10
Dimension Reduction and the JL Lemma

For a set of n points $\{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^D$, can we map them into some lower dimensional space $\mathbb{R}^k$ and still maintain the Euclidean distances between them? We can always take $k \leq n-1$, since any set of n points lies on a $n-1$-dimensional subspace. And this is (existentially) tight, e.g., if $x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1$ are all orthogonal vectors.

But what if we were fine with distances being approximately preserved? There can only be k orthogonal unit vectors in $\mathbb{R}^k$, but there are as many as $\exp(c\varepsilon^2 k)$ 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! (Indeed, we will see this in a homework exercise.)

This near-orthogonality of the unit vectors means that distances are also approximately preserved. Indeed, for any two $a, b \in \mathbb{R}^k$,

$$\|a-b\|_2^2 = \langle a-b, a-b\rangle = \langle a, a\rangle + \langle b, b\rangle - 2\langle a, b\rangle = \|a\|_2^2 + \|b\|_2^2 - 2\langle a, b\rangle,$$

so the squared Euclidean distance between any pair of the points defined by these ε-orthogonal vectors falls in the range $2(1 \pm \varepsilon)$. So, if we wanted n points at exactly the same (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

$$k = O\left(\frac{\log n}{\varepsilon^2}\right)$$

dimensions.

Having $n \geq \exp(c\varepsilon^2 k)$ vectors in d dimensions means the dimension is $k = O(\log n/\varepsilon^2)$.

10.1 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 10.1. *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 $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ such that*

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

Moreover, such a map can be computed in expected $\text{poly}(n, D, 1/\varepsilon)$ *time.*

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 not difficult to show that we need at least $\Omega(\log n)$ dimensions in such a result, using a packing argument. Noga Alon showed a lower bound of $\Omega(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon})$, and then Kasper Green Larson and Jelani Nelson showed a tight and matching lower bound of $\Omega(\frac{\log n}{\varepsilon^2})$ dimensions for any dimensionality reduction scheme from n dimensions that preserves pairwise distances.

The JL Lemma was first considered in the area of metric embeddings, for applications like fast near-neighbor searching; today we use it to speed up algorithms for problems like spectral sparsification of graphs, and solving linear programs fast.

Given n points with Euclidean distances in $(1 \pm \varepsilon)$, the balls of radius $\frac{1-\varepsilon}{2}$ around these points must be mutually disjoint, by the minimum distance, and they are contained within a ball of radius $(1 + \varepsilon) + \frac{1-\varepsilon}{2}$ around x_0. Since volumes of balls in $\mathbb{R}^k$ of radius r behave like $c_k r^k$, we have

$$n \cdot c_k \left(\frac{1-\varepsilon}{2}\right)^k \leq c_k \left(\frac{3+\varepsilon}{2}\right)^k$$

or $k \geq \Omega(\log n)$ for $\varepsilon \leq 1/2$.

Alon (2003)

Larson and Nelson (2017)

10.2 The Construction

The JL lemma is pretty surprising, but the construction of the map is perhaps even more surprising: it is a super-simple randomized 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.

Since $A(x)$ is a linear map and satisfies $\alpha A(x) + \beta A(y) = A(\alpha x + \beta y)$, it is enough to show the following lemma:

Lemma 10.2. *[Distributional Johnson-Lindenstrauss] 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*

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

To prove Lemma 10.1, set $\delta = 1/n^2$, and hence $k = O(\varepsilon^{-2} \log n)$. Now for each $x_i, x_j \in X$, use linearity of $A(\cdot)$ to infer

$$\frac{\left\|A(x_i) - A(x_j)\right\|^2}{\left\|x_i - x_j\right\|^2} = \frac{\left\|A(x_i - x_j)\right\|^2}{\left\|x_i - x_j\right\|^2} = \left\|A(v_{ij})\right\|^2 \in (1 \pm \varepsilon)$$

with probability at least $1 - 1/n^2$, where v_{ij} is the unit vector in the direction of $x_i - x_j$. By a union bound, all $\binom{n}{2}$ pairs of distances in $\binom{X}{2}$ are maintained with probability at least $1 - \binom{n}{2}\frac{1}{n^2} \geq 1/2$. 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) is ***data-oblivious***: it 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.

10.3 Intuition for the Distributional JL Lemma

Let us recall some basic facts about Gaussians distributions. 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.

The fact that the means and the variances take on the claimed values should not be surprising; this is true for all r.v.s. The surprising part is that the resulting variables are also Gaussians.

Proposition 10.3. *If $G_1 \sim N(\mu_1, \sigma_1^2)$ and $G_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then for $c \in \mathbb{R}$,*

$$c\,G_1 \sim N(c\mu_1, c^2\,\sigma_1^2) \tag{10.1}$$

$$G_1 + G_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{10.2}$$

Now, here's the main idea in the proof of Lemma 10.2. Imagine that the vector x is the elementary unit vector $e_1 = (1, 0, \ldots, 0)$. Then $M\,e_1$ is just the first column of M, which is a vector with independent and identical Gaussian values.

$$M\,e_1 = \begin{bmatrix} G_{1,1} & G_{1,2} & \cdots & G_{1,D} \\ G_{2,1} & G_{2,2} & \cdots & G_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ G_{k,1} & G_{k,2} & \cdots & G_{k,D} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} G_{1,1} \\ G_{2,1} \\ \vdots \\ G_{k,1} \end{bmatrix}.$$

$A(x)$ is a scaling-down of this vector by $\sqrt{k}$: every entry in this random vector $A(x) = A(e_1)$ is distributed as

$$1/\sqrt{k} \cdot N(0,1) = N(0, 1/k) \qquad \text{(by (10.1))}.$$

Thus, the expected squared length of $A(x) = A(e_1)$ is

$$\mathbb{E}\left[\|A(x)\|^2\right] = \mathbb{E}\left[\sum_{i=1}^{k} A(x)_i^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[A(x)_i^2\right] = \sum_{i=1}^{k} \frac{1}{k} = 1.$$

If G has mean μ and variance σ^2, then $\mathbb{E}[G^2] = \mathrm{Var}[G] + \mathbb{E}[G]^2 = \sigma^2 + \mu^2$.

So the expectation of $\|A(x)\|^2$ is 1; the heart is in the right place! Now to show that $\|A(x)\|^2$ does not deviate too much from the mean—i.e., to show a concentration result. Indeed, $\|A(x)\|^2$ is a sum of independent $N(0, 1/k)^2$ random variables, so if these $N(0, 1/k)^2$ variables were bounded, we would be done by the Chernoff bounds of the previous chapter. Sadly, they are not. However, their tails are fairly "thin", so if we squint hard enough, these random variables can be viewed as "pretty much bounded", and the Chernoff bounds can be used.

Of course this is very vague and imprecise. Indeed, the Laplace distribution with density function $f(x) \propto e^{-\lambda|x|}$ for $x \in \mathbb{R}$ also has pretty thin tails—"exponential tails". But using a matrix with Laplace entries does not work the same, no matter how hard we squint. It turns out you need the entries of M, the matrix used to define $A(x)$, to have "sub-Gaussian tails". The Gaussian entries have precisely this property.

We now make all this precise, and also remove the assumption that the vector $x = e_1$. In fact, we do this in two ways. First we give a direct proof: it has several steps, but each step is elementary, and you are mostly following your nose. The second proof formally defines the notion of sub-Gaussian random variables, and builds some general machinery for concentration bounds.

10.4 *The Direct Proof of Lemma 10.2*

Recall that we want to argue about the squared length of $A(x) \in \mathbb{R}^k$, where $A(x) = \frac{1}{\sqrt{k}} Mx$, and x is a unit vector. To start off, observe that the i^{th} coordinate of the vector Mx is the inner product of a row of M with the vector x. This is distributed as

$$Y_i \sim \langle G_1, G_2, \ldots, G_D \rangle \cdot x = \sum_j x_j G_j$$

where the G_j's are the i.i.d. $N(0,1)$ r.v.s on the i^{th} row of M. Now Proposition 10.3 tells us that $Y_i \sim N(0, x_1^2 + x_2^2 + \ldots + x_D^2)$. Since x is a unit length vector, we get

$$Y_i \sim N(0,1).$$

So, each of the k coordinates of Mx behaves just like an independent Gaussian!

10.4.1 The Expectation

Given the observation above, the squared length of $A(x) = \frac{1}{\sqrt{k}}Mx$ is

$$Z := \|A(z)\|^2 = \sum_{i=1}^{k} \frac{1}{k} \cdot Y_i^2$$

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

10.4.2 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. Let's start down the usual path for a Chernoff bound, for the upper tail, say:

$$\Pr[Z \geq 1+\varepsilon] \leq \Pr[e^{tkZ} \geq e^{tk(1+\varepsilon)}] \leq \mathbb{E}[e^{tkZ}]/e^{tk(1+\varepsilon)} \tag{10.3}$$

$$= \prod_i \left(\mathbb{E}[e^{tY_i^2}]/e^{t(1+\varepsilon)} \right) \tag{10.4}$$

The easy way out is to observe that the squares of Gaussians are chi-squared r.v.s, the sum of k of them is χ^2 *with k degrees of freedom*, and the internet conveniently has tail bounds for these things. But if you don't recall these facts, and don't have internet connectivity and cannot check Wikipedia, things are not that difficult.

for every $t > 0$. Now $\mathbb{E}[e^{tG^2}]$, the moment-generating function for G^2, where $G \sim N(0,1)$ is easy to calculate for $t < 1/2$:

$$\frac{1}{\sqrt{2\pi}} \int_{g \in \mathbb{R}} e^{tg^2} e^{-g^2/2} dg = \frac{1}{\sqrt{2\pi}} \int_{z \in \mathbb{R}} e^{-z^2/2} \frac{dz}{\sqrt{1-2t}} = \frac{1}{\sqrt{1-2t}}. \tag{10.5}$$

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

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

Let's just focus on part of this expression:

$$\left(\frac{1}{e^t \sqrt{1-2t}} \right) = \exp\left(-t - \frac{1}{2}\log(1-2t) \right) \tag{10.6}$$

$$= \exp\left((2t)^2/4 + (2t)^3/6 + \cdots \right) \tag{10.7}$$

$$\leq \exp\left(t^2(1 + 2t + 2t^2 + \cdots) \right) \tag{10.8}$$

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

Plugging this back, we get

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

$$\leq \exp(kt^2/(1-2t) - kt\varepsilon) \leq e^{-k\varepsilon^2/8},$$

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

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

for $k = \frac{8}{\varepsilon^2} \ln \frac{2}{\delta}$. A similar calculation bounds the lower tail, and finishes the proof of Lemma 10.2.

The JL Lemma was first proved by 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 above is some combinations of those by Piotr Indyk and Rajeev Motwani, and Sanjoy Dasgupta and myself.

Johnson and Lindenstrauss (1982)

Indyk and Motwani (1998)

Dasgupta and Gupta (2004)

10.5 *Subgaussian Random Variables*

While Gaussians have all kinds of nice properties, they are real-valued and hence require more randomness to generate. What other classes of r.v.s could give us bounds that are comparable? E.g., what about setting each $M_{ij} \in_R \{-1, +1\}$?

A random sign is also called a ***Rademacher random variable***, the name Bernoulli being already taken for a random bit in $\{0, 1\}$.

It turns out that Rademacher r.v.s also suffice, and we can prove this with some effort. But instead of giving a proof from first principles, let us abstract out the process of proving Chernoff-like bounds, and give a proof using this abstraction.

Recall the basic principle of a Chernoff bound: to bound the upper tail of an r.v. V with mean μ, we can choose any $t \geq 0$ to get

$$\Pr[V - \mu \geq \lambda] = \Pr[e^{t(V-\mu)} \geq e^{t\lambda}] \leq \mathbb{E}[e^{t(V-\mu)}] \cdot e^{-t\lambda}.$$

Now if we define the (centered) log-MGF of V as

$$\psi(t) := \ln \mathbb{E}[e^{t(V-\mu)}],$$

we get that for any $t \geq 0$,

$$\Pr[V - \mu \geq \lambda] \leq e^{-(t\lambda - \psi(t))}.$$

The best upper bound is obtained when the expression $t\lambda - \psi(t)$ is the largest. The ***Legendre dual*** of the function $\psi(t)$ is defined as

Exercise: if $\psi_1(t) \geq \psi_2(t)$ for all $t \geq 0$, then $\psi_1^*(\lambda) \leq \psi_2^*(\lambda)$ for all λ.

$$\psi^*(\lambda) := \inf_{t \geq 0} \{t\lambda - \psi(t)\},$$

so we get the concise statement for a generic Chernoff bound:

Bounds for the lower tail follow from the arguments applied to the r.v. $-X$.

$$\Pr[V - \mu \geq \lambda] \leq \exp(-\psi^*(\lambda)). \tag{10.9}$$

This abstraction allows us to just focus on bounds on the dual log-MGF function $\psi^*(\lambda)$, making the arguments cleaner.

10.5.1 A Couple of Examples

Let's do an example: suppose $V \sim N(\mu, \sigma^2)$, then

$$\begin{aligned}\mathbb{E}[e^{t(V-\mu)}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{x\in\mathbb{R}} e^{tx}e^{-\frac{x^2}{2\sigma^2}}dx \\ &= \frac{1}{\sqrt{2\pi}\sigma}e^{t^2\sigma^2/2}\int_{x\in\mathbb{R}} e^{-\frac{(x-t\sigma^2)^2}{2\sigma^2}}dx = e^{t^2\sigma^2/2}. \end{aligned}\tag{10.10}$$

Hence, for $N(\mu, \sigma^2)$ r.v.s, we have

$$\psi(t) = \frac{t^2\sigma^2}{2} \qquad \text{and} \qquad \psi^*(\lambda) = \frac{\lambda^2}{2\sigma^2},$$

the latter by basic calculus. Now the generic Chernoff bound for says that for normal $N(\mu, \sigma^2)$ variables,

$$\Pr[V - \mu \geq \lambda] \leq e^{-\frac{\lambda^2}{2\sigma^2}}. \tag{10.11}$$

How about a Rademacher $\{-1, +1\}$-valued r.v. V? The MGF is

$$\mathbb{E}[e^{t(V-\mu)}] = \frac{e^t + e^{-t}}{2} = \cosh t = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots \leq e^{t^2/2},$$

so

$$\psi(t) = \frac{t^2}{2} \qquad \text{and} \qquad \psi^*(\lambda) = \frac{\lambda^2}{2}.$$

Note that

$$\psi_{\text{Rademacher}}(t) \leq \psi_{N(0,1)}(t) \implies \psi^*_{\text{Rademacher}}(\lambda) \geq \psi^*_{N(0,1)}(\lambda).$$

This means the upper tail bound for a single Rademacher is at least as strong as that for the standard normal.

10.5.2 Defining Subgaussian Random Variables

Definition 10.4. A random variable V with mean μ is ***subgaussian with parameter*** σ if $\psi(t) \leq \frac{\sigma^2 t^2}{2}$.

By the generic Chernoff bound (10.9), such an r.v. has tails that are smaller than those of a normal r.v. with variance σ^2. The following fact is an analog of Proposition 10.3.

Lemma 10.5. *If $V_1, V_2, \ldots$ are independent and σ_i-subgaussian, and $x_1, x_2, \ldots$ are reals, then $V = \sum_i x_i V_i$ is $\sqrt{\sum_i x_i^2\sigma_i^2}$-subgaussian.*

Proof.

$$\mathbb{E}[e^{t(V-\mu)}] = \mathbb{E}[e^{t\sum_i x_i(V_i-\mu_i)}] = \prod_i \mathbb{E}[e^{tx_i(V_i-\mu_i)}] = \prod_i e^{tx_i(V_i-\mu_i)}.$$

Now taking logarithms, $\psi_V(t) = \sum_i \psi_{V_i}(tx_i) \leq \sum_i \frac{t^2x_i^2\sigma_i^2}{2}$. □

10.6 JL Matrices using Rademachers and Subgaussian-ness

Suppose we choose each $M_{ij} \in_R \{-1, +1\}$ and let $A(x) = \frac{1}{\sqrt{k}} Mx$ again? We want to show that

$$Z := \|A(x)\|^2 = \frac{1}{k} \sum_{i=1}^{k} \left(\sum_{j=1}^{D} M_{ij} \cdot x_j \right)^2. \tag{10.12}$$

has mean $\|x\|^2$, and is concentrated sharply around that value.

10.6.1 The Expectation

To keep subscripts to a minimum, consider the inner sum for index i in (10.12), which looks like

$$Y_i := \left(\sum_j M_j \cdot x_j \right). \tag{10.13}$$

where M_js are *pairwise independent*, with *mean zero* and *unit variance*.

$$\begin{aligned}
\mathbb{E}\left[Y_i^2\right] &= \mathbb{E}\left[(\sum_j M_j x_j)(\sum_l M_l x_l)\right] \\
&= \mathbb{E}\left[\sum_j M_j^2 x_j^2 + \sum_{j \neq l} M_j M_l x_j x_l\right] \\
&= \sum_j \mathbb{E}\left[M_j^2\right] x_j^2 + \sum_{j \neq l} \mathbb{E}\left[M_j M_l\right] x_j x_l = \sum_j x_j^2.
\end{aligned}$$

Here $\mathbb{E}[M_j^2] = \mathrm{Var}(M_j) + \mathbb{E}[M_j]^2 = 1$, and moreover $\mathbb{E}[M_j M_l] = \mathbb{E}[M_j]\mathbb{E}[M_l] = 0$ by pairwise independence. Plugging this into (10.12),

$$\mathbb{E}[Z] = \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[Y_i^2] = \frac{1}{k} \sum_{i=1}^{k} \sum_j x_j^2 = \|x\|_2^2. \tag{10.14}$$

So the expectation is what we want!

10.6.2 Concentration

Now to show concentration: the direct proof from §10.4 showed the Y_is were themselves Gaussian with variance $\|x\|^2$. Since the Rademachers are 1-subgaussian, Lemma 10.5 shows that Y_i is subgaussian with parameter $\|x\|^2$. Next, we need to consider Z, which is the average of ***squares of*** k independent Y_is. The following lemma shows that the MGF of squares of *symmetric* σ-subgaussians are bounded above by the corrresponding Gaussians with variance σ^2.

An r.v. X is ***symmetric*** if it is distributed the same as $R|X|$, where R is an independent Rademacher.

Lemma 10.6. *If V is symmetric mean-zero σ-subgaussian r.v., and $W \sim N(0, \sigma^2)$, then $\mathbb{E}[e^{tV^2}] \leq \mathbb{E}[e^{tW^2}]$ for $t > 0$.*

Proof. Using the calculation in (10.10) in the "backwards" direction

$$\mathbb{E}_V[e^{tV^2}] = \mathbb{E}_{V,W}[e^{\sqrt{2t}(|V|/\sigma)\,W}].$$

(Note that we've just introduced W into the mix, without any provocation!) Since W is also symmetric, we get $|V|W = V|W|$. Hence, rewriting

$$\mathbb{E}_{V,W}[e^{\sqrt{2t}(|V|/\sigma)\,W}] = \mathbb{E}_W[\mathbb{E}_V[e^{(\sqrt{2t}|W|/\sigma)V}]],$$

we can use the σ-subgaussian behavior of V in the inner expectation to get an upper bound of

$$\mathbb{E}_W[e^{\sigma^2(\sqrt{2t}|W|/2)^2/2}] = E_W[e^{tW^2}]. \qquad \square$$

Excellent. Now the tail bound for sums of squares of symmetric mean-zero σ-subgaussians follows from that of gaussians. Hence we get the same tail bounds as in §10.4.2, and hence that the Rademacher matrix also has the distributional JL property, while using far fewer random bits!

In general one can use other σ-subgaussian distributions to fill the matrix M—using σ different than 1 may require us to rework the proof from §10.4.2 since the linear terms in (10.6) don't cancel any more, see works by Indyk and Naor or Matousek for details.

Indyk and Naor (2008)
Matoušek (2008)

10.6.3 *The Fast JL Transform*

A different direction to consider is getting fast algorithms for the JL Lemma: 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 might be zero with high probability? Achlioptas showed that having a random two-thirds of the entries of A being zero still works fine: 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 paper from SOSA 2018 by Michael Cohen, T.S. Jayram, and Jelani Nelson. Jelani Nelson also has some notes on the Fast JL Transform.

Ailon and Chazelle

Cohen, Jayram, and Nelson (2018)

10.7 *Optional: Compressive Sensing*

To rewrite. In an attempt to build a better machine to take MRI scans, we decrease the number of sensors. Then, instead of the signal x we

intended to obtain from the machine, we only have a small number of measurements of this signal. Can we hope to recover x from the measurements we made if we make sparsity assumptions on x? We use the term r-sparse signal for a vector with at most r nonzero entries.

Formally, x is a n-dimensional vector, and a measurement of x with respect to a vector a is a real number given by $\langle x, a \rangle$. The question we want to answer is how to reconstruct x with r nonzero entries satisfying $Ax = b$ if we are given $k \times n$ matrix A and n dimensional vector b. This is often written as

$$\min \left\{ \|x\|_0 \mid Ax = b. \right\}$$

Here the ℓ_0 "norm" is the total number of non-zeros in the vector x.

Unfortunately, it turns out that the problem as formulated is NP-hard: but this is only assuming A and b are contrived by an adversary. Our setting is a bit different. x is some r-sparse signal out there that we want to determine. We have a handle over A and can choose it to be any matrix we like, and we are provided with appropriate $b = Ax$, from which we attempt to reconstruct x.

Consider the following similar looking problem called the ***basis pursuit*** (BP) problem:

$$\min \left\{ \|x\|_1 \mid Ax = b. \right\}$$

This problem can be formulated as a linear program as follows, and hence can be efficiently solved. Introduce n new variables $y_1, y_2, \ldots, y_n$ under the constraints

$$\min \left\{ \sum_i y_i \mid Ax = b, -y_i \leq x_i \leq y_i \right\}.$$

Definition 10.7. We call a matrix A as *BP-exact* if for all $b = Ax$ such that x^* is an r-sparse solution, x^* is also the unique solution to basis pursuit.

Call a distribution $\mathcal{D}$ over $k \times n$ matrices a **distributional JL family** if Lemma 10.2 is true when A is drawn from $\mathcal{D}$.

Theorem 10.8 (Donoho, Candes-Tao)**.** *If we pick $A \in \mathbb{R}^{k \times D}$ from a distributional JL family with $k \geq \Omega\left(r \log\left(\frac{D}{r}\right)\right)$, then with high probability A is BP-exact.*

We note that the $r \log \frac{D}{r}$ comes from $\log \binom{D}{r} \approx \log \left(\frac{D}{r}\right)^r = r \log \left(\frac{D}{r}\right)$. The last ingredient that one would use to show Theorem 10.8 is the *Restricted Isometry Property* (RIP) of such a matrix A.

Definition 10.9. A matrix A is (t, ε)-RIP if for all unit vectors x with $\|x\|_0 \leq t$, we have $\|Ax\|_2^2 \in [1 \pm \varepsilon]$.

See Chapter 4 of Ankur Moitra's book for more on compressed sensing, sparse recovery and basis pursuit. 10.8 comes from this paper by Emmanuel Candes and Terry Tao.

10.8 Some Facts about Balls in High-Dimensional Spaces

Consider the unit ball $\mathbb{B}_d := \{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1\}$. Here are two facts, whose proofs we sketch. These sketches can be made formal (since the approximations are almost the truth), but perhaps the style of arguments are more illuminating.

Theorem 10.10 (Heavy Shells). *At least $1-\varepsilon$ of the mass of the unit ball in $\mathbb{R}^d$ lies within a $\Theta(\frac{\log 1/\varepsilon}{d})$-width shell next to the surface.*

Proof. (Sketch) The volume of a radius-r ball in $\mathbb{R}^d$ goes as r^d, so the fraction of the volume *not* in the shell of width w is $(1-w)^d \approx e^{-wd}$, which is ε when $w \approx \frac{\log 1/\varepsilon}{d}$. □

Given any hyperplane $H = \{x \in \mathbb{R}^d \mid a \cdot x = b\}$ where $\|a\| = 1$, the width-w slab around it is $K = \{x \in \mathbb{R}^d \mid b - w \leq a \cdot x \leq b + w\}$.

Theorem 10.11 (Heavy Slabs). *At least $(1-\varepsilon)$ of the mass of the unit ball in $\mathbb{R}^d$ lies within $\Theta(1/\sqrt{d})$ slab around any hyperplane that passes through the origin.*

Proof. (Sketch) By spherical symmetry we can consider the hyperplane $\{x_1 = 0\}$. The volume of the ball within $\{-w \leq x_1 \leq w\}$ is at

$$\int_{y=0}^{w} (\sqrt{1-y^2})^{d-1} dy \approx \int_{y=0}^{w} e^{-y^2 \cdot \frac{d-1}{2}} dy.$$

If we define $\sigma^2 = \frac{4}{d-1}$, this is

$$\int_{y=0}^{w} e^{-\frac{y^2}{2\sigma^2}} dy \approx \Pr[G \leq w],$$

where $G \sim N(0, \sigma^2)$. But we know that $\Pr[G \geq w] \leq e^{-w^2/2\sigma^2}$ by our generic Chernoff bound for Gaussians (10.11). So setting that tail probability to be ε gives

$$w \approx \sqrt{2\sigma^2 \log(1/\varepsilon)} = O\Big(\sqrt{\frac{\log(1/\varepsilon)}{d}}\Big).$$

□

This may seem quite counter-intuitive: that 99% of the volume of the sphere is within $O(1/d)$ of the surface, yet 99% is within $O(1/\sqrt{d})$ of ***any*** central slab! This challenges our notion of the ball "looking like" the smooth circular object, and more like a very spiky sea-urchin. Finally, a last observation:

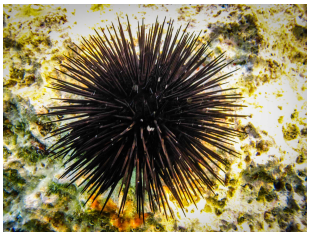

Figure 10.1: Sea Urchin (from uncommoncaribbean.com)

Corollary 10.12. *If we pick two random vectors from the surface of the unit ball in $\mathbb{R}^d$ (i.e., from the sphere), then they are nearly orthogonal with high probability. In particular, their dot-product is smaller than $O(\sqrt{\frac{\log(1/\varepsilon)}{d}})$ with probability $1 - \varepsilon$.*

Proof. Fix u. Then the dot-product $|u \cdot v| \leq w$ if v lies in the slab of width w around the hyperplane $\{x \cdot u = 0\}$. Now using Theorem 10.11 completes the argument. □

This means that if we pick n random vectors in $\mathbb{R}^d$, and set $\varepsilon = 1/n^2$, a union bound gives that all have dot-product $O(\sqrt{\frac{\log n}{d}})$. Setting this dot-product to ε gives us $n = \exp(\varepsilon^2 d)$ unit vectors with mutual dot-products at most ε, exactly as in the calculation at the beginning of the chapter.