Nearly Optimal Linear Embeddings into Very Low Dimensions

Elyot Grant, Chinmay Hegde, Piotr Indyk
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

***Abstract*—We propose algorithms for constructing linear embeddings of a finite dataset $V \subset \mathbb{R}^d$ into a k-dimensional subspace with provable, nearly optimal distortions. First, we propose an exhaustive-search-based algorithm that yields a k-dimensional linear embedding with distortion at most $\epsilon_{opt}(k)+\delta$, for any $\delta > 0$ where $\epsilon_{opt}(k)$ is the smallest achievable distortion over all possible orthonormal embeddings. This algorithm is space-efficient and can be achieved by a single pass over the data V. However, the runtime of this algorithm is exponential in k. Second, we propose a convex-programming-based algorithm that yields an $\mathcal{O}(k/\delta)$-dimensional orthonormal embedding with distortion at most $(1+\delta)\epsilon_{opt}(k)$. The runtime of this algorithm is polynomial in d and independent of k. Several experiments demonstrate the benefits of our approach over conventional linear embedding techniques, such as principal components analysis (PCA) or random projections.**

I. Introduction

In applications dealing with high dimensional metric spaces, an extremely useful tool is the notion of an *embedding* into a space of low dimension [1]. Such embeddings constitute a concise, yet faithful representation of the original metric space, and consequently enable the use of very efficient algorithmic tools and techniques in the smaller space.

In this paper, we consider inputs consisting of a set V of n vectors in $\mathbb{R}^d$, and seek embeddings of the form $f : \mathbb{R}^d \to \mathbb{R}^k$, where $k \ll d, n$. Here, we will consider the domain $\mathbb{R}^d$ and codomain $\mathbb{R}^k$ as being equipped with the Euclidean norm. Our goal is to minimize the *distortion* of the embedding f. Specifically, a function $f : \mathbb{R}^d \to \mathbb{R}^k$ is said to have distortion $\epsilon > 0$ if, for every $x \in V$,

$$1 - \epsilon \leq \frac{\|f(x)\|_2^2}{\|x\|_2^2} \leq 1 + \epsilon. \tag{1}$$

Our goal is to construct embeddings with as small a value of ϵ as possible. As intuition would suggest, there is a tradeoff between k, the number of dimensions of the embedding, and ϵ, the distortion. A celebrated result by Johnson and Lindenstrauss [2] states that given any set V of n vectors in $\mathbb{R}^d$ and $\epsilon > 0$, if $k = O(\log n/\epsilon^2)$, then there exists an embedding $f : \mathbb{R}^d \to \mathbb{R}^k$ that satisfies (1). In fact, the embedding in this theorem is constructed by choosing a random orthonormal projection from $\mathbb{R}^d$ to $\mathbb{R}^k$, scaled by an appropriate factor; such functions f may be called *J-L embeddings*. The random dimensionality reduction technique plays a foundational role in several areas, including high-dimensional similarity search and compressive sensing. (See Section I-B for more details.)

Unfortunately, the bound guaranteed by a J-L embedding cannot be improved (by much): by the result of [3], there exist sets of n points that necessarily require $\Omega(\log n/\epsilon^2 \log(1/\epsilon))$ dimensions in order to be embedded with distortion at most ϵ. However, real-world data often exhibit some low-dimensional structure, which can be potentially exploited to obtain an embedding of distortion ϵ using fewer dimensions. Thus, an intriguing question emerges: given a *particular* set V possessing some hidden low-dimensional structure, is it possible to examine V and find an embedding into *fewer* than $O(\log n/\epsilon^2)$ dimensions, while still obtaining distortion ϵ? This is a question that we address in this paper.

A. Our Contributions

In this paper, we focus on *orthonormal* embeddings. These are embeddings that correspond to an orthonormal projection of $\mathbb{R}^d$ on a k-dimensional subspace. Note that for such embeddings, the right-hand-side inequality of Eq. 1 trivially holds, as projection does not expand distances by the definition. It therefore suffices to focus on the left-hand-side inequality (i.e., the lower bound) in Eq. 1. Also, we can assume without loss of generality assume that all vectors in V have unit norm.

For a fixed set set $V \in \mathbb{R}^d$, we define $\epsilon_{opt}(k)$ to be the smallest achievable distortion over all possible orthonormal embeddings of V into $\mathbb{R}^k$. Our specific contributions are as follows:

- We propose an exhaustive-search-based algorithm that finds a linear embedding with distortion at most $\epsilon_{opt}(k) + \delta$, for any $\delta > 0$. This algorithm is space-efficient and can be achieved by a single pass over the data V. However, the runtime of this algorithm is exponential in k and $1/\delta$.
- We propose a convex-programming-based algorithm that yields an $\mathcal{O}(k)$-dimensional orthonormal embedding with distortion at most $(1+\delta)\epsilon_{opt}(k)$. Concretely, our algorithm produces a $2k$-dimensional orthonormal embedding with distortion at most $2\epsilon_{opt}(k)$. The runtime of this algorithm is polynomial in n and d, and independent of k.
- Several numerical experiments demonstrate the benefits of our approach over conventional linear embedding techniques, such as principal components analysis (PCA) and random projections.

B. Applications

The so-called *curse of dimensionality* poses a central challenge in various signal processing problems including sensing, storage, transmission, and inference. Therefore, an efficient method to transform high dimensional data into more compact

representations would have an impact on a number of applications that currently use random projections. This includes:

- *Neighborhood-preserving projections and hashes*: Suppose that given a high dimensional data set and a set of queries, one wishes to find the nearest neighbors of those queries in the data set. For a set V of n points in $\mathbb{R}^d$, a naïve point query would incur $\mathcal{O}(nd)$ computations. On the other hand, one can bring this computational cost down to $\mathcal{O}(nk)$, if one can reduce the dimensionality from d to k. This approach forms the core of *locality sensitive hashing* (LSH), a popular technique for pattern recognition and information retrieval [4]. Indeed, random linear functions that resemble J-L embeddings play a pivotal role in building space- and time-efficient hash functions.
- *Compressive signal acquisition*: Instead of acquiring (or recording) a high dimensional signal (or image) $x \in \mathbb{R}^d$, the technique of compressive sensing (CS) prescribes recording only a few linear projections (or *measurements*) $y = \Phi x$. A rich, extensive theory specifying the types of allowable projection matrices Φ, as well as efficient algorithms for recovering x from y, has been developed; see, for example, the seminal papers of [5, 6]. Here too, the concept of low-dimensional random projections plays a pivotal role.

C. Prior Work

The classical method to construct lower dimensional data representations is principal components analysis (PCA) [7], which involves orthogonally projecting a dataset into the subspace spanned by the top few eigenvectors of its covariance matrix. However, a global spectral technique such as PCA can potentially contract specific local distances, and hence cannot offer near-optimal distortion guarantees in general.

The optimal trade-offs between distortion and dimensions have been studied for many metrics [1]. Although most of those results were focused on the worst case distortion (along the line of the J-L theorem), there have been several works focused on designing algorithms that approximate the best distortion (see [8] and references therein). However, this research was focused on minimizing the distortion of *non-linear* embeddings, which is a much harder task. In particular, the minimum distortion of a non-linear embedding into a fixed-dimensional space is NP-hard to approximate, even up to a polynomial factor [9]. In contrast, our focus on linear and orthonormal embeddings enables us to obtain strong algorithmic results.

II. Exhaustive Search

First, we develop an algorithm that yields linear embeddings in ℓ_2^k having distortion arbitrarily close to the optimal distortion of any orthonormal embedding. Since the running time of the algorithm is exponential in k, its applications may be limited to cases where an embedding into a very small number of dimensions is desired. However, the algorithm exhibits many characteristics of a polynomial time approximation scheme (PTAS), so it is of theoretical interest.

Formally, our goal shall be to construct a linear embedding f into $\mathbb{R}^k$ having distortion at most $\epsilon_{opt}(k)+\delta$ for an arbitrary $\delta > 0$. Our embedding will not necessarily be orthonormal; it will instead be the the composition of a random J-L embedding and another linear embedding. We establish the following:

Theorem 2.1: Given a set V consisting of n points in $\mathbb{R}^d$, a positive integer $k < d$, and a parameter $\delta > 0$, there exists an algorithm $\mathcal{A}$ that returns an embedding f of V into $\mathbb{R}^k$ having distortion at most $\epsilon_{opt}(k) + \delta$, in time $O(n^2)(k/\delta)^{O(k^2 \log(n)/\delta^2)}$.

Proof: Our algorithm is similar to that used by Badoiu et al., who solve a variety of geometric optimization problems by first reducing the dimension of the input, and then performing a brute force search on the lower dimensional space [10]. Define U to be a k-dimensional subspace of $\mathbb{R}^d$ such that an orthonormal projection into U yields an embedding with the optimal distortion $\epsilon_{opt}(k)$. We let $\{u_1, \ldots, u_k\}$ be an orthonormal basis for U. The first step of our algorithm is to perform a regular J-L embedding $g : \mathbb{R}^d \to \mathbb{R}^q$ on the input. We need to ensure that g does not distort the angles between vectors in U and V too much; specifically, it suffices to obtain the following for every unit basis vector u_i and each unit vector $v \in V$:

$$\langle g(u_i), g(v)\rangle^2 = \langle u_i, v\rangle^2 \pm \frac{\delta}{2k}. \tag{2}$$

Here, the '$\pm$' symbol is used to denote worst case deviations. Such a mapping g can be performed on V with high probability of success, using a codomain having $q = \Theta(\log(n)k/\delta^2)$ dimensions. Note that the bound still holds for squared inner products, because U and V consist entirely of unit vectors. Note also that the high probability of success holds even though we don't know what U is.

Next, we do a brute force search over the unit sphere of $\mathbb{R}^q$ to approximately guess the transformed basis $\{g(u_1), \ldots, g(u_k)\}$. This may seem formidable, but fortunately for our purposes, it suffices to consider only k-tuples of candidates in a $\frac{\delta}{4k}$-net N over unit vectors in $\mathbb{R}^q$. A standard volume-packing argument states that it is possible to construct N with cardinality at most $\left(\frac{4k}{\delta}\right)^{Cq}$ for some absolute constant C. We simply iterate over all possible k-tuples of vectors in N. Suppose $\mathcal{W} = (w_1, \ldots, w_k)$ are the vectors considered in a particular iteration of the search, and define $M_{\mathcal{W}}$ to be the $k \times q$ matrix whose rows are the vectors $(w_1, \ldots, w_k)$. Among all such k-tuples $\mathcal{W} \in N^k$, we identify the k-tuple that minimizes the maximum of the *right-side distortion*

$$\text{RightDistortion}(\mathcal{W}) = \max_{v \in V} \|M_{\mathcal{W}} \cdot g(v)\|_2^2 - 1$$

and the *left-side distortion*

$$\text{LeftDistortion}(\mathcal{W}) = \max_{v \in V} \left(1 - \|M_{\mathcal{W}} \cdot g(v)\|_2^2\right).$$

We let $\mathcal{W}^* = (w_1^*, \ldots, w_k^*)$ be the minimizing set of vectors in N^k, and let M^* be the corresponding matrix. Our algorithm shall output the final linear transformation $f(v) = M^* \cdot g(v)$, the composition of the linear transformation implied by M^* with the J-L mapping g.

We now show that f has distortion at most $\epsilon_{opt}(k) + \delta$. For all i, define w_i' to be the element of N that is closest in

direction to $g(u_i)$. Vector w_i' is then a unit vector whose angle from $g(u_i)$ is at most $\frac{\delta}{4k}$, since N is a $\frac{\delta}{4k}$-net. It follows that

$$\langle w_i', g(v)\rangle = \langle g(u_i), g(v)\rangle \pm \frac{\delta}{4k},$$

and hence, for all $v \in V$,

$$\langle w_i', g(v)\rangle^2 = \langle g(u_i), g(v)\rangle^2 \pm \frac{\delta}{2k} = \langle u_i, v\rangle^2 \pm \frac{\delta}{k},$$

where the latter equality uses the bound in (2). Summing over all values of i, we see that

$$\|M^* \cdot g(v)\|_2^2 = \sum_{i=1}^{k} \langle w_i', g(v)\rangle^2 = \sum_{i=1}^{k} \langle u_i, v\rangle^2 \pm \delta.$$

By our choice of U and the fact that orthonormal projections are contractive, the value of $\sum_{i=1}^{k} \langle u_i, v\rangle^2$ must lie in the range $[1 - \epsilon_{opt}(k), 1]$, and hence:

$$1 - \epsilon_{opt}(k) - \delta \leq \|f(v)\|_2^2 \leq 1 + \delta.$$

From this, it follows that f has distortion at most $\epsilon_{opt}(k) + \delta$.

The time complexity is dominated by the time required to compute the worst case stretch and shinkage for each k-tuple of vectors $(w_1, \ldots, w_k)$ in our $\delta/4k$-net N. Naïvely, there are $O(n)$ vectors in V, and $O((k/\delta)^{qkC})$ k-tuples for some constant $C \in O(1)$, giving a total running time of $O(n)(k/\delta)^{O(k^2 \log(n)/\delta^2)}$. The running time could be potentially reduced by pruning the brute-force search (for example, by only considering k-tuples of vectors in N that are approximately mutually orthogonal), but we do not pursue that direction here. ∎

III. Convex Programming

Next, we develop an alternate approach for constructing orthonormal embeddings that are close to optimal. We start by recasting the problem as an optimization program. Recall that $V = \{v_1, \ldots, v_n\} \subset \mathbb{R}^d$ is a given set of unit vectors. We wish to find $\Phi \in \mathbb{R}^{k \times d}$ with orthonormal rows such that $\max_{i \in [1,\ldots,n]} \left| \|\Phi v_i\|^2 - 1 \right|$ is minimized. This problem is highly non-convex and its exact solution appears to require exponential time. However, we can achieve various *approximate* solutions in polynomial time as follows.

Since Φ is non-expansive, we can drop the absolute value constraint and stipulate the matrix of minimal distortion as the solution of the optimization

$$\begin{aligned} \epsilon_{opt}(k) \quad &= \quad \min \epsilon \\ \text{subject to} \quad &\quad 1 - \epsilon \leq \|\Phi v_i\|_2^2, \ i \in [n] \\ &\quad \Phi\Phi^T = I_{k \times k}. \end{aligned} \tag{3}$$

Consider the semidefinite program

$$\begin{aligned} \gamma_{opt}(k) \quad &= \quad \min \gamma \\ \text{subject to} \quad &\quad v_i^T X v_i \leq \gamma, \quad i \in [n] \\ &\quad \text{trace}(X) = d - k, \quad 0 \preceq X \preceq I. \end{aligned} \tag{4}$$

Clearly, $\gamma_{opt}(k) \leq \epsilon_{opt}(k)$. This is easily proved as follows: consider any orthonormal Φ that is feasible to (3) and construct an ortho-basis $\Phi_\perp$ spanning the nullspace of Φ. Then, $X = \Phi_\perp^T \Phi_\perp$ is feasible to (4) and the result holds. However, the program (4) is precisely the program used to estimate the *outer* $(d-k)$*-radius* of the point set V [11]. Moreover, the authors of [11] also propose an efficient scheme, based on *randomized rounding* (RR), that produces a matrix $\widehat{\Phi}_\perp$ with $d-k$ orthonormal rows, with cost γ at most $\mathcal{O}(\log n) \cdot \gamma_{opt}(k)$.

Immediately, we obtain the following algorithm (that we call *SDP+RR*): (i) solve the semidefinite program (4); (ii) apply the randomized rounding scheme of [11] to the solution of (4) to obtain an orthonormal $(d-k) \times d$ matrix $\widehat{\Phi}_\perp$; (iii) construct any ortho-basis of the nullspace of $\widehat{\Phi}_\perp$, for example, using Gram-Schmidt orthogonalization. This yields a matrix Φ that achieves a $\mathcal{O}(\log n)$-approximation factor for the original problem (3). Formally, we have:

Lemma 1: Given a set V with n points in $\mathbb{R}^d$, there exists a polynomial-time algorithm $\mathcal{A}$ that returns an orthonormal embedding f of V into $\mathbb{R}^k$ with distortion at most $\mathcal{O}(\log n) \cdot \epsilon_{opt}(k)$.

This guarantee is vacuous when $\epsilon_{opt}(k)$ and $\log(n)$ are large. However, we prove the following guarantee, which is somewhat stronger when k is small.

Theorem 3.1: Given a set V with n points in $\mathbb{R}^d$, there exists a polynomial time algorithm $\mathcal{A}$ that returns an orthonormal embedding f of V into $\mathbb{R}^k$ with distortion at most $(k+1)\epsilon_{opt}(k)$.

Proof: Let $X^* = \sum_{j=1}^{d} \lambda_j u_j u_j^*$ be any optimizer in (4). Suppose that $1 \geq \lambda_1 \geq \ldots \geq \lambda_d \geq 0$ be the n eigenvalues of X^*. By construction, we have $\sum_{j=1}^{d} \lambda_j = d - k$, and for all $v \in V$, we have $v^T X^* v = \sum_{j=1}^{d} \lambda_j \langle u_j, v\rangle^2 \leq \gamma_{opt}(k)$. First, observe that the largest eigenvalues $\lambda_{d-k-i} \geq \frac{1}{k+1}$ for $i \geq 1$. If this were not the case, then we would have $\lambda_{d-k+j} < \frac{1}{k+1}$ for all $j \geq 0$; however, this would imply that

$$\begin{aligned} \sum_j \lambda_j \quad &= \quad \sum_{j=1}^{d-k-1} \lambda_j + \sum_{j=d-k}^{d} \lambda_j \\ &< \quad d - k - 1 + (k+1)\frac{1}{k+1} = d - k, \end{aligned}$$

violating the trace constraint in the optimization. Therefore, we obtain the following relations:

$$\sum_{j=1}^{d} \lambda_j \langle u_j, v\rangle^2 \geq \sum_{j=1}^{d-k} \lambda_j \langle u_j, v\rangle^2 \geq \sum_{j=1}^{d-k} \frac{1}{k+1} \langle u_j, v\rangle^2,$$

$$\text{or} \sum_{j=1}^{d-k} \langle u_j, v\rangle^2 \leq (k+1)\gamma_{opt}(k).$$

In other words, any ortho-basis spanning the range of the k least significant eigenvectors of X^* yields a distortion at most $(k+1) \cdot \gamma_{opt}$, and by construction, at most $(k+1) \cdot \epsilon_{opt}$. ∎

We call this algorithm *SDP+DR*, where the suffix stands for *deterministic rounding*. Again, this approximation guarantee becomes vacuous when ϵ_{opt} or k are large. But however, notice that $\lambda_{d-2k-i} \geq \frac{1}{2}$ for $i \geq 0$. Else, as above, we would have

$$\sum_j \lambda_j < d - 2k - 1 + (2k+1)\frac{1}{2} < d - k,$$

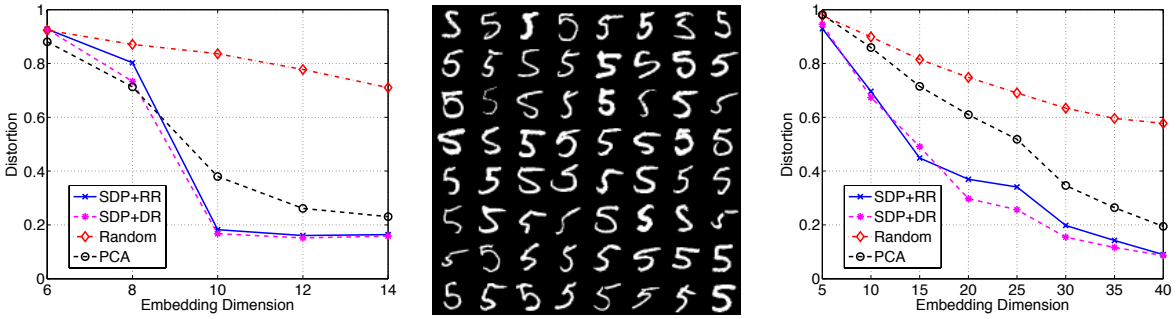

Fig. 1. Experimental results with orthonormal embeddings. (a) Subspace dataset: variation of embedding dimension m vs. distortion δ produced by different types of embeddings. (b) Example images from the MNIST dataset. (c) MNIST dataset: variation of m vs. δ.

violating the trace constraint. Therefore, as above, we have

$$\sum_{j=1}^{d} \lambda_j \langle u_j, v\rangle^2 \geq \sum_{j=1}^{d-2k} \frac{1}{2}\langle u_j, v\rangle^2, \text{implying that}$$
$$\sum_{j=1}^{d-2k} \langle u_j, v\rangle^2 \leq 2\gamma_{opt}(k).$$

As above, construct an ortho-basis spanning the range of the smallest $2k$ eigenvectors and we are done. Essentially, this result states that we can deterministically achieve, in polynomial time, an orthonormal embedding that offers at most twice the optimal distortion $\epsilon_{opt}(k)$ *provided we pay a factor of 2 in the embedding dimension*. In fact, the above method to obtain the bicriteria (two-sided) guarantee can be generalized to the case of asymmetric constants on either side. We omit the proof due to space constraints and summarize the result:

Theorem 3.2: Given a set V with n points in $\mathbb{R}^d$ and $\delta > 0$, there exists a polynomial time algorithm $\mathcal{A}$ that returns an orthonormal embedding f of V into $\mathbb{R}^q$ with distortion $(1+\delta)\epsilon_{opt}(k)$, where $q = \lceil (1+1/\delta)k \rceil$.

IV. Experiments

We present two sets of numerical experiments. First, let $d = 64$; then, we construct a simple synthetic dataset $V \in \mathbb{R}^d$ as follows: we sample $q = 500$ vectors from a random *subspace* of dimension $r = 10$, add d-dimensional Gaussian perturbations of small magnitude, and scale the vectors to have unit ℓ_2-norm. We then construct orthonormal embeddings of V into m dimensions using the SDP-RR and SDP-DR methods; for both, we solve (4) using a variant of *NuMax*, an efficient algorithm for solving SDP's with rank-1 constraints [12]. We also construct an m-dimensional orthonormal embedding by considering the first m normalized principal components of V, as well as an m-dimensional orthonormal J-L embedding. Figure 1(a) displays the variation of the distortion δ with increasing values of m. We observe that SDP-DR achieves the lowest distortion (in particular, far lower than PCA and random projections). Further, we see a levelling-off effect for $m \geq 10$ for our proposed algorithms, implying that we have correctly recovered the underlying subspace structure of the dataset.

Next, we consider a more challenging experiment with real-world data. The MNIST dataset [13] contains a large number of digital images of handwritten digits of size 28×28, and is commonly used as a benchmark for various machine learning algorithms. We collect $n = 200$ images of the digit '5', and construct the set V by calculating all $\binom{n}{2}$ pairwise difference vectors (normalized to unit norm). We then orthonormally embed this set into dimension m using different projection methods, and measure the distortion. The results of this experiment are plotted in Fig. 1(c). Once again, we observe that SDP-DR offers the lowest distortion among the different methods.

References

[1] P. Indyk and J. Matousek, "Low distortion embeddings of finite metric spaces," *Handbook of Discrete and Comp. Geom.*, vol. 273, pp. 177–196, 2004.

[2] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Proc. Conf. Modern Anal. and Prob.*, New Haven, CT, Jun. 1982.

[3] N. Alon, "Problems and results in extremal combinatorics," *Discrete Math.*, vol. 273, no. 1, pp. 31–53, 2003.

[4] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proc. ACM Symp. Theory of Comput.*, New York, NY, 1998, pp. 604–613.

[5] E. Candès, "Compressive sampling," in *Proc. Int. Cong. Math.*, Madrid, Spain, 2006, vol. 3, pp. 1433–1452.

[6] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Automat. Control*, vol. 26, no. 1, pp. 17–32, 1981.

[8] A. Sidiropoulos, *Computational Metric Embeddings*, Ph.D. thesis, Massachusetts Instt. Tech., May 2008.

[9] J. Matousek and A. Sidiropoulos, "Inapproximability of metric embeddings into $\mathbb{R}^d$," *Trans. Amer. Math. Soc*, vol. 362, no. 12, pp. 6341–6365, 2010.

[10] M. Badoiu, S. Har-Peled, and P. Indyk, "Approximate clustering via core sets," in *Proc. ACM Symp. Theory of Comput.*, 2002, pp. 250–257.

[11] K. Varadarajan, S. Venkatesh, Y. Ye, and J. Zhang, "Approximating the radii of point sets," *SIAM J. Comput.*, vol. 36, no. 6, pp. 1764–1776, 2007.

[12] C. Hegde, A. Sankaranarayanan, W. Yin, and R. Baraniuk, "A convex approach for learning near-isometric linear embeddings," Preprint, Nov. 2012.

[13] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 1998, available online at http://yann.lecun.com/exdb/mnist.