Random encoding of quantized finite frame expansions

Mark Iwen[a] and Rayan Saab[b]

[a] Department of Mathematics & Department of Electrical and Computer Engineering, Michigan State University;
[b]Department of Mathematics, University of California, San Diego.

ABSTRACT

Frames, which in finite dimensions are spanning sets of vectors, generalize the notion of bases and provide a useful tool for modeling the measurement (or sampling) process in several modern signal processing applications. In the digital era, the measurement process is typically followed by a quantization, or digitization step that allows for storage, transmission, and processing using digital devices. One family of quantization methods, popular for its robustness to errors caused by circuit imperfections and for its ability to act on the measurements progressively, is Sigma-Delta quantization. In the finite frame setting, Sigma-Delta quantization, unlike scalar quantization, has recently been shown to exploit the redundancy in the measurement process leading to a more efficient rate-distortion performance. Nevertheless, on its own, it is not known whether Sigma-Delta quantization can provide *optimal* rate-distortion performance. In this note, we show that a simple post-processing step consisting of a discrete, random Johnson-Lindenstrauss embedding of the resulting bit-stream yields near-optimal rate distortion performance, with high probability. In other words, it near optimally compresses the resulting bit-stream. Our result holds for a wide variety of frames, including smooth frames and random frames.

Keywords: Frames, quantization, random matrices, Sigma-Delta

1. INTRODUCTION

A collection of vectors $\{f_i\}_{i=1}^N \subset \mathbb{R}^d$, where N is finite, is a finite frame for $\mathbb{R}^d$ if there are constants $0 < A \leq B < \infty$ so that for any $x \in \mathbb{R}^d$, we have

$$A\|x\|_2^2 \leq \sum_{i=1}^{N} |\langle f_i, x\rangle|^2 \leq B\|x\|_2^2. \tag{1}$$

Denote by f_i the rows of the $N \times d$ *frame matrix* F, and note that $\sum_{i=1}^N |\langle f_i, x\rangle|^2 = \|Fx\|_2^2$. Thus, (1) implies that for $\{f_i\}_{i=1}^N$ to be a frame, it is necessary and sufficient for F to be full rank. Finite frames have been investigated as research topics in their own right, but also in the context of applications, where their redundancy ($N \geq d$) has proven particularly useful. Among other areas, they have found application in digital transmission(e.g., Refs. 2, 3), phase retrieval from magnitude data (e.g., Refs. 4, 5), and compressed sensing.[6,7] They have proven particularly useful at modeling the sampling (i.e., measurement) process in data acquisition systems. For example, in imaging applications, multiplex systems[8] collect linear combinations $y_i := \langle f_i, x\rangle$, $i \in \{1, ..., N\}$ of the pixels of interest x, viewed as a vector in $\mathbb{R}^d$. When one collects more measurements than the ambient dimension of the signal, the collection $\{f_i\}_{i=1}^N$ may constitute a frame and the vector of measurements can often be expressed as

$$y := Fx \in \mathbb{R}^N. \tag{2}$$

For example, imaging systems where the vector of measurements can the be represented using (2) have been devised using coded apertures (see, e.g., Ref. 9), as well as digital micro-mirror arrays (e.g., Ref. 10). Similarly, systems that acquire finite dimensional signals using filter banks also allow their measurement process to be modeled via (2).[11]

Further author information:
M.I.: E-mail: iwenmark@msu.edu R.S.: E-mail: rsaab@ucsd.edu.
An expanded version of this note,[1] including proofs, will appear in the Journal of Fourier Analysis and Applications.

While (2) provides an idealized model of the measurement process, it does not take into account the process of digitizing the measurements. This digitialization process consists of replacing elements of the vector y with elements from a finite set called the quantization alphabet, which we refer to as $\mathcal{A}$, and then assigning binary codewords to the elements of $\mathcal{A}^N$. Since the resulting vector is of length N, we require at most $N \log_2 |\mathcal{A}|$ bits to represent a quantized vector of measurements. Of course, such a quantization process is inherently lossy, as the quantization map

$$\mathcal{Q} : \mathbb{R}^N \to \mathcal{A}^N \tag{3}$$

is a many-to-one map. Even by tayloring the quantization process to the finite-frame scenario where the signals x belong to a compact set $\mathcal{X} \subset \mathbb{R}^d$, (4) becomes

$$\mathcal{Q} : F\mathcal{X} \to \mathcal{A}^N, \tag{4}$$

but remains generally a many-to-one map. In this paper, we are interested in $\mathcal{X} = B_2^d$, the Euclidean ball in d-dimensions.

Despite the inherently lossy nature of quantization, one still seeks reconstruction schemes

$$\Delta : \mathcal{A}^N \to \mathbb{R}^d \tag{5}$$

that allow for good approximation of a signal x from its quantized vector of measurements $q := \mathcal{Q}(Fx)$. In particular, one seeks quantization schemes and reconstruction maps whereby the error decreases as the resources committed to the measurement and/or quantization process increase. For example, a reasonable requirement is for the error to decreases as one increases the size of the alphabet $\mathcal{A}$ and the number of measurements N. In this paper, we assume a fixed alphabet $\mathcal{A}$ and are interested in the tradeoff between the number of bits associated with increasing the number of measurements, and the reconstruction error.

Additionally, in many applications, practical considerations dictate that one must work with quantization schemes that act progressively on the measurements. In other words, the quantization scheme does not have access to the full vector of measurements y. Rather, at time i, it only has access to y_i and possibly to some of its immediate predecessors and to state variables that are updated on the fly. This practical requirement is generally due to the difficulty associated with building circuitry that accurately stores analog quantities for long periods of time. Importantly, such a constraint precludes the possibility of employing vector quantizers that require access to the entire vector y. For example, one cannot apply a left inverse of F to y, thereby recovering x, and then proceed to encode x directly.

Sigma-Delta ($\Sigma\Delta$) quantization schemes (see, e.g., Refs.[12–14]) are a family of quantization schemes that satisfy the practical constraint of operating on the measurements progressively. At time i, they act on y_i as well as previous values of a state variable u to produce $q_i \in \mathcal{A}$. Furthermore, in the finite frame setting these schemes have been shown to exploit the redundancy associated with increasing the number of measurements.[15–18] On the other hand, volume arguments show that in the $\mathcal{X} = B_2^d$ case, the optimal relationship between the number of bits $\mathcal{R}$ needed to encode the ball and the associated (worst case) error $\mathcal{D}$ over all $x \in B_2^d$ is given by $\mathcal{D} = O(e^{-c\mathcal{R}/d})$ for some absolute constant c.[19] This optimal rate-distortion relationship can be achieved by directly encoding the vectors x, see Ref. 1 for more details. In many scenarios, one does not have direct access to x (for example when one can only measure $y = Fx$ for some $N \times d$ matrix F with $N \geq d$), and so is unable to recover x before the digitization process. In such a case, one must quantize first and recover an estimate of x later. Of course, ideally, one wishes to attain the optimal rate-distortion relationship in this case as well. Unfortunately, this optimal relationship is not known to be attained for $\Sigma\Delta$ schemes under the constraints of this paper: a fixed alphabet, and the inability to invert the frame matrix F prior to quantization. The best known rate-distortion relationship, for progressive quantizers is achieved with $\Sigma\Delta$ schemes and is of the form $\mathcal{D} = O(e^{-c\sqrt{\mathcal{R}/d}})$.[18, 20]

2. CONTRIBUTIONS

To help remedy this situation and bridge the gap to the optimal performance, this paper introduces a potentially *lossy encoding* stage, consisting of the map

$$\mathcal{E} : \mathcal{A}^N \to \mathcal{C},$$

where $\mathcal{C}$ is such that $|\mathcal{C}| \ll |\mathcal{A}^N|$. Thus, $\log_2 |\mathcal{C}|$ bits are sufficient for digitally representing the output of this encoder. To accommodate the additional encoding, the reconstruction is modified to approximate x directly from $\mathcal{C}$. In particular, we present a decoder

$$\Delta : \mathcal{C} \to \mathbb{R}^d,$$

where both the proposed decoder, Δ, and the proposed encoding map, $\mathcal{E}$, are *linear*, hence computationally efficient.

For stable $\Sigma\Delta$ quantization schemes, we show that there exists an encoding scheme $\mathcal{E}$ acting on the output $\mathcal{Q}(Fx)$ of the quantization, and a decoding scheme Δ, such that

$$\left.\begin{array}{lcl} \mathcal{D}_{\Sigma\Delta} & := & \max\limits_{x \in B_2^d} \left\| x - \Delta\Big(\mathcal{E}\big(\mathcal{Q}(Fx)\big)\Big)\right\|_2 \leq CN^{-\alpha} \\ \mathcal{R}_{\Sigma\Delta} & := & \ln|\mathcal{C}| \leq C' d \ln N, \end{array}\right\} \Longrightarrow \mathcal{D}_{\Sigma\Delta} \leq \exp\left(-c\frac{\mathcal{R}_{\Sigma\Delta}}{d}\right).$$

where α, C, C', and c are positive constants that depend on the $\Sigma\Delta$ scheme and d. More specifically:

1. We show that there exist frames (the Sobolev self-dual frames), for which encoding by random subsampling of the integrated $\Sigma\Delta$ bit-stream (and labeling the output) yields an *essentially optimal rate-distortion tradeoff up to logarithmic factors of d.*

2. We show that random *Bernoulli* matrices in $\mathbb{R}^{m \times d}$, with $m \approx d$, are *universal* encoders. Provided one has a good frame for $\Sigma\Delta$ quantization, such Bernoulli matrices yield an *optimal rate-distortion tradeoff, up to constants.*

3. We show that in both cases above, the decoding can be done linearly and we provide an explicit expression for the decoder.

These contributions are quantified and made explicit in Theorems 1 and 2.

3. SIGMA-DELTA QUANTIZATION

For concreteness, we now define some terminology and introduce $\Sigma\Delta$ quantization.

Fix an alphabet $\mathcal{A}$ and an integer $r \in \mathbb{Z}^+$, and initialize a state variable u via $u_0 = u_{-1} = \cdots = u_{1-r} = 0$. A scalar quantizer is a map $Q : \mathbb{R} \to \mathcal{A}$, defined via its action

$$Q(v) = \arg\min_{q \in \mathcal{A}} |q - v|.$$

An r^{th}-order $\Sigma\Delta$ quantization scheme with quantization rule $\rho : \mathbb{R}^{r+1} \to \mathbb{R}$ and scalar quantizer Q computes $q \in \mathcal{A}^N$ via the recursion

$$q_i = Q\left(\rho(y_i, u_{i-1}, u_{i-2}, \ldots, u_{i-r})\right), \tag{6}$$

$$u_i = y_i - q_i - \sum_{j=1}^{r} \binom{r}{j} (-1)^j u_{i-j} \tag{7}$$

for all $i \in [N]$. Thus, when $y_i = \langle f_i, x\rangle$, the relationship between x, u, and q can be written in matrix-vector notation as

$$D^r u \ = \ Fx - q. \tag{8}$$

Here the $N \times N$ matrix D is the first-order difference matrix with entries given by

$$D_{i,j} := \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

Henceforth, we focus on *midrise* alphabets of the form

$$\mathcal{A}_K^\delta = \{\pm(2n-1)\delta/2 : n \in [K]\}, \tag{10}$$

where δ denotes the quantization step size. We restrict our attention to *stable* r^{th}*-order schemes*, i.e., schemes for which (6) and (7) produce vectors $u \in \mathbb{R}^N$ with $\|u\|_\infty \le C_{\rho,Q}(r)$ for all $N \in \mathbb{N}$, and $y \in \mathbb{R}^N$ with $\|y\|_\infty \le 1$. Moreover, for our definition of stability we require that $C_{\rho,Q} : \mathbb{N} \mapsto \mathbb{R}^+$ be entirely independent of both N and y. For example, stable r^{th}-order $\Sigma\Delta$ schemes with $C_{\rho,Q}(r) = O(r^r)$ exist,[14,21] even when $\mathcal{A}$ is a 1-bit alphabet. Of particular interest to us is the *single-bit greedy first order* $\Sigma\Delta$ *scheme*, characterized by $r = 1$, $\mathcal{A} = \{\pm 1\}$ and $\rho : (y_i, u_{i-1}) \mapsto y_i + u_{i-1}$.

From a mathematical point of view, $\Sigma\Delta$ schemes have been investigated extensively in the case of bandlimited functions (see, e.g. Refs. 13, 14, 21) where it has been shown[14,21] that for certain quantization rules ρ, they can yield (near optimal) reconstruction errors that decay exponentially fast with the sampling rate. These results rely on choosing the order r as a function of the sampling rate. In the finite-frame scenario the best known analogous results (e.g., Refs. 18, 20) yield reconstruction errors that decay root-exponentially fast with the number of measurements, also by choosing the order r as a function of the sampling rate.

4. MAIN RESULTS

In what follows we will need a few definitions. We begin with the singular value decomposition of D, essentially computed by von Neumann,[22] and given by $D = U\Sigma V^T$, where

$$U_{i,j} = \sqrt{\frac{2}{N+1/2}}\cos\left(\frac{2(i-1/2)(N-j+1/2)\pi}{2N+1}\right), \tag{11}$$

$$\Sigma_{i,j} = \delta_{i,j}\sigma_j(D) = 2\delta_{i,j}\cos\left(\frac{j\pi}{2N+1}\right), \tag{12}$$

and

$$V_{i,j} = (-1)^{j+1}\sqrt{\frac{2}{N+1/2}}\sin\left(\frac{2ij}{2N+1}\pi\right). \tag{13}$$

Note that the difference matrix, D, is full rank (e.g., see (12)).

Definition 1. The 1st order Sobolev self-dual frame[18] is a frame formed from the (renormalized) last d columns of U,

$$F := \sqrt{\frac{N}{2d}}\left(U_{N-d+1} \dots U_N\right). \tag{14}$$

Definition 2. A random selector matrix is a matrix $R \in \{0,1\}^{m\times N}$ with exactly one nonzero entry per row, which is selected uniformly at random.

For a full-rank $N \times d$ matrix A with $N \ge d$, we denote its Moore-Penrose left-inverse by $A^\dagger := (A^*A)^{-1}A^*$.

Theorem 1. Let $\epsilon, p \in (0,1)$, and let $R \in \{0,1\}^{m\times N}$ be a random selector matrix. Denote by F the first order Sobolev self-dual frame and by q the vector resulting from single-bit greedy first order $\Sigma\Delta$ quantization applied to Fx, where $x \in \mathbb{R}^d$. Then

$$\left\|\mathbf{x} - \left(RD^{-1}F\right)^\dagger RD^{-1}q\right\|_2 \le \frac{\sqrt{2}\pi}{\sqrt{1-\epsilon}}\left(\frac{d^{\frac{3}{2}}}{N}\right)$$

for all $x \in B_2^d \subset \mathbb{R}^d$ with probability at least $1-p$, provided that $m \ge (16/3)\epsilon^{-2}d\ln(2d/p)$. Furthermore, $RD^{-1}q$ can always be encoded using $b \le m(\log_2 N + 1)$ bits.

Proof. See Ref. 1 for the full proof. □

Theorem 1 provides the desired exponentially decaying rate-distortion bounds. To see this, note that the encoding can be computed linearly via

$$\mathcal{E} : q \mapsto RD^{-1}q. \tag{15}$$

Since D^{-1} is a discrete integrator, the encoding consists of simply tracking the running sum of the bits and storing a few of them, chosen at random. The reconstruction is performed by applying $(RD^{-1}F)^{\dagger}$ to the encoded bit-stream. Moreover, by choosing the smallest integer $m \geq (16/3)\epsilon^{-2}d\ln(2d/p)$, the rate is

$$\mathcal{R} = m(\log_2 N + 1),$$

and the distortion is

$$\mathcal{D} = \frac{\sqrt{2}\pi d^{3/2}}{N\sqrt{1-\epsilon}}.$$

Expressing the distortion in terms of the rate, we obtain

$$\mathcal{D}(\mathcal{R}) = \frac{2\sqrt{2}\pi d^{3/2}}{\sqrt{1-\epsilon}} \cdot 2^{-\mathcal{R}/m} = C_1(\epsilon) \cdot d^{3/2} \exp\left(-\frac{\mathcal{R}}{C_2(\epsilon) d \ln(2d/p)}\right). \tag{16}$$

Above, $C_1(\epsilon) = 2\pi \cdot \sqrt{\frac{2}{1-\epsilon}}$ and $\frac{16}{3\ln 2 \cdot \epsilon^2} + \frac{1}{d \ln 2 \cdot \ln(2d/p)} \geq C_2(\epsilon) \geq \frac{16}{3 \ln 2 \cdot \epsilon^2}$.

While Theorem 1 provides an instance where $\Sigma\Delta$ quantization coupled with simple encoding and decoding steps yields near-optimal rate-distortion performance, there is still room for improvement. Specifically, the theorem pertains to first order $\Sigma\Delta$ and thus requires one to collect N measurements to obtain a reconstruction error of $O(N^{-1})$. Moreover, the theorem is tailored to Sobolev self-dual frames. To generalize Theorem 1 to more general frames and $\Sigma\Delta$ schemes of arbitrary order, we now introduce some notation.

Definition 3. We will call a frame matrix $F \in \mathbb{R}^{N \times d}$ an (r, C, α)-frame if

1. $\|Fx\|_\infty \leq 1$ for all $x \in B_2^d$, and
2. $\sigma_d\left(D^{-r}F\right) \geq C \cdot N^{\alpha}$.

Roughly speaking, the first condition of Definition 3 ensures that the frame F is uniformly bounded, while the second condition can be interpreted as a type of smoothness requirement. We now may state a more general theorem.

Theorem 2. Let $\epsilon, p \in (0,1)$, $B \in \{-1,1\}^{m \times N}$ be a Bernoulli random matrix, and $F \in \mathbb{R}^{N \times d}$ be an (r, C, α)-frame with $r \in \mathbb{N}$, $\alpha \in (1, \infty)$, and $C \in \mathbb{R}^+$. Consider q, the quantization of Fx via a stable r^{th}-order scheme with alphabet $\mathcal{A}_A^{2\mu}$ and stability constant $C_{\rho,Q}(r) \in \mathbb{R}^+$ (see (6), (7), (10) and the subsequent discussion). Then, the following are true.

(i) The reconstruction error (i.e., the distortion) satisfies

$$\left\|\mathbf{x} - \left(BD^{-r}F\right)^{\dagger} BD^{-r}q\right\|_2 \leq \frac{C_{\rho,Q}(r) \cdot N^{1-\alpha}}{C \cdot (1-\epsilon)}$$

for all $x \in B_2^d \subset \mathbb{R}^d$ with probability at least $1-p$, provided that $m \geq \frac{4d\ln(12/\epsilon) + 2\ln(1/p)}{\epsilon^2/8 - \epsilon^3/24}$.

(ii) $BD^{-r}q$ can always be encoded using $b \leq m[(r+1)\log_2 N + \log_2 A + 1]$ bits.

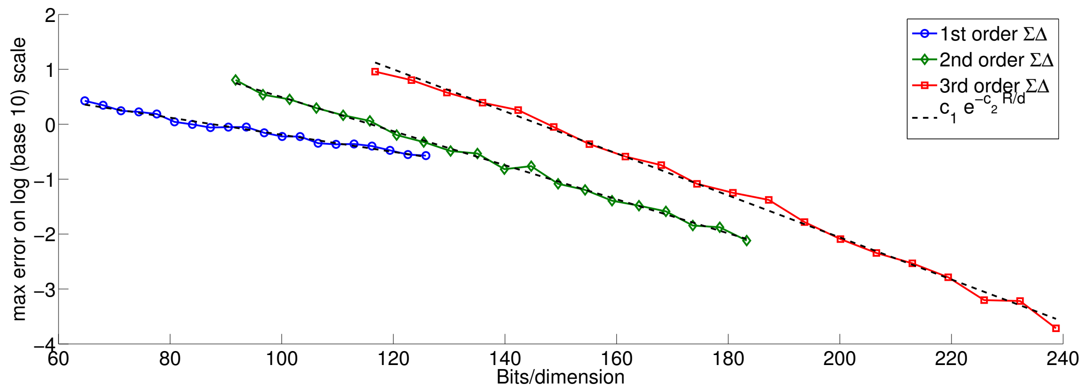

Figure 1. The maximum ℓ_2-norm error (in $\log_{10}$ scale) plotted against the number of bits per dimension (b/d). Here $d = 20$ and $\Sigma\Delta$ schemes with $r = 1, 2$ and 3 are used to quantize the frame coefficients. A Bernoulli matrix is used for encoding.

Theorem 2 provides exponentially decaying rate-distortion bounds with the encoder $\mathcal{E} : q \mapsto BD^r q$ and decoder $\Delta : z \mapsto (BD^{-r}F)^{\dagger} z$. By choosing the smallest integer $m \geq \frac{4d\ln(12/\epsilon)+2\ln(1/p)}{\epsilon^2/8-\epsilon^3/24}$, the rate is

$$\mathcal{R} = m[(r+1)\log_2 N + \log_2 A + 1],$$

and the distortion is

$$\mathcal{D} = \frac{C_{\rho,Q}(r) \cdot N^{1-\alpha}}{C \cdot (1-\epsilon)}.$$

Expressing the distortion in terms of the rate, we obtain

$$\mathcal{D}(\mathcal{R}) = \frac{C_{\rho,Q}(r) \cdot (2A)^{(\alpha-1)/(r+1)}}{C \cdot (1-\epsilon)} \cdot 2^{-(\mathcal{R}(\alpha-1)/m(r+1))} \;\leq\; \bar{C}_{\rho,Q}(A, \epsilon, \alpha, r) \cdot \exp\left(-\frac{\mathcal{R}}{d \cdot C_3(\epsilon, p)}\right). \tag{17}$$

Above, $\bar{C}_{\rho,Q}(A, \epsilon, \alpha, r) = \frac{C_{\rho,Q}(r)\cdot(2A)^{(\alpha-1)/(r+1)}}{C\cdot(1-\epsilon)}$ and $\frac{4(r+1)\ln(12/\epsilon p)}{\ln 2\cdot(\alpha-1)(\epsilon^2/8-\epsilon^3/24)} + \frac{1}{d} \geq C_3(\epsilon, p) > 0$.

It is worth noting that there are several examples of (r, C, α) frames for which the above theorem yields reconstruction errors that decay exponentially with the number of bits used for encoding. These include random frames whose entries are drawn independently from sub-Gaussian distributions, as well as smooth frames with frame matrices whose columns are sampled from piecewise smooth functions.[17] For more details, see Ref. 1.

5. NUMERICAL EXPERIMENT

To illustrate our results, we perform a numerical experiment whereby we fix $d = 20$, and generate 5000 points uniformly from B_2^d. For various N, we draw an $N \times d$ matrix F whose entries F_{ij} are Bernoulli random variables. We use the $\Sigma\Delta$ schemes of Ref. 21 with $r = 1, 2,$ and 3 to quantize the frame coefficients Fx. Finally, we use Bernoulli matrices with $m = 5d$ to encode. In Figure 1 we show the maximum error (over the 5000 instances of x) versus the bit-rate. Note the different slopes corresponding to $r = 1, 2,$ and 3.

ACKNOWLEDGMENTS

M.I. was supported in part by NSA grant H98230-13-1-0275. R.S. was supported in part by a Banting Postdoctoral Fellowship, administered by the Natural Sciences and Engineering Research Council of Canada (NSERC). The majority of the work reported on herein was completed while the authors were visiting assistant professors at Duke University.

REFERENCES

[1] Iwen, M. and Saab, R., "Near-optimal encoding for sigma-delta quantization of finite frame expansions," *arXiv preprint arXiv:1307.2136* (2013).

[2] Casazza, P. G. and Kovačević, J., "Equal-norm tight frames with erasures," *Advances in Computational Mathematics* **18**(2-4), 387–430 (2003).

[3] Goyal, V. K., Kovačević, J., and Kelner, J. A., "Quantized frame expansions with erasures," *Applied and Computational Harmonic Analysis* **10**(3), 203–233 (2001).

[4] Balan, R., Casazza, P., and Edidin, D., "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis* **20**(3), 345–356 (2006).

[5] Candes, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V., "Phase retrieval via matrix completion," *SIAM Journal on Imaging Sciences* **6**(1), 199–225 (2013).

[6] Candès, E. J., Romberg, J., and Tao, T., "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.* **59**, 1207–1223 (2006).

[7] Donoho, D., "Compressed sensing.," *IEEE Trans. Inform. Theory* **52**(4), 1289–1306 (2006).

[8] Brady, D. J., "Multiplex sensors and the constant radiance theorem," *Optics Letters* **27**(1), 16–18 (2002).

[9] Kohman, T. P., "Coded-aperture x-or γ-ray telescope with least-squares image reconstruction. i. design considerations," *Review of scientific instruments* **60**(11), 3396–3409 (1989).

[10] Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F., and Baraniuk, R. G., "Single-pixel imaging via compressive sampling," *Signal Processing Magazine, IEEE* **25**(2), 83–91 (2008).

[11] Fickus, M., Massar, M. L., and Mixon, D. G., "Finite frames and filter banks," in [*Finite Frames*], 337–379, Springer (2013).

[12] Inose, H. and Yasuda, Y., "A unity bit coding method by negative feedback," *Proceedings of the IEEE* **51**(11), 1524–1535 (1963).

[13] Daubechies, I. and DeVore, R., "Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order," *Ann. Math.* **158**(2), 679–710 (2003).

[14] Güntürk, C., "One-bit sigma-delta quantization with exponential accuracy," *Communications on Pure and Applied Mathematics* **56**(11), 1608–1630 (2003).

[15] Benedetto, J., Powell, A., and Yılmaz, Ö., "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inform. Theory* **52**(5), 1990–2005 (2006).

[16] Bodmann, B., Paulsen, V., and Abdulbaki, S., "Smooth frame-path termination for higher order sigma-delta quantization," *J. Fourier Anal. and Appl.* **13**(3), 285–307 (2007).

[17] Blum, J., Lammers, M., Powell, A., and Yılmaz, Ö., "Sobolev duals in frame theory and sigma-delta quantization," *J. Fourier Anal. and Appl.* **16**(3), 365–381 (2010).

[18] Krahmer, F., Saab, R., and Ward, R., "Root-exponential accuracy for coarse quantization of finite frame expansions," *IEEE Trans. Inform. Theory* **58**, 1069 –1079 (February 2012).

[19] Lorentz, G., von Golitschek, M., and Makovoz, Y., [*Constructive approximation: advanced problems*], Grundlehren der mathematischen Wissenschaften, Springer (1996).

[20] Krahmer, F., Saab, R., and Yılmaz, Ö., "Sigma-delta quantization of sub-Gaussian frame expansions and its application to compressed sensing," *Preprint, arXiv:1306.4549* (2013).

[21] Deift, P., Krahmer, F., and Güntürk, C., "An optimal family of exponentially accurate one-bit sigma-delta quantization schemes," *Communications on Pure and Applied Mathematics* **64**(7), 883–919 (2011).

[22] Von Neumann, J., "Distribution of the ratio of the mean square successive difference to the variance," *The Annals of Mathematical Statistics* **12**(4), 367–395 (1941).