Note: These lecture notes are closely based on lecture notes by Sanjeev Arora [1] and Matt Weinberg [5].

1 Curse and Blessing of Dimensionality

High dimensional vectors are common in data mining and machine learning (e.g. items purchased by a Amazon customer, gene expression data). The phrase "curse of dimensionality" refers to the fact that algorithms are frequently harder to design in high-dimensional space - we've seen this with the convex hull algorithm. But, there is sometimes a flip side called "blessing of dimensionality", wherein high-dimensional spaces can sometimes make life easier to analyze. For example, we can pack vectors more tightly in high-dimensional space, it is easier to route around obstacles there, and many random samples are more likely to be tightly clustered around a mean (e.g. via Chernoff bounds).

The fact is that high dimensional spaces behave differently than our intuition suggests (living as we are in 3-dimensional space). Following are some examples, but first some notation.

For a vector $x \in \mathbb{R}^d$, its ℓ_2-norm is $|x|_2 = (\sum_i x_i^2)^{1/2}$ and ℓ_1-norm is $|x|_1 = (\sum_i |x_i|)$. For any two vectors x, y, their Euclidean distance is $|x - y|_2$ and their Manhattan distance is $|x - y|_1$.

Some generalizations of geometric objects to higher dimensions:

- The unit *d-cube* in $\mathbb{R}^d$: $\{(x_1, \ldots x_d : 0 \leq x_i \leq 1\}$. In $\mathbb{R}^4$, if you are looking at one of the faces, say where $x_1 = 1$, then you are looking at a cube in $\mathbb{R}^3$. The volume of the n-cube is 1.
- The unit *d-ball* in $\mathbb{R}^d$: $B_d = \{(x_1, \ldots x_d : \sum_i x_i^2 \leq 1\}$. In $\mathbb{R}^4$, if you slice through it with a hyperplane, say $x_1 = 1/2$, then this slice is a ball in $\mathbb{R}^3$ with radius of $\sqrt{1 - 1/2^2} = \sqrt{3}/2$. Every parallel slice also gives a ball. The volume of B_d is $\frac{\pi^{d/2}}{(d/2)!}$ (assuming d even). This is $\frac{1}{d^{\Theta(d)}}$

1.1 High Dimensionality Weirdness

1.1.1 Unit Ball

What is the ratio of the unit ball to its circumscribing cube (cube of side length 2)? In $\mathbb{R}^2$, it is $\pi/4$ or about .78. In $\mathbb{R}^3$ it is $\pi/6$ or about .52. In d dimensions, it is $\frac{1}{d^{\Theta(d)}}/2^d = d^{-cd}$ for some constant $c > 0$.

1.2 Near Orthogonal Vectors

How many "almost orthogonal" unit vectors can we have such that all pairwise angles lie between say 89 and 91 degrees? In $\mathbb{R}^2$, the answer is 2. In $\mathbb{R}^3$, it is 3. In $\mathbb{R}^d$, it is e^{cd} for some constant $c > 0$. Intuitively, to see this note that to get the angle close to 90, we just need to get the dot product of all vector pairs "close" to 0. When there are many entries in the vector, this is much easier to do. (more on this later).

2 Some Probability

Some tools from probability will be surprisingly useful for us to both get intuition about high dimensional geometry and also to do our projections to lower dimensional spaces. To start recall that a random variable (rv), X is informally a variable whose value depends on the outcome of some random phenomena. Typically, random variables have a finite number of possible values in the real numbers, and we let X also refer to the set of possible outcomes. In this case, the expectation of a random variable, $E(X)$, is defined as $E(X) = \sum_{x \in X} xPr(X = x)$.

First we prove linearity of expectation. Note that in the following lemma and proof, the random variables do *not* need to be independent. This makes the result extremely powerful.

Lemma 1. *(Linearity of Expectation) Given a set of random variables $X_1, \dots X_n$, $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$.*

Proof: We first prove this for two random variables X and Y.

$$\begin{aligned} E(X+Y) &= \sum_{x \in X} \sum_{y \in Y} (x+y) Pr(X = x, Y = y) \\ &= \sum_{x \in X} \sum_{y \in Y} x \cdot Pr(X = x, Y = y) + \sum_{y \in Y} \sum_{x \in X} y \cdot Pr(X = x, Y = y) \\ &= \sum_{x \in X} x \cdot Pr(X = x) + \sum_{y \in Y} y \cdot Pr(Y = y) \\ &= E(X) + E(Y) \end{aligned}$$

The general result for n random variables now follows by induction. □

Lemma 2. *(Markov's Inequality) Let X be a random variable that only takes on nonnegative values (i.e. $X \geq 0$ always). Then for any $\lambda > 0$,*

$$Pr(X \geq \lambda) \leq \frac{E(X)}{\lambda}.$$

Proof: Assume not. Then for some value $\lambda > 0$, $Pr(X \geq \lambda) > \frac{E(X)}{\lambda}$. If this is true, then the expected value of X can be bounded as:

$$\begin{aligned} E(X) &\geq \sum_{i \geq \lambda} iPr(X = i) \\ &\geq \sum_{i \geq \lambda} \lambda Pr(X = i) \\ &= \lambda Pr(X \geq \lambda) \\ &> \lambda \frac{E(X)}{\lambda} \\ &= E(X) \end{aligned}$$

But this sequence of inequalities implies that $E(X) > E(X)$, which is clearly a contradiction. □

2.1 Chernoff Bounds

The following important bound only works for independent random variables. We prove it for 0/1-valued random variables, which only take on the values 0 or 1, and we prove an upper bound. The lemma generalizes easily to also bound the probability of deviation below the mean.

Lemma 3. *(Chernoff bounds) Let $X_1, \ldots, X_n$ be independent 0/1-valued random variables and let $p_i = E(X_i)$, where $0 \leq p_i < 1$ for all i. Then the sum $X = \sum_i X_i$, which has mean $\mu = E(X) = \sum_i p_i$ satisfies*

$$Pr(X \geq (1+\delta)\mu) \leq (c_\delta)^\mu,$$

where $c_\delta = \frac{e^\delta}{(1+\delta)^{1+\delta}}$.

Proof: Consider an arbitrary positive constant t, to be set later, and consider the random variable e^{tX}. (If $X = 2$, say, this rv is e^{2t}.). A nice property of this random variable is the following:

$$\begin{aligned} E(e^{tX}) &= E\left(e^{t\sum_i X_i}\right) \\ &= E\left(\prod_i e^{tX_i}\right) \\ &= \prod_i E(e^{tX_i}) \end{aligned}$$

The last inequality holds since the X_i random variables are independent, and hence so are the e^{tX_i} random variables; and since $E(XY) = E(X)E(Y)$ if X and Y are independent. Note that

$$E(e^{tX_i}) = (1-p_i) + p_i e^t.$$

Thus, we have:

$$\begin{aligned} \prod_i E(e^{tX_i}) &= \prod_i [1 + p_i(e^t - 1)] \\ &\leq \prod_i e^{p_i(e^t-1)} \\ &\leq e^{\mu(e^t-1)} \end{aligned}$$

In the above, the second step holds by the inequality $1 + x \leq e^x$ (via Taylor expansion of e. Recall that $e^x = 1 + x + x^2/2! + x^3/3! + \ldots$). Now, we apply Markov's inequality to the random e^{tX} to get:

$$\begin{aligned} Pr(X \geq (1+\delta)\mu) &= Pr(e^{tX} \geq e^{t(1+\delta)\mu}) \\ &\leq \frac{e^{\mu(e^t-1)}}{e^{t(1+\delta)\mu}} \\ &\leq e^{\mu((e^t-1)-t(1+\delta))} \end{aligned}$$

Recall that Markov's inequality says that for any positive random variable Y, and any $\lambda > 0$,

$$Pr(Y \geq \lambda) \leq E(Y)/\lambda.$$

We let $Y = e^{tX}$, and note that $E(Y) \leq e^{\mu(e^t-1)}$; and we let $\lambda = e^{t(1+\delta)\mu}$.

This holds for any positive t, and is minimized when $t = \ln(1+\delta)$ (to see this, differentiate to get the minimum). This gives the lemma statement. □

Using a symmetric argument, we can bound the probability of deviation below the mean. Combining the results and using some approximations gives the following extremely useful lemma.

Lemma 4. *Let $X_1, \ldots X_n$ be independent Poisson trials such that $P(X_i = 1) = p_i$. Let $X = \sum_i X_i$ and $\mu = E(X)$. Then for $0 \leq \delta \leq 1$,*

$$Pr(|X - \mu| \leq \delta\mu) \leq 2e^{-\mu\delta^2/3}$$

Another concentration bound that is either called Chernoff or Hoeffding Bound, which is proven in [4]

Lemma 5. *[*Hoeffding bound*] Suppose X_i are independent random variables and $\ell_i \leq X_i \leq h_i$ for all $i \in [n]$. Then for all $t > 0$,*

$$Pr(|X - E(X)| > t) \leq 2e^{-\frac{2t^2}{\sum_i (h_i - \ell_i)^2}}$$

2.1.1 Using Chernoff Bounds

Assume we flip a fair coin n times and let X be the number of heads. Note that $E(X) = n/2$. Then by Chernoff bounds, we have that:

$$Pr(|X - n/2| \leq \delta n/2) \leq 2e^{-n\delta^2/6}$$

Q: What is the smallest value of δ that still ensures that we have polynomially small probability?
A: To ensure this, need $2e^{-n\delta^2/6} \leq n^{-1}$, which means that $-n\delta^2/6 \leq -\ln n$.
How about $\delta = 1$: we get $-n1/6 \leq -\ln n$ which works
How about $\delta = 1/\sqrt{n}$: we get $-n(1/n)/6 = \Theta(1)$
How about $\delta = \sqrt{(\ln n)/n}$: we get $-n(\ln n)/n/6 = \Theta(-\ln n)$. That works!

2.1.2 Bernstein's Inequality

A related inequality, which is even more closely like the central limit theorem, is the Bernstein inequality (see [2]). Below is one version of it.

Theorem 1. *Let $X_1, \ldots X_n$ be independent random variables with $E(X_i) = 0$ and $|X_i| \leq 1$ for all i. Let $X = \sum_i X_i$, $\sigma_i^2 = E(X_i^2) - (E(X_i))^2$ and $\sigma^2 = \sum_i \sigma_i^2$. Then for all $0 \leq k \leq \sigma$ we have:*

$$Pr(|X| \geq k\sigma) \leq 2e^{-k^2/4}$$

Proof: We'll bound the $X \geq k\sigma$ direction, the theorem follows by using a symmetric proof to bound the $X \leq k\sigma$ direction and then using a union bound.

First note by Markov's inequality that:

$$\begin{aligned} Pr(X \geq k\sigma) &= Pr(e^{\lambda X} \geq e^{\lambda k\sigma}) \\ &\leq \frac{E(e^{\lambda X})}{e^{\lambda k\sigma}}. \end{aligned}$$

So how big is $E(e^{tX})$?

$$\begin{aligned}
E(e^{\lambda X}) &= \prod_i E(e^{\lambda X_i}) \\
&\leq \prod_i E(1 + \lambda X_i + (\lambda X_i)^2) \\
&= \prod_i \left(1 + (\lambda \sigma_i)^2\right) \\
&\leq e^{\sum_i (\lambda \sigma_i)^2} \\
&= e^{(\lambda \sigma)^2}
\end{aligned}$$

The first step holds by independency of the X_i. The second step holds by the fact that $e^{\lambda X_i} \leq 1 + \lambda X_i + (\lambda X_i)^2$ holds for $0 < |\lambda X_i| < 1$ by the Taylor expansion of $e^{\lambda X_i}$. The third step holds since $E(X_i) = 0$, and $E(X_i^2) = \sigma_i^2$. The fourth step holds since $1 + x \leq e^x$ for all x. The final step holds by definition of σ.

Now, we can plug this back into Markov's inequality to get that:

$$\begin{aligned}
Pr(X \geq k\sigma) &= Pr(e^{\lambda X} \geq e^{\lambda k \sigma}) \\
&\leq \frac{e^{(\lambda\sigma)^2}}{e^{\lambda k \sigma}} \\
&= e^{t\sigma(\lambda\sigma - k)} \\
&\leq e^{-k^2/4}
\end{aligned}$$

The last step holds by choosing $\lambda = k/(2\sigma)$, which minimizes the right hand side over all values of λ. Then, the constraint that $0 < |\lambda X_i| < 1$ always holds if $0 \leq k \leq \sigma$.

□

Below is a tighter version of Bernstein's. It's proof is similar to the one above.

Theorem 2. *(Bernstein Inequality) Let $X_1, \ldots X_n$ be independent random variables with $|X_i - E(X_i)| \leq b$ for each $i \in [n]$. Let $X = \sum_i X_i$, and $\sigma^2 = \sum_i \sigma_i^2$ be the variance of X. Then for any $t > 0$,*

$$Pr(|X - E(X)| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2(1+bt/3\sigma^2)}}$$

2.1.3 Geometry of Chernoff Bounds

The following discussion about connections between Chernoff bounds and geometry is from Kelner [3]. Following is a simple generalization of Chernoff bounds.

Theorem 3. *Let $x_i \in \{\pm 1\}^n$ be independent rv's with $Pr(x_i = 1) = .5$, and let x be a vector with entries equal to the x_i. Let a be any unit vector. Then for all $0 < \lambda < \sqrt{n}$.*

$$Pr\left(\left|\sum_{i=1}^{n} a \cdot x\right| > \lambda\right) \leq 2e^{-\lambda^2/2}$$

If we let a be the unit vector where all entries have weight $1/\sqrt{n}$, this is exactly the bound we just proved. It's a useful exercise to reprove the bound for any arbitrary unit vector a.

Next, we can change things so that $x_i \in [-1/2, +1/2]$ and they are uniformly and independently distributed in that range. Then the bound above still holds up to some constants.

What does this mean geometrically? First, note the following:

Fact 1. *$a \cdot x$ is the distance from x to the hyperplane $H_a = \{x : a \cdot x = 0\}$*

Now we can think of Chernoff bounds in the following way. First, pick any unit vector, a and let H_a be the hyperplane orthogonal to that vector. Next, pick any point x in the unit hypercube. Then, by Theorem 3 and Fact 1, the probability that point x is "far" (i.e. distance $> \lambda$) from H_a is small (i.e. $2e^{-\lambda^2/2}$).

In particular, if S is the set of all points within distance λ of H_a, and C is the entire hypercube, then, we have

$$\frac{Vol(S)}{Vol(C)} \geq 1 - 2e^{-6\lambda^2}.$$

Hence, for any hyperplane that cuts through the origin, almost all of the hypercube is "close" to that plane! (Note that the constant in the exponent improved since the points are in $[-1/2, +1/2]$ (unit hypercube) instead of $[-1, +1]$ (as in Berstein's).)

2.2 Union Bounds

The following tool is frequently useful in conjunction with Chernoff bounds.

Lemma 6. *(Union Bounds) Consider n events $\xi_1, \ldots \xi_n$. Then we have that*

$$Pr(\cup_i \xi_i) \leq \sum_{i=1}^{n} Pr(\xi_i)$$

Proof: We'll show this for two events, the lemma statement then holds by an inductive argument. Let ξ_1 and ξ_2 be any two events. Then we have that

$$\begin{aligned} Pr(\xi_1 \cup \xi_2) &= Pr(\xi_1) + Pr(\xi_2) - Pr(\xi_1 \cap \xi_2) \\ &\leq Pr(\xi_1) + Pr(\xi_2) \end{aligned}$$

□

3 Number of Almost Orthogonal Vectors

One of the *benefits* of high-dimensional spaces are that they are very "roomy". For example, we now show that there are $\Theta(e^d)$ vectors in $\mathbb{R}^d$ that are "almost" orthogonal. Recall that the angle, θ, between two vectors can be found via the identity $\cos(\theta) = \frac{x \cdot y}{|x||y|}$, where $|\cdot|$ is the 2-norm.

Lemma 7. *Let a be a unit vector in $\mathbb{R}^n$. Let $x = (x_1, \ldots x_n)$ be a unit vector in $\mathbb{R}^n$ created by choosing each x_i independently and uniformly in $\{\frac{-1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}$. Let $X = a \cdot x = \sum_i a_i x_i$. Then for all $0 \leq t \leq 1$,*

$$Pr(|X| > t) < 2e^{-nt^2/2}.$$

Proof: Note that $E(X) = E(\sum_i a_i x_i) = 0$. This is true since $E(a_i x_i) = \frac{1}{\sqrt{n}} a_i - \frac{1}{\sqrt{n}} a_i = 0$.

We will use the Hoeffding bound so we must compute $\sum_i (h_i - \ell_i)^2$, where ℓ_i and h_i are upper and lower bounds for each summand $a_i x_i$. Note that $-a_i/\sqrt{n} \leq a_i x_i \leq a_i/\sqrt{n}$, and so letting $\ell_i = -a_i/\sqrt{n}$ and $h_i = a_i/\sqrt{n}$, we have that

$$\begin{aligned}\sum_{i=1}^{n}(h_i - \ell_i)^2 &= \sum_{i=1}^{n}((a_i/\sqrt{n} + a_i/\sqrt{n}))^2 \\ &= 4/n \sum_{i=1}^{n} a_i^2 \\ &= 4/n\end{aligned}$$

Where the last step holds since a is a unit vector. Thus, the Hoeffding bound(Lemma 5) gives that:

$$\begin{aligned}Pr(|X| > t) &< 2e^{-\frac{2t^2}{\sum_i(\ell_i - h_i)^2}} \\ &< 2e^{-t^2 n/2}\end{aligned}$$

□

From the above, the dot product of any unit vector $x \in \mathbb{R}^n$ with a "randomly chosen" vector is "small" with high probability. Since the cosine of two unit vectors x and y equals $x \cdot y$, we have the following:

Lemma 8. *Let $\epsilon > 0$ be a fixed constant. Consider a set S of $e^{\epsilon^2 n/10}$ vectors in $\mathbb{R}^n$, where each entry is independently and uniformly chosen in $\{\frac{-1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}$. For any pair of vectors $x, y \in S$, let $\theta_{x,y}$ be the angle between x and y. Then for all $x, y \in S$,*

$$Pr(|\cos\theta_{x,y}| > \epsilon) \leq e^{-\epsilon^2 n/5}$$

Proof: Consider some fixed pair of vectors $x, y \in S$. Let $\xi_{x,y}$ be the event that $x \cdot y > \epsilon$. Note that $Pr(|\cos\theta_{x,y}| > \epsilon) = Pr(|x \cdot y| > \epsilon)$ Thus, by Lemma 7,

$$Pr(|\cos\theta_{x,y}| > \epsilon) < 2e^{-\epsilon^2 n/2}$$

Now let ξ be the event that *any* pair of vertices violates the bound. In particular, $\xi = \cup_{x,y \in S}\, \xi_{x,y}$. Then by a Union bound, we have:

$$\begin{aligned}Pr(\xi) &\leq \sum_{x,y \in S, x \neq y} Pr(\xi_{x,y}) \\ &\leq |S|^2 2e^{-\epsilon^2 n/2} \\ &\leq 2e^{\epsilon^2 n/5} e^{-\epsilon^2 n/2} \\ &\leq 2e^{-\epsilon^2 n/4} \\ &\leq e^{-\epsilon^2 n/5}\end{aligned}$$

where the last step holds for n sufficiently large. □

4 Dimension Reduction

In a typical dimension reduction problem, we're given n points $v_1, \dots v_n \in \mathbb{R}^d$ and a fixed $\epsilon > 0$. We want to find a function $f : \mathbb{R}^d \to \mathbb{R}^m$, where $m << d$ such that for all i and j:

$$|v_i - v_j| \leq |f(v_i) - f(v_j)| \leq (1 + \epsilon)|v_i - v_j|$$

In other words, the distances between points are (approximately) preserved.

Note that many naive ideas fail to achieve this such as: (1) taking a random sample of m coordinates out of d; and (2) partitioning coordinates into m subsets and add up the values in each subset.

Idea 1 fails for the case where we have vector $x = (0, 0, \dots, 1)$ and $y = (1, 0, 0 \dots, 0)$. Note that $|x - y| = 1$, but any random sample of coordinates is unlikely to find the 1 entry in either of these vectors. Idea 2 fails for the case that $x = (0, 1, 0, 1, \dots)$ and $y = (1, 0, 1, 0, \dots)$. Note that $|x - y|$ is large but these sums would be very close.

4.1 Johnson-Lindenstrauss Projection

Let G be a m by d matrix where each entry is a Normal random variable, i.e. $G_{i,j} \sim \mathcal{N}(0, 1)$. Let $\Pi = \frac{1}{\sqrt{m}} G$ and let

$$f(x) = \Pi x.$$

So each entry in $f(v)$ equals $v \cdot g$ for some vector g filled with scaled Normal random variables (note that Gaussian and Normal are synonmous). Other (simpler) approaches also work (See Section **??** below).

4.2 Analysis

4.3 Reduction to Norm Preservation

Our main lemma is below. Note that, by taking square roots, this theorem implies that

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

(For example, by Theorem 4 we have that:

$$\sqrt{(1 - \epsilon)}|x| \leq |\Pi x|$$

This implies that

$$(1 - \epsilon)|x| \leq \sqrt{(1 - \epsilon)}|x| \leq |\Pi x|$$

Distance Preservation: Then to prove distance preservation, we note that by the linearity of $f = \Pi$,

$$|f(x) - f(y)| = |\Pi x - \Pi y| = |\Pi(x - y)|$$

So with probability $1 - \delta$, we preserve the distance of one pair by Theorem 4. Then we'll do a union bound over all pairs, which will increase the error probability by $\binom{n}{2}$.

4.4 Main Theorem

Theorem 4. *(The (ϵ, δ)-JL property) If $m = 9\log(1/\delta)/\epsilon^2$ then, with probability $1-\delta$, for any vector x,*

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

Proof: Let $w = \Pi x$. Then we have:

$$|w|^2 = |\Pi x|^2 = |\frac{1}{\sqrt{m}} Gx|^2 = \frac{1}{m}\sum_{i=1}^{m} w_i^2.$$

In the above, we define w_i as:

$$w_i = \sum_{j=1}^{d} x_j g_j$$

where each $g_j \sim \mathcal{N}(0,1)$. So $E(w_i) = \sum_{j=1}^{d} x_j E(g_j) = 0$. Recall that $\mathrm{Var}(X) = E(X^2) - E^2(X)$. Thus $\mathrm{Var}(w_i) = E(w_i^2)$, and so:

$$\mathrm{Var}(w_i) = E(w_i^2) = \sum_{j=1}^{d} \mathrm{Var}(x_j g_j) = \sum_{j=1}^{d} x_j^2 \mathrm{Var}(g_j) = \sum_{j=1}^{d} x_j^2 = |x|^2.$$

The above follows since for independent random variables X and Y, $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$. So now we have that:

$$E(|w|^2) = E\left(\frac{1}{m}\sum_{i=1}^{m} w_i^2\right) = \frac{1}{m}\sum_{i=1}^{m} E(w_i^2) = \frac{1}{m}\sum_{i=1}^{m} |x|^2 = |x|^2$$

So at least things work in expectation. What about concentration? To address this, we make use of the following fact about normal random variables:

Fact 1: If X and Y are independent and $X \sim \mathcal{N}(0, a^2)$ and $Y \sim \mathcal{N}(0, b^2)$, then $X + Y \sim \mathcal{N}(0, a^2 + b^2)$. The property that the sum of Normal distributions remains normal is known as *stability.*

By this fact, $w_i \sim \mathcal{N}(0, |x|^2)$. It follows that w_i^2 is a χ^2 *(chi-squared)* random variable, and that $|w|^2 = \frac{1}{m}\sum_{i=1}^{m} w_i^2$ is a chi-squared random variable with m degrees of freedom. These random variables are very well studied and they concentrate around their mean essentially as well as a Normal random variable. In particular, if $X = \frac{1}{m}\sum_{i=1}^{m} w_i^2$, then we have the following[1] for any positive ϵ:

$$P(|X - E(X)| \geq \epsilon E(X)) \leq 2e^{-m\epsilon^2/8}.$$

So if we set $m = 9\log(1/\delta)/\epsilon^2$, then we get:

$$\begin{aligned} P(|X - E(X)| \geq \epsilon E(x)) &\leq 2e^{-(9\log(1/\delta)/\epsilon^2)(\epsilon^2/8)} \\ &= 2e^{-((9/8)\log(1/\delta))} \\ &= 2(\delta)^{9/8} \\ &\leq \delta \end{aligned}$$

The last step above holds when $2\delta^{9/8} \leq \delta$, or $\delta < 1/256$. □

[1] See, e.g., https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf

Now we can prove the main theorem.

Theorem 5. *Assume we are given n points $v_1, \dots v_n \in \mathbb{R}^d$ and a fixed $\epsilon > 0$. Let $m = O(\log n/\epsilon^2)$ and set $f = \Pi$, where Π is a m by d matrix of independent $\mathcal{N}(0,1)$ random variables. Then, with probability $1 - 1/n$, for any i and j, $1 \le i < j \le n$:*

$$(1-\epsilon)|v_i - v_j| \le |\Pi v_i - \Pi v_j)| \le (1+\epsilon)|v_i - v_j|$$

Proof: Set $\delta = 1/n^3$ and $m = 27 \log n/\epsilon^2$. For any fixed pair of points v_i and v_j, let $\xi_{i,j}$ be the (bad) event that the following does not hold:

$$|v_i - v_j| \le |\Pi(v_i) - \Pi(v_j)| \le (1+\epsilon)|v_i - v_j|$$

By Theorem 4, $Pr(\xi_{i,j}) \le 1/n^3$. Let ξ be the (bad) event that $\xi_{i,j}$ occurs for any v_i and v_j. Then, by a Union bound, we know that

$$\begin{aligned} Pr(\xi) &\le \sum_{i,j} Pr(\xi_{i,j}) \\ &= \binom{n}{2} \frac{1}{n^3} \\ &\le 1/n. \end{aligned}$$

□

Interestingly, this bound is tight. Noga Alon has shown (in 2018) that there are point sets that can't be embedded in less than $O(\log n/\epsilon^2)$ dimensions if we want to preserve pairwise distances.

References

[1] Sanjeev Arora. Advanced Algorithm Design Class, Princeton University, 2013. https://www.cs.princeton.edu/courses/archive/fall15/cos521/.

[2] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[3] Jonathan Kelner. An Algorithmist's Toolkit, 2009. https://ocw.mit.edu/courses/mathematics/18-409-topics-in-theoretical-computer-science-an-algorithmists-toolkit-fall-2009/lecture-notes/MIT18_409F09_scribe15.pdf.

[4] Lafferty, Liu, and Wasserman. Concentration of Measure, 2010. https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf.

[5] Matt Weinberg. Dimensionality Reduction and the Johnson-Lindenstrauss Lemma, 2019. https://www.cs.princeton.edu/~smattw/Teaching/Fa19Lectures/lec9/lec9.pdf.