Note: These lecture notes are closely based on lecture notes by Sanjeev Arora [1] and Matt Weinberg [3].

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 *n-cube* in $\mathbb{R}^d$: $\{(x_1, \dots x_d : 0 \leq x_i \leq 1\}$. 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 *n-ball* in $\mathbb{R}^d$: $B_d = \{(x_1, \dots x_d : \sum_i x_i^2 \leq 1\}$. 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 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.

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 $c > 0$.

2 Some Probability

Some tools from probability will be surprisingly useful 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 $X_1, \ldots X_n$, $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$.*

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:

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

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

A related inequality, which is related to the central limit theorem, is the Bernstein inequality below. (From https://www.cs.princeton.edu/~smattw/Teaching/Fa19Lectures/lec3/lec3.pdf and http://cseweb.ucsd.edu/~klevchen/techniques/chernoff.pdf. The following is closely based on the original proof due to Van Vu at UCSD, local copy available at https://www.cs.unm.edu/~saia/classes/506-s20/lec/bernstein.pdf)

Theorem 1. *(Bernstein Inequality) Let $X_1, \ldots X_n$ be discrete, independent random variables with $E(X_i) = 0$ and $|X_i| \leq 1$ for all $i \in [1, n]$. Let $X = \sum_{i \in [1,n]} X_i$, $\sigma_i^2 = E(X_i^2) - (E(X_i))^2$ and $\sigma^2 = \sum_{i \in [1,n]} \sigma_i^2$. Then, $\forall \lambda \in [0, \sigma/2]$:*

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

Equivalently, $\forall t \in [0, \sigma^2/2]$*:*

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

Proof: We'll show

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

The argument is symmetric for $Pr(-X \geq \lambda\sigma)$. Let t be a real number between 0 and 1 that we'll tune later. Then:

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

where the last step follows by Markov's inequality.

Next we bound $E(e^{tZ})$ for $-1 \leq Z \leq 1$, and $E(Z) = 0$, and $t \leq 1$. In the following, let z_j for $j \in [1, m]$ be the values in the sample space for Z and let $p_j = Pr(Z = z_j)$. By definition of expectation and using the Taylor expansion of e^x,

$$\begin{aligned} E(e^{tZ}) &= \sum_{j \in [1,m]} p_j e^{tz_j} \\ &= \sum_{j \in [1,m]} p_j \left(1 + tz_j + \frac{1}{2!}(tz_j)^2 + \frac{1}{3!}(tz_j)^3 + \dots\right) \\ &= \sum_{j \in [1,m]} p_j + t \sum_{j \in [1,m]} p_j z_j + \sum_{j \in [1,m]} p_j \left(\frac{1}{2!}(tz_j)^2 + \frac{1}{3!}(tz_j)^3 + \dots\right) \\ &= 1 + E(Z) + \sum_{j \in [1,m]} p_j \left(\frac{1}{2!}(tz_j)^2 + \frac{1}{3!}(tz_j)^3 + \dots\right) \\ &\leq 1 + \sum_{j \in [1,m]} p_j (tz_j)^2 \left(\sum_{i \in [2,\infty]} \frac{1}{i!}\right) \\ &\leq 1 + \sum_{j \in [1,m]} p_j (tz_j)^2 \\ &\leq 1 + t^2 \text{Var}(Z) \end{aligned}$$

Above, the fifth line holds since $E(Z) = 0$. Returning to our claim:

$$\begin{aligned}
E(e^{tX}) &= E(e^{t\sum_{i\in[1,n]} X_i}) \\
&= E(\prod_{i\in[1,n]} e^{tX_i}) \\
&= \prod_{i\in[1,n]} E(e^{tX_i}) && \text{Independence of } X_i \\
&= \prod_{i\in[1,n]} 1 + t^2\text{Var}(X_i) \\
&\le \prod_{i\in[1,n]} e^{t^2\text{Var}(X_i)} && \text{Since } 1+\alpha \le e^{\alpha} \text{ for } \alpha > 0 \\
&\le e^{t^2\sigma^2} && \text{By independence of } X_i
\end{aligned}$$

Plugging this back into our initial bound, we get

$$\begin{aligned}
Pr(X \ge \lambda\sigma) &\le \frac{E(e^{tX})}{e^{t\lambda\sigma}} \\
&\le \frac{e^{t^2\sigma^2}}{e^{t\lambda\sigma}} \\
&\le e^{t\sigma(t\sigma-\lambda)} \\
&\le e^{-\lambda^2/4} && \text{Optimizing } t \text{ to } t = \lambda/(2\sigma)
\end{aligned}$$

□

What does this tell us about high-dimensional geometry? Let X be the sum of the coordinates of a point chosen uniformly at random on a n dimensional hypercube of length 1 that touches the origin and where the coordinates of all vertices are non-negative. In particular, each coordinate of the random point is chosen independently and uniformly in the range $[0, 1]$. We can use Bernstein's inequality on the sum of the differences between each coordinate and the value $1/2$. Then $\sigma^2 = n/6$, and Bernstein's inequality tells us that the sum of all coordinates will be $n/2 + O(\sqrt{n}\lg n)$ with high probability, by setting $\lambda = \log n$.

What does this mean? If we think in terms of L_1 distances, it means that almost all of the volume of the hypercube is in a ball centered at the origin with radius $n/2 + O(\sqrt{n \lg n})$. This is true even though a vertex of the hypercube is at distance of $\Theta(n)$ from the origin. We can get a similar result in terms of L_2 distances: almost all of the volume of the cube is in a ball centered at the origin with radius $(1 + o(1))\sqrt{n/2}$.

2.1 Union Bounds

The following tool is simple to prove but surprisingly useful.

Lemma 3. *(Union Bounds) Consider n events $\xi_1, \ldots \xi_n$. Then we have that*

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

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, ϕ, between two vectors can be found via the identity $\cos(\phi) = \frac{x \cdot y}{|x||y|}$, where $|\cdot|$ is the 2-norm.

Lemma 4. *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 $X = a \cdot x = \sum_{i \in [1,n]} a_i x_i$. Then for all $\lambda > 0$,*

$$Pr(|X| > \lambda) < 2e^{-n\lambda^2/4}.$$

Proof: Note that $E(X) = E(\sum_{i \in [1,n]} a_i x_i) = 0$. This is true since $E(a_i x_i) = \frac{1}{2}(a_i \frac{-1}{\sqrt{n}}) + \frac{1}{2}(a_i \frac{1}{\sqrt{n}}) = 0$. Since $\sigma^2 = E(X^2) - (E(X))^2 = E(X^2)$, we have:

$$\begin{aligned} \sigma^2 &= E\left(\left(\sum_{i=1}^{n} a_i x_i\right)^2\right) \\ &= E\left(\sum_{1 \leq i,j \leq n} a_i a_j x_i x_j\right) \\ &= \sum_{1 \leq i,j \leq n} a_i a_j E(x_i x_j) \\ &= \sum_{1 \leq i \leq n} a_i^2 E(x_i^2) + \sum_{1 \leq i \neq j \leq n} a_i a_j E(x_i x_j) \\ &= \sum_{1 \leq i \leq n} a_i^2 (1/n) \\ &= 1/n. \end{aligned}$$

For the second to last step, note that if $i \neq j$, $E(x_i x_j) = \frac{1}{2} \cdot \frac{1}{n} + \frac{1}{2} \cdot \frac{-1}{n} = 0$, and if $i = j$, $E(x_i^2) = 1/n$. Thus, using Bernstein's inequality, we get:

$$Pr(|X| > t) < 2e^{-(t/\sigma)^2/4} \leq 2e^{-nt^2/4}$$

□

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 5. *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 $\phi_{x,y}$ be the angle between x and y. Then for all $x, y \in S$,*

$$Pr(|\cos\phi_{x,y}| > \epsilon) \leq e^{-\epsilon^2 n/21}$$

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 4,

$$Pr(|\cos\phi_{x,y}| > \epsilon) < 2e^{-\epsilon^2 n/4}$$

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:

$$\begin{aligned} Pr(\xi) &\leq \sum_{x \neq y \in S} Pr(\xi_{x,y}) \\ &\leq |S|^2 2e^{-\epsilon^2 n/4} \\ &\leq 2e^{\epsilon^2 n/5} e^{-\epsilon^2 n/4} \\ &\leq 2e^{-\epsilon^2 n/20} \\ &\leq e^{-\epsilon^2 n/21} \end{aligned}$$

where the last step holds for n sufficiently large. □

4 Dimension Reduction

We're given n points $v_1, \ldots v_n \in \mathbb{R}^d$ 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:

$$|f(v_i) - f(v_j)| \in (1 \pm \epsilon)|v_i - v_j|$$

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) partition coordinates into m subsets and add up the values in each subset.

Idea 1 fails for the case where we have vector $x = (0, 0, \ldots, 1)$ and $y = (1, 0, 0 \ldots, 0)$. 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 $x = (0, 1, 0, 1, \ldots)$ and $y = (1, 0, 1, 0, \ldots)$. 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 4.4 below).

4.2 Reduction to Norm Preservation

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

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

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

4.3 JL Theorem

Theorem 2. *Let x be any fixed vector in $\mathbb{R}^n$, $0 < \delta \leq 1/256$, and $\epsilon > 0$. Then, for $m = 9\ln(1/\delta)/\epsilon^2$, with probability $1 - \delta$:*

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

Proof: Let $w = \Pi x$. Then,

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

where

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

and each $g_j \sim \mathcal{N}(0, 1)$.

So $E(w_i) = \sum_{j=1}^{d} x_j E(g_j) = 0$. Recall $var(X) = E(X^2) - E^2(X)$. Thus,

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

The above follows since for independent random variables X and Y, $\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)$. Thus,

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

Now 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[1] . In particular, if $X = \frac{1}{m}\sum_{i=1}^{m} w_i^2$, then for any positive ϵ: $P(|X - E(X)| \geq \epsilon E(X)) \leq 2e^{-m\epsilon^2/8}$. For us, that gives:

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

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

If we let $m = 9\log(1/\delta)/\epsilon^2$, we get:

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

where the last step holds for δ sufficiently small. In particular, we want $2\delta^{9/8} \leq \delta$. Dividing both sides by δ, we see that this holds when $2\delta^{1/8} \leq 1$ or $\delta^{1/8} \leq 1/2$ or $\delta \leq (1/2)^8$ or $\delta < 1/256$.

Hence, we have

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

Taking square roots of the above inequality, we have:

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

Assuming $\epsilon \in (0,1)$, we have $\sqrt{1-\epsilon} \leq 1-\epsilon$ and also $\sqrt{1+\epsilon} \leq 1+\epsilon$. Thus:

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

□

The main theorem now holds essentially by union bounds as follows.

Theorem 3. *Assume we are given n points $v_1, \ldots v_n \in \mathbb{R}^d$ and a fixed $\epsilon > 0$. Let $m = (27\log n)/\epsilon^2$ 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, $1 \leq i < j \leq n$:*

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

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

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

Then by Theorem 2, $Pr(\xi_{i,j}) \leq 1/n^3$. Let ξ be the event that $\xi_{i,j}$ occurs for any v_i and v_j.

Then, by a Union bound, we know that

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

□

Interestingly, this bound is tight. There are point sets that can't be embedded in less than $O(\log n/\epsilon^2)$ dimensions if we want to approximately preserve pairwise distances [2].

4.4 Simpler Johnson-Lidenstrauss

Here is a simpler Johnson-Lindenstrauss projection that also works.

1. $x_1, \ldots x_m \leftarrow$ vectors in $\mathbb{R}^m$ chosen as follows. Each coordinate is chosen independently and randomly from $\left\{\sqrt{\frac{1}{m}}, -\sqrt{\frac{1}{m}}\right\}$
2. $u_i[j] \leftarrow x_i \cdot u_i$ for all $i : 1 \leq i \leq n$ and $j : 1 \leq j \leq m$

In other words, $u_i = (z_i \cdot x_1, \ldots z_i \cdot x_m)$ for $i = 1, \ldots m$. Note that we can think of this as a linear transformation $u = Az$ where A is a matrix with random and independent entries in $\left\{\sqrt{\frac{1}{m}}, -\sqrt{\frac{1}{m}}\right\}$.

4.5 Analysis

We now do a "sketch" of the analysis. The following lemma shows that things work out well in expectation.

Lemma 6. *For any* $1 \leq i < j \leq n$, $E(|u_i - u_j|^2) = |z_i - z_j|^2$

Proof: According to the projection, we have the following for any $1 \leq i < j \leq n$:

$$|u_i - u_j|^2 = \sum_{k=1}^{m} \left(\sum_{\ell=1}^{n} (z_i[\ell] - z_j[\ell]) x_k[\ell] \right)^2$$

Fix i and j. Let $z = z_i - z_j$ and let $u = u_i - u_j$. Then for any $1 \leq k \leq m$, we have

$$\begin{aligned} E(|u \cdot x_k|^2) &= E\left(\left(\sum_{\ell=1}^{n} (z[\ell] x_k[\ell] \right)^2 \right) \\ &= \sum_{\ell} \sum_{\ell'} E\left(z[\ell] x_k[\ell] z[\ell'] x_k[\ell'] \right) \\ &= \sum_{\ell=1}^{n} E\left((z[\ell] x_k[\ell])^2 \right) \\ &= \frac{1}{m} |z|^2 \end{aligned}$$

Hence, by linearity of expectation $E(|u|^2)) = |z|^2$. □

The rest of the analysis follows similar to that in Theorem 3. First, one establishes a (harder) tail-bound around this expectation and then does a union bound over all pairs of points. In this way, we can get the same result as Theorem 3.

5 Applications of JL Projection

- Approximate all-pairs distances in $O(n^2 \log n + nd)$ vs $O(n^2 d)$ time
- Approximate distance-based clustering
- Approximate support vector machine (SVM) classification

- Approximate Linear Regression

Note: For some of these Machine Learning type applications, we need it to be the case that distances are approximately preserved across *all* (infinite) vectors in the vector space. Thus, a simple union bound won't work and instead we need to make use of a technique called *ϵ-nets*. We discuss this technique below.

6 Linear Regression and ϵ-Nets

The following is the classic least-squares regression problem.

Given: n data vectors $a_1, \dots a_n \in \mathbb{R}^d$, and n response values $y_1, \dots, y_n \in \mathbb{R}$. Let A be a $n \times d$ matrix with rows $a_1, \dots, a_n$; let y be a length n vector with entries $y_1, \dots, y_n$.

Goal: Find $x \in \mathbb{R}^d$ to minimize

$$\sum_{i=1}^{n} (a_i \cdot x - y_i)^2 = |Ax - y|^2$$

Usually, this problem requires $O(nd^2)$ time to solve (for example, by using singular value decomposition). We now show how to speed it up by reducing n using Johnson-Lidenstrauss.[2]

Let Π be chosen from the family of matrices from Theorem 3. To obtain an approximate solution, we solve the "sketched" problem where we find $x \in \mathbb{R}^d$ to minimize:

$$|\Pi(Ax - y)|^2.$$

This can be solved in $O(md^2)$ time (once ΠA and Πy are computed - we haven't discussed this but there are JL transforms which are also fast, since they are sparse). We want to prove that a solution to this smaller problem is a good approximation to the big problem. In particular, we'd like the following inequality to hold for any input vector x.

$$(1 - \epsilon)|Ax - y|^2 \leq |\Pi(Ax - y)|^2 \leq (1 + \epsilon)|Ax - y|^2 \tag{1}$$

In particular, let x^* be the optimal solution for the original problem, and let $\tilde{x}^*$ be the solution for the sketched problem. Then if equation 1 holds, we have:

$$|A\tilde{x}^* - y|^2 \leq \frac{1}{1-\epsilon}|\Pi(A\tilde{x}^* - y)|^2 \leq \frac{1}{1-\epsilon}|\Pi(Ax^* - y)|^2 \leq \frac{1+\epsilon}{1-\epsilon}|Ax^* - y|^2.$$

In the above the first and last inequalities hold via equation 1, and the middle inequality holds by noting that $\tilde{x}^*$ minimizes $|\Pi(Ax - y)|$ over all vectors x.

If $\epsilon \leq 1/4$, then $\frac{1+\epsilon}{1-\epsilon} \leq 1 + 3\epsilon$, so we can get an approximation to the original regression problem. Q: Why do we need a bound for all x above??? The main problem is that $\tilde{x}^*$ depends on the projection Π, and so it's not fixed ahead of time. How do we extend equation 1 to all x? We can't use union bounds since there are an infinite number of possible vectors x.

[2] Note that we are reducing n (number of vectors) and not d (dimension). Since we only care about the matrix A, you could think of n as the dimension and d as the number of vectors.

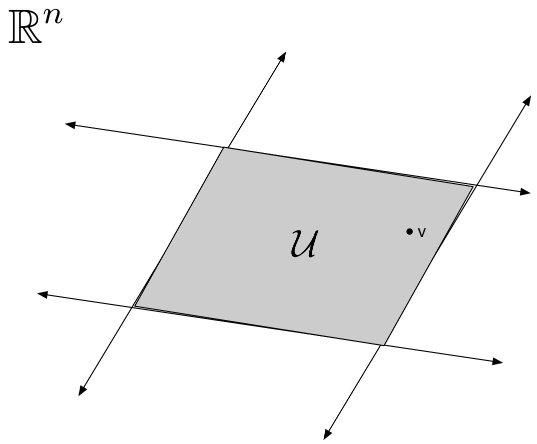

Figure 1. JL approximately preserves distances over any subspace $\mathcal{U}$ of dimension d contained in $\mathbb{R}^n$

7 Beyond Union Bounds

We need an approach that works for all vectors in certain subspace.

7.1 Subspace Embeddings

We will prove a more general statement that implies equation 1 for all the vectors we need, and is useful in other applications.

Theorem 4. *Let $\mathcal{U}$ be any d-dimensional linear subspace in $\mathbb{R}^n$, $0 < \delta \leq 1/256$, $0 < \epsilon \leq 2/5$, $m = (36d/\epsilon^2)\ln\left(\frac{8}{\delta\epsilon}\right)$, and Π be the matrix defined above for dimensions m by n. Then, with probability $1-\delta$, for all $v \in \mathcal{U}$:*

$$(1-\epsilon)|v| \leq |\Pi v| \leq (1+\epsilon)|v| \tag{2}$$

(Note that it's possible to prove a slightly tighter bound of $m = O(\frac{d+\log(1/\delta)}{\epsilon^2})$ that we won't discuss here.)

How does this theorem imply equation 1? We can apply it to the $d+1$ dimensional subspace spanned by the d columns of A and the vector y. Every vector formed by inputting some vector x into the linear equation $Ax - y$ lies in this $d+1$ dimensional subspace. In particular, we can approximately solve linear regression over $n >> d$ examples for the same amount of work as $O(d)$ examples, for fixed ϵ.

7.2 Reduction to a Sphere

We first note that Theorem 4 holds so long as equation 2 holds for all points on the unit sphere in $\mathcal{U}$. This is a consequence of linearity of the Euclidean norm. In particular, denote the sphere $\mathcal{S}_{\mathcal{U}}$ as

$$\mathcal{S}_{\mathcal{U}} = \{v \mid v \in \mathcal{U} \text{ and } |v| = 1\}.$$

Now any point $v \in \mathcal{U}$ can be written as cx for some scalar c and some point $x \in \mathcal{S}_{\mathcal{U}}$. Then, if $|\Pi x| \in (1 \pm \epsilon)|x|$, then $c|\Pi x| \in c(1 \pm \epsilon)|x|$ and so $|\Pi cx| \in (1 \pm \epsilon)|cx|$. The last inequality holds since $|cx| = \sqrt{\sum_i (cx)_i^2} = c\sqrt{\sum_i x_i^2} = c|x|$, since x was on the unit sphere.

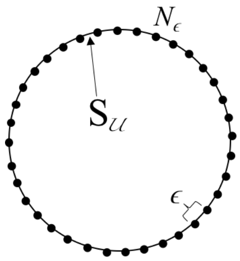

Figure 2. An ϵ-net N_ϵ for a sphere in a 2-dimensional subspace of $\mathcal{U}$

7.3 Constructing a Net

We prove Theorem 4 by showing that there is a large, but finite set of points $N_\epsilon \subset \mathcal{S}_{\mathcal{U}}$ such that if equation 2 holds for all $v \in N_\epsilon$, then it holds for all vectors v. The set N_ϵ is called an ϵ-net. In particular, we show:

Lemma 7. *For any positive $\epsilon \leq 2$, there exists a set $N_\epsilon \subset \mathcal{S}_{\mathcal{U}}$ with $|N_\epsilon| \leq \left(\frac{4}{\epsilon}\right)^d$ such that $\forall v \in \mathcal{S}_{\mathcal{U}}$,*

$$\min_{x \in N_\epsilon} |v - x| \leq \epsilon.$$

Proof: We use the following greedy procedure to construct N_ϵ; this construction is just for proof of existence, our algorithms do not need to implement this.

1. $N_\epsilon \leftarrow \{\}$
2. While there is a point $v \in \mathcal{S}_{\mathcal{U}}$ with distance greater than ϵ from any point in N_ϵ, add v to N_ϵ.

After running this procedure, we have $|N_\epsilon|$ points such that $\min_{x \in N_\epsilon} |v - x| \leq \epsilon$ for all $v \in \mathcal{S}_{\mathcal{U}}$. So we just need to bound $|N_\epsilon|$.

To do so, we first lower bound the volume taken up by balls around points in $N_\epsilon = \{x_1, x_2, \ldots, x_{|N_\epsilon|}\}$. In particular, note that for all $i \neq j$, $|x_i - x_j| \geq \epsilon$. If not, then either x_i or x_j would not have been added to N_ϵ by our greedy algorithm. So if we place balls of radius $\epsilon/2$ around each x_i:

$$B(x_1, \epsilon/2) \ldots B(x_{|N_\epsilon|}, \epsilon/2)$$

then for all $i \neq j$, $B(x_i, \epsilon/2)$ does not intersect $B(x_j, \epsilon/2)$.

So how do we now set up an inequality to bound $|N_\epsilon|$??? The volume of a d dimensional ball of radius r is cr^d for some fixed constant c. Thus, the amount of space taken up by all the balls surrounding points in N_ϵ is $c|N_\epsilon|(\epsilon/2)^d$.

Next, the amount of space that these balls can exist in is at most the volume of a d dimensional sphere with radius $1 + \epsilon/2$. This volume is $c(1 + \epsilon/2)^d$. Thus, we have that

$$|N_\epsilon| c (\epsilon/2)^d \leq c(1 + \epsilon/2)^d$$

Solving for $|N_\epsilon|$:

$$\begin{aligned} |N_\epsilon| &\leq \frac{(1 + \epsilon/2)^d}{(\epsilon/2)^d} \\ &\leq \left(\frac{4}{\epsilon}\right)^d \end{aligned}$$

The last line holds assuming $0 < \epsilon \leq 2$. □

7.4 Proving Theorem 4

We can now prove Theorem 4, by using the ϵ-net.

Proof: Let $\epsilon' = \epsilon/3$. We first choose m sufficiently large so that Equation 2 holds with value $\epsilon \leftarrow \epsilon'$ for all the at most $\left(\frac{4}{\epsilon'}\right)^d$ vectors in $N_{\epsilon'}$. Theorem 2 and a union bound tell us this works if we choose $\delta' = \delta\left(\frac{\epsilon'}{4}\right)^d$ in $m = 9\ln(1/\delta')/\epsilon'^2$. Plugging in gives $m = 9\ln\left(\left(\frac{4}{\epsilon'}\right)^d/\delta\right)/\epsilon'^2 = (9d/\epsilon'^2)\ln\left(\frac{4}{\delta\epsilon'}\right) = (81d/(\epsilon')^2)\ln\left(\frac{12}{\delta\epsilon'}\right)$

Now consider any $v \in \mathcal{S}_{\mathcal{U}}$. We claim that for some sequence $x_0, x_1, x_2, \ldots \in N_{\epsilon'}$ that we can write v as:

$$v = x_0 + c_1x_1 + c_2x_2 + \ldots$$

for constants $c_0 = 1, c_1, c_2, \ldots$ where $|c_i| \leq \epsilon^i$. To see this, note that there is some point x_0 within distance ϵ' of v. Next, we need to represent $v - x_0$, which has norm at most ϵ'. So, we can write the point $\frac{v-x_0}{|v-x_0|}$, which has norm 1 and multiply the resulting coefficients by ϵ'. Again there is some point x_1 within distance ϵ of *this* point. Continuing this process ad infinitum gives the claim. This sequence can possibly be infinite and possibly have repeats. Let I be the set of (possibly infinite) indices in this sequence.

Now, we consider $|\Pi v|$ and use the triangle inequality to get:

$$\begin{aligned}
|\Pi v| &= \left|\Pi\left(\sum_{i\in I} c_i x_i\right)\right| \\
&\leq \sum_{i\in I} |\Pi(c_i x_i)| \\
&\leq \sum_{i\in I} (1+\epsilon')c_i|x_i| \\
&\leq (1+\epsilon')\sum_{i=0}^{\infty}(\epsilon')^i \\
&= \frac{1+\epsilon'}{1-\epsilon'} \\
&\leq 1+\epsilon
\end{aligned}$$

In the above, the second step follows by the triangle inequality. The third step follows by the fact that each $x_i \in N'_\epsilon$, and so Equation 2 holds for each of them. The last step holds since $\frac{1+\epsilon'}{1-\epsilon'} \leq 1+\epsilon$, when $1+\epsilon' \leq (1+\epsilon)(1-\epsilon')$ or $\epsilon' \leq \frac{\epsilon}{2+\epsilon}$, which always since $\epsilon \leq 2/5$ and $\epsilon' \leq \epsilon/3$.

The other direction of the proof is symmetric, and is included below for completeness.

$$\begin{aligned}
|\Pi v| &= \left|\Pi\left(\sum_{i\in I} c_i x_i\right)\right| \\
&\geq |\Pi(c_0 x_0)| - \sum_{i\in I/\{0\}} |\Pi(c_i x_i)| \\
&\geq |\Pi x_0| - \sum_{i\in I/\{0\}} (\epsilon')^i |\Pi x_i| \\
&\geq (1-\epsilon') - (1+\epsilon')\sum_{i=1}^{\infty}(\epsilon')^i \\
&\geq (1-\epsilon') - (1+\epsilon')\frac{\epsilon'}{1-\epsilon'} \\
&\geq (1-\epsilon') - \frac{5}{4}(\epsilon' + (\epsilon')^2) && \text{For } \epsilon' \leq 1/5 \\
&\geq 1 - \epsilon' - (5/4)\epsilon' - (1/4)\epsilon' \\
&\geq 1 - \epsilon && \text{For } \epsilon' \leq (1/3)\epsilon
\end{aligned}$$

The first step holds since by the triangle inequality, $|y| + |x-y| \geq |y + x - y| = |x|$; moving the $|y|$ term gives $|x-y| \geq |x| - |y|$. Finally, plugging in $y = -y'$ gives $|x + y'| \geq |x| - |y'|$ for all vectors x and y'. □

7.5 Other Applications of JL

Speed up Machine Learning algorithms by projecting "training data"? Depends. If classifier is linear then yes. If classifier is low dimensional polynomial, probably.

Approximate solutions to System of Linear equations? Sometimes

Finding an ϵ-approximate convex hull? Sometimes

8 Appendix

This appendix discusses Chernoff bounds, which are related to Bernstein's inequality but are less general, and so sometime tighter.

8.1 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 8. *(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:

$$\begin{aligned} E(e^{tX}) &= E(e^{t\sum_i X_i}) \\ &= E(\prod_{i\in[1,n} e^{tX_i}) \\ &= \prod_{i\in[1,n]} E(e^{tX_i}) \end{aligned}$$

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:

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

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 9. *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 $0 \leq \delta \leq 1$,*

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

8.2 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!

References

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

[2] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

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