An Overview of ϵ-Distortion Random Embeddings

Ryan Marren
Johns Hopkins University
EN.600.469 Final Project

Abstract

Geometric algorithms operating on a subset of Euclidian space $X \subseteq \mathbb{R}^D$ pervade Computational Geometry, Computer Vision, and Machine Learning. Often, the geometric information needed from X (distances, angles, norms, etc.) can be approximated well by some set $Y \subseteq \mathbb{R}^k$ related to X by an embedding $f : X \to Y$. Here, we explore such mappings f which are both *oblivious* and *random*, meaning that they are constructed randomly and independently of the set X. In particular, we construct mappings f in the case that X is a finite point set, generalize these results to the case where X is a linear subspace, and discuss the computational trade-offs between different methods of constructing f.

1 Embeddings with ϵ-Distortion

Suppose $X \subseteq \mathbb{R}^D$ is a subset of Euclidian space.

Definition 1. *An* ***Embedding with ϵ-distortion*** *of a set X into a set Y is a map $f : X \to Y$ such that Y is a subset of Euclidian space and*

$$(1-\epsilon)\|x\|^2 \leq \|f(x)\|^2 \leq (1+\epsilon)\|x\|^2$$

Such an embedding f can be useful in the following way: if the *ambient dimension* of X is larger than the *intrinsic dimension* of X, and if the *ambient dimension* of Y is closer to the *intrinsic dimension* of X than the *ambient dimension* of X is, it may be beneficial to find $Y = f(X)$ and perform computations in Y rather than X.

We offer an ideal example as motivation.

Example 1.1. Suppose $X = \{x_i\}_{i=1}^n$ is a subset of $\mathbb{R}^D$ of cardinality n and V is the linear subspace spanned by the vectors in X. Let X_d be the subspace spanned by the first d singular vectors $U_d = \{u_i\}_{i=1}^d$ of $A = \big(x_1|x_2|\ldots|x_n\big) \in \mathbb{R}^{D\times n}$. Since U_d is an orthonormal set by definition of the *SVD* , we have that the linear map $P_{X_d} : X \to X_d$ is an embedding of X into X_d. Suppose we choose $d = dim(V) = rank(A)$, the *intrinsic dimension* of the space spanned by the point set X. If $A_d = \big(P_{X_d}(x_1)|P_{X_d}(x_2)|\ldots|P_{X_d}(x_n)\big)$, it can be shown that $\|A - A_d\|_F^2 = \sum_{i=1}^n \|x_i - P_{X_d}x_i\|^2 = 0$. So P_{X_d} is an embedding of X into X_d with distortion 0.

Here, the construction of $f = P_{X_d}$ is deterministic and dependent on the data X.

In a sense, 1.1 is the best we can do. We can't get a distortion less than zero, and embeddings $f : X \to Y$ where $dim(span(Y))$ is less than the intrinsic $dim(span(X))$ do not exist with $\epsilon < 1$.

Lemma 2. *There cannot exist an embedding f of a set $X \subseteq \mathbb{R}^D$ with $dim(span(X)) = d$ into a space Y of dimension $k < d$ with ϵ-distortion of $\epsilon < 1$.*

Proof. Suppose $f : X \to Y$ is a linear map with $dim(span(Y)) = k < dim(span(X)) = d$. By the rank-nullity theorem, $dim(span(X)) = dim(span(Y)) + dim(ker(f))$. Since $dim(span(X)) = d, dim(span(Y)) = k$, and $k < d$, we have $dim(ker(f)) > 0$, so $\exists x \in X$ such that $f(x) = 0$ and

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

$x \neq 0$. Thus $\|x\|^2 > 0$ since $x \neq 0$ but $\|f(x)\|^2 = 0$ since $f(x) = 0$, meaning the only way to satisfy $(1-\epsilon)\|x\|^2 \leq \|f(x)\|^2 \ \forall x \in X$ is $\epsilon = 1$. □

The problem is not completely solved here, since computing a *SVD* of a matrix $A \in \mathbb{R}^{D \times n}$ can be very expensive when D is large (think of each vector $x \in X$ being a bag-of-words representation in an NLP problem), and improvements can be made when $d << D$.

In section 2, we will explore such improvements for cases where $|X| = n$, e.g. X is a finite cardinality set. In section 3, we will generalize these results to X with infinite cardinality, and explore connections with the results from section 2. Finally, in section 4 we will briefly discuss applications of the theory presented.

2 Random Embeddings of Finite Sets

2.1 Inner Products with Random Vectors, Weak Unbiased Estimators

We begin with an alternate presentation of a well known concentration of measure phenomenon. Consider a multivariate standard normal distribution $p = \mathcal{N}(0, I_D)$, and consider the embedding

$$g : \{x\} \to \mathbb{R}, \ g(x) \mapsto \langle r, x \rangle$$

from the cardinality one set $\{x\} \subseteq \mathbb{R}^D$ into $\mathbb{R}$. The map g simply takes its input to the inner product of that input with a random vector r distributed p. Making the note that each entry r_i of r is itself a random variable distributed $r_i \sim \mathcal{N}(0,1)$, and that $cr_i \sim \mathcal{N}(0, c^2)$, we see that $g(x) = \langle r, x \rangle = \sum_{i=1}^D x_i r_i$ is the sum of D normal random variables with variances corresponding to the squared values of the entries of the vector x. E.g., g is a random, oblivious embedding with $g(x) = \langle r, x \rangle \sim \mathcal{N}(0, \sum_{i=1}^D x_i^2)$ and g not dependent on $\{x\}$.

There are some useful facts about the random embedding g:

1. $g(x)^2 = \langle r, x \rangle^2$ is an unbiased estimator for $\|x\|^2$: we see that $\mathbf{E}[g(x)^2] = \mathbf{E}[\langle r, x \rangle^2] = \sum_{i=1}^D x_i^2 = \|x\|^2$.
2. Since $g(x) = \langle r, x \rangle \sim \mathcal{N}(0, \|x\|^2)$, if we denote $Z = \mathcal{N}(0,1)$ as the standard normal distribution, we have that $\frac{g(x)}{\|x\|} \sim Z$.

Using these facts, we can analyze the probability that g is an embedding from $\{x\}$ into $\mathbb{R}$ with ϵ-distortion, which can be computed by

$$\Pr\left[(1-\epsilon)\|x\|^2 \leq g(x)^2 \leq (1+\epsilon)\|x\|^2\right] = \Pr\left[\frac{\langle r, x \rangle^2}{\|x\|^2} \in (1-\epsilon, 1+\epsilon)\right] = \Pr\left[\left|\frac{\langle r, x \rangle^2}{\|x\|^2} - 1\right| \leq \epsilon\right].$$

Using fact 2, we see that $Z^2 = \frac{\langle r, x \rangle^2}{\|x\|^2}$ is a $\mathcal{X}_1^2$ random variable (chi-squared with 1 degree of freedom). In general $\mathcal{X}_m^2$ random variables are sub-exponential, so if $Y \sim \mathcal{X}_m^2$ we have that $\Pr[|Y - 1| \geq \epsilon] \leq \exp\left(\frac{-\epsilon^2 m}{8}\right)$. This means that

$$\Pr\left[\left|\frac{\langle r, x \rangle^2}{\|x\|^2} - 1\right| \geq \epsilon\right] \leq \exp\left(\frac{-\epsilon^2}{8}\right)$$

so

$$\Pr\left[\left|\frac{\langle r, x \rangle^2}{\|x\|^2} - 1\right| \leq \epsilon\right] \geq 1 - \exp\left(\frac{-\epsilon^2}{8}\right)$$

In general, a Random Embedding from D-dimensional space into 1-dimensional space is not enough, as shown by the following example:

Example 2.1. Consider a modest $\epsilon = \frac{1}{100}$, and notice that

$$\Pr\left[\left|\frac{\langle r, x \rangle^2}{\|x\|^2} - 1\right| \leq \frac{1}{100}\right] \geq 1 - \exp\left(\frac{-(\frac{1}{100})^2}{8}\right) \approx 0.0001$$

is a very uninformative bound.

The problem here is that while the squared norm of the embedding $g(x)^2 = \langle r, x\rangle^2$ is an unbiased estimator for $\|x\|^2$, it is a weak estimator in the sense that its variance is quite high, giving it 'fat tails' and making the bound we see above very loose.

2.2 Boosting Weak Unbiased Estimators

To alleviate the weakness seen in 2.1, we note that the average of unbiased estimators for a random variable is itself an unbiased estimator. In this spirit, we construct a new embedding

$$f : \{x\} \subseteq \mathbb{R}^D \to \mathbb{R}^k, f(x) \mapsto \bigg[g_1(x), g_2(x), \dots, g_k(x)\bigg]^T$$

where each g_i is an embedding from $\{x\} \subseteq \mathbb{R}^D$ into $\mathbb{R}$, i.e. $g_i(x) = \langle r_i, x\rangle$ where $r_i \sim \mathcal{N}(0, I_D)$. In this case, $\frac{1}{k}\|f(x)\|^2 = \frac{1}{k}\sum_{i=1}^k g_i(x)^2$ is an unbiased estimator of $\|x\|^2$ since

$$\mathbf{E}[\frac{1}{k}\|f(x)\|^2] = \frac{1}{k}\mathbf{E}[\sum_{i=1}^k g_i(x)^2] = \frac{1}{k}\mathbf{E}[\sum_{i=1}^k \langle r_i, x\rangle^2] = \frac{1}{k}\sum_{i=1}^k \mathbf{E}[\langle r_i, x\rangle^2] = \frac{1}{k}\sum_{i=1}^k \|x\|^2 = \|x\|^2$$

We note that it is intuitive to view $\frac{1}{k}\|f(x)\|^2$ as a strong estimator of $\|x\|^2$ constructed by averaging the contributions of k *weak* estimators $\{g_i(x)\}_{i=1}^k$.

We expect the random variable $\|f(x)\|^2$ to have a lower variance than $g(x)^2$, which should result better measure concentration improving the bound on the failure probability of f being an ϵ-distortion.

We note that

$$\frac{\|f(x)\|^2}{\|x\|^2} = \frac{\sum_{i=1}^k \langle r_i, x\rangle^2}{\|x\|^2}$$

where $\frac{\langle r_i, x\rangle}{\|x\|} \sim \mathcal{N}(0, 1)$, meaning that $\frac{\|f(x)\|^2}{\|x\|^2} \sim \mathcal{X}_k^2$. This gives the following:

$$\Pr\left[\left|\frac{\|f(x)\|^2}{\|x\|^2} - 1\right| \le \epsilon\right] \ge 1 - \exp\frac{-k\epsilon^2}{8} \tag{1}$$

Now we improve the result from 2.1

Example 2.2. Suppose we set $k = \frac{80}{\epsilon^2}$, and we desired the same $\epsilon = \frac{1}{100}$ as we did in 2.1, the probability that the embedding f from $\{x\}$ into $\mathbb{R}^k$ is an ϵ-distortion is given by:

$$\Pr\left[\left|\frac{\|f(x)\|^2}{\|x\|^2} - 1\right| \le \epsilon\right] \ge 1 - e^{-10} \approx 0.99995$$

Thus, with high probability, f is an ϵ-distortion. This improvement can be thought of intuitively as a 'boosting' operation, in that we have taken a set of weak estimators (set of randomly generated g's) and averaged their predictions of $\|x\|^2$ to form one strong estimator f.

2.3 Randomized Embeddings of Single Points

We can view $f : \{x\} \subseteq \mathbb{R}^D \to R^k, f(x) \mapsto [\langle r_1, x\rangle, \dots, \langle r_k, x\rangle]^T$ as a linear map embedding a D-dimensional vector into a k-dimensional subspace spanned by random vectors. Indeed, if we randomly sampled the set of vectors $\{r_i\}_{i=1}^k, r_i \sim \mathcal{N}(0, I_D)$, $\langle r_1, \dots, r_k\rangle$ is a k-dimensional random subspace. If $R_k = (r_1|r_2|\dots|r_k)$, then $f(x) = R_k^T x$.

Taking this view, we derive a more application-friendly version of 1.

$$
\begin{aligned}
\Pr\left[\left|\frac{\|f(x)\|^2}{\|x\|^2}-1\right|\geq\epsilon\right] &= \Pr\left[\left|\frac{\|R_k^T x\|^2}{\|x\|^2}-1\right|\geq\epsilon\right] \\
&= \Pr\left[\left|\|R_k^T x\|^2-\|x\|^2\right|\geq\|x\|^2\epsilon\right] \\
&= 1-\Pr\left[-\|x\|^2\epsilon\leq\|R^T x\|^2-\|x\|^2\leq\|x\|^2\epsilon\right] \\
&= 1-\Pr\left[(1-\epsilon)\|x\|^2\leq\|R_k^T x\|^2\leq(1+\epsilon)\|x\|^2\right]\leq\exp\frac{-k\epsilon^2}{8}
\end{aligned}
$$

Giving us

$$
\Pr\left[(1-\epsilon)\|x\|^2\leq\|R_k^T x\|^2\leq(1+\epsilon)\|x\|^2\right]\geq 1-\exp\left(\frac{-k\epsilon^2}{8}\right). \tag{2}
$$

To recap: we began with some vector $x\in\mathbb{R}^D$ where D can be an arbitrarily large natural number. Then, we constructed the matrix R_k though a simple random sampling procedure, and we have the fact that for arbitrarily small ϵ that the embedding $R_k^T:\{x\}\subseteq\mathbb{R}^D\rightarrow\langle r_1,\ldots,r_k\rangle$ is an ϵ-distortion with probability $1-\exp\left(\frac{-k\epsilon^2}{8}\right)$.

This is an interesting fact because 2 is not dependent whatsoever on the ambient dimension D, and is useful in cases where we can make $k << D$ while maintaining a good probability bound. In fact, a result called the *Johnson–Lindenstrauss Lemma* explores such cases and gives guarantees for the dimensionality reduction technique of random embeddings, which is widely used in Theoretical Computer Science and Machine Learning.

2.4 The Johnson–Lindenstrauss Lemma

While the above analysis of embedding a single point is a simple way to understand Random Embeddings, in most use cases we would like to embed many more than one point. This idea is explored in the *Johnson–Lindenstrauss Lemma* .

Lemma 3. *(Johnson–Lindenstrauss)*
Suppose that $X=\{x_i\}_{i=1}^n$ is a set of vectors $x_i\in\mathbb{R}^D$, and that $R_k=(r_1|r_2|\ldots|r_k)$ where $\{r_i\}_{i=1}^k$ is a set of k random vectors drawn from $\mathcal{N}(0,I_D)$. Then if ϵ is an arbitrarily small constant and $k=c\frac{\log n}{\epsilon^2}$ where c is a constant, then

$$
\Pr\left[(1-\epsilon)\|x\|^2\leq\|R_k^T x\|^2\leq(1+\epsilon)\|x\|^2\ \forall\ x\in X\right]\geq 1-O(n^{-c+1}).
$$

The reason we used more advanced concentration inequalities in 1 and 2 is that our end goal is to embed a large set of points rather than a single point, meaning that it would be helpful to have an exponentially small failure probability for the embedding of one point in order to use a union bound argument to get a result for all points. This is the main idea of the proof of the *Johnshon–Lindenstrauss Lemma*.

Proof.

$$\Pr\left[(1-\epsilon)\|x\|^2 \leq \|R_k^T x\|^2 \leq (1+\epsilon)\|x\|^2 \ \forall\ x \in X\right] =$$

$$1 - \Pr\left[\exists\ x \in X : \|R_k^T x\|^2 \notin [(1-\epsilon)\|x\|^2, (1+\epsilon)\|x\|^2]\right] \geq$$

$$1 - \sum_{x \in X} \Pr\left[\|R_k^T x\|^2 \notin [(1-\epsilon)\|x\|^2, (1+\epsilon)\|x\|^2]\right] =$$

$$1 - \sum_{i=1}^{n} \Pr\left[\|R_k^T x_i\|^2 \notin [(1-\epsilon)\|x_i\|^2, (1+\epsilon)\|x_i\|^2]\right] =$$

$$1 + \sum_{i=1}^{n} \left(\Pr\left[(1-\epsilon)\|x_i\|^2 \leq \|R_k^T x_i\|^2 \leq (1+\epsilon)\|x_i\|^2\right] - 1\right) =$$

$$1 - n + \sum_{i=1}^{n} \left(\Pr\left[(1-\epsilon)\|x_i\|^2 \leq \|R_k^T x_i\|^2 \leq (1+\epsilon)\|x_i\|^2\right]\right) \geq$$

$$1 - n + n(1 - \exp\left(\frac{k\epsilon^2}{8}\right)) = 1 - n\exp\left(\frac{-clog(n)}{8}\right) = 1 - e^{1/8} n e^{\log(n^- c)} = 1 - e^{1/8} n^{-c+1}$$

□

This lemma gives us a more powerful statement than did 2. Here, we have that, with high probability, if we sample a matrix R_k as we did above, the embedding

$$R_k^T : X \subseteq \mathbb{R}^D \to \langle r_1, \ldots r_k\rangle \subseteq \mathbb{R}^k, R_k^T(x) \mapsto [\langle r_1, x\rangle, \ldots, \langle r_k, x\rangle]^T$$

from the set X into $\langle r_1, \ldots r_k\rangle$ is an ϵ-distortion.

Additionally, a useful corollary results from this lemma:

Corollary 4. *Suppose that $X = \{x_i\}_{i=1}^n$ is a set of vectors $x_i \in \mathbb{R}^D$, and that $R_k = (r_1|r_2|\ldots|r_k)$ where $\{r_i\}_{i=1}^k$ is a set of k random vectors drawn from $\mathcal{N}(0, I_D)$. Then if ϵ is an arbitrarily small constant and $k = 2c\frac{\log n}{\epsilon^2}$ where c is a constant,*

$$\Pr\left[(1-\epsilon)\|u-v\|^2 \leq \|R_k^T(u-v)\|^2 \leq (1+\epsilon)\|u-v\|^2 \ \forall\ u, v \in X\right] \geq 1 - O(n^{-c+1}).$$

Proof. We construct a new set of vectors $Z = \{u - v : u, v \in X\}$. $|X| = n$, so $|Z| = \binom{n}{2} = \frac{n(n-1)}{2} = O(n^2)$. By 3, we have that

$$\Pr\left[(1-\epsilon)\|u-v\|^2 \leq \|R_k^T x\|^2 \leq (1+\epsilon)\|u-v\|^2 \ \forall\ (u-v) \in Z\right] \geq 1 - O(n^{-c+1}).$$

where $k = c\frac{\log n^2}{\epsilon^2} = 2c\frac{\log n}{\epsilon^2}$. This is sufficient since there is a one to one mapping between elements $z \in Z$ and pairs $(u, v) \in X$. □

The applications of this lemma and its corollary are numerous. Any geometric algorithm with a dependence on D the dimension of input vectors can be sped up drastically with only a small loss in precision. Further, there are many algorithms with an exponential dependence on the dimension D (nearest neighbor, minimum spanning tree, point location) which can be made tractable with a Random Embedding preprocessing step.

3 Oblivious Subspace Embeddings

3.1 Embeddings of Infinite Sets

The main result from section 1 was that given a *finite* set of vectors $X = \{x_i\}_{i=1}^n$, we can find an embedding f from X into a random subspace $\langle r_1, \ldots r_k\rangle$ spanned by $\{r_i\}_{i=1}^k, r_i \sim \mathcal{N}(0, I_D)$ which is an ϵ-distortion with high probability.

In our definition 1, we purposely did not specify the cardinality of X, suggesting that perhaps an ϵ-distortion can exist for a set X with infinite cardinality. Indeed, we give a trivial example of such an embedding to motivate the search for more:

Example 3.1. Suppose $X \subseteq \mathbb{R}^d$ is a d-dimensional linear subspace with an orthonormal basis $\{s_i\}_{i=1}^d$. Then $P : X \to P(X) \subseteq \mathbb{R}^d, P(x) = Bx$ where $B = (s_1|s_2|\dots|s_d)$ is an embedding of X into a d dimensional space with distortion 0. This is simply because B is an orthogonal matrix, so $\|P(x)\|^2 = \|Bx\|^2 = \|x\|^2$.

Unlike the case where X was a finite set, if we have a subspace X and know that it is d-dimensional, it is impossible to achieve a dimensionality reduction result lowering d. Indeed, by 2, there cannot exist an embedding from X into Y where $dim(Y) < d$ with ϵ-distortion, $\epsilon < 1$. Intuitively, this is because by describing X as a d-dimensional subspace, we are inherently speaking of its intrinsic dimension which is already minimal.

Random Embeddings still have use here in the following sense: suppose $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^D$ is a finite set of vectors, and we know that all the vectors $x_i \in S$ where S is a d-dimensional subspace. It would be nice to perform geometric computations in S rather than X if $d << D$. Unfortunately, we do not know the subspace S, but recall from 1.1 that we could recover S up to a rotation by using the *singular value decomposition* method. However, if D is large, it is likely that the *SVD* could be too expensive to compute. Thus, if we settle for an approximation $\tilde{S}$ of dimension k to S, we could construct an embedding $f : S \to \tilde{S}$ such that

$$(1-\epsilon)\|x\|^2 \leq \|f(x)\|^2 \leq (1+\epsilon)\|x\| \ \forall \ x \in S,$$

i.e. f is an embedding from S into $\tilde{S}$ with ϵ-distortion with high probability. Then, we could perform computations in $\tilde{S}$ of dimension k rather than S of dimension d, which is still much better than performing computations in $\mathbb{R}^D$ of dimension D. Note that here, $d \leq k$ necessarily, but the benefit comes from results that show $d \leq k << D$.

Such approximations $\tilde{S}$ can found using a technique called *oblivious subspace embeddings*.

Definition 5. *A* $(D, d, k, \epsilon, \delta)$***-OSE*** *is a random linear operator* f *represented by a matrix* $\Pi \in \mathbb{R}^{k \times D}$ *such that for any fixed d-dimensional subspace* S *equivalent to the span of vectors* $\{x_i\}_{i=1}^n$ *of ambient dimension D, the map*

$$f : S \to \tilde{S}, f(x) \to \Pi x$$

is an embedding from a subspace S *of dimension* d *into* $\tilde{S}$ *of dimension* k *with* ϵ*-distortion with probability at least* $1 - \delta$.

3.2 Connections to Random Embeddings of Single Points

Recall that in 2.3 we show there exists a random embedding f

$$f : \{x\} \subseteq \mathbb{R}^D \to \mathbb{R}^k, f(x) \to \bigg[\langle r_1, x\rangle, \dots, \langle r_k, x\rangle\bigg]^T$$

from a set $\{x\}$ of cardinality one into a set $\mathbb{R}^k$ of dimension k which is an ϵ-distortion with probability at least $1 - \exp\left(\frac{-k\epsilon^2}{8}\right)$.

We now note that if you have an embedding of ϵ-distortion for one point $x \in \mathbb{R}^D$, then you have an embedding of ϵ-distortion for the entire 1-dimensional subspace spanned by x, since f is a linear map implying that

$$(1-\epsilon)\|x\|^2 \leq \|f(x)\|^2 \leq (1+\epsilon)\|x\|^2 \Rightarrow (1-\epsilon)\|cx\|^2 \leq \|f(cx)\|^2 \leq (1+\epsilon)\|cx\|^2 \ \forall \ c \in \mathbb{R}.$$

Thus, the random linear operator f is also a $(D, 1, k, \epsilon, \exp\left(\frac{-k\epsilon^2}{8}\right))$-OSE.

3.3 Generalizing to Subspaces

Suppose we had a $(D, d, k, \epsilon, \delta)$-OSE f represented by a matrix Π for a d-dimensional subspace S. By the definition of an ϵ-distortion, we have that

$$(1-\epsilon)\|x\|^2 \leq \|\Pi x\|^2 \leq (1+\epsilon)\|x\|^2$$

for all $x \in S$.

We recall from the previous section that if we are able to embed a single point x with ϵ-distortion, then we can embed every point in the span of x for free since f is a linear operator. Since any $x \in S$ is a vector in $\mathbb{R}^k$ and can be written as $\|x\| \frac{x}{\|x\|}$, it suffices to show that we can embed all unit vectors in S. Further, since S is a linear subspace, there must be some orthonormal basis $\{s_i\}_{i=1}^k$ and matrix $U = (s_1 | \dots | s_k)$ such that for each unit vector $x \in S$, there is some unit vector $x' \in \mathbb{S}^d = \{x \in \mathbb{R}^d : \|x\| = 1\}$ such that $x = Ux'$. In summary, if we are able to embed with ϵ-distortion all vectors in the set $\{x : x = Ux' \text{ for } x' \in \mathbb{S}^d\}$, then we have successfully embedded all points in the set $\{y : y \in im(U)\} = S$.

To use this fact, we can take a different view of the distortion inequality.

$$(1-\epsilon)\|x\|^2 \leq \|\Pi x\|^2 \leq (1+\epsilon)\|x\|^2$$
$$\|x\|^2 - \epsilon\|x\|^2 \leq \|\Pi x\|^2 \leq \|x\|^2 + \epsilon\|x\|^2$$
$$-\epsilon\|x\|^2 \leq \|\Pi x\|^2 - \|x\|^2 \leq \epsilon\|x\|^2$$
$$-\epsilon \leq \frac{\|\Pi x\|^2 - \|x\|^2}{\|x\|^2} \leq \epsilon$$
$$\left| \frac{\|\Pi x\|^2 - \|x\|^2}{\|x\|^2} \right| \leq \epsilon$$

From the previous discussion, we are only worried with embedding x such that $x = Ux'$, so we can re-write this as

$$\left| \frac{\|\Pi Ux'\|^2 - \|Ux'\|^2}{\|Ux'\|^2} \right| \leq \epsilon$$

and since U is orthogonal and x' is a unit vector, we get that $\|Ux'\|^2 = \|x'\|^2$, giving the inequality

$$|\|\Pi Ux'\|^2 - \|x'\|^2| \leq \epsilon$$

Finally, we expand the norms to reveal

$$|x'^T U^T \Pi^T \Pi U^T x' - x'^T x'| \leq \epsilon$$
$$|x'^T M x'| \leq \epsilon$$

where we define $M := U^T \Pi^T \Pi U - I$. This simplifies the problem to bounding $|x'^T M x'|$ for all $x' \in \mathbb{S}^d$. This can be done by bounding the operator norm of M, which is equal to $\sup_{x' \in \mathbb{S}^d} \|Mx'\|^2$. To do this, we consider a subset of points of $\mathbb{S}^d$ as a representative set of the entire space by using a net argument.

Definition 6. *A maximal γ-net $\Sigma \subseteq \mathbb{S}^d$ is a set of points where:*

1. *For all $x, y \in \Sigma$ where $x \neq y$, $\|x - y\| \geq \gamma$.*
2. *For all $z \in \mathbb{S}^d$, there is some $x \in \Sigma$ with $\|x - z\| < \gamma$*

Lemma 7. *A maximal $\frac{1}{4}$-net of $\mathbb{S}^d$ has $|\Sigma| \leq 6^d$.*

Proof. Suppose we have Σ a maximal $\frac{1}{4}$ net of $\mathbb{S}^d$. Imagine that around each point $x \in \Sigma$ we took the ball $B_{1/5}(x)$ to get the set $\{B_{1/5}(x)\}_{x \in \Sigma}$. These balls are in one-to-one correspondence with the points $x \in \Sigma$, so we can instead count these balls. All of these balls are disjoint (by property 1 of the $\frac{1}{4}$ net), so we know that if we have some set Q which contains all the balls (e.g. $Q \cap \bigcup_{x \in \Sigma} B_{1/5}(x) = \bigcup_{x \in \Sigma} B_{1/5}(x)$), then $|\Sigma| \leq \frac{Vol(B_{1/5}(\xi))}{Vol(Q)}$ where ξ is any point, since the volume will be the same regardless.

Noticing that $Q = B_{6/5}(\xi)$ is a valid set (since all the balls $\{B_{1/5}(x)\}_{x \in \Sigma}$ are contained within the $\frac{6}{5}$-sphere), we have $|\Sigma| \leq \frac{Vol(B_{6/5}(\xi))}{Vol(B_{1/5}(\xi))} \leq 6^d$. □

Now, we show that Σ is 'representative' of the set $\mathbb{S}^d$.

Lemma 8. *If A is a symmetric matrix, and Σ is a $\frac{1}{4}$-net of $\mathbb{S}^d$, then*

$$\|A\|_{op} \leq 4\left|\max_{x\in\Sigma} x^T Ax\right|$$

Proof. There must be some unit vector z where $|z^T Az| = \|A\|_{op}$. By the definition of the $\frac{1}{4}$-net, there must be some $x \in \Sigma$ such that $\|x - z\| \leq \frac{1}{4}$.

By writing

$$z^T Az = ((z-x)+x)^T A((z-x)+x) = (z-x)^T A(z-x) + (z-x)^T Ax + x^T A(z-x) + x^T Ax$$

We get that

$$\begin{aligned} |x^T Ax| &\geq |z^T Az| - |(z-x)^T A(z-x)| - |(z-x)^T Ax| - |x^T A(z-x)| \\ &\geq \|A\|_{op} - \frac{\|A\|_{op}}{4} - \frac{\|A\|_{op}}{4} - \frac{\|A\|_{op}}{16} \geq \frac{\|A\|_{op}}{4} \end{aligned}$$

the last step being a result of the following inequalities:

$$\begin{aligned} |(z-x)^T Ax| \leq \|(z-x)^T A\|\|x\| \leq \|A\|_{op}\|x-z\| < \frac{\|A\|_{op}}{4} \\ |x^T A(z-x)| \leq \|A(z-x)\|\|x\| \leq \|A\|_{op}\|x-z\| < \frac{\|A\|_{op}}{4} \\ |(x-z)^T A(z-x)| \leq \|A(z-x)\|\|z-x\| \leq \|A\|_{op}\|x-z\|^2 < \frac{\|A\|_{op}}{16} \end{aligned}$$

Since this is an example of one x such that $\|A\|_{op} \leq 4|x^T Ax|$, the claim holds. □

Due to this lemma, since M is symmetric we know that $\|M\|_{op} \leq 4\left|\max_{x'\in\Sigma} x'^T Mx'\right|$. Thus, if we are able to bound the distortion of embeddings of 6^d points, then we get that distortion is bounded for all points in S. Recalling the *Johnson–Lindenstrauss* lemma, using a matrix constructed identically to the R_k matrix discussed in that lemma, we can embed all 6^d points into a random subspace $\tilde{S}$ with dimension $k = 192\frac{d}{\epsilon^2}$ with distortion $\frac{\epsilon}{4}$ with $\delta = 6^{-d}$. Note again that this result is entirely independent of the ambient dimension D.

4 Random Embeddings in Practice

4.1 Sparsity of Π

Upon generating a random matrix $\Pi \in \mathbb{R}^{D\times k}$, an embedding of vectors in a matrix X (or the embedding of a subspace V spanned by the vectors in the columns of X) can be computed by simply taking the matrix product $Y = \Pi^T X$. If the matrix Π is dense (which it would be if we naively drew its columns from a multivariate normal distribution like in the above analyses) this matrix product would take $O(kDn)$. There are distributions from which you can draw vectors to span a random subspace which are not Gaussian, and particularly are not dense.

One simple way to do this is to independently choose a row of each column and set that value to 1 or -1 with 50% chance each, setting the un-chosen values to 0. This yields a fully sparse matrix with a much improved matrix multiplication time, but in the Oblivious Subspace Embedding case, this will make the dependence of k the dimension of the embedding quadratic in the intrinsic dimension d rather than linear [5].

A compromise between density and sparsity can be found in an approach called OSNAP [8], which is similar to the sparse case except we independently choose s rows from each column and set each of these values to $\pm\frac{1}{\sqrt{s}}$ with 50% chance each, setting unchosen values to 0. This approach allows you to change the parameter s to trade off between run time and embedding dimension.

4.2 Uses in Numerical Linear Algebra

We have seen that essentially any matrix X can be compressed significantly by embedding columns into a lower dimensional space, without significantly changing the geometric properties of the original matrix. This is an incredible versatile tool, and has lead to numerous advances in numerical linear algebra including speedups in least-squared regression, low rank approximation, and graph sparsification. See [9] for a very detailed survey of these advances.

References

[1] 6.854/18.415 Advanced Algorithms, Spring 2016. URL http://people.csail.mit.edu/moitra/854.html.

[2] CS/CMS 139: Advanced Algorithms. URL http://courses.cms.caltech.edu/cs139/.

[3] Princeton University CS Dept COS521: Advanced Algorithm Design Fall 2015. URL http://www.cs.princeton.edu/courses/archive/fall15/cos521/.

[4] Sanjeev Arora, Elad Hazan, and Satyen Kale. A Fast Random Sampling Algorithm for Sparsifying Matrices. 2006. URL https://ie.technion.ac.il/~ehazan/papers/AHKsample2006.pdf.

[5] Kenneth L Clarkson and David P Woodruff. Low Rank Approximation and Regression in Input Sparsity Time. 2013. URL https://arxiv.org/pdf/1207.6365.pdf.

[6] Michael B Cohen. Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities. 2015. URL http://delivery.acm.org/10.1145/2890000/2884456/p278-cohen.pdf?ip=128.220.160.202&id=2884456&acc=ACTIVE%20SERVICE&key=7777116298C9657D%2E34B115928DB6308C%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=763818006&CFTOKEN=49673000&__acm__=1495014864_a9b5940c926933c5e96fceae43de1a40.

[7] Xiangrui Meng and Michael W Mahoney. Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression. URL https://arxiv.org/pdf/1210.3135.pdf.

[8] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. 11 2012. URL http://arxiv.org/abs/1211.1002.

[9] David P Woodruff. Sketching as a Tool for Numerical Linear Algebra *. *Theoretical Computer Science " series*, 10:1–2, 2014. URL https://arxiv.org/pdf/1411.4357.pdf.