Nearest Neighbor Preserving Embeddings

Piotr Indyk
MIT

Assaf Naor
Microsoft Research

Abstract

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.

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 became 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 [20, 26], to give efficient approximate nearest neighbor algorithms in high-dimensional spaces.

More recently (e.g., in [19]), 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*, 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 aformentioned 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 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*

[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 [20, 16] 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 nearest neighbor, not just the near neighbor.

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 [19]. 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 [8, 30]. 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 [15]). See [9] for a survey of nearest neighbor algorithms for doubling sets.

   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 [3]), 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 fascinating 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

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

mapping is linear, an embedding satisfying the conditions of this problem cannot be in general linear (this is discussed in section 4).

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 $(\mathbb{R}^k, \|\cdot\|_2)$ [16, 4], 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 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 [25], 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 [20, 26] 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 [20], while being much 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 $\{S_1 \ldots S_n\}$, where $S_i \subset \mathbb{R}^d$, $i = 1\ldots n$. Let $\lambda = \max_{i=1\ldots n}\lambda_{S_i}$. 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 Gq in $\{GS_1 \ldots GS_n\}$, then, with probability $1/2$, it is also a $(1+O(\epsilon))$-approximate nearest neighbor of q in $\{S_1 \ldots S_n\}$. This corollary provides a strong generalization of a result of [31], 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_∞ [18].

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, $B_X(x,\varepsilon)$ 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 [17]), 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 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, 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 [28]). 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 [5, 6, 7]. 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 [28]) 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 [35] 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 [24, 17] 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 [36, 37] states that for every $X\subseteq\ell_2$, $E_X=\Theta(\gamma_2(X))$ (we refer to [14] 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*

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

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 [13]:

Theorem 3.2 (Gordon [13]). *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 $x \in X$,*

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

Remark 3.1. Since the proof of the above theorem uses probabilistic method (specifically, Gordon [13] 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 [34, 33]), 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 [23] 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 satisfy the ψ_2 condition $\mathbb{E}\, e^{\gamma_{ij}^2/C^2} \le 2$ for some constant C. In this case the implied constants may depend on C. A particular case of interest is when Γ is a random ± 1 matrix- just like in Achlioptas' variant [1] of the Johnson-Lindenstrauss lemma [22], we obtain "database friendly" versions of Theorem 4.1. It should be pointed out here that although several papers [12, 10, 21, 1] 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. ψ_2 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 $\mathbb{R}^d$, $E_Z = O(\sqrt{\log n})$).

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 $\widetilde{X} = \left\{\frac{x-y}{\|x-y\|_2} :\ x, y \in X,\ \|x-y\|_2 \ge \delta\right\}$. Then

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

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

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 [2] (see also [32]) 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 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 $\{g_{ij} : \ 1 \le i \le k, \ 1 \le j \le n\}$ be i.i.d. standard gaussian random variables. Denoting $G = \frac{1}{\sqrt{k}}(g_{ij})$, by standard arguments (see [11]) 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 $D > 0$

$$\Pr\left[\big|\|Ga\|_2 - 1\big| \ge D\right] \le e^{-kD^2/8} \quad \text{and} \quad \Pr[\|Ga\|_2 \le 1/D] \le \left(\frac{3}{D}\right)^k. \tag{5}$$

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. *$d(Gx_0, G(X \setminus \{x_0\})) \le (1 + \varepsilon)d(x_0, X \setminus \{x_0\})$*
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 [13, 34, 33], 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 $k \ge C \log \lambda_X + 1$ and $D > 1$,*

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

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 $|S_t| \le \lambda_X$ 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 $\|t_{j+1}(x) - t_j(x)\|_2 \le 2^{-j+1}$, we get

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

provided that $k \geq C \log \lambda_X + 1$ and using the first estimate in (5). $\square$

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 (5) we get that $\Pr[\|Gy\|_2 \geq 1 + \varepsilon] \leq e^{-k\varepsilon^2/8}$. Thus for $k \geq C/\varepsilon^2$ 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

$$\Pr\left[\exists s \in S \, \exists x \in B(s, \varepsilon/4) \cap X_i, \ \|Gs - Gx\|_2 \geq \frac{\varepsilon\sqrt{i}}{4}\right] \leq \lambda_X^{\log_2(16 r_i/\varepsilon)} \cdot e^{-cki} \leq e^{-c'ki}. \tag{6}$$

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

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

Hence, by (5)

$$\begin{aligned} \Pr\left[\exists s \in S \ \ \|Gs\|_2 \leq 1 + \varepsilon + \frac{\varepsilon\sqrt{i}}{4}\right] &\leq \begin{cases} \lambda_X^{\log_2(16 r_i/\varepsilon)} e^{-c'' k \varepsilon^2} & i \leq 1/\varepsilon^2 \\ \lambda_X^{\log_2(16 r_i/\varepsilon)} \cdot (3C/\sqrt{i})^k & i > 1/\varepsilon^2 \end{cases} \\ &\leq \begin{cases} e^{-c''' k \varepsilon^2} & i \leq 1/\varepsilon^2 \\ i^{-c''' k} & i > 1/\varepsilon^2. \end{cases} \end{aligned} \tag{7}$$

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

Now, from (6) and (7) 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,

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

for large enough k. This completes the proof of Theorem 4.1. $\square$

Remark 4.1. Since $\gamma \dim(X) = O(\log \lambda_X)$ Theorem 4.1 sheds some light on the following well known problem (which is folklore, but apparently has been first stated explicitly in print in [27]): 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 the analogous result in ℓ_1 is known to be false [29]. The following example shows that more work needs to be done towards this interesting problem: 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

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

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

[3] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing scheme based on stable distributions. In *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*. MIT Press, 2005.

[4] S. Arya and T. Malamatos. Linear-size approximate voronoi diagrams. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 147–155, 2002.

[5] A. Barvinok. Approximate counting via random optimization. *Random Structures Algorithms*, 11(2):187–198, 1997.

[6] A. Barvinok and A. Samorodnitsky. The distance approach to approximate combinatorial counting. *Geom. Funct. Anal.*, 11(5):871–899, 2001.

[7] A. Barvinok and A. Samorodnitsky. Random weighting, asymptotic counting, and inverse isoperimetry. 2004. Preprint.

[8] B. Brinkman and M. Charikar. On the impossibility of dimension reduction in ℓ_1. *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.

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

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

[11] R. Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

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

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

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

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

[16] S. Har-Peled. A replacement for voronoi diagrams of near linear size. *Annual Symposium on Foundations of Computer Science*, 2001.

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

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

[19] P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. *Annual Symposium on Foundations of Computer Science*, 2000.

[20] P. Indyk and R. Motwani. Approximate nearest neighbor: towards removing the curse of dimensionality. *Proceedings of the Symposium on Theory of Computing*, 1998.

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

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

[23] B. Klartag and S. Mendelson. Empirical processes and random projections. *J. Funct. Anal.* To appear.

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

[25] R. Krauthgamer and J. R. Lee. Navigating nets: Simple algorithms for proximity search. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[26] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *Proceedings of the Thirtieth ACM Symposium on Theory of Computing*, pages 614–623, 1998.

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

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

[29] J. R. Lee, M. Mendel, and A. Naor. Metric structures in L_1: Dimension, snowflakes, and average distortion. *European J. Combin.* To appear.

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

[31] A. Magen. Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications. *Proceedings of RANDOM*, 2002.

[32] J. Matoušek. *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.

[33] G. Schechtman. Two observations regarding embedding subsets of Euclidean spaces in normed spaces. *Adv. Math.* To appear.

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

[35] M. Talagrand. Regularity of Gaussian processes. *Acta Math.*, 159(1-2):99–149, 1987.

[36] M. Talagrand. Majorizing measures: the generic chaining. *Ann. Probab.*, 24(3):1049–1103, 1996.

[37] M. Talagrand. Majorizing measures without measures. *Ann. Probab.*, 29(1):411–417, 2001.