Today we'll talk about dimensionality reduction, and some related topics in data streaming.

1 Dimension Reduction

Suppose we are given a set of n points $\{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^D$. How small can we make D and still maintain the Euclidean distances between the points? Clearly, we can always make $D = n - 1$, since any set of n points lies on a $n-1$-dimensional subspace. And this is (existentially) tight: e.g., the case when $x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1$ are all orthogonal vectors.

But what if we were OK with the distances being approximately preserved? In HW#5, you will see that while there could only be D orthogonal unit vectors in $\mathbb{R}^D$, there could be as many as $\exp(c\varepsilon^2 D)$ unit vectors which are ε-orthogonal—i.e., whose mutual inner products all lie in $[-\varepsilon, \varepsilon]$. Near-orthogonality allows us to pack exponentially more vectors!

Put another way, note that

$$\|\vec{a} - \vec{b}\|_2^2 = \langle \vec{a} - \vec{b}, \vec{a} - \vec{b} \rangle = \langle \vec{a}, \vec{a} \rangle + \langle \vec{b}, \vec{b} \rangle - 2\langle \vec{a}, \vec{b} \rangle = \|\vec{a}\|_2^2 + \|\vec{b}\|_2^2 - 2\langle \vec{a}, \vec{b} \rangle.$$

And hence the squared Euclidean distance between any pair of the points defined by these ε-orthogonal vectors falls in $2(1 \pm \varepsilon)$. So, if we wanted n points exactly at unit (Euclidean) distance from each other, we would need $n-1$ dimensions. (Think of a triangle in 2-dims.) But if we wanted to pack in n points which were at distance $(1 \pm \varepsilon)$ from each other, we could pack them into

$$O\left(\tfrac{\log n}{\varepsilon^2}\right)$$

dimensions.

1.1 The Johnson Lindenstrauss lemma

The Johnson Lindenstrauss "flattening" lemma says that such a claim is true not just for equidistant points, but for any set of n points in Euclidean space:

Lemma 19.1. *Let $\varepsilon \in (0, 1/2)$. Given any set of points $X = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^D$, there exists a map $A : \mathbb{R}^D \to \mathbb{R}^k$ with $k = O(\varepsilon^{-2} \log n)$ such that*

$$1 - \varepsilon \leq \frac{\|A(x_i) - A(x_j)\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \varepsilon.$$

Note that the target dimension k is independent of the original dimension D, and depends only on the number of points n and the accuracy parameter ε.

It is easy to see that we need at least $\Omega(\frac{1}{\varepsilon} \log n)$ using a packing argument. Noga Alon showed a lower bound of $\Omega(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon})$. Recently, Larsen and Nelson showed that any *linear* dimensionality reduction scheme must require $\Omega(\varepsilon^2 \log n)$ dimensions for some data sets.

1.2 The construction

The JL lemma is pretty surprising, but the construction of the map is perhaps even more surprising: it is a super-simple random construction. Let M be a $k \times D$ matrix, such that every entry of M is filled with an i.i.d. draw from a standard normal $N(0,1)$ distribution (a.k.a. the "Gaussian" distribution). For $x \in \mathbb{R}^D$, define

$$A(x) = \frac{1}{\sqrt{k}} Mx.$$

That's it. You hit the vector x with a Gaussian matrix M, and scale it down by $\sqrt{k}$. That's the map A. Note that it is a linear map: $A(x) + A(y) = A(x+y)$. So suppose we could show the following lemma:

Lemma 19.2. *Let $\varepsilon \in (0, 1/2)$. If A is constructed as above with $k = c\varepsilon^{-2} \log \delta^{-1}$, and $x \in \mathbb{R}^D$ is a unit vector, then*

$$\mathbf{Pr}[\|A(x)\|_2^2 \in 1 \pm \varepsilon] \geq 1 - \delta.$$

Then we'd get a proof of Lemma 19.1. Indeed, set $\delta = 1/n^2$, and hence $k = O(\varepsilon^{-2} \log n)$. Now for each $x_i, x_j \in X$ we get that the squared length of $x_i - x_j$ is maintained to within $1 \pm \varepsilon$ with probability at least $1 - 1/n^2$. By a union bound, all $\binom{n}{2}$ pairs of distances in $\binom{X}{2}$ are maintained with probability at least $1 - \binom{n}{2}\frac{1}{n^2} \geq 1/2$. This proves Lemma 19.1.

A few comments about this construction:

- The above proof shows not only the existence of a good map, we also get that a random map as above works with constant probability! In other words, a Monte-Carlo randomized algorithm for dimension reduction. (Since we can efficiently check that the distances are preserved to within the prescribed bounds, we can convert this into a Las Vegas algorithm.) Or we can also get deterministic algorithms: see here.
- The algorithm (at least the Monte Carlo version) *does not even look* at the set of points X: it works for any set X with high probability. Hence, we can pick this map A before the points in X arrive.

1.3 The proof

Now, on to the proof of Lemma 19.2. Here's the main idea. Imagine that the vector we're considering is just the elementary unit vector $e_1 = (1, 0, \ldots, 0)$. Then $M e_1$ is just a vector with independent and identical Gaussian values, and we're interested in its length—the sum of squares of these Gaussians. If these were bounded r.v.s, we'd be done—but they are not. However, their tails are very small, so things should work out

But what's a Gaussian $N(0,1)$? Well, it looks like this:

Which is not too different from this (bounded) random variable, if you squint a bit:

Which has constant mean. So, if we take a sum of a bunch of such random variables (actually of their squares), it should behave pretty much like its mean (which is $\propto k$), because of a Chernoff-like argument. And so the expected length is close to $\sqrt{k}$, which explains the division by $\sqrt{k}$.

Of course this is very vague and imprecise. In fact, while the Laplace distribution with distribution $f(x) \propto e^{-\lambda|x|}$ for $x \in \mathbb{R}$ also has pretty thin tails—"exponential tails", this won't work the same, even if you squint as hard as you like. It turns out you need "sub-Gaussian tails". So we just need to make all this precise, and remove the assumption that the vector was just e_1. That's what the rest of the formal proof does: it has a few steps, but each of them is fairly elementary.

1.4 The proof, this time for real

We'll be using basic facts about Gaussians, let's just recall them. The probability density function for the Gaussian $N(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}.$$

We also use the following; the proof just needs some elbow grease.

Proposition 19.3. *If $G_1 \sim N(\mu_1, \sigma_1^2)$ and $G_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then for $c \in \mathbb{R}$,*

$$c\,G_1 \sim N(c\mu_1, c^2\,\sigma_1^2) \tag{19.1}$$

$$G_1 + G_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{19.2}$$

Recall that we want to argue about the squared length of $A(x) \in \mathbb{R}^k$. To start off, observe that each coordinate of the vector Mx behaves like

$$Y \sim \langle G_1, G_2, \ldots, G_D \rangle \cdot x = \sum x_i\, G_i$$

where the G_i's are i.i.d. $N(0,1)$ r.v.s. But then Proposition 19.3 tells us that $Y \sim N(0, x_1^2 + x_2^2 + \ldots + x_D^2)$. And since x is a unit length vector, this is $N(0,1)$. So, each of the k coordinates of Mx behaves just like an independent Gaussian!

1.4.1 The Expectation

What is the squared length of $A(x) = \frac{1}{\sqrt{k}} Mx$, then? It is

$$Z := \sum_{i=1}^{k} \frac{1}{k} \cdot G_i^2$$

where each $G_i \sim N(0,1)$, independent of the others. And since $E[G_i^2] = \mathrm{Var}(G_i) + E[G_i]^2 = 1$, we get $E[Z] = 1$.

1.4.2 Concentration about the Mean

Now to show that Z does not deviate too much from 1. And Z is the sum of a bunch of independent and identical random variables. If only the G_i's were all bounded, we could have used a Chernoff bound and be done. But these are not bounded, so this is finally where we'll need to do a little work. [1] So let's start down the ye olde Chernoff path, for the upper tail, say:

$$\mathbf{Pr}[Z \geq 1 + \varepsilon] \leq \mathbf{Pr}[e^{tkZ} \geq e^{tk(1+\varepsilon)}] \leq E[e^{tkZ}]/e^{tk(1+\varepsilon)} = \prod_i \left(E[e^{tG_i^2}]/e^{t(1+\varepsilon)} \right) \tag{19.3}$$

[1] Note: we could take the easy way out, observe that the squares of Gaussians are chi-squared r.v.s, the sum of k of them is *chi-squared with k degrees of freedom*, and the internets conveniently has tail bounds for these things. But we digress.

for every $t > 0$. And what is $E[e^{tG^2}]$ for $G \sim N(0,1)$? Let's calculate it:

$$\frac{1}{\sqrt{2\pi}} \int_{g \in \mathbb{R}} e^{tg^2} e^{-g^2/2} dg = \frac{1}{\sqrt{2\pi}} \int_{z \in \mathbb{R}} e^{-z^2/2} \frac{dz}{\sqrt{1-2t}} = \frac{1}{\sqrt{1-2t}}. \tag{19.4}$$

for $t < 1/2$. So our current bound on the upper tail is that for all $t \in (0, 1/2)$ we have

$$\mathbf{Pr}[Z \geq (1+\varepsilon)] \leq \left(\frac{1}{e^{t(1+\varepsilon)}\sqrt{1-2t}} \right)^k.$$

Let's just focus on part of this expression:

$$\left(\frac{1}{e^{t}\sqrt{1-2t}} \right) = \exp\left(-t - \frac{1}{2}\log(1-2t) \right) \tag{19.5}$$

$$= \exp\left((2t)^2/4 + (2t)^3/6 + \cdots \right) \leq \exp\left(t^2(1 + 2t + 2t^2 + \cdots) \right) \tag{19.6}$$

$$= \exp(t^2/(1-2t)).$$

Plugging this back, we get

$$\mathbf{Pr}[Z \geq (1+\varepsilon)] \leq \left(\frac{1}{e^{t(1+\varepsilon)}\sqrt{1-2t}} \right)^k$$
$$\leq \exp(kt^2/(1-2t) - kt\varepsilon) \leq e^{-k\varepsilon^2/8},$$

if we set $t = \varepsilon/4$ and use the fact that $1 - 2t \geq 1/2$ for $\varepsilon \leq 1/2$. (Note: this setting of t also satisfies $t \in (0, 1/2)$, which we needed from our previous calculations.)

Almost done: let's take stock of the situation. We observed that $\|A(x)\|_2^2$ was distributed like an average of squares of Gaussians, and by a Chernoff-like calculation we proved that

$$\mathbf{Pr}[\|A(x)\|_2^2 > 1 + \varepsilon] \leq \exp(-k\varepsilon^2/8) \leq \delta/2$$

for $k = \frac{8}{\varepsilon^2} \ln \frac{2}{\delta}$. A similar calculation bounds the lower tail, and finishes the proof of Lemma 19.2.

Citations: The JL Lemma was first proved in this paper of Bill Johnson and Joram Lindenstrauss. There have been several proofs after theirs, usually trying to tighten their results, or simplify the algorithm/proof (see citations in some of the newer papers): the proof follows some combinations of the proofs in this STOC '98 paper of Piotr Indyk and Rajeev Motwani, and this paper by Sanjoy Dasgupta and myself.

2 Using Random Signs instead of Gaussians

While Gaussians have all kinds of nice properties, they are real-valued distributions and hence require attention to precision. How about populating A with draws from other, simpler distributions? How about setting each $M_{ij} \in_R \{-1, +1\}$, and letting $A = \frac{1}{\sqrt{k}} M$? (A random sign is also called a *Rademacher random variables*, the name Bernoulli being already taken for a random bit in $\{0, 1\}$.)

Now, we want to study the properties of

$$Z := \sum_{i=1}^{k} \left(\sum_{j=1}^{D} A_{ij} \cdot x_j \right)^2. \tag{19.7}$$

To keep subscripts to a minimum, consider the inner sum for index i, which looks like

$$Y_i := \left(\sum_j R_j \cdot x_j \right) \tag{19.8}$$

each R_j being an independent Rademacher variable.

$$\begin{aligned} E[Y_i^2] &= E[(\sum_j R_j x_j)(\sum_l R_l x_l)] \\ &= E[\sum_j R_j^2 x_j^2 + \sum_{j \neq l} R_j R_l x_j x_l] \\ &= \sum_j E[R_j^2] x_j^2 + \sum_{j \neq l} E[R_j R_l] x_j x_l = \sum_j x_j^2. \end{aligned}$$

if the R_j's are pairwise independent, since $R_j^2 = 1$ and $E[R_j R_l] = E[R_j]E[R_l] = 0$ by independence. Plugging this into (19.7) and recalling that $A_{ij} \in \{-\frac{1}{\sqrt{k}}, +\frac{1}{\sqrt{k}}\}$, we get

$$E[Z] = \sum_{i=1}^{k} \frac{1}{k} E[Y_i^2] = \sum_{i=1}^{k} \frac{1}{k} \sum_j x_j^2 = \|x\|_2^2. \tag{19.9}$$

Just what we like! Now we just need to show that $\mathbf{Pr}[Z \in (1 \pm \varepsilon)\|x\|_2^2] \geq 1 - \delta$ as long as $k = \Omega(\varepsilon^{-2} \log \delta^{-1})$.

2.1 Concentration Around the Mean via Subgaussian-ness

Let's look over the proof in Section 1.4.2, and see what properties of Gaussians we used. We used that for $t \in (0, 1/2)$,

$$(\star)\ E[e^{tG^2}] \leq \frac{1}{\sqrt{1-2t}}$$

but the rest of the Chernoff-like proof of Section 1.4 did not use any other facts about Gaussians. We can prove $(\star)$ for Rademacher random variables using explicit calculations, but instead let's give a useful abstraction:

Definition 19.4. A random variable V is said to be *subgaussian with parameter c* and for all real s, we have $E[e^{sV}] \leq e^{cs^2}$.

(You can define subgaussian-ness alternatively as satisfying $\mathbf{Pr}[|V| > v] \leq Ce^{-cv^2}$ for suitable constants c, C; these notes by Roman Vershynin show the two definitions are equivalent for symmetric distributions.) Here are some useful facts:

Fact 19.5. *The following facts hold:*

(i) (Gaussian) For $G \sim N(0,1)$, then $E[e^{sG}] = e^{s^2/2}$; i.e., it is $1/2$-subgaussian.

(ii) (Rademacher) A Rademacher random variable is $1/2$-subgaussian.

(iii) (Sums) If V_i's are independent and c-subgaussian, and $\|x\|_2 = 1$, then $V = \sum_i x_i V_i$ is also c-subgaussian.

Proof. The first fact about Gaussians is a simple calcuation:

$$E[e^{sG}] = \frac{1}{\sqrt{2\pi}} \int_{x\in\mathbb{R}} e^{sx} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} e^{s^2/2} \int_{x\in\mathbb{R}} e^{-\frac{(x-s)^2}{2}} dx = e^{s^2/2}.$$

(Good — it's heartening to know that a Gaussian is also subgaussian!)

For the second fact, observe that:

$$E[e^{sR}] = \frac{e^s + e^{-s}}{2} = \cosh s = 1 + \frac{s^2}{2!} + \frac{s^4}{4!} + \cdots \leq e^{s^2/2}.$$

Finally,

$$E[e^{sV}] = E[e^{\sum_i (sx_i)V_i}] \leq \prod_i e^{c(sx_i)^2} = e^{cs^2 \sum_i x_i^2} = e^{cs^2}.$$

The inequality in the middle uses the definition of subgaussian-ness. □

Lemma 19.6. *If V is subgaussian with parameter c, then $E[e^{sV^2}] \leq \frac{1}{\sqrt{1-4cs}}$ for $s > 0$.*

Proof. Well, suppose $G \sim N(0,1)$ is an independent Gaussian, then

$$E_V[e^{sV^2}] = E_{G,V}[e^{\sqrt{2s}VG}]$$

by the calculation we just did for Gaussians. (Note that we've just introduced a Gaussian into the mix, without any provocation! But it will all work out.) Let just rewrite that

$$E_{G,V}[e^{\sqrt{2s}VG}] = E_G[E_V[e^{(\sqrt{2s}G)V}]].$$

Using the c-subgaussian behavior of V we bound this by

$$E_G[e^{c(\sqrt{2s}|G|)^2}] = E_G[e^{2csG^2}].$$

Finally, the calculation (19.4) gives this to be $\frac{1}{\sqrt{1-4cs}}$. □

Excellent. Note that Y_i is a weighted sum of Rademachers (as defined in (19.8)); by Fact 19.5(ii) and (iii), each R_i's are 1/2-subgaussian, so $Y_i = \sum_i x_i R_i$ is too. And hence $E[e^{tY_i^2}] \leq \frac{1}{\sqrt{1-2t}}$ for $\{-1,+1\}$-random variables as well. And now doing the same calculations as for the Gaussian case, from Section 1.4.2, we get that the Rademacher matrix also has the JL property!

Note that the Rademacher JL matrix A now just requires us to pick $kD = O(D\varepsilon^{-2}\log\delta^{-1})$ random bits (instead of kD random Gaussians); also, there are fewer precision issues to worry about. One can consider other distributions to stick into the matrix A—all you need to show is that Z has the right mean, and that the entries are subgaussian.[2]

Citations: The scheme of using Rademacher matrices instead of Gaussian matrices for JL was first proposed in this paper by Dimitris Achlioptas. The idea of extending it to subgaussian distributions appears in this paper of Indyk and Naor, and this paper of Matousek. The paper of Klartag and Mendelson generalizes this even further.

Fast J-L: Do we really need to plug in non-zero values into every entry of the matrix A? What if most of A is filled with zeroes? The first problem is that if x is a very sparse vector, then Ax

[2] If the $E[e^{tG^2}] \leq \frac{1}{\sqrt{1-at}}$ for $a \neq 2$, you will need to redo the proof from Section 1.4.2 since the linear terms in (19.5) don't cancel any more to give (19.6). See, e.g., Indyk and Naor or Matousek for details.

might be zero with high probability? Achlioptas showed that having a random two-thirds of the entries of A being zero still works fine: the paper of Nir Ailon and Bernard Chazelle showed that if you first hit x with a suitable matrix P which caused Px to be "well-spread-out" whp, and then $\|APx\| \approx \|x\|$ would still hold for a much sparser A. Moreover, this P requires much less randomless, and furthermore, the computations can be done faster too! There has been much work on fast and sparse versions of JL: see, e.g., this SODA 11 paper of Ailon and Edo Liberty, and this arxiv preprint by Daniel Kane and Jelani Nelson. Jelani has some notes on the Fast JL Transform.

Compressive Sensing: Finally, the J-L lemma is closely related to compressive sensing: how to reconstruct a sparse signal using very few measurements. See these notes by Jiri Matousek, or these by Baraniuk and others for a proof of the beautiful connection. Hopefully I will say more about compressive sensing in a later lecture.