\sigma} \|\boldsymbol{\xi} - \boldsymbol{\xi}^M\|_2^2 \, .
\end{aligned}
$$

Plugging these estimates into the second term in (15.31) and using that $\mathbf{x}^{-1} = \mathbf{x}^0$ yields

$$
\begin{aligned}
&\tau^2 \sum_{n=1}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau^2 \mathbf{x}^n, \boldsymbol{\xi} - \boldsymbol{\xi}^n \rangle) \\
&\leq \frac{\sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2}}{2\tau} \sum_{n=1}^{M-1} \|\mathbf{x}^n - \mathbf{x}^{n-1}\|_2^2 + \frac{\sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2}}{2\sigma} \sum_{n=1}^{M} \|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2^2 \\
&+ \frac{1}{2\tau} \|\mathbf{x}^M - \mathbf{x}^{M-1}\|_2^2 + \frac{\tau\sigma\|\mathbf{A}\|_{2\to 2}^2}{2\sigma} \|\boldsymbol{\xi} - \boldsymbol{\xi}^M\|_2^2 \, .
\end{aligned}
$$

Together with inequality (15.31) we arrive at the claim. □

Corollary 15.14. *Let* $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ *be a primal dual optimum, that is, a saddle point of* (15.15)*. Then the iterates of the primal dual algorithm with* $\theta = 1$ *and* $\sigma\tau\|\mathbf{A}\|_{2\to 2} < 1$ *satisfy*

$$\frac{1}{2\sigma}\|\boldsymbol{\xi}^\sharp - \boldsymbol{\xi}^M\|_2^2 + \frac{1}{2\tau}\|\mathbf{x}^\sharp - \mathbf{x}^M\|_2^2 \le C\left(\frac{1}{2\sigma}\|\boldsymbol{\xi}^\sharp - \boldsymbol{\xi}^0\|_2^2 + \frac{1}{2\tau}\|\mathbf{x}^\sharp - \mathbf{x}^0\|_2^2\right),$$

where $C = (1-\sigma\tau\|\mathbf{A}\|_{2\to 2}^2)^{-1}$. *In particular, the iterates* $(\mathbf{x}^n, \boldsymbol{\xi}^n)$ *are bounded.*

Proof. For a saddle point $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ the terms $L(\mathbf{x}^n, \boldsymbol{\xi}^\sharp) - L(\mathbf{x}^\sharp, \boldsymbol{\xi}^n)$ are nonnegative so that all terms on the left hand side of (15.30) are positive. In particular,

$$\frac{1}{2\tau}\|\mathbf{x} - \mathbf{x}^M\|_2^2 + \frac{1-\sigma\tau\|\mathbf{A}\|_{2\to 2}^2}{2\sigma}\|\boldsymbol{\xi} - \boldsymbol{\xi}^M\|_2^2 \le \frac{1}{2\tau}\|\mathbf{x} - \mathbf{x}^0\|_2^2 + \frac{1}{2\sigma}\|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_2^2 .$$

This yields the claim. □

We are now in the position to complete the convergence proof for our primal dual algorithm.

Proof (of Theorem 15.8). We start with the proof of (b), where $\mathbf{x}_M = M^{-1}\sum_{n=1}^M \mathbf{x}_n$ and $\boldsymbol{\xi}_M = M^{-1}\sum_{n=1}^M \boldsymbol{\xi}_n$. Convexity of G and F^* together with (15.30) yield, for arbitrary $(\mathbf{x}, \boldsymbol{\xi})$,

$$\begin{aligned}
&L(\mathbf{x}_M, \boldsymbol{\xi}) - L(\mathbf{x}, \boldsymbol{\xi}_M) \\
&= (\mathrm{Re}(\langle \mathbf{A}\mathbf{x}_M, \boldsymbol{\xi}\rangle) + G(\mathbf{x}_M) - F^*(\boldsymbol{\xi})) - (\mathrm{Re}(\langle \mathbf{A}\mathbf{x}, \boldsymbol{\xi}_M\rangle) + G(\mathbf{x}) - F^*(\boldsymbol{\xi}_M)) \\
&\le \frac{1}{M}\sum_{n=1}^M (\mathrm{Re}(\langle \mathbf{A}\mathbf{x}^n, \boldsymbol{\xi}\rangle) + G(\mathbf{x}^n) - F^*(\boldsymbol{\xi})) \\
&\quad - \frac{1}{M}\sum_{n=1}^M (\mathrm{Re}(\langle \mathbf{A}\mathbf{x}, \boldsymbol{\xi}^n\rangle) + G(\mathbf{x}) - F^*(\boldsymbol{\xi}^n)) \\
&\le \frac{1}{M}\left(\frac{1}{2\tau}\|\mathbf{x} - \mathbf{x}^0\|_2^2 + \frac{1}{2\sigma}\|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_2^2\right) .
\end{aligned}$$

Taking the supremum over all $(\mathbf{x}, \boldsymbol{\xi}) \in B_1 \times B_2$ establishes (15.26).

For (a) we first note that the boundedness of the sequence $(\mathbf{x}^n, \boldsymbol{\xi}^n)$ established in Corollary 15.14 implies that there exists a convergent subsequence, say $(\mathbf{x}^{n_k}, \boldsymbol{\xi}^{n_k})_k \to (\mathbf{x}^\circ, \boldsymbol{\xi}^\circ)$ as $k \to \infty$. Choosing $(\mathbf{x}, \boldsymbol{\xi})$ to be a saddle point $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ in (15.30) makes all terms positive, and we conclude in particular that

$$\frac{1-\sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2}}{2\sigma}\sum_{n=1}^{M-1}\|\mathbf{x}^n - \mathbf{x}^{n-1}\|_2^2 \le \frac{1}{2\tau}\|\mathbf{x} - \mathbf{x}^0\|_2^2 + \frac{1}{2\sigma}\|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_2^2 .$$

Since the right hand side is independent of M, and since $\sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2} < 1$ we conclude that $\|\mathbf{x}^n - \mathbf{x}^{n-1}\|_2 \to 0$ as $n \to \infty$. Similarly, $\lim_{n\to\infty}\|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2 = 0$. In particular, also $(\mathbf{x}^{n_k-1}, \boldsymbol{\xi}^{n_k-1})$ converges to $(\mathbf{x}^\circ, \boldsymbol{\xi}^\circ)$. It follows that $(\mathbf{x}^\circ, \boldsymbol{\xi}^\circ)$ is a fixed point of the primal dual algorithm, so that by Proposition 15.6 it is a primal dual optimal point (or saddle point).

We choose $(\mathbf{x}, \boldsymbol{\xi}) = (\mathbf{x}^\circ, \boldsymbol{\xi}^\circ)$ in (15.29) so that $L(\mathbf{x}^n, \boldsymbol{\xi}^\circ) - L(\mathbf{x}^\circ, \boldsymbol{\xi}^n) \geq 0$. We proceed now similarly as in the proof of Lemma 15.13. Summing (15.29) from $n = n_k$ to $n = M > n_k$ results in

$$\begin{aligned}
&\frac{1}{2\sigma}(\|\boldsymbol{\xi}^\circ - \boldsymbol{\xi}^M\|_2^2 - \|\boldsymbol{\xi}^\circ - \boldsymbol{\xi}^{n_k}\|_2^2) + \frac{1}{2\tau}(\|\mathbf{x}^\circ - \mathbf{x}^M\|_2^2 - \|\mathbf{x}^\circ - \mathbf{x}^{n_k}\|_2^2) \\
&+\frac{1}{2\sigma}\sum_{n=n_k}^{M} \|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2^2 + \frac{1}{2\tau}\sum_{n=n_k}^{M} \|\mathbf{x}^n - \mathbf{x}^{n-1}\|_2^2 \\
&\leq \tau^2 \sum_{n=n_k}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau^2 \mathbf{x}^n, \boldsymbol{\xi}^\circ - \boldsymbol{\xi}^n \rangle) \,. \qquad (15.33)
\end{aligned}$$

Discrete integration by parts (15.28) yields

$$\begin{aligned}
&\tau^2 \sum_{n=n_k}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau^2 \mathbf{x}^n, \boldsymbol{\xi}^\circ - \boldsymbol{\xi}^n \rangle) \\
&= \sigma\tau \sum_{n=n_k}^{M} \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^{n-1}, \Delta_\sigma \boldsymbol{\xi}^n \rangle) + \tau \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^M, \boldsymbol{\xi}^\circ - \boldsymbol{\xi}^M \rangle) \\
&- \tau \operatorname{Re}(\langle \mathbf{A}\Delta_\tau \mathbf{x}^{n_k - 1}, \boldsymbol{\xi}^\circ - \boldsymbol{\xi}^{n_k} \rangle) \,.
\end{aligned}$$

Inequality (15.32) therefore implies

$$\begin{aligned}
&\frac{1}{2\sigma}\|\boldsymbol{\xi}^\circ - \boldsymbol{\xi}^M\|_2^2 + \frac{1}{2\tau}\|\mathbf{x}^\circ - \mathbf{x}^M\|_2^2 + \frac{1 - \sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2}}{2\sigma} \sum_{n=n_k}^{M} \|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2^2 \\
&+ \frac{1 - \sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2}}{2\tau} \sum_{n=n_k}^{M-1} \|\mathbf{x}^n - \mathbf{x}^{n-1}\|_2^2 \\
&+ \frac{1}{2\tau}\left(\|\mathbf{x}^M - \mathbf{x}^{M-1}\|_2^2 - \sqrt{\sigma\tau}\|\mathbf{A}\|_{2\to 2}\|\mathbf{x}^{n_k - 1} - \mathbf{x}^{n_k - 2}\|_2^2\right) \\
&+ \operatorname{Re}(\langle \mathbf{A}(\mathbf{x}^M - \mathbf{x}^{M-1}), \boldsymbol{\xi}^\circ - \boldsymbol{\xi}^M \rangle) - \operatorname{Re}(\langle \mathbf{A}(\mathbf{x}^{n_k - 1} - \mathbf{x}^{n_k - 2}), \boldsymbol{\xi}^\circ - \boldsymbol{\xi}^{n_k} \rangle) \\
&\leq \frac{1}{2\sigma}\|\boldsymbol{\xi}^\circ - \boldsymbol{\xi}^{n_k}\|_2^2 + \frac{1}{2\tau}\|\mathbf{x}^\circ - \mathbf{x}^{n_k}\|_2^2 \,.
\end{aligned}$$

Since $\lim_{n\to\infty} \|\mathbf{x}^n - \mathbf{x}^{n-1}\|_2 = \lim_{n\to\infty} \|\boldsymbol{\xi}^n - \boldsymbol{\xi}^{n-1}\|_2 = 0$ and $\lim_{k\to\infty} \|\mathbf{x}^\circ - \mathbf{x}^{n_k}\|_2 = \lim_{k\to\infty} \|\boldsymbol{\xi}^\circ - \boldsymbol{\xi}^{n_k}\|_2 = 0$ it follows that $\lim_{M\to\infty} \|\mathbf{x}^\circ - \mathbf{x}^M\|_2 = \lim_{M\to\infty} \|\boldsymbol{\xi}^\circ - \boldsymbol{\xi}^M\|_2 = 0$. We have established the claim. □

15.3 Iteratively Reweighted Least Squares

We now turn to an iterative algorithm that serves as a proxy for ℓ_1-minimization. It does not always compute the solution of an ℓ_1-minimization problem, but

provides similar error estimates under the null space property as the ones for ℓ_1-minimization in Chapter 4.

The starting point is the trivial observation that $|t| = \frac{|t|^2}{|t|}$ for $t \neq 0$. Therefore, an ℓ_1-minimization can be recast into a weighted ℓ_2-minimization in the following sense. Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ with $m \leq N$. If $\mathbf{x}^\sharp$ is a minimizer of

$$\min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} \tag{15.34}$$

and $x_j^\sharp \neq 0$ for all $j \in [N]$, then $\mathbf{x}^\sharp$ is also a minimizer of the weighted ℓ_2-problem

$$\min_{\mathbf{x} \in \mathbb{C}^N} \sum_{j=1}^N |x_j|^2 |x_j^\sharp|^{-1} \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} \, .$$

The advantage of this reformulation consists in the fact that minimizing the smooth quadratic function $|t|^2$ is an easier task than the minimization of the nonsmooth function $|t|$. However, the obvious drawbacks are that we neither dispose of $\mathbf{x}^\sharp$ a priori (this is the vector we would like to compute!) nor we can expect that $x_j^\sharp \neq 0$ for all $j = 1, \ldots, N$, since one expects s-sparse solutions. In fact by Theorem 3.1, the ℓ_1-minimizer is actually always m-sparse in the real case provided it is unique.

Nevertheless, the above observation motivates to iteratively solve weighted ℓ_1-minimization problems, where the weight in the next iterate is computed from the solution of the weighted least squares problem of the previous step.

Key to the formulation and analysis of the algorithm is the functional

$$\mathcal{J}(\mathbf{x}, \mathbf{w}, \varepsilon) = \frac{1}{2} \left[\sum_{j=1}^N |x_j|^2 w_j + \sum_{j=1}^N (\varepsilon^2 w_j + w_j^{-1}) \right] , \tag{15.35}$$

where $\mathbf{x} \in \mathbb{C}^N$, $\varepsilon \geq 0$ and $\mathbf{w} \in \mathbb{R}^N$ is a positive weight vector, $w_j > 0$ for all $j \in [N]$. The formulation of our algorithm below uses the nonincreasing rearrangement $(\mathbf{x}^n)^* \in \mathbb{R}^N$ of the iterate $\mathbf{x}^n \in \mathbb{C}^N$, see Definition 2.4.

Iteratively reweighted least squares (IRLS)

Parameter: $\gamma > 0$, $s \in [N]$
Initialization: $\mathbf{w}^0 = (1, 1, \ldots, 1)^T \in \mathbb{R}^N$, $\varepsilon_0 := 1$.
Iteration: repeat until $\varepsilon_n = 0$ or stopping criterion is met at $n = \bar{n}$:

$$\mathbf{x}^{n+1} := \arg\min_{\mathbf{z}\in\mathbb{C}^N} \mathcal{J}(\mathbf{z}, \mathbf{w}^n, \varepsilon_n) \quad \text{subject to } \mathbf{Az} = \mathbf{y} \,, \tag{IRLS$_1$}$$

$$\varepsilon_{n+1} := \min\{\varepsilon_n, \gamma\,(\mathbf{x}^{n+1})^*_{s+1}\} \,, \tag{IRLS$_2$}$$

$$\mathbf{w}^{n+1} := \arg\min_{\mathbf{w}>0} \mathcal{J}(\mathbf{x}^{n+1}, \mathbf{w}, \varepsilon_{n+1}) \,. \tag{IRLS$_3$}$$

Output: A solution $\mathbf{x}^\sharp = \mathbf{x}^{\bar{n}}$ of $\mathbf{Ax} = \mathbf{y}$, approximating the sparsest solution.

Since $\mathbf{w}^n$ and ε_n are fixed in the minimization problem in (IRLS$_1$), the second sum in the definition (15.35) of $\mathcal{J}$ is constant, so that $\mathbf{x}^{n+1}$ is the minimizer of the weighted least squares problem

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_{2,\mathbf{w}^n} = \left(\sum_{j=1}^N |z_j|^2 w_j^n\right)^{1/2} \qquad \text{subject to } \mathbf{Az} = \mathbf{y} \,.$$

By (A.35) the minimizer $\mathbf{x}^{n+1}$ is given explicitly by the formula $\mathbf{x}^{n+1} = \mathbf{D}_{\mathbf{w}^n}^{-1/2}(\mathbf{AD}_{\mathbf{w}^n}^{-1/2})^\dagger \mathbf{y}$, where $(\mathbf{AD}_{\mathbf{w}^n}^{-1/2})^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{AD}_{\mathbf{w}^n}^{-1/2}$, see Definition A.18, and $\mathbf{D}_{\mathbf{w}^n} = \operatorname{diag}(\mathbf{w}^n) = \operatorname{diag}(w_j^n, j \in [N])$ is the diagonal matrix determined by the weight $\mathbf{w}^n$. If $\mathbf{A}$ has full rank (which will usually be the case in the setting of compressive sensing) then also $\mathbf{AD}_{\mathbf{w}^n}$ has full rank by positivity of the weight $\mathbf{w}^n$ and (A.36) yields

$$\mathbf{x}^{n+1} = \mathbf{D}_{\mathbf{w}^n}^{-1} \mathbf{A}^* (\mathbf{AD}_{\mathbf{w}^n}^{-1} \mathbf{A}^*)^{-1} \mathbf{y}$$

Clearly, $\mathbf{D}_{\mathbf{w}^n}^{-1} = \operatorname{diag}(1/w_j^n, j \in [N])$. In particular, we can write

$$\mathbf{x}^{n+1} = \mathbf{D}_{\mathbf{w}^n}^{-1} \mathbf{A}^* \mathbf{v} \quad \text{where} \quad \mathbf{AD}_{\mathbf{w}^n}^{-1} \mathbf{A}^* \mathbf{v} = \mathbf{y} \tag{15.36}$$

so that computing $\mathbf{x}^{n+1}$ involves solving the linear system for the vector $\mathbf{v}$ above. We refer to Appendix A.3 for more basic information on least squares and weighted least squares problems.

Remark 15.15. If $\mathbf{A}$ possesses a fast matrix multiplication algorithm as in situations described in Chapter 12 then it is usually not advisable to solve the linear system of equations in (15.36) by a direct method such as Gaussian elimination because such a method cannot exploit fast forward transforms. Instead, one preferably works with iterative methods such as conjugate gradients, which use only forward applications of $\mathbf{A}$ and $\mathbf{A}^*$ in order to approximately solve for $\mathbf{x}^{n+1}$. It is then, however, a subtle problem to determine the

accuracy required in each step in order to ensure overall convergence, see also the Notes section.

The minimization in (IRLS_3) can be performed explicitly,

$$w_j^{n+1} = \frac{1}{\sqrt{|x_j^{n+1}|^2 + \varepsilon_{n+1}^2}}, \quad j \in [N] \, . \tag{15.37}$$

This formula also illustrates the role of ε_n. While in the naive definition $w_j^{n+1} = |x_j^{n+1}|^{-1}$ motivated above, the weight may grow unboundedly when x_j^{n+1} approaches zero, the introduction of ε_{n+1} regularizes $\mathbf{w}^{n+1}$; in particular, $\|\mathbf{w}^{n+1}\|_\infty \le \varepsilon_{n+1}^{-1}$. Nevertheless, during the iterations we aim at approaching the ℓ_1-minimizer, which requires that ε_n decreases with n. The choice (IRLS_2) indeed ensures that ε_n does not grow, and when $\mathbf{x}^n$ tends to a s-sparse vector then ε_n tends to zero. In particular, the parameter s of the algorithm controls the desired sparsity.

We note that other update rules for ε_{n+1}, as well as for the weight $\mathbf{w}^{n+1}$ are possible as well, see also the Notes section.

The formulation of the main result on convergence of the algorithm requires to introduce, for $\varepsilon > 0$, the auxiliary functional

$$F_\varepsilon(\mathbf{x}) := \sum_{\ell=1}^N \sqrt{x_\ell^2 + \varepsilon^2} \tag{15.38}$$

and the optimization problem

$$\min_{\mathbf{z} \in \mathbb{C}^N} F_\varepsilon(\mathbf{z}) \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y} \, . \tag{15.39}$$

We denote by $\mathbf{x}^{(\varepsilon)}$ its minimizer, which is unique by strict convexity of F_ε.

The recovery theorem for iteratively reweighted least squares below is based on the notion of stable null space property in Definition 4.10, and closely resembles the corresponding statements for ℓ_1-minimization. Recall that it is proven directly in Section (9.3) that Gaussian random matrices satisfy the null space property with high probability under appropriate conditions. Also, by Theorem 6.12, the restricted isometry property implies the null space property, so that the various other matrices described in this book also satisfy the stable null space property under appropriate conditions.

Theorem 15.16. *Assume that $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the stable null space property of order s and parameter $\rho < 1$. Let $\mathbf{x} \in \mathbb{C}^N$ and form $\mathbf{y} = \mathbf{A}\mathbf{x}$.*

Consider the IRLS algorithm with parameters s and $\gamma = 1/N$. Then the sequence $(\mathbf{x}^n)_n$ converges to a vector $\mathbf{x}^\sharp \in \mathbb{C}^N$ as $n \to \infty$, whose non-increasing rearrangement $(\mathbf{x}^\sharp)^*$ satisfies $(\mathbf{x}^\sharp)_s^* = N \lim_{n\to\infty} \varepsilon_n$. Moreover, the following holds:

(a) If $\lim_{n\to\infty} \varepsilon_n = 0$*, then* $\mathbf{x}^\sharp$ *is* s*-sparse and a solution of the* ℓ_1*-minimization problem* (15.34)*. If also* $\mathbf{x}$ *is* s*-sparse then* $\mathbf{x} = \mathbf{x}^\sharp$*. More generally,*

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \le \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x})_1 \,. \tag{15.40}$$

(b) If $\varepsilon := \lim_{n\to\infty} \varepsilon_n > 0$ *then* $\mathbf{x}^\sharp = \mathbf{x}^{(\varepsilon)}$*. In this case, if* ρ *satisfies the tighter bound* $\rho < 1 - \frac{2}{s+2}$*, then for any* $\widetilde{s} < s - \frac{2\rho}{1-\rho}$

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \le c\sigma_{\widetilde{s}}(\mathbf{x})_1 \quad \text{with } c = \frac{2(1+\rho)}{1-\rho} \frac{s - \widetilde{s} + 3/2}{s - \widetilde{s} - \frac{2\rho}{1-\rho}} \,. \tag{15.41}$$

In particular, $\lim_{n\to\infty} \varepsilon_n > 0$ *is excluded if* $\mathbf{x}$ *is* $\widetilde{s}$*-sparse for* $\widetilde{s} < s - \frac{2\rho}{1-\rho}$*.*

Remark 15.17. (a) The specific parameter choice $\gamma = 1/N$ allows to prove the stated theorem, but larger choices such as $\gamma = 1$ seem to be favorable in practice. However, a theoretical guarantee for convergence for such different choices is presently not available.

(b) The constant c above is usually very reasonable. If for instance, $\rho \le 1/2$ and $\widetilde{s} \le s - 4$ then $c \le 16.5$.

We develop the proof of this theorem in several steps. We start with some properties of the iterates $\mathbf{x}^n, \mathbf{w}^n$.

Lemma 15.18. *Let* $\mathbf{x}^n$*,* $\mathbf{w}^n$ *be the iterates of the IRLS algorithm. Then, for* $n \in \mathbb{N}$*,*

$$\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) = \sum_{j=1}^N \sqrt{|x_j^n|^2 + \varepsilon_n^2} = F_{\varepsilon_n}(\mathbf{x}^n) \,, \tag{15.42}$$

and

$$\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) \le \mathcal{J}(\mathbf{x}^n, \mathbf{w}^{n-1}, \varepsilon_n) \le \mathcal{J}(\mathbf{x}^n, \mathbf{w}^{n-1}, \varepsilon_{n-1}) \tag{15.43}$$

$$\le \mathcal{J}(\mathbf{x}^{n-1}, \mathbf{w}^{n-1}, \varepsilon_{n-1}) \,. \tag{15.44}$$

Moreover, the sequence $\mathbf{x}^n$ *is bounded,*

$$\|\mathbf{x}^n\|_1 \le \mathcal{J}(\mathbf{x}^1, \mathbf{w}^0, \varepsilon_0) =: B \,, \quad n \in \mathbb{N}, \tag{15.45}$$

and the weights $\mathbf{w}^n$ *are bounded from below,*

$$w_j^n \ge B^{-1} \,, \quad j \in [N],\ n \in \mathbb{N} \,. \tag{15.46}$$

Proof. The relation (15.42) is derived from (15.37) by an easy calculation.

The first inequality in (15.43) follows from the minimization property defining $\mathbf{w}^n$, the second from $\varepsilon_{n+1} \le \varepsilon_n$, and the inequality (15.44) is a consequence of the minimization property that defines $\mathbf{x}^n$.

It follows from (15.42) that

$$\|\mathbf{x}^n\|_1 \leq \sum_{j=1}^{N} \sqrt{|x_j^n|^2 + \varepsilon_n^2} = F_{\varepsilon_n}(\mathbf{x}^n) = \mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) \leq \mathcal{J}(\mathbf{x}^1, \mathbf{w}^0, \varepsilon_0) = B \,,$$

where the last inequality uses (15.43). This establishes (15.45). Finally,

$$(w_j^n)^{-1} = \sqrt{|x_j^n|^2 + \varepsilon_n^2} \leq \mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) \leq B \,, \quad j \in [N] \,,$$

yields (15.46). □

Note that (15.44) tells us that each iteration decreases the value of the functional $\mathcal{J}$. As the next step, we establish that the difference of subsequent iterates converges to zero.

Lemma 15.19. *The iterates of the IRLS algorithm satisfy*

$$\sum_{j=1}^{\infty} \|\mathbf{x}^{n+1} - \mathbf{x}^n\|_2^2 \leq 2B^2 \,,$$

where B is the constant in (15.45)*. Consequently,* $\lim_{n\to\infty}(\mathbf{x}^{n+1} - \mathbf{x}^n) = \mathbf{0}$.

Proof. The monotonicity property in (15.43) yields

$$\begin{aligned}
&2\left(\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) - \mathcal{J}(\mathbf{x}^{n+1}, \mathbf{w}^{n+1}, \varepsilon_{n+1})\right) \\
&\geq 2\left(\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) - \mathcal{J}(\mathbf{x}^{n+1}, \mathbf{w}^n, \varepsilon_n)\right) \\
&= \sum_{j=1}^{N} (|x_j^n|^2 - |x_j^{n+1}|^2) w_j^n = \mathrm{Re}\left(\langle \mathbf{x}^n + \mathbf{x}^{n+1}, \mathbf{x}^n - \mathbf{x}^{n+1} \rangle_{\mathbf{w}^n}\right) \,,
\end{aligned}$$

where we have used the notion of the weighted inner product $\langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{w}} = \sum_{j=1}^{N} x_j \overline{z}_j w_j$. By their definition in (IRLS$_1$) both $\mathbf{x}^n$ and $\mathbf{x}^{n+1}$ satisfy $\mathbf{A}\mathbf{x}^n = \mathbf{y} = \mathbf{A}\mathbf{x}^{n+1}$, so that $\mathbf{x}^n - \mathbf{x}^{n+1} \in \ker \mathbf{A}$. The characterization in (A.37) of the minimizer of a weighted least squares problem implies that $\mathrm{Re}(\langle \mathbf{x}^{n+1}, \mathbf{x}^n - \mathbf{x}^{n+1} \rangle_{\mathbf{w}^n}) = 0$. Therefore, with the above inequality

$$\begin{aligned}
&2\left(\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) - \mathcal{J}(\mathbf{x}^{n+1}, \mathbf{w}^{n+1}, \varepsilon_{n+1})\right) \\
&\geq \mathrm{Re}\left(\langle \mathbf{x}^n + \mathbf{x}^{n+1}, \mathbf{x}^n - \mathbf{x}^{n+1} \rangle_{\mathbf{w}^n}\right) - 2\,\mathrm{Re}(\langle \mathbf{x}^{n+1}, \mathbf{x}^n - \mathbf{x}^{n+1} \rangle_{\mathbf{w}^n}) \\
&= \mathrm{Re}\left(\langle \mathbf{x}^n - \mathbf{x}^{n+1}, \mathbf{x}^n - \mathbf{x}^{n+1} \rangle_{\mathbf{w}^n}\right) \\
&= \|\mathbf{x}^n - \mathbf{x}^{n+1}\|_{2,\mathbf{w}^n}^2 = \sum_{j=1}^{N} w_j^n |x_j^n - x_j^{n+1}|^2 \geq B^{-1} \|\mathbf{x}^n - \mathbf{x}^{n+1}\|_2^2 \,,
\end{aligned}$$

where we have used (15.46) in the last step. Summing these inequalities over n shows that

$$\begin{aligned}
\sum_{n=1}^{\infty} \|\mathbf{x}^n - \mathbf{x}^{n+1}\|_2^2 &\leq 2B \sum_{n=1}^{\infty} \left(\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) - \mathcal{J}(\mathbf{x}^{n+1}, \mathbf{w}^{n+1}, \varepsilon_{n+1})\right) \\
&\leq 2B\mathcal{J}(\mathbf{x}^1, \mathbf{w}^1, \varepsilon_1) \leq 2B^2 \,,
\end{aligned}$$

by Lemma 15.18. □

We further require a characterization of the minimizer $\mathbf{x}^{(\varepsilon)}$ of F_ε in (15.38), see also (15.39).

Lemma 15.20. *Let $\varepsilon > 0$ and $\mathbf{z} \in \mathbb{C}^N$ such that $\mathbf{A}\mathbf{z} = \mathbf{y}$. Then $\mathbf{z} = \mathbf{x}^{(\varepsilon)}$ if and only if $\mathrm{Re}(\langle \mathbf{z}, \mathbf{v} \rangle_{\mathbf{w}_{\mathbf{z},\varepsilon}}) = 0$ for all $\mathbf{v} \in \ker \mathbf{A}$, where $(\mathbf{w}_{\mathbf{z},\varepsilon})_j = (|z_j|^2 + \varepsilon^2)^{-1/2}$.*

Proof. First assume that $\mathbf{z} = \mathbf{x}^{(\varepsilon)}$ is the minimizer of (15.39). Let $\mathbf{v} \in \ker \mathbf{A}$ be arbitrary and consider the differentiable function

$$G(t) = F_\varepsilon(\mathbf{z} + t\mathbf{v}) - F_\varepsilon(\mathbf{z}) \,, \quad t \in \mathbb{R} \,.$$

Clearly $G(0) = 0$. By the minimizing property and $\mathbf{A}(\mathbf{z}+t\mathbf{v}) = \mathbf{y}$ for all $t \in \mathbb{R}$, we have $G(t) \geq 0$ for all $t \in \mathbb{R}$, so that $G'(0) = 0$. By a direct calculation

$$G'(0) = \sum_{j=1}^{N} \frac{\mathrm{Re}(z_j v_j)}{\sqrt{|z_j|^2 + \varepsilon^2}} = \mathrm{Re}\left(\langle \mathbf{z}, \mathbf{v} \rangle_{\mathbf{w}_{\mathbf{z},\varepsilon}} \right) \,,$$

and consequently $\mathrm{Re}(\langle \mathbf{z}, \mathbf{v} \rangle_{\mathbf{w}_{\mathbf{z},\varepsilon}}) = 0$ for all $\mathbf{v} \in \ker \mathbf{A}$.

Conversely, assume that $\mathbf{z}$ satisfies $\mathbf{A}\mathbf{z} = \mathbf{y}$ and $\langle \mathbf{z}, \mathbf{v} \rangle_{\mathbf{w}_{\mathbf{z},\varepsilon}} = 0$ for all $\mathbf{v} \in \ker \mathbf{A}$. By convexity of the function $f(u) := \sqrt{|u|^2 + \varepsilon^2}, u \in \mathbb{C}$, and Proposition B.9(a), we have for any $u, u_0 \in \mathbb{C}$,

$$\sqrt{|u|^2 + \varepsilon^2} \geq \sqrt{|u_0|^2 + \varepsilon^2} + \frac{\mathrm{Re}(u_0(\overline{u - u_0}))}{\sqrt{|u_0|^2 + \varepsilon^2}} \,.$$

Therefore, for any $\mathbf{v} \in \ker \mathbf{A}$ we have

$$F_\varepsilon(\mathbf{z} + \mathbf{v}) \geq F_\varepsilon(\mathbf{z}) + \sum_{j=1}^{N} \frac{\mathrm{Re}(z_j \overline{v_j})}{\sqrt{|z_j|^2 + \varepsilon^2}} = F_\varepsilon(\mathbf{z}) + \mathrm{Re}\left(\langle \mathbf{z}, \mathbf{v} \rangle_{\mathbf{w}_{\mathbf{z},\varepsilon}} \right) = F_\varepsilon(\mathbf{z}) \,.$$

Since $\mathbf{v} \in \ker \mathbf{A}$ is arbitrary it follows that $\mathbf{z} = \mathbf{x}^{(\varepsilon)}$ is a minimizer of (15.39). □

Now we are in the position to prove Theorem 15.16 on the convergence of the iteratively reweighted least squares algorithm.

Proof (of Theorem 15.16). First note that by $0 \leq \varepsilon_{n+1} \leq \varepsilon_n$ the sequence $(\varepsilon_n)_{n \in \mathbb{N}}$ always converges. We denote by $\varepsilon := \lim_{n \to \infty} \varepsilon_n$ its limit.

(a) Case $\varepsilon = 0$: First assume that $\varepsilon_{n_0} = 0$ for some $n_0 \in \mathbb{N}$. Then the algorithm stops by definition and we can set $\mathbf{x}^n = \mathbf{x}^{n_0}$ for $n \geq n_0$ so that $\lim_{n \to \infty} \mathbf{x}^n = \mathbf{x}^{n_0} = \mathbf{x}^\sharp$. By definition of ε it follows that the nonincreasing rearrangement $(\mathbf{x}^{n_0})^*_{s+1} = 0$ so that $\mathbf{x}^\sharp = \mathbf{x}^{n_0}$ is s-sparse. It follows from the null space property of order s of $\mathbf{A}$ that $\mathbf{x}^\sharp$ is the unique ℓ_1-minimizer of (15.34). If in addition, $\mathbf{x}$ is s-sparse, then also $\mathbf{x}$ is the unique ℓ_1-minimizer so that $\mathbf{x} = \mathbf{x}^\sharp$. For a general $\mathbf{x} \in \mathbb{C}^N$, not necessarily being s-sparse, it follows from Theorem 4.11 that

$$\|\mathbf{x} - \mathbf{x}^\sharp\|_1 \le \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x})_1 \,.$$

Now assume that $\varepsilon_n > 0$ for all $n \in \mathbb{N}$. Since $\lim_{n\to\infty} \varepsilon_n = 0$, there must exist an increasing sequence of indices (n_j) such that $\varepsilon_{n_j} < \varepsilon_{n_j-1}$ for all $j \in \mathbb{N}$. By definition (IRLS$_2$) of ε_n and the choice $\gamma = 1/N$ this implies that the nonincreasing rearrangement of $\mathbf{x}^{n_j}$ satisfies

$$(\mathbf{x}^{n_j})^*_{s+1} < N\varepsilon_{n_j-1}, \quad j \in \mathbb{N} \,.$$

By (15.45) the sequence $(\mathbf{x}^n)$ is bounded. Therefore, there exists a subsequence (n_{j_ℓ}) of (n_j) such that $(\mathbf{x}^{n_{j_\ell}})_\ell$ converges to some $\mathbf{x}^\sharp$ satisfying $\mathbf{A}\mathbf{x}^\sharp = \mathbf{y}$. It follows from the Lipschitz property (2.1) of the nonincreasing rearrangement that also $(\mathbf{x}^{n_{j_\ell}})^*$ converges to $(\mathbf{x}^\sharp)^*$, so that

$$(\mathbf{x}^\sharp)^*_{s+1} = \lim_{\ell\to\infty} (\mathbf{x}^{n_{j_\ell}})^*_{s+1} \le \lim_{j\to\infty} N\varepsilon_{n_{j_\ell}} = 0 \,.$$

This implies that $\mathbf{x}^\sharp$ is s-sparse. As above the null space property of order s implies that $\mathbf{x}^\sharp$ is the unique ℓ_1-minimizer. We still need to show that the full sequence $(\mathbf{x}^n)$ converges to $\mathbf{x}^\sharp$. Since $\mathbf{x}^{n_{j_\ell}} \to \mathbf{x}^\sharp$ and $\varepsilon_{n_{j_\ell}} \to 0$ as $\ell \to \infty$ the identity (15.42) implies that

$$\lim_{\ell\to\infty} \mathcal{J}(\mathbf{x}^{n_{j_\ell}}, \mathbf{w}^{n_{j_\ell}}, \varepsilon_{n_{j_\ell}}) = \|\mathbf{x}^\sharp\|_1 \,.$$

It follows from the monotonicity properties in (15.43) and (15.44) that $\lim_{n\to\infty} \mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) = \|\mathbf{x}^\sharp\|_1$. Again by (15.42) we conclude that

$$\mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) - N\varepsilon_n \le \|\mathbf{x}^n\|_1 \le \mathcal{J}(\mathbf{x}^n, \mathbf{w}^n, \varepsilon_n) \,,$$

so that $\lim_{n\to\infty} \|\mathbf{x}^n\|_1 = \|\mathbf{x}^\sharp\|_1$. By the stable null space property and Theorem 4.13 we finally obtain

$$\limsup_{n\to\infty} \|\mathbf{x}^n - \mathbf{x}^\sharp\|_1 \le \frac{1+\rho}{1-\rho}\big(\lim_{n\to\infty} \|\mathbf{x}^n\|_1 - \|\mathbf{x}^\sharp\|_1 + 2\,\sigma_s(\mathbf{x}^\sharp)_1\big) = 0 \,,$$

which shows that $\mathbf{x}^n \to \mathbf{x}^\sharp$. The error estimate (15.40) follows from the stable null space property as above.

(b) Case $\varepsilon > 0$: We first show that $\mathbf{x}^n \to \mathbf{x}^{(\varepsilon)}$, where $\mathbf{x}^{(\varepsilon)}$ is the minimizer of (15.39). By Lemma 15.19 the sequence $(\mathbf{x}^n)$ is bounded, so that it has accumulation points. Let $\mathbf{x}^{n_j}$ be a convergent subsequence with limit $\mathbf{x}^\sharp$. We claim that $\mathbf{x}^\sharp = \mathbf{x}^{(\varepsilon)}$, which by uniqueness of $\mathbf{x}^{(\varepsilon)}$ implies that *every* convergent subsequence converges to $\mathbf{x}^\sharp$. This means in turn that $\mathbf{x}^\sharp$ is the unique accumulation point of $(\mathbf{x}^n)_n$ and therefore, $\mathbf{x}^n$ converges to $\mathbf{x}^\sharp$ as $n \to \infty$.

Since $w_j^n = (|x_j^n|^2 + \varepsilon^2)^{-1/2} \le \varepsilon^{-1}$ it follows that

$$\lim_{j\to\infty} w_j^{n_j} = (|x_j^\sharp|^2 + \varepsilon^2)^{-1/2} = (w_{\mathbf{z},\varepsilon})_j =: w_j^\sharp \,, \quad j \in [N] \,,$$

where we have used the same notation as in Lemma 15.20. It follows from Lemma 15.19 that also $\mathbf{x}^{n_j+1}$ converges to $\mathbf{x}^\sharp$. By the orthogonality relation (A.37) and the minimizing property (IRLS_1) of $\mathbf{x}^{n_j+1}$ we have, for every $\mathbf{v} \in \ker \mathbf{A}$,

$$\mathrm{Re}\left(\langle \mathbf{x}^\sharp, \mathbf{v}\rangle_{\mathbf{w}^\sharp}\right) = \lim_{j\to\infty} \mathrm{Re}\left(\langle \mathbf{x}^{n_j+1}, \mathbf{v}\rangle_{\mathbf{w}^{n_j}}\right) = 0 \,.$$

The characterization in Lemma 15.20 implies that $\mathbf{x}^\sharp = \mathbf{x}^{(\varepsilon)}$.

Now we show the error estimate (15.41). For our $\mathbf{x} \in \mathbb{C}^N$ with $\mathbf{A}\mathbf{x} = \mathbf{y}$ we have by the minimizing property of $\mathbf{x}^{(\varepsilon)}$ that

$$\|\mathbf{x}^{(\varepsilon)}\|_1 \le F_\varepsilon(\mathbf{x}^{(\varepsilon)}) \le F_\varepsilon(\mathbf{x}) = \sum_{j=1}^N \sqrt{|x_j|^2+\varepsilon^2} \le N\varepsilon + \sum_{j=1}^N |x_j| = N\varepsilon + \|\mathbf{x}\|_1 \,.$$

It follows from the stable null space property of order $\widetilde{s} \le s$ and Theorem 4.13 that

$$\|\mathbf{x}^{(\varepsilon)} - \mathbf{x}\|_1 \le \frac{1+\rho}{1-\rho}\left(\|\mathbf{x}^{(\varepsilon)}\|_1 - \|\mathbf{x}\|_1 + 2\sigma_{\widetilde{s}}(\mathbf{x})_1\right) \le \frac{1+\rho}{1-\rho}\big(N\varepsilon + 2\sigma_{\widetilde{s}}(\mathbf{x})_1\big) \,. \tag{15.47}$$

The definition (IRLS_2) of ε_n with $\gamma = 1/N$ and the property (2.1) of the nonincreasing rearrangement yield

$$N\varepsilon = \lim_{n\to\infty} N\varepsilon_n \le \lim_{n\to\infty} (\mathbf{x}^n)^*_{s+1} = (\mathbf{x}^{(\varepsilon)})^*_{s+1} \,.$$

Invoking (2.3) gives

$$\begin{aligned}(s+1-\widetilde{s})N\varepsilon &\le (s+1-\widetilde{s})(\mathbf{x}^{(\varepsilon)})^*_{s+1} \le \|\mathbf{x}^{(\varepsilon)} - \mathbf{x}\|_1 + \sigma_{\widetilde{s}}(\mathbf{x})_1 \\ &\le \frac{1+\rho}{1-\rho}\big(N\varepsilon + 2\sigma_{\widetilde{s}}(\mathbf{x})_1\big) + \sigma_{\widetilde{s}}(\mathbf{x})_1 \,, \end{aligned} \tag{15.48}$$

where we have also applied (15.47). Equivalently,

$$\left(s+1-\widetilde{s}-\frac{1+\rho}{1-\rho}\right)\big(N\varepsilon + 2\sigma_{\widetilde{s}}(\mathbf{x})_1\big) \le 2(1/2+s+1-\widetilde{s})\sigma_{\widetilde{s}}(\mathbf{x})_1 \,.$$

By assumption, we have $s - \widetilde{s} > \frac{2\rho}{1-\rho}$ so that $s+1-\widetilde{s} > (1+\rho)/(1-\rho)$ and

$$N\varepsilon + 2\sigma_{\widetilde{s}}(\mathbf{x})_1 \le \frac{2(s-\widetilde{s})+3}{(s-\widetilde{s})-\frac{2\rho}{1-\rho}}\sigma_{\widetilde{s}}(\mathbf{x})_1 \,.$$

Plugging this into (15.47) completes the proof (15.41).

Finally, assume that $\mathbf{x}$ is $\widetilde{s}$-sparse with $\widetilde{s} \le s - \frac{2\rho}{1-\rho}$. Then by (15.41) we have $\mathbf{x}^\sharp = \mathbf{x}$, so that also $\mathbf{x}^\sharp$ is $\widetilde{s}$ sparse. But this implies that $(\mathbf{x}^\sharp)^*_{s+1} = 0$, so that $\varepsilon = \lim_{n\to\infty} \varepsilon_n = 0$ by definition (IRLS_2) of ε_n. □

We conclude this section with an estimate of the rate of convergence of the iteratively reweighted least squares algorithm. In contrast to (15.26) concerning the primal dual algorithm of the previous section, we actually estimate the ℓ_1-distance of the iterates $\mathbf{x}^n$ to their limit $\mathbf{x}^\sharp$ rather than an auxiliary primal dual gap, which does not allow conclusions on such distance. However, the estimated rate kicks in only when the iterates are close enough to the limit. Nothing is said about the initial phase, although practical experience shows that the initial phase does not take overly long. The estimate for the exactly sparse case below shows linear convergence in ℓ_1.

Theorem 15.21. *Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfy the stable null space property of order s with constant $\rho < 1 - \frac{2}{s+2}$. Let $\widetilde{s} < s - \frac{2\rho}{1-\rho}$, $0 < \kappa < 1$ be such that*

$$\mu := \frac{\rho(1+\rho)}{1-\kappa}\left(1 + \frac{1}{s+1-\widetilde{s}}\right) < 1 \,.$$

Let $\mathbf{x} \in \mathbb{C}^N$ be $\widetilde{s}$-sparse with $S = \operatorname{supp}\mathbf{x}$. Let $\mathbf{x}^\sharp$ be the limit of the sequence $(\mathbf{x}^n)_n$ generated by the IRLS algorithm with parameters s and $\gamma = 1/N$. Then by Theorem 15.16 $\mathbf{x} = \mathbf{x}^\sharp$. Let $n_0 \in \mathbb{N}$ be such that

$$\|\mathbf{x}^\sharp - \mathbf{x}^{n_0}\|_1 \le R := \kappa \min_{j\in S} |x_j| \,.$$

Then for all $n \ge n_0$ we have

$$\|\mathbf{x}^\sharp - \mathbf{x}^{n+1}\|_1 \le \mu \|\mathbf{x}^\sharp - \mathbf{x}^n\|_1 \,. \tag{15.49}$$

Consequently, $\mathbf{x}^n$ converges linearly to $\mathbf{x}^\sharp$, that is, $\|\mathbf{x}^n - \mathbf{x}^\sharp\|_1 \le \mu^{n-n_0}\|\mathbf{x}^{n_0} - \mathbf{x}^\sharp\|_1$ for all $n > n_0$.

Remark 15.22. Note that if ρ is sufficiently small, i.e., $\rho(1+\rho) < 2/3$ then for any $\widetilde{s} \le s - 1$ there is always a $\kappa > 0$ such that $\mu < 1$, so that $\mathbf{x}^n$ always converges linearly to $\mathbf{x}^\sharp$ whenever $\mathbf{x}$ is $(s-1)$-sparse.

Proof. Denote $\mathbf{v}^n = \mathbf{x}^n - \mathbf{x}^\sharp \in \ker \mathbf{A}$. By the minimizing property (IRLS_1) of $\mathbf{x}^{n+1}$ and the characterization of the minimizer in (A.37) we have

$$0 = \operatorname{Re}\left(\langle \mathbf{x}^{n+1}, \mathbf{v}^{n+1}\rangle_{\mathbf{w}^n}\right) = \operatorname{Re}\left(\langle \mathbf{x}^\sharp + \mathbf{v}^{n+1}, \mathbf{v}^{n+1}\rangle_{\mathbf{w}^n}\right) .$$

Using that $\operatorname{supp}\mathbf{x}^\sharp = \operatorname{supp}\mathbf{x} = S$ and rearranging terms gives

$$\sum_{j=1}^N |v_j^{n+1}|^2 w_j^n = -\operatorname{Re}\left(\sum_{j\in S} x_j^\sharp v_j^{n+1} w_j^n\right) = -\operatorname{Re}\left(\sum_{j\in S} \frac{x_j^\sharp}{\sqrt{|x_j^n|^2 + \varepsilon_n^2}} v_j^{n+1}\right) . \tag{15.50}$$

Now let $n \ge n_0$ so that $E_n := \|\mathbf{x}^\sharp - \mathbf{x}^n\|_1 \le R$. Then, for $j \in S$,

$$|v_j^n| \le \|\mathbf{v}^n\|_1 = E_n \le \kappa |x_j^\sharp| \,,$$

so that

$$\frac{|x_j^\sharp|}{\sqrt{|x_j^n|^2+\varepsilon_n^2}} \le \frac{|x_j^\sharp|}{|x_j^n|} = \frac{|x_j^\sharp|}{|x_j^\sharp|+v_j^n|} \le \frac{1}{1-\kappa} \, . \tag{15.51}$$

By combining (15.50) and (15.51) with the stable null space property we reach

$$\sum_{j=1}^N |v_j^{n+1}|^2 w_j^n \le \frac{1}{1-\kappa} \|\mathbf{v}_S^{n+1}\|_1 \le \frac{\rho}{1-\kappa} \|\mathbf{v}_{\overline{S}}^{n+1}\|_1 \, .$$

The Cauchy-Schwarz inequality yields

$$\begin{aligned} \|\mathbf{v}_{\overline{S}}^{n+1}\|_1^2 &\le \left(\sum_{j \in \overline{S}} |v_j^{n+1}|^2 w_j^n \right) \left(\sum_{j \in \overline{S}} \sqrt{|x_j^n|^2+\varepsilon_n^2} \right) \\ &\le \left(\sum_{j=1}^N |v_j^{n+1}|^2 w_j^n \right) (\|\mathbf{v}^n\|_1 + N\varepsilon_n) \\ &\le \frac{\rho}{1-\kappa} \|\mathbf{v}_{\overline{S}}^{n+1}\|_1 (\|\mathbf{v}^n\|_1 + N\varepsilon_n) \, . \end{aligned} \tag{15.52}$$

If $\mathbf{v}_{\overline{S}}^{n+1} = \mathbf{0}$ then $\mathbf{x}_{\overline{S}}^{n+1} = \mathbf{0}$, so that $\mathbf{x}^{n+1}$ is $\widetilde{s}$-sparse and the algorithm has stopped by definition. Since $\mathbf{x}^{n+1} - \mathbf{x}^\sharp \in \ker \mathbf{A}$, which does not contain $\widetilde{s}$-sparse elements different from $\mathbf{0}$ by the null space property we have obtained the solution $\mathbf{x}^{n+1} = \mathbf{x}^\sharp$ so that $E_{n+1} = 0$ and (15.49) is trivially satisfied.

If $\mathbf{v}_{\overline{S}}^{n+1} \neq \mathbf{0}$ then we may divide by $\|\mathbf{v}_{\overline{S}}^{n+1}\|_1$ in inequality (15.52) to obtain

$$\|\mathbf{v}_{\overline{S}}^{n+1}\|_1 \le \frac{\rho}{1-\kappa} (\|\mathbf{v}^n\|_1 + N\varepsilon_n) \, .$$

Using once more the stable null space property we arrive at

$$\|\mathbf{v}^{n+1}\|_1 = \|\mathbf{v}_S^{n+1}\|_1 + \|\mathbf{v}_{\overline{S}}^{n+1}\|_1 \le (1+\rho) \|\mathbf{v}_{\overline{S}}^{n+1}\|_1 \le \frac{\rho(1+\rho)}{1-\kappa} (\|\mathbf{v}^n\|_1 + N\varepsilon_n) \, .$$

By the update rule (IRLS$_2$) for ε_{n+1} and (2.3) we have

$$N\varepsilon_n \le (\mathbf{x}^n)_{s+1}^* \le \frac{1}{s+1-\widetilde{s}} (\|\mathbf{x}^n - \mathbf{x}^\sharp\|_1 + \sigma_{\widetilde{s}}(\mathbf{x}^\sharp)_1) = \frac{\|\mathbf{v}^n\|_1}{s+1-\widetilde{s}}$$

because $\sigma_s(\mathbf{x}^\sharp)_1 = 0$ by $\|\mathbf{x}^\sharp\|_0 = \widetilde{s}$. Altogether we get the bound

$$E_{n+1} = \|\mathbf{v}^{n+1}\|_1 \le \frac{\rho(1+\rho)}{1-\kappa} \left(1 + \frac{1}{s+1-\widetilde{s}} \right) \|\mathbf{v}^n\|_1 = \mu E_n \, .$$

Since $\mu < 1$ by assumption we have also $E_{n+1} \le R$. We conclude that $E_{n+1} \le \mu E_n$ for all $n \ge n_0$. □

Remark 15.23. The precise update rule (IRLS$_2$) for ε_n is not very important for this analysis. When $\|\mathbf{x}^\sharp - \mathbf{x}^{n_0}\|_1 \le R$ then the estimate $\|\mathbf{x}^\sharp - \mathbf{x}^{n+1}\|_1 \le \mu_0 (\|\mathbf{x}^\sharp - \mathbf{x}^n\|_1 + N\varepsilon_n)$ with $\mu_0 = \rho(1+\rho)/(1-\kappa)$ is always valid. The rule (IRLS$_2$) only guarantees that $\|\mathbf{x}^\sharp - \mathbf{x}^{n_0}\|_1 \le R$ will actually be satisfied for some n_0 by Theorem 15.16.

Notes

Background on general convex optimization methods, in particular, interior point methods can be found in various textbooks including [59, 318].

The homotopy method – or modified LARS – was introduced and analyzed in [326, 325, 154, 147]. Theorem 15.2 was shown in [154]. Concerning the unlikely case that the maximum in (15.6) or the minimum in (15.9) or (15.10) is simultaneously attained at more than one index the reader is referred to [154].

The adaptive inverse scale space method [66] is another fast ℓ_1-minimization algorithm, which is similar to the homotopy method. It also builds up the support successively. At each step, however, one solves a least squares problem with a positivity constraint instead of a system of linear equations. Like the homotopy method, the inverse scale space method seems to apply only for the real-valued case.

The primal dual algorithm of Section 15.2 was first introduced for a special case in [342]. In full generality it was presented and analyzed by A. Chambolle and T. Pock in [91]. For the case, that either F^* or G is strongly convex with known strong convexity constant γ in (B.6), a modification of the algorithm where θ, τ, σ are varying throughout the iterations is introduced in [91]. This variant has an improved convergence rate $\mathcal{O}(1/n^2)$. On the other hand, it was proved by Nesterov [316] that the convergence rate $\mathcal{O}(1/n)$ cannot be improved for general convex functions F, G, so that in this sense the rate of Theorem 15.8 is optimal. Also note that for the basis pursuit problems in Examples 15.7 (a) and (b), the strong convexity assumptions fail and only the described basic primal dual algorithm applies.

The proof technique involving the discrete derivative, see Lemma 15.10, was introduced by S. Bartels in [25]. The main motivation of Chambolle and Pock for their algorithm were total variation and related minimization problems appearing in imaging applications [91]. The parameter choice $\theta = 0$ in (PD_3) yields the Arrow-Hurwicz method [16, 454]. While empirically the corresponding algorithm converges as well this point has not yet been verified theoretically. Chambolle and Pock's algorithm is also related to Douglas-Rachford splitting methods [283], see [91] for more details on this relation. Further algorithms for ℓ_1-minimization based on so-called Bregman iterations, including the inexact Uzawa algorithm, are discussed in [453, Section 5], see also [452].

Iterative thresholding algorithms [28, 115, 177, 176] can be viewed as predecessors to Chambolle and Pock's primal dual algorithm. Consider the ℓ_1-regularized functional (15.3),

$$F_\lambda(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1 \,, \tag{15.53}$$

which is equivalent to (15.24) after an obvious transformation of the regularization parameter. Using the soft thresholding operator $\mathcal{S}_\lambda$ in (15.21), (15.20), the minimizer $\mathbf{x}^\sharp$ of F_λ satisfies the fixed point equation

$$\mathbf{x}^\sharp = \mathcal{S}_\lambda(\mathbf{x}^\sharp + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^\sharp)) \,, \tag{15.54}$$

see Exercise 15.2. This motivates to consider the fixed point iteration

$$\mathbf{x}^{n+1} = \mathcal{S}_\lambda(\mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)) \tag{15.55}$$

with some initial point $\mathbf{x}^0$ as an algorithm for the minimization of F_λ. Without the soft thresholding operator $\mathcal{S}_\lambda$, that is, $\mathbf{x}^{n+1} = \mathbf{x}^n + \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^n)$, this is called a Landweber iteration, and it is well-known that it converges to the solution of the corresponding ℓ_2-regularized functional provided $\|\mathbf{A}\|_{2\to 2} < 1$, see e.g. [162]. The iteration (15.55) is sometimes called thresholded Landweber iteration. It was shown in [115] that $\mathbf{x}^n$ converges to the minimizer $\mathbf{x}^\sharp$ of the functional (15.53) if $\|\mathbf{A}\|_{2\to 2} < 1$. (Note that $\|\mathbf{A}\|_{2\to 2} < 1$ can always be achieved by renormalizing the whole functional.) In practice, the convergence of (15.55) is rather slow. Therefore, acceleration methods were introduced [28, 91, 117], of which Chambolle and Pock's primal dual algorithm is one variant. Another variant called FISTA is introduced and analyzed in [28], which uses only a primal variable. Its convergence speed is much faster than the one of the thresholded Landweber iteration (15.55).

The iteratively reweighted least squares algorithm of Section 15.3 was introduced and analyzed in [116]. A version of the convergence rate estimate in Theorem 15.21, which applies also to approximately sparse vectors is Theorem 6.4 in [116]. This paper contains also a variant where the update rule for the weight is motivated by ℓ_p-minimization with $p < 1$. Although it is not known in general whether the corresponding algorithm always converges, it is shown that once it converges, then it converges superlinearly in a neighborhood of the limit. In [448] a version of iteratively reweighted least squares is developed and analyzed that uses the conjugate gradient method in [262] in order to approximately solve the weighted least squares problem in (IRLS_2). An estimate of the accuracy required in each iteration is provided that ensures overall convergence.

A variant of iteratively reweighted least squares for low rank matrix recovery is contained in [178]. Translating the corresponding algorithm back to the vector case, this paper considers a slightly different update rule for the weight, namely

$$w_j^{n+1} = \min\{|x_j^{n+1}|^{-1}, \varepsilon_n^{-1}\} \,.$$

Convergence results can also be shown for this variant, see [178] for precise statements. Versions of iteratively reweighted least squares methods appeared also earlier in [100, 276, 324].

Further information on numerical methods for sparse recovery can be found in [175]. Other optimization methods specialized to ℓ_1-minimization can be found in [29, 66, 261, 171, 430, 222].

As outlined in Chapter 3, the basic ℓ_1-minimization problem (BP) is equivalent to the linear optimization problem (P'_1) in the real case, and to the second order cone problem ($\mathrm{P}'_{1,\eta}$) (with $\eta = 0$) in the complex case. For such

problems, general purpose optimization algorithms apply. While for linear optimization problems the well-known simplex method [318] applies, so-called interior point methods are an efficient alternative (in both the real and complex case). We refer to [59, 318] for more information.

Exercises

15.1. Verify formula (B.17) for the soft-thresholding operator. Show that

$$S_\tau(y)^2 = \min_{|x|\le\tau} (x-y)^2 \, .$$

15.2. Show that the minimizer $\mathbf{x}^\sharp$ of F_λ in (15.53) satisfies (15.54).

15.3. Implement one or more of the algorithms of this Chapter. Choose $\mathbf{A} \in \mathbb{R}^{m\times N}$ as Gaussian random matrix, or $\mathbf{A} \in \mathbb{C}^{m\times N}$ as partial random Fourier matrix. In the latter case exploit the Fast Fourier Transform. Test the algorithm on randomly generated s-sparse signals, where first the support is chosen at random and then the nonzero coefficients. By varying m, s, N evaluate the empirical success probability of recovery. Compare the runtime of the algorithms for small and medium sparsity s.

A

Matrix Analysis

This appendix collects useful background from linear algebra and matrix analysis, such as vector and matrix norms, singular value decompositions, Gershgorin's disc theorem and matrix functions. Much more material than listed here, can be found in various books on the subject including [38, 41, 198, 235, 242, 243, 413].

A.1 Vector and Matrix Norms

We work with real or complex vector spaces X, usually $X = \mathbb{R}^n$ or $X = \mathbb{C}^n$. We will usually write the vectors in $\mathbb{C}^n$ in boldface, $\mathbf{x}$, while their entries will be denoted x_j, $j \in [n]$, where $[n] := \{1, \ldots, n\}$. The canonical unit vectors in $\mathbb{R}^n$ will be denoted by $\mathbf{e}_\ell$ with entries

$$(\mathbf{e}_\ell)_j = \delta_{\ell,j} = \begin{cases} 1 & \text{if } j = \ell\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

We denote by $\mathbb{R}_+ := \{x \in \mathbb{R}, x \geq 0\}$ the non-negative reals.

Definition A.1. *A non-negative function $\|\cdot\| : X \to \mathbb{R}_+$ is called a norm if*

(a) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$ (definiteness).
(b) $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$ for all scalars λ and $\mathbf{x} \in X$ (homogeneity).
(c) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

If only (b) and (c) hold, so that $\|\mathbf{x}\| = 0$ does not necessarily imply $\mathbf{x} = 0$, then $\|\cdot\|$ is called a semi-norm.
If (a) and (b) hold, but (c) is replaced by the weaker quasi-triangle inequality

$$\|\mathbf{x} + \mathbf{y}\| \leq C(\|\mathbf{x}\| + \|\mathbf{y}\|)$$

for some constant $C \geq 1$, then $\|\cdot\|$ is called a quasi-norm. The constant C is called its quasi-norm constant.

A space X endowed with a norm $\|\cdot\|$ is called a normed space.

Definition A.2. *Let X be a set. A function $d : X \times X \to \mathbb{R}_+$ is called a metric if*

(a) $d(x,y) = 0$ *if and only if* $x = y$.
(b) $d(x,y) = d(y,x)$ *for all* $x, y \in X$.
(c) $d(x,z) \le d(x,y) + d(y,z)$ *for all* $x, y, z \in X$.

If only (b) and (c) hold then d is called a pseudo-metric.

The set X endowed with a metric d is called a metric space. Clearly, a norm $\|\cdot\|$ on X induces a metric on X by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| .$$

A semi-norm induces a pseudo-metric.

The ℓ_p-norm (or simply p-norm) on $\mathbb{R}^n$ or $\mathbb{C}^n$ is defined as

$$\|\mathbf{x}\|_p := \big(\sum_{j=1}^{n} |x_j|^p\big)^{1/p} , \quad 0 < p < \infty ,$$
$$\|\mathbf{x}\|_\infty := \max_{j\in[n]} |x_j| .$$

If $1 \le p \le \infty$ then $\|\cdot\|_p$ is a norm. For $0 < p < 1$ it is only a quasi-norm with quasi-norm constant $C = 2^{1/p} - 1$ that satisfies the p-triangle inequality

$$\|\mathbf{x} + \mathbf{y}\|_p^p \le \|\mathbf{x}\|_p^p + \|\mathbf{y}\|_p^p .$$

Therefore, the ℓ_p-norm induces a metric via $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p^p$ for $0 < p < 1$.

Remark A.3. It is known that for any quasi-norm one can find an equivalent quasi-norm that satisfies the p-triangle inequality for some $0 < p \le 1$, see e.g. [124].

We define a ball of radius $t > 0$ around a point $\mathbf{x}$ in a metric space (X, d) by

$$B(\mathbf{x}, t) = B_d(\mathbf{x}, t) = \{\mathbf{z} \in X, d(\mathbf{x}, \mathbf{z}) \le t\} .$$

If the metric is induced by a norm $\|\cdot\|$ on a vector space then we also write

$$B_{\|\cdot\|}(\mathbf{x}, t) \ = \{\mathbf{z} : \|\mathbf{x} - \mathbf{z}\| \le t\} . \tag{A.1}$$

If $\mathbf{x} = 0$ is the zero vector and $t = 1$ then $B = B_{\|\cdot\|} = B_{\|\cdot\|}(0, 1)$ is called unit ball.

The inner product on $\mathbb{C}^n$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{n} x_j \overline{y_j} , \quad \mathbf{x}, \mathbf{y} \in \mathbb{C}^n .$$

On $\mathbb{R}^n$ it is given by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^n x_j y_j$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The ℓ_2-norm is related to the inner product by

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \, .$$

The Cauchy-Schwarz inequality states that

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{C}^n \, .$$

More generally, for $p, q \in [1, \infty]$ such that $1/p + 1/q = 1$ (with the convention that $1/\infty = 0$ and $1/0 = \infty$), we have Hölder's inequality

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \le \sum_{j=1} |x_j||y_j| \le \|\mathbf{x}\|_p \|\mathbf{y}\|_q \quad \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{C}^n \, .$$

We note the easy but important special case $|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty$. Denoting by $\mathbf{1} \in \mathbb{C}^n$ the vector having all entries equal to 1, Hölder's inequality implies that, for $1 \le p \le \infty$,

$$\|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j| \le \|\mathbf{1}\|_q \|\mathbf{x}\|_p = n^{1/q} \|\mathbf{x}\|_p = n^{1-1/p} \|\mathbf{x}\|_p \, , \tag{A.2}$$

where $1/q + 1/p = 1$. We note the important special cases

$$\|\mathbf{x}\|_1 \le \sqrt{n} \|\mathbf{x}\|_2 \quad \text{ and } \quad \|\mathbf{x}\|_1 \le n \|\mathbf{x}\|_\infty \, .$$

More generally, if $0 < p \le q \le \infty$ applying inequality (A.2) with p replaced by q/p gives

$$\|\mathbf{x}\|_p^p = \sum_{j=1}^n |x_j|^p \le n^{1-p/q} \big(\sum_{j=1}^n (|x_j|^p)^{q/p} \big)^{p/q} \, .$$

By taking the p-th root we reach

$$\|\mathbf{x}\|_p \le n^{1/p - 1/q} \|\mathbf{x}\|_q \, . \tag{A.3}$$

If $\mathbf{x}$ has actually at most s non-zero entries, $\|\mathbf{x}\|_0 = \text{card}(\{\ell, x_\ell \ne 0\}) \le s$, then the above inequalities become $\|\mathbf{x}\|_p \le s^{1/p - 1/q} \|\mathbf{x}\|_q$, in particular,

$$\|\mathbf{x}\|_1 \le \sqrt{s} \|\mathbf{x}\|_2 \le s \|\mathbf{x}\|_\infty .$$

We also have reversed inequalities, for $0 < p < q \le \infty$,

$$\|\mathbf{x}\|_q \le \|\mathbf{x}\|_p \, . \tag{A.4}$$

In particular, $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$. Indeed, the bound $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_p$ is obvious. For $p < q < \infty$,

$$\|\mathbf{x}\|_q^q = \sum_{j=1}^n |x_j|^q = \sum_{j=1}^n |x_j|^{q-p} |x_j|^p \le \|\mathbf{x}\|_\infty^{q-p} \sum_{j=1}^n |x_j|^p \le \|\mathbf{x}\|_p^{q-p} \|\mathbf{x}\|_p^p = \|\mathbf{x}\|_p^q \, .$$

Both bounds (A.3) and (A.4) are sharp in general. Indeed, equality holds in (A.3) for a vector with constant entries, while equality holds in (A.4) for (scalar multiples of) a canonical unit vector.

Definition A.4. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ or $\mathbb{C}^n$. Its dual norm $\|\cdot\|_*$ is defined by*

$$\|\mathbf{x}\|_* := \sup_{\|\mathbf{y}\|\le 1} |\langle \mathbf{y}, \mathbf{x}\rangle| \ .$$

In the real case, the dual norm may equivalently be defined via

$$\|\mathbf{x}\|_* = \sup_{\mathbf{y}\in\mathbb{R}^n, \|\mathbf{y}\|\le 1} \langle \mathbf{y}, \mathbf{x}\rangle \ ,$$

while in the complex case

$$\|\mathbf{x}\|_* = \sup_{\mathbf{y}\in\mathbb{C}^n, \|\mathbf{y}\|\le 1} \mathrm{Re}(\langle \mathbf{y}, \mathbf{x}\rangle) \ .$$

The dual of the dual norm $\|\cdot\|_*$ is the "primal" norm $\|\cdot\|$. In particular, we have

$$\|\mathbf{x}\| = \sup_{\|\mathbf{y}\|_*\le 1} |\langle \mathbf{x}, \mathbf{y}\rangle| = \sup_{\|\mathbf{y}\|_*\le 1} \mathrm{Re}(\langle \mathbf{x}, \mathbf{y}\rangle|) \ . \tag{A.5}$$

The dual of $\|\cdot\|_p$ is $\|\cdot\|_q$ with $1/p + 1/q = 1$. In particular, $\|\cdot\|_2$ is self-dual,

$$\|\mathbf{x}\|_2 = \sup_{\|\mathbf{y}\|_2\le 1} |\langle \mathbf{y}, \mathbf{x}\rangle| \ , \tag{A.6}$$

while $\|\cdot\|_\infty$ is the dual of $\|\cdot\|_1$ and vice versa.

Given a subspace W of a vector space X, the *quotient space* X/W consists of the residue classes

$$[\mathbf{x}] := \mathbf{x} + W = \{\mathbf{x} + \mathbf{w}, \mathbf{w} \in W\} \ , \quad \mathbf{x} \in X \ .$$

The *quotient map* is the surjective linear map $\mathbf{x} \mapsto [\mathbf{x}] = \mathbf{x} + W \in X/W$. The *quotient norm* on X/W is defined by

$$\|[\mathbf{x}]\|_{X/W} := \inf\{\|\mathbf{v}\|, \mathbf{v} \in [\mathbf{x}] = \mathbf{x} + W\}, \qquad [\mathbf{x}] \in X/W \ .$$

Next we consider matrices $\mathbf{A} \in \mathbb{C}^{m\times n}$ (or more generally, linear mappings between normed spaces). The entries of $\mathbf{A}$ will be denoted A_{jk}, $j \in [m], k \in [n]$. The columns of $\mathbf{A}$ will be denoted $\mathbf{a}_k$, so that $\mathbf{A} = (\mathbf{a}_1|\dots|\mathbf{a}_n)$. The transpose of $\mathbf{A} \in \mathbb{C}^{m\times n}$ is the matrix $\mathbf{A}^T \in \mathbb{C}^{n\times m}$ with entries $(\mathbf{A}^T)_{kj} = A_{jk}$. A matrix $\mathbf{B} \in \mathbb{C}^{n\times n}$ is called symmetric if $\mathbf{B}^\top = \mathbf{B}$. The adjoint (or Hermitian transpose) of $\mathbf{A} \in \mathbb{C}^{m\times n}$ is the matrix $\mathbf{A}^* \in \mathbb{C}^{n\times m}$ with entries $(\mathbf{A}^*)_{kj} = \overline{A_{jk}}$. For $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{y} \in \mathbb{C}^m$ we have $\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle$. A matrix $\mathbf{B} \in \mathbb{C}^{n\times n}$ is called self-adjoint (or Hermitian) if $\mathbf{B}^* = \mathbf{B}$. The identity matrix on $\mathbb{C}^n$ will be denoted $\mathbf{Id}$ or $\mathbf{Id}_n$. A matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ is called unitary if $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{Id}$. A self-adjoint matrix $\mathbf{B}$ possesses an eigenvalue decomposition of the form $\mathbf{B} = \mathbf{U}^*\mathbf{D}\mathbf{U}$, where $\mathbf{U}$ is a unitary matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ containing the real eigenvalues $\lambda_1, \dots, \lambda_n$ of $\mathbf{B}$.

Definition A.5. *Let $\mathbf{A} : X \to Y$ be a linear map between two normed spaces $(X, \|\cdot\|)$ and $(Y, \|\cdot\|)$. The operator norm of $\mathbf{A}$ is defined as*

$$\|\mathbf{A}\| := \sup_{\|\mathbf{x}\| \le 1} \|A\mathbf{x}\| \,. \tag{A.7}$$

In particular, for a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, and $1 \le p, q \le \infty$ we define the matrix norm, or operator norm, between ℓ_p and ℓ_q as

$$\|\mathbf{A}\|_{p\to q} := \sup_{\|\mathbf{x}\|_p \le 1} \|\mathbf{A}\mathbf{x}\|_q \,. \tag{A.8}$$

By definition for $\mathbf{A} : X \to Y$ and $\mathbf{x} \in X$, we have $\|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$. Note that for a real matrix $\mathbf{A}$ it does not matter whether the supremum is taken over real or complex $\mathbf{x}$ in (A.7). Also, we may restrict to unit vectors, that is,

$$\|\mathbf{A}\|_{p\to q} = \sup_{\|\mathbf{x}\|_p = 1} \|\mathbf{A}\mathbf{x}\|_q \,.$$

We summarize explicit expressions for the matrix norm $\|\mathbf{A}\|_{p\to q}$ for some special choices of p, q. The lemma below also refers to the singular values of a matrix, which will be covered in the next section.

Lemma A.6. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$.*

(a) We have $\|\mathbf{A}\|_{2\to 2} = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})} = \sigma_{\max}(\mathbf{A})$, where $\lambda_{\max}(\mathbf{A}^*\mathbf{A})$ is the largest eigenvalues of $\mathbf{A}^*\mathbf{A}$, and $\sigma_{\max}(\mathbf{A})$ the largest singular value of $\mathbf{A}$.
(b) For $1 \le p \le \infty$ we have $\|\mathbf{A}\|_{1\to p} = \max_{k\in[n]} \|\mathbf{a}_k\|_p$. In particular,

$$\|\mathbf{A}\|_{1\to 1} = \max_{k\in[n]} \sum_{j=1}^{m} |A_{jk}| \,, \tag{A.9}$$

and

$$\|\mathbf{A}\|_{1\to 2} = \max_{k\in[n]} \|\mathbf{a}_k\|_2 \,. \tag{A.10}$$

(c) $\|\mathbf{A}\|_{\infty\to\infty} = \max_{j\in[m]} \sum_{k=1}^{n} |A_{jk}|$.

Proof. (a) Since $\mathbf{A}^*\mathbf{A} \in \mathbb{C}^{n\times n}$ is self-adjoint it can be diagonalized, $\mathbf{A}^*\mathbf{A} = \mathbf{U}^*\mathbf{D}\mathbf{U}$ with unitary $\mathbf{U}$ and diagonal $\mathbf{D}$ containing the eigenvalues λ_ℓ of $\mathbf{A}^*\mathbf{A}$ on the diagonal. For $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\|_2 = 1$ we have

$$\begin{aligned}\|\mathbf{A}\mathbf{x}\|_2^2 &= \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}\rangle = \langle \mathbf{U}^*\mathbf{D}\mathbf{U}\mathbf{x}, \mathbf{U}^*\mathbf{D}\mathbf{U}\mathbf{x}\rangle = \langle \mathbf{D}\mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{U}^*\mathbf{D}\mathbf{U}\mathbf{x}\rangle \\ &= \langle \mathbf{D}\mathbf{U}\mathbf{x}, \mathbf{D}\mathbf{U}\mathbf{x}\rangle = \|\mathbf{D}\mathbf{U}\mathbf{x}\|_2^2 \,.\end{aligned}$$

Since $\mathbf{U}$ is unitary, we have $\|\mathbf{U}\mathbf{x}\|_2^2 = \langle \mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{x}\rangle = \langle \mathbf{x}, UU^*\mathbf{x}\rangle = \|\mathbf{x}\|_2^2 = 1$. Moreover, for an abritrary vector $\mathbf{z} \in \mathbb{C}^n$ we have

$$\|\mathbf{D}\mathbf{z}\|_2^2 = \sum_{j=1}^{n} \lambda_j^2 |z_j|^2 \le \max_{j\in n} \lambda_j^2 \sum_{j=1}^{n} |z_j|^2 = \lambda_{\max}(\mathbf{A}^*\mathbf{A})\|\mathbf{z}\|_2^2 \,.$$

Combining these facts establishes the inequality $\|\mathbf{A}\|_{2\to 2} \leq \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})}$. Now choose $\mathbf{x}$ to be an eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$ of $\mathbf{A}^*\mathbf{A}$, that is, $\mathbf{A}^*\mathbf{A}\mathbf{x} = \lambda_{\max}\mathbf{x}$. Then

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}\rangle = \langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \lambda_{\max}\langle \mathbf{x}, \mathbf{x}\rangle = \lambda_{\max}\|\mathbf{x}\|_2^2 \,.$$

This shows that the inequality derived above is sharp, and this completes the proof by noting that $\sigma_{\max}(\mathbf{A}) = \sqrt{\lambda(\mathbf{A}^*\mathbf{A})}$ by definition.

(b) For $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\|_1 = 1$, the triangle inequality gives

$$\|\mathbf{A}\mathbf{x}\|_p = \|\sum_{j=1}^n x_j \mathbf{a}_k\|_p \leq \sum_{k=1}^n |x_k| \|\mathbf{a}_k\|_p \leq \|\mathbf{x}\|_1 \max_{k\in[n]} \|\mathbf{a}_k\|_p \,. \tag{A.11}$$

This shows that $\|\mathbf{A}\|_{1\to p} \leq \max_{k\in[n]} \|\mathbf{a}_k\|_p$. Now choose $\mathbf{x} = \mathbf{e}_k$ to be the kth canonical unit vector with k being the index that realizes the maximum in the expression above. Then $\|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{a}_k\|_p = \max_{k\in[n]} \|\mathbf{a}_k\|_p$. This establishes the statement.

(c) For $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\|_\infty = 1$ we have

$$\|\mathbf{A}\mathbf{x}\|_\infty = \max_{j\in[m]} |\sum_{k=1}^n A_{jk}x_k| \leq \max_{j\in[m]} \sum_{k=1}^n |A_{jk}||x_k| \leq \|\mathbf{x}\|_\infty \max_{j\in[m]} \sum_{k=1}^n |A_{jk}| \,.$$

To see that this inequality is sharp in general, we choose an index $j \in [m]$ that realizes the maximum in the previous expression, and set $x_k = \operatorname{sgn}(\overline{A_{jk}}) = \overline{A_{jk}}/|A_{jk}|$ if $A_{jk} \neq 0$ and $x_k = 0$ if $A_{jk} = 0$. Then $\|\mathbf{x}\|_\infty = 1$ (unless $\mathbf{A} = 0$, in which case the statement is trivial) and

$$(\mathbf{A}\mathbf{x})_j = \sum_{k=1}^n A_{jk}x_k = \sum_{j=1}^n |A_{jk}| = \max_{j\in[m]} \sum_{k=1}^n |A_{jk}| \,.$$

Together with the inequality established above this shows the claim. □

Remark A.7. (a) The general identity

$$\|\mathbf{A}\|_{p\to q} = \|\mathbf{A}^*\|_{p'\to q'}$$

where $1/p + 1/p' = 1 = 1/q + 1/q'$ shows that (c) (as well as some more general statements) can also be deduced from (b).

(b) Computing the operator norms $\|\mathbf{A}\|_{\infty\to 1}$, $\|\mathbf{A}\|_{2\to 1}$ and $\|\mathbf{A}\|_{\infty\to 2}$ is known to be an NP hard problem, see [368]. (The cases $\|\mathbf{A}\|_{2\to 1}$ and $\|\mathbf{A}\|_{\infty\to 2}$, though not treated explicitly in this paper, follow from similar considerations.)

Lemma A.8. *For $\mathbf{A} \in \mathbb{C}^{m\times n}$ we have*

$$\|\mathbf{A}\|_{2\to 2} = \sup_{\|\mathbf{y}\|_2\leq 1} \sup_{\|\mathbf{x}\|_2\leq 1} |\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle| = \sup_{\|\mathbf{y}\|_2\leq 1} \sup_{\|\mathbf{x}\|_2\leq 1} \operatorname{Re}(\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle) \,. \tag{A.12}$$

If $\mathbf{B} \in \mathbb{C}^{n\times n}$ *is self-adjoint then*

$$\|\mathbf{B}\|_{2\to 2} = \sup_{\|\mathbf{x}\|_2\le 1} |\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle| \, .$$

Proof. The first statement follows immediately from (A.6). For the second claim, let $\mathbf{B} = \mathbf{U}^*\mathbf{D}\mathbf{U}$ be the eigenvalue decomposition of $\mathbf{B}$ with unitary $\mathbf{U}$ and diagonal matrix $\mathbf{D}$ with the eigenvalues λ_j of $\mathbf{B}$ on the diagonal. Then

$$\begin{aligned}\sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle| &= \sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{U}\mathbf{D}\mathbf{U}^*\mathbf{x}, \mathbf{x}\rangle| = \sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{D}\mathbf{U}^*\mathbf{x}, \mathbf{U}^*\mathbf{x}\rangle| \\ &= \sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{D}\mathbf{x}, \mathbf{x}\rangle| = \sup_{\|\mathbf{x}\|_2=1} |\sum_{j=1}^n \lambda_j |x_j|^2| = \max_{j\in[n]} |\lambda_j| \\ &= \|\mathbf{B}\|_{2\to 2} \, .\end{aligned}$$

For the identity $\sup_{\|\mathbf{x}\|_2=1} |\sum_{j=1}^n \lambda_j |x_j|^2| = \max_{j\in[n]} |\lambda_j|$ above, we observe on the one hand

$$|\sum_{j=1}^n \lambda_j |x_j|^2| \le \max_{j\in[n]} |\lambda_j| \sum_{j=1}^n |x_j|^2 \, .$$

One the other hand, if $\mathbf{x} = \mathbf{e}_{j_0}$ is the canonical unit vector corresponding to the index j_0 where $|\lambda_j|$ is maximal, we have $|\sum_{j=1}^n \lambda_j |x_j|^2| = |\lambda_{j_0}| = \max_{j\in[n]} |\lambda_j|$. This point completes the proof. □

Specializing the above identity to the rank-1 matrix $\mathbf{B} = \mathbf{u}\mathbf{u}^*$ for a vector $\mathbf{u} \in \mathbb{C}^n$ yields

$$\begin{aligned}\|\mathbf{u}\mathbf{u}^*\|_{2\to 2} &= \sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{u}\mathbf{u}^*\mathbf{x}, \mathbf{x}\rangle| = \sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{u}^*\mathbf{x}, \mathbf{u}^*\mathbf{x}\rangle| \\ &= \sup_{\|\mathbf{x}\|_2=1} |\langle \mathbf{x}, \mathbf{u}\rangle|^2 = \|\mathbf{u}\|_2^2 \, , \end{aligned} \tag{A.13}$$

where we also applied (A.6).

Lemma A.9. *(Schur test) Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$*. Then*

$$\|\mathbf{A}\|_{2\to 2} \le \sqrt{\|\mathbf{A}\|_{1\to 1} \|\mathbf{A}\|_{\infty\to\infty}} \, .$$

In particular, for a self-adjoint matrix $\mathbf{B} = \mathbf{B}^* \in \mathbb{C}^{n\times N}$

$$\|\mathbf{B}\|_{2\to 2} \le \|\mathbf{B}\|_{1\to 1} \, .$$

Proof. The statement follows immediately from the Riesz-Thorin interpolation theorem. For readers not familiar with interpolation theory we give a more elementary proof. By the Cauchy-Schwarz inequality, the jth entry of $\mathbf{A}\mathbf{x}$ satisfies

$$|(\mathbf{A}\mathbf{x})_j| \le \sum_{k=1}^n |x_k||A_{jk}| \le \big(\sum_{k=1}^n |x_k|^2 |A_{jk}|\big)^{1/2} \big(\sum_{\ell=1}^n |A_{j\ell}|\big)^{1/2} \, .$$

Summing this inequality yields

$$\begin{aligned}\|\mathbf{A}\mathbf{x}\|_2^2 &= \sum_{j=1}^m |(\mathbf{A}\mathbf{x})_j|^2 \le \sum_{j=1}^m \big(\sum_{k=1}^n |x_k|^2 |A_{jk}|\big)\big(\sum_{\ell=1}^n |A_{j\ell}|\big) \\ &\le \sum_{j=1}^m \big(\max_{k\in[n]} |A_{jk}| \sum_{k=1}^n |x_k|^2\big) \sum_{\ell=1}^n |A_{j\ell}| \\ &\le \big(\max_{j\in[m]} \sum_{k=1}^n |A_{jk}|\big)\big(\max_{\ell\in[n]} \sum_{j=1}^n |A_{j\ell}|\big) \sum_{k=1}^n |x_k|^2 \\ &= \|\mathbf{A}\|_{\infty\to\infty} \|\mathbf{A}\|_{1\to 1} \|\mathbf{x}\|_2^2\end{aligned}$$

by Lemma (A.6). This establishes the first claim. If $\mathbf{B} = \mathbf{B}^*$ is self-adjoint then $\|\mathbf{B}\|_{1\to 1} = \|\mathbf{B}\|_{\infty\to\infty}$ by Lemma (A.6), which implies the second claim. □

We note that the above inequality may rather be crude for certain matrices, which are important in the context of this book. For a general matrix, however, it cannot be improved further.

Lemma A.10. *The operator norm of a submatrix is bounded by the one of the whole matrix. More precisely, if* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *has the form*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(2)} \\ \mathbf{A}^{(3)} & \mathbf{A}^{(4)} \end{pmatrix}$$

for matrices $\mathbf{A}^{(\ell)}$*, then* $\|\mathbf{A}^{(\ell)}\|_{2\to 2} \le \|\mathbf{A}\|_{2\to 2}$ *for* $\ell = 1, \dots, 4$*. In particular, any entry of* $\mathbf{A}$ *satisfies* $|A_{jk}| \le \|\mathbf{A}\|_{2\to 2}$*.*

Proof. We give the proof for $\mathbf{A}^{(1)}$. The other cases are analogous. Let $\mathbf{A}^{(1)}$ be of size $m_1 \times n_1$. Then for $\mathbf{x}^{(1)} \in \mathbb{C}^{n_1}$ we have

$$\|\mathbf{A}^{(1)}\mathbf{x}^{(1)}\|_2^2 \le \|\mathbf{A}^{(1)}\mathbf{x}^{(1)}\|_2^2 + \|\mathbf{A}^{(3)}\mathbf{x}^{(1)}\|_2^2 = \left\| \begin{pmatrix} \mathbf{A}^{(1)} \\ \mathbf{A}^{(3)} \end{pmatrix} \mathbf{x}^{(1)} \right\|_2^2 = \left\| \mathbf{A} \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{0} \end{pmatrix} \right\|_2^2 .$$

The set T_1 of vectors $\begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{0} \end{pmatrix} \in \mathbb{C}^n$ with $\|\mathbf{x}^{(1)}\|_2 \le 1$ is contained in the set $T := \{\mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_2 \le 1\}$. Therefore, the supremum over $\mathbf{x}^{(1)} \in T_1$ above is bounded by $\sup_{\mathbf{x}\in T} \|\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{A}\|_{2\to 2}^2$. This concludes the proof. □

Remark A.11. The same result and proof also holds for the operator norms $\|\cdot\|_{p\to q}$, $1 \le p, q \le \infty$.

Gershgorin's disc theorem stated next provides information about the eigenvalues of a square matrix.

Theorem A.12. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a square matrix, and let λ be an eigenvalue. Then there exists an index $j \in [n]$ such that*

$$|\lambda - A_{jj}| \le \sum_{\ell\in[n]\setminus\{j\}} |A_{j\ell}| \,.$$

Proof. Let $\mathbf{u} \in \mathbb{C}^n$ be an eigenvector associated with λ and let $j \in [n]$ such that $|u_j|$ is maximal, i.e., $|u_j| = \|\mathbf{u}\|_\infty$. Then $\sum_{\ell\in[n]} A_{j\ell} = \lambda u_j$ and a rearrangement gives $\sum_{\ell\in[n]\setminus\{j\}} A_{j\ell}u_\ell = \lambda u_j - A_{jj}u_j$. The triangle inequality yields

$$|\lambda - A_{jj}|\,|u_j| \le \sum_{\ell\in[n]\setminus\{j\}} |A_{j\ell}|\,|u_\ell| \le \|\mathbf{u}\|_\infty \sum_{\ell\in[n]\setminus\{j\}} |A_{j\ell}| = |u_j| \sum_{\ell\in[n]\setminus\{j\}} |A_{j\ell}| \,.$$

Dividing by $|u_j|$ (which is nonzero by construction) yields the desired statement. □

More information on Gershgorin's theorem and its variations can be found, for instance, in the monograph [433].

The trace of a square matrix $\mathbf{B} \in \mathbb{C}^{n\times n}$ is the sum of its diagonal elements,

$$\operatorname{tr}(\mathbf{B}) = \sum_{j=1}^{n} B_{jj} \,.$$

The trace is cyclic, $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ for matrices of matching dimensions. It induces an inner product on the set of matrices by

$$\langle \mathbf{A}, \mathbf{B}\rangle_F := \operatorname{tr}(\mathbf{AB}^*) \,. \tag{A.14}$$

The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ is defined as

$$\|\mathbf{A}\|_F := \sqrt{\operatorname{tr}(\mathbf{A}^*\mathbf{A})} = \sqrt{\operatorname{tr}(\mathbf{AA}^*)} = \left(\sum_{j\in[m],k\in[n]} |A_{jk}|^2 \right)^{1/2} . \tag{A.15}$$

After identifying matrices on $\mathbb{C}^{m\times n}$ with vectors in $\mathbb{C}^{nm}$, the Frobenius norm is an ℓ_2-norm.

The operator norm on ℓ_2 is bounded by the Frobenius norm,

$$\|\mathbf{A}\|_{2\to 2} \le \|\mathbf{A}\|_F \,. \tag{A.16}$$

Indeed, for $\mathbf{x} \in \mathbb{C}^n$, the Cauchy-Schwarz inequality yields

$$\|\mathbf{Ax}\|_2^2 = \sum_{j=1}^{m} \big(\sum_{k=1}^{n} A_{jk}x_j \big)^2 \le \sum_{j=1}^{m} \big(\sum_{k=1}^{n} |x_j|^2 \big) \big(\sum_{\ell=1}^{n} |A_{j\ell}|^2 \big) = \|\mathbf{x}\|_2^2 \|\mathbf{A}\|_F^2 \,.$$

Next we consider bounds for the operator norm of the inverse of a square matrix.

Lemma A.13. *Let* $\mathbf{B} \in \mathbb{C}^{n\times n}$ *such that, for some* $\eta \in [0,1)$,

$$\|\mathbf{B} - \mathbf{Id}\|_{2\to 2} \leq \eta \,.$$

Then $\mathbf{B}$ *is invertible and* $\|\mathbf{B}^{-1}\|_{2\to 2} \leq (1-\eta)^{-1}$.

Proof. We first note that, for $\mathbf{H} = \mathbf{Id} - \mathbf{B}$, the Neumann series $\sum_{k=0}^{\infty} \mathbf{H}^k$ converges. Indeed, by the triangle inequality

$$\|\sum_{k=0}^{\infty} \mathbf{H}^k\|_{2\to 2} \leq \sum_{k=0}^{\infty} \|\mathbf{H}\|_{2\to 2}^k \leq \sum_{k=0}^{\infty} \eta^k \leq \frac{1}{1-\eta} \,.$$

Now observe that

$$(\mathbf{Id} - \mathbf{H}) \sum_{k=0}^{\infty} \mathbf{H}^k = \sum_{k=0}^{\infty} \mathbf{H}^k - \sum_{k=1}^{\infty} \mathbf{H}^k = \mathbf{Id}$$

by convergence of the Neumann series, and similarly $\sum_{k=0}^{\infty} \mathbf{H}^k(\mathbf{Id} - \mathbf{H}) = \mathbf{Id}$. Therefore, $\mathbf{Id} - \mathbf{H}$ is invertible and

$$(\mathbf{Id} - \mathbf{H})^{-1} = \sum_{k=0}^{\infty} \mathbf{H}^k \,.$$

This establishes the claim. □

A.2 The Singular Value Decomposition

While the concept of eigenvalues and eigenvectors applies only to square matrices, every (possibly rectangular) matrix possesses a singular value decomposition.

Proposition A.14. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$. *Then there exist unitary matrices* $\mathbf{U} \in \mathbb{C}^{m\times m}, \mathbf{V} \in \mathbb{C}^{n\times n}$, *and uniquely defined non-negative numbers* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0$, *called singular values, such that*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \quad \mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}}) \in \mathbb{R}^{m\times n} \,.$$

Remark A.15. Writing $\mathbf{U} = (\mathbf{u}_1|\cdots|\mathbf{u}_m)$ and $\mathbf{V} = (\mathbf{v}_1|\ldots|\mathbf{v}_n)$, the vectors $\mathbf{u}_\ell$ are called left singular vectors, while the $\mathbf{v}_\ell$ are called right singular vectors.

Proof. Let $\mathbf{v}_1 \in \mathbb{C}^n$ be a vector with $\|\mathbf{v}_1\|_2 = 1$ that realizes the maximum in the definition (A.8) of the operator norm $\|\mathbf{A}\|_{2\to 2}$, and set $\sigma_1 = \|\mathbf{A}\|_{2\to 2}$,

$$\|\mathbf{A}\mathbf{v}_1\|_2 = \|\mathbf{A}\|_{2\to 2} = \sigma_1 \,.$$

By compactness of the sphere $S^{n-1} = \{\mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_2 = 1\}$ such a vector $\mathbf{v}_1$ always exists. If $\sigma_1 = 0$ then $\mathbf{A} = 0$, and we can set $\sigma_\ell = 0$ for all

$\ell = 1, \ldots, \min\{m, n\}$, and $\mathbf{U}, \mathbf{V}$ to be arbitrary unitary matrices. Therefore, we assume $\sigma_1 > 0$ and set

$$\mathbf{u}_1 = \sigma_1^{-1} \mathbf{A} \mathbf{v}_1 \, .$$

We can extend $\mathbf{u}_1, \mathbf{v}_1$ to orthonormal bases in order to find unitary matrices $\mathbf{U}_1 = (\mathbf{u}_1 | \widetilde{\mathbf{U}}_1)$, $\mathbf{V}_1 = (\mathbf{v}_1 | \widetilde{\mathbf{V}}_1)$. Since $\widetilde{\mathbf{U}}_1^* \mathbf{A} \mathbf{v}_1 = \sigma_1 \widetilde{\mathbf{U}}_1^* \mathbf{u}_1 = 0$ the matrix $\mathbf{A}_1 = \mathbf{U}_1^* \mathbf{A} \mathbf{V}_1$ takes the form

$$\mathbf{A}_1 = \begin{pmatrix} \sigma_1 & \mathbf{b}^* \\ 0 & \mathbf{B} \end{pmatrix} \, ,$$

where $\mathbf{b}^* = \mathbf{u}_1^* \mathbf{A} \widetilde{\mathbf{V}}_1$ and $\mathbf{B} = \widetilde{\mathbf{U}}_1^* \mathbf{A} \widetilde{\mathbf{V}}_1 \in \mathbb{C}^{(m-1)\times(n-1)}$. It follows from

$$\|\mathbf{A}_1\|_{2\to 2} \sqrt{\sigma_1^2 + \|\mathbf{b}\|_2^2} \geq \left\| \mathbf{A}_1 \begin{pmatrix} \sigma_1 \\ \mathbf{b} \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \sigma_1^2 + \|\mathbf{b}\|_2^2 \\ \mathbf{B}\mathbf{w} \end{pmatrix} \right\|_2 \geq \sigma_1^2 + \|\mathbf{b}\|_2^2$$

that $\|\mathbf{A}_1\|_{2\to 2} \geq \sqrt{\sigma_1^2 + \|\mathbf{b}\|_2^2}$. But since $\mathbf{U}, \mathbf{V}$ are unitary we have $\|\mathbf{A}_1\|_{2\to 2} = \|\mathbf{A}\|_{2\to 2} = \sigma_1$, and therefore $\mathbf{b} = 0$. In conclusion

$$\mathbf{A}_1 = \mathbf{U}_1^* \mathbf{A} \mathbf{V}_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \mathbf{B} \end{pmatrix} \, .$$

With the same arguments we can further decompose $\mathbf{B} \in \mathbb{C}^{(m-1)\times(n-1)}$, and by induction we arrive at the stated singular value decomposition. □

From the previous proof it follows that the largest and smallest singular values satisfy

$$\begin{aligned} \sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A}) = \|\mathbf{A}\|_{2\to 2} &= \max_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 \, , \\ \sigma_{\min}(\mathbf{A}) = \sigma_p(\mathbf{A}) &= \min_{\|\mathbf{x}\| = 1} \|\mathbf{A}\mathbf{x}\|_2 \, . \end{aligned}$$

If $\mathbf{A}$ has rank r then its singular values satisfy $\sigma_1, \ldots, \sigma_r > 0$ while $\sigma_{r+1} = \sigma_{r+2} = \ldots = 0$. Sometimes it is more convenient to work with the reduced singular value decomposition. For $\mathbf{A}$ of rank r with (full) singular value decompositon $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ we take the submatrices $\widetilde{\mathbf{U}} \in \mathbb{C}^{m\times r}, \widetilde{\mathbf{V}} \in \mathbb{C}^{n\times r}$ such that $\mathbf{U} = (\widetilde{\mathbf{U}}|*)$ and $\mathbf{V} = (\widetilde{\mathbf{V}}|*)$, and $\widetilde{\mathbf{\Sigma}} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$. Writing $\mathbf{U} = (\mathbf{u}_1 | \ldots | \mathbf{u}_m)$, $\mathbf{V} = (\mathbf{v}_1 | \ldots | \mathbf{v}_n)$, we have

$$\mathbf{A} = \widetilde{\mathbf{U}} \widetilde{\mathbf{\Sigma}} \widetilde{\mathbf{V}}^* = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^* \, .$$

Given $\mathbf{A} \in \mathbb{C}^{m\times n}$ with reduced singular value decomposition $\mathbf{A} = \widetilde{\mathbf{U}}\mathbf{\Sigma}\widetilde{\mathbf{V}}^*$ we observe that

$$\begin{aligned} \mathbf{A}^*\mathbf{A} &= \widetilde{\mathbf{V}}\mathbf{\Sigma}\widetilde{\mathbf{U}}^*\widetilde{\mathbf{U}}\mathbf{\Sigma}\widetilde{\mathbf{V}}^* = \widetilde{\mathbf{V}}\mathbf{\Sigma}^2\widetilde{\mathbf{V}}^* \, , \\ \mathbf{A}\mathbf{A}^* &= \widetilde{\mathbf{U}}\mathbf{\Sigma}\widetilde{\mathbf{V}}^*\widetilde{\mathbf{V}}\mathbf{\Sigma}\widetilde{\mathbf{U}}^* = \widetilde{\mathbf{U}}\mathbf{\Sigma}^2\widetilde{\mathbf{U}}^* \, . \end{aligned}$$

Thus, we obtain the (reduced) eigenvalue decompositions of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$. In particular, the singular values $\sigma_j = \sigma_j(\mathbf{A})$ satisfy

$$\sigma_j(\mathbf{A}) = \sqrt{\lambda_j(\mathbf{A}^*\mathbf{A})} = \sqrt{\lambda_j(\mathbf{A}\mathbf{A}^*)}, \quad j = 1, \dots, \min\{m, n\}\,, \tag{A.17}$$

where $\lambda_1(\mathbf{A}^*\mathbf{A}) \geq \lambda_2(\mathbf{A}^*\mathbf{A}) \geq \cdots$ are the eigenvalues of $\mathbf{A}^*\mathbf{A}$ in decreasing order. Moreover, the left and right singular vectors listed in $\mathbf{U}, \mathbf{V}$ can be obtained from the eigenvalue decomposition of the positive semidefinite matrices $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$. (One can also prove existence of the singular value decomposition via the eigenvalue decompositions of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$.)

For the purposes in this book, the following observation is very useful.

Proposition A.16. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m \geq n$. If, for some $\delta \in [0, 1]$,*

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{Id}\|_{2\to 2} \leq \delta\,, \tag{A.18}$$

then the largest and smallest singular value of $\mathbf{A}$ satisfy

$$\sigma_{\max}(\mathbf{A}) \leq \sqrt{1+\delta}, \quad \sigma_{\min}(\mathbf{A}) \geq \sqrt{1-\delta}\,. \tag{A.19}$$

Conversely, if both inequalities in (A.19) *hold then* (A.18) *follows.*

Proof. By (A.17) the eigenvalues of $\mathbf{A}^*\mathbf{A}$ are the squared singular values of $\mathbf{A}$, $\lambda_j(\mathbf{A}^*\mathbf{A}) = \sigma_j^2(\mathbf{A})$, $j = 1, \dots, n$. The eigenvalues of $\mathbf{A}^*\mathbf{A} - \mathbf{Id}$ are given by $\sigma_j^2(\mathbf{A}) - 1$, and by (A.18)

$$\max\{\sigma_{\max}^2(\mathbf{A}) - 1, 1 - \sigma_{\min}^2(\mathbf{A})\} = \|\mathbf{A}^*\mathbf{A} - \mathbf{Id}\|_{2\to 2} \leq \delta\,.$$

This establishes the claim. □

The largest and smallest singular values are 1-Lipschitz functions with respect to the operator norm and the Frobenius norm.

Proposition A.17. *The smallest and largest singular values $\sigma_{\min}$, $\sigma_{\max}$, satisfy for all matrices $\mathbf{A}, \mathbf{B}$ of the same dimension,*

$$|\sigma_{\max}(\mathbf{A}) - \sigma_{\max}(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|_{2\to 2} \leq \|\mathbf{A} - \mathbf{B}\|_F, \tag{A.20}$$

$$|\sigma_{\min}(\mathbf{A}) - \sigma_{\min}(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|_{2\to 2} \leq \|\mathbf{A} - \mathbf{B}\|_F. \tag{A.21}$$

Proof. By the identification of the largest singular value with the operator norm we have

$$|\sigma_{\max}(\mathbf{A}) - \sigma_{\max}(\mathbf{B})| = |\|\mathbf{A}\|_{2\to 2} - \|\mathbf{B}\|_{2\to 2}| \leq \|\mathbf{A} - \mathbf{B}\|_{2\to 2}\,.$$

The inequality for the smallest singular is deduced as follows,

$$\begin{aligned}\sigma_{\min}(\mathbf{A}) &= \inf_{\|x\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \leq \inf_{\|\mathbf{x}\|_2=1} (\|\mathbf{B}\mathbf{x}\|_2 + \|(\mathbf{A}-\mathbf{B})\mathbf{x}\|_2) \\ &\leq \inf_{\|\mathbf{x}\|_2=1} (\|\mathbf{B}\mathbf{x}\|_2 + \|\mathbf{A}-\mathbf{B}\|_{2\to 2}) = \sigma_{\min}(\mathbf{B}) + \|\mathbf{A}-\mathbf{B}\|_{2\to 2}\,.\end{aligned}$$

Therefore, $\sigma_{\min}(\mathbf{A}) - \sigma_{\min}(\mathbf{B}) \leq \|\mathbf{A} - \mathbf{B}\|_{2\to 2}$ and (A.21) follows by symmetry. The estimates by the Frobenius norm in (A.20), (A.21) follow from the domination (A.16) of the operator norm by the Frobenius norm. □

Next we introduce the Moore-Penrose pseudo-inverse, which generalizes the usual inverse of a square matrix, but exists for any (possibly rectangular) matrix.

Definition A.18. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *of rank* r *with reduced singular value decomposition*

$$\mathbf{A} = \widetilde{\mathbf{U}}\mathbf{\Sigma}\widetilde{\mathbf{V}}^* = \sum_{j=1}^{r} \sigma_j(\mathbf{A})\mathbf{u}_j\mathbf{v}_j^* \,.$$

Then its Moore-Penrose pseudo-inverse $\mathbf{A}^\dagger \in \mathbb{C}^{n\times m}$ *is defined as*

$$\mathbf{A}^\dagger = \widetilde{\mathbf{V}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{U}}^* = \sum_{j=1}^{r} \sigma_j(\mathbf{A})^{-1}\mathbf{v}_j\mathbf{u}_j^* \,.$$

Note that the singular values satisfy $\sigma_j(\mathbf{A}) > 0$, $j = 1, \dots, r = \operatorname{rank}(\mathbf{A})$, so that $\mathbf{A}^\dagger$ is well-defined. If $\mathbf{A}$ is an invertible square matrix, then one easily checks that $\mathbf{A}^\dagger = \mathbf{A}^{-1}$. It follows immediately from the definition that $\mathbf{A}^\dagger$ has the same rank r as $\mathbf{A}$, and that

$$\sigma_{\max}(\mathbf{A}^\dagger) = \|\mathbf{A}^\dagger\|_{2\to 2} = \sigma_r^{-1}(\mathbf{A}) \,.$$

In particular, if $\mathbf{A}$ has full rank then

$$\|\mathbf{A}^\dagger\|_{2\to 2} = \sigma_{\min}^{-1}(\mathbf{A}) \,. \tag{A.22}$$

Moreover, $\sigma_r(\mathbf{A}^\dagger) = \sigma_{\max}^{-1}(\mathbf{A})$.

If $\mathbf{A}^*\mathbf{A} \in \mathbb{C}^{n\times n}$ is invertible (implying $m \geq n$) then

$$\mathbf{A}^\dagger = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* \,. \tag{A.23}$$

Indeed,

$$(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = (\widetilde{\mathbf{V}}\mathbf{\Sigma}^2\widetilde{\mathbf{V}}^*)^{-1}\widetilde{\mathbf{V}}\mathbf{\Sigma}\widetilde{\mathbf{U}}^* = \widetilde{\mathbf{V}}\mathbf{\Sigma}^{-2}\widetilde{\mathbf{V}}^*\widetilde{\mathbf{V}}\mathbf{\Sigma}\widetilde{\mathbf{U}}^* = \widetilde{\mathbf{V}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{U}}^* = \mathbf{A}^\dagger \,.$$

Similarly, if $\mathbf{A}\mathbf{A}^* \in \mathbb{C}^{m\times m}$ is invertible (so that necessarily $n \geq m$) then

$$\mathbf{A}^\dagger = \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1} \,. \tag{A.24}$$

The Moore-Penrose pseudo-inverse is closely connected to least squares problems as stated in Proposition A.21 below.

The singular values $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_{\min}(\mathbf{A}) \geq 0$ of a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ obey the useful variational characterization

$$\sigma_k(\mathbf{A}) = \max_{\mathcal{M}\subset\mathbb{C}^d, \dim\mathcal{M}=k} \; \min_{\mathbf{x}\in\mathcal{M}, \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \,.$$

This follows from the characterization of the eigenvalues $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ of a self-adjoint matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, known as *Courant–Fischer minimax theorem* or simply *minimax principle*, namely

$$\lambda_k(\mathbf{A}) = \max_{\mathcal{M}\subset\mathbb{C}^n, \dim\mathcal{M}=k} \min_{\mathbf{x}\in\mathcal{M}, \|\mathbf{x}\|_2=1} \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \,. \tag{A.25}$$

This characterization generalizes to the *Wielandt's minimax principle* for sums of eigenvalues stated next.

Lemma A.19. *If $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ are the eigenvalues of a self-adjoint matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, then, for any $1 \leq i_1 < \cdots < i_k \leq n$,*

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) = \max_{\substack{\mathcal{M}_1\subset\cdots\subset\mathcal{M}_k\subset\mathbb{C}^n \\ \dim\mathcal{M}_j=i_j}} \min_{\substack{(\mathbf{x}_1,\ldots,\mathbf{x}_k) \text{ orthonormal} \\ \mathbf{x}_j\in\mathcal{M}_j}} \sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \,.$$

Proof. Let $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ be an orthonormal basis of eigenvectors for the eigenvalues $\lambda_1 := \lambda_1(\mathbf{A}), \ldots, \lambda_n := \lambda_n(\mathbf{A})$. With $\mathcal{M}'_j := \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{i_j})$, we have $\dim\mathcal{M}'_j = i_j$. Setting $\mathbf{x}_j =: \sum_{i=1}^{i_j} \alpha_{i,j}\mathbf{u}_i \in \mathcal{M}'_i$ yields

$$\langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle = \sum_{i=1}^{i_j} \lambda_i \alpha_{i,j}^2 \geq \lambda_{i_j} \sum_{i=1}^{i_j} \alpha_{i,j}^2 = \lambda_{i_j} \|\mathbf{x}_j\|_2^2 \,.$$

It follows that

$$\begin{aligned} \max_{\substack{\mathcal{M}_1\subset\cdots\subset\mathcal{M}_k\subset\mathbb{C}^n \\ \dim\mathcal{M}_j=i_j}} \min_{\substack{(\mathbf{x}_1,\ldots,\mathbf{x}_k) \text{ orthonormal} \\ \mathbf{x}_j\in\mathcal{M}_j}} \sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \\ \geq \min_{\substack{(\mathbf{x}_1,\ldots,\mathbf{x}_k) \text{ orthonormal} \\ \mathbf{x}_j\in\mathcal{M}'_j}} \sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \geq \sum_{j=1}^{k} \lambda_{i_j} \,. \end{aligned}$$

For the reverse inequality, we need to prove that, given $\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_k \subset \mathbb{C}^n$ with $\dim\mathcal{M}_j = i_j$, $j \in [k]$, there exists an orthonormal system $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ with $\mathbf{x}_j \in \mathcal{M}_j$, $j \in [k]$, such that $\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \leq \sum_{j=1}^{k} \lambda_{i_j}$. For any $j \in [k]$, a dimensional argument guarantees the existence of a vector $\mathbf{v}_j$ in $\mathcal{M}_j \cap \operatorname{span}(\mathbf{u}_{i_j}, \ldots, \mathbf{u}_n)$. Applying the Gram–Schmidt orthonormalization process to the ordered system $(\mathbf{v}_1, \ldots, \mathbf{v}_k)$, we obtain an orthonormal basis $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ of $\mathcal{V} := \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ with $\mathbf{x}_j \in \mathcal{M}_j$ for all $j \in [k]$. Moreover, applying the Gram–Schmidt orthonormalization process to the ordered system $(\mathbf{v}_k, \ldots, \mathbf{v}_1)$, we obtain an orthonormal basis $(\mathbf{y}_1, \ldots, \mathbf{y}_k)$ of $\mathcal{V} = \operatorname{span}(\mathbf{v}_k, \ldots, \mathbf{v}_1)$ such that $\mathbf{y}_j \in \operatorname{span}(\mathbf{u}_{i_j}, \ldots, \mathbf{u}_n)$ for any $j \in [k]$. We denote by $\mathbf{A}_{|\mathcal{V}} : \mathcal{V} \to \mathcal{V}$ the restriction of the operator $\mathbf{A}$ to the subspace $\mathcal{V}$. Then $\operatorname{tr}(\mathbf{A}_{|\mathcal{V}})$ is the sum of the eigenvalues of $\mathbf{A}_{|\mathcal{V}}$. We obtain

$$\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle = \operatorname{tr}(\mathbf{A}_{|\mathcal{V}}) = \sum_{j=1}^{k} \langle \mathbf{A}\mathbf{y}_j, \mathbf{y}_j\rangle \leq \sum_{j=1}^{k} \lambda_{i_j} \,.$$

With $\mathbf{y}_j =: \sum_{i=i_j}^{n} \beta_{i,j}\mathbf{u}_i$ the last inequality follows from

$$\langle \mathbf{A}\mathbf{y}_j, \mathbf{y}_j \rangle = \sum_{i=i_j}^{n} \lambda_i \beta_{i,j}^2 \leq \lambda_{i_j} \sum_{i=i_j}^{n} \beta_{i,j}^2 = \lambda_{i_j} \|\mathbf{y}_j\|_2^2 = \lambda_{i_j} \ .$$

The proof is complete. □

Note that the case $i_1 = 1, \ldots, i_k = k$ of Wielandt's minimax principle reads

$$\sum_{j=1}^{k} \lambda_j(\mathbf{A}) = \max_{(\mathbf{x}_1,\ldots,\mathbf{x}_k) \text{ orthonormal}} \sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j \rangle \ . \tag{A.26}$$

With this observation at hand, we can deduce from Wielandt's minimax principle that, for any $1 \leq i_1 < \cdots < i_k \leq n$,

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}+\mathbf{B}) \leq \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) + \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) \ , \tag{A.27}$$

where $(\lambda_j(\mathbf{A}))_{j\in[n]}$, $(\lambda_j(\mathbf{B}))_{j\in[n]}$, $(\lambda_j(\mathbf{A}+\mathbf{B}))_{j\in[n]}$ denote the eigenvalues of the self-adjoint matrices $\mathbf{A}, \mathbf{B}, \mathbf{A}+\mathbf{B} \in \mathbb{C}^{n\times n}$ arranged in nonincreasing order. This inequality, known as *Lidskii's inequality* (*Weyl's inequality* in the case $k = 1$), allows to establish the following lemma, where we assume $m \geq n$ without loss of generality.

Lemma A.20. *If* $\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_n(\mathbf{X}) \geq 0$ *and* $\sigma_1(\mathbf{Y}) \geq \cdots \geq \sigma_n(\mathbf{Y}) \geq 0$, *are the singular values of* $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{m\times n}$, $m \geq n$, *then, for any* $k \in [n]$,

$$\sum_{j=1}^{k} |\sigma_j(\mathbf{X}) - \sigma_j(\mathbf{Y})| \leq \sum_{j=1}^{k} \sigma_j(\mathbf{X}-\mathbf{Y}).$$

Proof. The *self-adjoint dilations* $S(\mathbf{X}), S(\mathbf{Y}) \in \mathbb{C}^{(m+n)\times(m+n)}$ defined by

$$S(\mathbf{X}) = \begin{bmatrix} 0 & \mathbf{X} \\ \mathbf{X}^* & 0 \end{bmatrix} \quad \text{and} \quad S(\mathbf{Y}) = \begin{bmatrix} 0 & \mathbf{Y} \\ \mathbf{Y}^* & 0 \end{bmatrix}$$

have eigenvalues

$$\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_n(\mathbf{X}) \geq 0 = \cdots = 0 \geq -\sigma_n(\mathbf{X}) \geq \cdots \geq -\sigma_1(\mathbf{X}),$$
$$\sigma_1(\mathbf{Y}) \geq \cdots \geq \sigma_n(\mathbf{Y}) \geq 0 = \cdots = 0 \geq -\sigma_n(\mathbf{Y}) \geq \cdots \geq -\sigma_1(\mathbf{Y}).$$

Therefore, for any $k \in [n]$, there exists a subset I_k of $[m+n]$ with size k such that

$$\sum_{j=1}^{k} |\sigma_j(\mathbf{X}) - \sigma_j(\mathbf{Y})| = \sum_{j\in I_k} \big(\lambda_j(S(\mathbf{X})) - \lambda_j(S(\mathbf{Y}))\big).$$

Using (A.27) with $\mathbf{A} = S(\mathbf{Y})$ and $\mathbf{B} = S(\mathbf{X}-\mathbf{Y})$, so that $\mathbf{A}+\mathbf{B} = S(\mathbf{X})$, yields

$$\sum_{j=1}^{k} |\sigma_j(\mathbf{X}) - \sigma_j(\mathbf{Y})| \leq \sum_{j \in I_k} \lambda_j(S(\mathbf{X}-\mathbf{Y})) \leq \sum_{j \in I_k} \sigma_j(\mathbf{X}-\mathbf{Y}).$$

The proof is complete. □

Lemma A.20 implies in particular the triangle inequality

$$\sum_{j=1}^{\min\{m,n\}} \sigma_j(\mathbf{A}+\mathbf{B}) \leq \sum_{j=1}^{\min\{m,n\}} \sigma_j(\mathbf{A}) + \sum_{j=1}^{\min\{m,n\}} \sigma_j(\mathbf{B})$$

for all $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times n}$. Moreover, it is easy to verify that $\sum_{j=1}^{\min\{m,n\}} \sigma_j(\mathbf{A}) = 0$ if and only if $\mathbf{A} = 0$ and that $\sum_{j=1}^{n} \sigma_j(\lambda\mathbf{A}) = |\lambda| \sum_{j=1}^{n} \sigma_j(\mathbf{A})$. These three properties show that the expression

$$\|\mathbf{A}\|_* := \sum_{j=1}^{\min\{m,n\}} \sigma_j(\mathbf{A}), \qquad \mathbf{A} \in \mathbb{C}^{m\times n}\,, \tag{A.28}$$

defines a norm on $\mathbb{C}^{m\times n}$, called the *nuclear norm*. It is also referred to as the Schatten 1-norm, in view of the fact that, for all $1 \leq p \leq \infty$, the expression

$$\|\mathbf{A}\|_{S_p} := \left[\sum_{j=1}^{\min\{m,n\}} \sigma_j(\mathbf{A})^p\right]^{1/p}, \qquad \mathbf{A} \in \mathbb{C}^{m\times n},$$

defines a norm on $\mathbb{C}^{m\times n}$, called the Schatten p-norm. We note in particular that it reduces to the Frobenius norm for $p = 2$ and to the operator norm for $p = \infty$.

A.3 Least Squares Problems

Let us first connect least-squares problems with the Moore-Penrose pseudo-inverse in Definition A.18.

Proposition A.21. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, $\mathbf{y} \in \mathbb{C}^m$. Define $\mathcal{M} \subset \mathbb{C}^n$ to be the set of minimizers of $\mathbf{x} \mapsto \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2$. The optimization problem*

$$\min_{\mathbf{x}\in\mathcal{M}} \|\mathbf{x}\|_2 \tag{A.29}$$

has the unique solution $\mathbf{x}^\sharp = \mathbf{A}^\dagger\mathbf{y}$.

Proof. Let $r = \operatorname{rank}(\mathbf{A})$. Then the (full) singular value decomposition of $\mathbf{A}$ can be written $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$ with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{(r)} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times n}\,,$$

where $\mathbf{\Sigma}^{(r)} \in \mathbb{R}^{r\times r}$ is the diagonal matrix containing the non-zero singular values $\sigma_1(\mathbf{A}), \ldots, \sigma_r(\mathbf{A})$ on its diagonal. We introduce

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \mathbf{V}^*\mathbf{x}, \quad \mathbf{z}_1 \in \mathbb{C}^r \ ,$$
$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \mathbf{U}^*\mathbf{y}, \quad \mathbf{b}_1 \in \mathbb{C}^r \ .$$

Since the ℓ_2-norm is invariant under orthogonal transformations we have

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{U}^*(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2 = \|\mathbf{\Sigma}\mathbf{V}^*\mathbf{x} - \mathbf{b}\|_2 = \left\| \begin{pmatrix} \mathbf{\Sigma}^{(r)}\mathbf{z}_1 - \mathbf{b}_1 \\ -\mathbf{b}_2 \end{pmatrix} \right\|_2 .$$

This ℓ_2-norm is minimized for $\mathbf{z}_1 = (\mathbf{\Sigma}^{(r)})^{-1}\mathbf{b}_1$ and arbitrary $\mathbf{z}_2$. Fixing $\mathbf{z}_1$, by unitarity of $\mathbf{V}$, $\|\mathbf{x}\|_2^2 = \|\mathbf{V}^*\mathbf{x}\|_2^2 = \|\mathbf{z}\|_2^2 = \|\mathbf{z}_1\|_2^2 + \|\mathbf{z}_2\|_2^2$ is minimized for $\mathbf{z}_2 = 0$. Altogether, the minimizer $\mathbf{x}^\sharp$ of (A.29) is given by

$$\mathbf{x} = \mathbf{V}\begin{pmatrix} \mathbf{z}_1 \\ 0 \end{pmatrix} = \mathbf{V}\begin{pmatrix} (\mathbf{\Sigma}^{(r)})^{-1} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{U}^*\mathbf{y} = \mathbf{A}^\dagger\mathbf{y}$$

by definition of the Moore-Penrose pseudo-inverse. □

Let us highlight two special cases.

Corollary A.22. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m \geq n$ be of full rank, $\mathbf{y} \in \mathbb{C}^m$. Then the least squares problem*

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \tag{A.30}$$

has the unique solution $\mathbf{x}^\sharp = \mathbf{A}^\dagger\mathbf{y}$.

Note that the minimizer of (A.30) is the orthogonal projection of $\mathbf{y}$ onto the range of $\mathbf{A}$. Consequently, $\mathbf{A}\mathbf{A}^\dagger$ is the orthogonal projection onto the range of $\mathbf{A}$. Since $\mathbf{A}$ is assumed to be of full rank and $m \geq n$, the matrix $\mathbf{A}^*\mathbf{A}$ is invertible so that by (A.23) $\mathbf{A}^\dagger = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$. Therefore, $\mathbf{x}^\sharp = \mathbf{A}^\dagger\mathbf{y}$ satisfies the normal equation

$$\mathbf{A}^*\mathbf{A}\mathbf{x}^\sharp = \mathbf{A}^*\mathbf{y} \ . \tag{A.31}$$

Corollary A.23. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, $n \geq m$ be of full rank, $\mathbf{y} \in \mathbb{C}^m$. Then the least squares problem*

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{x}\|_2 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} \tag{A.32}$$

has the unique solution $\mathbf{x}^\sharp = \mathbf{A}^\dagger\mathbf{y}$.

By (A.24) we have $\mathbf{A}^\dagger = \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}$ if $\mathbf{A}$ is of full rank (and $n \geq m$). Therefore, in the situation of the previous corollary the minimizer $\mathbf{x}^\#$ of (A.32) satisfies the normal equation of the second kind

$$\mathbf{x}^\sharp = \mathbf{A}^*\mathbf{b}, \quad \text{where } \mathbf{A}\mathbf{A}^*\mathbf{b} = \mathbf{y} \,. \tag{A.33}$$

We can also treat the weighted ℓ_2-minimization problem

$$\min_{\mathbf{z}\in\mathbb{C}^n} \|\mathbf{z}\|_{2,\mathbf{w}} = \left(\sum_{j=1}^n |z_j|^2 w_j\right)^{-1/2} \quad \text{subject to } \mathbf{A}\mathbf{z} = \mathbf{y} \,, \tag{A.34}$$

where $\mathbf{w} = (w_j)$ is a sequence of positive weights $w_j > 0$. Introducing the diagonal matrix $\mathbf{D}_\mathbf{w} = \operatorname{diag}(w_j, j \in [n]) \in \mathbb{R}^{n\times n}$, and making the substitution $\mathbf{x} = \mathbf{D}_\mathbf{w}^{1/2}\mathbf{z}$, the minimizer $\mathbf{z}^\#$ of (A.34) is related to the minimizer $\mathbf{x}^\sharp$ of

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{x}\|_2 \quad \text{subject to } \mathbf{A}\mathbf{D}_\mathbf{w}^{-1/2}\mathbf{x} = \mathbf{y} \,,$$

via

$$\mathbf{z}^\sharp = \mathbf{D}_\mathbf{w}^{-1/2}\mathbf{x}^\# = \mathbf{D}_\mathbf{w}^{-1/2}(\mathbf{A}\mathbf{D}_\mathbf{w}^{-1/2})^\dagger\mathbf{y} \,. \tag{A.35}$$

In particular, if $n \geq m$ and $\mathbf{A}$ is of full rank then

$$\mathbf{z}^\sharp = \mathbf{D}_\mathbf{w}^{-1}\mathbf{A}^*(\mathbf{A}\mathbf{D}_\mathbf{w}^{-1}\mathbf{A}^*)^{-1}\mathbf{y} \,. \tag{A.36}$$

Proposition A.24. *A vector $\mathbf{x} \in \mathbb{C}^n$ is a minimizer of* (A.34) *if and only if*

$$\operatorname{Re}(\langle \mathbf{x}, \mathbf{v}\rangle_\mathbf{w}) = 0 \quad \textit{for all } \mathbf{v} \in \ker \mathbf{A} \,, \tag{A.37}$$

where $\langle \mathbf{x}, \mathbf{v}\rangle_\mathbf{w} = \sum_{j=1}^n x_j \overline{v}_j w_j$.

Proof. Take $\mathbf{x}$ with $\mathbf{A}\mathbf{x} = \mathbf{y}$. Then a vector $\mathbf{z} \in \mathbb{C}^n$ is feasible for (A.34) if and only if it can be written as $\mathbf{z} = \mathbf{x} + \mathbf{v}$ with $\mathbf{v} \in \ker \mathbf{A}$. For $t \in \mathbb{R}$ and $\mathbf{v} \in \ker \mathbf{A}$ consider

$$\|\mathbf{x} + t\mathbf{v}\|_{2,\mathbf{w}}^2 = \|\mathbf{x}\|_{2,\mathbf{w}}^2 + t^2\|\mathbf{v}\|_{2,\mathbf{w}}^2 + 2t\operatorname{Re}(\langle \mathbf{x}, \mathbf{v}\rangle_\mathbf{w}) \,. \tag{A.38}$$

Therefore, if $\operatorname{Re}(\langle \mathbf{x}, \mathbf{v}\rangle_\mathbf{w}) = 0$ then $t = 0$ is the minimizer of $t \mapsto \|\mathbf{x} + t\mathbf{v}\|_{2,\mathbf{w}}$. Consequently, (A.37) implies that $\mathbf{x}$ is a minimizer of (A.32). Conversely, if $\mathbf{x}$ is a minimizer of (A.32) then $t = 0$ is a minimizer of $t \mapsto \|\mathbf{x} + t\mathbf{v}\|_{2,\mathbf{w}}$ for all $\mathbf{v} \in \ker A$. However, if $\operatorname{Re}(\langle \mathbf{x}, \mathbf{v}\rangle_\mathbf{w})$ would be non-zero then by (A.38) we could find a non-zero t sufficiently close to 0 and of opposite sign as $\operatorname{Re}(\langle \mathbf{x}, \mathbf{v}\rangle_\mathbf{w})$ such that $\|\mathbf{x} + t\mathbf{v}\|_{2,\mathbf{w}} < \|\mathbf{x}\|_2$, a contradiction to $\mathbf{x}$ being a minimizer. □

While (A.31) and (A.33) suggest that one may solve least squares problems simply via solving the normal equations with any method, for instance, Gauss elimination, it is of advantage for numerical reasons to use specialized methods for least squares problems. The reference [41] provides an overview on various approaches. We shortly mention the prominent method of solving least squares problems via the QR decomposition.

For any matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$, $m \geq n$, there exists a unitary matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{m\times m}$ with non-negative diagonal entries such that

$$\mathbf{A} = \mathbf{Q} \begin{pmatrix} \mathbf{R} \\ 0 \end{pmatrix} .$$

We refer to [41, Theorem 1.3.1] for existence, and to [41] in general for methods of computing the QR decomposition. Consider the least squares problem (A.30). By unitarity of $\mathbf{Q}$ we have

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{Q}^*\mathbf{A}\mathbf{x} - \mathbf{Q}^*\mathbf{y}\|_2 = \left\| \begin{pmatrix} \mathbf{R} \\ 0 \end{pmatrix} \mathbf{x} - \mathbf{Q}^*\mathbf{y} \right\|_2 .$$

Partitioning $\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \mathbf{Q}^*\mathbf{y}$ with $\mathbf{b}_1 \in \mathbb{C}^m$, we solve the equation $\mathbf{R}\mathbf{x}_1 = \mathbf{b}_1$, $\mathbf{x}_1 \in \mathbb{C}^m$ via simple backward elimination (recall that $\mathbf{R}$ is upper triangular). Set $\mathbf{x}_2 = \mathbf{b}_2 \in \mathbb{C}^{n-m}$ and $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$. Then $\mathbf{x}$ solves (A.30).

In the context of orthogonal matching pursuit we encounter sequences of optimization problems of the type (A.30), where in each step a new column is added to $\mathbf{A}$. In such a situation it is beneficial to keep the QR-decomposition of $\mathbf{A}$. It is numerically cheap to update the QR-decomposition when a new column is added, see [41, Section 3.2.4] for details.

Also the least squares problem (A.32) maybe solved via the QR decomposition of $\mathbf{A}^*$, see [41, Theorem 1.3.3] for details.

If $\mathbf{A}$ and $\mathbf{A}^*$ have fast matrix vector multiplication algorithms (for instance, if one can make use of the Fast Fourier transform, or if $\mathbf{A}$ is sparse), then iterative algorithms for least squares problems are fast alternatives to QR decompositions. Conjugate gradients [41, 198] and especially the variant in [262] fall into this class of algorithms.

A.4 Vandermonde matrices

The *Vandermonde* matrix associated with $x_0, x_1, \ldots, x_n \in \mathbb{C}$ is defined as

$$\mathbf{V} := \mathbf{V}(x_0, x_1, \ldots, x_n) := \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} . \tag{A.39}$$

Theorem A.25. *The determinant of the Vandermonde matrix* (A.39) *equals*

$$\det(\mathbf{V}) = \prod_{0 \le k < \ell \le n} (x_\ell - x_k).$$

Proof. The proof can be done by induction on $n \ge 1$. For $n = 1$, the result is clear. For $n \ge 2$, we remark that $\det(\mathbf{V}(x_0, x_1, \ldots, x_n))$ is a polynomial in x_n which has degree at most n and which vanishes at $x_0, \ldots, x_{n-1}$. Therefore, we have

$$\det(\mathbf{V}(x_0, x_1, \ldots, x_n)) = c \prod_{0 \le k < n} (x_n - x_k) \tag{A.40}$$

for some constant c depending on $x_1, \ldots, x_{n-1}$. We notice that the constant c is the coefficient of x_n^n in $\det(V(x_0, x_1, \ldots, x_n))$. We then observe, by expanding the determinant of $V(x_0, x_1, \ldots, x_n)$ with respect to its last row, that $c = \det(V(x_0, x_1, \ldots, x_{n-1}))$. Using the induction hypothesis to substitute the value of c in (A.40) concludes the proof. □

Let us now establish the more involved result on the total positivity of Vandermonde matrices when $x_n > \cdots x_1 > x_0 > 0$.

Theorem A.26. *The Vandermonde matrix* (A.39) *is* totally positive, *i.e., for any sets $I, J \subseteq [n+1]$ of same cardinality,*

$$\det \mathbf{V}_{I,J} > 0,$$

where $\mathbf{V}_{I,J}$ represents the submatrix of $\mathbf{V}$ with rows indexed by I and columns indexed by J.

We start with the following lemma, known as *Descartes' rule of signs.*

Lemma A.27. *For a polynomial $p(x) = a_n x^n + \cdots + a_1 x + a_0 \neq 0$, the number $Z(p)$ of positive zeros of p and the number $S(a) := \mathrm{card}(\{i \in [n] : a_{i-1} a_i < 0\})$ of sign changes of $a = (a_0, a_1, \ldots, a_n)$ satisfy*

$$Z(p) \le S(a).$$

Proof. We proceed by induction on $n \ge 1$. For $n = 1$, the desired result is clear. Let us now assume that the result holds up to an integer $n-1$, $n \ge 2$. We want to establish that, given $p(x) = a_n x^n + \cdots + a_1 x + a_0 \neq 0$, we have $Z(p) \le S(a)$. We suppose that $a_0 \neq 0$, otherwise the result is clear from the induction hypothesis. Changing p in $-p$ if necessary, we may assume $a_0 > 0$. Now let k be the smallest positive integer such that $a_k \neq 0$ — the result is clear of no such k exists. We separate two cases.

1. $a_0 > 0$ and $a_k < 0$.
   The result follows from Rolle's theorem and the induction hypothesis via
   $$Z(p) \le Z(p') + 1 \le S(a_1, \ldots, a_n) + 1 = S(a_0, a_1, \ldots, a_n).$$
2. $a_0 > 0$ and $a_k > 0$.
   Let t denote the smallest positive zero of p — again te result is clear if no such t exists. Let us assume that p' does not vanish on $(0, t)$, Since $p'(0) = k a_k > 0$, we derive that $p'(x) > 0$ for all $x \in (0, t)$. If follows that $a_0 = p(0) < p(t) = 0$, which is absurd. Therefore, there must be a zero of p' in $(0, t)$. Taking into account the zeros of p' guaranteed by Rolle's theorem, the result follows from the induction hypothesis via
   $$Z(p) \le Z(p') \le S(a_1, \ldots, a_n) = S(a_0, a_1, \ldots, a_n).$$

This concludes the inductive proof. □

Proof (of Theorem A.26). We will prove by induction on $1 \leq k \leq n$ that

$$\det \begin{bmatrix} x_{i_1}^{j_1} & x_{i_1}^{j_2} & \cdots & x_{i_1}^{j_k} \\ x_{i_2}^{j_1} & x_{i_2}^{j_2} & \cdots & x_{i_2}^{j_k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{i_k}^{j_1} & x_{i_k}^{j_2} & \cdots & x_{i_k}^{j_k} \end{bmatrix} > 0$$

for all $0 < x_0 < x_1 < \cdots < x_n$ and for all $0 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $0 \leq j_1 < j_2 < \cdots < j_k \leq n$. For $k = 1$, this is nothing else than the positivity of the x_i's. Let us now suppose that the result holds up to an integer $k-1$, $2 \leq k \leq n$, and assume that

$$\det \begin{bmatrix} x_{i_1}^{j_1} & x_{i_1}^{j_2} & \cdots & x_{i_1}^{j_k} \\ x_{i_2}^{j_1} & x_{i_2}^{j_2} & \cdots & x_{i_2}^{j_k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{i_k}^{j_1} & x_{i_k}^{j_2} & \cdots & x_{i_k}^{j_k} \end{bmatrix} \leq 0 \tag{A.41}$$

for some $0 < x_0 < x_1 < \cdots < x_n$, and for some $0 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $0 \leq j_1 < j_2 < \cdots < j_k \leq n$. We introduce the polynomial p defined by

$$p(x) := \det \begin{bmatrix} x_{i_1}^{j_1} & x_{i_1}^{j_2} & \cdots & x_{i_1}^{j_k} \\ x_{i_2}^{j_1} & x_{i_2}^{j_2} & \cdots & x_{i_2}^{j_k} \\ \vdots & \vdots & \cdots & \vdots \\ x^{j_1} & x^{j_2} & \cdots & x^{j_k} \end{bmatrix}.$$

Expanding with respect to the last row and invoking Descartes' rule of signs, we observe that $Z(p) \leq k-1$. Since the polynomial p vanishes at the positive points $x_{i_1}, \ldots, x_{i_{k-1}}$, it cannot vanish anywhere else. The assumption (A.41) then implies that $p(x) < 0$ for all $x > x_{i_{k-1}}$. But this contradicts the induction hypothesis, because

$$\lim_{x \to +\infty} \frac{p(x)}{x^{j_k}} = \det \begin{bmatrix} x_{i_1}^{j_1} & x_{i_1}^{j_2} & \cdots & x_{i_1}^{j_{k-1}} \\ x_{i_2}^{j_1} & x_{i_2}^{j_2} & \cdots & x_{i_2}^{j_{k-1}} \\ \vdots & \vdots & \cdots & \vdots \\ x_{i_{k-1}}^{j_1} & x_{i_{k-1}}^{j_2} & \cdots & x_{i_{k-1}}^{j_{k-1}} \end{bmatrix} > 0.$$

Thus, we have shown that the desired result holds for the integer k, and this concludes the inductive proof. □

A.5 Matrix Functions

In this section we consider functions of self-adjoint matrices and some of their basic properties. We recall that a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is called self-adjoint

if $\mathbf{A} = \mathbf{A}^*$, and positive semidefinite, if additionally $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \geq 0$ for all $\mathbf{x} \in \mathbb{C}^n$. If $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle > 0$ for all $\mathbf{x} \neq 0$ then $\mathbf{A}$ is called positive definite. For two self-adjoint matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ we write $\mathbf{A} \preccurlyeq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive semidefinite, and $\mathbf{A} \prec \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive definite. Moreover, we also use the notation $\mathbf{B} \succcurlyeq \mathbf{A}$ if $\mathbf{A} \preccurlyeq \mathbf{B}$ and $\mathbf{B} \succ \mathbf{A}$ if $\mathbf{A} \prec \mathbf{B}$.

A self-adjoint matrix $\mathbf{A}$ possesses an eigenvalue decomposition of the form

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^* \,,$$

where $\mathbf{U} \in \mathbb{C}^{n\times n}$ is unitary and $\mathbf{D} = \mathrm{diag}(\lambda)$ is a diagonal matrix formed with the eigenvalues $\lambda_j \in \mathbb{R}, j \in [n]$, of $\mathbf{A}$ (repeated according to their multiplicities). For a function $f : I \to \mathbb{R}$, $I \subset \mathbb{R}$ such that I contains the eigenvalues of $\mathbf{A}$, we define $f(\mathbf{A}) \in \mathbb{C}^{n\times n}$ via the spectral mapping

$$f(\mathbf{A}) = \mathbf{U} f(\mathbf{D}) \mathbf{U}^* \,, \quad f(\mathbf{D}) = \mathrm{diag}\left(f(\lambda_j), j = 1, \ldots, n\right) \,. \tag{A.42}$$

It is simple to check that for polynomials f, this definition coincides with the natural one. For instance, if $f(t) = t^2$, then by unitarity,

$$f(\mathbf{A}) = \mathbf{U}\mathbf{D}^2\mathbf{U}^* = \mathbf{U}\mathbf{D}\mathbf{U}^*\mathbf{U}\mathbf{D}\mathbf{U}^* = \mathbf{A}^2 \,.$$

Clearly, $f(\mathbf{A})$ is a self-adjoint matrix again. Moreover, if $f(x) \leq g(x)$ for all $x \in [a, b]$ then

$$f(\mathbf{A}) \preccurlyeq g(\mathbf{A}) \,, \tag{A.43}$$

for all $\mathbf{A}$ with eigenvalues contained in $[a, b]$. It is a straightforward consequence of the definition that for a block diagonal matrix with self-adjoint blocks $\mathbf{A}_j$, $j \in [L]$ on the diagonal

$$f\begin{pmatrix} \mathbf{A}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{A}_2 & 0 & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{A}_L \end{pmatrix} = \begin{pmatrix} f(\mathbf{A}_1) & 0 & \cdots & 0 \\ 0 & f(\mathbf{A}_2) & 0 & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & f(\mathbf{A}_L) \end{pmatrix} \,. \tag{A.44}$$

Moreover, if $\mathbf{A}$ commutes with $\mathbf{B}$, i.e., $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$, then also $f(\mathbf{A})$ commutes with $\mathbf{B}$, $f(\mathbf{A})\mathbf{B} = \mathbf{B}f(\mathbf{A})$.

The *matrix exponential function* of a self-adjoint matrix $\mathbf{A}$ maybe defined by applying (A.42) with the function $f(x) = e^x$, or equivalently via the power series

$$e^{\mathbf{A}} := \exp(\mathbf{A}) := \mathbf{Id} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{A}^k \,. \tag{A.45}$$

(The power series definition actually applies to any square, not necessarily self-adjoint, matrix.) The matrix exponential of a self-adjoint matrix is always positive definite by (A.43). Moreover, it follows from $1 + x \leq e^x$ and again (A.43) that, for self-adjoint $\mathbf{A}$,

$$\mathbf{Id} + \mathbf{A} \preccurlyeq \exp(\mathbf{A}) \,. \tag{A.46}$$

Lemma A.28. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *commute, i.e.,* $\mathbf{AB} = \mathbf{BA}$*, then*

$$\exp(\mathbf{A} + \mathbf{B}) = \exp(\mathbf{A}) \exp(\mathbf{B}) \; .$$

Proof. If $\mathbf{A}$ and $\mathbf{B}$ commute then

$$\frac{1}{k!}(\mathbf{A} + \mathbf{B})^k = \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} \mathbf{A}^j \mathbf{B}^{k-j} = \sum_{j=0}^{k} \frac{\mathbf{A}^j}{j!} \frac{\mathbf{B}^{k-j}}{(k-j)!} \; .$$

Therefore,

$$\begin{aligned} \exp(\mathbf{A} + \mathbf{B}) &= \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{A} + \mathbf{B})^k = \sum_{k=0}^{\infty} \sum_{j=0}^{k} \frac{\mathbf{A}^j}{j!} \frac{\mathbf{B}^{k-j}}{(k-j)!} \\ &= \sum_{j=0}^{\infty} \sum_{k=j}^{\infty} \frac{\mathbf{A}^j}{j!} \frac{\mathbf{B}^{k-j}}{(k-j)!} = \sum_{j=0}^{\infty} \frac{1}{j!} \mathbf{A}^j \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \mathbf{B}^{\ell} = \exp(\mathbf{A}) \exp(\mathbf{B}) \; . \end{aligned}$$

This yields the claim. □

This lemma fails in the general case, when $\mathbf{A}$ and $\mathbf{B}$ do not commute.

Corollary A.29. *The matrix exponential* $\exp(\mathbf{A})$ *is invertible for any square matrix* $\mathbf{A}$*, and*

$$\exp(\mathbf{A})^{-1} = \exp(-\mathbf{A}) \; .$$

Proof. Clearly, $\mathbf{A}$ and $-\mathbf{A}$ commute, so that by the previous lemma

$$\exp(\mathbf{A}) \exp(-\mathbf{A}) = \exp(\mathbf{A} - \mathbf{A}) = \exp(\mathbf{0}) = \mathbf{Id} \; ,$$

and similarly $\exp(-\mathbf{A}) \exp(\mathbf{A}) = \mathbf{Id}$. □

Of special interest is the *trace exponential*

$$\operatorname{tr} \exp : \mathbf{A} \mapsto \operatorname{tr} \exp(\mathbf{A}) \; . \tag{A.47}$$

The trace exponential is monotone with respect to the semidefinite order. Indeed, for selfadjoint $\mathbf{A}, \mathbf{B}$ we have

$$\operatorname{tr} \exp \mathbf{A} \leq \operatorname{tr} \exp \mathbf{B} \quad \text{whenever } \mathbf{A} \preccurlyeq \mathbf{B} \; . \tag{A.48}$$

This fact follows from a more general statement.

Proposition A.30. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a non-decreasing function, and* $\mathbf{A}, \mathbf{B}$ *be self-adjoint matrices. Then* $\mathbf{A} \preccurlyeq \mathbf{B}$ *implies*

$$\operatorname{tr} f(\mathbf{A}) \leq \operatorname{tr} f(\mathbf{B}) \; .$$

Proof. It follows from the minimax principle in Lemma A.19 that the ordered eigenvalues $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots$ and $\lambda_1(\mathbf{B}) \geq \lambda_2(\mathbf{B}) \geq \cdots$ of $\mathbf{A}$ and $\mathbf{B}$ satisfy

$$\begin{aligned}\lambda_k(\mathbf{A}) &= \max_{\mathcal{M}\subset\mathbb{C}^d,\dim\mathcal{M}=k} \min_{\mathbf{x}\in\mathcal{M},\|\mathbf{x}\|_2=1} \langle \mathbf{A}\mathbf{x},\mathbf{x}\rangle \\ &\leq \max_{\mathcal{M}\subset\mathbb{C}^d,\dim\mathcal{M}=k} \min_{\mathbf{x}\in\mathcal{M},\|\mathbf{x}\|_2=1} \langle \mathbf{B}\mathbf{x},\mathbf{x}\rangle = \lambda_k(\mathbf{B}) \;,\end{aligned}$$

because $\mathbf{A} \preccurlyeq \mathbf{B}$ by assumption. Since f is non-decreasing it follows that

$$\operatorname{tr} f(\mathbf{A}) = \sum_{k=1}^{n} f(\lambda_k(\mathbf{A})) \leq \sum_{k=1}^{n} f(\lambda_k(\mathbf{B})) = \operatorname{tr} f(\mathbf{B}) \;.$$

This completes the proof. □

Next we show that certain inequalities for scalar functions extend to traces of matrix-valued functions [330].

Theorem A.31. *Let $f_\ell, g_\ell : [a,b] \to \mathbb{R}$, $\ell = 1,\ldots,M$ be functions such that for some $c_\ell \in \mathbb{R}$ and all $x, y \in [a,b]$*

$$\sum_{\ell=1}^{M} c_\ell f_\ell(x) g_\ell(y) \geq 0$$

Then for all self-adjoint matrices $\mathbf{A},\mathbf{B}$ with eigenvalues in $[a,b]$

$$\operatorname{tr}\left(\sum_{\ell=1}^{M} c_\ell f_\ell(\mathbf{A}) g_\ell(\mathbf{B})\right) \geq 0 \;.$$

Proof. Let $\mathbf{A} = \sum_{k=1}^{n} \lambda_k u_k u_k^*$, $\mathbf{B} = \sum_{k=1}^{n} \eta_k v_k v_k^*$ be the eigenvalue decompositions of $\mathbf{A},\mathbf{B}$; in particular, $\mathbf{u}_k, \mathbf{v}_k \in \mathbb{C}^n$ are eigenvectors of $\mathbf{A},\mathbf{B}$. Then

$$\begin{aligned}&\operatorname{tr}\left(\sum_{\ell=1}^{M} c_\ell f_\ell(\mathbf{A}) g_\ell(\mathbf{B})\right) = \operatorname{tr}\left(\sum_{\ell=1}^{M} c_\ell \sum_{j,k=1}^{d} f_\ell(\lambda_j) g_\ell(\eta_k) \mathbf{u}_j \mathbf{u}_j^* \mathbf{v}_k \mathbf{v}_k^*\right) \\ &= \sum_{j,k=1}^{n} \sum_{\ell=1}^{M} c_\ell f_\ell(\lambda_j) g_\ell(\eta_k) \operatorname{tr}\left(\mathbf{u}_j \mathbf{u}_j^* \mathbf{v}_k \mathbf{v}_k^*\right) = \sum_{j,k=1}^{n} \sum_{\ell=1}^{M} c_\ell f_\ell(\lambda_j) g_\ell(\eta_k) |\langle \mathbf{u}_j, \mathbf{v}_k\rangle|^2 \\ &\geq 0\end{aligned}$$

by assumption. Hereby, we have also used the cyclicity of the trace in the last step. □

A function f is called *matrix monotone* (or operator monotone) if $\mathbf{A} \preccurlyeq \mathbf{B}$ implies

$$f(\mathbf{A}) \preccurlyeq f(\mathbf{B}) \;. \tag{A.49}$$

It may come as a surprise that the extension of a monotonically increasing function $f : \mathbb{R} \to \mathbb{R}$ to a matrix function via (A.42) may fail to be matrix monotone. A simple example is the function $f(t) = t^2$.

In order to study matrix monotonicity for some specific functions below, we first state an easy observation.

Lemma A.32. *If $\mathbf{A} \preccurlyeq \mathbf{B}$, then for all matrices $\mathbf{Y}$ of matching dimensions we have $\mathbf{Y}^*\mathbf{A}\mathbf{Y} \preccurlyeq \mathbf{Y}^*\mathbf{B}\mathbf{Y}$. If in addition $\mathbf{Y}$ is invertible and $\mathbf{A} \prec \mathbf{B}$ then $\mathbf{Y}^*\mathbf{A}\mathbf{Y} \prec \mathbf{Y}^*\mathbf{B}\mathbf{Y}$.*

Proof. For every vector $\mathbf{x}$ it holds

$$\langle \mathbf{Y}^*\mathbf{A}\mathbf{Y}\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}\mathbf{Y}\mathbf{x}, \mathbf{Y}\mathbf{x}\rangle \leq \langle \mathbf{B}\mathbf{Y}\mathbf{x}, \mathbf{Y}\mathbf{x}\rangle = \langle \mathbf{Y}^*\mathbf{B}\mathbf{Y}\mathbf{x}, \mathbf{x}\rangle,$$

which shows the first part. The second part requires only minor changes in the proof. □

Next, we show the matrix monotonicity of the negative inverse map.

Proposition A.33. *The matrix function $f(\mathbf{A}) = -\mathbf{A}^{-1}$ is matrix monotone on the set of positive definite matrices.*

Proof. Let $0 \prec \mathbf{A} \preccurlyeq \mathbf{B}$. Then the matrix $\mathbf{B}^{-1/2}$ exists (and may be defined via (A.42)). It follows from Lemma A.32 that

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \preccurlyeq \mathbf{B}^{-1/2}\mathbf{B}\mathbf{B}^{-1/2} = \mathbf{Id}\,.$$

The matrix $\mathbf{C} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ has an eigenvalue decomposition $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with unitary $\mathbf{U}$ and diagonal $\mathbf{D}$, and the above relation implies by Lemma A.32 that $0 \prec \mathbf{D} \preccurlyeq \mathbf{Id}$. Therefore, $\mathbf{Id} \preccurlyeq \mathbf{D}^{-1}$ and again by Lemma A.32

$$\mathbf{Id} \preccurlyeq \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^* = \mathbf{C}^{-1} = (\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})^{-1} = \mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}\,.$$

Applying Lemma A.32 another time shows that $\mathbf{B}^{-1} = \mathbf{B}^{-1/2}\mathbf{Id}\mathbf{B}^{-1/2} \preccurlyeq \mathbf{A}^{-1}$. □

The *matrix logarithm* can be defined for positive definite matrices via the spectral mapping formula (A.42) with $f(x) = \ln(x)$. It is the inverse of the matrix exponential,

$$\exp(\ln(\mathbf{A})) = \mathbf{A}\,. \tag{A.50}$$

Remark A.34. The definition of the matrix logarithm can be extended to invertible, not necessarily self-adjoint, matrices; similarly as the matrix exponential extends to all square matrices via the power series expansion (A.45). Like the extension of the logarithm to the complex numbers, one encounters the fact that the logarithm is no longer uniquely defined via (A.50). One usually chooses the principal branch, which requires to restrict to matrices with eigenvalues not contained on the negative real line.

In contrast to the matrix exponential the matrix logarithm is matrix monotone.

Proposition A.35. *Let* $\mathbf{A}, \mathbf{B}$ *be self-adjoint, positive definite matrices. Then*

$$\ln(\mathbf{A}) \preccurlyeq \ln(\mathbf{B}) \quad \textit{whenever } \mathbf{A} \preccurlyeq \mathbf{B} \, .$$

Proof. We first claim that the (scalar) logarithm satisfies

$$\ln(x) = \int_0^\infty \left(\frac{1}{t+1} - \frac{1}{t+x} \right) dt \, , \quad x > 0 \, . \tag{A.51}$$

Indeed, a simple integral transformation shows that, for $R > 0$,

$$\int_0^R \frac{1}{t+x} dt = \ln(x+R) - \ln(x) \, .$$

and $\int_0^R \frac{1}{t+1} dt = \ln(R+1)$. We obtain

$$\begin{aligned} &\int_0^\infty \left(\frac{1}{t+1} - \frac{1}{t+x} \right) dt = \lim_{R \to \infty} \ln(x) - \ln(x+R) + \ln(R+1) \\ &= \ln(x) + \lim_{R \to \infty} \ln\left(\frac{R+1}{x+R} \right) = \ln(x) \, . \end{aligned}$$

It follows from Proposition A.33 that, for $t \geq 0$ the matrix function

$$g_t(\mathbf{A}) := \frac{1}{t+1} \mathbf{Id} - (t\mathbf{Id} + \mathbf{A})^{-1}$$

is matrix monotone on the set of self-adjoint positive definite matrices. By (A.51) and by the definition of the matrix logarithm via (A.42) we conclude that, for a self-adjoint positive definite matrix $\mathbf{A}$,

$$\ln(\mathbf{A}) = \int_0^\infty g_t(\mathbf{A}) dt \, .$$

Therefore, the matrix logarithm is matrix monotone, since integrals preserve the semidefinite ordering. □

Also the square-root function $\mathbf{A} \mapsto \mathbf{A}^{1/2}$ is matrix monotone on the set of positive semidefinite matrices [38].

We continue the discussion of matrix function in Section B.6, where we treat convexity issues.

B

Convex Analysis

This appendix provides a short overview on convex analysis and convex optimization. Much more information can be found in various books on the subject such as [59, 155, 237, 254, 366, 367].

For the purpose of this exposition on convexity, we work on real vector spaces $\mathbb{R}^N$, and treat sets in and functions on $\mathbb{C}^N$ by identifying $\mathbb{C}^N$ with $\mathbb{R}^{2N}$. In order to reverse this identification in some of the statements and definitions below, one needs to replace the inner product $\langle \mathbf{x}, \mathbf{z} \rangle$ by $\mathrm{Re}\langle \mathbf{x}, \mathbf{z} \rangle$ for complex $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$.

B.1 Convex Sets

Let us start with the basic definition.

Definition B.1. *A subset $K \subset \mathbb{R}^N$ is called convex if for all $\mathbf{x}, \mathbf{z} \in K$ the line segment connecting $\mathbf{x}, \mathbf{z}$ is also completely contained in K, that is,*

$$t\mathbf{x} + (1-t)\mathbf{z} \in K \quad \text{for all } t \in [0,1] \, .$$

It is straightforward to check that a set $K \in \mathbb{R}^N$ is convex if and only if for all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in K$ and $t_1, \ldots, t_n \geq 0$ such that $\sum_{j=1}^n t_j = 1$ the convex combination $\sum_{j=1}^n t_j \mathbf{x}_j$ is also contained in K.

Definition B.2. *Let $T \subset \mathbb{R}^N$ be a set. Its convex hull $\mathrm{conv}(T)$ is the smallest convex set containing T.*

It is well-known [366, Theorem 2.3] that the convex hull of T consists of the convex combinations of T,

$$\mathrm{conv}(T) = \left\{ \sum_j t_j \mathbf{x}_j : t_j \geq 0, \sum_j t_j = 1, \mathbf{x}_j \in T \right\} \, .$$

Simple examples of convex sets include subspaces, affine spaces, half spaces, polygons or norm balls $B(\mathbf{x}, t)$, see (A.1). The intersection of convex sets is again convex.

Definition B.3. *A $K \subset \mathbb{R}^N$ is called a cone if for all $\mathbf{x} \in K$ and all $t \geq 0$ also $t\mathbf{x}$ is contained in K. If, in addition, K is convex, then K is called a convex cone.*

Obviously, the zero vector is contained in every cone. A set K is a convex cone if for all $\mathbf{x}, \mathbf{z} \in K$ and $t, s \geq 0$ also $s\mathbf{x} + t\mathbf{z}$ is contained in K.

Simple examples of convex cones include the positive orthant $\mathbb{R}^N_+ = \{\mathbf{x} \in \mathbb{R}^N : x_i \geq 0 \text{ for all } i \in [N]\}$ or the set of positive semidefinite matrices in $\mathbb{R}^{N \times N}$. Another important example of a convex cone is the second order cone

$$\left\{ \mathbf{x} \in \mathbb{R}^{N+1} : \sqrt{\sum_{j=1}^{N} x_j^2} \leq x_{N+1} \right\} . \tag{B.1}$$

For a cone $K \subset \mathbb{R}^N$, its dual cone K^* is defined via

$$K^* := \left\{ \mathbf{z} \in \mathbb{R}^N : \langle \mathbf{x}, \mathbf{z} \rangle \geq 0 \text{ for all } \mathbf{x} \in K \right\} . \tag{B.2}$$

As the intersection of half spaces, K^* is closed and convex, and it is straightforward to check that K^* is again a cone. If K is a closed cone then $K^{**} = K$. Moreover, if $H, K \subset \mathbb{R}^N$ are cones such that $H \subset K$ then $K^* \subset H^*$. As an example, the dual cone of the positive orthant $\mathbb{R}^N_+$ is $\mathbb{R}^N_+$ itself; in other words, $\mathbb{R}^N_+$ is self-dual. Note that the dual cone is closely related to the polar cone, which is defined by

$$K^\circ := \left\{ \mathbf{z} \in \mathbb{R}^N : \langle \mathbf{x}, \mathbf{z} \rangle \leq 0 \text{ for all } \mathbf{x} \in K \right\} = -K^* . \tag{B.3}$$

The conic hull $\operatorname{cone}(T)$ of a set $T \subset \mathbb{R}^N$ is the smallest convex cone containing T. It can be described as

$$\operatorname{cone}(T) = \left\{ \sum t_j \mathbf{x}_j : t_j \geq 0, \mathbf{x}_j \in T \right\} . \tag{B.4}$$

Convex sets can be separated by hyperplanes as stated next.

Theorem B.4. *Let $K_1, K_2 \subset \mathbb{R}^N$ be convex sets such their interiors have empty intersection. Then there exists a vector $\mathbf{w} \in \mathbb{R}^N$ and a scalar λ such that*

$$K_1 \subset \{\mathbf{x} \in \mathbb{R}^N : \langle \mathbf{x}, \mathbf{w} \rangle \leq \lambda\} ,$$
$$K_2 \subset \{\mathbf{x} \in \mathbb{R}^N : \langle \mathbf{x}, \mathbf{w} \rangle \geq \lambda\} .$$

Remark B.5. The theorem applies in particular when $K_1 \cap K_2 = \emptyset$ or when K_1, K_2 intersect in one point, $K_1 \cap K_2 = \{\mathbf{x}_0\}$. In the latter case, one chooses $\lambda = \langle \mathbf{x}_0, \mathbf{w} \rangle$. If K_2 is a subset of a hyperplane then we can choose $\mathbf{w}$ and λ such that $K_2 \subset \{\mathbf{x} \in \mathbb{R}^N : \langle \mathbf{x}, \mathbf{w} \rangle = \lambda\}$.

Next we consider the notion of extreme points.

Definition B.6. *Let $K \subset \mathbb{R}^N$ be a convex set. A point $\mathbf{x} \in K$ is called an extreme point of K if $\mathbf{x} = t\mathbf{y} + (1-t)\mathbf{z}$ for $\mathbf{y}, \mathbf{z} \in K$ and $t \in (0,1)$ implies $\mathbf{x} = \mathbf{y} = \mathbf{z}$.*

Compact convex sets can be described via their extremal points as stated next (see for instance [366, Corollary 18.5.1] or [237, Theorem 2.3.4]).

Theorem B.7. *A compact convex set is the convex hull of its extreme points.*

If K is a polygon then its extreme points are the zero-dimensional faces of K, and the above statement is clearly intuitive.

B.2 Convex Functions

We work with extended valued functions $F : \mathbb{R}^N \to (-\infty, \infty] = \mathbb{R} \cup \{+\infty\}$. Sometimes we also consider an additional extension of the values to $-\infty$. Addition, multiplication and inequalities in $(-\infty, \infty]$ are understood in the "natural" sense; for instance, $x + \infty = \infty$ for all $x \in \mathbb{R}$, $\lambda \cdot \infty = \infty$ for $\lambda > 0$, $x < \infty$ for all $x \in \mathbb{R}$, $\infty \leq \infty$. The domain of an extended-valued function F is defined as

$$\mathrm{dom}(F) = \{\mathbf{x} \in \mathbb{R}^N, F(\mathbf{x}) \neq \infty\} .$$

A function with $\mathrm{dom}(F) \neq \emptyset$ is called proper. A function $F : K \to \mathbb{R}$ on a subset $K \subset \mathbb{R}^N$ can be extended to an extended valued function by setting $F(\mathbf{x}) = \infty$ for $\mathbf{x} \notin K$. Then clearly $\mathrm{dom}(F) = K$, and we call this extension the canonical one.

Definition B.8. *An extended valued function $F : \mathbb{R}^N \to (-\infty, \infty]$ is called convex if for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$ and $t \in [0,1]$,*

$$F(t\mathbf{x} + (1-t)\mathbf{z}) \leq tF(\mathbf{x}) + (1-t)F(\mathbf{z}) . \tag{B.5}$$

F is called strictly convex if

$$F(t\mathbf{x} + (1-t)\mathbf{z}) < tF(\mathbf{x}) + (1-t)F(\mathbf{z})$$

for all $\mathbf{x} \neq \mathbf{z}$ and all $t \in (0,1)$.
F is called strongly convex with parameter $\gamma > 0$ if for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$ and $t \in [0,1]$

$$F(t\mathbf{x} + (1-t)\mathbf{z}) \leq tF(\mathbf{x}) + (1-t)F(\mathbf{z}) - \frac{\gamma}{2} t(1-t)\|\mathbf{x} - \mathbf{z}\|_2^2 . \tag{B.6}$$

A function $F : \mathbb{R}^N \to [-\infty, \infty)$ is called (strictly, strongly) concave if $-F$ is (strictly, strongly) convex.

Obviously, a strongly convex function is strictly convex.

The domain of a convex function is convex, and a function $F : K \to \mathbb{R}^N$ on a convex subset $K \subset \mathbb{R}^N$ is called convex if its canonical extension to $\mathbb{R}^N$ is convex (or alternatively, $\mathbf{x}, \mathbf{z}$ in the definition (B.5) are assumed to be in K). A function F is convex if and only if its epigraph

$$\operatorname{epi}(F) = \{(\mathbf{x}, r) : r \geq F(\mathbf{x})\} \subset \mathbb{R}^N \times \mathbb{R}$$

is a convex set.

As for convex sets we may also consider general convex combinations: A function $F : \mathbb{R}^N \to [-\infty, \infty]$ is convex if and only if for all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^N$ and $t_1, \ldots, t_n \geq 0$ such that $\sum_{j=1}^n t_j = 1$,

$$F\big(\sum_{j=1}^n t_j \mathbf{x}_j\big) \leq \sum_{j=1}^n t_j F(\mathbf{x}_j) \, .$$

For differentiable functions we have the following characterizations of convexity.

Proposition B.9. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be differentiable.*

(a) F is convex if and only if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$

$$F(\mathbf{x}) \geq F(\mathbf{y}) + \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \, ,$$

where the gradient $\nabla F(\mathbf{y}) = (\frac{\partial F}{\partial y_1}(\mathbf{y}), \ldots, \frac{\partial F}{\partial y_N}(\mathbf{y}))^\top$ as usual.

(b) F is strongly convex with parameter $\gamma > 0$ if and only if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$

$$F(\mathbf{x}) \geq F(\mathbf{y}) + \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \, .$$

(c) Assume that F is twice differentiable. Then F is convex if and only if

$$\nabla^2 F(\mathbf{x}) \succcurlyeq \mathbf{0}$$

for all $\mathbf{x} \in \mathbb{R}^N$, where $\nabla^2 F$ is the Hessian of F.

Let us summarize some results on the composition of convex functions.

Proposition B.10. *(a) Let F, G be convex functions on $\mathbb{R}^N$. Then, for $\alpha, \beta \geq 0$ the function $\alpha F + \beta G$ is convex.*

(b) Let $F : \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing, and $G : \mathbb{R}^N \to \mathbb{R}$ be convex. Then the function $H(\mathbf{x}) = F(G(\mathbf{x}))$ is convex.

Proof. (a) is straightfoward to verify. For (b) take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and $t \in [0, 1]$. Then

$$\begin{aligned} H(t\mathbf{x} + (1-t)\mathbf{y}) &= F(G(t\mathbf{x} + (1-t)\mathbf{y})) \leq F(tG(\mathbf{x}) + (1-t)G(\mathbf{y})) \\ &\leq tF(G(\mathbf{x})) + (1-t)F(G(\mathbf{y})) = tH(\mathbf{x}) + (1-t)H(\mathbf{y}), \end{aligned}$$

where we have applied convexity of G and monotonicity of F in the first step and convexity of F in the second step. □

Let us give some examples of convex functions.

Example B.11. (a) For $p \geq 1$, the function $F(x) = |x|^p$, $x \in \mathbb{R}$, is convex.
(b) Every norm $\|\cdot\|$ on $\mathbb{R}^N$ is a convex function. This follows from the triangle inequality and homogeneity.
(c) The ℓ_p-norms $\|\cdot\|_p$ are strictly convex if $1 < p < \infty$, and they are *not* strictly convex if $p = 1$ or $p = \infty$.
(d) For a non-decreasing convex function $F : \mathbb{R} \to (-\infty, \infty]$ and a norm $\|\cdot\|$ on $\mathbb{R}^N$, the function $H(\mathbf{x}) = F(\|\mathbf{x}\|)$ is convex. This follows from (a) and Proposition B.10(b). In particular, the function $\mathbf{x} \mapsto \|\mathbf{x}\|^p$ is convex provided that $p \geq 1$.
(e) For a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, the function $F(\mathbf{x}) = \mathbf{x}^*\mathbf{A}\mathbf{x}$ is convex. If $\mathbf{A}$ is positive definite then F is strongly convex.
(f) For a convex set K the characteristic function

$$\chi_K(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in K \ , \\ \infty & \text{if } \mathbf{x} \notin K \end{cases} \tag{B.7}$$

is convex.

We continue with the discussion of continuity properties.

Proposition B.12. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex function. Then F is continuous on $\mathbb{R}^N$.*

The treatment of extended valued functions requires the notion of lower semicontinuity.

Definition B.13. *A function $F : \mathbb{R}^N \to (-\infty, \infty]$ is called lower semicontinuous if for all $\mathbf{x} \in \mathbb{R}^N$ and every sequence $(\mathbf{x}_j)_{j\in\mathbb{N}} \subset \mathbb{R}$ converging to $\mathbf{x}$ it holds*

$$\liminf_{j\to\infty} F(\mathbf{x}_j) \geq F(\mathbf{x}) \ .$$

Clearly, a continuous function $F : \mathbb{R}^N \to \mathbb{R}$ is lower semicontinuous. A nontrivial example is the characteristic function χ_K of a proper subset $K \subset \mathbb{R}^N$ defined in (B.7). Clearly, χ_K is not continuous, but it is lower semicontinuous if and only if K is closed.

A function is lower semicontinuous if and only if its epigraph is closed.

(We remark that the notion of lower semicontinuity is particularly useful in infinite-dimensional Hilbert spaces, where for instance the norm $\|\cdot\|$ is not continuous with respect to the weak topology but is still lower semicontinuous with respect to the weak topology, see for instance [155]).

Convex functions have nice properties related to minimization. A (global) minimum (or minimizer) of a function $F : \mathbb{R}^N \to (-\infty, \infty]$ is a point $\mathbf{x} \in \mathbb{R}^N$ satisfying $F(\mathbf{x}) \leq F(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^N$. A local minimum of F is a point $\mathbf{x} \in \mathbb{R}^N$ such that there exists $\varepsilon > 0$ and $F(\mathbf{x}) \leq F(\mathbf{y})$ for all $\mathbf{y}$ satisfying $\|\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon$. (The Euclidean norm $\|\cdot\|_2$ can be replaced by any other norm $\|\cdot\|$ in this definition.)

Proposition B.14. *Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be convex.*

(a) A local minimum of F is a global minimum.
(b) The set of minima of F is convex.
(c) If F is strictly convex then the minimum is unique.

Proof. (a) Let $\mathbf{x}$ be a local minimum and $\mathbf{z} \in \mathbb{R}^N$ be arbitrary. Let $\varepsilon > 0$ be the parameter appearing in the definition of a local minimum. Then there exists $\mathbf{y} \in \mathbb{R}^N$ such $\|\mathbf{x}-\mathbf{y}\|_2 \leq \varepsilon$ and such that $\mathbf{y} = t\mathbf{x}+(1-t)\mathbf{z}$ for a suitable $t \in (0,1)$. Then $F(\mathbf{x}) \leq F(\mathbf{y})$ and by convexity $F(\mathbf{y}) = F(t\mathbf{x} + (1-t)\mathbf{z}) \leq tF(\mathbf{x}) + (1-t)F(\mathbf{z})$. Therefore,

$$(1-t)F(\mathbf{z}) \geq F(\mathbf{y}) - tF(\mathbf{x}) \geq F(\mathbf{x}) - tF(\mathbf{x}),$$

which by $t < 1$ implies $F(\mathbf{z}) \geq F(\mathbf{x})$. Since $\mathbf{z}$ was arbitrary, it follow that $\mathbf{x}$ is a global minimum.

(b) Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ be two minima, i.e., $F(\mathbf{x}) = F(\mathbf{y}) = \inf_{\mathbf{z}} F(\mathbf{z})$. Then, for $t \in [0,1]$,

$$F(t\mathbf{x} + (1-t)\mathbf{y}) \leq tF(\mathbf{x}) + (1-t)F(\mathbf{y}) = \inf_{\mathbf{z}} F(\mathbf{z}),$$

so that $t\mathbf{x} + (1-t)\mathbf{y}$ is a minimum as well.

(c) Suppose that $\mathbf{x} \neq \mathbf{y}$ are both minima of F. Then, for $t \in (0,1)$,

$$F(t\mathbf{x} + (1-t)\mathbf{y}) < tF(\mathbf{x}) + (1-t)F(\mathbf{y}) = \inf_{\mathbf{z}} F(\mathbf{z}),$$

which is a contradiction. □

The fact that local minima of convex functions are automatically global minima is the essential reason why efficient optimization methods are available for convex optimization problems.

We say that a function $f(\mathbf{x}, \mathbf{y})$ of two arguments $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ is jointly convex if it is convex as a function of the variable $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Partial minimization of a jointly convex function in one variable yields again a convex function as stated next.

Theorem B.15. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to (-\infty, \infty]$ be a jointly convex function. Then the function $g(\mathbf{x}) := \inf_{\mathbf{y} \in \mathbb{R}^m} f(\mathbf{x}, \mathbf{y}), \mathbf{x} \in \mathbb{R}^n$, is convex.*

Proof. For simplicity we assume that the infimum is always attained. The general case has to be treated with an ε-argument.

Given $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, there exist $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^m$ such that

$$f(\mathbf{x}_1, \mathbf{y}_1) = \min_{\mathbf{y} \in \mathbb{R}^m} f(\mathbf{x}_1, \mathbf{y}) \quad f(\mathbf{x}_2, \mathbf{y}_2) = \min_{\mathbf{y} \in \mathbb{R}^m} f(\mathbf{x}_2, \mathbf{y}) \ .$$

For $t \in [0,1]$ the joint convexity implies that

$$\begin{aligned} g(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) &\leq f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2, t\mathbf{y}_1 + (1-t)\mathbf{y}_2) \\ &\leq tf(\mathbf{x}_1, \mathbf{y}_1) + (1-t)f(\mathbf{x}_2, \mathbf{y}_2) = tg(\mathbf{x}_1) + (1-t)g(\mathbf{x}_2) \ . \end{aligned}$$

This point finishes the argument. □

Clearly, the previous theorem shows as well that partial maximization of a jointly concave function yields a concave function.

Next we consider the maximum of a convex function over a convex set.

Theorem B.16. *Let $K \subset \mathbb{R}^N$ be a compact convex set, and $F : K \to \mathbb{R}$ be a convex function. Then F attains its maximum at an extreme point of K.*

Proof. Let $\mathbf{x} \in K$ such that $F(\mathbf{x}) \geq F(\mathbf{z})$ for all $\mathbf{z} \in K$. Since by Theorem B.7 K is the convex hull of its extreme points, we can write $\mathbf{x} = \sum_{j=1}^m t_j \mathbf{x}_j$ for some m, $t_j > 0$, $\sum_{j=1}^m t_j = 1$ and $\mathbf{x}_j$, $j = 1, \ldots, m$ being extreme points of K. By convexity

$$F(\mathbf{x}) = F\big(\sum_{j=1}^m t_j \mathbf{x}_j\big) \leq \sum_{j=1}^m t_j F(\mathbf{x}_j) \leq \sum_{j=1}^m t_j F(\mathbf{x}) = F(\mathbf{x})$$

because by definition $F(\mathbf{x}_j) \leq F(\mathbf{x})$. Therefore, all inequalities actually hold with equality, which is only possible if $F(\mathbf{x}_j) = F(\mathbf{x})$ for all $j = 1, \ldots, m$. Therefore, the maximum of F is attained at an extreme point of K. □

B.3 The Convex Conjugate

The convex conjugate is a very useful concept in convex analysis and optimization.

Definition B.17. *Let $F : \mathbb{R}^N \to (-\infty, \infty]$. Then its convex conjugate (or Fenchel dual) function $F^* : \mathbb{R}^N \to (-\infty, \infty]$ is defined by*

$$F^*(\mathbf{y}) := \sup_{\mathbf{x} \in \mathbb{R}^N} \{\langle \mathbf{x}, \mathbf{y} \rangle - F(\mathbf{x})\} .$$

The convex conjugate F^* is always a convex function, no matter whether the function F is convex or not. The definition of F^* immediately gives the Fenchel (or Young, or Fenchel-Young) inequality

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq F(\mathbf{x}) + F^*(\mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N . \tag{B.8}$$

Let us summarize some properties of convex conjugate functions.

Proposition B.18. *Let $F : \mathbb{R}^N \to (-\infty, \infty]$.*

(a) The convex conjugate F^* is lower semicontinuous.
(b) The biconjugate F^{**} is the largest lower semicontinuous convex function satisfying $F^{**}(\mathbf{x}) \leq F(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^N$. In particular, if F is convex and lower semicontinuous then $F = F^{**}$.
(c) For $\tau \neq 0$ let $F_\tau(\mathbf{x}) := F(\tau \mathbf{x})$. Then $(F_\tau)^*(\mathbf{y}) = F^*(\mathbf{y}/\tau)$.
(d) For $\tau > 0$, $(\tau F)^*(\mathbf{y}) = \tau F^*(\mathbf{y}/\tau)$.

(e) For $\mathbf{z} \in \mathbb{R}^N$ *let* $F^{(\mathbf{z})} := F(\mathbf{x} - \mathbf{z})$. *Then* $(F^{(\mathbf{z})})^*(\mathbf{y}) = \langle \mathbf{z}, \mathbf{y} \rangle + F^*(\mathbf{y})$.

Proof. For (a) and (b) we refer to [366, Corollary12.1.1 and Theorem 12.2]. For (d) a substitution gives

$$(\tau F)^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbb{R}^N} \{\langle \mathbf{x}, \mathbf{y} \rangle - \tau F(\mathbf{x})\} = \tau \sup_{\mathbf{x} \in \mathbb{R}^N} \{\langle \mathbf{x}, \tau^{-1}\mathbf{y} \rangle - F(\mathbf{x})\} = \tau F^*(\mathbf{y}/\tau) \, .$$

The statements (c) and (e) follow from simple calculations. □

The biconjugate F^{**} is sometimes also called the convex relaxation of F because of (b).

Let us compute the convex conjugate for some examples.

Example B.19. (a) Let $F(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, $\mathbf{x} \in \mathbb{R}^N$. Then $F^*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2 = F(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^N$. Indeed, since

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2 \tag{B.9}$$

we have

$$F^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbb{R}^N} \{\langle \mathbf{x}, \mathbf{y} \rangle - F(\mathbf{x})\} \leq \frac{1}{2}\|\mathbf{y}\|_2^2 \, .$$

For the converse inequality, we just set $\mathbf{x} = \mathbf{y}$ in the definition of the convex conjugate to obtain

$$F^*(\mathbf{y}) \geq \|\mathbf{y}\|_2^2 - \frac{1}{2}\|\mathbf{y}\|_2^2 = \frac{1}{2}\|\mathbf{y}\|_2^2 \, .$$

Note that this example is the only function F on $\mathbb{R}^N$ satisfying $F = F^*$.

(b) Let $F(x) = \exp(x)$, $x \in \mathbb{R}$. The function $x \mapsto xy - \exp(x)$ takes its maximum at $x = \ln y$ if $y > 0$ so that

$$F^*(y) = \sup_{x \in \mathbb{R}} \{xy - e^x\} = \begin{cases} y \ln y - y & \text{if } y > 0 \, , \\ 0 & \text{if } y = 0 \, , \\ \infty & \text{if } y < 0 \, . \end{cases}$$

The Young inequality for this particular pair reads

$$xy \leq e^x + y \ln(y) - y \quad \text{for all } x > 0, y \in \mathbb{R} \, . \tag{B.10}$$

(c) Let $F(\mathbf{x}) = \|\mathbf{x}\|$ for some norm on $\mathbb{R}^N$. Let $\|\cdot\|_*$ be its dual norm, see Definition A.4. Then the convex conjugate is the characteristic function of the dual norm ball, that is,

$$F^*(\mathbf{y}) = \chi_{B_{\|\cdot\|_*}}(\mathbf{y}) = \begin{cases} 0 & \text{if } \|\mathbf{y}\|_* \leq 1 \, , \\ \infty & \text{otherwise} \quad . \end{cases}$$

Indeed, by the definition of the dual norm $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{y}\|_* \|\mathbf{x}\|$, so that in this case

$$F^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathbb{R}^N} \{\langle \mathbf{x}, \mathbf{y}\rangle - \|\mathbf{x}\|\} \le \sup_{\mathbf{x}\in\mathbb{R}^N} \{(\|\mathbf{y}\|_* - 1)\|\mathbf{x}\|\}$$

so that $F^*(\mathbf{y}) \le 0$ if $\|\mathbf{y}\|_* \le 1$. The choice $\mathbf{x} = \mathbf{0}$ shows that $F^*(\mathbf{y}) = 0$ in this case. If $\|\mathbf{y}\|_* > 1$ then there exists $\mathbf{x}$ such that $\langle \mathbf{x}, \mathbf{y}\rangle > \|\mathbf{x}\|$. Replacing $\mathbf{x}$ by $\lambda\mathbf{x}$ for $\lambda > 0$ and letting $\lambda \to \infty$ shows that $F^*(\mathbf{y}) = \infty$ in this case.

(d) Let $F = \chi_K$ be the characteristic function of a convex set K, see (B.7). Its convex conjugate is given by

$$F^*(\mathbf{y}) = \sup_{\mathbf{x}\in K} \langle \mathbf{x}, \mathbf{y}\rangle \ .$$

B.4 The Subdifferential

The subdifferential generalizes the gradient for not necessarily differentiable functions.

Definition B.20. *The subdifferential of a convex function $F : \mathbb{R}^N \to \mathbb{R}$ at a point $\mathbf{x} \in \mathbb{R}^N$ is defined by*

$$\partial F(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^N : F(\mathbf{z}) - F(\mathbf{x}) \ge \langle \mathbf{v}, \mathbf{z} - \mathbf{x}\rangle \ \text{ for all } \mathbf{z} \in \mathbb{R}^N\} \ . \qquad \text{(B.11)}$$

The elements of $\partial F(\mathbf{x})$ are called subgradients of F at $\mathbf{x}$.

The subdifferential $\partial F(\mathbf{x})$ of a convex function F is always non-empty. If F is differentiable in $\mathbf{x}$ then $\partial F(\mathbf{x})$ contains only the gradient,

$$\partial F(\mathbf{x}) = \{\nabla F(\mathbf{x})\} \ ,$$

see Proposition B.9(a) for one direction. A simple example of a function with a non-trivial subdifferential is the absolute value $F(x) = |x|$, for which

$$\partial F(x) = \begin{cases} \{\mathrm{sgn}(x)\} & \text{if } x \ne 0 \ , \\ [-1, 1] & \text{if } x = 0 \ , \end{cases}$$

where $\mathrm{sgn}(x) = +1$ for $x > 0$ and $\mathrm{sgn}(x) = -1$ for $x < 0$ as usual.

The subdifferential allows a simple characterization of minimizers of convex functions.

Theorem B.21. *A vector $\mathbf{x}$ is a minimum of F if and only if $\mathbf{0} \in \partial F(\mathbf{x})$.*

Proof. This is obvious from the definition of the subdifferential. □

Convex conjugate functions and subdifferentials are related in the following way.

Theorem B.22. *Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be a convex function and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. The following conditions are equivalent*

(a) $\mathbf{y} \in \partial F(\mathbf{x})$,
(b) $F(\mathbf{x}) + F^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$.

If additionally, F is lower semicontinuous then (a) and (b) are equivalent to

(c) $\mathbf{x} \in \partial F^*(\mathbf{y})$.

Proof. Condition (a) reads by definition of the subgradient

$$\langle \mathbf{x}, \mathbf{y} \rangle - F(\mathbf{x}) \geq \langle \mathbf{z}, \mathbf{y} \rangle - F(\mathbf{z}) \quad \text{for all } \mathbf{z} \in \mathbb{R}^N \,. \tag{B.12}$$

Therefore, the function $\mathbf{z} \mapsto \langle \mathbf{z}, \mathbf{y} \rangle - F(\mathbf{z})$ attains its maximum in $\mathbf{x}$. By definition of the convex conjugate this implies that $F^*(\mathbf{y}) \leq \langle \mathbf{x}, \mathbf{y} \rangle - F(\mathbf{x})$ or $F^*(\mathbf{y}) + F(\mathbf{x}) \leq \langle \mathbf{x}, \mathbf{y} \rangle$. But the reversed inequality holds always due to Fenchel's inequality (B.8). This shows that (a) implies (b). Conversely, condition (b) implies by Fenchel's inequality and the definition of the convex conjugate that the function $\mathbf{z} \mapsto \langle \mathbf{z}, \mathbf{y} \rangle - F(\mathbf{z})$ attains its maximum in $\mathbf{x}$, which is nothing else than (B.12). It follows from the definition of the subdifferential that $\mathbf{y} \in \partial F(\mathbf{x})$.

Now if F is lower semicontinuous then $F^{**} = F$ by Proposition B.18(b) so that (b) is equivalent to $F^{**}(\mathbf{x}) + F^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$. Using the equivalence of (a) and (b) with F replaced by F^* concludes the proof. □

As a consequence, if F is a convex lower semicontinuous function then ∂F is the inverse of ∂F^* in the sense that $\mathbf{x} \in \partial F^*(\mathbf{y})$ if and only if $\mathbf{y} \in \partial F(\mathbf{x})$.

Next we consider the so-called proximal mapping (also called proximation or resolvent operator). Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be a convex function. Then, for $\mathbf{z} \in \mathbb{R}^N$ the function

$$\mathbf{x} \mapsto F(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2$$

is strictly convex due to the strict convexity of $\mathbf{x} \mapsto \|\mathbf{x}\|_2^2$. By Proposition B.14(b) its minimizer is unique. The mapping

$$P_F(\mathbf{z}) := \arg\min \left\{ F(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 \;:\; \mathbf{x} \in \mathbb{R}^N \right\} , \tag{B.13}$$

is called the proximal mapping associated with F. In the special case that $F = \chi_K$ is the characteristic function of a convex set K defined in (B.7), then $P_K := P_{\chi_K}$ is the orthogonal projection onto K, that is,

$$P_K(\mathbf{z}) = \arg\min_{\mathbf{x} \in K} \|\mathbf{x} - \mathbf{z}\|_2 \,.$$

If K is a subspace of $\mathbb{R}^N$ then it is the usual orthogonal projection onto K, and in particular, a linear map.

The proximal mapping can be expressed via subdifferentials as shown in the next statement.

Proposition B.23. *Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be a convex function. Then $\mathbf{x} = P_F(\mathbf{z})$ if and only if $\mathbf{z} \in \mathbf{x} + \partial F(\mathbf{x})$.*

Proof. By Theorem B.21 $\mathbf{x} = P_F(\mathbf{z})$ if and only if

$$0 \in \partial\big(\frac{1}{2}\|\cdot - \mathbf{z}\|_2^2 + F\big)(\mathbf{x}) \; .$$

The function $\mathbf{x} \mapsto \frac{1}{2}\|\cdot - z\|_2^2$ is differentiable with gradient $\nabla\big(\frac{1}{2}\|\cdot - z\|_2^2\big)(\mathbf{x}) = \mathbf{x} - \mathbf{z}$ so that the above condition reads $0 \in \mathbf{x} - \mathbf{z} + \partial F(\mathbf{x})$, which is equivalent to $\mathbf{z} \in \mathbf{x} + \partial F(\mathbf{x})$. □

The previous proposition justifies to write

$$P_F = (\mathbf{Id} + \partial F)^{-1} \; .$$

Moreau's identity relates the proximal mappings of F and F^*.

Theorem B.24. *(Moreau's identity) Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be a lower semicontinuous convex function. Then, for all $\mathbf{z} \in \mathbb{R}^N$,*

$$P_F(\mathbf{z}) + P_{F^*}(\mathbf{z}) = \mathbf{z} \; .$$

Proof. Let $\mathbf{x} = P_F(\mathbf{z})$ and set $\mathbf{y} := \mathbf{z} - \mathbf{x}$. By Proposition B.23 we have $\mathbf{z} \in \mathbf{x} + \partial F(\mathbf{x})$, that is, $\mathbf{y} = \mathbf{z} - \mathbf{x} \in \partial F(\mathbf{x})$. Since F is lower semicontinuous it follows from Theorem B.22 that $\mathbf{x} \in \partial F^*(\mathbf{y})$ or $\mathbf{z} \in \mathbf{y} + \partial F^*(\mathbf{y})$. By uniqueness of the minimizer in the definition of P_{F^*} it follows again from Proposition B.23 that $\mathbf{y} = P_{F^*}(\mathbf{z})$. In particular, we have shown that $P_F(\mathbf{z}) + P_{F^*}(\mathbf{z}) = \mathbf{x} + \mathbf{y} = \mathbf{z}$ by definition of $\mathbf{y}$. □

If P_F is easy to compute then the previous result shows that also $P_{F^*}(\mathbf{z}) = \mathbf{z} - P_F(\mathbf{z})$ is easy to compute. It is useful to note that applying Moreau's identity to the function τF for some $\tau > 0$ shows that

$$P_{\tau F}(\mathbf{z}) + \tau P_{\tau^{-1} F^*}(\mathbf{z}/\tau) = \mathbf{z} \; . \tag{B.14}$$

Indeed, $P_{\tau F}(\mathbf{z}) + P_{(\tau F)^*}(\mathbf{z}) = \mathbf{z}$, so that it remains to show that $P_{(\tau F)^*}(\mathbf{z}) = \tau P_{\tau^{-1} F}(\mathbf{z}/\tau)$. This follows from Proposition B.18(d),

$$\begin{aligned} P_{(\tau F)^*}(\mathbf{z}) &= \arg\min_{\mathbf{x} \in \mathbb{R}^N} \big\{ \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + (\tau F)^*(\mathbf{x}) \big\} \\ &= \arg\min_{\mathbf{x} \in \mathbb{R}^N} \big\{ \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + \tau F^*(\mathbf{x}/\tau) \big\} \\ &= \arg\min_{\mathbf{x} \in \mathbb{R}^N} \big\{ \tau^2 \big(\frac{1}{2}\|\mathbf{x}/\tau - \mathbf{z}/\tau\|_2^2 + \tau^{-1} F^*(\mathbf{x}/\tau) \big) \big\} = \tau P_{\tau^{-1} F^*}(\mathbf{z}/\tau) \; . \end{aligned}$$

Since for a lower semicontinuous function $F = F^{**}$ we can apply (B.14) to F^* in place of F to obtain

$$P_{\tau F^*}(\mathbf{z}) + \tau P_{\tau^{-1} F}(\mathbf{z}/\tau) = \mathbf{z} \; . \tag{B.15}$$

Theorem B.25. *For a convex function $F : \mathbb{R}^N \to (-\infty, \infty]$ the proximal mapping P_F is a contraction,*

$$\|P_F(\mathbf{z}) - P_F(\mathbf{z}')\|_2 \le \|\mathbf{z} - \mathbf{z}'\|_2 \quad \text{for all } \mathbf{z}, \mathbf{z}' \in \mathbb{R}^N .$$

Proof. Set $\mathbf{x} = P_F(\mathbf{z})$ and $\mathbf{x}' = P_F(\mathbf{z}')$. By Proposition B.23 we have $\mathbf{z} \in \mathbf{x} + \partial F(\mathbf{x})$, so that with $\mathbf{y} = \mathbf{z} - \mathbf{x}$ we have $\mathbf{y} \in \partial F(\mathbf{x})$. Theorem B.22 shows that $F(\mathbf{x}) + F^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$, and similarly, we can find $\mathbf{x}', \mathbf{y}'$ such that $\mathbf{z}' = \mathbf{x}' + \mathbf{y}'$ and $F(\mathbf{x}') + F^*(\mathbf{y}') = \langle \mathbf{x}', \mathbf{y}' \rangle$. It follows that

$$\|\mathbf{z} - \mathbf{z}'\|_2^2 = \|\mathbf{x} - \mathbf{x}'\|_2^2 + \|\mathbf{y} - \mathbf{y}'\|_2^2 + 2\langle \mathbf{x} - \mathbf{x}', \mathbf{y} - \mathbf{y}' \rangle . \tag{B.16}$$

Note that by Fenchel's inequality (B.8) we have

$$\langle \mathbf{x}, \mathbf{y}' \rangle \le F(\mathbf{x}) + F^*(\mathbf{y}') \quad \text{and} \quad \langle \mathbf{x}', \mathbf{y} \rangle \le F(\mathbf{x}') + F^*(\mathbf{y}) .$$

Therefore,

$$\begin{aligned} \langle \mathbf{x} - \mathbf{x}', \mathbf{y} - \mathbf{y}' \rangle &= \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}'\mathbf{y}' \rangle - \langle \mathbf{x}', \mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y}' \rangle \\ &= F(\mathbf{x}) + F^*(\mathbf{y}) + F(\mathbf{x}') + F^*(\mathbf{y}') - \langle \mathbf{x}', \mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y}' \rangle \ge 0 . \end{aligned}$$

Together with (B.16) this shows that $\|\mathbf{x} - \mathbf{x}'\|_2^2 \le \|\mathbf{z} - \mathbf{z}'\|_2^2$. □

Let us conclude with an important example of a proximal mapping. Let $F(x) = |x|$, $x \in \mathbb{R}$, be the absolute value function. Then a straightforward computation shows that, for $\tau > 0$,

$$\begin{aligned} P_{\tau F}(y) = \arg\min_{x \in \mathbb{R}} \big\{ \frac{1}{2}(x - y)^2 + \tau |x| \big\} &= \begin{cases} y - \tau & \text{if } y \ge \tau , \\ 0 & \text{if } |y| \le \tau , \\ y + \tau & \text{if } y \le -\tau \end{cases} \\ &=: S_\tau(y) . \end{aligned} \tag{B.17}$$

The function $S_\tau(y)$ is called soft thresholding or shrinkage operator. More generally, if $F(\mathbf{x}) = \|\mathbf{x}\|_1$ is the ℓ_1-norm on $\mathbb{R}^N$ then the minimization problem defining the proximal operator decouples and $P_{\tau F}(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^N$ is given entrywise by

$$P_{\tau F}(\mathbf{y})_\ell = S_\tau(y_\ell), \quad \ell \in [N] . \tag{B.18}$$

B.5 Convex Optimization Problems

An optimization problem is of the form

$$\underset{\mathbf{x} \in \mathbb{R}^N}{\text{minimize}}\ F_0(\mathbf{x}) \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} , \tag{B.19}$$

$$F_j(\mathbf{x}) \le b_j, \quad j \in [M] , \tag{B.20}$$

where the function $F_0 : \mathbb{R}^N \to (-\infty, \infty]$ is called *objective function*, the functions $F_1, \ldots, F_M : \mathbb{R}^N \to (-\infty, \infty]$ are called *constraint functions*, and $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m$ provide the equality constraint. A point $\mathbf{x}$ satisfying the constraints is called feasible, and (B.19) is called feasible if there exists a feasible point. A feasible point $\mathbf{x}^\sharp$ for which the minimum is attained, that is, $F_0(\mathbf{x}^\sharp) \leq F(\mathbf{x})$ for all feasible $\mathbf{x}$, is called a minimizer or optimal point, and $F_0(\mathbf{x}^\sharp)$ is the optimal value.

We note that the equality constraint maybe removed and represented by inequality constraints of the form $F_j(\mathbf{x}) \leq \mathbf{y}_j$, $-F_j(\mathbf{x}) \leq -\mathbf{y}_j$ with $F_j(\mathbf{x}) = \langle \mathbf{A}_j, \mathbf{x} \rangle$ where $\mathbf{A}_j \in \mathbb{R}^N$ is a row of $\mathbf{A}$.

The set of feasible points described by the constraints is given by

$$K = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{A}\mathbf{x} = \mathbf{y}, F_j(\mathbf{x}) \leq b_j, j \in [M]\} \,. \tag{B.21}$$

Two optimization problems are said to be equivalent if given the solution of one problem the solution to other problem can be "easily" computed. For the purpose of this short exposition we leave it at this rather vague definition of equivalence, and hope that it will be clear in concrete situations what is meant.

The optimization problem (B.19) is equivalent to the problem of minimizing F_0 over K,

$$\min_{\mathbf{x} \in K} F_0(\mathbf{x}) \,. \tag{B.22}$$

Introduce the characteristic function

$$\chi_K(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in K \,, \\ \infty & \text{if } \mathbf{x} \notin K \,. \end{cases}$$

Then our optimization problem becomes as well equivalent to the unconstrained optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^N} F_0(\mathbf{x}) + \chi_K(\mathbf{x}) \,.$$

A *convex optimization problem* (or *convex program*) is a problem of the form (B.19), in which the objective function F_0 and the constraint functions F_i are convex. In this case, the set of feasible points K defined in (B.21) is convex. The convex optimization problem becomes then equivalent to the unconstrained optimization problem of finding the minimum of the convex function

$$F(\mathbf{x}) = F_0(\mathbf{x}) + \chi_K(\mathbf{x}) \,.$$

Due to this equivalence we may freely switch between constrained and unconstrained optimization problems. Clearly, the statements of Proposition B.14 carry therefore over to constrained optimization problems. We only note that in constrained optimization problems the function F_0 is usually taken to be finite, i.e., $\mathrm{dom}(F) = \mathbb{R}^N$.

A *linear optimization problem* (or *linear program*) is one, where the objective function F_0 and all constraint functions $F_1, \dots, F_M$ are linear. Clearly, this is a special case of a convex optimization problem.

The *Lagrange function* of an optimization problem of the form (B.19) is defined for $\mathbf{x} \in \mathbb{R}^N, \boldsymbol{\xi} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^M$, $\nu_\ell \geq 0$, $\ell \in [M]$, by

$$L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) := F_0(\mathbf{x}) + \boldsymbol{\xi}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(\mathbf{x}) - b_\ell) \; . \tag{B.23}$$

For an optimization problem without inequality constraints we clearly set

$$L(\boldsymbol{\xi}) := F_0(\mathbf{x}) + \boldsymbol{\xi}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) \; . \tag{B.24}$$

The variables $\boldsymbol{\xi}, \boldsymbol{\nu}$ are called *Lagrange multipliers*. For ease of notation we write $\boldsymbol{\nu} \succcurlyeq \mathbf{0}$ if $\nu_\ell \geq 0$ for all $\ell \in [M]$. The *Lagrange dual function* is defined by

$$H(\boldsymbol{\xi}, \boldsymbol{\nu}) := \inf_{\mathbf{x} \in \mathbb{R}^N} L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) \; , \quad \boldsymbol{\xi} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^M, \boldsymbol{\nu} \succcurlyeq \mathbf{0} \; .$$

If $\mathbf{x} \mapsto L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu})$ is unbounded from below, then we set $H(\boldsymbol{\xi}, \boldsymbol{\nu}) = -\infty$. Again, if there are no inequality constraints then

$$H(\boldsymbol{\xi}) := \inf_{\mathbf{x} \in \mathbb{R}^N} L(\mathbf{x}, \boldsymbol{\xi}) = \inf_{\mathbf{x} \in \mathbb{R}^N} \big\{ F_0(\mathbf{x}) + \boldsymbol{\xi}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) \big\} \; , \quad \boldsymbol{\xi} \in \mathbb{R}^m \; .$$

The dual function is always concave because it is the pointwise infimum of a family of affine functions, even if the original problem (B.19) is not convex. The dual function provides a bound on the optimal value of $F_0(\mathbf{x}^\sharp)$ of the minimization problem (B.19),

$$H(\boldsymbol{\xi}, \boldsymbol{\nu}) \leq F(\mathbf{x}^\sharp) \quad \text{for all } \boldsymbol{\xi} \in \mathbb{R}^m, \boldsymbol{\nu} \succcurlyeq \mathbf{0} \; . \tag{B.25}$$

Indeed, if $\mathbf{x}$ is a feasible point for (B.19) then $\mathbf{A}\mathbf{x} - \mathbf{y} = 0$ and $F_\ell(\mathbf{x}) \leq 0$, $\ell = 1, \dots, M$, so that, for all $\boldsymbol{\xi} \in \mathbb{R}^m$ and $\boldsymbol{\nu} \succcurlyeq \mathbf{0}$,

$$\boldsymbol{\xi}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(\mathbf{x}) - b_\ell) \leq 0 \; .$$

Therefore,

$$L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) = F_0(\mathbf{x}) + \boldsymbol{\xi}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(\mathbf{x}) - b_\ell) \leq F_0(\mathbf{x}) \; .$$

Taking the infimum over all $\mathbf{x} \in \mathbb{R}^N$ on the left hand side, and over all feasible $\mathbf{x}$ on the right hand side shows (B.25). We would like this lower bound to be as tight as possible. This motivates to consider the optimization problem

$$\text{maximize } H(\boldsymbol{\xi}, \boldsymbol{\nu}) \quad \text{subject to } \boldsymbol{\nu} \succcurlyeq \mathbf{0} \; . \tag{B.26}$$

This optimization problem is called the *dual problem* to (B.19), which in this context is sometimes called primal problem. Since H is concave this problem is equivalent to the convex optimization problem of minimizing the convex function $-H$ subject to the positivity constraint $\boldsymbol{\nu} \succcurlyeq \mathbf{0}$. A pair $(\boldsymbol{\xi}, \boldsymbol{\nu})$, $\boldsymbol{\xi} \in \mathbb{R}^m$, $\boldsymbol{\nu} \succcurlyeq \mathbf{0}$ is called dual feasible. A (feasible) maximizer $(\boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp)$ of (B.26) is referred to as *dual optimal* or optimal Lagrange multipliers. If $\mathbf{x}^\sharp$ is optimal for the primal problem (B.19) then the triple $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp)$ is called primal dual optimal. Inequality (B.25) shows that always

$$H(\boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp) \leq F(\mathbf{x}^\sharp) \; . \tag{B.27}$$

This inequality is called *weak duality*. For most (but not all) convex optimization problems even strong duality holds,

$$H(\boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp) = F(\mathbf{x}^\sharp) \; . \tag{B.28}$$

Slater's constraint qualification, which we state in a simplified form below, provides a condition ensuring strong duality.

Theorem B.26. *(Slater's constraint qualification) Assume that $F_0, F_1, \ldots, F_M$ are convex functions with $\mathrm{dom}(F_0) = \mathbb{R}^N$. If there exists $\mathbf{x} \in \mathbb{R}^N$ such that $\mathbf{A}\mathbf{x} = \mathbf{y}$ and $F_\ell(\mathbf{x}) < 0$ for all $\ell \in [M]$, then strong duality holds for the optimization problem* (B.19)*.*
In case that there are no inequality constraints, strong duality holds provided there exists $\mathbf{x}$ with $\mathbf{A}\mathbf{x} = \mathbf{y}$, that is, if (B.19) *is feasible.*

Proof. See for instance [59, Section 5.3.2] or [254, Satz 8.1.7]. □

Given primal and dual feasible $(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu})$, that is, $\mathbf{x}$ is feasible for (B.19) and $\xi \in \mathbb{R}^m$, $\boldsymbol{\nu} \succcurlyeq \mathbf{0}$ the primal dual gap

$$E(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) = F(\mathbf{x}) - H(\boldsymbol{\xi}, \boldsymbol{\nu}) \tag{B.29}$$

can be taken as a measure on how close $\mathbf{x}$ is to the minimizer $\mathbf{x}^*$ of the primal problem (B.19), and how close $(\boldsymbol{\xi}, \boldsymbol{\nu})$ is to the maximizer of the dual problem (B.26). If $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp)$ is primal dual optimal, and strong duality holds then $E(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp) = 0$. The primal dual gap is often taken as a stopping criterion in iterative optimization methods.

For illustration let us compute the dual problem of the ℓ_1-minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y} \; . \tag{B.30}$$

The Lagrange function for this problem takes the form

$$L(\mathbf{x}, \boldsymbol{\xi}) = \|\mathbf{x}\|_1 + \boldsymbol{\xi}^*(\mathbf{A}\mathbf{x} - \mathbf{y}) \; .$$

The Lagrange dual function is

$$H(\boldsymbol{\xi}) = \inf_{\mathbf{x}\in\mathbb{R}^N} \{\|\mathbf{x}\|_1 + \langle \mathbf{A}^*\boldsymbol{\xi}, \mathbf{x}\rangle - \langle \boldsymbol{\xi}, \mathbf{y}\rangle\} \,.$$

If $\|\mathbf{A}^*\boldsymbol{\xi}\|_\infty > 1$ then there exists $\mathbf{x} \in \mathbb{R}^N$ such that $\langle \mathbf{A}^*\boldsymbol{\xi}, x\rangle < -\|\mathbf{x}\|_1$. Replacing $\mathbf{x}$ by $\lambda\mathbf{x}$ and letting $\lambda \to \infty$ shows that $H(\boldsymbol{\xi}) = -\infty$ in this case. If $\|\mathbf{A}^*\boldsymbol{\xi}\|_\infty \le 1$ then $\|\mathbf{x}\|_1 + \langle \mathbf{A}^*\boldsymbol{\xi}, x\rangle \ge 0$. The choice $\mathbf{x} = 0$ yields therefore the infimum, and $H(\boldsymbol{\xi}) = -\langle \boldsymbol{\xi}, \mathbf{y}\rangle$. In conclusion,

$$H(\boldsymbol{\xi}) = \begin{cases} -\langle \boldsymbol{\xi}, \mathbf{y}\rangle & \text{if } \|\mathbf{A}^*\boldsymbol{\xi}\|_\infty \le 1 \,, \\ -\infty & \text{otherwise} \,. \end{cases}$$

Clearly, it is enough to maximize over the points $\boldsymbol{\xi}$ for which $H(\boldsymbol{\xi}) > -\infty$. Making this constraint explicit, the dual program to (B.30) is given by

$$\max_{\boldsymbol{\xi}\in\mathbb{R}^m} -\langle \boldsymbol{\xi}, \mathbf{y}\rangle \quad \text{subject to } \|\mathbf{A}^*\boldsymbol{\xi}\|_\infty \le 1 \,. \tag{B.31}$$

By Theorem B.26 strong duality holds for this pair of primal and dual optimization problems provided the primal problem (B.30) is feasible.

Remark B.27. In the complex case $\mathbf{A} \in \mathbb{C}^{m\times N}$, $\mathbf{y} \in \mathbb{C}^m$ the inner product has to be replaced by the real inner product $\mathrm{Re}(\langle \mathbf{x}, \mathbf{y}\rangle)$ as noted in the beginning of this Chapter. Following the derivation above we see that the dual program of (B.30), where the minimum now ranges over $\mathbf{x} \in \mathbb{C}^N$, is given by

$$\max_{\boldsymbol{\xi}\in\mathbb{C}^m} -\mathrm{Re}(\langle \boldsymbol{\xi}, \mathbf{y}\rangle) \quad \text{subject to } \|\mathbf{A}^*\boldsymbol{\xi}\|_\infty \le 1 \,. \tag{B.32}$$

A *conic optimization problem* is of the form

$$\begin{aligned} \text{minimize}_{\mathbf{x}\in\mathbb{R}^N} F_0(\mathbf{x}) \quad \text{subject to } & \mathbf{x} \in K \,, \\ & F_\ell(\mathbf{x}) \le b_\ell, \quad \ell \in [M] \,, \end{aligned} \tag{B.33}$$

where K is a convex cone, and the F_ℓ are convex functions. If K is a second order cone, see (B.1) (possibly in a subset of variables, or the intersection of second order cones in different variables), then the above problem is called a second order cone problem. If K is the cone of positive semidefinite matrices then the above optimization problem is called a semidefinite program.

Also conic programs have their duality theory. The Lagrange function of a conic program of the above form is given by

$$L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) := F_0(\mathbf{x}) - \langle \mathbf{x}, \boldsymbol{\xi}\rangle + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(\mathbf{x}) - b_\ell), \quad \boldsymbol{\xi} \in K^*, \nu_\ell \ge 0 \,,$$

where K^* is the dual cone of K defined in (B.2). (If there are no inequality constraints then the last term above is omitted, of course.) The Lagrange dual function is then defined as

$$H(\boldsymbol{\xi}, \boldsymbol{\nu}) := \min_{\mathbf{x}\in\mathbb{R}^N} L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) \,, \quad \boldsymbol{\xi} \in K^*, \nu_\ell \ge 0 \,.$$

Similarly to (B.25) the minimizer $\mathbf{x}^\sharp$ of (B.25) satisfies the lower bound

$$H(\boldsymbol{\xi}, \boldsymbol{\nu}) \leq F_0(\mathbf{x}^\sharp) \,, \quad \text{for all } \boldsymbol{\xi} \in K^*, \boldsymbol{\nu} \succcurlyeq \mathbf{0} \,. \tag{B.34}$$

Indeed, if $\mathbf{x} \in K$ and $F_\ell(\mathbf{x}) \leq 0$, $\ell = 1, \ldots, M$, then $\langle \mathbf{x}, \boldsymbol{\xi} \rangle \geq 0$ for all $\boldsymbol{\xi} \in K^*$ by definition (B.2) of the dual cone and with $\boldsymbol{\nu} \succcurlyeq \mathbf{0}$,

$$-\langle \mathbf{x}, \boldsymbol{\xi} \rangle + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(\mathbf{x}) - b_\ell) \leq 0 \,.$$

Therefore,

$$L(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\nu}) = F_0(\mathbf{x}) - \langle \mathbf{x}, \boldsymbol{\xi} \rangle + \sum_{\ell=1}^{M} \nu_\ell F_\ell(\mathbf{x}) \leq F_0(\mathbf{x}) \,.$$

This point establishes (B.34). The dual program of (B.33) is then defined as

$$\text{maximize } H(\boldsymbol{\xi}, \boldsymbol{\nu}) \quad \text{subject to } \boldsymbol{\xi} \in K^*, \boldsymbol{\nu} \succcurlyeq \mathbf{0} \,. \tag{B.35}$$

Denote by $(\boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp)$ a dual optimum, that is, a maximizer of this program, and $\mathbf{x}^\sharp$ a minimum of the primal program (B.33). The triple $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp)$ is again called a primal dual optimum. The above arguments establish weak duality,

$$H(\boldsymbol{\xi}^\sharp, \boldsymbol{\nu}^\sharp) \leq F(\mathbf{x}^\sharp) \,. \tag{B.36}$$

If there is actually equality, then we say that strong duality holds. Similar conditions as in Slater's constraint qualification (Theorem B.26) ensure strong duality for conic programs; for instance, if there exists a point in the interior of K such that all inequality constraints hold strictly, see e.g. [59, Section 5.9].

Let us illustrate duality for conic programs with an example relevant to Section 9.2. For a convex cone K and a vector $\mathbf{g} \in \mathbb{R}^N$ we consider the optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^N} \langle \mathbf{x}, \mathbf{g} \rangle \quad \text{subject to } \mathbf{x} \in K \,, \|\mathbf{x}\|_2^2 \leq 1 \,.$$

Its Lagrange function is given by

$$L(\mathbf{x}, \boldsymbol{\xi}, \nu) = \langle \mathbf{x}, \mathbf{g} \rangle - \langle \boldsymbol{\xi}, \mathbf{x} \rangle + \nu(\|\mathbf{x}\|_2^2 - 1), \quad \boldsymbol{\xi} \in K^*, \nu \geq 0 \,.$$

The minimum with respect to $\mathbf{x}$ of the Lagrange function is attained at $\mathbf{x} = (2\nu)^{-1}(\boldsymbol{\xi} - \mathbf{g})$. By plugging this value into L, the Lagrange dual function turns out to be

$$\begin{aligned} H(\boldsymbol{\xi}, \nu) &= \min_{\mathbf{x} \in \mathbb{R}^N} \langle \mathbf{x}, \mathbf{g} \rangle - \langle \boldsymbol{\xi}, \mathbf{x} \rangle + \nu(\|\mathbf{x}\|_2^2 - 1) \\ &= -\nu - \frac{1}{4\nu} \|\mathbf{g} - \boldsymbol{\xi}\|_2^2 \,. \end{aligned}$$

This leads to the dual program

$$\max_{\boldsymbol{\xi},\nu} \; -\nu - \frac{1}{4\nu}\|\mathbf{g}-\boldsymbol{\xi}\|_2^2 \quad \text{subject to } \boldsymbol{\xi} \in K^*, \nu \geq 0 \; .$$

Solving this optimization program with respect to ν gives $\nu = \frac{1}{2}\|\mathbf{g}-\boldsymbol{\xi}\|_2$, so that we obtain the dual program

$$\max_{\boldsymbol{\xi}} -\|\mathbf{g}-\boldsymbol{\xi}\|_2 \quad \text{subject to } \boldsymbol{\xi} \in K^* \; . \tag{B.37}$$

Note that the minimizer is the orthogonal projection of $\mathbf{g}$ onto the dual cone K^*, which always exists since K^* is convex and closed. Weak duality for this case reads $\max_{\boldsymbol{\xi}\in K^*} -\|\mathbf{g}-\boldsymbol{\xi}\|_2 \leq \min_{\mathbf{x}\in K, \|\mathbf{x}\|_2 \leq 1} \langle \mathbf{g}, \mathbf{x}\rangle$, or

$$\max_{\mathbf{x}\in K, \|\mathbf{x}\|_2 \leq 1} \langle -\mathbf{g}, \mathbf{x}\rangle \leq \min_{\boldsymbol{\xi}\in K^*} \|\mathbf{g}-\boldsymbol{\xi}\|_2 \; . \tag{B.38}$$

In fact, often strong duality holds, that is, equality above; for instance, if K has non-empty interior. Note that the inequality above can be rewritten with the polar cone $K^\circ = -K^*$, see (B.3),

$$\max_{\mathbf{x}\in K, \|\mathbf{x}\|_2 \leq 1} \langle \mathbf{g}, \mathbf{x}\rangle \leq \min_{\boldsymbol{\xi}\in K^\circ} \|\mathbf{g}-\boldsymbol{\xi}\|_2 \; . \tag{B.39}$$

Lagrange duality has a saddle-point interpretation. For ease of exposition we consider (B.19) without inequality constraints, but note that extensions that include inequality constraints or conic programs are easily derived in the same way.

Let $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ be primal dual optimal. Recalling the definition of the Lagrange function L we have

$$\begin{aligned} \sup_{\boldsymbol{\xi}\in\mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\xi}) &= \sup_{\boldsymbol{\xi}\in\mathbb{R}^m} F_0(\mathbf{x}) + \boldsymbol{\xi}^*(A\mathbf{x}-\mathbf{y}) \\ &= \begin{cases} F_0(\mathbf{x}) & \text{if } \mathbf{A}\mathbf{x} = \mathbf{y} \; , \\ \infty & \text{otherwise} \; . \end{cases} \end{aligned} \tag{B.40}$$

In other words, the above supremum is ∞ if $\mathbf{x}$ is not feasible. The (feasible) minimizer $\mathbf{x}^\sharp$ of the primal problem (B.19) satisfies therefore

$$F_0(\mathbf{x}^\sharp) = \inf_{\mathbf{x}\in\mathbb{R}^N} \sup_{\boldsymbol{\xi}\in\mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\xi}) \; .$$

On the other hand, a dual optimal vector $\boldsymbol{\xi}^\sharp$ satisfies

$$H(\boldsymbol{\xi}^\sharp) = \sup_{\boldsymbol{\xi}\in\mathbb{R}^m} \inf_{\mathbf{x}\in\mathbb{R}^N} L(\mathbf{x}, \boldsymbol{\xi})$$

by definition of the Lagrange dual function. Weak duality implies therefore,

$$\sup_{\boldsymbol{\xi}\in\mathbb{R}^m} \inf_{\mathbf{x}\in\mathbb{R}^N} L(\mathbf{x},\boldsymbol{\xi}) \leq \inf_{\mathbf{x}\in\mathbb{R}^N} \sup_{\boldsymbol{\xi}\in\mathbb{R}^m} L(\mathbf{x},\boldsymbol{\xi}) \,,$$

while strong inequality reads

$$\sup_{\boldsymbol{\xi}\in\mathbb{R}^m} \inf_{\mathbf{x}\in\mathbb{R}^N} L(\mathbf{x},\boldsymbol{\xi}) = \inf_{\mathbf{x}\in\mathbb{R}^N} \sup_{\boldsymbol{\xi}\in\mathbb{R}^m} L(\mathbf{x},\boldsymbol{\xi}) \,.$$

In other words, the order of minimization and maximization can be interchanged in the case of strong duality. This property is called the strong *max-min property* or *saddle point property*. Indeed, in this case, a primal dual optimal $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ is a saddle point of the Lagrange function,

$$L(\mathbf{x}^\sharp,\boldsymbol{\xi}) \leq L(\mathbf{x}^\sharp,\boldsymbol{\xi}^\sharp) \leq L(\mathbf{x},\boldsymbol{\xi}^\sharp) \quad \text{for all } \mathbf{x}\in\mathbb{R}^N, \boldsymbol{\xi}\in\mathbb{R}^m \,. \tag{B.41}$$

Jointly optimizing the primal and dual problem is therefore equivalent to finding a saddle point of the Lagrange function provided that strong duality holds.

Based on these findings let us show the following theorem on the relation between certain optimization problems relevant to this book.

Theorem B.28. *Let $\|\cdot\|, \|\!|\cdot|\!\|$ be two norms on $\mathbb{R}^N$. For $\mathbf{A}\in\mathbb{R}^{m\times N}$, $\mathbf{y}\in\mathbb{R}^m$ and $\eta>0$ consider the optimization problem*

$$\min_{\mathbf{x}\in\mathbb{R}^N} |\!|\!|\mathbf{x}|\!|\!| \quad \textit{subject to } \|\mathbf{A}\mathbf{x}-\mathbf{y}\| \leq \eta \,. \tag{B.42}$$

Assume that (B.42) *is strictly feasible in the sense that there exists $\mathbf{x}\in\mathbb{R}^N$ such that $\|\mathbf{A}\mathbf{x}-\mathbf{y}\|<\eta$. Consider a minimizer $\mathbf{x}^\sharp$ of* (B.42). *Then there exists a parameter $\lambda\geq 0$ such that $\mathbf{x}^\sharp$ is also the minimizer of the optimization problem*

$$\min_{\mathbf{x}\in\mathbb{R}^N} |\!|\!|\mathbf{x}|\!|\!| + \lambda\|\mathbf{A}\mathbf{x}-\mathbf{y}\| \,. \tag{B.43}$$

Conversely, for $\lambda>0$ let $\mathbf{x}^\sharp$ be a minimizer of (B.43). *Then there exists $\eta\geq 0$ such that $\mathbf{x}^\sharp$ is a minimizer of* (B.42).

Clearly, this theorem holds also in the complex setting by simply interpreting $\mathbb{C}^N=\mathbb{R}^{2N}$.

Proof. The Lagrange function of (B.42) is given by

$$L(\mathbf{x},\xi) = |\!|\!|\mathbf{x}|\!|\!| + \xi(\|\mathbf{A}\mathbf{x}-\mathbf{y}\|-\eta)$$

By Theorem B.26 strong duality holds for (B.42). Therefore, there exists a dual optimal $\xi^\sharp\geq 0$. The saddle point property (B.41) implies that $L(\mathbf{x}^\sharp,\xi^\sharp)\leq L(\mathbf{x},\xi^\sharp)$ for all $\mathbf{x}\in\mathbb{R}^N$. Therefore, $\mathbf{x}^\sharp$ is also a minimizer of $\mathbf{x}\mapsto L(\mathbf{x},\xi^\sharp)$. Since the constant term $-\xi^\sharp\eta$ does not affect the minimizer the conclusion follows with $\lambda=\xi^\sharp$.

For the converse statement, let $\mathbf{x}^\sharp$ be the minimizer of (B.43) and set $\xi=\lambda$. Choose $\eta=\|\mathbf{A}\mathbf{x}^\sharp-\mathbf{y}\|$. Then the Lagrange dual function H satisfies $H(\xi)=L(\mathbf{x}^\sharp,\xi)=|\!|\!|\mathbf{x}^\sharp|\!|\!|$. By weak duality $H(\xi)\leq |\!|\!|\mathbf{x}|\!|\!|$ for all feasible $\mathbf{x}$. Since $\mathbf{x}^\sharp$ is feasible by the choice of η, it follows that $\mathbf{x}^\sharp$ is a minimizer of (B.42). □

Remark B.29. (a) The same type of statement and proof is valid for the pair of optimization problems (B.43) and

$$\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{A}\mathbf{x}-\mathbf{y}\| \quad \text{subject to } |\!|\!|\mathbf{x}|\!|\!| \le t\,, t>0\,. \tag{B.44}$$

Note that in this case strict feasibility always holds because the zero vector $\mathbf{x}=\mathbf{0}$ satisfies $\|\mathbf{x}\|<t$.

(b) The equivalence of the problems (B.42), (B.43), (B.44) is somewhat implicit because the parameter transformation between λ and η or λ and t is implicit in the sense that it depends on the respective minimizer. Therefore, it can only be performed after solving the optimization problem, which makes this equivalence somewhat theoretical for practical purposes.

For the remainder of this section we consider a convex optimization problem of the form

$$\min_{\mathbf{x}\in\mathbb{R}^N} F(\mathbf{A}\mathbf{x})+G(\mathbf{x})\,, \tag{B.45}$$

with $\mathbf{A}\in\mathbb{R}^{m\times N}$ and convex functions $F:\mathbb{R}^m\to(-\infty,\infty]$, $G:\mathbb{R}^N\to(-\infty,\infty]$. All the relevant optimization problems appearing in this book fall into this class, see Section 15.2. For instance, the choice $G(\mathbf{x})=\|\mathbf{x}\|_1$ and $F=\chi_{\{\mathbf{y}\}}$, the characteristic function (B.7) of the singleton $\{\mathbf{y}\}$, yields the ℓ_1-minimization problem (B.30).

The substitution $\mathbf{z}=\mathbf{A}\mathbf{x}$ yields the equivalent problem

$$\min_{\mathbf{x}\in\mathbb{R}^N,\mathbf{z}\in\mathbb{R}^m} F(\mathbf{z})+G(\mathbf{x}) \quad \text{subject to } \mathbf{A}\mathbf{x}-\mathbf{z}=\mathbf{0}\,. \tag{B.46}$$

The Lagrange dual function to this problem is given by

$$\begin{aligned} H(\boldsymbol{\xi}) &= \inf_{\mathbf{x},\mathbf{z}} \{F(\mathbf{z})+G(\mathbf{x})+\langle \mathbf{A}^*\boldsymbol{\xi},\mathbf{x}\rangle - \langle \boldsymbol{\xi},\mathbf{z}\rangle\} \\ &= -\sup_{\mathbf{z}\in\mathbb{R}^m}\{\langle\boldsymbol{\xi},\mathbf{z}\rangle - F(\mathbf{z})\} - \sup_{\mathbf{x}\in\mathbb{R}^N}\{\langle\mathbf{x},-\mathbf{A}^*\boldsymbol{\xi}\rangle - G(\mathbf{x})\} \\ &= -F^*(\boldsymbol{\xi}) - G^*(-\mathbf{A}^*\boldsymbol{\xi})\,, \end{aligned} \tag{B.47}$$

where F^* and G^* are the convex conjugate functions of F and G, respectively. Therefore, the dual problem of (B.45) is

$$\max_{\boldsymbol{\xi}\in\mathbb{R}^m} -F^*(\boldsymbol{\xi}) - G^*(-\mathbf{A}^*\boldsymbol{\xi})\,. \tag{B.48}$$

Since the maximal values of (B.45) and (B.46) coincide we refer to (B.48) also as the dual problem of (B.45) – although strictly speaking (B.48) is not the dual to (B.45) in the sense described above. (Indeed, an unconstrained optimization problem does not introduce dual variables in the Lagrange function. In general, equivalent problems may have non-equivalent duals.)

The following theorem states strong duality of the problems (B.45) and (B.48).

Theorem B.30. *Let* $\mathbf{A} \in \mathbb{R}^{m\times N}$ *and* $F : \mathbb{R}^m \to (-\infty, \infty]$, $G : \mathbb{R}^N \to (-\infty, \infty]$ *be proper convex functions such that either* $\operatorname{dom}(F) = \mathbb{R}^m$ *or* $\operatorname{dom}(G) = \mathbb{R}^N$ *and there exists* $\mathbf{x}$ *such that* $\mathbf{Ax} \in \operatorname{dom}(F)$. *Assume that the optima in* (B.45) *and* (B.48) *are attained. Then strong duality holds in the form*

$$\min_{\mathbf{x}\in\mathbb{R}^N} F(\mathbf{Ax}) + G(\mathbf{x}) = \max_{\boldsymbol{\xi}\in\mathbb{R}^m} -F^*(\boldsymbol{\xi}) - G^*(-\mathbf{A}^*\boldsymbol{\xi}) \; .$$

Furthermore, a primal dual optimum $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ *is a solution to the saddle point problem*

$$\min_{\mathbf{x}\in\mathbb{R}^N} \max_{\boldsymbol{\xi}\in\mathbb{R}^m} \langle \mathbf{Ax}, \boldsymbol{\xi}\rangle + G(\mathbf{x}) - F^*(\boldsymbol{\xi}) \; , \tag{B.49}$$

where F^* *is the convex conjugate of* F.

Proof. The first statement follows from Fenchel's duality theorem, see e.g. [366, Theorem 31.1]. Strong duality implies the saddle point property (B.41) of the Lagrange function. By (B.47) the value of the Lagrange function in the primal dual optimal point is the optimal value of the min-max problem

$$\begin{aligned}
&\min_{\mathbf{x},\mathbf{z}\in\mathbb{R}^N} \max_{\boldsymbol{\xi}\in\mathbb{R}^m} F(\mathbf{z}) + G(\mathbf{x}) + \langle \mathbf{A}^*\boldsymbol{\xi}, \mathbf{x}\rangle - \langle \boldsymbol{\xi}, \mathbf{z}\rangle \\
&= \min_{\mathbf{x}\in\mathbb{R}^N} \max_{\boldsymbol{\xi}\in\mathbb{R}^m} -\big(\min_{\mathbf{z}\in\mathbb{R}^m} \langle \boldsymbol{\xi}, \mathbf{z}\rangle - F(\mathbf{z})\big) + \langle \mathbf{A}^*\boldsymbol{\xi}, \mathbf{x}\rangle + G(\mathbf{x}) \\
&= \min_{\mathbf{x}\in\mathbb{R}^N} \max_{\boldsymbol{\xi}\in\mathbb{R}^m} \langle \mathbf{Ax}, \boldsymbol{\xi}\rangle + G(\mathbf{x}) - F^*(\boldsymbol{\xi})
\end{aligned}$$

by definition of the convex conjugate function. The interchange of the minimum and maximum above is justified due to the fact that if $((\mathbf{x}^\sharp, \mathbf{z}^\sharp), \boldsymbol{\xi}^\sharp)$ is a saddle point of $L((\mathbf{x}, \mathbf{z}), \boldsymbol{\xi})$ then $(\mathbf{x}^\sharp, \boldsymbol{\xi}^\sharp)$ is a saddle point of $H(\mathbf{x}, \boldsymbol{\xi}) = \min_{\mathbf{z}} L((\mathbf{x}, \mathbf{z}), \boldsymbol{\xi})$. □

The condition $\operatorname{dom}(F) = \mathbb{R}^m$ or $\operatorname{dom}(G) = \mathbb{R}^N$ above maybe relaxed, see e.g. [366, Theorem 31.1].

B.6 Matrix Convexity

We recall the notion of matrix functions from Section A.5, in particular, of the matrix exponential and matrix logarithm. The main goal of this section will be to show the following concavity theorem due to Lieb [281], which is a key ingredient in the proof of the noncommutative Bernstein inequality in Section 8.5.

Theorem B.31. *Let* $\mathbf{H}$ *be a self-adjoint matrix. Then the function*

$$\mathbf{X} \mapsto \operatorname{tr} \exp(\mathbf{H} + \ln(\mathbf{X}))$$

is concave on the set of positive definite matrices.

While the original proof [281], and variants [163, 379, 380] use complex analysis, we proceed as in [423], see also [40]. This requires to introduce some background from matrix convexity and some concepts from quantum information theory.

Given a function $f : I \to \mathbb{R}$ on an interval $I \subset \mathbb{R}$, we recall that f is extended to self-adjoint matrices $\mathbf{A}$ with eigenvalues contained in I by (A.42). Similarly to the definition (A.49) of operator monotonicity, we say that f is *matrix convex* (or operator convex) if for any $n \in \mathbb{N}$, for all self-adjoint matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ with eigenvalues in I and for all $t \in [0,1]$

$$f(t\mathbf{A} + (1-t)\mathbf{B}) \preccurlyeq tf(\mathbf{A}) + (1-t)f(\mathbf{B}) \,. \tag{B.50}$$

Equivalently, f is matrix convex if for all n, and all $\mathbf{x} \in \mathbb{C}^n$ the scalar-valued function $\mathbf{A} \mapsto \langle f(\mathbf{A})\mathbf{x}, \mathbf{x}\rangle$ is convex on the set of self-adjoint matrices in $\mathbb{C}^{n\times n}$ with eigenvalues in I. As for matrix monotonicity, matrix convexity is a much stronger property than the usual scalar convexity.

We start with a simple characterization, for which we recall that a self-adjoint matrix $\mathbf{P}$ is called a projection if $\mathbf{P}^2 = \mathbf{P}$. Here and in the following, when matrix dimensions are not specified, they are arbitrary, but the matrices are assumed to be of matching dimension so that matrix multiplication is well-defined.

Theorem B.32. *Let $I \subset \mathbb{R}$ be an interval containing 0, and $f : I \to \mathbb{R}$. Then f is matrix convex and $f(0) \leq 0$ if and only if $f(\mathbf{PAP}) \preccurlyeq \mathbf{P}f(\mathbf{A})\mathbf{P}$ for all projections $\mathbf{P}$ and all self-adjoint matrices $\mathbf{A}$ with eigenvalues in I.*

Proof. We only prove matrix convexity based on the given condition as the converse direction will not be needed in the sequel. (A proof of the other direction and more equivalences can be found in [38, Theorem V.2.3] or [223, Theorem 2.1].)

Let $\mathbf{A}, \mathbf{B}$ be self-adjoint matrices of the same dimension with eigenvalues in I, and $t \in [0,1]$. Define $\mathbf{T}, \mathbf{P}, \mathbf{V}_t$ to be the block matrices

$$\mathbf{T} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \,, \quad P = \begin{pmatrix} \mathbf{Id} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \,, \quad \mathbf{V}_t = \begin{pmatrix} \sqrt{t}\,\mathbf{Id} & -\sqrt{1-t}\,\mathbf{Id} \\ \sqrt{1-t}\,\mathbf{Id} & \sqrt{t}\,\mathbf{Id} \end{pmatrix} \,.$$

The matrix $\mathbf{V}_t$ is unitary and $\mathbf{P}$ is a projection. Observe that

$$\mathbf{P}\mathbf{V}_t^*\mathbf{T}\mathbf{V}_t\mathbf{P} = \begin{pmatrix} t\mathbf{A} + (1-t)\mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \,.$$

Therefore, by (A.44) and by the hypothesis we have

$$\begin{aligned} \begin{pmatrix} f(t\mathbf{A} + (1-t)\mathbf{B}) & \mathbf{0} \\ \mathbf{0} & f(\mathbf{0}) \end{pmatrix} &= f(\mathbf{P}\mathbf{V}_t^*\mathbf{T}\mathbf{V}_t\mathbf{P}) \preccurlyeq \mathbf{P}f(\mathbf{V}_t^*\mathbf{T}\mathbf{V}_t)\mathbf{P} \\ &= \mathbf{P}\mathbf{V}_t^* f(\mathbf{T})\mathbf{V}_t\mathbf{P} = \begin{pmatrix} tf(\mathbf{A}) + (1-t)f(\mathbf{B}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \,. \end{aligned}$$

The first equality in the second line is valid due to unitarity of $\mathbf{V}_t$. This shows that f is matrix convex and $f(0) \leq 0$. □

Theorem B.33. *Let f be a continuous function on $[0,\infty)$. Then f is matrix convex and $f(0) \le 0$ if and only if the function $g(t) = f(t)/t$ is matrix monotone on $(0,\infty)$.*

Proof. We only prove that f is matrix convex with $f(0) \le 0$ if g is matrix monotone because we will require only this part of theorem. For the converse statement we refer to [38, Theorem V.2.9] or [223, Theorem 2.4].

Let $\mathbf{A}$ be an arbitrary self-adjoint matrix with eigenvalues in $(0,\infty)$ and $\mathbf{P}$ be an arbitrary projection (of the same dimension as $\mathbf{A}$). Let $\varepsilon > 0$. Since $\mathbf{P} + \varepsilon\mathbf{Id} \preccurlyeq (1+\varepsilon)\mathbf{Id}$ it follows that $\mathbf{A}^{1/2}(\mathbf{P}+\varepsilon\mathbf{Id})\mathbf{A}^{1/2} \preccurlyeq (1+\varepsilon)\mathbf{A}$ by Lemma A.32. The matrix monotonicity of g implies then

$$\left(\mathbf{A}^{1/2}(\mathbf{P}+\varepsilon\mathbf{Id})\mathbf{A}^{1/2}\right)^{-1} f(\mathbf{A}^{1/2}(\mathbf{P}+\varepsilon\mathbf{Id})\mathbf{A}^{1/2}) \preccurlyeq (1+\varepsilon)^{-1}\mathbf{A}^{-1} f((1+\varepsilon)\mathbf{A}) \,.$$

By multiplying with $(\mathbf{P}+\varepsilon\mathbf{Id})\mathbf{A}^{1/2}$ on the left, and $\mathbf{A}^{1/2}(\mathbf{P}+\varepsilon\mathbf{Id})$ on the right hand side, and using that $f((1+\varepsilon)\mathbf{A})$ commutes with $\mathbf{A}^{-1/2}$, we reach

$$\begin{aligned}&\mathbf{A}^{-1/2} f(\mathbf{A}^{1/2}(\mathbf{P}+\varepsilon\mathbf{Id})\mathbf{A}^{1/2})\mathbf{A}^{1/2}(\mathbf{P}+\varepsilon\mathbf{Id})\\ &\preccurlyeq (1+\varepsilon)^{-1}(\mathbf{P}+\varepsilon\mathbf{Id}) f((1+\varepsilon)\mathbf{A})(\mathbf{P}+\varepsilon\mathbf{Id}) \,.\end{aligned}$$

Letting ε tend to zero we obtain by continuity of f that

$$\mathbf{A}^{-1/2} f(\mathbf{A}^{1/2}\mathbf{P}\mathbf{A}^{1/2})\mathbf{A}^{1/2}\mathbf{P} \preccurlyeq \mathbf{P} f(\mathbf{A})\mathbf{P} \,. \tag{B.51}$$

The identity

$$f(\mathbf{A}^{1/2}\mathbf{P}\mathbf{A}^{1/2})\mathbf{A}^{1/2}\mathbf{P} = \mathbf{A}^{1/2}\mathbf{P} f(\mathbf{P}\mathbf{A}\mathbf{P}) \tag{B.52}$$

holds for all monomials $f(t) = t^n$, $n \in \mathbb{N}$. By the Weierstrass approximation theorem it extends to all continuous functions on any compact interval containing the eigenvalues of the involved matrices, and therefore to all continuous functions on $[0,\infty)$. Plugging (B.52) into (B.51) shows that

$$\mathbf{P} f(\mathbf{P}\mathbf{A}\mathbf{P}) \preccurlyeq \mathbf{P} f(\mathbf{A})\mathbf{P} \,. \tag{B.53}$$

Note that (scalar) monotonicity of $f(t)/t$ on $(0,\infty)$ implies that $f(0) \le 0$. Indeed, for $0 < s \le t$ we have $f(s)/s \le f(t)/t$, hence $tf(s) \le sf(t)$. Letting $s \to 0$ shows that, for any $t > 0$, $tf(0) \le 0$, hence, $f(0) \le 0$.

Since $\mathbf{P} \preccurlyeq \mathbf{Id}$ for a projection it follows that for the zero matrix $f(\mathbf{0}) \preccurlyeq \mathbf{P} f(\mathbf{0})$. Set $g(t) = f(t) - f(0)$. Then on a compact interval containing the eigenvalues of $\mathbf{PAP}$, the function g can be approximated by a linear combination of the monomials $g_n(t) = t^n$, $n \ge 1$, by the Weierstrass approximation theorem. (Note that the constant function $g_0(t) = 1$ is not needed as $g(0) = 0$.) For such monomials $g_n(\mathbf{PAP}) = (\mathbf{PAP})^n = \mathbf{P}(\mathbf{PAP})^n = \mathbf{P} g_n(\mathbf{PAP})$ whenever $n \ge 1$. Therefore, this property extends to g, i.e., $g(\mathbf{PAP}) = \mathbf{P} g(\mathbf{PAP})$. We obtain

$$f(\mathbf{PAP}) = g(\mathbf{PAP}) - f(\mathbf{0}) \preccurlyeq g(\mathbf{PAP}) + \mathbf{P}f(\mathbf{0}) = \mathbf{P}(g(\mathbf{PAP}) + f(\mathbf{0})) \\ = \mathbf{P}f(\mathbf{PAP}) \; .$$

Combine this with (B.53) to conclude $f(\mathbf{PAP}) \preccurlyeq \mathbf{P}f(\mathbf{A})\mathbf{P}$. Since $\mathbf{P}$ and $\mathbf{A}$ were arbitrary f is matrix convex by Theorem B.32. □

We are particularly interested in the following special case.

Corollary B.34. *The continuous function $\phi(x) = x\ln(x)$, $x > 0$, $\phi(0) = 0$ is matrix convex on $[0, \infty)$.*

Proof. Combine Proposition A.35 with Theorem B.33.

Next we state the affine version of the Hansen-Pedersen-Jensen inequality.

Theorem B.35. *Let f be matrix convex on some interval $I \subset \mathbb{R}$, and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be square matrices such that $\sum_{j=1}^n \mathbf{X}_j^* \mathbf{X}_j = \mathbf{Id}$. Then, for self-adjoint matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ with eigenvalues in I,*

$$f\left(\sum_{j=1}^n \mathbf{X}_j^* \mathbf{A}_j \mathbf{X}_j\right) \preccurlyeq \sum_{j=1}^n \mathbf{X}_j^* f(\mathbf{A}_j) \mathbf{X}_j \; . \tag{B.54}$$

The converse of this theorem holds as well in the sense that if (B.54) holds for arbitrary choices of $\mathbf{A}_j$, $\mathbf{X}_j$ then f is matrix convex [224, Theorem 2.1]. The proof requires the following auxiliary lemma.

Lemma B.36. *Let $\mathbf{B}_{jk} \in \mathbb{C}^{m\times m}$, $j, k \in [n]$, be a double sequence of square matrices and form the block matrix $\mathbf{B} = (\mathbf{B}_{ij}) \in \mathbb{C}^{mn\times mn}$. Set $\omega = e^{2\pi i/n}$ and let $\mathbf{E} \in \mathbb{C}^{mn\times mn}$ be the unitary block diagonal matrix $\mathbf{E} = \operatorname{diag}(\omega^j \mathbf{Id}, j \in [n])$. Then*

$$\frac{1}{n}\sum_{k=1}^n \mathbf{E}^{-k}\mathbf{B}\mathbf{E}^k = \operatorname{diag}(\mathbf{B}_{11}, \mathbf{B}_{22}, \ldots, \mathbf{B}_{nn}) \; .$$

Proof. A direct computation shows that

$$(\mathbf{E}^{-k}\mathbf{B}\mathbf{E}^k)_{j\ell} = (\omega^{k(\ell-j)}\mathbf{B}_{j\ell})_{j\ell}.$$

Since by the formula for geometric sums, for $j, \ell \in [n]$,

$$\sum_{k=1}^n \omega^{k(\ell-j)} = \sum_{k=1}^n e^{2\pi i(\ell-j)k/n} = \begin{cases} n & \text{if } \ell = j \; , \\ 0 & \text{otherwise} \end{cases} ,$$

this establishes the claim. □

Proof (of Theorem B.35). Define the block matrices

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \dots \\ \mathbf{X}_n \end{pmatrix}$$

and

$$\mathbf{U} = \begin{pmatrix} \mathbf{Id} - \mathbf{X}\mathbf{X}^* & \mathbf{X} \\ -\mathbf{X}^* & \mathbf{0} \end{pmatrix} .$$

We have

$$\mathbf{U}\mathbf{U}^* = \begin{pmatrix} (\mathbf{Id} - \mathbf{X}\mathbf{X}^*)^2 + \mathbf{X}\mathbf{X}^* & \mathbf{X}\mathbf{X}^*\mathbf{X} - \mathbf{X} \\ -\mathbf{X}^* + \mathbf{X}^*\mathbf{X}\mathbf{X}^* & \mathbf{X}^*\mathbf{X} \end{pmatrix} = \mathbf{Id}$$

because by assumption $\mathbf{X}^*\mathbf{X} = \mathbf{Id}$, and similarly $\mathbf{U}^*\mathbf{U} = \mathbf{Id}$. Therefore, $\mathbf{U}$ is a unitary matrix, also called the unitary dilation of $\mathbf{X}$. Divide $\mathbf{U} = (\mathbf{U}_{jk})_{j,k\in[n+1]}$ into blocks so that $\mathbf{U}_{k,n+1} = \mathbf{A}_k$ for $k \in [n]$ and $\mathbf{U}_{n+1,n+1} = 0$. Further, let $\mathbf{A}$ be the block diagonal matrix $\mathbf{A} = \operatorname{diag}(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_n, 0)$. Using Lemma B.36 with n replaced by $n+1$ together with the matrix convexity of f we obtain

$$\begin{aligned} f\left(\sum_{j=1}^{n} \mathbf{X}_j^* \mathbf{A}_j \mathbf{X}_j\right) &= f\left((\mathbf{U}^*\mathbf{A}\mathbf{U})_{n+1,n+1}\right) \\ &= f\left(\left(\frac{1}{n+1}\sum_{k=1}^{n+1} \mathbf{E}^{-k}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{E}^k\right)_{n+1,n+1}\right) \\ &= \left(f\left(\frac{1}{n+1}\sum_{k=1}^{n+1} \mathbf{E}^{-k}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{E}^k\right)\right)_{n+1,n+1} \\ &\preccurlyeq \left(\frac{1}{n+1}\sum_{k=1}^{n+1} f\left(\mathbf{E}^{-k}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{E}^k\right)\right)_{n+1,n+1} \end{aligned}$$

The equality in the third line is due to the fact that the matrix in the argument of f in the second line is block diagonal by Lemma B.36. By unitarity of $\mathbf{E}$ and $\mathbf{U}$ and by the definition of f on self-adjoint matrices the previous term equals

$$\begin{aligned} \left(\frac{1}{n}\sum_{k=1}^{n+1} \mathbf{E}^{-k}\mathbf{U}^* f\left(\mathbf{A}\right)\mathbf{U}\mathbf{E}^k\right)_{n+1,n+1} &= \left(\mathbf{U}^* f(\mathbf{A})\mathbf{U}\right)_{n+1,n+1} \\ &= \sum_{j=1}^{n} \mathbf{X}_j^* f(\mathbf{A}_j)\mathbf{X}_j \ . \end{aligned}$$

This completes the proof. □

Our next tool is the *perspective*. In the scalar case, given a convex function f on some convex set $K \subset \mathbb{R}^n$ it is defined via $g(x,t) = tf(x/t)$, $t > 0$ and $x/t \in K$. It is straightforward to check that g is jointly convex in (x,t), that is, g is convex function in the variable $\mathbf{y} = (x,t)$. As an important example, the perspective of the convex function $f(x) = x \ln x$, $x \geq 0$, yields the jointly convex function

$$g(x,t) = x \ln x - x \ln t \; . \tag{B.55}$$

Now given a matrix convex function $f : (0,\infty) \to \mathbb{R}$ we define its perspective on positive definite matrices $\mathbf{A}, \mathbf{B}$ via

$$g(\mathbf{A},\mathbf{B}) = \mathbf{B}^{1/2} f(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})\mathbf{B}^{1/2} \; . \tag{B.56}$$

By the next theorem due to Effros [153] g is jointly matrix convex in $(\mathbf{A},\mathbf{B})$.

Theorem B.37. *Let $f : (0,\infty) \to \mathbb{R}$ be a matrix convex function. Then the perspective g defined by* (B.56) *is jointly matrix convex in the sense that for all positive definite $\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}_1, \mathbf{B}_2$ (of matching dimension) and $t \in [0,1]$,*

$$g(t\mathbf{A}_1 + (1-t)\mathbf{A}_2, t\mathbf{B}_1 + (1-t)\mathbf{B}_2) \preccurlyeq tg(\mathbf{A}_1,\mathbf{B}_1) + (1-t)g(\mathbf{A}_2,\mathbf{B}_2) \; .$$

Proof. Let $\mathbf{A} := t\mathbf{A}_1 + (1-t)\mathbf{A}_2$ and $\mathbf{B} := t\mathbf{B}_1 + (1-t)\mathbf{B}_2$. The matrices $\mathbf{X}_1 := (t\mathbf{B}_1)^{1/2}\mathbf{B}^{-1/2}$ and $\mathbf{X}_2 := ((1-t)\mathbf{B}_2)^{1/2}\mathbf{B}^{-1/2}$ satisfy

$$\mathbf{X}_1^*\mathbf{X}_1 + \mathbf{X}_2^*\mathbf{X}_2 = t\mathbf{B}^{-1/2}\mathbf{B}_1\mathbf{B}^{-1/2} + (1-t)\mathbf{B}^{-1/2}\mathbf{B}_2\mathbf{B}^{-1/2} = \mathbf{Id} \; .$$

Theorem B.35 together with Lemma A.32 implies then that

$$\begin{aligned}
g(\mathbf{A},\mathbf{B}) &= \mathbf{B}^{1/2} f\left(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\right)\mathbf{B}^{1/2} \\
&= \mathbf{B}^{1/2} f\left(\mathbf{X}_1^*\mathbf{B}_1^{-1/2}\mathbf{A}_1\mathbf{B}_1^{-1/2}\mathbf{X}_1 + \mathbf{X}_2^*\mathbf{B}_2^{-1/2}\mathbf{A}_2\mathbf{B}_2^{-1/2}\mathbf{X}_2\right)\mathbf{B}^{1/2} \\
&\preccurlyeq \mathbf{B}^{1/2}\left(\mathbf{X}_1^* f(\mathbf{B}_1^{-1/2}\mathbf{A}_1\mathbf{B}_1^{-1/2})\mathbf{X}_1 + \mathbf{X}_2^* f(\mathbf{B}_2^{-1/2}\mathbf{A}_2\mathbf{B}_2^{-1/2})\mathbf{X}_2\right)\mathbf{B}^{1/2} \\
&= t\mathbf{B}_1^{1/2} f(\mathbf{B}_1^{-1/2}\mathbf{A}_1\mathbf{B}_1^{-1/2})\mathbf{B}_1^{1/2} + (1-t)\mathbf{B}_2^{1/2} f(\mathbf{B}_2^{-1/2}\mathbf{A}_2\mathbf{B}_2^{-1/2})\mathbf{B}_2^{1/2} \\
&= tg(\mathbf{A}_1,\mathbf{B}_1) + (1-t)g(\mathbf{A}_2,\mathbf{B}_2) \; .
\end{aligned}$$

This concludes the proof. □

Next we introduce a concept from quantum information theory [321, 331].

Definition B.38. *For two positive definite matrices $\mathbf{A}, \mathbf{B}$ the* quantum relative entropy *is defined as*

$$\mathscr{D}(\mathbf{A},\mathbf{B}) := \operatorname{tr}\left(\mathbf{A}\ln\mathbf{A} - \mathbf{A}\ln\mathbf{B} - (\mathbf{A}-\mathbf{B})\right) \; .$$

If $\mathbf{A}, \mathbf{B}$ are scalars then the above definition reduces to the usual scalar relative entropy (B.55) (up to the term $\mathbf{A} - \mathbf{B}$).

The quantum relative entropy is non-negative, a fact that is also known as *Klein's inequality*.

Theorem B.39. *Let* $\mathbf{A}, \mathbf{B}$ *be positive definite matrices. Then*

$$\mathscr{D}(\mathbf{A}, \mathbf{B}) \geq 0 \, .$$

Proof. The scalar function $\phi(x) = x \ln x$, $x > 0$, is convex (even matrix convex by Corollary B.34). It follows from Proposition B.9 that

$$x \ln x = \phi(x) \geq \phi(y) + \phi'(y)(x-y) = y \ln y + (1 + \ln y)(x-y) = x \ln y + (x-y) \, ,$$

so that $x \ln x - x \ln y - (x-y) \geq 0$. Theorem A.31 shows that $\mathscr{D}(\mathbf{A}, \mathbf{B}) \geq 0$. □

As a consequence we obtain a variational formula for the trace.

Corollary B.40. *Let* B *be a positive definite matrix. Then*

$$\operatorname{tr} \mathbf{B} = \max_{\mathbf{A} \succ \mathbf{0}} \operatorname{tr} \left(\mathbf{A} \ln \mathbf{B} - \mathbf{A} \ln \mathbf{A} + \mathbf{A} \right) .$$

Proof. By definition of the quantum relative entropy and Theorem B.39

$$\operatorname{tr} \mathbf{B} \geq \operatorname{tr} \left(\mathbf{A} \ln \mathbf{B} - \mathbf{A} \ln \mathbf{A} + \mathbf{A} \right) .$$

Choosing $\mathbf{A} = \mathbf{B}$ yields equality above and establishes the claim. □

Generalizing the convexity of the standard relative entropy (B.55) (or Kullback-Leibler divergence), the quantum relative entropy is jointly convex. This fact goes back to Lindblad [282], see also [348, 429]. Our proof based on the perspective was proposed by Effros in [153].

Theorem B.41. *The quantum relative entropy* $\mathscr{D}$ *is jointly convex on pairs of positive definite matrices.*

Proof. Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n \times n}$ be positive definite matrices. We associate to these matrices operators acting on $\mathbb{C}^{n \times n}$ endowed with the inner product structure (A.14) induced by the Frobenius norm, $\langle \mathbf{A}, \mathbf{B} \rangle_F = \operatorname{tr} (\mathbf{A}\mathbf{B}^*)$. We set

$$\mathbf{L}_\mathbf{A} \mathbf{X} := \mathbf{A}\mathbf{X}, \quad \mathbf{R}_\mathbf{B} \mathbf{X} := \mathbf{X}\mathbf{B}, \quad \mathbf{X} \in \mathbb{C}^{n \times n} \, .$$

By associativity of matrix multiplication the operators $\mathbf{L}_\mathbf{A}$ and $\mathbf{R}_\mathbf{B}$ commute, and by positivity of $\mathbf{A}, \mathbf{B}$ they are positive. Indeed, $\langle \mathbf{L}_\mathbf{A}(\mathbf{X}), \mathbf{X} \rangle_F = \operatorname{tr} (\mathbf{A}\mathbf{X}\mathbf{X}^*) = \|\mathbf{A}^{1/2}\mathbf{X}\|_F^2 > 0$ for non-zero $\mathbf{X}$. The function $\phi(x) = x \ln x$, $x > 0$, is operator convex by Corollary B.34 and due to commutativity its perspective g is given by

$$\begin{aligned} g(\mathbf{L}_\mathbf{A}, \mathbf{R}_\mathbf{B}) &= \mathbf{R}_\mathbf{B} \phi(\mathbf{R}_\mathbf{B}^{-1} \mathbf{L}_\mathbf{A}) = \mathbf{R}_\mathbf{B} (\mathbf{R}_\mathbf{B}^{-1} \mathbf{L}_\mathbf{A}) \ln (\mathbf{R}_\mathbf{B}^{-1} \mathbf{L}_\mathbf{A}) \\ &= \mathbf{L}_\mathbf{A} (\ln \mathbf{L}_\mathbf{A} - \ln \mathbf{R}_\mathbf{B}) \, . \end{aligned}$$

By joint matrix convexity of the perspective (Theorem B.37), the scalar-valued function

$$h(\mathbf{A}, \mathbf{B}) := \langle g(\mathbf{L}_\mathbf{A}, \mathbf{R}_\mathbf{B}) \mathbf{Id}, \mathbf{Id} \rangle_F$$

is jointly convex in $(\mathbf{A}, \mathbf{B})$. Further, $f(\mathbf{L_A})(\mathbf{Id}) = f(\mathbf{A})$ and $f(\mathbf{R_B})(\mathbf{Id}) = f(\mathbf{B})$ for any continuous function f. Indeed, these relations are easily checked for monomials $f(x) = x^n$, and by the Weierstrass approximation theorem extend to any continuous f. Therefore, h takes the form

$$\begin{aligned} h(\mathbf{A}, \mathbf{B}) &= \langle g(\mathbf{L_A}, \mathbf{R_B})\mathbf{Id}, \mathbf{Id}\rangle_F = \operatorname{tr}\left(g(\mathbf{L_A}, \mathbf{R_B})\mathbf{Id}\right) \\ &= \operatorname{tr}\left(\mathbf{L_A}(\ln \mathbf{L_A} - \ln \mathbf{R_B})\mathbf{Id}\right) = \operatorname{tr}\left(\mathbf{A}(\ln \mathbf{A} - \ln \mathbf{B})\right) . \end{aligned}$$

Therefore,

$$\mathscr{D}(\mathbf{A}, \mathbf{B}) = h(\mathbf{A}, \mathbf{B}) - \operatorname{tr}(\mathbf{A} - \mathbf{B})$$

is jointly convex in $(\mathbf{A}, \mathbf{B})$. □

We are finally in the position to prove Lieb's concavity theorem.

Proof (of Theorem B.31). Setting $\mathbf{B} = \exp(\mathbf{H} + \ln \mathbf{X})$ in Corollary B.40 yields

$$\begin{aligned} \operatorname{tr} \exp(\mathbf{H} + \ln \mathbf{X}) &= \max_{\mathbf{A} \succ \mathbf{0}} \operatorname{tr}\left(\mathbf{A}(\mathbf{H} + \ln \mathbf{X}) - \mathbf{A} \ln \mathbf{A} + \mathbf{A}\right) \\ &= \max_{\mathbf{A} \succ \mathbf{0}} \left(\operatorname{tr}(\mathbf{AH}) + \operatorname{tr} \mathbf{X} - \mathscr{D}(\mathbf{A}, \mathbf{X})\right) . \end{aligned}$$

For each self-adjoint matrix $\mathbf{H}$ the term in the bracket is a jointly concave function in the self-adjoint matrices $\mathbf{X}$ and $\mathbf{A}$ by Theorem B.41. It follows from Theorem B.15 that partial maximization of a jointly concave function yields a concave function, so that $\mathbf{X} \mapsto \operatorname{tr} \exp(\mathbf{H} + \ln \mathbf{X})$ is concave on the set of positive definite matrices $\mathbf{X}$. □

C

Miscellanea

C.1 Fourier Analysis

This section recalls some simple facts from Fourier analysis. We cover the finite-dimensional analog of Shannon's sampling theorem as mentioned in Section 1.2 as well as basic facts on the Fourier matrix and the Fast Fourier Transform (FFT). More background on Fourier and harmonic analysis can be found in various books on the subject including [30, 172, 203, 233, 234, 345, 337, 391, 442].

Finite-Dimensional Sampling Theorem

We consider trigonometric polynomials of degree at most M, that is, functions of the form

$$f(t) = \sum_{k=-M}^{M} c_k e^{2\pi i k t} , \quad t \in [0,1] . \tag{C.1}$$

The numbers c_k are called Fourier coefficients and they are given in terms of f by

$$c_k = \int_0^1 f(t) e^{-2\pi i k t} dt .$$

The Dirichlet kernel is defined as

$$D_M(t) := \sum_{k=-M}^{M} e^{2\pi i k t} = \begin{cases} \dfrac{\sin(\pi(2M+1)t)}{\sin(\pi t)} & \text{if } t \neq 0 , \\ 2M+1 & \text{if } t = 0 . \end{cases}$$

The expression for $t \neq 0$ follows from the geometric sum identity and simplifying. The finite-dimensional version of Shannon's sampling theorem reads as follows.

Theorem C.1. *Let f be a trigonometric polynomial of degree at most M. Then, for all $t \in [0,1]$,*

$$f(t) = \frac{1}{2M+1} \sum_{j=0}^{2M} f\left(\frac{j}{2M+1}\right) D_M\left(t - \frac{j}{2M+1}\right), \quad t \in [0,1]\,. \tag{C.2}$$

Proof. Let $f(t) = \sum_{k=-M}^{M} c_k e^{2\pi i k t}$. We evaluate the expression on the right hand side of (C.2) as

$$\begin{aligned}
&\sum_{j=0}^{2M} f\left(\frac{j}{2M+1}\right) D_M\left(t - \frac{j}{2M+1}\right) \\
&= \sum_{j=0}^{2M} \sum_{k=-M}^{M} c_k e^{2\pi i k j/(2M+1)} \sum_{\ell=-M}^{M} e^{2\pi i \ell (t - j/(2M+1))} \\
&= \sum_{k=-M}^{M} c_k \sum_{\ell=-M}^{M} \sum_{j=0}^{2M} e^{2\pi i (k-\ell) j/(2M+1)} e^{2\pi i \ell t}\,.
\end{aligned}$$

The identity $\sum_{j=0}^{2M} e^{2\pi i (k-\ell) j/(2M+1)} = (2M+1)\delta_{k,\ell}$ completes the proof. □

The Fast Fourier Transform

The Fourier matrix $\mathbf{F} \in \mathbb{C}^{N \times N}$ has entries

$$F_{\ell,k} = \frac{1}{\sqrt{N}} e^{2\pi i (\ell-1)(k-1)/N}, \quad \ell, k \in [N]\,. \tag{C.3}$$

The application of $\mathbf{F}$ to a vector $\mathbf{x} \in \mathbb{C}^N$ is called the Fourier transform of $\mathbf{x}$ and denoted by

$$\widehat{\mathbf{x}} = \mathbf{F}\mathbf{x}\,.$$

Intuitively, the coefficient $\widehat{x}_j$ reflects the frequency content of $\mathbf{x}$ corresponding to the monomials $j \mapsto e^{2\pi i (j-1)(k-1)/N}$. The Fourier transform arises, for instance, when evaluating a trigonometric polynomial of the form (C.1) at the points $j/(2M+1)$, $j = -M, \ldots, M$.

The Fourier matrix is unitary, i.e., $\mathbf{F}^*\mathbf{F} = \mathbf{Id}$ so that $\mathbf{F}^{-1} = \mathbf{F}^*$, see (12.1). This reflects the fact that its columns form an orthonormal basis of $\mathbb{C}^N$.

A naive implementation of the Fourier transform requires $\mathcal{O}(N^2)$ operations. The Fast Fourier Transform (FFT) is an algorithm that evaluates the Fourier transform much quicker, namely in $\mathcal{O}(N \ln N)$ operations. It is basically this fact which makes the FFT one of the most widely used algorithms and many devices of modern technology would not work without it.

Let us give the main idea of the FFT algorithm. Assume that N is even. Then the Fourier transform of $\mathbf{x} \in \mathbb{C}^N$ has entries

$$\begin{aligned}\widehat{x}_j &= \frac{1}{\sqrt{N}} \sum_{k=1}^{N} x_k e^{2\pi i(j-1)(k-1)/N} \\ &= \frac{1}{\sqrt{N}} \left(\sum_{\ell=1}^{N/2} x_{2\ell} e^{2\pi i(j-1)(2\ell-1)/N} + \sum_{\ell=1}^{N/2} x_{2\ell-1} e^{2\pi i(j-1)(2\ell-2)/N} \right) \\ &= \frac{1}{\sqrt{N}} \left(e^{2\pi i(j-1)/N} \sum_{\ell=1}^{N/2} x_{2\ell} e^{2\pi i(j-1)(\ell-1)/(N/2)} \right. \\ &\qquad \left. + \sum_{\ell=1}^{N/2} x_{2\ell-1} e^{2\pi i(j-1)(\ell-1)/(N/2)} \right) .\end{aligned}$$

We have basically reduced the evaluation of $\widehat{\mathbf{x}} \in \mathbb{C}^N$ to the evaluation of two Fourier transforms in dimension $N/2$, namely to the one of $(x_{2\ell})_{\ell=1}^{N/2}$ and of $(x_{2\ell-1})_{\ell=1}^{N/2}$. If $N = 2^n$ then in this way we can recursively reduce the evaluation of the Fourier transform to the ones of half dimension until we reach the dimension 2. This requires n recursion steps and altogether $\mathcal{O}(n2^n) = \mathcal{O}(N \log N)$ algebraic operations. The resulting algorithm is named after Cooley and Tukey [105], and often simply called the Fast Fourier Transform. For other composite numbers $N = pq$ similar reduction steps can be made. We refer to [431, 443] for details.

C.2 Covering Numbers

Let T be a subset of a metric space (X, d). For $t > 0$ the covering number $\mathcal{N}(T, d, t)$ is defined as the smallest integer $\mathcal{N}$ such that T can be covered with balls $B(\mathbf{x}_\ell, t) = \{\mathbf{x} \in X, d(\mathbf{x}, \mathbf{x}_\ell) \leq t\}$, $\mathbf{x}_\ell \in T$, $\ell = 1, \ldots, \mathcal{N}$, that is

$$T \subset \bigcup_{\ell=1}^{\mathcal{N}} B(\mathbf{x}_\ell, t) .$$

The set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mathcal{N}\}$ is then called a t-covering. (Note that some authors only require $\mathbf{x}_\ell \in X$, so that the points are not necessarily elements of T.)

The packing number $\mathcal{P}(T, d, t)$, for $t > 0$, is defined as the maximal integer $\mathcal{P}$ such that there are points $\mathbf{x}_\ell \in T$, $\ell = 1, \ldots, \mathcal{P}$ which are t-separated, that is $d(\mathbf{x}_\ell, \mathbf{x}_k) > t$ for all $0\ k, \ell = 1, \ldots, \mathcal{P}$, $k \neq \ell$.

If $X = \mathbb{R}^n$ is a vector space and the metric is induced by a norm, $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$, we also write $\mathcal{N}(T, \|\cdot\|, t)$ and $\mathcal{P}(T, \|\cdot\|, t)$.

Let us first state some obvious properties of the covering numbers. The packing numbers satisfy precisely the same properties. For arbitrary sets $S, T \subset X$,

$$\mathcal{N}(S \cup T, d, t) \le \mathcal{N}(S, d, t) + \mathcal{N}(T, d, t) \; . \tag{C.4}$$

For some $\alpha > 0$ it holds

$$\mathcal{N}(T, \alpha d, t) = \mathcal{N}(T, d, t/\alpha) \; . \tag{C.5}$$

If $X = \mathbb{R}^n$ and d is induced by a norm $\|\cdot\|$ then furthermore

$$\mathcal{N}(\alpha T, d, t) = \mathcal{N}(T, d, \alpha^{-1} t) \; . \tag{C.6}$$

Moreover, if d' is another metric on X that satisfies $d'(\mathbf{x}, \mathbf{y}) \le d(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in T$ then

$$\mathcal{N}(T, d', t) \le \mathcal{N}(T, d, t) \; . \tag{C.7}$$

There is the following simple relation between covering and packing numbers.

Lemma C.2. *Let T be a subset of a metric space (X, d) and $t > 0$. Then*

$$\mathcal{P}(T, d, 2t) \le \mathcal{N}(T, d, t) \le \mathcal{P}(T, d, t).$$

Proof. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mathcal{P}\}$ be a $2t$ separated set and $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_\mathcal{N}\}$ be a t-covering. Then we can assign to each point $\mathbf{x}'_\ell$ a point $\mathbf{x}_j$ with $d(\mathbf{x}'_\ell, \mathbf{x}_j) \le t$. This is assignment is unique since the points $\mathbf{x}_j$ are $2t$-separated. Indeed, the assumption that two points $\mathbf{x}_j, \mathbf{x}_k$, $j \ne k$, can be assigned to a point $\mathbf{x}'_\ell$ would lead to a contradiction by the triangle inequality: $d(\mathbf{x}_j, \mathbf{x}_k) \le d(\mathbf{x}_j, \mathbf{x}'_\ell) + d(\mathbf{x}'_\ell, \mathbf{x}_k) \le 2t$. It follows that $\mathcal{P} \le \mathcal{N}$.

Now let $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mathcal{N}\}$ be a maximal t-packing. Then it is also a t-covering. Indeed, if there were a point $\mathbf{x}$, which is not covered by a ball $B(\mathbf{x}_\ell, t)$, $\ell = 1, \ldots, \mathcal{N}$, then $d(\mathbf{x}, \mathbf{x}_\ell) > t$ for all $\ell \in [\mathcal{N}]$. This means that we could add $\mathbf{x}$ to the t-packing. But this would be a contradiction to the maximality. □

The following proposition estimates the packing number of a norm-sphere in a finite-dimensional space.

Proposition C.3. *Let $\|\cdot\|$ be some norm on $\mathbb{R}^n$ and let U be a subset of the unit ball $B = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| \le 1\}$. Then the packing and covering numbers satisfy, for $t > 0$,*

$$\mathcal{N}(U, \|\cdot\|, t) \le \mathcal{P}(U, \|\cdot\|, t) \le \left(1 + \frac{2}{t}\right)^n \; . \tag{C.8}$$

Proof. Lemma C.2 shows the first inequality. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mathcal{P}\} \subset U$ be a maximal t-packing of U. Then the balls $B(\mathbf{x}_\ell, t/2)$ do not intersect and they are contained in the scaled unit ball $(1 + t/2)B$. By comparing volumes (that is, Lebesgue measures) of the involved balls we get

$$\operatorname{vol}\left(\bigcup_{\ell=1}^{\mathcal{P}} B(\mathbf{x}_\ell, t/2)\right) = \mathcal{P} \operatorname{vol}\left((t/2)B\right) \le \operatorname{vol}\left((1 + t)B\right) \; .$$

On $\mathbb{R}^n$ the volume satisfies $\operatorname{vol}(tB) = t^n \operatorname{vol}(B)$, hence, $\mathcal{P}(t/2)^n \operatorname{vol}(B) \le (1 + t/2)^n \operatorname{vol}(B)$ or $\mathcal{P} \le (1 + 2/t)^n$. □

C.3 The Gamma Function and Stirling's Formula

The Gamma function is defined for $x > 0$ via

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \; . \tag{C.9}$$

It interpolates the factorial function in the sense that, for positive integers n,

$$\Gamma(n) = (n-1)! \; . \tag{C.10}$$

It follows from integration by parts that the Gamma function satisfies the functional equation

$$\Gamma(x+1) = x\Gamma(x) \; , \quad x > 0 \; . \tag{C.11}$$

Its value at the point $1/2$ is given by $\Gamma(1/2) = \sqrt{\pi}$.

Stirling's formula states that

$$\Gamma(x) = \sqrt{2\pi} x^{x-1/2} e^{-x} \exp\left(\frac{\theta(x)}{12x}\right) \tag{C.12}$$

for positive x and $0 \le \theta(x) \le 1$. Using (C.10) and applying the formula (C.12) shows that the factorial satisfies

$$n! = \sqrt{2\pi n}\, n^n e^{-n} e^{R_n} \; . \tag{C.13}$$

Wallis' inequality states that, for each integer $n \ge 1$,

$$\frac{2^{2n}}{\sqrt{\pi(n+1/2)}} \le \binom{2n}{n} \le \frac{2^{2n}}{\sqrt{\pi n}} \; .$$

A simple proof consists in determining the quantities $I_n := \int_0^{\pi/2} \sin^n(x) dx$ inductively, using integration by parts and the values $I_0 = \pi/2$ and $I_1 = 1$. Wallis' inequality is then a consequence of $I_{2n+1} \le I_{2n} \le I_{2n-1}$.

We also need the following technical lemma about the quantities $\sqrt{2}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma(m/2)}$, which asymptotically behave like $\sqrt{m}$ as $m \to \infty$.

Lemma C.4. *For $m, s \in \mathbb{N}$ with $m > s$ it holds*

$$\sqrt{2}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma(m/2)} - \sqrt{2}\frac{\Gamma\left(\frac{s+1}{2}\right)}{\Gamma(s/2)} \ge \sqrt{m} - \sqrt{s} \; .$$

Proof. It is sufficient to show that

$$d_m := \sqrt{m} - \sqrt{2}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma(m/2)} = \sqrt{m} - \frac{m}{\sqrt{2}}\frac{\Gamma((m+1)/2)}{\Gamma((m+2)/2)}$$

decreases with m. Set

$$c_m := \frac{1}{\sqrt{2}} \frac{\Gamma\left((m+1)/2\right)}{\Gamma\left((m+2)/2\right)} \, .$$

From the functional equation $z\Gamma(z) = \Gamma(z+1)$ it follows that

$$c_{m+2} = \frac{m+1}{m+2} c_m$$

and

$$c_{m+1} c_m = \frac{1}{2} \frac{\Gamma((m+1)/2)}{\Gamma((m+3)/2)} = \frac{1}{m+1}. \tag{C.14}$$

Introduce $I_m := \int_0^{\pi/2} \sin^m(\theta) d\theta$. We claim that $c_m = \sqrt{\frac{2}{\pi}} I_m$. Indeed, $c_0 = \sqrt{\frac{\pi}{2}} I_0$, $c_1 = \sqrt{\frac{2}{\pi}} I_1$ and integration by parts yields

$$\begin{aligned} I_{m+2} &= \int_0^{\pi/2} \sin^{n+2}(\theta) d\theta = (m+1) \int_0^{\pi/2} \cos^2(\theta) \sin^m(\theta) d\theta \\ &= (m+1) \int_0^{\pi/2} (1 - \sin^2(\theta)) \sin^m(\theta) d\theta = (m+1) I_m + (m+1) I_{m+2} \, . \end{aligned}$$

Therefore,

$$I_{m+2} = \frac{m+1}{m+2} I_m \tag{C.15}$$

and I_m satisfies the same recursion as c_m, so that $c_m = \sqrt{\frac{2}{\pi}} I_m$. The Cauchy-Schwarz inequality gives

$$I_m = \int_0^{\pi/2} \sqrt{\sin^{m-1}(\theta)} \sqrt{\sin^{m+1}(\theta)} d\theta \leq \sqrt{I_{m-1}} \sqrt{I_{m+1}} = \sqrt{\frac{m}{m+1}} I_{m-1} \, ,$$

where we also used (C.15). Therefore,

$$c_m \leq \sqrt{\frac{m}{m+1}} c_{m-1} \, .$$

Together with (C.14) it follows that

$$\sqrt{\frac{m+1}{m}} c_m^2 \leq c_m c_{m-1} = \frac{1}{m} \leq \sqrt{\frac{m}{m+1}} c_{m-1}^2 \, ,$$

and therefore,

$$\sqrt{\frac{m+2}{m+1}} \frac{1}{m+1} \leq c_m^2 \leq \sqrt{\frac{m}{m+1}} \frac{1}{m} \, .$$

Recall that we would like to show that $d_m = \sqrt{m} - m c_m$ decreases with m, i.e., $d_{m+1} \leq d_m$ for all m, or that,

$$(m+1)c_{m+1} - mc_m \geq \sqrt{m+1} - \sqrt{m}\ .$$

Multiplying by c_m and using (C.14) this is equivalent to

$$1 - mc_m^2 \geq (\sqrt{m+1} - \sqrt{m})c_m\ ,$$

or to $p_m(c_m) \leq 0$, where $p_m(x) := mx^2 + (\sqrt{m+1} - \sqrt{m})x - 1$. Since

$$c_m \leq b_m := \left(\frac{m}{m+1}\right)^{1/4} \frac{1}{\sqrt{m}}\ ,$$

it is enough to show that $p_m(b_m) \leq 0$. Setting $\alpha := \left(\frac{m}{m+1}\right)^{1/4} < 1$, we have

$$\begin{aligned} p_m(b_m) &= m\sqrt{\frac{m}{m+1}}\frac{1}{m} + (\sqrt{m+1} - \sqrt{m})\left(\frac{m}{m+1}\right)^{1/4}\frac{1}{\sqrt{m}} - 1 \\ &= \alpha^2 + \left(\frac{1}{\alpha^2} - 1\right)\alpha - 1 = \alpha^{-2}(\alpha^4 - \alpha^3 + \alpha - 1) \\ &= \alpha^{-2}(\alpha^3 + 1)(\alpha - 1) < 0\ . \end{aligned}$$

This concludes the proof. □

C.4 The Multinomial Theorem

The multinomial theorem is concerned with the expansion of a power of a sum. It states that, for $n \in \mathbb{N}$,

$$\Big(\sum_{\ell=1}^{m} x_\ell\Big)^n = \sum_{k_1+k_2+\ldots+k_m=n} \frac{n!}{k_1!k_2!\cdots k_m!} \prod_{j=1}^{m} x_j^{k_j}\ .$$

The sum is taken over all possible m-tuples of nonnegative integers $k_1, \ldots, k_m$ that sum up to n. This formula can be proved, for instance, with the binomial theorem and induction on n.

C.5 Some Elementary Estimates

Lemma C.5. *For integers $n \geq k \geq 0$,*

$$\Big(\frac{n}{k}\Big)^k \leq \binom{n}{k} \leq \Big(\frac{en}{k}\Big)^k\ .$$

Proof. For the upper bound, we use

$$e^k = \sum_{\ell=0}^{\infty} \frac{k^\ell}{\ell!} \geq \frac{k^k}{k!}$$

to derive the inequality

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \le \frac{n^k}{k!} = \frac{k^k}{k!}\frac{n^k}{k^k} \le e^k \frac{n^k}{k^k}.$$

As for the lower bound, we write

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \prod_{\ell=1}^{k} \frac{n-k+\ell}{\ell} \ge \Big(\frac{n}{k}\Big)^k,$$

having used that $(n-k+\ell)/\ell = (n-k)/\ell + 1$ decreases with $\ell \ge 1$. □

Lemma C.6. *Given integers $N, m, s \ge 1$ and prescribed constants $c, d > 0$, if $m \ge cs\ln(dN/s)$ and $m \ge s$, then $m \ge cs\ln(dN/m)$. As a partial converse, if $m \ge c's\ln(c'N/m)$ with $c' = c(1+d/e)$, then $m \ge cs\ln(dN/s)$. Moreover, if $m \ge cs\ln(dN/m)$, then $m \ge c''s\ln(dN/s)$ with $c'' = ec/(e+c)$ or better $c'' = c/(\ln(ec))$ provided $c, d \ge e$.*

Proof. The first statement simply follows from $m \ge s$. For the second and third statements, let us assume that $m \ge \gamma s\ln(\delta N/m)$ for some $\gamma, \delta > 0$. For any $\delta' > 0$, we then have,

$$m \ge \gamma s \ln\Big(\frac{\delta' N}{s}\Big) + \gamma s \ln\Big(\frac{\delta s}{\delta' m}\Big) = \gamma s \ln\Big(\frac{\delta' N}{s}\Big) + \frac{\gamma\delta'}{\delta} m \frac{\delta s}{\delta' m} \ln\Big(\frac{\delta s}{\delta' m}\Big).$$

We notice that the function $f(x) := x\ln(x)$ is decreasing on $(0, 1/e)$ and increasing on $(1/e, +\infty)$, with a minimum value of $-1/e$, to derive

$$m \ge \gamma s \ln\Big(\frac{\delta' N}{s}\Big) - \frac{\gamma\delta'}{e\delta} m, \quad \text{i.e.,} \quad \Big(1 + \frac{\gamma\delta'}{e\delta}\Big) m \ge \gamma s \ln\Big(\frac{\delta' N}{s}\Big).$$

The second statement is obtained by taking (among other possible choices) $\gamma = \delta = c(1+d/e)$ and $\delta' = d$, while the first part of the third statement is obtained by taking $\gamma = c$ and $\delta = \delta' = d$. As for the second part of this statement, where $c, d \ge e$, it follows from $s/m \le 1/(c\ln(dN/s)) \le 1/c$, hence $f(s/m) \ge f(1/c) = -\ln(c)/c$. The same choice of γ, δ, δ' yields

$$m \ge cs \ln\Big(\frac{dN}{s}\Big) - \ln(c) m, \quad \text{i.e.,} \quad (1 + \ln(c)) m \ge cs \ln\Big(\frac{dN}{s}\Big),$$

which is a rewriting of the desired conclusion. □

In Chapter 12 we also need the following similar statement.

Lemma C.7. *Given prescribed $c > 0, d_1, d_2 \ge 1$ and $s, m \ge 1$, if $m \ge c'\tilde{c}s\ln^3(\tilde{c}s)$ with $\tilde{c} = \max\{c(1 + d_1(1+e^{-1})^2/e), ed_2\}$ and $c' = (1+2/e)^3$ then $m/\ln(d_1 m) \ge cs\ln^2(d_2 s)$.*

Proof. Let $\tilde{c} > 0$, $d_2' \geq 1$ to be determined later and suppose that $m \geq \tilde{c}s\ln^3(d_2's)$. Then

$$\begin{aligned} m &\geq \tilde{c}\ln^2(d_2's)\big(\ln(d_1m) + \ln\big(\frac{d_2's}{d_1m}\big)\big) \\ &= \tilde{c}s\ln^2(d_2's)\ln(d_1m) + \tilde{c}\ln^2(d_2's)\frac{d_2's}{d_1m}\ln\big(\frac{d_2's}{d_1m}\big)\frac{d_1m}{d_2'} \\ &\geq \tilde{c}s\ln^2(d_2's)\ln(d_1m) - \frac{\tilde{c}d_1}{d_2'e}\ln^2(d_2's)m\ , \end{aligned}$$

where we used again that $-e^{-1}$ is the minimum of the function $x \mapsto x\ln(x)$. Now we choose $d_2' = \tilde{c}\ln^2(\tilde{c}s)$. Then

$$\frac{\ln^2(d_2's)}{d_2'} = \frac{\ln^2(\tilde{c}s\ln^2(\tilde{c}s))}{\tilde{c}\ln^2(\tilde{c}s)} = \frac{1}{\tilde{c}}\big(1 + \frac{\ln(\ln(\tilde{c}s))}{\ln(\tilde{c}s)}\big)^2 \leq \frac{1}{\tilde{c}}\big(1+e^{-1}\big)^2\ ,$$

because e^{-1} is the maximum of the function $x \mapsto \ln(x)/x$. We obtain $m \geq \tilde{c}s\ln^2(d_2's)\ln(d_1m) - \frac{d_1}{e}(1+e^{-1})^2m$, or

$$\begin{aligned} m &\geq \frac{\tilde{c}}{1 + d_1(1+e^{-1})^2/e}s\ln^2(\tilde{c}s\ln^2(\tilde{c}s))\ln(d_1m) \\ &\geq \frac{\tilde{c}}{1 + d_1(1+e^{-1})^2/e}s\ln^2(\tilde{c}\ln^2(\tilde{c})s)\ln(d_1m) \\ &\geq \frac{\tilde{c}}{1 + d_1(1+e^{-1})^2/e}s\ln^2(d_2s)\ln(d_1m)\big(1 + \frac{\ln\big(\frac{\tilde{c}}{d_2}\ln^2(\tilde{c}s)\big)}{\ln^2(d_2s)}\big)^2 \\ &\geq \frac{\tilde{c}}{1 + d_1(1+e^{-1})^2/e}s\ln^2(d_2s)\ln(d_1m)\ , \end{aligned}$$

where the last inequality is valid if $\tilde{c} \geq ed_2 \geq e$. Moreover, note that

$$\begin{aligned} \ln^3(d_2's) = \ln^3(\tilde{c}\ln^2(\tilde{c}s)) &= (\ln(\tilde{c}s) + 2\ln(\ln(\tilde{c}s)))^3 \leq (\ln(\tilde{c}s) + 2e^{-1}\ln(\tilde{c}s))^3 \\ &= (1+2/e)^3\ln^3(\tilde{c}s), \end{aligned}$$

where it is used once more that e^{-1} is the maximum of the function $x \mapsto \ln(x)/x$. In particular, if $m \geq (1+2/e)^3\tilde{c}\ln^3(\tilde{c}s)$ with

$$\tilde{c} = \max\{c(1 + d_1(1+e^{-1})^2/e), ed_2\}$$

then $m/\ln(d_1m) \geq cs\ln^2(d_2s)$ as claimed. □

C.6 Estimates of Some Integrals

Next we provide some useful estimates of certain integrals. The first two lemmas are related to estimating the tail of a Gaussian random variable from above and below.

Lemma C.8. *For $u > 0$ it holds*

$$\int_u^\infty e^{-t^2/2} dt \leq \min\left\{\sqrt{\frac{\pi}{2}}, \frac{1}{u}\right\} \exp(-u^2/2).$$

Proof. A change of variables yields

$$\int_u^\infty e^{-t^2/2} dt = \int_0^\infty e^{-\frac{(t+u)^2}{2}} dt = e^{-u^2/2} \int_0^\infty e^{-tu} e^{-t^2/2} dt. \tag{C.16}$$

On the one hand, using that $e^{-tu} \leq 1$ for $t, u \geq 0$, we get

$$\int_u^\infty e^{-t^2/2} dt \leq e^{-u^2/2} \int_0^\infty e^{-t^2/2} dt = \sqrt{\frac{\pi}{2}} e^{-u^2/2}.$$

On the other hand, using that $e^{-t^2} \leq 1$ for $t \geq 0$ yields

$$\int_u^\infty e^{-t^2/2} dt \leq e^{-u^2/2} \int_0^\infty e^{-tu} dt = \frac{1}{u} e^{-u^2/2}. \tag{C.17}$$

This shows the desired estimate. □

Lemma C.9. *For $u > 0$ it holds*

$$\int_u^\infty e^{-t^2/2} dt \geq \max\left\{\frac{1}{u} - \frac{1}{u^3}, \sqrt{\frac{\pi}{2}} - u\right\} e^{-u^2/2}.$$

Proof. We use (C.16) together with $\exp(-t^2/2) \geq 1 - t^2/2$ to obtain

$$\int_u^\infty e^{-t^2/2} dt \geq e^{-u^2/2} \int_0^\infty \big(1 - \frac{t^2}{2}\big) e^{-tu} dt = e^{-u^2/2} \big(\frac{1}{u} - \frac{1}{u^3}\big).$$

Using instead $\exp(-ut) \geq 1 - ut$ in (C.16) yields

$$\int_u^\infty e^{-t^2/2} dt \geq e^{-u^2/2} \int_0^\infty e^{-t^2/2} (1 - ut) dt = e^{-u^2/2} \big(\sqrt{\frac{\pi}{2}} - u\big).$$

This completes the proof. □

Lemma C.10. *For $\alpha > 0$ it holds*

$$\int_0^\alpha \sqrt{\ln(1 + t^{-1})} dt \leq \alpha \sqrt{\ln(e(1 + \alpha^{-1}))}. \tag{C.18}$$

Proof. First apply the Cauchy-Schwarz inequality to obtain

$$\int_0^\alpha \sqrt{\ln(1 + t^{-1})} dt \leq \sqrt{\int_0^\alpha 1 dt \int_0^\alpha \ln(1 + t^{-1}) dt}.$$

A change of variables and integration by parts yields

$$\int_0^\alpha \ln(1+t^{-1})dt = \int_{\alpha^{-1}}^\infty u^{-2}\ln(1+u)du$$
$$= -u^{-1}\ln(1+u)\big|_{\alpha^{-1}}^\infty + \int_{\alpha^{-1}}^\infty u^{-1}\frac{1}{1+u}du \le \alpha\ln(1+\alpha^{-1}) + \int_{\alpha^{-1}}^\infty \frac{1}{u^2}du$$
$$= \alpha\ln(1+\alpha^{-1}) + \alpha.$$

Combining the above estimates concludes the proof. □

C.7 Hahn-Banach Theorems

The *Hahn–Banach extension theorem* says that if λ is a continuous linear functional defined on a subspace Y of a real or complex normed space X, then there exists a continuous linear functional $\widetilde{\lambda}$ defined on X such that $\widetilde{\lambda}(\mathbf{y}) = \lambda(\mathbf{y})$ for all $\mathbf{y} \in Y$ and $\|\widetilde{\lambda}\|_{X^*} = \|\lambda\|_{Y^*}$, where X^*, Y^* denote the dual spaces of X and Y. In fact, it says more generally that if a linear functional λ defined on a subspace Y of a vector space X satisfies $|\lambda(\mathbf{y})| \le p(\mathbf{y})$ for all $\mathbf{y} \in Y$, where p is a seminorm on X, then there exists a linear functional $\widetilde{\lambda}$ defined on X such that $\widetilde{\lambda}(\mathbf{y}) = \lambda(\mathbf{y})$ for all $\mathbf{y} \in Y$ and $|\lambda(\mathbf{x})| \le p(\mathbf{x})$ for all $\mathbf{x} \in X$.

The *Hahn–Banach separation theorem* says that if C and D are two disjoint nonempty convex subset of a normed space X and if C is open, then there exists a linear functional λ defined on X and a real number t such that $\mathrm{Re}(\lambda(\mathbf{x})) < t$ for all $\mathbf{x} \in C$ and $\mathrm{Re}(\lambda(\mathbf{x})) \ge t$ for all $\mathbf{x} \in D$.

C.8 Smoothing Lipschitz functions

The proof of the concentration of measure results, Theorems 8.35 and 8.38 require to approximate a Lipschitz function by a smooth Lipschitz function. The following result establishes this rigorously.

Theorem C.11. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant $L = 1$, see (8.72). For $\varepsilon > 0$ and $x \in \mathbb{R}^n$ denote by $B_\varepsilon(x) = \{y \in \mathbb{R}^n : \|y - x\|_2 \le \varepsilon\}$ the ball of radius ϵ around x, and $|B_\varepsilon(x)|$ its volume. Define $g : \mathbb{R}^n \to \mathbb{R}$ to be the function*

$$g(x) = \frac{1}{|B_\epsilon(x)|}\int_{B_\varepsilon(x)} f(y)dy\ .$$

Then the function g is differentiable and $\|\nabla f(x)\|_2 \le L$ for all $x \in \mathbb{R}^n$ (so that also g is Lipschitz with constant $L = 1$). Furthermore, we have

$$|f(x) - g(x)| \le \frac{\varepsilon n}{n+1} \le \varepsilon \quad \textit{for all } x \in \mathbb{R}^n\ .$$

Proof. We start with the case $n = 1$. Then g is defined via

$$g(x) = \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} g(y)dy \, .$$

Therefore,

$$g'(x) = \frac{f(x+\varepsilon) - f(x-\epsilon)}{2\varepsilon},$$

and since f is 1-Lipschitz it follows that $|f'(x)| \leq 1$. Moreover,

$$|f(x) - g(x)| = \left| \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} f(x) - f(y)dy \right| \leq \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} |f(x) - f(y)|dy$$
$$\leq \int_{x-\varepsilon}^{x+\varepsilon} |x-y|dy = \frac{2}{2\varepsilon} \int_0^{\epsilon} tdt = \frac{\varepsilon^2}{2\varepsilon} = \varepsilon/2 \, .$$

Assume now $n > 1$. We choose a unit vector $u \in \mathbb{R}^n$ and some $x \in \mathbb{R}^n$ and show that the function $\psi(t) = g(x + tu)$ is differentiable with $|\psi'(t)| \leq 1$, which is equivalent to $|\nabla g \cdot u| \leq 1$. As u and x will be arbitrary, this establishes then that f is differentiable with $\|\nabla g(x)\|_2 \leq 1$ for all $x \in \mathbb{R}^n$. Without loss of generality we assume that $x = 0$ and that $u = (0, \ldots, 0, 1)$. Then the orthogonal complement $u^{\perp}$ can be identified with $\mathbb{R}^{n-1}$. Let $D_\varepsilon = \{w = (z, 0) : z \in \mathbb{R}^{n-1},\ \|z\|_2 \leq \varepsilon\}$. Then for any $(z, 0) \in D_\epsilon$ the intersection of the line through $(z, 0)$ in the direction of u with $B_\varepsilon(0)$ is an interval with endpoints of the form $(z, -a(z))$ and $(z, a(z))$, where $a(z) > 0$. Then it follows that

$$|B_\varepsilon(0)| = 2 \int_{D_\varepsilon} a(z)dz \, . \tag{C.19}$$

Now we estimate the derivative of ψ. For $\tau \in \mathbb{R}$, we have

$$\psi(\tau) = g\tau u) = \frac{1}{|B_\varepsilon(0)|} \int_{B_\varepsilon(\tau u)} f(y)dy = \frac{1}{|B_\varepsilon(0)|} \int_{D_\varepsilon} \int_{-a(z)+\tau} f(z,t)dtdz \, , s$$

and, hence,

$$\psi'(0) = \frac{1}{|B_\varepsilon(0)|} \int_{D_\varepsilon} f(z, a(z)) - f(z, -a(z))dz \, .$$

Since f is 1-Lipschitz we have $|f(z, a(z)) - f(z, -a(z))| \leq 2a(z)$ so that by (C.19)

$$|\psi'(0)| = |\nabla g \cdot u| \leq 1.$$

The approximation property follows similarly as in the case $n = 1$,

$$|f(0) - g(0)| \leq \int \frac{1}{|B_\varepsilon(0)|} \int_{B_\varepsilon(0)} |f(0) - f(y)|dy \leq \frac{1}{|B_\varepsilon(0)} \int_{B_\varepsilon(0)} \|y\|_2 dy$$
$$= \frac{|S^{n-1}|}{|B_\epsilon(0)|} \int_0^{\varepsilon} r^n dr = \frac{\varepsilon |S^{n-1}|}{(n+1)|B_1(0)|} \, , \tag{C.20}$$

where $|S^{n-1}|$ is the surface area of the sphere $S^{n-1} = \{x \in \mathbb{R}^n, \|x\|_2 = 1\}$. Denoting $s_n(r)$ the surface area of the sphere of radius r in $\mathbb{R}^{n-}$ and $v_n(r)$ the volume of the corresponding ball we have the relation

$$v_n(r) = \int_0^r s_n(r)dr \; . \tag{C.21}$$

The volume in $\mathbb{R}^n$ satisfies $v_n(r) = |B_1(0)|r^n$ for some constant C_n. Differentiating (C.21) shows that $s_n(r) = n|B_1(0)|r^{n-1}$, so that $|S^{n-1}| = s_n(1) = n|B_1(0)|$. Plugging into (C.20) completes the proof. □

C.9 Weak and Distributional Derivatives

The concept of weak derivative generalizes the classical derivative. Given a measurable function $f : \mathbb{R} \to \mathbb{R}$ we say that $v : \mathbb{R} \to \mathbb{R}$ is a weak derivative of f if for all infinitely differentiable functions $\phi : \mathbb{R} \to \mathbb{R}$ with compact support

$$\int_{-\infty}^{\infty} f(x)\phi'(x)dx = -\int_{-\infty}^{\infty} v(x)\phi(x)dx \; . \tag{C.22}$$

In this case we write $v = f' = \frac{d}{dx}f$. If f is continuously differentiable in the classical sense then it follows from integration by parts that the classical derivative f' is a weak derivative. If v and w are weak derivatives of f then they are equal almost everywhere, and in this sense, the weak derivative is unique. If a weak derivative exists then we say that f is weakly differentiable.

If the function f is defined only on a compact subinterval $[a, b] \subset \mathbb{R}$ then the integrals in (C.22) are only defined on $[a, b]$, and the functions ϕ are assumed to vanish on the boundary, $\phi(a) = \phi(b) = 0$.

This concept generalizes to the multivariate case and to higher derivatives derivatives in an obvious way. For a function $f : \mathbb{R}^n \to \mathbb{R}$, and a multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\alpha_j \in \mathbb{N}_0$, we set $|\alpha| = \sum_j \alpha_j$ and $D^\alpha f = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} f$. Then $v : \mathbb{R}^n \to \mathbb{R}$ is a weak-derivative of order α if

$$\int_{\mathbb{R}^d} f(\mathbf{x})D^\alpha\phi(\mathbf{x})d\mathbf{x} = (-1)^{|\alpha|}\int_{\mathbb{R}^d} v(\mathbf{x})\phi(\mathbf{x})d\mathbf{x}$$

for all infinitely differentiable functions $\phi : \mathbb{R}^n \to \mathbb{R}$ with compact support. We write $v = D^\alpha f$ in this case.

Distributional derivatives. The concept of weak derivative can be further generalized. We denote $\mathcal{D}$ the space of all infinitely differentiable functions with compact support. A distribution is a functional on $\mathcal{D}$, that is, a linear mapping from $\mathcal{D}$ into the scalars. A function f on $\mathbb{R}^m$, which is bounded on every compact subset of $\mathbb{R}^m$ (or at least locally integrable), induces a distribution via $f(\phi) = \int_{\mathbb{R}^m} f(\mathbf{x})\phi(\mathbf{x})d\mathbf{x}$ for $\phi \in \mathcal{D}$. The distributional derivative of a distribution f is defined via

$$\frac{\partial}{\partial x_j} f(\phi) = -f\left(\frac{\partial}{\partial x_j}\phi\right) , \quad \phi \in \mathcal{D} .$$

The distributional derivative exists always. If f can be identified with a function then it is the functional

$$\frac{\partial}{\partial x_j} f(\phi) = -\int_{\mathbb{R}^m} f(\mathbf{x}) \frac{\partial}{\partial x_j}\phi(\mathbf{x}) d\mathbf{x} .$$

If f possesses a weak derivative, then the distributional derivative can be identified with it by (C.22). If f is even differentiable then both distributional and weak derivative can be identified with the classical derivative.

We say that a distribution f is nonnegative if $f(\phi) \geq 0$ for all nonnegative functions $\phi \in \mathcal{D}$. In this sense, also nonnegativity of distributional derivatives is understood. For instance, we write $\frac{\partial f}{\partial x_j} \geq 0$ for a function f if, for all nonnegative $\phi \in \mathcal{D}$,

$$\int_{\mathbb{R}^m} f(\mathbf{x}) \frac{\partial}{\partial x_j}\phi(\mathbf{x}) d\mathbf{x} \geq 0 .$$

C.10 Differential Inequalities

The following lemma bounds the solution of a differential inequality by the solution of a corresponding differential equation.

Lemma C.12. *Let $f, g, h : [0, \infty) \to \mathbb{R}$ be continous functions with $g(x) \geq 0$ and $f(x) > 0$ for all $x \in [0, \infty)$. Assume that $L_0 : [0, \infty) \to \mathbb{R}$ is such that*

$$f(x)L_0'(x) - g(x)L_0(x) = h(x) , \quad x \in [0, \infty) , \tag{C.23}$$

while L satisfies the differential inequality

$$f(x)L'(x) - g(x)L(x) \leq h(x) , \quad x \in [0, \infty) . \tag{C.24}$$

If $L(0) = L_0(0)$ and $L'(0) = L_0'(0)$ then $L(x) \leq L_0(x)$ for all $x \in [0, \infty)$.

Proof. We first consider the differential inequality

$$f(x)L'(x) - g(x)L(x) \leq 0 . \tag{C.25}$$

Let $L : [0, \infty) \to \mathbb{R}$ be a continuously differentiable function satisfying (C.25) and $L(0) = L'(0) = 0$. We distinguish the following three cases that, for some sufficiently small $a > 0$,

- $L(x) = 0$ for all $x \in [0, a]$;
- $L(x) < 0$ for all $x \in [0, a]$;
- $L(x) > 0$ for all $x \in [0, a]$.

In the first case, either $L(x) = 0$ for all $x \in [0, \infty)$ or we end up in one of the other cases by translation.

In the second case, since $g(x)$ is positive and because of (C.25), $L'(x) \leq 0$ for $x \in (0, a)$ and thus L is non-increasing so that it always stays strictly negative. By continuity this implies that $L(x) \leq 0$ for all $x \in [0, \infty)$.

In the third case, we can rewrite (C.25) as

$$\frac{L'(x)}{L(x)} \leq \frac{g(x)}{f(x)} \, .$$

Integration on $[c, x_0]$ for $[c, x_0] \subset [0, a]$ shows that

$$\ln\left(L(x_0)/L(c)\right) = \int_c^{x_0} \frac{L'(x)}{L(x)} \leq \int_c^{x_0} \frac{g(x)}{f(x)} dx \, ,$$

or equivalently,

$$L(x_0) \leq L(c) \exp\left(\int_c^{x_0} \frac{f(x)}{g(x)} dx \right) \, .$$

Letting c tend to zero shows that $L(x_0) \leq 0$ by continuity of L, which is a contradiction to $L(x) > 0$ for all $x \in [0, a]$, so that the third case is not possible. Altogether, $L(x) \leq 0$ for all $x \in [0, \infty)$.

Now let L_0 be a solution to the differential equation (C.23), and let L satisfy (C.24) with $L(0) = L_0(0)$ and $L'(0) = L_0'(0)$. Then $L_1 := L - L_0$ satisfies (C.25) with $L_1(0) = 0$ and $L_1'(0) = 0$. By the above reasoning $L_1(x) \leq 0$ for all $x \in [0, \infty)$, which is equivalent to $L \leq L_0$. □

C.11 Sequences of Minimization Problems

In Chapter 15 we encounter sequences of minimization problems. A natural questions concerns the convergence of the minimizers. Coerciveness is an important concept in this context.

Definition C.13. *A function $F : K \to \mathbb{R}$, $K \subset \mathbb{R}^n$ is called coercive if the level sets $K_t := \{\mathbf{x} \in \mathbb{R}^n, F(\mathbf{x}) \leq t\}$ are compact.*

We have the following result on the convergence of minimizers.

Proposition C.14. *Let $F_k : K \to \mathbb{R}$, $K \subset \mathbb{R}^n$ be a decreasing sequence of continuous functions converging pointwise to $F : K \to \mathbb{R}$, that is, $F_{k+1}(\mathbf{x}) \leq F_k(\mathbf{x})$ and $\lim_{k\to\infty} F_k(\mathbf{x}) = F(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. Assume that F is continuous and coercive. Suppose that $\mathbf{x}_k$ minimizes F_k over K. Then the accumulation points of the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ are minimizers of F. If the minimizer of F is unique, then $\mathbf{x}_k$ converges to it as $k \to \infty$.*

Proof. Let $(\mathbf{z}_k)$ be a sequence converging to some $\mathbf{z} \in K$. From continuity and pointwise convergence of the F_k it follows easily that

$$F(\mathbf{z}) = \lim_{k\to\infty} F_k(\mathbf{z}_k) \geq \lim_{k\to\infty} F_k(\mathbf{x}_k) \;, \tag{C.26}$$

since $\mathbf{x}_k$ minimizes F_k. Since $\mathbf{z}$ was arbitrary, it follows that

$$\inf_{\mathbf{z}\in K} F(\mathbf{z}) \geq \lim_{k} F_k(\mathbf{x}_k) \;. \tag{C.27}$$

Since $F_k \geq F$, we have $\{\mathbf{x}, F_k(\mathbf{x}) \leq t\} \subset \{\mathbf{x}, F(\mathbf{x}) \leq t\}$ for all t and the latter is contained in a compact set by coerciveness. Since $\mathbf{x}_k$ minimizes F_k the sequence $\mathbf{x}_k$ is contained in a compact set. Hence, we can extract a subsequence $\mathbf{x}_{k_j}$ which converges to one of the accumulation points $\mathbf{x}'$ of $\mathbf{x}_k$. Then inequality (C.26) together with (C.27) yields

$$\inf_{\mathbf{x}\in X} F(\mathbf{x}) \leq F(\mathbf{x}') = \lim_{j} F_{k_j}(\mathbf{x}_{k_j}) = \lim_{k_j}(\min_{\mathbf{x}\in X} F_{k_j}(\mathbf{x})) \leq \inf_{\mathbf{z}\in K} F(\mathbf{z})$$

so that

$$\inf_{\mathbf{x}\in X} F(\mathbf{x}) = \lim_{j} F_{k_j}(\mathbf{x}_{k_j}) = F(\mathbf{x}') \;.$$

This means that $\mathbf{x}'$ minimizes F and we showed that all accumulation points of the sequence $(\mathbf{x}_k)$ are minimizers of F. Now if the minimizer of F is unique then with the same argument as above it follows that every subsequence of $\mathbf{x}_k$ contains another subsequence that converges to $\mathbf{x}'$. But then $\mathbf{x}_k$ itself must converge to $\mathbf{x}'$. □

This proof is in the spirit of the theory of Γ-convergence, see for instance [111].

Solutions

Hints for Chapter 2

2.1 For $0 < p < 1$, start by proving $(a+b)^p \leq a^p + b^p$, $a, b \geq 0$. Derive $\|\mathbf{x}^1 + \cdots + \mathbf{x}^k\|_p \leq k^{1/p-1}(\|\mathbf{x}^1\|_p + \cdots + \|\mathbf{x}^k\|_p)$ from $\|\mathbf{x}^1 + \cdots + \mathbf{x}^k\|_p^p \leq \|\mathbf{x}^1\|_p^p + \cdots + \|\mathbf{x}^k\|_p^p$ by applying Hölder's inequality to the vector $[\|\mathbf{x}^1\|_p^p, \ldots, \|\mathbf{x}^p\|_p^p]^\top$.

2.3 For the inequality $\|\mathbf{u}+\mathbf{v}\|_{1,\infty} \leq \|\mathbf{u}\|_{1,\infty} + \|\mathbf{v}\|_{1,\infty}$, adapt the proof of Proposition 2.7, and take $\mathbf{u} = [1, 0, \ldots, 0]^\top$, $\mathbf{v} = [0, 1, 0, \ldots, 0]^\top$ to observe that the inequality is sharp. For the inequality $\|\mathbf{u}+\mathbf{v}\|_{1,\infty} \geq \|\mathbf{u}\|_{1,\infty}$, say, notice that $\mathrm{card}(\{j : |u_j| \geq t\}) \leq \mathrm{card}(\{j : |u_j + v_j| \geq t\}) \leq \|\mathbf{u}+\mathbf{v}\|_{1,\infty}/t$ for all $t > 0$, and take $\mathbf{v} = 0$ to observe that the inequality is sharp.

2.4 With $\|\mathbf{x}\|_{p,\infty} = 1$, one has $|x_k^*| \leq 1/k^{1/p}$, and $\|\mathbf{x}\|_p^p = \sum_{j=1}^N |x_k^*|^p \leq \sum_{j=1}^N 1/k \leq 1 + \int_1^N dx/x = 1 + \ln(N)$.

2.6 Observe that $B_i^n(x_j) = \binom{n}{i}(1-x_j)u_j^i$, $u_j := x_j/(1-x_j)$, so that $[B_i^n(x_j)]_{i,j=0}^n = DVD'$ for two diagonal matrices D and D' and for the totally positive Vandermonde matrix $V = [u_j^i]_{i,j=0}^n$.

2.7 Use Cauchy–Binet formula.

2.9 The map $F : (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^s \times \mathbb{R}^s \mapsto f(\mathbf{u}, 0_{\mathbb{R}^{N-s}}) - f(0_{\mathbb{R}^s}, \mathbf{v}, 0_{\mathbb{R}^{N-2s}}) \in \mathbb{R}^m$ is continuous and antipodal from a space of dimension $2s$ into a space of dimension m. If $m < 2s$, then Borsuk–Ulam theorem gives a nonzero $(\mathbf{u}, \mathbf{v})$ such that $F(\mathbf{u}, \mathbf{v}) = 0$. The injectivity of f on sparse vectors implies $\mathbf{u} = \mathbf{v} = 0$, a contradiction.

Hints for Chapter 3

3.1 Let $(\mathbf{e}_1, \ldots, \mathbf{e}_N)$ denote the canonical basis of $\mathbb{K}^N$, and fix $j \in [N]$, suppose that for all $\mathbf{z} \in \mathbb{K}^N$ satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{e}_j$ we have $\|\mathbf{z}\|_q^q \geq \|\mathbf{e}_j\|_q^q = 1$,

considering a vector $\mathbf{v} \in \ker A \backslash \{0\}$ and a real number $t \neq 0$ with $|t| < 1/\|\mathbf{v}\|_\infty$, we obtain

$$1 \leq \|\mathbf{e}_j + t\mathbf{v}\|_q^q = |1 + t\,v_j|^q + \sum_{k=1,\,k\neq j}^{N} |t\,v_k|^q = (1 + t\,v_j)^q + t^q \sum_{k=1,\,k\neq j}^{N} |v_k|^q \underset{t\to 0}{\sim} 1 + q\,t\,v_j,$$

which implies $v_j = 0$, this is a contradiction since it holds for all $j \in [N]$.

3.3 Reproduce the argument of Theorem 3.1.

3.4 Apply Theorem 3.1.

3.7 Express the minimization problem as the classical least-square problem $\underset{\mathbf{z}\in\mathbb{C}^N}{\text{minimize}}\ \|B\mathbf{z} - \mathbf{u}\|_2$ for properly chosen $B \in \mathbb{C}^{2m\times N}$ and $\mathbf{u} \in \mathbb{C}^{2m}$.

3.8 [TO BE WRITTEN]

Hints for Chapter 4

4.2 One can take $A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$ and $D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

4.4 Observe that $\mathbf{A}$ has the real s-th order null space property if and only if

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\overline{S}}\|_1 \quad \text{and} \quad \|(\mathbf{v}_S + t\mathbf{w})\|_1 < \|(\mathbf{v} + t\mathbf{w})_{\overline{S}}\|_1$$

for all $t \in \mathbb{R}$ and all $S \subseteq [N]$ with $\operatorname{card}(S) = s$. Now observe that the piecewise linear function $t \in \mathbb{R} \mapsto \|(\mathbf{v}_S + t\mathbf{w})\|_1 - \|(\mathbf{v} + t\mathbf{w})_{\overline{S}}\|_1$ is negative if and only if it is negative at $\pm\infty$ and at each breakpoint $-v_i/w_i$, $i \in [N]$.

4.6 To obtain (ii) from (i), we take $\mathbf{v} = \mathbf{x} - \mathbf{z}$ and $S = \operatorname{supp}(\mathbf{x})$. To obtain (i) from (ii), we take $\mathbf{x} = \mathbf{v}_S$ and $\mathbf{z} = -\mathbf{v}_{\overline{S}}$. The converse of Proposition 4.13 then reads: if there is a constant $C > 1$ such that

$$\|\mathbf{z} - \mathbf{x}\|_1 \leq C\,[\|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + 2\,\sigma_s(\mathbf{x})_1]$$

for all vectors $\mathbf{x}, \mathbf{z} \in \mathbb{K}^N$ satisfying $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$, then the matrix $\mathbf{A}$ has the s-th order null space property with constant $\rho = (C-1)/(C+1)$.

4.7 Apply Lemma 4.14.

4.8 For the direct implication, suppose that $\mathbf{v}_S \geq 0$ for some $\mathbf{v} \in \ker \mathbf{A} \setminus \{0\}$ and $S \subseteq [N]$ with $\operatorname{card}(S) = s$, if S^- and S^+ are the index sets of negative and nonnegative components of $\mathbf{v}$, then $\|\mathbf{v}_{S^-}\|_1 < \|\mathbf{v}_{S^+}\|_1$ because the vectors $-\mathbf{v}_{S^-}$ are $\mathbf{v}_{S^+}$ are both nonnegative and have the same image and because $-\mathbf{v}_{S^-}$ is s-sparse. For the reverse implication, suppose that $\mathbf{x}$ is a vector

supported on S, $\mathrm{card}(S) \leq s$, and let $\mathbf{x}^\star$ be a solution of the optimization problem, if $\mathbf{v} = \mathbf{x}^\star - \mathbf{x} \in \ker \mathbf{A}$ is nonzero, then $\sum v_j > 0$ because $\mathbf{v}_{\overline{S}} \geq 0$, i.e. $\|\mathbf{v}_{S^+}\|_1 > \|\mathbf{v}_{S^-}\|_1$, but this is contradiction with Lemma 4.14.

4.9 For a vector $\mathbf{x} \geq 0$ supported on S and a vector $\mathbf{z} \geq 0$ such that $\mathbf{Az} = \mathbf{Ax}$, suppose that $\mathbf{v} := \mathbf{z} - \mathbf{x} \neq 0$, note that $\mathbf{v} \in \ker \mathbf{A}$ and that $\mathbf{v}_{\overline{S}} \geq 0$, so that $S^- \subseteq S$, where S^- and S^+ are defined as above, then $-\mathbf{v}_{S^-} \geq 0$ and $\mathbf{A}(-\mathbf{v}_{S^-}) = \mathbf{A}(\mathbf{v}_{S^+})$ imply $\|-\mathbf{v}_{S^-}\|_1 < \|\mathbf{v}_{S^+}\|_1$, i.e. $-\sum_{j\in S^-} v_j < \sum_{j\in S^+} v_j$, a contradiction with $\sum_{j=1}^N v_j = 0$.

4.10 Use $\|\mathbf{MAv}\|_2 \leq \|M\|_{2\to 2}\|\mathbf{Av}\|_2$ to prove that $\mathbf{A}$ satisfies the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau\|\mathbf{M}\|_{2\to 2} > 0$.

4.12 Choose a matrix with a null space spanned by the vector with s entries equal to $(1 + 1/s)^q$ and $(s + 1)$ entries equal to 1.

4.17 Let $\mathbf{M} = \mathbf{Z} - \mathbf{X} \in \ker \mathcal{A} \setminus \{\mathbf{0}\}$. (If $\mathbf{M} = \mathbf{0}$ then there is nothing to do.) Use Lemma A.20 to show that $\sum_{j=1}^r \sigma_j(\mathbf{X}) \leq \sum_{j=1}^r (\sigma_j(\mathbf{M}) + \sigma_j(\mathbf{Z}))$. Further, apply (A.27) in the same way as in the proof of Lemma A.20 to obtain $\sum_{j=r+1}^{\min\{n_1,n_2\}} \sigma_j(\mathbf{M}) \leq \sum_{j=r+1}^{\min\{n_1,n_2\}} (\sigma_j(\mathbf{X}) + \sigma_j(\mathbf{Z}))$. Use these inequalities to obtain the analog of Lemma 4.14, that is,

$$\sum_{j=r+1}^{\min\{n_1,n_2\}} \sigma_j(\mathbf{M}) \leq \|\mathbf{Z}\|_* - \|\mathbf{X}\|_* + \sum_{j=1}^{r} \sigma_j(\mathbf{M}) + 2 \sum_{j=r+1}^{\min\{n_1,n_2\}} \sigma_j(\mathbf{X}) \, .$$

Conclude the proof analogously to the one of Theorem 4.13.

4.18 Combine the proof of Exercise 4.17 with analog arguments as in Section 4.3.

4.19 (a) Let the singular value decomposition of $\mathbf{X}$ be given by $\sum_{\ell=1}^r \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^*$ where r is the rank of $\mathbf{X}$. Cyclicity of the trace and the Cauchy-Schwarz inequality yields, for $\mathbf{Y} \in \mathbb{C}^{n_1 \times n_2}$,

$$\begin{aligned} |\langle \mathbf{X}, \mathbf{Y}\rangle_F| &= \left| \mathrm{tr}\left(\sum_{\ell=1}^r \sigma_r \mathbf{v}_\ell \mathbf{u}_\ell^* \mathbf{Y}^* \right) \right| = |\sum_{\ell=1}^r \sigma_r \mathrm{tr}\,(\mathbf{u}_\ell^* \mathbf{Y}^* \mathbf{v}_\ell)| = |\sum_{\ell=1}^r \sigma_r \langle \mathbf{v}_\ell, \mathbf{Y}\mathbf{u}_\ell\rangle| \\ &\leq \sum_{\ell=1}^r \sigma_r \|\mathbf{v}_\ell\|_2 \|\mathbf{Y}\mathbf{u}_\ell\|_2 \leq \|\mathbf{Y}\|_{2\to 2} \sum_{\ell=1}^r \sigma_\ell = \|\mathbf{Y}\|_{2\to 2}\|\mathbf{X}\|_* \, . \end{aligned}$$

The matrix $\mathbf{Y} = \sum_{\ell=1}^r \mathbf{u}_\ell \mathbf{v}_\ell^*$ satisfies $\|\mathbf{Y}\|_{2\to 2} = 1$ and $\langle \mathbf{X}, \mathbf{Y}\rangle_F = \|\mathbf{X}\|_*$. (Different proofs can be found in [38] and [362].)

(b) Consider the singular value decompositions $\mathbf{X} = \sum_{\ell=1}^{r_1} \sigma_\ell^1 \mathbf{u}_\ell^1 (\mathbf{v}_\ell^1)^*$ and $\mathbf{Y} = \sum_{\ell=1}^{r_2} \sigma_\ell^2 \mathbf{u}_\ell^2 (\mathbf{v}_\ell^2)^*$. The relations $\mathbf{XY}^* = \mathbf{0}$ and $\mathbf{X}^*\mathbf{Y} = \mathbf{0}$ imply that the vectors $\mathbf{u}_1^1, \ldots, \mathbf{u}_{r_1}^1, \mathbf{u}_1^2, \ldots, \mathbf{u}_{r_2}^2$ as well as the vectors $\mathbf{v}_1^1, \ldots, \mathbf{v}_{r_1}^1, \mathbf{v}_1^2, \ldots, \mathbf{v}_{r_2}^2$ are orthonormal. This implies that we have the singular value decomposition

$$\mathbf{X} + \mathbf{Y} = \sum_{\ell=1}^{r_1} \sigma_\ell^1 \mathbf{u}_\ell^1 (\mathbf{v}_\ell^1)^* + \sum_{\ell=1}^{r_2} \sigma_\ell^2 \mathbf{u}_\ell^2 (\mathbf{v}_\ell^2)^*,$$

so that $\|\mathbf{X}+\mathbf{Y}\|_* = \sum_{\ell=1}^{r_1}\sigma_\ell^1 + \sum_{\ell=1}^{r_2}\sigma_\ell^2 = \|\mathbf{X}\|_* + \|\mathbf{Y}\|_*$.

(d) Let $\mathbf{Z} \neq \mathbf{X}$ such that $\mathcal{A}(\mathbf{Z}) = \mathcal{A}(\mathbf{X})$. We have $\mathbf{X} \in T$ by construction and $\langle \mathbf{X}, \mathcal{P}_{T^\perp}(\mathbf{M})\rangle = 0$. Since $\mathcal{P}_T(\mathbf{M}) = \sum_{\ell=1}^r \mathbf{u}_\ell \mathbf{v}_\ell^*$ we have

$$\begin{aligned}\|\mathbf{X}\|_* &= \langle \mathbf{X}, \mathcal{P}_T(\mathbf{M})\rangle_F = \langle \mathbf{X}, \mathbf{M}\rangle_F = \langle \mathbf{X}, \mathcal{A}^*\mathbf{h}\rangle_F = \langle \mathcal{A}(\mathbf{X}), \mathbf{h}\rangle = \langle \mathcal{A}(\mathbf{Z}), \mathbf{h}\rangle \\ &= \langle \mathbf{Z}, \mathcal{A}^*\mathbf{h}\rangle_F = \langle \mathcal{P}_T(\mathbf{Z}), \mathcal{P}_T(\mathbf{M})\rangle_F + \langle \mathcal{P}_{T^\perp}(\mathbf{Z}), \mathcal{P}_{T^\perp}(\mathbf{M})\rangle_F \,.\end{aligned}$$

Observe that $\mathcal{P}_T(\mathbf{M}) = \mathbf{P}_U\mathcal{P}_T(\mathbf{M})\mathbf{P}_V$ by assumption so that by cyclicity of the trace

$$\begin{aligned}\langle \mathcal{P}_T(\mathbf{Z}), \mathcal{P}_T(\mathbf{M})\rangle_F &= \operatorname{tr}\left(\mathcal{P}_T(\mathbf{Z})\mathbf{P}_V\mathbf{P}_T(\mathbf{M})^*\mathbf{P}_U\right) = \operatorname{tr}\left(\mathbf{P}_U\mathcal{P}_T(\mathbf{Z})\mathbf{P}_V\mathcal{P}_T(\mathbf{M})^*\right) \\ &= \langle \mathbf{P}_U\mathbf{Z}\mathbf{P}_V, \mathcal{P}_T(\mathbf{M})\rangle_F \,.\end{aligned}$$

Therefore, by (a)

$$\begin{aligned}\|\mathbf{X}\|_* &\leq \langle \mathbf{P}_U\mathbf{Z}\mathbf{P}_V, \mathcal{P}_T(\mathbf{M})\rangle_F + \langle \mathcal{P}_{T^\perp}(\mathbf{Z}), \mathcal{P}_{T^\perp}(\mathbf{M})\rangle_F \\ &\leq \|\mathbf{P}_U\mathbf{Z}\mathbf{P}_V\|_*\|\mathcal{P}_T(\mathbf{M})\|_{2\to 2} + \|\mathcal{P}_{T^\perp}(\mathbf{Z})\|_*\|\mathcal{P}_{T^\perp}(\mathbf{M})\|_{2\to 2} \\ &< \|\mathbf{P}_U\mathbf{Z}\mathbf{P}_V\|_* + \|\mathcal{P}_{T^\perp}(\mathbf{Z})\|_* \,.\end{aligned}$$

In the last inequality, the assumption on $\mathbf{M}$ was applied. Moreover, the strict inequality holds because $\mathcal{A}$ restricted to T is injective so that $\mathbf{Z}$ is not contained in T and $\mathcal{P}_{T^\perp}(\mathbf{M}) \neq \mathbf{0}$. Set $T^\circ = \operatorname{span}\{\mathbf{u}_\ell\mathbf{v}_\ell^* : \ell \in [r]\}$ and $\widehat{T}$ to be the orthogonal complement of T° in T. Denote by $\mathcal{P}_{\widehat{T}}$ the orthogonal projection onto $\widehat{T}$. Further, $(\mathbf{P}_U\mathbf{Z}\mathbf{P}_V)\mathcal{P}_{T^\perp}(\mathbf{Z})^* = \mathbf{0}$ and $(\mathbf{P}_U\mathbf{Z}\mathbf{P}_V)^*\mathcal{P}_{T^\perp}(\mathbf{Z}) = \mathbf{0}$ since $\mathcal{P}_{T^\perp}(\mathbf{Z}) = (\mathbf{Id}-\mathbf{P}_U)\mathbf{Z}(\mathbf{Id}-\mathbf{P}_V)$ by (c). Therefore, (b) and the duality shown in (a) yields

$$\begin{aligned}\|\mathbf{X}\|_* &< \|\mathbf{P}_U\mathbf{Z}\mathbf{P}_V + \mathcal{P}_{T^\perp}(\mathbf{Z})\|_* \\ &= \sup_{\mathbf{Y}\in\mathbb{C}^{n_1\times n_2}:\|\mathbf{Y}\|_{2\to 2}\leq 1} |\langle \mathbf{P}_U\mathbf{Z}\mathbf{P}_V + \mathcal{P}_{T^\perp}(\mathbf{Z}), \mathbf{Y}\rangle_F| \\ &= \sup_{\mathbf{Y}\in\mathbb{C}^{n_1\times n_2}:\|\mathbf{Y}\|_{2\to 2}\leq 1} |\langle \mathbf{Z}, \mathbf{P}_U\mathbf{Y}\mathbf{P}_V + \mathcal{P}_{T^\perp}(\mathbf{Y})\rangle_F| \\ &= \sup_{\mathbf{Y}\in\mathbb{C}^{n_1\times n_2}:\|\mathbf{Y}\|_{2\to 2}\leq 1, \mathcal{P}_{\widehat{T}}(\mathbf{Y})=\mathbf{0}} |\langle \mathbf{Z}, \mathbf{Y}\rangle_F| \\ &\leq \sup_{\mathbf{Y}\in\mathbb{C}^{n_1\times n_2}:\|\mathbf{Y}\|_{2\to 2}\leq 1} |\langle \mathbf{Z}, \mathbf{Y}\rangle_F| = \|\mathbf{Z}\|_* \,.\end{aligned}$$

(An alternative proof based on the subdifferential of the nuclear norm can be found in [76].)

Hints for Chapter 5

5.1 The inequality $\mu(\mathbf{U},\mathbf{V}) \leq \sqrt{m}$ follows from $|\langle \mathbf{u}_i, \mathbf{u}_j\rangle| \leq \|\mathbf{u}_i\|_2\|\mathbf{v}_j\|_2 = 1$, it is sharp since $\mu(\mathbf{U},\mathbf{U}) = \sqrt{m}$ for any orthonormal basis $\mathbf{U}$; the inequality

$\mu(\mathbf{U},\mathbf{V}) \geq 1$ follows from $1 = \|\mathbf{u}_i\|_2^2 = \sum_{j=1}^m |\langle \mathbf{u}_i, \mathbf{v}_j\rangle|^2 \leq m\,\mu(\mathbf{U},\mathbf{V})^2$, it is sharp since $\mu(\mathbf{E},\mathbf{F}) = 1$ for the canonical and Fourier bases $\mathbf{E}$ and $\mathbf{F}$ of $\mathbb{C}^m$.

5.2 The implication $(ii) \Rightarrow (i)$ is clear; for the implication $(i) \Rightarrow (ii)$, use a polarization formula to derive $\langle \mathbf{x},\mathbf{y}\rangle = \lambda \sum_{j=1}^N \langle \mathbf{x},\mathbf{a}_j\rangle\langle \mathbf{a}_j,\mathbf{y}\rangle$ for all $\mathbf{x},\mathbf{y} \in \mathbb{K}^m$, so that $\mathbf{x} = \lambda \sum_{j=1}^N \langle \mathbf{x},\mathbf{a}_j\rangle \mathbf{a}_j$ for all $\mathbf{x} \in \mathbb{K}^m$; for the equivalence $(ii) \Leftrightarrow (iii)$, observe that $\langle \mathbf{x},\mathbf{y}\rangle = \lambda \sum_{j=1}^N \langle \mathbf{x},\mathbf{a}_j\rangle\langle \mathbf{a}_j,\mathbf{y}\rangle$ for the choice $\mathbf{x} = \mathbf{e}_\ell$ and $\mathbf{y} = \mathbf{e}_k$ reads $\delta_{k,\ell} = \lambda \sum_{j=1}^N \overline{A}_{\ell,j} A_{k,\ell}$; for ℓ_2-normalized vectors $\mathbf{a}_1,\dots,\mathbf{a}_N$, we find $\lambda = m/N$ by taking the trace in (iii).

5.3 Use $\|\mathbf{A}_S^*\mathbf{A}_S - I\|_1 = \|\mathbf{A}_S^*\mathbf{A}_S - I\|_\infty = \max_{i\in S} \sum_{j\in S} \big|(\mathbf{A}_S^*\mathbf{A}_S - I)_{i,j}\big| = \max_{i\in S} \sum_{j\in S, j\neq i} |\langle \mathbf{a}_i,\mathbf{a}_j\rangle|$.

5.4 Interpret the $(m+1)$-vertices if a regular m-simplex as the orthogonal projections of the canonical basis $(\mathbf{e}_1,\dots,\mathbf{e}_{m+1})$ of $\mathbb{R}^{m+1}$ onto $[1,\dots,1]^\top$, express them as $\mathbf{e}_j - [1,\dots,1]^\top/(m+1)$, and compute an inner product between normalized vectors equal to $-1/m$.

5.5 The magnitude of the inner products of $[1,\pm c,0]^\top$ with other columns only gives the values c and $1-c^2$, which are equal; since the shifts of $[1,\pm 1,0,\pm 1,0,0,0]^\top$ only have one common nonzero entry, and since two different $[1,\pm 1,0,\pm 1,0,0,0]^\top$ have an inner product of ± 1, the magnitude of all possible inner products is always 1.

5.8 If $\mathbf{A} = [I|F]$ is the concatenation of the identity and the Fourier matrices, then the unknown vector is $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{C}^m$ where $\mathbf{x} \in \mathbb{C}^{2m}$ is s-sparse, calculate $\mu = 1/\sqrt{m}$ and apply Theorems 5.14 and 5.15.

5.9 Follow the proof of Theorem 5.13 to obtain, for any $\mathbf{v} \in \mathbb{C}^N$, $\|\mathbf{v}_S\|_1 \leq \mu_1(s)\|\mathbf{v}_{\overline{S}}\|_1 + \mu_1(s-1)\|\mathbf{v}_S\|_1 + s\|\mathbf{A}\mathbf{v}\|_2 \leq \nu\|\mathbf{v}_{\overline{S}}\|_1 + \nu\|\mathbf{v}_S\|_1 + s\|\mathbf{A}\mathbf{v}\|_2$, i.e., $\|\mathbf{v}_S\|_1 \leq \nu/(1-\nu)\|\mathbf{v}_{\overline{S}}\|_1 + s/(1-\nu)\|\mathbf{A}\mathbf{v}\|_2$, and apply Theorem 4.18 to deduce the required result with $C = 2/(1-2\nu)$ and $D = 4/(1-2\nu)$.

Hints for Chapter 6

6.1 Use $\|(\mathbf{A}_S^*\mathbf{A}_S)^{-1}\|_{2\to 2} = \lambda_{\max}((\mathbf{A}_S^*\mathbf{A}_S)^{-1}) = 1/\lambda_{\min}((\mathbf{A}_S^*\mathbf{A}_S)^{-1})$ and $\|\mathbf{A}_S^\dagger\|_{2\to 2}^2 = \lambda_{\max}(\mathbf{A}_S^\dagger(\mathbf{A}_S^\dagger)^*) = \lambda_{\max}((\mathbf{A}_S^*\mathbf{A}_S)^{-1})$, together with the fact that the eigenvalues of $\mathbf{A}_S^*\mathbf{A}_S$ are contained in $[1-\delta_s, 1+\delta_s]$.

6.6 If S_0 is an index set of s largest entries of $\mathbf{x}$ in modulus, S_1 an index set of next s largest entries, and so on, write $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}_{S_0}\|_2 + \|\mathbf{A}\mathbf{x}_{S_1}\|_2 + \|\mathbf{A}\mathbf{x}_{S_2}\|_2 + \cdots \leq \sqrt{1+\delta_s}(\|\mathbf{x}_{S_0}\|_2 + \|\mathbf{x}_{S_1}\|_2 + \|\mathbf{x}_{S_2}\|_2 + \cdots) \leq \sqrt{1+\delta_s}(\|\mathbf{x}\|_2 + \|\mathbf{x}_{S_0}\|_1/\sqrt{s} + \|\mathbf{x}_{S_1}\|_1/\sqrt{s} + \cdots)$.

6.7 The first inclusion is obvious. For the second observe that $\|\mathbf{x}\|_1 \leq \sqrt{s}\|\mathbf{x}\|_2 \leq \sqrt{s}$ for a vector $\mathbf{x} \in D_{s,N}$. By considering a convex combination of

elements in $D_{s,N}$ this estimate extends to $\operatorname{conv}(D_{s,N})$ by the triangle inequality. For the inclusion $\sqrt{s}B_1^N \cap B_2^N \subset 2\operatorname{conv}(D_{s,N})$ one proceeds similarly as in the proof of Theorem 6.8. Let $\mathbf{x}$ with $\|\mathbf{x}\|_2 \le 1$ and $\|\mathbf{x}\|_1 \le \sqrt{s}$. Partition $[N] = S_1 \cup S_2 \cup \cdots$ such that S_1 corresponds to the s largest absolute entries of $\mathbf{x}$, S_2 to the next s largest entries and so on. Write $\mathbf{x} = \sum_{j\ge 1} \|\mathbf{x}_{S_j}\|_2 \frac{\mathbf{x}_{S_j}}{\|\mathbf{x}_{S_j}\|_2}$ (the sum ranging only over the non-zero $\mathbf{x}_{S_j}$). and observe that $\mathbf{x}_{S_j}/\|\mathbf{x}_{S_j}\|_2$ is contained in $D_{s,N}$. Then it suffices to show that $\sum_{j\ge 1} \|\mathbf{x}_{S_j}\|_2 \le 2$. Due to Lemma 6.9, $\|\mathbf{x}_{S_j}\|_2 \le \|\mathbf{x}_{S_{j-1}}\|_1/\sqrt{s}$ and since $\|\mathbf{x}_{S_1}\|_2 \le \|\mathbf{x}\|_2 \le 1$, we obtain

$$\sum_{j\ge 1} \|\mathbf{x}_{S_j}\|_2 \le (1 + \sum_{j\ge 2} \|\mathbf{x}_{S_{j-1}}\|_1/\sqrt{s}) \le (1 + \|\mathbf{x}\|_1/\sqrt{s}) \le 2.$$

6.8 Apply the polarization formula to $\mathbf{x} = \mathbf{A}(\mathbf{u}/\|\mathbf{u}\|_2)$ and $\mathbf{y} = e^{i\theta}\mathbf{A}(\mathbf{v}/\|\mathbf{v}\|_2)$ for a properly chosen θ.

6.21 To prove the condition of Proposition 3.5, notice that, for $\operatorname{supp}(\mathbf{z}) \subseteq S$, $(1-\delta_s)\|\mathbf{z}\|_2^2 \le \|\mathbf{A}\mathbf{z}\|_2^2 = \sum_{j\in S} z_j(\mathbf{A}^*\mathbf{A}\mathbf{z})_j \le \sqrt{s}\|\mathbf{z}\|_2 \max_{j\in S} |(\mathbf{A}^*\mathbf{A}\mathbf{z})_j|$, and that $|(\mathbf{A}^*\mathbf{A}\mathbf{z})_\ell| = \langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{e}_\ell\rangle \le \delta_{s+1}\|\mathbf{z}\|_2$ if $\ell \in \overline{S}$.

6.24 (a) First observe that if T is a subspace of $\mathbb{C}^{n_1\times n_2}$ consisting of matrices of rank at most r and $\mathcal{A}_T$ denotes the restriction of $\mathcal{A}$ to T then

$$\delta_r \ge \sup_{\mathbf{X}\in T} \big|\|\mathcal{A}(\mathbf{X})\|_2^2 - \|\mathbf{X}\|_F^2\big| = \sup_{\mathbf{X}\in T} |\langle (\mathcal{A}_T^*\mathcal{A}_T - \mathbf{Id})\mathbf{X}, \mathbf{X}\rangle| = \|\mathcal{A}_T^*\mathcal{A}_T - \mathbf{Id}\|_{F\to F} .$$

For a matrix $\mathbf{X} \in \mathbb{C}^{n_1\times n_2}$ of rank r_1 with singular value decomposition $\mathbf{X} = \sum_{j=1}^{r_1} \sigma_j \mathbf{u}_j\mathbf{v}_j^*$ we define the subspace $T(\mathbf{X})$ spanned by $\{\mathbf{u}_j\mathbf{v}_j^*, j \in [r_1]\}$. Given $\mathbf{X}, \mathbf{Z} \in \mathbb{C}^{n_1\times n_2}$ with $\operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\mathbf{Z}) \le r$ the linear space $T = T(\mathbf{X}) + T(\mathbf{Z})$ contains only matrices of rank at most r, and furthermore, $\mathbf{X}, \mathbf{Z} \in T$. Assuming $\langle \mathbf{X}, \mathbf{Z}\rangle_F = 0$ the same argument as in Proposition (6.3) gives

$$\begin{aligned}|\langle \mathcal{A}(\mathbf{X}), \mathcal{A}(\mathbf{Z})\rangle| &= |\langle \mathcal{A}_T(\mathbf{X}), \mathcal{A}_T(\mathbf{Z})\rangle - \langle \mathbf{X}, \mathbf{Z}\rangle_F| \le \|\mathcal{A}_T^*\mathcal{A}_T - \mathbf{Id}\|_{F\to F}\|\mathbf{X}\|_F\|\mathbf{Z}\|_F \\ &\le \delta_r\|\mathbf{X}\|_F\|\mathbf{Z}\|_F .\end{aligned}$$

(b) Given a matrix $\mathbf{M} \in \ker\mathcal{A} \setminus \{\mathbf{0}\}$ with singular value decomposition $\mathbf{M} = \sum_{j=1}^{\min\{n_1,n_2\}} \sigma_j\mathbf{u}_j\mathbf{v}_j^*$ define matrices $\mathbf{M}_k$, $k = 0, 1, \ldots$, via $\mathbf{M}_k = \sum_{j=kr+1}^{(k+1)r} \sigma_j\mathbf{u}_j\mathbf{v}_j^*$. Then $\mathbf{M} = \sum_{k\ge 0} \mathbf{M}_k$ and $\langle \mathbf{M}_k, \mathbf{M}_0\rangle_F = 0$ for $k \ge 1$. Proceed then analogously as in the proof of Theorem 6.8 using also (a).

(c) Use the result of Exercise 4.18 and proceed analogously to the proof of Theorem 6.12.

Hints for Chapter 8

7.3 Use Hölder's inequality $\mathbb{E}(\xi\chi_{\{\xi>t\}}) \le \mathbb{E}(\xi^p)^{(p-1)/p}\mathbb{E}(\chi_{\{\xi>t\}})^{1/p^*}$ in the proof of Lemma 7.16; then apply the result to $\xi := \big(\sum_{i=1}^n x_i\xi_i\big)^2$ and use symmetrization, Khintchine inequality, and the convexity of $t \mapsto t^p$ to observe

$$\mathbb{E}(\xi^p) \le C_p \mathbb{E}_\xi \mathbb{E}_\epsilon \Big(\big(\sum \epsilon_i x_i \xi_i \big)^{2p} \Big) \le C'_p \mathbb{E}_\xi \Big(\big(\sum x_i^2 \xi_i^2 \big)^p \Big)$$
$$= C'_p \|\mathbf{x}\|_2^{2p} \mathbb{E}_\xi \Big(\big(\sum \theta_i \xi_i^2 \big)^p \Big) \le C'_p \|\mathbf{x}\|_2^{2p} \mathbb{E}_\xi \Big(\sum \theta_i |\xi_i|^{2p} \Big) \le C'_p \|\mathbf{x}\|_2^{2p} \mu^{2p}.$$

8.5 (a) Use Hölder's inequality to estimate $\mathbb{E}\|A\mathbf{g}\|_2 \le \sqrt{\mathbb{E}\|A\mathbf{g}\|_2} = \|A\|_F$. The triangle inequality and the definition of the operator norm shows that $|\|A\mathbf{x}\|_2 - \|A\mathbf{y}\|_2| \le \|A(\mathbf{x}-\mathbf{y})\|_2 \le \|A\|_{2\to 2}\|\mathbf{x}-\mathbf{y}\|_2$ so that the Lipschitz constant of the function $x \mapsto \|A\mathbf{x}\|_2$ is bounded by (actually equal to) $\|A\|_{2\to 2}$. The tail bound follows then from concentration of measure, Theorem 8.38.

(b) With $B = A^*A$ write $\|A\boldsymbol{\epsilon}\|_2 = \boldsymbol{\epsilon}^* B \boldsymbol{\epsilon}$. Noting that B is positive semidefinite the proof of Theorem 8.13, see (8.22), shows that

$$\mathbb{E}\exp(\theta\|A\boldsymbol{\epsilon}\|_2) \le \exp\left(\frac{\theta\|A\|_F^2}{1-8\theta\|A\|_{2\to 2}^2}\right) \qquad \text{for } 0 < \kappa < 1/(8\|A\|_{2\to 2}^2)\,.$$

The choice $\theta = 1/(16\|A\|_{2\to 2}^2)$ yields $\mathbb{E}\exp(\theta\|A\boldsymbol{\epsilon}\|_2) \le \exp(2\theta\|A\|_F^2)$. Use Markov's inequality to deduce that

$$\mathbb{P}(\|A\boldsymbol{\epsilon}\|_2^2 \ge 2\|A\|_F^2 + 16t\|A\|_{2\to 2}^2)$$
$$\le \mathbb{E}\exp(\theta\|A\boldsymbol{\epsilon}\|_2)\exp(-2\theta\|A\|_F^2 - 16\theta t\|A\|_{2\to 2}^2) \le e^{-t}\,.$$

Apply the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ to get

$$\mathbb{P}\left(\|A\boldsymbol{\epsilon}\|_2 \ge \sqrt{2}\|A\|_F + 4t\|A\|_{2\to 2}\right) \le \mathbb{P}\left(\|A\boldsymbol{\epsilon}\|_2 \ge \sqrt{2\|A\|_F^2 + 16t^2\|A\|_{2\to 2}^2}\right)$$
$$\le e^{-t^2}\,.$$

Hints for Chapter 9

9.4 Use that $\sigma_{\min}(\frac{1}{\sqrt{m}}\mathbf{B}) \le \sqrt{1-\delta} \le 1-\delta$ if $\|m^{-1}\mathbf{B}^*\mathbf{B} - \mathbf{Id}\|_2 \le \delta$, and apply (9.15).

Hints for Chapter 10

10.1 $d^1(B_1^2, \ell_2^2) = 1/\sqrt{2}$, $d^1(B_1^3, \ell_2^3) = \sqrt{2/3}$, $d^2(B_1^3, \ell_2^3) = 1/\sqrt{3}$.

10.9 Original proof of the lower bound
(a) For $d_m(C,X) < \alpha < \varepsilon/2$, consider a subspace X_m of X with $\dim(X_m) \le m$ such that $\sup_{\mathbf{x}\in C}\inf_{\mathbf{z}\in X_m}\|\mathbf{x}-\mathbf{z}\| < \alpha$; let $\{\mathbf{u}_1,\dots,\mathbf{u}_n\}$ be a maximal ε-separating set for $C \cap tB_X$, so that $n = P(\varepsilon, C\cap tB_X, X)$, choose $\mathbf{z}_i \in X_m$ with $\|\mathbf{u}_i - \mathbf{z}_i\| < \alpha$; for $i \ne j$, observe that $\|\mathbf{z}_i\| \le \|\mathbf{u}_i\| + \|\mathbf{u}_i - \mathbf{z}_i\| \le t + \alpha$ and $\|\mathbf{z}_i - \mathbf{z}_j\| \ge \|\mathbf{u}_i - \mathbf{u}_j\| - \|\mathbf{u}_i - \mathbf{z}_i\| - \|\mathbf{u}_j - \mathbf{z}_j\| \ge \varepsilon - 2\alpha$; thus $\{\mathbf{z}_1,\dots,\mathbf{z}_n\}$ is an $(\varepsilon-2\alpha)$-separating set for $(t+\alpha)B_{X_m}$; use Appendix ?? and let α tend

to $d_m(C,X)$.
(b) The set $\big\{\mathbf{x} \in \{-k^{-1/p}, 0, k^{-1/p}\}^N : \mathrm{card}(\{j : x_j \neq 0\}) = k\big\}$ is a subset of $B_p^N \cap k^{-1/p} B_\infty^N$ with $\geq 2^k \binom{N}{k}$ elements that are ε-separated in ℓ_∞^N.
(c) Assume that $\varepsilon := 3\, d_m(B_p^N, \ell_\infty^N) < 1$ and choose $k \geq 1$ as the largest integer smaller than $1/\varepsilon^p$, so that $1/(2\varepsilon^p) \leq k < 1/\varepsilon^p$; apply 1. and 2. with $C = B_p^N$, $X = \ell_\infty^N$, $t := k^{-1/p}$ — note that 2. can be applied in view of Exercise 10.7 — to get

$$2^k \binom{N}{k} \leq \left(1 + 2\frac{k^{-1/p} + \varepsilon/3}{\varepsilon/3}\right)^m ;$$

the right-hand side can be bounded by $15^m \leq e^{3m}$ because $k^{-1/p} \leq 2^{1/p}\varepsilon \leq 2\varepsilon$, for the left-hand side, use $2^k \binom{N}{k} \geq (2N/k)^k \geq (2\varepsilon^p N)^{1/(2\varepsilon^p)}$; take the logarithm and observe that $\varepsilon^p \geq 3^p/(m+1) \geq 3/(2m)$ according to Exercise 10.7.

Hints for Chapter 11

11.4 Adapt the proof of Theorem 11.4.

11.5 Take $\mathbf{x} = 0$ and $\mathbf{x} = \mathbf{v} \in \ker \mathbf{A}$ to obtain $\|\mathbf{v}\| \leq C/\sqrt{s}\|\mathbf{v}\|_1$; then, whenever $|T| < t$, $\|\mathbf{v}_T\|_2 \leq \|\mathbf{v}\|_2 \leq \rho/(2\sqrt{t})\|\mathbf{v}_1\|_1$; this is the ℓ_2-robust null space property; conclude using Theorem ??.

11.7 Use Exercise 4.14.

11.8 To obtain $\|\mathbf{y}\|_*^{(\alpha)} \geq \inf\{\|\mathbf{y}'\|_2 + \|\mathbf{y}''\|_1/\alpha,\ \mathbf{y}' + \mathbf{y}'' = \mathbf{y}\}$, reproduce the arguments of the proof of Lemma (11.20); To obtain reversed inequality, write $|\langle \mathbf{y}' + \mathbf{y}'', \mathbf{u}\rangle| \leq \|\mathbf{y}'\|_2\|\mathbf{u}\|_2 + \|\mathbf{y}''\|_1\|\mathbf{u}\|_\infty$, take the supremum over $\mathbf{u} \in \mathbb{C}^m$ with $\|\mathbf{u}\|^{(\alpha)} \leq 1$, then the infimum over $\mathbf{y}', \mathbf{y}'' \in \mathbb{C}^m$ with $\mathbf{y}' + \mathbf{y}'' = \mathbf{y}$.

11.2 We first isolate the case $s = 1$. The mixed (ℓ_q, ℓ_p)-instance optimality of order 1 with constant C reads $\|\mathbf{x} - \Delta(\mathbf{A}\mathbf{x})\|_q \leq C\, \sigma_1(\mathbf{x})_p$ for all $\mathbf{x} \in \mathbb{C}^N$. The mixed $(\ell_q, \ell_{p'})$-instance optimality of order 1 with constant C then simply follows from the inequality $\sigma_1(\mathbf{x})_p \leq \sigma_1(\mathbf{x})_{p'}$ when $p \geq p'$. Next, we deal with the case $s \geq 2$. For any $\mathbf{x} \in \mathbb{C}^N$, Proposition 2.3 yields the inequality

$$\sigma_s(\mathbf{x})_p \leq \frac{1}{\lfloor s/2 \rfloor^{1/p' - 1/p}} \sigma_{\lceil s/2 \rceil}(\mathbf{x})_{p'}.$$

We now use the mixed (ℓ_q, ℓ_p)-instance optimality of order s with constant C, while noticing that $s \geq 3\lceil s/2 \rceil/2$ and that $\lfloor s/2 \rfloor \geq \lceil s/2 \rceil/2$, to write

$$\begin{aligned}\|\mathbf{x}-\Delta(\mathbf{A}\mathbf{x})\|_q &\le \frac{C}{s^{1/p-1/q}}\sigma_s(\mathbf{x})_p \le \frac{C}{s^{1/p-1/q}}\frac{1}{\lfloor s/2\rfloor^{1/p'-1/p}}\sigma_{\lceil s/2\rceil}(\mathbf{x})_{p'}\\ &\le \frac{2^{1/p-1/q}C}{(3\lceil s/2\rceil)^{1/p-1/q}}\frac{2^{1/p'-1/p}}{\lceil s/2\rceil^{1/p'-1/p}}\sigma_{\lceil s/2\rceil}(\mathbf{x})_{p'}\\ &\le \frac{2\,C}{\lceil s/2\rceil^{1/p'-1/q}}\sigma_{\lceil s/2\rceil}(\mathbf{x})_{p'}.\end{aligned}$$

This is the mixed (ℓ_q,ℓ_p)-instance optimality of order $\lceil s/2\rceil$ with constant $C'=2\,C$.

Hints for Chapter 12

12.10 Combine Theorem 9.34 with Theorem 12.32.

Hints for Chapter 14

14.2 [TO BE WRITTEN]

14.6 See Exercise 4.9, we need to prove that $\sum_{j=1}^N v_j = 0$ for every $\mathbf{v}\in\ker\mathbf{A}$, which is seen from $0=\sum_{i=1}^m(\mathbf{A}\mathbf{v})_i=\sum_{\overline{j\,i}\in E} v_j = d\sum_{j=1}^N v_j$.

14.3 For each n, define $R_n(J)$ as the set of right vertices connected to exactly n left vertices in J, from $\operatorname{card}(R_{\ge 2}(J))=\operatorname{card}(R_2(J))+\operatorname{card}(R_3(J))+\cdots$ and $\operatorname{card}(R(J))=\operatorname{card}(R_1(J))+\operatorname{card}(R_2(J))+\operatorname{card}(R_3(J))+\cdots$, we derive by counting the edges in $E(J)$ that $d\operatorname{card}(J)=\operatorname{card}(R_1(J))+2\operatorname{card}(R_2(J))+3\operatorname{card}(R_3(J))+\cdots\ge\operatorname{card}(R(J))+\operatorname{card}(R_{\ge 2}(J))\ge(1-\theta)d\operatorname{card}(J)+\operatorname{card}(R_{\ge 2}(J))$, hence the result.

List of Symbols

δ_s	restricted isometry constant of order s
μ	coherence
μ_1	ℓ_1-coherence function
$\mathbf{A}$	usually the measurement matrix
$\mathbf{A}^\top$	transpose of matrix $\mathbf{A}$, $(\mathbf{A}^\top)_{jk} = A_{jk}$
$\mathbf{A}^*$	Hermitian conjugate of matrix $\mathbf{A}$, i.e., $(\mathbf{A}^*)_{jk} = \overline{A}_{jk}$
$\mathbf{a}_j$	jth column of the matrix $\mathbf{A}$
$\mathbf{A}_S$	submatrix of $\mathbf{A}$ obtained by selecting the columns indexed by S
$M_{I,J}$	submatrix of M with rows indexed by I and columns indexed by J (p. 45)
B_p^N	unit ball of the (quasi)normed space ℓ_p^N (p. 38)
$\mathrm{card}(S)$	cardinality of the set S
cone	conic hull (p. 490)
conv	convex hull (p. 489)
ℓ_p^N	$\mathbb{C}^N$ equipped with the ℓ_p-norm
$\overline{S}$	complement of the set S, often $\overline{S} = [N] \setminus S$
$\mathbf{x}$	usually the vector in $\mathbb{C}^N$ to be recovered
$\mathbf{x}_S$	either the vector in $\mathbb{C}^N$ equal to $\mathbf{x}$ on S and to zero on $\bar{S}$, or the vector in $\mathbb{C}^S$, which is the restriction of $\mathbf{x}$ to the entries in S
$\mathbf{y}$	usually the measurement vector, $\mathbf{y} = \mathbf{A}\mathbf{x}$
$\mathbf{Id}$	identity matrix
I_m	the $m \times m$ identity matrix
J_N	the $N \times N$ matrix with all entries equal to one
$\mathrm{sgn}(a)$	sign of $a \in \mathbb{C}$, p. 81
$\mathrm{sgn}(\mathbf{x})$	componentwise sign of the vector $\mathbf{x} \in \mathbb{C}^N$, p. 81
$\mathrm{supp}(\mathbf{x})$	support of the vector $\mathbf{x}$
$(\mathbf{e}_1, \dots, \mathbf{e}_n)$	canonical basis of $\mathbb{K}^N$
$\|\mathbf{x}\|_p$	ℓ_p-norm, $0 < p \leq \infty$ (p. 37, p. 464)
$\|\mathbf{x}\|_{p,\infty}$	weak ℓ_p-quasinorm of a vector $\mathbf{x}$ (p. 40)
$\|\mathbf{x}\|_0$	number of nonzero entries of a vector $\mathbf{x}$ (p. 37)
$\|\mathbf{A}\|_{p\to q}$	operator norm between ℓ_p and ℓ_q of the matrix $\mathbf{A}$ (p. 467)

$\|\mathbf{A}\|_{2\to 2}$	operator norm (largest singular value) of the matrix $\mathbf{A}$ on ℓ_2 (p. 467)
$\|\mathbf{A}\|_F$	Frobenius norm of the matrix $\mathbf{A}$ (p. 471)
$\sigma_s(\mathbf{x})_p$	error of best s-term approximation to a vector $\mathbf{x}$ (p. 38)
$\mathbf{x}^*$	nonincreasing rearrangement of the vector $\mathbf{x}$ (p. 38)
$\mathbb{N}$	natural numbers $\{1, 2, \ldots\}$
$\mathbb{N}_0$	natural numbers including 0, $\{0, 1, 2, \ldots\}$
$\mathbb{Z}$	integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Q}$	rational numbers
$\mathbb{R}$	real numbers
$\mathbb{R}_+$	subset of $\mathbb{R}$ consisting of the nonnegative real numbers (p. 38)
$\mathbb{C}$	complex numbers
$\mathbb{R}^N$	N-dimensional real vector space
$\mathbb{C}^N$	N-dimensional complex vector space
$\mathbb{C}^S$	the space of vectors $\mathbf{x}$ indexed by the set S, isomorphic to $\mathbb{C}^{\mathrm{card}(S)}$
$\mathbb{K}$	field $\mathbb{R}$ or $\mathbb{C}$
$\mathbb{E}$	expectation
$\mathbb{P}$	probability of an event
$[N]$	the set $\{1, 2, \ldots, N\}$ of the natural integers not exceeding N (p. 37)

References

1. P. Abrial, Y. Moudden, J. Starck, J. Fadili, J. Delabrouille, and M. Nguyen. CMB data analysis and sparsity. *Stat. Methodol.*, 5:289–298, 2008. (Cited on p. 388.)
2. R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13(34):1000–1034, 2008. (Cited on p. 243.)
3. F. Affentranger and R. Schneider. Random projections of regular simplices. *Discrete Comput. Geom.*, 7(3):219–226, 1992. (Cited on p. 280.)
4. M. Aharon, M. Elad, and A. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322, 2006. (Cited on p. 36.)
5. R. Ahlswede and A. Winter. Strong converse for identification via quantum channels,. *IEEE Trans. Inform. Theory*, 48(3):569 –579, 2002. (Cited on p. 240.)
6. N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009. (Cited on p. 392.)
7. N. Ailon and E. Liberty. Almost optimal unrestricted fast Johnson-Lindenstrauss transform. In *Symposium on Discrete Algorithms (SODA)*, 2011. (Cited on p. 392.)
8. K. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.*, 12(4):1041–1067, 1984. (Cited on p. 243.)
9. W. O. Alltop. Complex sequences with low periodic correlations. *IEEE Trans. Inform. Theory*, 26(3):350–354, 1980. (Cited on p. 115.)
10. G. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2010. (Cited on p. 277.)
11. F. Andersson, M. Carlsson, and M. V. de Hoop. Nonlinear approximation of functions in two dimensions by sums of exponentials. *Appl. Comput. Harmon. Anal.*, 29(2):198–213, 2010. (Cited on p. 52.)
12. F. Andersson, M. Carlsson, and M. V. de Hoop. Sparse approximation of functions using sums of exponentials and AAK theory. *J. Approx. Theory*, 163(2):213–248, 2011. (Cited on p. 52.)

13. G. Andrews, R. Askey, and R. Roy. *Special Functions*, volume 71 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1999. (Cited on pp. 386, 388.)
14. M. Anthony and P. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, 1999. (Cited on p. 34.)
15. S. Arora and B. Barak. *Computational complexity*. Cambridge University Press, Cambridge, 2009. A modern approach. (Cited on pp. 52, 428.)
16. K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-linear Programming*. : Stanford University Press. 229 p., 1958. (Cited on p. 460.)
17. U. Ayaz and H. Rauhut. Nonuniform sparse recovery with subgaussian matrices. *Preprint*, 2011. (Cited on p. 277.)
18. J.-M. Azais and M. Wschebor. *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons Inc., 2009. (Cited on p. 241.)
19. F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012. (Cited on p. 34.)
20. B. Bah and J. Tanner. Improved bounds on restricted isometry constants for Gaussian matrices. *SIAM J. Matrix Anal. Appl.*, 31(5):2882–2898, 2010. (Cited on p. 281.)
21. Z. Bai and J. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, Second edition, 2010. (Cited on p. 277.)
22. W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak. Toeplitz-structured compressed sensing matrices. In *Proc. IEEE Stat. Sig. Proc. Workshop*, pages 294–298, 2007. (Cited on p. 390.)
23. W. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak. Compressed Channel Sensing: A New Approach to Estimating Sparse Multipath Channels. *Proc. IEEE*, 98(6):1058–1076, June 2010. (Cited on p. 34.)
24. R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008. (Cited on p. 277.)
25. S. Bartels. Total variation minimization with finite elements: convergence and iterative solution. *preprint*, 2011. (Cited on p. 460.)
26. A. Barvinok. Measure concentration, 2005. (Cited on p. 242.)
27. I. Bechar. A Bernstein-type inequality for stochastic processes of quadratic forms of Gaussian variables. *preprint*, 2009. (Cited on p. 241.)
28. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009. (Cited on pp. 460, 461, 461.)
29. S. Becker, J. Bobin, and E. J. Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM J. Imaging Sci.*, 4(1):1–39, 2011. (Cited on p. 461.)
30. J. J. Benedetto and P. J. S. G. Ferreira. *Modern Sampling Theory. Mathematics and Applications*. Birkhäuser, 2001. (Cited on pp. 32, 517.)
31. G. Bennett. Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.*, 57:33–45, 1962. (Cited on p. 181.)
32. R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. *preprint*, 2008. (Cited on pp. 32, 35, 429, 429.)

33. R. Berinde, P. Indyk, and M. Ržić. Practical near-optimal sparse recovery in the L1 norm. In *Proc. Allerton*, 2008. (Cited on p. 429.)
34. S. Bernstein. Sur une modification de l'inéqualité de Tchebichef. *Annals Science Insitute Sav. Ukraine, Sect. Math. I*, 1924. (Cited on p. 181.)
35. S. Bernstein. *Theory of Probability.* 1927. (Cited on p. 181.)
36. G. Beylkin and L. Monzón. On approximation of functions by exponential sums. *Appl. Comput. Harmon. Anal.*, 19(1):17–48, 2005. (Cited on p. 52.)
37. G. Beylkin and L. Monzón. Approximation by exponential sums revisited. *Appl. Comput. Harmon. Anal.*, 28(2):131–149, 2010. (Cited on p. 52.)
38. R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, 1997. (Cited on pp. 463, 488, 510, 511, 535.)
39. P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. (Cited on p. 34.)
40. I. Bjelakovic and R. Siegmund Schultze. Quantum Stein's lemma revisited, inequalities for quantum entropies, and a concavity theorem of Lieb. *preprint*, 2012. (Cited on p. 510.)
41. A. Björck. *Numerical Methods for Least Squares Problems.* SIAM, 1996. (Cited on pp. 463, 480, 481, 481, 481, 481, 481.)
42. R. Blahut. *Algebraic Codes for Data Transmission.* Cambridge Univ. Press, Cambridge, U.K., 2003. (Cited on p. 52.)
43. J. Blanchard, C. Cartis, and J. Tanner. Compressed sensing: how sharp is the restricted isometry property? *SIAM Rev.*, 53(1):105–125, 2011. (Cited on p. 281.)
44. J. Blanchard and A. Thompson. On support sizes of restricted isometry constants. *Appl. Comput. Harmon. Anal.*, 29(3):382–390, 2010. (Cited on p. 153.)
45. T. Blu, P. Marziliano, and M. Vetterli. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.*, 50(6):1417–1428, 2002. (Cited on p. 52.)
46. T. Blumensath and M. Davies. Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.*, 14:629–654, 2008. (Cited on p. 153.)
47. T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009. (Cited on p. 153.)
48. T. Blumensath and M. Davies. Normalized iterative hard thresholding: guaranteed stability and performance. *IEEE J. Select Topics Signal Processing*, 4(2):298–309, april , 2010. (Cited on p. 153.)
49. S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005. (Cited on p. 243.)
50. S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003. (Cited on p. 243.)
51. J. Bourgain. Bounded orthogonal systems and the $\Lambda(p)$-set problem. *Acta Math.*, 162(3-4):227–245, 1989. (Cited on p. 389.)
52. J. Bourgain. Λ_p-sets in analysis: results, problems and related aspects. In *Handbook of the Geometry of Banach Spaces, Vol I*, pages 195–232. North-Holland, 2001. (Cited on p. 389.)
53. J. Bourgain, S. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova. Breaking the k^2-barrier for explicit RIP matrices. In *STOC'11*, pages 637–644, 2011. (Cited on p. 154.)

54. J. Bourgain, S. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova. Explicit constructions of RIP matrices and related problems. *Duke Math. J.*, 159(1):145–185, 2011. (Cited on p. 154.)
55. J. Bourgain and L. Tzafriri. Invertibility of 'large' submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math.*, 57(2):137–224, 1987. (Cited on pp. 240, 408.)
56. J. Bourgain and L. Tzafriri. On a problem of Kadison and Singer. *J. Reine Angew. Math.*, 420:1–43, 1991. (Cited on pp. 408, 409.)
57. O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R., Math., Acad. Sci. Paris*, 334(6):495–500, 2002. (Cited on p. 243.)
58. O. Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications*, volume 56 of *Progr. Probab.*, pages 213–247. Birkhäuser, 2003. (Cited on p. 243.)
59. S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge Univ. Press, 2004. (Cited on pp. 64, 460, 462, 489, 503, 505.)
60. A. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 51(1):34–81, 2009. (Cited on p. 33.)
61. A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Math. Ann.*, 319:1–16, 2001. (Cited on p. 240.)
62. A. Buchholz. Optimal constants in Khintchine type inequalities for fermions, Rademachers and q-Gaussian operators. *Bull. Pol. Acad. Sci. Math.*, 53(3):315–321, 2005. (Cited on p. 240.)
63. P. Bühlmann and d. van. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, 2011. (Cited on p. 34.)
64. H. Buhrman, P. Miltersen, J. Radhakrishnan, and S. Venkatesh. Are bitvectors optimal? In *STOC '00: Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, pages 449–458. ACM, 2000. (Cited on p. 299.)
65. V. Buldygin and Y. Kozachenko. *Metric Characterization of Random variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs.* American Mathematical Society, Providence, RI, 2000. Translated from the 1998 Russian original by V. Zaiats. (Cited on p. 181.)
66. M. Burger, M. Moeller, M. Benning, and S. Osher. An adaptive inverse scale space method for compressed sensing. Technical Report 11-08, UCLA, 2011. (Cited on pp. 460, 461.)
67. N. Burq, S. Dyatlov, R. Ward, and M. Zworski. Weighted eigenfunction estimates with applications to compressed sensing. *Arxiv preprint arXiv:1111.2383*, 2011. (Cited on p. 388.)
68. T. Cai, L. Wang, and G. Xu. New bounds for restricted isometry constants. *IEEE Trans. Inform. Theory*, 56(9):4388–4394, 2010. (Cited on p. 153.)
69. T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse vectors. *IEEE Trans. Signal Process.*, 58(3):1300–1308, 2010. (Cited on p. 153.)
70. E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris S'er. I Math.*, 346:589–592, 2008. (Cited on p. 153.)
71. E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31(1):59 – 73, 2011. (Cited on p. 277.)

72. E. J. Candès, J., T. Tao, and J. Romberg. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006. (Cited on pp. 31, 32, 32, 34, 93, 383, 384.)
73. E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009. (Cited on pp. 407, 408.)
74. E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory*, 57(11):7235 – 7254, 2011. (Cited on p. 383.)
75. E. J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011. (Cited on p. 277.)
76. E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, 2009. (Cited on pp. 34, 536.)
77. E. J. Candès and B. Recht. Simple bounds for low-complexity model reconstruction. *Preprint*, 2011. (Cited on p. 277.)
78. E. J. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Found. Comput. Math.*, 6(2):227–254, 2006. (Cited on pp. 33, 389, 389, 408.)
79. E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007. (Cited on p. 383.)
80. E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006. (Cited on pp. 93, 153.)
81. E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005. (Cited on pp. 33, 152, 153.)
82. E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006. (Cited on pp. 32, 152, 277, 384, 384.)
83. E. J. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007. (Cited on pp. 34, 64.)
84. E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010. (Cited on p. 34.)
85. P. Casazza and G. Pfander. A two-sided restricted invertibility theorem. in preparation. (Cited on p. 408.)
86. P. Casazza and J. Tremain. Revisiting the Bourgain-Tzafriri restricted invertibility theorem. *Oper. Matrices*, 3(1):97–110, 2009. (Cited on p. 409.)
87. D. Chafaï, O. Guédon, G. Lecué, and A. Pajor. *Interactions between compressed sensing, random matrices and high dimensional geometry.* to appear. (Cited on p. 384.)
88. A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. In *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Ser. Comput. Appl. Math.*, pages 263–340. Walter de Gruyter, Berlin, 2010. (Cited on p. 34.)
89. A. Chambolle, R. A. DeVore, N.-y. Lee, and B. J. Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.*, 7(3):319–335, 1998. (Cited on p. 33.)

90. A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997. (Cited on p. 34.)
91. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40:120–145, 2011. (Cited on pp. 34, 460, 460, 460, 460, 461.)
92. V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Preprint*, 2010. (Cited on p. 277.)
93. S. Chen, S. Billings, and W. Luo. Orthogonal least squares methods and their application to nonlinear system identification. *Intl. J. Contr.*, 50(5):18731896, 1989. (Cited on p. 65.)
94. S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999. (Cited on pp. 31, 33, 64.)
95. H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952. (Cited on p. 181.)
96. T. Chihara. *An introduction to orthogonal polynomials.* Gordon and Breach Science Publishers, 1978. (Cited on p. 386.)
97. S. Chrétien and S. Darses. Invertibility of random submatrices via tail decoupling and a matrix Chernoff inequality. *preprint*, 2011. (Cited on p. 407.)
98. O. Christensen. *An Introduction to Frames and Riesz Bases.* Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, 2003. (Cited on pp. 116, 390.)
99. O. Christensen. *Frames and Bases. An Introductory Course.* Applied and Numerical Harmonic Analysis. Basel Birkhäuser, 2008. (Cited on p. 116.)
100. A. Cline. Rate of convergence of Lawson's algorithm. *Math. Commun.*, 26:167–176, 1972. (Cited on p. 461.)
101. A. Cohen. *Numerical Analysis of Wavelet Methods.* North-Holland, 2003. (Cited on pp. 33, 386.)
102. A. Cohen, W. Dahmen, and R. A. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009. (Cited on pp. 51, 93, 152, 299, 330.)
103. A. Cohen, R. DeVore, S. Foucart, and H. Rauhut. Recovery of functions of many variables via compressive sensing. In *Proc. SampTA 2011, Singapore*, 2011. (Cited on p. 384.)
104. R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Appl. Comput. Harmon. Anal.*, 10(1):27–44, 2001. (Cited on pp. 384, 385, 385.)
105. J. Cooley and J. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965. (Cited on p. 519.)
106. G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. In *CISS*, Princeton, 2006. (Cited on pp. 32, 35.)
107. H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Actual. sci. industr.*, 736:5–23, 1938. (Cited on p. 180.)
108. F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc., New Ser.*, 39(1):1–49, 2002. (Cited on p. 34.)
109. F. Cucker and D.-X. Zhou. *Learning theory: an approximation theory viewpoint.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007. (Cited on p. 34.)
110. W. Dai and O. Milenkovic. Subspace Pursuit for Compressive Sensing Signal Reconstruction. *preprint*, 2008. (Cited on pp. 65, 154.)

111. M. Dal. *An introduction to Γ-convergence*. Progress in Nonlinear Differential Equations and their Applications, 8. Birkhäuser Boston Inc., 1993. (Cited on p. 532.)
112. P. Daniel. Fast algorithms for discrete polynomial transforms on arbitrary grids. *Linear Algebra and its Applications*, 366:353 – 370, 2003. (Cited on p. 388.)
113. S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures Algorithms*, 22(1):60–65, 2003. (Cited on p. 277.)
114. I. Daubechies. *Ten Lectures on Wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, 1992. (Cited on pp. 33, 386.)
115. I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004. (Cited on pp. 34, 460, 461.)
116. I. Daubechies, R. DeVore, M. Fornasier, and C. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63(1):1–38, 2010. (Cited on p. 461.)
117. I. Daubechies, M. Fornasier, and I. Loris. Accelerated projected gradient methods for linear inverse problems with sparsity constraints. *J. Fourier Anal. Appl.*, 14(5-6):764–792, 2008. (Cited on p. 461.)
118. K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of the geometry of Banach spaces I*. Elsevier, 2001. (Cited on p. 277.)
119. M. Davies and Y. Eldar. Rank Awareness in Joint Sparse Recovery. *IEEE Trans. Inform. Theory*, 58(2):1135 –1146, 2012. (Cited on p. 35.)
120. M. Davies and R. Gribonval. Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. *IEEE Trans. Inform. Theory*, 55(5):2203–2214, 2009. (Cited on p. 153.)
121. G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Opt. Eng.*, 33(7):21832191, 1994. (Cited on p. 65.)
122. M. De, V. De, and L. Rosasco. Elastic-net regularization in learning theory. *J. Complexity*, 25(2):201–230, 2009. (Cited on p. 34.)
123. V. de la Peña and E. Giné. *Decoupling. From Dependence to Independence*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. (Cited on p. 240.)
124. R. A. DeVore and G. G. Lorentz. *Constructive Approximation.*, volume 303 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 1993. (Cited on p. 464.)
125. R. A. DeVore, G. Petrova, and P. Wojtaszczyk. Instance-optimality in probability with an ℓ_1-minimization decoder. *Appl. Comput. Harmon. Anal.*, 27(3):275–288, 2009. (Cited on p. 331.)
126. D. Donoho. De-noising by soft-thresholding. 41(3):613 –627, 1995. (Cited on p. 33.)
127. D. Donoho and B. Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(2):577–591, 1992. (Cited on p. 31.)
128. D. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proc. Nat. Acad. Sci. U.S.A.*, 106(45):18914–18919, 2009. (Cited on pp. 66, 281.)

129. D. L. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. *preprint*, 2005. (Cited on p. 93.)
130. D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006. (Cited on pp. 31, 299.)
131. D. L. Donoho. For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution. *Commun. Pure Appl. Anal.*, 59(6):797–829, 2006. (Cited on p. 153.)
132. D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.*, 35(4):617–652, 2006. (Cited on pp. 35, 278, 278, 278.)
133. D. L. Donoho and M. Elad. Optimally sparse representations in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc. Nat. Acad. Sci.*, 100:2197–2202, 2002. (Cited on pp. 33, 51, 93.)
134. D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ell^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202, 2003. (Cited on p. 31.)
135. D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006. (Cited on p. 33.)
136. D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001. (Cited on pp. 31, 33, 33, 93, 383.)
137. D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998. (Cited on p. 33.)
138. D. L. Donoho and G. Kutyniok. Microlocal analysis of the geometric separation problem. *Comm. Pure Appl. Math.*, to appear. (Cited on p. 33.)
139. D. L. Donoho and Michael Elad. On the stability of the basis pursuit in the presence of noise. *Signal Processing*, 86(3):511–532, 2006. (Cited on p. 33.)
140. D. L. Donoho, J.-L. Starck, and E. J. Candès. The curvelet transform for image denoising. *IEEE Trans. Image Process.*, 11(6):670–684, 2002. (Cited on p. 33.)
141. D. L. Donoho and P. Stark. Recovery of a sparse signal when the low frequency information is missing. Technical report, Dept. Stat., UCB., June 1989. (Cited on p. 33.)
142. D. L. Donoho and P. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 48(3):906–931, 1989. (Cited on pp. 33, 383, 389.)
143. D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27):9452–9457, 2005. (Cited on pp. 35, 278, 278, 280.)
144. D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proc. Nat. Acad. Sci.*, 102(27):9446–9451, 2005. (Cited on pp. 35, 278, 278.)
145. D. L. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.*, 22(1):1–53, 2009. (Cited on pp. 33, 35, 278, 278, 278, 280.)
146. D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4273–4293, 2009. (Cited on p. 280.)

147. D. L. Donoho and Y. Tsaig. Fast solution of l1-norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54(11):4789–4812, 2008. (Cited on p. 460.)
148. D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies. Data compression and harmonic analysis. *IEEE Trans. Inform. Theory*, 44(6):2435–2476, 1998. (Cited on p. 33.)
149. R. Dorfman. The detection of defective members of large populations. *Ann. Statist.*, 14:436–440, 1943. (Cited on p. 32.)
150. D.-Z. Du and F. Hwang. *Combinatorial group testing and its applications*. World Scientific, Singapore, 1993. (Cited on p. 32.)
151. M. Duarte, M. Davenport, D. Takhar, J. Laska, S. Ting, K. Kelly, and R. G. Baraniuk. Single-Pixel Imaging via Compressive Sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, March , 2008. (Cited on p. 32.)
152. R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967. (Cited on p. 241.)
153. E. Effros. A matrix convexity approach to some celebrated quantum inequalities. *Proc. Natl. Acad. Sci. USA*, 106(4):1006–1008, 2009. (Cited on pp. 514, 515.)
154. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. (Cited on p. 460.)
155. I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. SIAM, 1999. (Cited on pp. 489, 493.)
156. M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010. (Cited on p. 33.)
157. M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12):3736–3745, 2006. (Cited on p. 33.)
158. M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 48(9):2558–2567, 2002. (Cited on pp. 31, 33, 93, 383.)
159. Y. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inform. Theory*, 55(11):5302–5316, 2009. (Cited on p. 35.)
160. Y. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Inform. Theory*, 56(1):505–519, 2010. (Cited on pp. 35, 239, 408, 408.)
161. J. Ender. On compressive sensing applied to radar. *Signal Processing*, 90(5):1402 – 1414, 2010. (Cited on p. 33.)
162. H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Springer-Verlag, 1996. (Cited on pp. 34, 461.)
163. H. Epstein. Remarks on two theorems of E. Lieb. *Comm. Math. Phys.*, 31:317–325, 1973. (Cited on p. 510.)
164. A. Fannjiang, P. Yan, and T. Strohmer. Compressed remote sensing of sparse objects. *SIAM J. Imag. Sci.*, 3(3):596–618, 2010. (Cited on p. 33.)
165. M. Fazel. *Matrix rank minimization with applications*. PhD thesis, 2002. (Cited on p. 34.)
166. H. G. Feichtinger, F. Luef, and T. Werther. A guided tour from linear algebra to the foundations of Gabor analysis. In *Gabor and Wavelet Frames*, volume 10

of *Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, pages 1–49. World Sci. Publ., Hackensack, 2007. (Cited on p. 390.)
167. H. G. Feichtinger and T. Strohmer. *Gabor Analysis and Algorithms. Theory and Applications.* Birkhäuser, 1998. (Cited on p. 390.)
168. X. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École d'Été de Probabilités de Saint-Flour, IV-1974*, pages 1–96. Lecture Notes in Math., Vol. 480. Springer, 1975. (Cited on pp. 241, 242.)
169. X. Fernique. *Fonctions Al'eatoires Gaussiennes, Vecteurs Al'eatoires Gaussiens.* Université de Montréal Centre de Recherches Mathématiques, 1997. (Cited on p. 241.)
170. P. J. S. G. Ferreira and J. R. Higgins. The establishment of sampling as a scientific principle -A striking case of multiple discovery. *Notices of the American Mathematical Society*, 58(10):1446–1450, November 2011. (Cited on p. 32.)
171. M. A. T. Figueiredo, R. D. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Proces.*, 1(4):586–598, 2007. (Cited on p. 461.)
172. G. B. Folland. *Fourier Analysis and its Applications.* Wadsworth and Brooks, 1992. (Cited on pp. 382, 517.)
173. G. B. Folland. *A Course in Abstract Harmonic Analysis.* CRC Press, 1995. (Cited on p. 382.)
174. G. B. Folland and A. Sitaram. The uncertainty principle: A mathematical survey. *J. Fourier Anal. Appl.*, 3(3):207–238, 1997. (Cited on p. 383.)
175. M. Fornasier. Numerical methods for sparse recovery. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, pages 93–200. deGruyter, 2010. (Cited on p. 461.)
176. M. Fornasier and H. Rauhut. Iterative thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 25(2):187 – 208, 2008. (Cited on p. 460.)
177. M. Fornasier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008. (Cited on pp. 35, 460.)
178. M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21(4):1614–1640, 2011. (Cited on p. 461.)
179. S. Foucart. A note on guaranteed sparse recovery via ℓ_1-minimization. *Appl. Comput. Harmon. Anal.*, 29(1):97–103, 2010. (Cited on p. 153.)
180. S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *preprint*, 2010. (Cited on p. 153.)
181. S. Foucart. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In *Proceedings of the 13th International Conference on Approximation Theory*, 2010. (Cited on pp. 51, 153.)
182. S. Foucart. Stability and robustness of ℓ_1-minimizations with Weibull matrices and redundant dictionaries. *preprint*, 2012. (Cited on p. 331.)
183. S. Foucart and R. Gribonval. Real vs. complex null space properties for sparse vector recovery. *preprint*, 2009. (Cited on p. 93.)
184. S. Foucart and M. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.*, 26(3):395–407, 2009. (Cited on p. 153.)

185. S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. *J. Complexity*, 26(6):629–640, 2010. (Cited on pp. 299, 330.)
186. J. Friedman and W. Stuetzle. Projection pursuit regressions. *J. Amer. Statist. Soc.*, 76:817823, 1981. (Cited on p. 65.)
187. J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory*, 50(6):1341–1344, 2004. (Cited on pp. 31, 93.)
188. R. Garg and R. Khandekar. Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *Proc. 26th Intern. Conf. Machine Learning*, 2009. (Cited on p. 153.)
189. A. Garnaev and E. Gluskin. On widths of the Euclidean ball. *Sov. Math., Dokl.*, 30:200–204, 1984. (Cited on pp. 32, 299.)
190. D. Ge, X. Jiang, and Y. Ye. A note on complexity of l_p minimization. *Relation*, 10:9526, 2010. (Cited on p. 52.)
191. Q. Geng and J. Wright. On the local correctness of ℓ^1-minimization for dictionary learning. *preprint*, 2011. (Cited on p. 36.)
192. A. Gilbert and M. Strauss. Analysis of Data Streams. *Technometrics*, 49(3):346–356, 2007. (Cited on p. 35.)
193. A. C. Gilbert, S. Muthukrishnan, S. Guha, P. Indyk, and M. Strauss. Near-Optimal Sparse Fourier Representations via Sampling. In *Proc. STOC'02*, pages 152 – 161. Association for Computing Machinery, 2002. (Cited on pp. 32, 35, 392.)
194. A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Baltimore, MD, USA, January 12-14, 2003*, pages 243–252. SIAM and Association for Computing Machinery, New York, NY; Philadelphia, PA, 2003. (Cited on pp. 31, 33.)
195. A. C. Gilbert, M. Strauss, J. A. Tropp, and R. Vershynin. One sketch for all: Fast algorithms for compressed sensing. *preprint*, 2006. (Cited on pp. 32, 35, 51, 429.)
196. E. Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Math. USSR-Sb.*, 48:173–182, 1984. (Cited on p. 32.)
197. E. Gluskin. Extremal properties of orthogonal parallelepipeds and their applications to the geometry of Banach spaces. *Mat. Sb. (N.S.)*, 136(178)(1):85–96, 1988. (Cited on p. 331.)
198. G. Golub and C. F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd ed. edition, 1996. (Cited on pp. 389, 463, 481.)
199. R. Gopinath. Nonlinear recovery of sparse signals from narrowband data. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1995 - Volume 02*, ICASSP '95, page 3. IEEE Computer Society, 1995. (Cited on p. 52.)
200. Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50(4):265–289, 1985. (Cited on p. 242.)
201. Y. Gordon. Elliptically contoured distributions. *Probab. Theory Related Fields*, 76(4):429–438, 1987. (Cited on p. 242.)
202. Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric aspects of functional analysis (1986/87)*, volume 1317 of *Lecture Notes in Math.*, pages 84–106. Springer, Berlin, 1988. (Cited on p. 277.)

203. L. Grafakos. *Modern Fourier analysis*, volume 250 of *Graduate Texts in Mathematics*. Springer, Second edition, 2009. (Cited on pp. 382, 517.)
204. R. Graham and N. Sloane. Lower bounds for constant weight codes. *IEEE Trans. Inform. Theory*, 26(1):37–43, 1980. (Cited on p. 299.)
205. R. Gribonval. Sparse Decomposition of Stereo Signals with Matching Pursuit and Application to Blind Separation of more than Two Sources from a Stereo Mixture. May 2002. (Cited on p. 33.)
206. R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325, 2003. (Cited on pp. 31, 93, 115.)
207. R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Appl. Comput. Harmon. Anal.*, 22(3):335–355, 2007. (Cited on p. 93.)
208. R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Fourier Anal. Appl.*, 14(5):655–687, 2008. (Cited on pp. 35, 408, 408.)
209. G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, New York, Third edition, 2001. (Cited on p. 180.)
210. K. Gröchenig. *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, MA, 2001. (Cited on pp. 33, 390, 391.)
211. D. Gross. Recovering low-rank matrices from few coefficients in any basis. *preprint*, 2009. (Cited on pp. 34, 383, 383.)
212. D. Gross, Y.-K. Liu, S. FlammiaT., S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *preprint*, 2009. (Cited on p. 34.)
213. L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975. (Cited on p. 242.)
214. O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Rev. Mat. Iberoam.*, 24(3):1075–1095, 2008. (Cited on pp. 383, 389.)
215. C. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö. Yilmaz. Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements. *preprint*, 2010. (Cited on p. 35.)
216. S. Gurevich, R. Hadani, and N. Sochen. On some deterministic dictionaries supporting sparsity. *J. Fourier Anal. Appl.*, 14:859–876, 2008. (Cited on p. 115.)
217. V. Guruswani, C. Umans, and S. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. In *IEEE Conference on Computational Complexity*, pages 237–246, 2007. (Cited on p. 428.)
218. M. Haacke, R. Brown, M. Thompson, R. Venkatesan, M. Haacke, R. Brown, M. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss, June 1999. (Cited on p. 33.)
219. U. Haagerup. The best constants in the Khintchine inequality. *Studia Math.*, 70(3):231–283 (1982), 1981. (Cited on p. 239.)
220. T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Inform. Process. Lett.*, 33(6):305–308, 1990. (Cited on p. 181.)
221. J. Haldar, D. Hernando, and Z. Liang. Compressed-sensing MRI with random encoding. *IEEE Trans. Med. Imaging*, 30(4):893–903, 2011. (Cited on p. 33.)

222. E. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130, 2008. (Cited on p. 461.)
223. F. Hansen and G. Pedersen. Jensen's inequality for operators and Löwner's theorem. *Math. Ann.*, 258(3):229–241, 1982. (Cited on pp. 510, 511.)
224. F. Hansen and G. Pedersen. Jensen's operator inequality. *Bull. London Math. Soc.*, 35(4):553–564, 2003. (Cited on p. 512.)
225. D. Hanson and F. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42:1079–1083, 1971. (Cited on p. 241.)
226. H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Nearly optimal sparse Fourier transform. In *STOC*, 2012. (Cited on pp. 32, 35, 429.)
227. H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Simple and practical algorithm for sparse Fourier transform. In *SODA*, 2012. (Cited on pp. 32, 429.)
228. J. Haupt, W. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *preprint*, 2008. (Cited on p. 390.)
229. J. Haupt, W. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Trans. Inform. Theory*, 56(11):5862–5875, 2010. (Cited on p. 390.)
230. D. Healy Jr., D. Rockmore, P. Kostelec, and S. Sean. FFTs for the 2-Sphere - Improvements and Variations. *The Journal of Fourier Analysis and Applications*, 9:341–385, 1996. (Cited on p. 388.)
231. W. Hendee and C. Morgan. Magnetic resonance imaging Part I - Physical principles. *West J. Med.*, 141(4):491–500, 1984. (Cited on p. 33.)
232. M. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.*, 57(6):2275–2284, 2009. (Cited on pp. 33, 391, 391.)
233. J. R. Higgins. *Sampling Theory in Fourier and Signal Analysis: Foundations.* Clarendon Press, 1996. (Cited on pp. 32, 517.)
234. J. R. Higgins and R. L. Stens. *Sampling Theory in Fourier and Signal Analysis. Vol. 2: Advanced Topics.* Oxford University Press, 1999. (Cited on pp. 32, 517.)
235. N. J. Higham. *Functions of Matrices. Theory and Computation.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2008. (Cited on p. 463.)
236. A. Hinrichs and J. Vybiral. Johnson-Lindenstrauss lemma for circulant matrices. *Random Struct. Algorithms*, 39(3):391–398, 2011. (Cited on p. 392.)
237. J.-B. Hiriart Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis.* Grundlehren Text Editions. Springer-Verlag, 2001. (Cited on pp. 489, 491.)
238. W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. (Cited on p. 180.)
239. J. Högborn. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astronom. and Astrophys.*, 15:417, 1974. (Cited on p. 65.)
240. D. Holland, M. Bostock, L. Gladden, and D. Nietlispach. Fast multidimensional NMR spectroscopy using compressed sensing. *Angew. Chem. Int. Ed.*, 50(29):6548–6551, 2011. (Cited on p. 33.)
241. S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc. (N.S.)*, 43(4):439–561 (electronic), 2006. (Cited on p. 428.)

242. R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990. (Cited on p. 463.)
243. R. Horn and C. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994. (Cited on p. 463.)
244. W. Huffman and V. Pless. *Fundamentals of error-correcting codes*. Cambridge University Press, 2003. (Cited on p. 33.)
245. R. Hunt. On $L(p, q)$ spaces. *Enseignement Math. (2)*, 12:249–276, 1966. (Cited on p. 51.)
246. P. Indyk and A. Gilbert. Sparse recovery using sparse matrices. *Proc. IEEE*, 98(6):937 – 947, 2010. (Cited on pp. 35, 429.)
247. P. Indyk and M. Ržić. Near-optimal sparse recovery in the L1 norm. In *Proc. FOCS*, 2008. (Cited on p. 429.)
248. M. Iwen. Combinatorial sublinear-time Fourier algorithms. *Found. Comput. Math.*, 10(3):303–338, 2010. (Cited on pp. 32, 392, 429.)
249. M. Iwen. Improved approximation guarantees for sublinear-time Fourier algorithms. *preprint*, 2010. (Cited on pp. 32, 392.)
250. M. Iwen, A. Gilbert, and M. Strauss. Empirical evaluation of a sub-linear time sparse DFT algorithm. *Commun. Math. Sci.*, 5(4):981–998, 2007. (Cited on pp. 35, 392.)
251. L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *preprint*, 2011. (Cited on p. 36.)
252. S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Trans. Inform. Theory*, to appear. (Cited on p. 429.)
253. R. James, M. Dennis, and N. Daniel. Fast discrete polynomial transforms with applications to data analysis for distance transitive graphs. *SIAM J. Comput.*, 26(4):1066–1099, 1997. (Cited on p. 388.)
254. F. Jarre and J. Stoer. *Optimierung*. Springer, 2004. (Cited on pp. 489, 503.)
255. A. J. Jerri. The Shannon sampling theorem - its various extensions and applications: a tutorial review. *Proc. IEEE*, 65(11):1565–1596, 1977. (Cited on p. 32.)
256. W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., 1984. (Cited on p. 277.)
257. W. B. Johnson and J. Lindenstrauss, editors. *Handbook of the Geometry of Banach Spaces Vol I*. North-Holland Publishing Co., Amsterdam, 2001. (Cited on p. 239.)
258. S. Karlin. *Total Positivity Vol. I*. Stanford University Press, 1968. (Cited on p. 52.)
259. B. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Math. USSR, Izv.*, 11:317–333, 1977. (Cited on pp. 32, 300.)
260. A. Khintchine. Über dyadische Brüche. *Math. Z.*, 18(1):109–116, 1923. (Cited on p. 239.)
261. S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics. *IEEE J. Sel. Top. Signal Proces.*, 4(1):606–617, 2007. (Cited on p. 461.)

262. J. King. A minimal error conjugate gradient method for ill-posed problems. *J. Optim. Theory Appl.*, 60(2):297–304, 1989. (Cited on pp. 461, 481.)
263. T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005. (Cited on p. 243.)
264. H. König and S. Kwapień. Best Khintchine type inequalities for sums of independent, rotationally invariant random vectors. *Positivity*, 5(2):115–152, 2001. (Cited on p. 239.)
265. N. Kôno. Sample path properties of stochastic processes. *J. Math. Kyoto Univ.*, 20(2):295–313, 1980. (Cited on p. 241.)
266. F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. in preparation. (Cited on pp. 390, 391.)
267. F. Krahmer, G. E. Pfander, and P. Rashkov. Uncertainty principles for time–frequency representations on finite abelian groups. *Appl. Comput. Harmon. Anal.*, 25(2):209–225, 2008. (Cited on p. 390.)
268. F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011. (Cited on pp. 277, 392, 392.)
269. I. Krasikov. On the Erdelyi-Magnus-Nevai conjecture for Jacobi polynomials. *Constr. Approx.*, 28(2):113–125, 2008. (Cited on p. 388.)
270. M. A. Krasnosel'skij and Y. B. Rutitskij. *Convex Functions and Orlicz Spaces.* Groningen-The Netherlands: P. Noordhoff Ltd. IX, 249 p., 1961. (Cited on p. 242.)
271. J. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.*, 18(2):95–138, 1977. (Cited on p. 51.)
272. P. Kuppinger, G. Durisi, and H. Bölcskei. Uncertainty relations and sparse signal recovery for pairs of general signal sets. *preprint*, 2011. (Cited on p. 383.)
273. M.-J. Lai and L. Y. Liu. The null space property for sparse recovery from multiple measurement vectors. *Applied and Computational Harmonic Analysis*, 30:402–406, 2011. (Cited on p. 93.)
274. J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. *Appl. Comput. Harmon. Anal.*, 31(3):429–443, 2011. (Cited on p. 35.)
275. J. Lawrence, G. E. Pfander, and D. Walnut. Linear independence of Gabor systems in finite dimensional vector spaces. *J. Fourier Anal. Appl.*, 11(6):715–726, 2005. (Cited on p. 390.)
276. C. Lawson. *Contributions to the Theory of Linear Least Maximum Approximation.* PhD thesis, University of California Los Angeles, 1961. (Cited on p. 461.)
277. J. Lederer and S. van de Geer. The Bernstein-Orlicz norm and deviation inequalities. *preprint*, 2011. (Cited on p. 243.)
278. M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87, 1996. (Cited on p. 243.)
279. M. Ledoux. *The Concentration of Measure Phenomenon.* AMS, 2001. (Cited on pp. 242, 243, 243, 243.)
280. M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, Berlin, Heidelberg, NewYork, 1991. (Cited on pp. 180, 239, 239, 241, 241, 242, 242, 242, 243.)

281. E. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Math.*, 11:267–288, 1973. (Cited on pp. 509, 510.)
282. G. Lindblad. Expectations and entropy inequalities for finite quantum systems. *Comm. Math. Phys.*, 39:111–119, 1974. (Cited on p. 515.)
283. P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16:964–979, 1979. (Cited on p. 460.)
284. A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195(2):491–523, 2005. (Cited on p. 331.)
285. Y. Liu. Universal low-rank matrix recovery from Pauli measurements. *preprint*, 2011. (Cited on p. 34.)
286. A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind audiovisual source separation based on sparse redundant representations. *IEEE Trans. Multimed.*, 12(5):358–371, August 2010. (Cited on p. 33.)
287. B. Logan. *Properties of High-Pass Signals*. PhD thesis, Columbia University, 1965. (Cited on p. 31.)
288. G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive approximation: advanced problems*. Springer, 1996. (Cited on p. 299.)
289. F. Lust-Piquard. Inégalites de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Acad. Sci. Paris S'er. I Math.*, 303:289–292, 1986. (Cited on p. 240.)
290. F. Lust-Piquard and G. Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991. (Cited on p. 240.)
291. M. Lustig, D. Donoho, and J. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007. (Cited on p. 33.)
292. L. Mackey, M. Jordan, R. Chen, B. Farrell, and J. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *preprint*, 2012. (Cited on p. 240.)
293. S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998. (Cited on p. 386.)
294. S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993. (Cited on pp. 31, 33, 33, 65.)
295. M. Marcus and L. Shepp. Sample behavior of Gaussian processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 423–441. Univ. California Press, 1972. (Cited on p. 242.)
296. S. Marple. *Digital Spectral Analysis with Applications*. Prentice - Hall, 1987. (Cited on pp. 31, 52.)
297. P. Massart. Rates of convergence in the central limit theorem for empirical processes. *Ann. Inst. H. Poincar'e Probab. Statist.*, 22(4):381–423, 1986. (Cited on p. 243.)
298. P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000. (Cited on p. 243.)
299. P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007. (Cited on pp. 241, 242, 242, 242, 242, 243.)
300. C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 195–248. Springer, 1998. (Cited on p. 181.)

301. S. Mendelson, A. Pajor, and M. Rudelson. The geometry of random $-1, 1$-polytopes. *Discr. Comput. Geom.*, 34(3):365–379, 2005. (Cited on p. 299.)
302. S. Mendelson, A. Pajor, and N. Tomczak Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2009. (Cited on p. 277.)
303. D. Middleton. Channel Modeling and Threshold Signal Processing in Underwater Acoustics: An Analytical Overview. *IEEE J. Oceanic Eng.*, 12(1):4–28, 1987. (Cited on p. 391.)
304. M. Mishali and Y. C. Eldar. From theory to practice: Sub-nyquist sampling of sparse wideband analog signals. *IEEE J. Sel. Topics Sig. Process.*, 4(2):375–391, April 2010. (Cited on p. 34.)
305. Q. Mo and S. Li. New bounds on the restricted isometry constant δ_{2k}. *Appl. Comput. Harmon. Anal.*, in press, 2011. DOI: 10.1016/j.acha.2011.04.005. (Cited on p. 153.)
306. Q. Mo and Y. Shen. Remarks on the restricted isometry property in orthogonal matching pursuit algorithm. *Preprint*, 2011. (Cited on p. 153.)
307. S. Montgomery Smith. The distribution of Rademacher sums. *Proc. Amer. Math. Soc.*, 109(2):517–522, 1990. (Cited on p. 331.)
308. T. K. Moon and W. C. Stirling. *Mathematical Methods and Algorithms for Signal Processing*. Upper Saddle River, NJ: Prentice Hall, 2000. (Cited on p. 52.)
309. M. Murphy, M. Alley, J. Demmel, K. Keutzer, S. Vasanawala, and M. Lustig. Fast ℓ_1-SPIRiT compressed sensing parallel imaging MRI: Scalable parallel implementation and clinically feasible runtime. *IEEE Trans. Med. Imaging*, to appear. (Cited on p. 33.)
310. B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1995. (Cited on pp. 31, 33, 52.)
311. F. Nazarov and A. Podkorytov. Ball, Haagerup, and distribution functions. In *Complex Analysis, Operators, and related Topics*, volume 113 of *Oper. Theory Adv. Appl.*, pages 247–267. Birkhäuser, 2000. (Cited on p. 239.)
312. D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26(3):301–321, 2008. (Cited on pp. 65, 154.)
313. D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.*, 9(3):317–334, 2009. (Cited on pp. 65, 154.)
314. D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J. Sel. Topics Sig. Process.*, 4(2):310 – 316, April 2010. (Cited on pp. 65, 154.)
315. D. Needell and R. Ward. Stable image reconstruction using total variation minimization. *preprint*, 2012. (Cited on p. 34.)
316. Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005. (Cited on p. 460.)
317. N. Noam and W. Avi. Hardness vs randomness. *Journal of Computer and System Sciences*, 49(2):149 – 167, 1994. (Cited on p. 299.)
318. J. Nocedal and S. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, Second edition, 2006. (Cited on pp. 64, 460, 462, 462.)
319. E. Novak. Optimal recovery and n-widths for convex classes of functions. *J. Approx. Theory*, 80(3):390–408, 1995. (Cited on pp. 32, 299.)

320. E. Novak and H. Woźniakowski. EMS Publishing House, 2008. (Cited on p. 299.)
321. M. Ohya and D. Petz. *Quantum entropy and its use*. Texts and monographs in physics. Springer, 2004. (Cited on p. 514.)
322. R. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *preprint*, 2009. (Cited on p. 240.)
323. R. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *preprint*, 2010. (Cited on p. 240.)
324. M. Osborne. *Finite Algorithms in Optimization and Data Analysis*. John Wiley & Sons, 1985. (Cited on p. 461.)
325. M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000. (Cited on p. 460.)
326. M. Osborne, B. Presnell, and B. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000. (Cited on p. 460.)
327. Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal Matching Pursuit: Recursive Function Approximation with with Applications to Wavelet Decomposition. In *1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, Nov. 1-3, 1993.*, pages 40 – 44. others, 1993. (Cited on p. 65.)
328. G. Peškir. Best constants in Kahane-Khintchine inequalities for complex Steinhaus functions. *Proc. Amer. Math. Soc.*, 123(10):3101–3111, 1995. (Cited on p. 239.)
329. G. Peškir and A. N. Shiryaev. The Khintchine inequalities and martingale expanding sphere of their action. *Russian Math. Surveys*, 50(5):849–904, 1995. (Cited on p. 239.)
330. D. Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory (Warsaw, 1992)*, volume 30 of *Banach Center Publ.*, pages 287–298. Polish Acad. Sci., 1994. (Cited on p. 486.)
331. D. Petz. *Quantum information theory and quantum statistics*. Theoretical and mathematical physics. Springer, 2008. (Cited on p. 514.)
332. G. Pfander, H. Rauhut, and J. Tropp. The restricted isometry property for time-frequency structured random matrices. *preprint*, 2011. (Cited on pp. 33, 391.)
333. G. E. Pfander, H. Rauhut, and J. Tanner. Identification of matrices having a sparse representation. *IEEE Trans. Signal Process.*, 56(11):5376–5388, 2008. (Cited on pp. 33, 391, 391.)
334. A. Pinkus. *n-Widths in Approximation Theory*. Springer-Verlag, Berlin, 1985. (Cited on p. 299.)
335. A. Pinkus. *On L^1-Approximation*, volume 93 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1989. (Cited on p. 93.)
336. A. Pinkus. *Totally Positive Matrices*, volume 181 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2010. (Cited on p. 52.)
337. M. Pinsky. *Introduction to Fourier analysis and wavelets*. Graduate Studies in Mathematics 102. Providence, RI: American Mathematical Society (AMS), 2009. (Cited on pp. 382, 517.)
338. G. Pisier. Conditions d'entropie assurant la continuité de certains processus et applications à l'analyse harmonique. In *Seminar on Functional Analysis, 1979–1980 (French)*, pages Exp. No. 13–14, 43. École Polytech., 1980. (Cited on p. 241.)

339. G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics. Cambridge University Press, 1999. (Cited on p. 241.)
340. Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *preprint*, 2011. (Cited on p. 36.)
341. Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *preprint*, 2012. (Cited on p. 36.)
342. T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *ICCV Proceedings*. Springer, 2009. (Cited on p. 460.)
343. L. Potter, E. Ertin, J. Parker, and M. Cetin. Sparsity and compressed sensing in radar imaging,. *Proc. IEEE*, 98(6):1006 –1020, 2010. (Cited on p. 33.)
344. D. Potts, G. Steidl, and M. Tasche. Fast algorithms for discrete polynomial transforms. *Math. Comp.*, 67:1577–1590, oct 1998. (Cited on p. 388.)
345. D. Potts, G. Steidl, and M. Tasche. Fast Fourier Transforms for Nonequispaced Data: A Tutorial. In J. Benedetto and P. Ferreira, editors, *Modern Sampling Theory: Mathematics and Applications*, chapter 12, pages 247 – 270. Birkhäuser, 2001. (Cited on pp. 382, 517.)
346. D. Potts and M. Tasche. Parameter estimation for exponential sums by approximate prony method. *Signal Process.*, 90(5):1631–1642, 2010. (Cited on pp. 31, 52.)
347. R. Prony. Essai expérimental et analytique sur les lois de la Dilatabilité des uides élastique et sur celles de la Force expansive de la vapeur de leau et de la vapeur de lalkool, à différentes températures. *J. École Polytechnique*, 1:24–76, 1795. (Cited on pp. 31, 52.)
348. W. Pusz and S. Woronowicz. Form convex functions and the WYDL and other inequalities. *Lett. Math. Phys.*, 2(6):505–512, 1977/78. (Cited on p. 515.)
349. S. Qian and D. Chen. Signal representation using adaptive normalized Gaussian functions. *Signal Process.*, 36(1):1–11, 1994. (Cited on p. 65.)
350. R. Ramlau and G. Teschke. Sparse recovery in inverse problems. In *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Ser. Comput. Appl. Math.*, pages 201–262. Walter de Gruyter, Berlin, 2010. (Cited on p. 34.)
351. M. Raphan and E. Simoncelli. Optimal denoising in redundant representation. *IEEE Trans. Image Process.*, 17(8):1342–1352, August 2008. (Cited on p. 33.)
352. H. Rauhut. Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.*, 22(1):16–42, 2007. (Cited on pp. 32, 384, 384.)
353. H. Rauhut. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process.*, 7(2):197–215, 2008. (Cited on pp. 32, 153, 384.)
354. H. Rauhut. Circulant and Toeplitz matrices in compressed sensing. In *Proc. SPARS'09*, Saint-Malo, France, 2009. (Cited on p. 390.)
355. H. Rauhut. Compressive Sensing and Structured Random Matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series Comp. Appl. Math.*, pages 1–92. deGruyter, 2010. (Cited on pp. 32, 180, 240, 240, 240, 241, 241, 383, 384, 390.)
356. H. Rauhut and G. E. Pfander. Sparsity in time-frequency representations. *J. Fourier Anal. Appl.*, 16(2):233–260, 2010. (Cited on pp. 33, 390, 391.)

357. H. Rauhut, J. K. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmon. Anal.*, 32(2):242–254, 2012. (Cited on pp. 390, 392.)
358. H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Trans. Inform. Theory*, 54(5):2210 – 2219, 2008. (Cited on p. 277.)
359. H. Rauhut and R. Ward. Sparse recovery for spherical harmonic expansions. In *Proc. SampTA 2011, Singapore*, 2011. (Cited on p. 388.)
360. H. Rauhut and R. Ward. Sparse Legendre expansions via ℓ_1-minimization. *J. Approx. Theory*, 164(5):517533, 2012. (Cited on pp. 32, 386, 388.)
361. B. Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, to appear. (Cited on pp. 34, 383.)
362. B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010. (Cited on pp. 34, 277, 535.)
363. J. Renes, R. Blume Kohout, A. Scott, and C. Caves. Symmetric informationally complete quantum measurements. *J. Math. Phys.*, 45(6):2171–2180, 2004. (Cited on p. 116.)
364. E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Related Fields*, 119(2):163–175, 2001. (Cited on p. 243.)
365. E. Rio. Une inégalité de Bennett pour les maxima de processus empiriques. *Ann. Inst. H. Poincar'e Probab. Statist.*, 38(6):1053–1057, 2002. (Cited on p. 243.)
366. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, reprint edition, 1997. (Cited on pp. 489, 491, 496, 509, 509.)
367. R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental*. Springer-Verlag, 1998. (Cited on p. 489.)
368. J. Rohn. Computing the norm $\|A\|_{\infty,1}$ is NP-hard. *Linear and Multilinear Algebra*, 47(3):195–204, 2000. (Cited on p. 468.)
369. S. Ross. *Introduction to Probability Models*. Academic Press, Nineth edition, 2006. (Cited on p. 180.)
370. R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: learning sparse dictionaries for sparse signal approximation. *IEEE Trans. Signal Process.*, 58(3, part 2):1553–1564, 2010. (Cited on p. 36.)
371. M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999. (Cited on p. 240.)
372. M. Rudelson and R. Vershynin. Geometric approach to error-correcting codes and reconstruction of signals. *Internat. Math. Res. Notices*, (64):4019–4041, 2005. (Cited on p. 33.)
373. M. Rudelson and R. Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *J. ACM*, 54(4):Art. 21, 19 pp. (electronic), 2007. (Cited on p. 408.)
374. M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008. (Cited on pp. 277, 384, 384.)
375. M. Rudelson and R. Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Adv. Math.*, 218(2):600–633, 2008. (Cited on p. 278.)

376. M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians*, volume III, pages 1576–1602. Hindustan Book Agency, 2010. (Cited on p. 278.)
377. L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992. (Cited on pp. 31, 34.)
378. W. Rudin. *Fourier Analysis on Groups*. Interscience Publishers, 1962. (Cited on p. 382.)
379. M. Ruskai. Inequalities for quantum entropy: a review with conditions for equality. *J. Math. Phys.*, 43(9):4358–4375, 2002. (Cited on p. 510.)
380. M. Ruskai. Erratum: "Inequalities for quantum entropy: a review with conditions for equality". *J. Math. Phys.*, 46(1):019901, 1, 2005. (Cited on p. 510.)
381. F. Santosa and W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7(4):1307–1330, 1986. (Cited on pp. 31, 34.)
382. G. Schechtman. Special orthogonal splittings of L_1^{2k}. *Israel J. Math.*, 139:337–347, 2004. (Cited on p. 300.)
383. K. Schnass and R. Gribonval. Dictionary identification - sparse matrix-factorisation via l_1-minimisation. *IEEE Trans. Inform. Theory*, 56(7):3523–3539, 2010. (Cited on p. 36.)
384. K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Trans. Signal Process.*, 56(5):1994–2002, 2008. (Cited on p. 115.)
385. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002. (Cited on p. 34.)
386. I. Segal and M. Iwen. Improved sparse Fourier approximation results: Faster implementations and stronger guarantees. *preprint*, 2012. (Cited on p. 392.)
387. Y. Shrot and L. Frydman. Compressed sensing and the reconstruction of ultrafast 2D NMR data: Principles and biomolecular applications. *J. Magn. Reson.*, 209(2):352–358, 2011. (Cited on p. 33.)
388. D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41:463–501, 1962. (Cited on p. 242.)
389. D. Spielman and N. Srivastava. An elementary proof of the restricted invertibility theorem. *Israel J. Math.*, to appear. (Cited on p. 408.)
390. J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse image and signal processing Wavelets, Curvelets, Morphological Diversity*. Cambridge: Cambridge University Press. xvii, 316 p., 2010. (Cited on p. 33.)
391. E. M. Stein and R. Shakarchi. *Functional Analysis: Introduction to further Topics in Analysis*. 2011. (Cited on pp. 382, 517.)
392. M. Stojanovic. Underwater Acoustic Communications. In M. Stojanovic and J. G. Webster, editors, *Encyclopedia of Electrical and Electronics Engineering*, volume 22, pages 688–698. John Wiley & Sons, 1999. (Cited on p. 391.)
393. M. Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *Preprint*, 2009. arXiv:0907.3666. (Cited on p. 277.)
394. T. Strohmer and B. Friedlander. Analysis of sparse MIMO radar. *preprint*, 2012. (Cited on p. 33.)
395. T. Strohmer and R. W. j. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003. (Cited on p. 115.)

396. S. Szarek. On Kashin's almost Euclidean orthogonal decomposition of l_n^1. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.*, 26(8):691–694, 1978. (Cited on p. 300.)
397. G. Szegő. *Orthogonal polynomials*. American Mathematical Society, Fourth edition, 1975. (Cited on pp. 386, 388.)
398. M. Talagrand. Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *Ann. Probab.*, 17(4):1546–1570, 1989. (Cited on p. 243.)
399. M. Talagrand. A new look at independence. *Ann. Probab.*, 24(1):1–34, 1996. (Cited on p. 242.)
400. M. Talagrand. Majorizing measures: the generic chaining. *Ann. Probab.*, 24(3):1049–1103, 1996. (Cited on p. 241.)
401. M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996. (Cited on p. 243.)
402. M. Talagrand. Selecting a proportion of characters. *Israel J. Math.*, 108:173–191, 1998. (Cited on pp. 383, 388, 388.)
403. M. Talagrand. Majorizing measures without measures. *Ann. Probab.*, 29(1):411–417, 2001. (Cited on p. 241.)
404. M. Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer-Verlag, 2005. (Cited on p. 241.)
405. M. Talagrand. *Mean Field Models for Spin Glasses. Volume I: Basic Examples*. Springer, 2010. (Cited on p. 242.)
406. G. Tauböck, F. Hlawatsch, D. Eiwen, and H. Rauhut. Compressive Estimation of Doubly Selective Channels in Multicarrier Systems: Leakage Effects and Sparsity-Enhancing Processing. *IEEE J. Sel. Topics Sig. Process.*, 4(2):255–271, 2010. (Cited on p. 34.)
407. H. Taylor, S. Banks, and J. McCoy. Deconvolution with the ℓ_1-norm. *Geophys. J. Internat.*, 44(1):39–52, 1979. (Cited on pp. 31, 34.)
408. V. Temlyakov. Nonlinear methods of approximation. *Found. Comput. Math.*, 3(1):33–107, 2003. (Cited on p. 65.)
409. V. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235–409, 2008. (Cited on p. 65.)
410. V. Temlyakov. *Greedy approximation*. Cambridge Monographs on Applied and Computational Mathematics (No. 20). Cambridge University Press, 2011. (Cited on pp. 65, 115.)
411. R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. (Cited on pp. 31, 34, 64.)
412. J. Traub, G. Wasilkowski, and H. Wo'zniakowski. *Information-based complexity*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1988. With contributions by A. G. Werschulz and T. Boult. (Cited on p. 31.)
413. L. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 2000. (Cited on p. 463.)
414. J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004. (Cited on pp. 31, 33, 65, 115.)
415. J. A. Tropp. Recovery of short, complex linear combinations via l_1 minimization. *IEEE Trans. Inform. Theory*, 51(4):1568–1570, 2005. (Cited on p. 93.)
416. J. A. Tropp. Algorithms for simultaneous sparse approximation: part II: Convex relaxation. *Signal Processing*, 86(3):589 – 602, 2006. (Cited on p. 35.)

417. J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 51(3):1030–1051, 2006. (Cited on pp. 31, 33.)
418. J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R., Math., Acad. Sci. Paris*, 346(23-24):1271–1274, 2008. (Cited on pp. 240, 407, 408.)
419. J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25:1–24, 2008. (Cited on pp. 239, 240, 383, 389, 391, 407, 407, 408, 408.)
420. J. A. Tropp. On the linear independence of spikes and sines. *J. Fourier Anal. Appl.*, 14(5-6):838–858, 2008. (Cited on pp. 33, 408.)
421. J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008. (Cited on p. 408.)
422. J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *ACM-SIAM Symp. Discrete Algorithms (SODA)*, pages 978–986, 2009. (Cited on p. 409.)
423. J. A. Tropp. From the joint convexity of quantum relative entropy to a concavity theorem of Lieb. *Proc. Amer. Math. Soc.*, to appear. (Cited on p. 510.)
424. J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, to appear. (Cited on pp. 240, 407.)
425. J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation: part I: Greedy pursuit. *Signal Processing*, 86(3):572 – 588, 2006. (Cited on p. 35.)
426. J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *IEEE Trans. Inform. Theory*, 56(1):520 –544, 2010. (Cited on pp. 34, 391, 392, 392.)
427. J. A. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk. Random filters for compressive sampling and reconstruction. *Proc. 2006 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 3:872–875, 2006. (Cited on p. 390.)
428. M. Tygert. Fast algorithms for spherical harmonic expansions, II. *J. Comput. Phys.*, 227(8):4260–4279, 2008. (Cited on p. 388.)
429. A. Uhlmann. Relative entropy and the Wigner-Yanase-Dyson-Lieb concavity in an interpolation theory. *Comm. Math. Phys.*, 54(1):21–32, 1977. (Cited on p. 515.)
430. E. van den Berg and M. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008. (Cited on p. 461.)
431. C. F. Van Loan. *Computational Frameworks for the Fast Fourier Transform.* SIAM, 1992. (Cited on p. 519.)
432. S. Varadhan. Large deviations and applications. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Math.*, pages 1–49. Springer, 1988. (Cited on pp. 180, 181.)
433. R. Varga. *Gershgorin and his Circles.* Springer Series in Computational Mathematics. Springer-Verlag, 2004. (Cited on p. 471.)
434. S. Vasanawala, M. Alley, B. Hargreaves, R. Barth, J. Pauly, and M. Lustig. Improved pediatric MR imaging with compressed sensing. *Radiology*, 256(2):607–616, 2010. (Cited on p. 33.)

435. A. Vershik and P. Sporyshev. Asymptotic behavior of the number of faces of random polyhedra and the neighborliness problem. *Sel. Math. Sov.*, 11(2):181–201, 1992. (Cited on p. 280.)
436. R. Vershynin. John's decompositions: selecting a large part. *Israel J. Math.*, 122:253–277, 2001. (Cited on p. 409.)
437. R. Vershynin. Frame expansions with erasures: an approach through the non-commutative operator theory. *Appl. Comput. Harmon. Anal.*, 18(2):167–176, 2005. (Cited on p. 240.)
438. R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge Univ Press, to appear. (Cited on pp. 181, 278.)
439. J. Vybiral. A variant of the Johnson-Lindenstrauss lemma for circulant matrices. *J. Funct. Anal.*, 260(4):1096–1105, 2011. (Cited on p. 392.)
440. M. Wakin. *The Geometry of Low-dimensional Signal Models*. PhD thesis, Rice University, 2006. (Cited on p. 51.)
441. S. Waldron. *An Introduction to Finite Tight Frames*. Birkhäuser Boston. (Cited on p. 116.)
442. J. Walker. Fourier analysis and wavelet analysis. *Notices Amer. Math. Soc.*, 44(6):658–670, 1997. (Cited on pp. 382, 517.)
443. J. S. Walker. *Fast Fourier transforms*. CRC Press, 1991. (Cited on p. 519.)
444. Y. Wiaux, L. Jacques, G. Puy, A. Scaife, and P. Vandergheynst. Compressed sensing imaging techniques for radio interferometry. *Monthly Notices of the Royal Astronomical Society*, 395(3):1733–1742, 2009. (Cited on p. 34.)
445. P. Wojtaszczyk. *A Mathematical Introduction to Wavelets*. Cambridge University Press, 1997. (Cited on pp. 33, 385, 386.)
446. P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *Found. Comput. Math.*, 10:1–13, 2010. (Cited on pp. 330, 331, 331.)
447. P. Wojtaszczyk. ℓ_1 minimisation with noisy data. *SIAM J. Numer. Anal.*, 50(2):458–467, 2012. (Cited on p. 331.)
448. S. Worm. Iteratively re-weighted least squares for compressed sensing, 2011. Diploma thesis, University of Bonn. (Cited on p. 461.)
449. G. Wright. Magnetic resonance imaging. *IEEE Signal Processing Magazine Magazine*, 14(1):56–66, 1997. (Cited on p. 33.)
450. J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. *IEEE Trans. Pattern Anal. Machine Intelligence*, 31(2):210 –227, 2009. (Cited on p. 34.)
451. T. Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inform. Theory*, 57(9):6215 – 6221, 2011. (Cited on p. 154.)
452. X. Zhang, M. Burger, X. Bresson, and S. Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM J. Imaging Sci.*, 3(3):253–276, 2010. (Cited on p. 460.)
453. X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *J. Sci. Comput.*, 46:20–46, 2011. (Cited on p. 460.)
454. M. Zhu and T. Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. Technical report, 2008. CAM Report 08–34, UCLA, Los Angeles, CA (2008). (Cited on p. 460.)

455. J. Zou, A. C. Gilbert, M. Strauss, and I. Daubechies. Theoretical and experimental analysis of a randomized algorithm for sparse Fourier transform analysis. *J. Comput. Phys.*, 211:572–595, 2005. (Cited on pp. 32, 392.)
456. A. Zymnis, S. Boyd, and E. J. Candés. Compressed sensing with quantized measurements. *IEEE Signal Process. Letters*, 17(2):149 –152, 2010. (Cited on p. 35.)

Index