Nearest Neighbor Preserving Embeddings

Piotr Indyk
MIT
and
Assaf Naor
Microsoft Research

In this paper we introduce the notion of nearest neighbor preserving embeddings. These are randomized embeddings between two metric spaces which preserve the (approximate) nearest neighbors. We give two examples of such embeddings, for Euclidean metrics with low "intrinsic" dimension. Combining the embeddings with known data structures yields the best known approximate nearest neighbor data structures for such metrics.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems

General Terms: Algorithms, Theory

Additional Key Words and Phrases: nearest neighbor, dimensionality reduction, embeddings, doubling spaces

1. INTRODUCTION

The nearest neighbor problem is defined as follows: Given a set X of points in $\mathbb{R}^d$, build a data structure which given any $q \in \mathbb{R}^d$, quickly reports the point in X that is (approximately) closest to q. This problem, and its approximate versions, are some of the central problems in computational geometry.

Since the late 1990's, it has become apparent that designing efficient approximate nearest neighbor algorithms, at least for high-dimensional data, is closely related to the task of designing *low-distortion embeddings*. A *bi-Lipschitz embedding* between two metric spaces (X, d_X) and $(X', d'_{X'})$ is a mapping $f : X \to X'$ such that for some scaling factor $C > 0$, for every $p, q \in X$ we have $Cd_X(p, q) \leq d'_{X'}(f(p), f(q)) \leq DCd_X(p, q)$, where the parameter $D \geq 1$ called the *distortion* of f. Of particular importance, in the context of approximate nearest neighbor, are low-distortion embeddings that map $X \subseteq \mathbb{R}^d$ into $\mathbb{R}^k$, where k is much smaller than d. For example, a well-known theorem of Johnson and Lindenstrauss guarantees that for any set $X \subseteq \mathbb{R}^d$ there is a $(1 + \varepsilon)$-distortion embedding of $(X, \|\cdot\|_2)$ into $(\mathbb{R}^k, \|\cdot\|_2)$ for $k = O(\log |X|/\varepsilon^2)$. This embedding and its variants have been utilized, e.g. in [Indyk and Motwani 1999; Kushilevitz et al. 2000], to give efficient approximate nearest neighbor algorithms in high-dimensional spaces.

More recently (e.g., in [Indyk 2000]), it has been realized that the approximate nearest neighbor problem requires embedding properties that are somewhat different from the above definition. One (obvious) difference is that the embedding must be *oblivious* to X, that is, well-defined over the whole space $\mathbb{R}^d$, not just the input data points X. This is because, in general, a query point $q \in \mathbb{R}^d$ does not belong to X. The aforementioned Johnson-Lindenstrauss lemma indeed satisfies this (stronger) property. The second difference is that the embedding does not need to preserve *all* interpoint distances. Instead, it suffices[1] that the embedding f

[1]If we consider the approximate *near* neighbor problem, i.e., the decision version of the approximate nearest neighbor, then the constraints that an embedding needs to satisfy are even weaker. Also, it is known [Indyk and Motwani 1999; Har-Peled 2001] that the approximate nearest neighbor can be reduced to its decision version. However, such reductions are non-trivial and introduce certain overhead in the query time and space. Thus, it is beneficial that the embedding preserves the approximate

The first author was supported in part by NSF ITR grant CCR-0220280, David and Lucille Packard Fellowship and Alfred P. Sloan Fellowship.

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
© 20YY ACM 0000-0000/20YY/0000-0001 $5.00

is randomized, and satisfies the following definition which we introduce:

DEFINITION 1.1. *Let $(Y, d_Y), (Z, d_Z)$ be metric spaces and $X \subseteq Y$. We say that a distribution over mappings $f : Y \to Z$ is a nearest neighbor preserving embedding (or NN-preserving) with distortion $D \geq 1$ and probability of correctness*[2] *$P \in [0, 1]$ if for every $c \geq 1$ and any $q \in Y$, with probability at least P, if $x \in X$ is such that $f(x)$ is a c-approximate nearest neighbor of $f(q)$ in $f(X)$ (i.e. $d(f(q), f(x)) \leq c \cdot d(f(q), f(X))$), then x is a $D \cdot c$ approximate nearest neighbor of q in X.*

This notion is the appropriate generalization of oblivious embeddings à la Johnson and Lindenstrauss: We want f to be defined on the entire space of possible query points Y, and we require much less than a bi-Lipschitz condition. Clearly, the Johnson-Lindenstrauss theorem is an example of a NN-preserving embedding. Another example of such a mapping is a (weak) dimensionality reduction in ℓ_1 norm given in [Indyk 2000]. It maps $(\mathbb{R}^d, \|\cdot\|_1)$ into $(\mathbb{R}^k, \|\cdot\|_1)$, where k is much smaller than d, and guarantees that, for any pair of points, the probability that the distance between the pair gets contracted is "very small", while the probability of the distance being expanded by a is at most $1/2$. It is easy to see that such mapping is a-NN-preserving. At the same time, the standard dimensionality reduction in ℓ_1 (that preserves all distances) is provably impossible [Brinkman and Charikar 2005; Lee and Naor 2004]. Thus, the definition of NN-preserving embeddings allows us to overcome the impossibility results for the stronger notion of bi-Lipschitz embeddings, while being sufficient for the purpose of the nearest neighbor and related problems.

In this paper we initiate a systematic study of NN-preserving embeddings into low-dimensional spaces. In particular, we prove that such embeddings exist for the following subsets X of the Euclidean space $(\mathbb{R}^d, \|\cdot\|_2)$:

(1) Doubling sets. The doubling constant of X, denoted λ_X, is the smallest integer λ such that for any $p \in X$ and $r > 0$, the ball $B(p, r)$ (in X) can be covered by at most λ balls of radius $r/2$ centered at points in X. It is also convenient to define the *doubling dimension* of X to be $\log \lambda_X$ (this terminology is used in [Gupta et al. 2003]). See [Clarkson 2005] for a survey of notions of dimension which are relevant to nearest neighbor search.
We give, for any $\varepsilon > 0, \delta \in (0, 1/2]$, a randomized mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ that is $(1 + \varepsilon)$-NN-preserving for X, with probability of correctness $1 - \delta$, where
$$k = O\left(\frac{\log(1/\varepsilon)}{\varepsilon^2} \cdot \log(1/\delta) \cdot \log \lambda_X\right).$$

(2) Sets with small aspect ratio and small γ-dimension. Consider sets X of diameter 1. The aspect ratio of X, Δ, is the inverse of the smallest interpoint distance in X. The γ-dimension of X, which is a natural notion motivated by the theory of Gaussian processes, is defined in Section 2. Here we just state that $\gamma(X) = O(\sqrt{\log \lambda_X})$ for all X.
We give, for any $\varepsilon > 0$, a randomized mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ that is $(1 + \varepsilon)$-NN-preserving for X, where $k = O(\Delta^2 \gamma^2 / \varepsilon^2)$.
Although quadratic dependence of the dimension on Δ might seem excessive, there exist natural high dimensional data sets with (effectively) small aspect ratios. For example, in the MNIST data set (investigated e.g., in [Andoni et al. 2005]), for all but 2% of points, the distances to nearest neighbors lie in the range $[0.19, 0.72]$.

The above two results are not completely disjoint. This is because, for metrics with constant aspect ratio, the γ dimension and doubling dimension coincide, up to constant factors. However, this is not the case for $\Delta = \omega(1)$.

Our investigation here is related to the following open problem in metric geometry: Is it true that doubling subsets of ℓ_2 embed bi-Lipschitzly into low dimensional Euclidean space? (see Section 4 for a precise formulation). This question is of great theoretical interest, but it is also clear that a positive answer to it will have algorithmic applications. Our result shows that for certain purposes, such as nearest neighbor search, a weaker notion of embedding suffices, and provably exists. It is worth noting that while our nearest neighbor preserving mapping is linear, a bi-Lipschitz embedding of a doubling Euclidean metric into a low dimensional Euclidean space cannot be in general linear (this is discussed in Section 4).

---

nearest neighbor, not just the near neighbor.

[2]Whenever P is not specified, it is assumed to be equal to $1/2$.

Algorithmic implications. Our NN-preserving embeddings have, naturally, immediate applications to efficient approximate nearest neighbor problems.

Our first application combines NN-preserving embeddings with efficient $(1+\varepsilon)$-approximate nearest neighbor data structures in the Euclidean space $(\mathbb{R}^k, \|\cdot\|_2)$ [Har-Peled 2001; Arya and Malamatos 2002], which use $O(|X|/\varepsilon^k)$ space and have $O(k\log(|X|/\varepsilon))$ query time (recall that we guarantee $k = O(\log\lambda_X \log(1/\varepsilon)/\varepsilon^2)$, and that we need to add $O(dk)$ to the query time to account for the time needed to embed the query point). This results in a very efficient $(1+\varepsilon)$-approximate nearest neighbor data structure. For comparison, the data structure of [Krauthgamer and Lee 2004; Har-Peled and Mendel 2006], which works for *general* metrics, suffers from query time exponential in $\log\lambda_X$. For the case of subsets of $(\mathbb{R}^d, \|\cdot\|_2)$ (which we consider here), their data structure can be made faster [Krauthgamer-Lee, personal communication] by using fast approximate nearest neighbor algorithm of [Indyk and Motwani 1999; Kushilevitz et al. 2000] as a subroutine. In particular, the query time becomes roughly $O(dk + k\cdot\log\Delta)$ and space $O(|X|/\varepsilon^k)$. However, unlike in our case, the query time of that algorithm depends on the aspect ratio Δ. Since for any X we have that $\lambda_X \le |X|$, it follows that our algorithm always uses space $|X|^{O([\log(1/\varepsilon)/\varepsilon]^2)}$ and has query time $O(d\log|X|\log(1/\varepsilon)/\varepsilon^2)$. Thus, our algorithm almost matches the bounds of the algorithm of [Indyk and Motwani 1999], while being more general.

Our second application involves approximate nearest neighbor where the data set consists of objects that are more complex than points. Specifically, we consider an arbitrary set X containing n sets $\{S_1 \dots S_n\}$, where $S_i \subset \mathbb{R}^d$, $i = 1 \dots n$. Let $\lambda = \max_{i=1\dots n}\lambda_{S_i}$. If we set $\delta = \frac{1}{2n}$, it follows that for any point $q \in \mathbb{R}^d$, a random mapping $G : \mathbb{R}^d \to \mathbb{R}^k$, $k = O(\log\lambda\cdot\log n\cdot\log(1/\varepsilon)/\varepsilon^2)$, preserves a $(1+\epsilon)$-nearest neighbor of q in $\cup_{i=1}^n S_i$ with probability at least $1/2$. Therefore, if we find a $(1+\epsilon)$-approximate nearest neighbor of $G(q)$ in $\{G(S_1)\dots G(S_n)\}$, then, with probability $1/2$, it is also a $(1+O(\epsilon))$-approximate nearest neighbor of q in $\{S_1 \dots S_n\}$. This corollary provides a strong generalization of a result of [Magen 2002], who showed this fact for the case when $S_i's$ are affine spaces (although our bound is weaker by a factor of $\log(1/\epsilon)$).

Our embedding-based approach to design of approximate nearest neighbor algorithms has the following benefits:

—Simplicity preservation: our data structure is as simple as the data structure we use as a subroutine.

—Modularity: any potential future improvements to algorithms for the approximate nearest neighbor problem in ℓ_2^k will, when combined with our embedding, automatically yield a better bound for the same problem in ℓ_2 metrics with low doubling constant.

Although in this paper we focused on embeddings into ℓ_2, it is interesting to design NN-preserving embeddings into any space which supports fast approximate nearest neighbor search, e.g., low dimensional ℓ_∞ [Indyk 1998].

2. BASIC CONCEPTS

In this section we introduce the basic concepts used in this paper. In particular, we define the doubling constant, the parameter E_X, and the γ-dimension. We also point out the relations between these parameters.

Doubling constant. Let (X, d_X) be a metric space. In what follows, $B_X(x,r)$ denotes the ball in X of radius r centered at $x \in X$, i.e. $B_X(x,r) = \{y \in X : \ d_X(x,y) < r\}$. The doubling constant of X (see [Heinonen 2001]), denoted λ_X, is the least integer $\lambda \ge 1$ such that for every $x \in X$ and $r > 0$ there is $S \subseteq X$ with $|S| \le \lambda$ such that

$$B_X(x,2r) \subseteq \bigcup_{s\in S} B_X(s,r).$$

The parameter E_X. Fix an integer N and denote by $\langle\cdot,\cdot\rangle$ the standard inner product in $\mathbb{R}^N$. In what follows $g = (g_1, ..., g_N)$ is a standard Gaussian vector in $\mathbb{R}^N$ (i.e. a vector with independent coordinates which are standard Gaussian random variables). Given $X \subseteq \ell_2^N$ we denote:

$$E_X = \mathbb{E}\sup_{x\in X}|\langle x, g\rangle| = \mathbb{E}\sup_{(x_1,\dots,x_N)\in X}\sum_{i=1}^N x_i g_i. \tag{1}$$

We observe that the parameter E_X of a given bounded set $X \subseteq \ell_2^d$ can be estimated very efficiently, that is, in time $O(d|X|)$. This follows directly from the definition of E_X, and the fact that for every $t > 0$,

$\Pr[|\sup_{x\in X}|\langle g,x\rangle| - E_X| > t] \le 2e^{-t^2/(4\max_{x\in X}\|x\|_2^2)}$ (this deviation inequality is a consequence of the fact that the mapping $g \mapsto \sup_{x\in X}|\langle g,x\rangle|$ is Lipschitz with constant $\max_{x\in X}\|x\|_2$, and the Gaussian isoperimetric inequality– see [Ledoux and Talagrand 1991]). In addition, even if X is large, e.g., has size exponential in d, E_X can often be computed in time polynomial in d [Barvinok 1997; Barvinok and Samorodnitsky 2001; 2004]. For example, this is the case when X is a set of all matchings in a given graph G, where each matching is represented by a characteristic vector of its edge set.

Doubling constant vs. the parameter E_X. We observe that for every bounded $X \subseteq \ell_2^N$:

$$E_X = O\left(\operatorname{diam}(X)\sqrt{\log \lambda_X}\right). \tag{2}$$

Indeed an, inequality of Dudley (see [Ledoux and Talagrand 1991]) states that

$$E_X \le 24\int_0^{\operatorname{diam}(X)} \sqrt{\log N(X,\varepsilon)}\, d\varepsilon,$$

where $N(X,\varepsilon)$ are the *entropy numbers* of X, namely the minimal number of balls of radius ε required to cover X. The doubling condition implies that for every $\varepsilon > 0$ we have that $N(X, \varepsilon\operatorname{diam}(X)) \le (2/\varepsilon)^{\log_2 \lambda_X}$, so

$$E_X \le 24\operatorname{diam}(X)\sqrt{\log_2 \lambda_X}\int_0^1 \sqrt{\log_2(2/\varepsilon)}\, d\varepsilon \le 80\operatorname{diam}(X)\sqrt{\log_2 \lambda_X}.$$

Another way to prove (2) is as follows. Let $\mathcal{B}(X)$ be the set of all Borel probability measures on X. The celebrated Majorizing Measure Theorem of Talagrand [Talagrand 1987] states that

$$E_X = \Theta\left(\inf_{\mu\in\mathcal{B}(X)}\sup_{x\in X}\int_0^\infty \sqrt{\log\left(\frac{1}{\mu(B_X(x,\varepsilon))}\right)}\, d\varepsilon\right). \tag{3}$$

A theorem of Konyagin and Vol′berg [Konyagin and Vol′berg 1987; Heinonen 2001] states that if X is a complete metric space there exists a Borel measure μ on X such that for every $x \in X$ and $r > 0$, $\mu(B_X(x,2r)) \le \lambda_X^2\mu(B_X(x,r))$. Now we just plug μ into (3) and obtain (2).

γ-dimension. The right-hand side of (3) makes sense in arbitrary metric spaces, not just subsets of ℓ_2. In fact, another equivalent formulation of (1) is based on Talagrand's γ_2 functional, defined as follows. Given a metric space (X, d_X) set

$$\gamma_2(X) = \inf\sup_{x\in X}\sum_{s=0}^\infty 2^{s/2} d_X(x, A_s), \tag{4}$$

where the infimum is taken over all choices of subsets $A_s \subseteq X$ with $|A_s| \le 2^{2^s}$. Talagrand's "generic chaining" version of the majorizing measures theorem [Talagrand 1996; 2001; 2005] states that for every $X \subseteq \ell_2$, $E_X = \Theta(\gamma_2(X))$ (we refer to [Guédon and Zvavitch 2003] for a related characterization). The parameter $\gamma_2(X)$ can be defined for arbitrary metric spaces (X, d_X) and it is straightforward to check that in general $\gamma_2(X) = O\left(\operatorname{diam}(X)\sqrt{\log \lambda_X}\right)$. Thus, it is natural to define the γ *dimension* of X to be:

$$\gamma\dim(X) \equiv \left[\frac{\gamma_2(X)}{\operatorname{diam}(X)}\right]^2.$$

3. THE CASE OF EUCLIDEAN SPACES WITH LOW γ-DIMENSION

We introduce the following useful modification of the notion of bi-Lipschitz embedding. We then use this notion to give NN-preserving embeddings of ℓ_2 submetrics with bounded aspect ratio and γ-dimension, into low-dimensional ℓ_2.

Definition 3.1 Bi-Lipschitz embeddings with resolution. *Let (X, d_X), (Y, d_Y) be metric spaces, and $\delta, D > 0$. A mapping $f : X \to Y$ is said to be D bi-Lipschitz with resolution δ if there is a (scaling factor) $C > 0$ such that*

$$\forall a, b \in X,\ d_X(a,b) > \delta \Longrightarrow\ Cd_X(a,b) \le d_Y(f(a), f(b)) \le CDd_X(a,b).$$

In what follows S^{d-1} denotes the unit Euclidean sphere centered at the origin. We will use the following theorem, which is due to Gordon [Gordon 1988]:

THEOREM 3.2 [GORDON 1988]. *Fix $X \subseteq S^{d-1}$ and $\varepsilon \in (0,1)$. Then there exists an integer $k = O\left(\frac{E_X^2}{\varepsilon^2}\right)$ and a linear mapping $T : \mathbb{R}^d \to \mathbb{R}^k$ such that for every $x \in X$,*

$$1 - \varepsilon \leq \|Tx\|_2 \leq 1 + \varepsilon.$$

REMARK 3.1. Since the proof of the above theorem uses probabilistic method (specifically, Gordon [Gordon 1988] proved that if Γ is a $k \times d$ matrix whose coordinates are i.i.d. standard Gaussian random variables then $T = \frac{1}{\sqrt{k}}\Gamma$ will satisfy the assertion of Theorem 3.2 with high probability- see also [Schechtman 1989; 2006]), it also follows that there exists a randomized embedding that satisfy the thesis of the above theorem with probability 1/2. Recently Klartag and Mendelson [Klartag and Mendelson 2005] showed that the same result holds true if the entries of $\Gamma = (\gamma_{ij})$ are only assumed to be i.i.d., have mean 0 and variance 1, and have sub-Gaussian tail bounds, i.e. $\Pr[|\gamma_{ij}| > u] \leq Ke^{-\delta u^2}$ for every $u > 0$ and some constants $K, \delta > 0$. In this case the implied constants may depend on K, δ. A particular case of interest is when Γ is a random ± 1 matrix- just like in Achlioptas' variant [Achlioptas 2003] of the Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss 1984], we obtain "database friendly" versions of Theorem 4.1.

It should be pointed out here that although several papers [Frankl and Maehara 1988; Dasgupta and Gupta 2003; Indyk and Motwani 1999; Achlioptas 2003] obtained alternative proofs of the Johnson-Lindenstrauss lemma using different types of random matrices, it turns out that the only thing that matters is that the entries are i.i.d. sub-Gaussian random variables (to see this just note that when $\delta = 0$ the set $\widetilde{X}$ contains at most $|X|^2$ points, and for any n-point subset Z of $\mathbb{R}^d$, $E_Z = O(\sqrt{\log n})$). Here is a simple proof of this fact using an elementary large-deviation argument (a similar argument was also recently obtained in [Matoušek 2006]). Let g be a standard Gaussian random variable, i.e. its density is $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. In what follows g will always be assumed to be independent of all the other random variables that appear in the proof. The only properties of g that we will need is that for all $t \in \mathbb{R}$ we have $\mathbb{E}e^{tg} = e^{t^2/2}$ and for all $t \in (0, 1/2)$ we have $\mathbb{E}e^{tg^2} = \frac{1}{\sqrt{1-2t}}$.

Now, let X be a symmetric random variable (i.e. X and $-X$ have the same distribution) such that $\mathbb{E}X^2$=1. Assume that X is sub-Gaussian, i.e. that $\mathbb{E}e^{uX} \leq e^{cu^2}$ for all $u \in \mathbb{R}$ and some constant $c > 0$ (note that this condition on the moment generating function easily follows from sub-Gaussian tail bounds). Fix a unit vector $a = (a_1, \ldots, a_n) \in S^{n-1}$. Let $X_1, \ldots, X_n$ be i.i.d. copies of X and denote $U = \sum_{j=1}^n a_j X_j$. Then $\mathbb{E}U^2 = 1$, and for every $0 \leq t \leq \frac{1}{8c}$ we have

$$\mathbb{E}e^{tU^2} = \mathbb{E}_U \mathbb{E}_g e^{\sqrt{2t}gU} = \mathbb{E}_g \left(\prod_{j=1}^n \mathbb{E}_X e^{\sqrt{2t}a_j g X} \right) \leq E_g \left[\prod_{j=1}^n e^{2cta_j^2 g^2} \right] = \mathbb{E}_g e^{2ctg^2} = \frac{1}{\sqrt{1-4ct}} \leq \sqrt{2}. \quad (5)$$

Therefore using (5) we see that for every $-\frac{1}{8c} \leq t \leq \frac{1}{8c}$ we have

$$\mathbb{E}e^{tU^2} = \sum_{m=0}^{\infty} \frac{t^m \mathbb{E}U^{2m}}{m!} \leq 1 + t + \sum_{m=2}^{\infty} \frac{(8c|t|)^m \left(\frac{1}{8c}\right)^m \mathbb{E}U^{2m}}{m!} \leq 1 + t + (8ct)^2 \sum_{m=2}^{\infty} \frac{\left(\frac{1}{8c}\right)^m \mathbb{E}U^{2m}}{m!}$$
$$\leq 1 + t + (8ct)^2 \mathbb{E}e^{U^2/(8c)} \leq 1 + t + 100c^2t^2 \leq e^{t+100c^2t^2}. \quad (6)$$

Hence if $U_1, \ldots, U_k$ are i.i.d. copies of U then for every $0 < \varepsilon \leq 25c$ we can apply (6) with $t = \frac{\varepsilon}{200c^2}$ to get that

$$\Pr\left[\frac{1}{k} \sum_{i=1}^k U_i^2 \geq 1 + \varepsilon \right] \leq e^{-k(1+\varepsilon)t} \left(\mathbb{E}e^{tU^2} \right)^k \leq e^{-k(1+\varepsilon)t + kt + 100kc^2t^2} = e^{-k\varepsilon^2/(400c^2)},$$

and

$$\Pr\left[-\frac{1}{k} \sum_{i=1}^k U_i^2 \geq -1 + \varepsilon \right] \leq e^{k(1-\varepsilon)t} \left(\mathbb{E}e^{-tU^2} \right)^k \leq e^{k(1-\varepsilon)t - kt + 100kc^2t^2} = e^{-k\varepsilon^2/(400c^2)}.$$

In summary, we proved that for every $0 < \varepsilon < 25c$ we have

$$\Pr\left[\left|\frac{1}{k}\sum_{i=1}^{k} U_i^2 - 1\right| \geq \varepsilon\right] \leq 2e^{-k\varepsilon^2/(400c^2)}.$$

Another way to phrase this inequality is that if $\{X_{ij} : \ i = 1, \ldots, k \ j = 1, \ldots, n\}$ are i.i.d. copies of X, and we consider the $k \times n$ random matrix $A := \frac{1}{\sqrt{k}}(X_{ij})$ then for every $x \in \mathbb{R}^n$ we have

$$\Pr\left[\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq \varepsilon\|x\|_2^2\right] \leq 2e^{-k\varepsilon^2/(400c^2)}.$$

An application of the union bound shows that this concentration inequality implies the Johnson-Lindenstrauss dimensionality reduction result for arbitrary random matrices with sub-Gaussian i.i.d. entries.

A simple corollary of Theorem 3.2 is the following theorem.

THEOREM 3.3. *Fix $\varepsilon, \delta > 0$ and a set $X \subseteq \mathbb{R}^d$. Then there exists an integer $k = O\left(\frac{E_X^2}{\delta^2\varepsilon^2}\right)$ such that X embeds $1+\varepsilon$ bi-Lipschitzly in $\mathbb{R}^k$ with resolution δ. Moreover, the embedding extends to a linear mapping defined on all of $\mathbb{R}^d$.*

PROOF. Consider the set $\widetilde{X} = \left\{\frac{x-y}{\|x-y\|_2} : \ x, y \in X, \ \|x-y\|_2 \geq \delta\right\}$. Then

$$E_{\widetilde{X}} = \mathbb{E}\left(\sup\left\{\frac{|\langle x-y, g\rangle|}{\|x-y\|_2} : \ x, y \in X \ \|x-y\|_2 \geq \delta\right\}\right) \leq \frac{1}{\delta}\,\mathbb{E}\sup_{x,y\in X}|\langle x-y, g\rangle| \leq \frac{2}{\delta}E_X.$$

So the required result is a consequence of Theorem 3.2 applied to $\widetilde{X}$. □

REMARK 3.2. *We can make Theorem 3.3 scale invariant by normalizing by* $\operatorname{diam}(X)$*, in which case we get a $1+\varepsilon$ bi-Lipschitz embedding with resolution $\delta \operatorname{diam}(X)$, where $k = O\left(\frac{\gamma \dim(X)}{\delta^2\varepsilon^2}\right)$.*

REMARK 3.3. Let $\{e_j\}_{j=1}^{\infty}$ be the standard basis of ℓ_2. Fix $\varepsilon, \delta \in (0, 1/2)$, an integer n and set $m = \lceil n^{1/\delta^2} \rceil$. Consider the set $X = \{e_1, \ldots, e_n, \delta e_{n+1}, \ldots, \delta e_{n+m}\}$. Then $E_X = \Theta(\sqrt{\log n} + \delta\sqrt{\log m}) = \Theta(\sqrt{\log n})$. Let $f : X \to \mathbb{R}^k$ be a (not necessarily linear) $1+\varepsilon$ bi-Lipschitz embedding with resolution δ (which is thus a $1+\varepsilon$ bi-Lipschitz embedding since the minimal distance in X is $\delta\sqrt{2}$). By a result of Alon [Alon 2003] (see also [Matoušek 2002]) we deduce that $k = \Omega\left(\frac{\log m}{\varepsilon^2 \log(1/\varepsilon)}\right) = \Omega\left(\frac{E_X^2}{\delta^2\varepsilon^2 \log(1/\varepsilon)}\right)$. Thus the value of k in Theorem 3.3 is nearly optimal.

4. THE CASE OF EUCLIDEAN DOUBLING SPACES

We recall some facts about random Gaussian matrices. Let $a \in S^{n-1}$ be a unit vector and let $\{g_{ij} : \ 1 \leq i \leq k, \ 1 \leq j \leq n\}$ be i.i.d. standard gaussian random variables. Denoting $G = \frac{1}{\sqrt{k}}(g_{ij})$, by standard arguments (see [Durrett 1996]) the random variable $\|Ga\|_2^2$ has distribution whose density is:

$$\frac{1}{2^{k/2}\Gamma(k/2)} \cdot x^{\frac{k}{2}-1}e^{-x/2}, \quad x > 0.$$

By a simple computation it follows that for $D > 0$

$$\Pr\left[\left|\|Ga\|_2 - 1\right| \geq D\right] \leq e^{-kD^2/8} \quad \text{and} \quad \Pr[\|Ga\|_2 \leq 1/D] \leq \left(\frac{3}{D}\right)^k. \tag{7}$$

The main result of this section is the following theorem:

THEOREM 4.1. *For $X \subseteq \mathbb{R}^d$, $\varepsilon \in (0,1)$ and $\delta \in (0, 1/2)$ there exists $k = O\left(\frac{\log(2/\varepsilon)}{\varepsilon^2} \cdot \log(1/\delta) \cdot \log \lambda_X\right)$ such that for every $x_0 \in X$ with probability at least $1-\delta$,*

(1) $d(Gx_0, G(X \setminus \{x_0\})) \leq (1+\varepsilon)d(x_0, X \setminus \{x_0\})$
(2) Every $x \in X$ with $\|x_0 - x\|_2 > (1+2\varepsilon)d(x_0, X \setminus \{x_0\})$ satisfies

$$\|Gx_0 - Gx\| > (1+\varepsilon)d(x_0, X \setminus \{x_0\}).$$

The following lemma can be proved using the methods of [Gordon 1988; Schechtman 1989; 2006], but the doubling assumption allows us to give a simple direct proof.

LEMMA 4.2. *Let $X \subseteq B(0,1)$ be a subset of the n-dimensional Euclidean unit ball. Then there exist universal constants $c, C > 0$ such that for $k \geq C \log \lambda_X + 1$ and $D > 1$,*

$$\Pr[\exists x \in X,\ \|Gx\|_2 \geq D] \leq e^{-ckD^2}.$$

PROOF. Without loss of generality $0 \in X$. We construct subsets $I_0, I_1, I_2, \ldots \subseteq X$ as follows. Set $I_0 = \{0\}$. Inductively, for every $t \in I_j$ there is a minimal $S_t \subseteq X$ with $|S_t| \leq \lambda_X$ such that $B(t, 2^{-j}) \cap X \subseteq \cup_{s \in S_t} B(s, 2^{-j-1}) \cap X$. We define $I_{j+1} = \cup_{t \in I_j} S_t$.

For $x \in X$ there is a sequence $\{0 = t_0(x), t_1(x), t_2(x), \ldots\} \subseteq X$ such that for all j we have $t_{j+1}(x) \in S_{t_j(x)}$, and $x = \sum_{j=0}^{\infty} [t_{j+1}(x) - t_j(x)]$. Now, using the fact that $\|t_{j+1}(x) - t_j(x)\|_2 \leq 2^{-j+1}$, we get

$$\begin{aligned}
\Pr[\exists x \in X,\ \|Gx\|_2 \geq D] &\leq \Pr\left[\exists x \in X\, \exists j \geq 0,\ \|G[t_{j+1}(x) - t_j(x)]\|_2 \geq \frac{D}{3}\left(\frac{3}{2}\right)^{-j}\right] \\
&\leq \sum_{j=0}^{\infty} \Pr\left[\exists t \in I_j\, \exists s \in S_t,\ \|G(t-s)\|_2 \geq \frac{D}{6}\left(\frac{4}{3}\right)^{j} \|t-s\|_2\right] \\
&\leq \sum_{j=0}^{\infty} \lambda_X^{2j} e^{-\frac{kD^2}{400}(4/3)^{2j}} \leq e^{-ckD^2},
\end{aligned}$$

provided that $k \geq C \log \lambda_X + 1$ and using the first estimate in (7). $\square$

PROOF PROOF OF THEOREM 4.1. Without loss of generality $x_0 = 0$ and $d(x_0, X \setminus \{x_0\}) = 1$. If $y \in X$ satisfies $\|y\|_2 = 1$ then by (7) we get that $\Pr[\|Gy\|_2 \geq 1 + \varepsilon] \leq e^{-k\varepsilon^2/8}$. Thus for $k \geq C \log(1/\delta)/\varepsilon^2$ we get that

$$\Pr[d(Gx_0, G(X \setminus \{x_0\})) > (1+\varepsilon) d(x_0, X \setminus \{x_0\})] < \delta/2.$$

Define $r_{-1} = 0$, $r_0 = 1$, $r_i = 1 + 2\varepsilon + \varepsilon(i-1)/4$, and consider the annuli

$$X_i = X \cap [B(0, r_i) \setminus B(0, r_{i-1})]\,.$$

Fix an integer $i \geq 1$, and use the doubling condition to find $S \subseteq X_i$ such that $X_i \subseteq \cup_{s \in S} B(s, \varepsilon/4)$ and $|S| \leq \lambda_X^{\log_2(16 r_i/\varepsilon)}$. Then by Lemma 4.2

$$\Pr\left[\exists s \in S\, \exists x \in B(s, \varepsilon/4) \cap X_i,\ \|Gs - Gx\|_2 \geq \frac{\varepsilon\sqrt{i}}{4}\right] \leq \lambda_X^{\log_2(16 r_i/\varepsilon)} \cdot e^{-cki} \leq e^{-c'ki}. \tag{8}$$

On the other hand fix $s \in S$. If $\|Gs\|_2 < 1 + \varepsilon + \frac{\varepsilon\sqrt{i}}{4}$ then there exists a universal constant $C > 0$ such that

$$\frac{\|Gs\|_2}{\|s\|_2} \leq \frac{1 + \varepsilon + \frac{\varepsilon\sqrt{i}}{4}}{1 + 2\varepsilon + \frac{\varepsilon(i-2)}{4}} \leq \begin{cases} 1 - \varepsilon/4 & i \leq 1/\varepsilon^2 \\ C/\sqrt{i} & i > 1/\varepsilon^2. \end{cases}$$

Hence, by (7)

$$\begin{aligned}
\Pr\left[\exists s \in S\ \ \|Gs\|_2 \geq 1 + \varepsilon + \frac{\varepsilon\sqrt{i}}{4}\right] &\leq \begin{cases} \lambda_X^{\log_2(16 r_i/\varepsilon)} e^{-c''k\varepsilon^2} & i \leq 1/\varepsilon^2 \\ \lambda_X^{\log_2(16 r_i/\varepsilon)} \cdot (3C/\sqrt{i})^k & i > 1/\varepsilon^2 \end{cases} \\
&\leq \begin{cases} e^{-c'''k\varepsilon^2} & i \leq 1/\varepsilon^2 \\ i^{-c'''k} & i > 1/\varepsilon^2. \end{cases}
\end{aligned} \tag{9}$$

provided $k \geq C \frac{\log(2/\varepsilon)}{\varepsilon^2} \cdot \log \lambda_X$ for a large enough constant C.

Now, from (8) and (9) we see that there exists a constant $\widetilde{c}$ such that

$$\Pr[\forall x \in X_i,\ \|Gx\|_2 > 1 + \varepsilon] \geq \begin{cases} 1 - 2e^{-\widetilde{c}k\varepsilon^2} & i \leq 1/\varepsilon^2 \\ 1 - 2i^{-\widetilde{c}k} & i > 1/\varepsilon^2. \end{cases}$$

Hence,

$$\Pr\left[\exists x \in X,\ \|x\|_2 > 1+2\varepsilon \wedge \|Gx\|_2 < 1+\varepsilon\right] \leq \sum_{i=1}^{\infty} \Pr[\exists x \in X_i,\ \|Gx\|_2 < 1+\varepsilon]$$
$$\leq \frac{2}{\varepsilon^2} e^{-\widetilde{c}k\varepsilon^2} + 2\sum_{i>1/\varepsilon^2} i^{-\widetilde{c}k} < \delta/2,$$

for large enough k. This completes the proof of Theorem 4.1. □

REMARK 4.1. Since $\gamma\dim(X) = O(\log\lambda_X)$ Theorem 4.1 sheds some light on the following problem (which is folklore, but apparently has been first stated explicitly in print in [Lang and Plaut 2001]): Is it true that any subset $X \subseteq \ell_2$ embeds into $\ell_2^{d(\lambda_X)}$ with distortion $D(\lambda_X)$, where $d(\lambda_X), D(\lambda_X)$ depend only on the doubling constant of X. Ideally $d(\lambda_X)$ should be $O(\log\lambda_X)$, but no bound depending only on λ_X is known. Moreover, as observed in [Gupta et al. 2003], the results of [Laakso 2000; Gupta et al. 2004] imply that the analogous result in ℓ_1 is false (see also [Lee et al. 2005] for a stronger negative result in ℓ_1). The following example shows that more work needs to be done towards solving this problem positively (if at all possible. In fact, we believe that this problem has a negative answer): Linear mappings cannot yield the required embedding without a positive lower bound on the resolution. Specifically, we claim that for every $D > 1$ there are arbitrarily large n-point subsets X_n of ℓ_2 which are doubling with constant 6, such that if there exists a linear mapping $T : \ell_2 \to \mathbb{R}^d$ which is D-bi-Lipschitz on X_n then $d \geq \frac{\log n}{\log D}$ (observe that by the Johnson-Lindenstrauss lemma **any** n point subset of ℓ_2 embeds with distortion D via a linear mapping into ℓ_2^k, with $k = O\left(\frac{\log n}{\log D}\right)$).

To see this fix $D > 1$ and an integer d. Let $\mathcal{N}$ be a $1/(4D)$ net in S^{d-1}. Write $n+1 = |\mathcal{N}|$ and $\mathcal{N} = \{x_1, \ldots, x_n\} \cup \{0\}$. Define $X = \{2^{-j}x_j\}_{j=1}^n$. Whenever $1 \leq i < j \leq n$ we have that

$$2^{-i} - 2^{-j} \leq \|2^{-i}x_i - 2^{-j}x_j\|_2 \leq 2^{-i} + 2^{-j} \leq 3(2^{-i} - 2^{-j}),$$

so X is embeddable into the real line with distortion 3. In particular, X is doubling with constant at most 6. However, X cannot be embedded into low dimensions using a linear mapping. Indeed, assume that $T : \mathbb{R}^d \to \mathbb{R}^k$ is a linear mapping such that for every $x, y \in X$, $\|x-y\|_2 \leq \|Tx-Ty\|_2 \leq D\|x-y\|_2$. Then for every i, $\|Tx_i\|_2 = 2^i\|T(2^{-i}x_i) - T(0)\|_2 \in [1, D]$. Take $x \in S^{d-1}$ for which $\|Tx\|_2 = \|T\| = \max_{y\in S^{d-1}}\|Ty\|_2$. There is $1 \leq i \leq n$ such that $\|x - x_i\|_2 \leq 1/(4D) \leq 1/2$. Then $\|T\| = \|Tx\|_2 \leq \|Tx_i\|_2 + \|T(x-x_i)\|_2 \leq D + \|T\| \cdot \|x - x_i\|_2 \leq D + \frac{1}{2}\|T\|$. Thus $\|T\| \leq 2D$. Now, for every $y \in S^{d-1}$ there is $1 \leq j \leq n$ for which $\|y - x_j\|_2 \leq 1/(4D)$. It follows that $\|Ty\|_2 \geq \|Tx_j\|_2 - \|T(y-x_j)\|_2 \geq 1 - \|T\|/(4D) \geq 1/2$. This implies that T is invertible, so necessarily $k \geq d$. This proves our claim since by standard volume estimates $|X| \leq (12D)^d$.

Acknowledgments

The authors would like to thank Mihai Badoiu, Robert Krauthgamer, James R. Lee and Vitali Milman, for discussions during the initial phase of this work.

REFERENCES

ACHLIOPTAS, D. 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. System Sci. 66,* 4, 671–687. Special issue on PODS 2001 (Santa Barbara, CA).

ALON, N. 2003. Problems and results in extremal combinatorics. I. *Discrete Math. 273,* 1-3, 31–53. EuroComb'01 (Barcelona).

ANDONI, A., DATAR, M., IMMORLICA, N., INDYK, P., AND MIRROKNI, V. 2005. Locality-sensitive hashing scheme based on stable distributions. In *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*. MIT Press.

ARYA, S. AND MALAMATOS, T. 2002. Linear-size approximate voronoi diagrams. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 147–155.

BARVINOK, A. 1997. Approximate counting via random optimization. *Random Structures Algorithms 11,* 2, 187–198.

BARVINOK, A. AND SAMORODNITSKY, A. 2001. The distance approach to approximate combinatorial counting. *Geom. Funct. Anal. 11,* 5, 871–899.

BARVINOK, A. AND SAMORODNITSKY, A. 2004. Random weighting, asymptotic counting, and inverse isoperimetry. Preprint.

BRINKMAN, B. AND CHARIKAR, M. 2005. On the impossibility of dimension reduction in l_1. *J. ACM 52,* 5, 766–788 (electronic).

Clarkson, K. 2005. Nearest-neighbor searching and metric space dimensions. In *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*. MIT Press.

Dasgupta, S. and Gupta, A. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures Algorithms 22,* 1, 60–65.

Durrett, R. 1996. *Probability: theory and examples*, Second ed. Duxbury Press, Belmont, CA.

Frankl, P. and Maehara, H. 1988. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Combin. Theory Ser. B 44,* 3, 355–362.

Gordon, Y. 1988. On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric aspects of functional analysis (1986/87)*. Lecture Notes in Math., vol. 1317. Springer, Berlin, 84–106.

Guédon, O. and Zvavitch, A. 2003. Supremum of a process in terms of trees. In *Geometric aspects of functional analysis*. Lecture Notes in Math., vol. 1807. Springer, Berlin, 136–147.

Gupta, A., Krauthgamer, R., and Lee, J. R. 2003. Bounded geometries, fractals, and low-distortion embeddings. *Annual Symposium on Foundations of Computer Science*, 534–543.

Gupta, A., Newman, I., Rabinovich, Y., and Sinclair, A. 2004. Cuts, trees and l_1-embeddings of graphs. *Combinatorica 24,* 2, 233–269.

Har-Peled, S. 2001. A replacement for voronoi diagrams of near linear size. *Annual Symposium on Foundations of Computer Science*, 94–103.

Har-Peled, S. and Mendel, M. 2006. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput. 35,* 5, 1148–1184 (electronic).

Heinonen, J. 2001. *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York.

Indyk, P. 1998. On approximate nearest neighbors in non-euclidean spaces. *Proceedings of the Symposium on Foundations of Computer Science*, 148–155.

Indyk, P. 2000. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000)*. IEEE Comput. Soc. Press, Los Alamitos, CA, 189–197.

Indyk, P. and Motwani, R. 1999. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98 (Dallas, TX)*. ACM, New York, 604–613.

Johnson, W. B. and Lindenstrauss, J. 1984. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*. Amer. Math. Soc., Providence, RI, 189–206.

Klartag, B. and Mendelson, S. 2005. Empirical processes and random projections. *J. Funct. Anal. 225,* 1, 229–245.

Konyagin, S. V. and Vol′berg, A. L. 1987. On measures with the doubling condition. *Izv. Akad. Nauk SSSR Ser. Mat. 51,* 3, 666–675.

Krauthgamer, R. and Lee, J. R. 2004. Navigating nets: simple algorithms for proximity search. In *SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 798–807.

Kushilevitz, E., Ostrovsky, R., and Rabani, Y. 2000. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput. 30,* 2, 457–474 (electronic).

Laakso, T. J. 2000. Ahlfors Q-regular spaces with arbitrary $Q > 1$ admitting weak Poincaré inequality. *Geom. Funct. Anal. 10,* 1, 111–123.

Lang, U. and Plaut, C. 2001. Bilipschitz embeddings of metric spaces into space forms. *Geom. Dedicata 87,* 1-3, 285–307.

Ledoux, M. and Talagrand, M. 1991. *Probability in Banach spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 23. Springer-Verlag, Berlin. Isoperimetry and processes.

Lee, J. R., Mendel, M., and Naor, A. 2005. Metric structures in L_1: dimension, snowflakes, and average distortion. *European J. Combin. 26,* 8, 1180–1190.

Lee, J. R. and Naor, A. 2004. Embedding the diamond graph in L_p and dimension reduction in L_1. *Geom. Funct. Anal. 14,* 4, 745–747.

Magen, A. 2002. Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications. In *Randomization and approximation techniques in computer science*. Lecture Notes in Comput. Sci., vol. 2483. Springer, Berlin, 239–253.

Matoušek, J. 2002. *Lectures on discrete geometry*. Graduate Texts in Mathematics, vol. 212. Springer-Verlag, New York.

Matoušek, J. 2006. On variants of the Johnson-Lindenstrauss lemma. Preprint.

Schechtman, G. 1989. A remark concerning the dependence on ϵ in Dvoretzky's theorem. In *Geometric aspects of functional analysis (1987–88)*. Lecture Notes in Math., vol. 1376. Springer, Berlin, 274–277.

Schechtman, G. 2006. Two observations regarding embedding subsets of Euclidean spaces in normed spaces. *Adv. Math. 200,* 1, 125–135.

Talagrand, M. 1987. Regularity of Gaussian processes. *Acta Math. 159,* 1-2, 99–149.

Talagrand, M. 1996. Majorizing measures: the generic chaining. *Ann. Probab. 24,* 3, 1049–1103.

Talagrand, M. 2001. Majorizing measures without measures. *Ann. Probab. 29,* 1, 411–417.

Talagrand, M. 2005. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.