Fast Subspace Approximation via Greedy Least-Squares

M. A. Iwen*
Department of Mathematics, Michigan State University
Department of Electrical and Computer Engineering, Michigan State University
Email: markiwen@math.msu.edu

Felix Krahmer
Institute for Numerical and Applied Mathematics, University of Göttingen
Email: f.krahmer@math.uni-goettingen.de

June 30, 2014

Abstract

In this note, we develop fast and deterministic dimensionality reduction techniques for a family of subspace approximation problems. We then utilize these dimensionality reduction techniques in order to help rapidly and accurately approximate the n-widths of point sets. Let $P \subset \mathbb{R}^N$ be a given set of M points. The techniques developed herein find an $O(n \log M)$-dimensional subspace that is guaranteed to always contain a near-best fit n-dimensional hyperplane $\mathcal{H}$ for P with respect to the cumulative projection error $\left(\sum_{\mathbf{x} \in P} \|\mathbf{x} - \Pi_{\mathcal{H}} \mathbf{x}\|_2^p\right)^{1/p}$, for any chosen $p > 2$. The deterministic algorithm runs in $\tilde{O}\left(MN^2\right)$-time, and can be randomized to run in only $\tilde{O}\left(MNn\right)$-time while maintaining its error guarantees with high probability. In the important $p = \infty$ case the dimensionality reduction techniques are then combined with efficient algorithms for computing the John ellipsoid of a data set in order to produce an n-dimensional subspace whose maximum ℓ_2-distance to any point in the convex hull of P is minimized. The resulting algorithm remains $\tilde{O}\left(MNn\right)$-time.

Keywords: Approximation algorithms, subspace approximation, n-widths, dimensionality reduction, greedy algorithms, least-squares

1 Introduction

Fitting a given point cloud with a low-dimensional affine subspace is a fundamental computational task in data analysis. In this paper we consider fast algorithms for approximating a given set of M points, $P \subset \mathbb{R}^N$, with an n-dimensional affine subspace $\mathcal{A} \subset \mathbb{R}^N$ that is a near-best fit. Here the fitness of $\mathcal{A}$ will be measured by $d^{(p)}(P, \mathcal{A}) := \sqrt[p]{\sum_{\mathbf{x} \in P} \left(d(\mathbf{x}, \mathcal{A})\right)^p}$, where $d(\mathbf{x}, \mathcal{A})$ is the Euclidean distance from $\mathbf{x}$ to $\mathcal{A}$, and $p \in \mathbb{R}^+$. Similarly, when $p = \infty$ the fitness measure will be $d^{(\infty)}(P, \mathcal{A}) := \max_{\mathbf{x} \in P} d(\mathbf{x}, \mathcal{A})$. An n-dimensional affine subspace $\mathcal{A} \subset \mathbb{R}^N$ is a *near-best fit for P* with respect to this fitness measure if there exists a small constant $C \in \mathbb{R}^+$ such that $d^{(p)}(P, \mathcal{A}) \leq C \cdot d^{(p)}(P, \mathcal{H})$ for all n-dimensional affine subspaces $\mathcal{H} \subset \mathbb{R}^N$.[1] In this paper we are

*Contact Author. Supported in part by NSA grant H98230-13-1-0275.
[1]The approximation constant C may depend (mildly) on both p and $|P| = M$.

interested in calculating near-best fit affine subspaces for large and high-dimensional point sets, $P \subset \mathbb{R}^N$, as rapidly as possible.

In the case $p = 2$ the problem above is the well known least-squares approximation problem. Mathematically, a near-best fit n-dimensional least-squares subspace can be obtained by computing the top n eigenvectors of XX^{T} for the matrix $X \in \mathbb{R}^{N \times M}$ whose columns are the points in P. Decades of progress related to the computational eigenvector problem has resulted in many efficient numerical schemes for this problem (see, e.g., [19, 7], and the references therein). The situation is more difficult when $p \neq 2$. None the less, a good deal of work has been done developing algorithms for other values of p as well.

Examples include methods for approximately solving the case $p = 1$, which has been proposed as a means of reducing the effects of statistical outliers on an approximating subspace (see, e.g., [15]). However, in this paper we are primarily interested in $p > 2$, with our main focus being on the important $p = \infty$ case. In particular, we develop fast dimensionality reduction techniques for the subspace approximation problem which can be used in combination with existing solution methods for any $p > 2$ [16, 2] in order to reduce their runtimes. For the important case $p = \infty$ these new dimensionality reduction methods yield a new fast approximation algorithm guaranteed to find near-optimal solutions.

1.1 Results and Previous Work for the $p = \infty$ Case

The case $p = \infty$ is closely related to several fundamental computational problems in convex geometry and has been widely studied (see, e.g., [6, 4, 8, 20, 1, 18], and references therein). Previous computational methods developed for this case can be grouped into two general categories: methods based on semi-definite programming relaxations (e.g., [20, 18]), and methods based on core-set techniques (e.g., [8, 1]). Both approaches have comparative strengths. The semidefinite programming approach leads to highly accurate approximations. In particular, [18] demonstrates a randomized approach which computes an n-dimensional subspace $\mathcal{A}$ that has $d^{(\infty)}(P, \mathcal{A}) \leq \sqrt{12 \log M} \cdot d^{(\infty)}(P, \mathcal{H})$ for all n-dimensional subspaces $\mathcal{H} \subset \mathbb{R}^N$ with high probability. Furthermore, the approximation factor $\sqrt{12 \log M}$ is shown to be close to the best achievable in polynomial time. However, the method requires the solution of a semi-definite program, and so has a runtime complexity that scales super-linearly in both M and N. This makes the technique intractable for large sets of points in high dimensional space.

The core-set approach achieves better runtime complexities for small values of n. In [1] a $\tilde{O}(MN2^n)$-time randomized approximation algorithm is developed for the $p = \infty$ case.[2] This algorithm has the advantage of being linear in both M and N, but quickly becomes computationally infeasible as the dimension of the approximating subspace, n, grows.

In this paper we develop an $\tilde{O}(MN^2)$-time deterministic algorithm which computes an n-dimensional subspace $\mathcal{A}$ that is guaranteed to have $d^{(\infty)}(P, \mathcal{A}) \leq C\sqrt{n \log M} \cdot d^{(\infty)}(P, \mathcal{H})$ for all n-dimensional subspaces $\mathcal{H} \subset \mathbb{R}^N$. Here $C \in \mathbb{R}^+$ is a small universal constant (e.g., it can be made less than 10). Furthermore, the algorithm can be randomized to run in only $\tilde{O}(MNn)$-time while still achieving the same accuracy guarantee with high probability. This improves on the runtime complexities of existing core-set approaches while simultaneously obtaining accuracies on the order of existing semi-definite programming methods for small n.

The approximation algorithms for the $p = \infty$ case developed in this paper are motivated by the following idea: The difficulty of approximating $P \subset \mathbb{R}^N$ with a subspace can be greatly reduced by first approximating (the convex hull of) P with an ellipsoid, and then approximating the resulting

[2] Herein, $\tilde{O}(\cdot)$-notation indicates that polylogarithmic factors have been dropped from the associated O-upper bounds for the sake of readability.

ellipsoid with an n-dimensional subspace. In fact, fast algorithms for approximating (the convex hull of) P by an ellipsoid are already known (see, e.g., [11, 14, 17]). And, it is straightforward to approximate an ellipsoid optimally with an n-dimensional subspace – one may simply use its n largest semi-axes as a basis. The only deficit in this simple approach is that the accuracy it guarantees is rather poor. The resulting n-dimensional subspace $\mathcal{A}$ may have $d^{(\infty)}(P, \mathcal{A})$ as large as $\sqrt{N} \cdot d^{(\infty)}(P, \mathcal{H})$ for some other n-dimensional subspace $\mathcal{H} \subset \mathbb{R}^N$. This guarantee can be improved, however, if N (i.e., the dimension of the point set P) is reduced before the approximating ellipsoid is computed. Motivated by this idea, we develop new dimensionality reduction algorithms for the subspace approximation problem below.

1.2 Dimensionality Reduction Results and Previous Work

An algorithm is a dimensionality reduction method for the subspace approximation problem if, for any $P \subset \mathbb{R}^N$, it finds a low-dimensional subspace that is guaranteed to contain a near-best fit n-dimensional hyperplane $\mathcal{H}$. Such dimensionality reduction methods can be regarded as a "weak" approximate solution methods for the subspace approximation problem in the following sense. They produce subspaces whose dimensions are larger than n (i.e., larger than the target dimension of the desired best-fit hyperplane), but solving the problem restricted to these subspaces will yield a near-optimal solution. Thus dimensionality reduction methods – when sufficiently fast – allow the subspace approximation problem to be simplified before more time intensive solution methods are employed. For example, if a low-dimensional subspace has been found, which still contains a near-best fit solution, high-dimensional data (i.e., with N large) can be projected onto that subspace in order to reduce its complexity before solving. Hence, fast dimensionality reduction algorithms can be used to help speed up existing solutions methods for $p > 2$ (e.g., by reducing the input problem sizes for methods based on solving convex programs [2].)

Several dimensionality reduction techniques have been developed for the subspace approximation problem over the past several years (see, e.g., [1, 3, 5] and references therein). These methods are all based on sampling techniques and either have runtime complexities that scale exponentially in n, or embedding subspace dimensions that scale exponentially in p. In [3], for example, an $MNn^{O(1)}$-time randomized algorithm is given which is guaranteed, with high probability, to return an $\tilde{O}(n^{p+3})$-dimensional subspace that itself contains another n-dimensional subspace, $\mathcal{A}$, whose fit, $d^{(p)}(P, \mathcal{A})$, is the near-best possible for any $p \in [1, \infty)$. Although useful for small p, these methods quickly become infeasible as p increases.

In this paper a different dimensionality reduction approach is taken that reduces the problem, for any $p \geq 2$, to a small number of least-squares problems. The idea is to greedily approximate a large portion of the input data P with a fast least-squares method. It turns out that a large portion of P is always well-approximated, *for any* $p > 2$, by P's best-fit n-dimensional least-squares subspace. Then, the previously worst-approximated points in P can be iteratively fit by least-squares subspaces until all of P has eventually been approximated well, with respect to any desired $p > 2$, by the union of $O(\log M)$ least-squares subspaces. Using this idea, a deterministic $\tilde{O}(MN^2)$-time algorithm can be developed which is always guaranteed to return an $O(n \log M)$-dimensional subspace that itself contains another n-dimensional subspace, $\mathcal{A}$, whose fit, $d^{(p)}(P, \mathcal{A})$, is the near-best possible for any $p \in [2, \infty]$. Furthermore, this algorithm can be randomized to run in only $\tilde{O}(MNn)$-time while still achieving the same accuracy guarantees as the deterministic variant with high probability.

1.3 Organization

The remainder of this paper is organized as follows: In Section 2 notation is established and necessary theory is reviewed. Then, in Section 3, the dimensionality reduction results are developed for any $p > 2$. Finally, in Section 4, our improved dimensionality reduction result for the case $p = \infty$ is used to illustrate a fast and simple subspace approximation algorithm for the $p = \infty$ subspace approximation problem.

2 Preliminaries: Notation and Setup

For any matrix $X \in \mathbb{R}^{N\times M}$ we will denote the j^{th} column of X by $\mathbf{X}_j \in \mathbb{R}^N$. The transpose of a matrix, $X \in \mathbb{R}^{N\times M}$, will be denoted by $X^{\mathrm{T}} \in \mathbb{R}^{M\times N}$, and the singular values of any matrix $X \in \mathbb{R}^{N\times M}$ will always be ordered as $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_{\min(N,M)}(X) \geq 0$. The Frobenius norm of $X \in \mathbb{R}^{N\times M}$ is defined as

$$\|X\|_F := \sqrt{\sum_{j=1}^{M}\sum_{i=1}^{N}|X_{i,j}|^2} = \sqrt{\sum_{l=1}^{\min(N,M)} \sigma_l^2(X)}. \tag{1}$$

A key ingredient of our results is the following perturbation bounds for singular values (see, e.g., [9]).

Theorem 1 (Weyl)**.** *Let $A, B \in \mathbb{R}^{M\times N}$, and $q = \min\{M, N\}$. Then,*

$$\sigma_{i+j-1}(A + B) \leq \sigma_i(A) + \sigma_j(B)$$

holds for all $i, j \in \{1, \ldots, q\}$ with $i + j \leq q + 1$.

Given an $\tilde{n}$-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^N$, we will denote the set of all n-dimensional affine subspaces of $\mathcal{S}$ by $\Gamma_n(\mathcal{S})$. Here, of course, we assume that $N \geq \tilde{n} \geq n$. Given an affine subspace $\mathcal{A} \in \Gamma_n(\mathcal{S})$, we will denote the offset of $\mathcal{A}$ by

$$\mathbf{a}_{\mathcal{A}} := \arg\min_{\mathbf{x}\in\mathcal{A}} \|\mathbf{x}\|_2, \tag{2}$$

and the n-dimensional subspace of $\mathcal{S}$ that is parallel to $\mathcal{A}$ by

$$\mathcal{S}_{\mathcal{A}} := \mathcal{A} - \mathbf{a}_{\mathcal{A}} := \left\{\mathbf{x} - \mathbf{a}_{\mathcal{A}} \;\middle|\; \mathbf{x} \in \mathcal{A}\right\}. \tag{3}$$

Note that $\mathbf{a}_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}^{\perp}$. Thus, we may define the projection operator onto $\mathcal{A}$, $\Pi_{\mathcal{A}} : \mathbb{R}^N \to \mathcal{A}$, by

$$\Pi_{\mathcal{A}}\mathbf{x} := \Pi_{\mathcal{S}_{\mathcal{A}}}\mathbf{x} + \mathbf{a}_{\mathcal{A}}. \tag{4}$$

Here $\Pi_{\mathcal{S}_{\mathcal{A}}}$ is the orthogonal projection onto $\mathcal{S}_{\mathcal{A}}$.

2.1 A Family of Distances

Given a subset $T \subset \mathbb{R}^N$ and an affine subspace $\mathcal{A} \in \Gamma_n(\mathcal{S})$ we will want to consider the "distance" of T from $\mathcal{A}$, defined by

$$d^{(\infty)}(T, \mathcal{A}) := \sup_{\mathbf{x}\in T} \|\mathbf{x} - \Pi_{\mathcal{A}}\mathbf{x}\|_2. \tag{5}$$

Let $\mathcal{S}$ be an $\tilde{n} \geq n$ subspace of $\mathbb{R}^N$. We can now define the Euclidean Kolmogorov n-width of T in this setting by

$$d_n^{(\infty)}(T, \mathcal{S}) := \inf_{\mathcal{A} \in \Gamma_n(\mathcal{S})} d^{(\infty)}(T, \mathcal{A}) = \inf_{\mathcal{A} \in \Gamma_n(\mathcal{S})} \sup_{\mathbf{x} \in T} \|\mathbf{x} - \Pi_{\mathcal{A}} \mathbf{x}\|_2. \tag{6}$$

Finally, we note that there will always be (at least one) optimal affine subspace, $\mathcal{A}_{\text{opt}} \in \Gamma_n(\mathcal{S})$, with

$$d^{(\infty)}(T, \mathcal{A}_{\text{opt}}) = d_n^{(\infty)}(T, \mathcal{S}) \tag{7}$$

when T is "sufficiently nice" (e.g., when T is either finite, or convex and compact).[3]

When $T = \{\mathbf{t}_1, \ldots, \mathbf{t}_M\} \subset \mathbb{R}^N$ is finite, we may define a vector $\mathbf{e}_{\mathcal{A}} \in \mathbb{R}^M$ for any given $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ by

$$\left(\mathbf{e}_{\mathcal{A}}\right)_j := \|\mathbf{t}_j - \Pi_{\mathcal{A}} \mathbf{t}_j\|_2 \,. \tag{8}$$

Thus, when T is finite we can see that

$$d_n^{(\infty)}(T, \mathcal{S}) = \inf_{\mathcal{A} \in \Gamma_n(\mathcal{S})} \|\mathbf{e}_{\mathcal{A}}\|_\infty \,, \tag{9}$$

and the least squares approximation error over all subspaces in $\Gamma_n(\mathcal{S})$ is given by

$$d_n^{(2)}(T, \mathcal{S}) = \inf_{\mathcal{A} \in \Gamma_n(\mathcal{S})} \|\mathbf{e}_{\mathcal{A}}\|_2 \,. \tag{10}$$

These two quantities can be seen as extreme instances of the infinite family of approximation errors given by

$$d_n^{(p)}(T, \mathcal{S}) := \inf_{\mathcal{A} \in \Gamma_n(\mathcal{S})} \|\mathbf{e}_{\mathcal{A}}\|_p \,, \tag{11}$$

for any parameter $2 \leq p \leq \infty$. Note that, analogously to (6), one has

$$d_n^{(p)}(T, \mathcal{S}) := \inf_{\mathcal{A} \in \Gamma_n(\mathcal{S})} d^{(p)}(T, \mathcal{A}), \tag{12}$$

where

$$d^{(p)}(T, \mathcal{A}) := \|\mathbf{e}_{\mathcal{A}}\|_p \,. \tag{13}$$

Finally, as above, we note that there will always be at least one optimal affine subspace, $\mathcal{A}_{\text{opt}} \in \Gamma_n(\mathcal{S})$, with

$$d^{(p)}(T, \mathcal{A}_{\text{opt}}) = d_n^{(p)}(T, \mathcal{S}) \tag{14}$$

when T is finite.

2.2 Symmetry, Ellipsoids, and Properties of n-widths

Let $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^N$, and define

$$\bar{\mathbf{p}} := \frac{1}{M} \cdot \sum_{j=1}^{M} \mathbf{p}_j. \tag{15}$$

[3]This follows from the fact that Stiefel manifolds are compact, together with the fact that only offsets, $\mathbf{a}_{\mathcal{A}} \in \mathbb{R}^N$, contained in the ball of radius $\sup_{\mathbf{x} \in T} \|x\|_2$ are ever relevant to minimizing $d^{(\infty)}(T, \cdot)$. Thus, the set of relevant affine subspaces under consideration is compact when T is bounded. Finally, $d^{(\infty)}(T, \cdot) : \Gamma_n(\mathcal{S}) \to \mathbb{R}^+$, $T \subset \mathbb{R}^N$ fixed, will be continuous when T is sufficiently well behaved (e.g., either finite, or compact and convex).

We will let $\bar{P} \subset \mathbb{R}^N$ denote the following symmetrized translation of P,

$$\bar{P} := (P - \bar{\mathbf{p}}) \cup (\bar{\mathbf{p}} - P) \cup \{\mathbf{0}\} := \left\{\mathbf{p}_j - \bar{\mathbf{p}} \mid \mathbf{p}_j \in P\right\} \cup \left\{\bar{\mathbf{p}} - \mathbf{p}_j \mid \mathbf{p}_j \in P\right\} \cup \{\mathbf{0}\}. \tag{16}$$

We will say that P is *symmetric* if and only if $P = \bar{P}$. Furthermore, we will denote the convex hull of P by $\mathrm{CH}(P)$. The following theorem due to Fritz John [10] guarantees the existence of an ellipsoid that approximates $\mathrm{CH}\left(\bar{P}\right)$ well.

Theorem 2 (John). *Let $K \subset \mathbb{R}^N$ be a compact and convex set with nonempty interior that is symmetric about the origin (so that $K = -K$). Then, there is an ellipsoid centered at the origin, $\mathcal{E} \subset \mathbb{R}^N$, such that $\mathcal{E} \subseteq K \subseteq \sqrt{N} \cdot \mathcal{E}$.*

Given $P \subset \mathbb{R}^N$, an ellipsoid which is nearly as good an approximation to $\mathrm{CH}\left(\bar{P}\right)$ as the ellipsoid guaranteed by Thoerem 2 can be computed in polynomial time (see, e.g., [11, 14, 17]). More specifically, one can compute an ellipsoid $\mathcal{E}$ such that $\mathcal{E} \subseteq \mathrm{CH}\left(\bar{P}\right) \subseteq \sqrt{(1+\epsilon)N} \cdot \mathcal{E}$ in $O(MN^2(\log N + 1/\epsilon))$-time for any $\epsilon \in (0, \infty)$ [17]. Finally, in the following Lemma, we summarize a few facts concerning the n-widths of finite sets, convex hulls, and ellipsoids that will be useful for establishing our results (proofs are included in Appendix A for the sake of completeness).

Lemma 1. *Let $P = \{\mathbf{p}_1, \dots, \mathbf{p}_M\} \subset \mathbb{R}^N$, and $\mathcal{E} \subset \mathbb{R}^N$ be the ellipsoid*

$$\left\{\mathbf{x} \in \mathbb{R}^N \mid \mathbf{x}^T Q \mathbf{x} \leq 1\right\},$$

where $Q \in \mathbb{R}^{N \times N}$ is symmetric and positive definite. Then,

1. $d_n^{(\infty)}\left(P - \mathbf{x}, \mathbb{R}^N\right) = d_n^{(\infty)}\left(P, \mathbb{R}^N\right)$ *for all $\mathbf{x} \in \mathbb{R}^N$, and $n = 1, \dots, N$.*
2. *$\bar{P}$ will have an optimal n-dimensional subspace (i.e., with $\mathbf{a}_{\mathcal{A}_{\mathrm{opt}}} = \mathbf{0}$) for all $n = 1, \dots, N$.*
3. $d_n^{(\infty)}\left(\bar{P}, \mathbb{R}^N\right) \leq 2 \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right)$ *for all $n = 1, \dots, N$.*
4. $d_n^{(\infty)}(B, \mathbb{R}^N) \leq d_n^{(\infty)}(C, \mathbb{R}^N)$ *for all $B \subseteq C \subset \mathbb{R}^N$, and $n = 1, \dots, N$.*
5. $d_n^{(\infty)}(\mathrm{CH}(P), \mathbb{R}^N) = d_n^{(\infty)}(P, \mathbb{R}^N)$ *for all $n = 1, \dots, N$.*
6. $d_n^{(\infty)}(\mathcal{E}, \mathbb{R}^N) = \sqrt{\frac{1}{\sigma_{N-n+1}(Q)}}$ *for all $n = 1, \dots, N$. Consequently, an optimal n-dimensional subspace for $\mathcal{E}$ is spanned by the eigenvectors of Q associated with $\sigma_N(Q), \dots, \sigma_{N-n+1}(Q)$.*

We will assume hereafter, without loss of generality, that $P = \{\mathbf{p}_0, \dots, \mathbf{p}_M\} \subset \mathbb{R}^N$ both spans $\mathbb{R}^N$ and is symmetric.[4] Note that these assumptions are rather mild in practice. If P is not initially symmetric we will simply approximate $\bar{P}$ by a subspace instead. A translation of our approximating subspace for $\bar{P}$ will then still approximate the original set P well by parts $(1) - (4)$ of Lemma 1. If P initially does not span $\mathbb{R}^N$, we will replace each element of P with the coordinates of its orthogonal projection into the span of P, reducing N accordingly. Any such change of basis for P will lead to no loss of accuracy in our solution. Finally, we will always denote $\mathbf{0}$ by $\mathbf{p}_0$ for notational convenience.

[4] Here $\mathbf{p}_0 := \mathbf{0}$ has been added to P, if not initially present, so that P contains its mean.

3 Dimensionality Reduction Results

In this section we establish our main theorems regarding dimensionality reduction. As we shall see, the main idea behind the proofs of both Theorems 3 and 4 below is to use existing fast least-squares methods in order to quickly approximate the point set P in a greedy fashion. To see how this works, note that P's best-fit least squares subspace will generally fail to approximate all of P to within $d_n^{(p)}\left(P, \mathbb{R}^N\right)$-accuracy when $p > 2$. However, it *will* generally approximate a large fraction of P sufficiently well. Furthermore, we can easily tell which portion of P is approximated best.

Hence, we may employ the following greedy approach: we (i) approximate P with its best-fit least squares subspace, (ii) identify the half of its points fit the best, (iii) remove them from P, and then (iv) repeat the process again on the remaining portion of P. After $O(\log M)$ repetitions we end up with a collection of at most $O(\log M)$ least squares subspaces whose collective span is guaranteed to contain a near-optimal n-dimensional approximation to all of P with respect to $d_n^{(p)}\left(P, \mathbb{R}^N\right)$.

We are now ready to begin proving Theorems 3 and 4. We start by proving Lemma 2, which demonstrates that ordering the points of P properly results in a predictable decay of their distances from the best-fit least squares subspace for P with respect to $d_n^{(p)}\left(P, \mathbb{R}^N\right)$. Thus, points which are well approximated with respect to $d_n^{(p)}\left(P, \mathbb{R}^N\right)$ by the best-fit least squares subspace for P are easy to identify via sorting. Next, Lemmas 3 and 4 use Lemma 2 to establish that a best-fit least squares subspace for P will approximate *most* of P near-optimally with respect to $d_n^{(p)}\left(P, \mathbb{R}^N\right)$ (Lemma 3 deals with $p = \infty$, and Lemma 4 with $p \in (2, \infty)$). Finally, Lemmas 3 and 4 are used in order to establish Theorems 3 and 4, respectively.

Lemma 2. *Let $P = \{\mathbf{p}_0 := \mathbf{0}, \mathbf{p}_1, \dots, \mathbf{p}_M\} \subset \mathbb{R}^N$ be symmetric, $n \in \{1, \dots, N\}$, and $p \in (2, \infty]$. Then there is an $O\left(MN^2\right)$-time[5] algorithm that outputs an n-dimensional subspace $\mathcal{S} \subset \mathbb{R}^N$ such that for $m \in \{1, \dots, M\}$ one has*

$$\|\mathbf{p}_{l_m} - \Pi_{\mathcal{S}} \mathbf{p}_{l_m}\|_2^2 \leq \frac{M^{1-\frac{2}{p}}}{M - m + 1} \cdot \left(d_n^{(p)}\left(P, \mathbb{R}^N\right)\right)^2, \tag{17}$$

where the $\ell_i > 0$, $i = 1, \dots, M$, are chosen to satisfy

$$0 = \|\mathbf{p}_0 - \Pi_{\mathcal{S}} \mathbf{p}_0\|_2 \leq \|\mathbf{p}_{l_1} - \Pi_{\mathcal{S}} \mathbf{p}_{l_1}\|_2 \leq \|\mathbf{p}_{l_2} - \Pi_{\mathcal{S}} \mathbf{p}_{l_2}\|_2 \leq \cdots \leq \|\mathbf{p}_{l_M} - \Pi_{\mathcal{S}} \mathbf{p}_{l_M}\|_2. \tag{18}$$

Proof: Denote the matrix whose columns are the points in P by $X \in \mathbb{R}^{N \times M}$. That is, let

$$X := (\mathbf{p}_1, \dots, \mathbf{p}_M). \tag{19}$$

Let $\mathcal{A}_{\text{opt}}^{(p)} \in \Gamma_n(\mathbb{R}^D)$ be an optimal n-dimensional subspace for P satisfying

$$d^{(p)}\left(P, \mathcal{A}_{\text{opt}}^{(p)}\right) = d_n^{(p)}\left(P, \mathbb{R}^N\right). \tag{20}$$

It is not difficult to see that we will have $X = Y + E$, where $Y, E \in \mathbb{R}^{N \times M}$ have the following properties: the column span of Y is contained in $\mathcal{A}_{\text{opt}}^{(p)}$, and the vector $\mathbf{e}$ whose entries are the ℓ^2-norms of the columns of E has ℓ^p-norm at most $d_n^{(p)}\left(P, \mathbb{R}^N\right)$. It follows from Hölder's inequality using $\frac{p}{2}$ and $\frac{p}{p-2}$ that

$$\sum_{l=1}^{\min(N,M)} \sigma_l^2(E) = \|E\|_F^2 = \|\mathbf{e}\|_2^2 \leq \|\mathbf{e}\|_p^2 \|\mathbb{I}\|_{1+\frac{2}{p-2}} \leq M^{1-\frac{2}{p}} \cdot \left(d_n^{(p)}\left(P, \mathbb{R}^N\right)\right)^2, \tag{21}$$

[5] We assume here that $M \geq N \geq \log M$. We also note that this runtime complexity can be improved substantially by utilizing randomized low-rank approximation algorithms. See Remark 1 for more details.

where $\mathbb{I} \in \mathbb{R}^M$ is the vector whose entries are all one. Note that Y has rank at most n so that

$$\sigma_{n+1}(Y) = \cdots = \sigma_{\min(N,M)}(Y) = 0. \tag{22}$$

Applying Theorem 1 we now learn that

$$\sigma_{n+l}(X) \leq \sigma_l(E) \tag{23}$$

for all $l \in \{1, \ldots, N-n\}$.

Let X_n be the best rank n approximation to X with respect to Frobenius norm,

$$X_n := \operatorname*{arg\,min}_{\substack{L \in \mathbb{R}^{N \times M} \\ \text{rank } L = \ n}} \|X - L\|_F. \tag{24}$$

Let $\mathcal{S}$ be the n-dimensional subspace spanned by the columns of X_n. We have that

$$\|X - X_n\|_F^2 = \sum_{l=n+1}^{\min(N,M)} \sigma_l^2(X) \leq M^{1-\frac{2}{p}} \cdot \left(d_n^{(p)}\left(P, \mathbb{R}^N\right)\right)^2 \tag{25}$$

due to (21) and (23). Thus, for each positive integer k there can be at most k (nonzero) columns of X, $\mathbf{p}_j \in P$, with the property that

$$\|\mathbf{p}_j - \Pi_{\mathcal{S}} \mathbf{p}_j\|_2^2 \geq \frac{M^{1-\frac{2}{p}}}{k} \cdot \left(d_n^{(p)}\left(P, \mathbb{R}^N\right)\right)^2. \tag{26}$$

Setting $k = M - m + 1$, we see that (17) must hold in order for (25) to hold.

To finish, we note that the subspace $\mathcal{S}$ above is spanned by the n left singular vectors of X associated with its n largest singular values. These can be computed deterministically in $O\left(NM \cdot \min\{N, M\}\right)$-time as part of the full singular value decomposition of X, although significantly faster (randomized) approximation algorithms exist (see, e.g., [19, 7]). The stated runtime complexity follows given our assumption that $M \geq N \geq \log M$. □

We may now use Lemma 2 to prove that a best-fit least squares subspace for P will also approximate most of P near-optimally with respect to $d^{(\infty)}$.

Lemma 3. *Let $\xi \in (1, \infty)$, $P = \{\mathbf{p}_0 := \mathbf{0}, \mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^N$ be symmetric, and $n \in \{1, \ldots, N\}$. Then, there is an $O\left(MN^2\right)$-time*[6] *algorithm which outputs both an n-dimensional subspace $\mathcal{S} \subset \mathbb{R}^N$, and a symmetric subset $P' \subset P$ with $|P'| \geq \lceil (1 - 1/\xi)M \rceil + 1$, such that*

$$d^{(\infty)}(P', \mathcal{S}) < \sqrt{\xi} \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right). \tag{27}$$

Proof: We first order the nonzero elements of P according to (18), and then set

$$P' := \left\{\mathbf{p}_0, \mathbf{p}_{l_1}, \mathbf{p}_{l_2}, \ldots, \mathbf{p}_{l_{\lceil (1-1/\xi)M \rceil}}\right\} \subset P. \tag{28}$$

If P' is not symmetric, continue to add additional points from P until it is (i.e., by adding the negation of each current point in P' to P'). Applying Lemma 2 with $m = \lceil (1 - 1/\xi)M \rceil$, we see that

$$\|\mathbf{p}_{\lceil (1-1/\xi)M \rceil} - \Pi_{\mathcal{S}} \mathbf{p}_{\lceil (1-1/\xi)M \rceil}\|_2^2 \leq \xi \cdot \left(d_n^{(\infty)}\left(P, \mathbb{R}^N\right)\right)^2. \tag{29}$$

[6] Again, we assume that $M \geq N \geq \log M$.

Thus there can be at most $\lfloor M/\xi \rfloor$ (nonzero) columns of X, $\mathbf{p}_j \in P$, with the property that

$$\|\mathbf{p}_j - \Pi_{\mathcal{S}} \mathbf{p}_j\|_2^2 \geq \xi \cdot \left(d_n^{(\infty)} \left(P, \mathbb{R}^N \right) \right)^2. \tag{30}$$

By the ordering (18), the associated indices j must be contained in $\{\ell_{\lceil (1-1/\xi)M \rceil + 1}, \ldots, \ell_M\}$, hence $P' \subset P$ will satisfy (27).

By Lemma 2, a suitable set $\mathcal{S}$ can be found in $O\left(NM \cdot \min\{N, M\}\right)$-time. Having computed (the singular value decomposition of) X_n, the ordering in (18) can then be determined in $O(NM + M \log M)$-time. Finally, the symmetry of P' can be ensured in $O(NM \log M)$-time by, e.g., ordering the points of P' lexicographically, and then performing a binary search for the negation of each point in order to ensure its inclusion. The stated runtime complexity follows given our assumption that $M \geq N \geq \log M$. □

Remark 1. *The runtime complexity quoted in Lemma 2 and consequently also Lemma 3 and Lemma 4 is dominated by the time required to compute* X_n (24) *via the full singular value decomposition of* X (19)*. However, computing* X_n *this way is computationally wasteful when* $n \ll \min\{N, M\}$*. Note that it suffices to find a* $O(n)$*-dimensional matrix,* $\tilde{X}_n \in \mathbb{R}^{N \times M}$*, with the property that*

$$\|X - \tilde{X}_n\|_F \leq C \cdot \|X - X_n\|_F \tag{31}$$

for a suitably small constant C*. Taking* $\tilde{\mathcal{S}}$ *to be the column span of* $\tilde{X}_n$ *in the proof of Lemma 3 then produces a similarly sized subset* $P' \subset P$ *satisfying* $d^{(\infty)}(P', \tilde{\mathcal{S}}) \leq C\sqrt{\xi} \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right)$*. A tremendous number of methods have been developed for rapidly computing an* $\tilde{X}_n$ *as above (see, e.g., [19, 7]). In particular, we note here that there exists a modest absolute constant* $C \in \mathbb{R}^+$ *such that a randomly constructed matrix* $\tilde{X}_n$ *of rank* $\max\{2n, 7\}$ *will satisfy* (31) *with probability* > 0.9.[7] *Furthermore, this matrix can always be constructed in* $O(NMn + Nn^2)$*-time.*

An argument similar to the proof of Lemma 3 now allows us to prove that a best-fit least squares subspace for P will also approximate most of P near-optimally with respect to $d^{(p)}$, for any $p \in (2, \infty)$.

Lemma 4. *Let* $p \in (2, \infty)$, $\xi \in (1, M/2]$, $P = \{\mathbf{p}_0, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^N$ *be symmetric, and* $n \in \{1, \ldots, N\}$*. Then, there is an* $O\left(MN^2\right)$*-time*[8] *algorithm which outputs both an* n*-dimensional subspace* $\mathcal{S} \subset \mathbb{R}^N$*, and a symmetric subset* $P' \subset P$ *with* $|P'| \geq \lceil (1 - 1/\xi)\, M \rceil + 1$*, such that*

$$d^{(p)}(P', \mathcal{S}) \leq \sqrt{2\xi} \cdot d_n^{(p)}\left(P, \mathbb{R}^N\right). \tag{32}$$

Proof: We again order the nonzero elements of P according to (18), and then define P' as above

[7] See Theorem 10.7 from [7] for more details concerning the constant C, etc.. Also, note that the probability of satisfying (31) can be boosted as close to 1 as desired by constructing several different $\tilde{X}_n$ matrices independently, and then choosing the most accurate one.

[8] Again, we assume that $M \geq N \geq \log M$.

in (28). From Lemma 2 with $m = \lceil (1 - 1/\xi)\, M \rceil$ we obtain that

$$(d^{(p)}(P', \mathcal{S}))^p = \sum_{j=1}^{m} \|\mathbf{p}_{\ell_j} - \Pi_{\mathcal{S}} \mathbf{p}_{\ell_j}\|_2^p \tag{33}$$

$$\leq \sum_{j=1}^{m} \left(\frac{M^{1-\frac{2}{p}}}{M - j + 1} \cdot \left(d_n^{(p)} \left(P, \mathbb{R}^N\right) \right)^2 \right)^{p/2} \tag{34}$$

$$= M^{\frac{p}{2}-1} \left(d_n^{(p)} \left(P, \mathbb{R}^N\right) \right)^p \sum_{j=M-m+1}^{M} j^{-p/2} \tag{35}$$

$$\leq M^{\frac{p}{2}-1} \left(d_n^{(p)} \left(P, \mathbb{R}^N\right) \right)^p \int_{M-m}^{M} x^{-p/2} dx \tag{36}$$

$$= \frac{\left(1 - \frac{m}{M}\right)^{1-\frac{p}{2}} - 1}{\frac{p}{2} - 1} \left(d_n^{(p)} \left(P, \mathbb{R}^N\right) \right)^p . \tag{37}$$

Set $\delta := m/M - (1 - 1/\xi) < 1/M$. It is not difficult to see that $1/\xi - \delta \in (0, 1)$ since $\xi \in (1, M/2]$. Thus,

$$\left(1 - \frac{m}{M}\right)^{1-\frac{p}{2}} = \left(\frac{1}{\xi} - \delta\right)^{1-\frac{p}{2}} < \left(\frac{\xi}{1 - \frac{\xi}{M}}\right)^{\frac{p}{2}-1} \leq (2\xi)^{\frac{p}{2}-1}, \tag{38}$$

which now allows us to bound (37) as follows:

$$(d^{(p)}(P', \mathcal{S}))^p \leq \frac{\left(1 - \frac{m}{M}\right)^{1-\frac{p}{2}} - 1}{\frac{p}{2} - 1} \left(d_n^{(p)} \left(P, \mathbb{R}^N\right) \right)^p < \frac{(2\xi)^{\frac{p}{2}-1} - 1}{\frac{p}{2} - 1} \cdot \left(d_n^{(p)} \left(P, \mathbb{R}^N\right) \right)^p . \tag{39}$$

Now let $f_\xi : [2, \infty) \to \mathbb{R}^+$ be defined by

$$f_\xi(p) := \begin{cases} \left(\frac{(2\xi)^{\frac{p}{2}-1} - 1}{\frac{p}{2} - 1} \right)^{\frac{1}{p}} & \text{if } p > 2 \\ \sqrt{\ln(2\xi)} & \text{if } p = 2 \end{cases} . \tag{40}$$

One can see that f_ξ is continuous on $[2, \infty)$ via l'Hopital's rule. Furthermore, the Taylor series expansion of $(2\xi)^{\frac{p}{2}-1}$ reveals that

$$f_\xi(p) = \left(\ln(2\xi) \cdot \sum_{n=0}^{\infty} \frac{\left((\frac{p}{2} - 1) \ln(2\xi) \right)^n}{(n+1)!} \right)^{\frac{1}{p}} \leq \left(\ln(2\xi) \cdot (2\xi)^{\frac{p}{2}-1} \right)^{\frac{1}{p}} = \left(\frac{\ln(2\xi)}{2\xi} \right)^{\frac{1}{p}} \cdot \sqrt{2\xi} \tag{41}$$

for all $p \in [2, \infty)$. Thus, (39) yields (32) as desired. As the set P' is constructed in the same way as in the proof of Lemma 3, the runtime analysis given there carries over directly. □

Remark 2. *Note that the ordered distances* (18) *between the points in* P *and the subspace* $\mathcal{S}$ *from Lemma 3 satisfy*

$$\left\| \mathbf{p}_{l_{\lceil (1-1/\xi)M \rceil}} - \Pi_{\mathcal{S}} \mathbf{p}_{l_{\lceil (1-1/\xi)M \rceil}} \right\|_2 \leq \sqrt{\xi} \cdot d_n^{(\infty)} \left(P, \mathbb{R}^N\right). \tag{42}$$

We can use this information to bound $d_n^{(\infty)} \left(P, \mathbb{R}^N\right)$ *from above and below. Set*

$$\alpha := \frac{\|\mathbf{p}_{l_M} - \Pi_{\mathcal{S}} \mathbf{p}_{l_M}\|_2}{\left\| \mathbf{p}_{l_{\lceil (1-1/\xi)M \rceil}} - \Pi_{\mathcal{S}} \mathbf{p}_{l_{\lceil (1-1/\xi)M \rceil}} \right\|_2}. \tag{43}$$

We now have

$$d_n^{(\infty)}\left(P,\mathbb{R}^N\right) \leq \|\mathbf{p}_{l_M} - \Pi_{\mathcal{S}}\mathbf{p}_{l_M}\|_2 = \alpha\cdot \left\|\mathbf{p}_{l_{\lceil(1-1/\xi)M\rceil}} - \Pi_{\mathcal{S}}\mathbf{p}_{l_{\lceil(1-1/\xi)M\rceil}}\right\|_2 \leq \alpha\sqrt{\xi}\cdot d_n^{(\infty)}\left(P,\mathbb{R}^N\right). \quad (44)$$

Thus, computing α allows us to estimate $d_n^{(\infty)}\left(P,\mathbb{R}^N\right)$. If α is sufficiently small, $\mathcal{S}$ will itself be a passible approximation to an optimal subspace $\mathcal{A}_{\rm opt}$. Similarly, if the $P' \subset P$ and $\mathcal{S}$ from Lemma 4 satisfy

$$d^{(p)}(P,\mathcal{S}) \leq \alpha \cdot d^{(p)}(P',\mathcal{S}) \quad (45)$$

for a modest $\alpha \in \mathbb{R}^+$, then we may infer that $\mathcal{S}$ is a near-optimal subspace for P.

Lemmas 3 and 4 now allow us to establish the main results of this section. We will first prove the main dimensionality reduction result for the $p = \infty$ case.

Theorem 3. *Let $\xi \in (1,\infty)$, $P = \{\mathbf{p}_0,\dots,\mathbf{p}_M\} \subset \mathbb{R}^N$ be symmetric, and $n \in \{1,\dots,N\}$. Then, there is an $O\left(\frac{\xi}{\xi-1}\cdot MN^2 + N\cdot n^2\log_\xi^2 M\right)$-time algorithm which outputs an at most $(n\cdot\lceil\log_\xi M\rceil)$-dimensional subspace $\mathcal{S} \subset \mathbb{R}^N$ with*

$$d_n^{(\infty)}(P,\mathcal{S}) \leq \left(1+\sqrt{\xi}\right)\cdot d_n^{(\infty)}\left(P,\mathbb{R}^N\right). \quad (46)$$

Proof: Let $\mathcal{S} \subset \mathbb{R}^D$ be an $\tilde{n}$-dimensional subspace with $\tilde{n} \geq n$, and $\mathcal{A} \in \Gamma_n(\mathbb{R}^D)$. We have that

$$\begin{aligned}
d_n^{(\infty)}(P,\mathcal{S}) &\leq \max_{\mathbf{p}_j\in P}\|\mathbf{p}_j - \Pi_{\mathcal{S}}\Pi_{\mathcal{A}}\mathbf{p}_j\|_2 && (47)\\
&\leq \max_{\mathbf{p}_j\in P}\left(\|\mathbf{p}_j - \Pi_{\mathcal{S}}\mathbf{p}_j\|_2 + \|\Pi_{\mathcal{S}}\mathbf{p}_j - \Pi_{\mathcal{S}}\Pi_{\mathcal{A}}\mathbf{p}_j\|_2\right) && (48)\\
&\leq \max_{\mathbf{p}_j\in P}\|\mathbf{p}_j - \Pi_{\mathcal{S}}\mathbf{p}_j\|_2 + \max_{\mathbf{p}_j\in P}\|\mathbf{p}_j - \Pi_{\mathcal{A}}\mathbf{p}_j\|_2. && (49)
\end{aligned}$$

The fact that this holds for all $\mathcal{A} \in \Gamma_n(\mathbb{R}^D)$ now immediately implies that

$$d_n^{(\infty)}(P,\mathcal{S}) \leq d^{(\infty)}(P,\mathcal{S}) + d_n^{(\infty)}\left(P,\mathbb{R}^N\right). \quad (50)$$

It remains to make a good choice for the subspace $\mathcal{S}$. More precisely, we would like to find a subspace $\mathcal{S}$ with $d^{(\infty)}(P,\mathcal{S}) \leq \sqrt{\xi}\cdot d_n^{(\infty)}\left(P,\mathbb{R}^N\right)$ so that we can obtain (46) from (50).

Appealing to Lemma 3, we note that we can find a sufficiently accurate n-dimensional subspace, $\mathcal{S}^1$, for a large symmetric subset $P' \subset P$ with $|P'| \geq \lceil(1-1/\xi)M\rceil + 1$. It remains to find a similarly accurate subspace for the rest of P. Set $P_2 := (P - P') \cup \{0\}$, noting that P_2 will be a symmetric point set with $|P_2| \leq M/\xi$. We may now apply Lemma 3 to P_2 in order to find a second n-dimensional subspace, $\mathcal{S}^2$, which approximates all but at most M/ξ^2 elements of P_2 to within the desired $\sqrt{\xi}\cdot d_n^{(\infty)}\left(P,\mathbb{R}^N\right)$-accuracy. More generally, we can see that iterating Lemma 3 at most $\lceil\log_\xi M\rceil$-times in this fashion will produce a collection of at most $\lceil\log_\xi M\rceil$ different n-dimensional subspaces, $\mathcal{S}^1,\dots,\mathcal{S}^{\lceil\log_\xi M\rceil}$, which will collectively approximate all of P to the desired $\sqrt{\xi}\cdot d_n^{(\infty)}\left(P,\mathbb{R}^N\right)$-accuracy. We now set

$$\mathcal{S} := \operatorname{span}\left(\mathcal{S}^1 \cup \cdots \cup \mathcal{S}^{\lceil\log_\xi M\rceil}\right). \quad (51)$$

It is not difficult to see that $\mathcal{S}$ will be at most $(n \cdot \lceil \log_\xi M \rceil)$-dimensional. Furthermore, the at most $\lceil \log_\xi M \rceil$ applications of Lemma 3 will induce a runtime of complexity of

$$O\left(\sum_{j=0}^{\lceil \log_\xi M \rceil - 1} \frac{NM \cdot \min\{N, M/\xi^j\}}{\xi^j} \right) = O\left(\frac{\xi}{\xi - 1} \cdot MN^2 \right). \tag{52}$$

Finally, we note that an orthonormal basis for $\mathcal{S}$ can be computed in $O\left(N \cdot n^2 \log_\xi^2 M\right)$-time via Gram–Schmidt. The stated result follows. □

A similar argument now allows us to prove a dimensionality reduction result for the $p \in (2, \infty)$ case.

Theorem 4. *Let $\xi \in (1, \infty)$, $P = \{\mathbf{p}_0, \dots, \mathbf{p}_M\} \subset \mathbb{R}^N$ be symmetric, and $n \in \{1, \dots, N\}$. Then, there is an $O\left(\frac{\xi}{\xi-1} \cdot MN^2 + N \cdot n^2 \log_\xi^2 M\right)$-time algorithm which outputs an at most $(n \cdot \lceil \log_\xi M \rceil)$-dimensional subspace $\mathcal{S} \subset \mathbb{R}^N$ such that*

$$d_n^{(p)}(P, \mathcal{S}) \leq \left(1 + \lceil \log_\xi M \rceil^{1/p} \sqrt{2\xi}\right) \cdot d_n^{(p)}\left(P, \mathbb{R}^N\right). \tag{53}$$

Proof: Let $\mathcal{S} \subset \mathbb{R}^D$ be an $\tilde{n}$-dimensional subspace with $\tilde{n} \geq n$, and $\mathcal{A} \in \Gamma_n(\mathbb{R}^D)$. We have that

$$d_n^{(p)}(P, \mathcal{S}) \leq \left(\sum_{\mathbf{p}_j \in P} \|\mathbf{p}_j - \Pi_{\mathcal{S}} \Pi_{\mathcal{A}} \mathbf{p}_j\|_2^p \right)^{1/p} \leq \left(\sum_{\mathbf{p}_j \in P} \|\mathbf{p}_j - \Pi_{\mathcal{S}} \mathbf{p}_j\|_2^p \right)^{1/p} + \left(\sum_{\mathbf{p}_j \in P} \|\mathbf{p}_j - \Pi_{\mathcal{A}} \mathbf{p}_j\|_2^p \right)^{1/p}.$$

The fact that this holds for all $\mathcal{A} \in \Gamma_n(\mathbb{R}^D)$ now again implies that

$$d_n^{(p)}(P, \mathcal{S}) \leq d^{(p)}(P, \mathcal{S}) + d_n^{(p)}\left(P, \mathbb{R}^N\right). \tag{54}$$

The subspace $\mathcal{S}$ is now chosen in the same fashion as in the proof of Theorem 3. That is, $\mathcal{S}$ is taken to be the span of the union of the at most $\lceil \log_\xi M \rceil$ recursively constructed subspaces $\mathcal{S}^1, \dots, \mathcal{S}^{\lceil \log_\xi M \rceil}$ discussed therein (i.e., see (51)).

Consider the recursive partition $P = \bigcup_{i=1}^{\lceil \log_\xi M \rceil} P_i$ used to construct the subspaces $\mathcal{S}^1, \dots, \mathcal{S}^{\lceil \log_\xi M \rceil}$ in the proof of Theorem 3. Each n-dimensional subspace $\mathcal{S}^i$ will approximate P_i well in the sense of $d^{(p)}$ by Lemma 4. That is,

$$d^{(p)}(P_i, \mathcal{S}^i) \leq \sqrt{2\xi} \cdot d_n^{(p)}\left(P_{i-1}, \mathbb{R}^N\right) \tag{55}$$

holds for all $1 \leq i \leq \lceil \log_\xi M \rceil$ (here, $P_0 := P$). Using (55) we can see that

$$(d^{(p)}(P, \mathcal{S}))^p = \sum_{i=1}^{\lceil \log_\xi M \rceil} (d^{(p)}(P_i, \mathcal{S}))^p \leq \sum_{i=1}^{\lceil \log_\xi M \rceil} (d^{(p)}(P_i, \mathcal{S}^i))^p \tag{56}$$

$$\leq (2\xi)^{p/2} \cdot \sum_{i=1}^{\lceil \log_\xi M \rceil} \left(d_n^{(p)}\left(P_{i-1}, \mathbb{R}^N\right)\right)^p \tag{57}$$

$$\leq \lceil \log_\xi M \rceil (2\xi)^{p/2} \cdot \left(d_n^{(p)}\left(P, \mathbb{R}^N\right)\right)^p. \tag{58}$$

The desired bound (53) now follows from (54) and (58). As the construction of $\mathcal{S}$ is the same as for Theorem 3, the runtime analysis there carries over. The stated result follows. □

Remark 3. *Recalling Remark 1, we note that the runtime complexities quoted in both Theorems 3 and 4 can be reduced by using faster randomized row-rank approximation methods in Lemmas 3 and 4, respectively. Furthermore, we point out that one can use the ideas from Remark 2 in order to guarantee a, e.g.,* $2\sqrt{\xi} \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right)$*-accurate approximation to* P *with potentially fewer than* $\lceil \log_\xi M \rceil$ *applications of Lemma 3. This can be achieved by terminating the iterative applications of Lemma 3 described in the proof of Theorem 3 once* α *from* (43) *falls below* 2*. Similarly, the iterative applications of Lemma 4 described in the proof of Theorem 4 can be terminated without seriously degrading accuracy as soon as* $\alpha := d^{(p)}(P, \mathcal{S})/d^{(p)}(P', \mathcal{S})$ *falls below a user prescribed threshold. Finally, it worth noting that the accuracy of Theorem 3 (and Theorem 4) can be improved in practice by replacing* $P \setminus P'$ *with* $(I - \Pi_{\mathcal{S}})(P \setminus P')$ *after each iteration of Lemma 3 (or Lemma 4). This allows subsequent iterations to strictly improve on the progress made in previous iterations.*

Remark 4. *It is interesting to note that the greedy method utilized in Section 3 is closely related to the meta algorithm outlined in [5] when* $p \in (2, \infty)$*. As a result, it may be possible to improve the* $\lceil \log_\xi M \rceil^{1/p}$*-factor in* (53) *by combining Lemma 4 with the proof techniques of Theorem 11.2 in [5]. Verifying this with a rigorous proof is left as future work.*

4 A Fast Algorithm for $p = \infty$ Subspace Approximation

In this section we demonstrate that the dimensionality reduction results developed above can be combined with computational techniques for computing the John ellipsoid of a point set in order to produce a fast approximation algorithm for the $p = \infty$ problem. The following result establishes the speed and accuracy of this approach.

Theorem 5. *Let* $P = \{\mathbf{p}_0, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^N$ *be symmetric, and* $n \in \{1, \ldots, N\}$*. Then, one can calculate an* $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ *with*

$$d^{(\infty)}(P, \mathcal{A}) \leq C\sqrt{n \cdot \log M} \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right) \tag{59}$$

in $O\left(MN^2 + Mn^2 \cdot \log^2 M \cdot \log(n \log M)\right)$*-time. Here* $C \in \mathbb{R}^+$ *is an absolute constant.*

Before proving Theorem 5 we will need an intermediate lemma. Lemma 5 shows that the projection of the dataset P onto its Theorem 3 subspace $\mathcal{S}$, $P' := \Pi_{\mathcal{S}} P$, will be approximated near-optimally by an n-dimensional subspace obtained from its John ellipsoid.

Lemma 5. *Suppose that* $P' = \{\mathbf{p}_0, \ldots, \mathbf{p}_M\} \subset \mathcal{S} \in \Gamma_{\tilde{m}}\left(\mathbb{R}^N\right)$ *is symmetric. Let* $\epsilon \in (0, \infty)$*,* $\xi \in (1, \infty)$*, and* $n \in \{1, \ldots, N\}$ *be such that* $n \leq \tilde{m} \leq n\lceil \log_\xi M \rceil$*. Then, one can calculate an* $\mathcal{H} \in \Gamma_n(\mathcal{S})$ *with*

$$d^{(\infty)}\left(P', \mathcal{H}\right) \leq \sqrt{(1 + \epsilon)\tilde{m}} \cdot d_n^{(\infty)}\left(P', \mathcal{S}\right) \tag{60}$$

in $O\left(MN \cdot n \log_\xi M + Mn^2 \cdot \log_\xi^2 M \cdot \left(\log(n \log_\xi M) + 1/\epsilon\right)\right)$*-time.*

Proof: Let $B_{\mathcal{S}}$ be an orthonormal basis of $\mathcal{S}$ (assumed to be provided). We will also work with P' expressed in terms of its $B_{\mathcal{S}}$ coordinates, $P'' \subset \mathbb{R}^{\tilde{m}}$. Compute an ellipsoid $\mathcal{E} := \left\{\mathbf{x} \mid \mathbf{x}^T Q \mathbf{x} \leq 1\right\} \subset \mathbb{R}^{\tilde{m}}$ such that

$$\mathcal{E} \subseteq \mathrm{CH}\left(P''\right) \subseteq \sqrt{(1 + \epsilon)\tilde{m}} \cdot \mathcal{E} \tag{61}$$

in $O\left(M\tilde{m}^2(\log \tilde{m} + 1/\epsilon)\right)$-time [17]. Next, let $\mathcal{A}'_{\mathcal{E}} \subset \mathbb{R}^{\tilde{m}}$ be the subspace spanned by the n eigenvectors of Q associated with $\sigma_{\tilde{m}}(Q), \ldots, \sigma_{\tilde{m}-n+1}(Q)$, and let $\mathcal{A}_{\mathcal{E}} \subset \mathcal{S} \subset \mathbb{R}^N$ be $\mathcal{A}'_{\mathcal{E}}$ re-expressed as an n-dimensional subspace of the span of $B_{\mathcal{S}}$. Finally, let $\mathcal{A}'_{\mathrm{opt}} \in \Gamma_n\left(\mathbb{R}^{\tilde{m}}\right)$ be an optimal subspace for $\mathrm{CH}(P'')$, so that $d^{(\infty)}\left(\mathrm{CH}(P''), \mathcal{A}'_{\mathrm{opt}}\right) = d_n^{(\infty)}\left(\mathrm{CH}(P''), \mathbb{R}^{\tilde{m}}\right)$.

We can now see that

$$
\begin{aligned}
d^{(\infty)}\left(P', \mathcal{A}_{\mathcal{E}}\right) &= d^{(\infty)}\left(P'', \mathcal{A}'_{\mathcal{E}}\right) && \text{(Change of Coordinates)}\\
&= d^{(\infty)}\left(\mathrm{CH}\left(P''\right), \mathcal{A}'_{\mathcal{E}}\right) && \text{(Proof of Lemma 1 Parts (4)\&(5))}\\
&\leq d^{(\infty)}\left(\sqrt{(1+\epsilon)\tilde{m}} \cdot \mathcal{E}, \mathcal{A}'_{\mathcal{E}}\right) && \text{(Proof of Part (4) of Lemma 1)}\\
&= \sqrt{(1+\epsilon)\tilde{m}} \cdot d^{(\infty)}\left(\mathcal{E}, \mathcal{A}'_{\mathcal{E}}\right) && \text{(Scalability of } d^{(\infty)}\text{)}\\
&\leq \sqrt{(1+\epsilon)\tilde{m}} \cdot d^{(\infty)}\left(\mathcal{E}, \mathcal{A}'_{\text{opt}}\right) && \text{(Part (6) of Lemma 1)}\\
&\leq \sqrt{(1+\epsilon)\tilde{m}} \cdot d^{(\infty)}\left(\mathrm{CH}\left(P''\right), \mathcal{A}'_{\text{opt}}\right) && \text{(Proof of Part (4) of Lemma 1)}\\
&= \sqrt{(1+\epsilon)\tilde{m}} \cdot d^{(\infty)}_n\left(P'', \mathbb{R}^{\tilde{m}}\right) && \text{(Part (5) of Lemma 1).}
\end{aligned} \tag{62}
$$

After noting that $d^{(\infty)}_n\left(P'', \mathbb{R}^{\tilde{m}}\right) = d^{(\infty)}_n\left(P', \mathcal{S}\right)$, we can see that (62) implies that

$$
d^{(\infty)}\left(P', \mathcal{A}_{\mathcal{E}}\right) \leq \sqrt{(1+\epsilon)\tilde{m}} \cdot d^{(\infty)}_n\left(P', \mathcal{S}\right). \tag{63}
$$

Thus, we have achieved (60).

The runtime complexity can be accounted for as follows: Computing P'' from P' can be done in $O(MN \cdot n \log_\xi M)$-time, after which $\mathcal{A}'_{\mathcal{E}}$ can be found in $O\left(M \cdot n^2 \log^2_\xi M \cdot \left(\log(n \log_\xi M) + 1/\epsilon\right)\right)$-time via [17]. Finally, a basis for $\mathcal{A}_{\mathcal{E}}$ can be computed in $O(N \cdot n^2 \log_\xi M)$-time once $\mathcal{A}'_{\mathcal{E}}$ is known. The stated runtime complexity follows. □

We are now prepared to prove Theorem 5.

Proof of Theorem 5: Choose $\epsilon \in (0, \infty)$ and $\xi \in (1, \infty)$. Compute $\mathcal{S} \in \Gamma_{\tilde{m}}\left(\mathbb{R}^N\right)$, with $n \leq \tilde{m} \leq n\lceil \log_\xi M \rceil$, via Theorem 3/Remark 3. Let $P' := \Pi_{\mathcal{S}} P \subset \mathcal{S} \subset \mathbb{R}^N$ be the projection of P onto $\mathcal{S}$. Finally, compute a subspace $\mathcal{H} \in \Gamma_n(\mathcal{S})$ satisfying (60) via Lemma 5.

Fix an arbitrary $\mathcal{A} \in \Gamma_n(\mathcal{S})$, noting that $\Pi_{\mathcal{A}}\Pi_{\mathcal{S}} = \Pi_{\mathcal{A}}$ since $\mathcal{A} \subset \mathcal{S}$. Then, there exists a $\mathbf{y} \in P$ such that

$$
\|\Pi_{\mathcal{S}}\mathbf{y} - \Pi_{\mathcal{A}}\mathbf{y}\|_2 = \|\Pi_{\mathcal{S}}\mathbf{y} - \Pi_{\mathcal{A}}\Pi_{\mathcal{S}}\mathbf{y}\|_2 = d^{(\infty)}\left(P', \mathcal{A}\right) \geq d^{(\infty)}_n\left(P', \mathcal{S}\right). \tag{64}
$$

Now fix an arbitrary $\mathbf{x} \in P$. Combining (60) and (64), we can see that

$$
\begin{aligned}
\|\Pi_{\mathcal{S}}\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\|_2^2 &= \|\Pi_{\mathcal{S}}\mathbf{x} - \Pi_{\mathcal{H}}\Pi_{\mathcal{S}}\mathbf{x}\|_2^2 && \text{(Since } \mathcal{H} \subset \mathcal{S}\text{)}\\
&\leq \left(d^{(\infty)}\left(P', \mathcal{H}\right)\right)^2 && \text{(Def. of } d^{(\infty)}\text{)}\\
&\leq (1+\epsilon)\tilde{m} \cdot \left(d^{(\infty)}_n\left(P', \mathcal{S}\right)\right)^2 && \text{(Using (60))}\\
&\leq (1+\epsilon)\tilde{m} \cdot \|\Pi_{\mathcal{S}}\mathbf{y} - \Pi_{\mathcal{A}}\mathbf{y}\|_2^2 && \text{(Using (64))}\\
&\leq (1+\epsilon)\tilde{m} \cdot \left(\|\Pi_{\mathcal{S}}\mathbf{y} - \Pi_{\mathcal{A}}\mathbf{y}\|_2^2 + \|\Pi_{\mathcal{S}^\perp}\mathbf{y}\|_2^2\right) && \\
&= (1+\epsilon)\tilde{m} \cdot \left(\|\Pi_{\mathcal{S}}\left(\mathbf{y} - \Pi_{\mathcal{A}}\mathbf{y}\right)\|_2^2 + \|\Pi_{\mathcal{S}^\perp}\left(\mathbf{y} - \Pi_{\mathcal{A}}\mathbf{y}\right)\|_2^2\right) && \text{(Since } \mathcal{A} \subset \mathcal{S}\text{)}\\
&= (1+\epsilon)\tilde{m} \cdot \|\mathbf{y} - \Pi_{\mathcal{A}}\mathbf{y}\|_2^2 && \text{(Pythagoras)}\\
&\leq (1+\epsilon)\tilde{m} \cdot \left(d^{(\infty)}\left(P, \mathcal{A}\right)\right)^2 && \text{(Def. of } d^{(\infty)}\text{).}
\end{aligned} \tag{65}
$$

The fact that (65) holds for all $\mathcal{A} \in \Gamma_n(\mathcal{S})$ now implies that

$$
\|\Pi_{\mathcal{S}}\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\|_2 \leq \sqrt{(1+\epsilon)\tilde{m}} \cdot d^{(\infty)}_n\left(P, \mathcal{S}\right). \tag{66}
$$

Continuing, we now have that

$$
\begin{aligned}
\|\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\|_2 &= \sqrt{\|\Pi_{\mathcal{S}}\left(\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\right)\|_2^2 + \|\Pi_{\mathcal{S}^\perp}\left(\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\right)\|_2^2} && \text{(Pythagoras)}\\
&= \sqrt{\|\Pi_{\mathcal{S}}\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\|_2^2 + \|\Pi_{\mathcal{S}^\perp}\mathbf{x}\|_2^2} && \text{(Since } \mathcal{H} \subset \mathcal{S}\text{)}\\
&\leq \sqrt{(1+\epsilon)\tilde{m} \cdot \left(d^{(\infty)}_n\left(P, \mathcal{S}\right)\right)^2 + \left(d^{(\infty)}\left(P, S\right)\right)^2} && \text{(By (66), Def. } d^{(\infty)}\text{).}
\end{aligned} \tag{67}
$$

Recalling that $\mathcal{S}$ was provided by Theorem 3, we obtain

$$\|\mathbf{x} - \Pi_{\mathcal{H}}\mathbf{x}\|_2 \leq \sqrt{(1+\epsilon)\left(1+\sqrt{\xi}\right)^2 \tilde{m} \cdot \left(d_n^{(\infty)}\left(P, \mathbb{R}^N\right)\right)^2 + \xi\left(d_n^{(\infty)}\left(P, \mathbb{R}^N\right)\right)^2}. \tag{68}$$

The fact that (68) holds for all $\mathbf{x} \in P$ yields (59).

The runtime complexity can be accounted for as follows: Computing $\mathcal{S}$ via Theorem 3 can be accomplished in $O\left(\frac{\xi}{\xi-1} \cdot MN^2 + N \cdot n^2 \log_\xi^2 M\right)$-time. Computing P' from P can be done in $O(MN \cdot n \log_\xi M)$-time. Finally, computing $\mathcal{H} \in \Gamma_n(\mathcal{S})$ via Lemma 5 can be accomplished in $O\left(MN \cdot n \log_\xi M + Mn^2 \cdot \log_\xi^2 M \cdot \left(\log(n \log_\xi M) + 1/\epsilon\right)\right)$-time. □

Remark 5. The more precise accuracy bound in terms of the parameters ϵ and ξ derived in the proof of the theorem predicts that one can find a set $\mathcal{A}$ that satisfies

$$d^{(\infty)}(P, \mathcal{A}) \leq \left(\sqrt{(1+\epsilon)\left(1+\sqrt{\xi}\right)^2 n \lceil \log_\xi M \rceil + \xi}\right) \cdot d_n^{(\infty)}\left(P, \mathbb{R}^N\right) \tag{69}$$

in $O\left(\frac{\xi}{\xi-1} \cdot MN^2 + Mn^2 \cdot \log_\xi^2 M \cdot \left(\log(n \log_\xi M) + 1/\epsilon\right)\right)$-time. Choosing ϵ small and ξ to minimize the accuracy bound to find that one can achieve $C < 10$. Finally, we note that the runtime complexity quoted in Theorem 5 can be reduced, along the lines of Remark 1, by using a fast randomized least-squares method instead of a deterministic SVD method.

Acknowledgements

The authors would like to thank Kasturi Varadarajan for the helpful advice he kindly provided to them during the initial preparation of this manuscript. The authors would also like to thank the anonymous reviewers for their many helpful comments and suggestions.

References

[1] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

[2] A. Deshpande, M. Tulsiani, and N. K. Vishnoi. Algorithms and hardness for subspace approximation. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 482–496. SIAM, 2011.

[3] A. Deshpande and K. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 641–650. ACM, 2007.

[4] U. Faigle, W. Kern, and M. Streng. Note on the computational complexity of j-radii of polytopes in $\mathbb{R}^n$. *Mathematical Programming*, 73(1):1–5, 1996.

[5] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011.

[6] P. Gritzmann and V. Klee. On the complexity of some basic problems in computational convexity: I. containment problems. *Discrete Mathematics*, 136(1):129–174, 1994.

[7] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[8] S. Har-Peled and K. Varadarajan. Projective clustering in high dimensions using core-sets. In *Proceedings of the eighteenth annual symposium on Computational geometry*, pages 312–318. ACM, 2002.

[9] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[10] F. John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays presented to R. Courant on his 60th Birthday*, pages 187–204, 1948.

[11] L. G. Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2):307–320, 1996.

[12] A. Kolmogorov. Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse. *The Annals of Mathematics*, 37(1):107–110, 1936.

[13] A. N. Kolmogorov, É. Charpentier, A. Lesne, and N. Kapitonovich. *Kolmogorov's Heritage in Mathematics*. Springer, 2007.

[14] P. Kumar and E. A. Yildirim. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and Applications*, 126(1):1–21, 2005.

[15] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. *CoRR*, abs/1202.4044, 2012.

[16] N. D. Shyamalkumar and K. Varadarajan. Efficient subspace approximation algorithms. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 532–540. Society for Industrial and Applied Mathematics, 2007.

[17] M. J. Todd and E. A. Yildirim. On khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. *Discrete Applied Mathematics*, 155(13):1731 – 1744, 2007.

[18] K. Varadarajan, S. Venkatesh, Y. Ye, and J. Zhang. Approximating the radii of point sets. *SIAM Journal on Computing*, 36(6):1764–1776, 2007.

[19] D. S. Watkins. *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*. SIAM, 2007.

[20] Y. Ye and J. Zhang. An improved algorithm for approximating the radii of point sets. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 178–187. Springer, 2003.

A Proof of Lemma 1

We present the proof of each part below:

1. This follows directly from the fact that $d^{(\infty)}(P - \mathbf{x}, \mathcal{A}) = d^{(\infty)}\left(P, \mathcal{A} - \Pi_{\mathcal{S}_{\mathcal{A}}^{\perp}} \mathbf{x}\right)$ for all $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ and $\mathbf{x} \in \mathbb{R}^N$.

2. Let $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ be such that $d^{(\infty)}(\bar{P}, \mathcal{A}) = d_n^{(\infty)}(\bar{P}, \mathbb{R}^N)$. Suppose $\mathbf{a}_{\mathcal{A}}$ is nonzero. Partition $\bar{P}$ into three parts:

   (a) $\bar{P}_1 := \left\{\mathbf{p} \in \bar{P} \mid \langle \mathbf{p}, \mathbf{a}_{\mathcal{A}} \rangle = 0\right\}$
   (b) $\bar{P}_2 := \left\{\mathbf{p} \in \bar{P} \mid \langle \mathbf{p}, \mathbf{a}_{\mathcal{A}} \rangle > 0\right\}$
   (c) $\bar{P}_3 := \left\{\mathbf{p} \in \bar{P} \mid \langle \mathbf{p}, \mathbf{a}_{\mathcal{A}} \rangle < 0\right\}$

   If $\mathbf{p} \in P_1$ then $\|\mathbf{p} - \Pi_{\mathcal{A}}\mathbf{p}\|_2^2 = \|\mathbf{p} - \Pi_{\mathcal{S}_{\mathcal{A}}}\mathbf{p}\|_2^2 + \|\mathbf{a}_{\mathcal{A}}\|_2^2$. This is minimized for all $\mathbf{p} \in P_1$ when $\|\mathbf{a}_{\mathcal{A}}\|_2 = 0$. Next, note that $\mathbf{p} \in P_3$ if and only if $-\mathbf{p} \in P_2$, and that $\mathbf{p} \in P_3$ means $\|\mathbf{p} - \Pi_{\mathcal{A}}\mathbf{p}\|_2 > \|(-\mathbf{p}) - \Pi_{\mathcal{A}}(-\mathbf{p})\|_2$. Thus, we can decrease $d^{(\infty)}(\bar{P}, \mathcal{A})$ by making $\mathbf{a}_{\mathcal{A}}$ shorter (a contradiction).

3. Let $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$ be such that $d^{(\infty)}(P, \mathcal{A}) = d_n^{(\infty)}(P, \mathbb{R}^N)$. We have that

$$\|\bar{\mathbf{p}} - \mathbf{p}_j - \Pi_{\mathcal{S}_{\mathcal{A}}}(\bar{\mathbf{p}} - \mathbf{p}_j)\|_2 = \|\mathbf{p}_j - \bar{\mathbf{p}} - \Pi_{\mathcal{S}_{\mathcal{A}}}(\mathbf{p}_j - \bar{\mathbf{p}})\|_2 = \left\|\mathbf{p}_j - \Pi_{\mathcal{S}_{\mathcal{A}}}\mathbf{p}_j - \Pi_{\mathcal{S}_{\mathcal{A}}^{\perp}}\bar{\mathbf{p}}\right\|_2 \quad (70)$$

$$\leq \|\mathbf{p}_j - \Pi_{\mathcal{A}}\mathbf{p}_j\|_2 + \|\bar{\mathbf{p}} - \Pi_{\mathcal{A}}\bar{\mathbf{p}}\|_2. \quad (71)$$

   Noting that $\|\bar{\mathbf{p}} - \Pi_{\mathcal{A}}\bar{\mathbf{p}}\|_2 \leq d^{(\infty)}(P, \mathcal{A})$ – see part five below for an analogous calculation – concludes the proof.

4. This follows directly from the fact that $d^{(\infty)}(B, \mathcal{A}) \leq d^{(\infty)}(C, \mathcal{A})$ for all $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$.

5. Part four implies $d_n^{(\infty)}(P, \mathbb{R}^N) \leq d_n^{(\infty)}(\mathrm{CH}(P), \mathbb{R}^N)$ since $P \subseteq \mathrm{CH}(P)$. To obtain the other inequality, we recall that every $\mathbf{x} \in \mathrm{CH}(P)$ has $\alpha_j \in [0, 1]$, $j = 1, \ldots, M$, such that

$$\mathbf{x} = \sum_{j=1}^{M} \alpha_j \cdot \mathbf{p}_j, \quad (72)$$

   and

$$\sum_{j=1}^{M} \alpha_j = 1. \quad (73)$$

   Hence, we can see that

$$\|\mathbf{x} - \Pi_{\mathcal{A}}\mathbf{x}\|_2 = \left\|\sum_{j=1}^{M} \alpha_j \cdot (\mathbf{p}_j - \Pi_{\mathcal{S}_{\mathcal{A}}}\mathbf{p}_j - \mathbf{a}_{\mathcal{A}})\right\|_2 \leq \sum_{j=1}^{M} \alpha_j \cdot \|\mathbf{p}_j - \Pi_{\mathcal{A}}\mathbf{p}_j\|_2 \leq d^{(\infty)}(P, \mathcal{A}) \quad (74)$$

   holds for all $\mathbf{x} \in \mathrm{CH}(P)$, and $\mathcal{A} \in \Gamma_n\left(\mathbb{R}^N\right)$. It now follows that $d_n^{(\infty)}(\mathrm{CH}(P), \mathbb{R}^N) \leq d_n^{(\infty)}(P, \mathbb{R}^N)$.

6. Part two tells us that there will be an optimal subspace, since $\mathcal{E}$ is symmetric. Thus, standard results concerning the n-widths of ellipsoids apply (see, e.g., [12, 13]).