8
Gaussians, Martingales, and Discrepancy

8.1 Introduction

In this lecture, we explore the interplay between Gaussian random variables, dimension reduction, martingale concentration, and their application to the problem of discrepancy minimization.

The plan for today is:

1. Basics about Gaussian random variables (r.v.s).
2. Dimension reduction via the Johnson-Lindenstrauss Lemma.
3. A recap of martingale concentration, particularly for Gaussians.
4. Discrepancy minimization using Random Walks, Martingales, and Gaussians.

8.2 Facts about Gaussian Random Variables

We start by defining the Gaussian distribution and recalling some fundamental properties.

Definition 8.1 (Gaussian (Normal) Distribution)**.** A random variable X follows a Gaussian (or Normal) distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if its probability density function (PDF) is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

If $\mu = 0$ and $\sigma^2 = 1$, it is called a **standard Gaussian**.

In this course, we often work with vectors of Gaussians.

Definition 8.2 (Multivariate Gaussian Distribution)**.** A random vector $\vec{X} \in \mathbb{R}^n$ follows a multivariate Gaussian distribution with mean

vector $\vec{\mu} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ (where Σ is positive definite), denoted $\vec{X} \sim N(\vec{\mu}, \Sigma)$, if its PDF is:

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right).$$

If $\vec{X}$ consists of n independent standard Gaussians, then $\vec{\mu} = 0$ and $\Sigma = I_n$ (the identity matrix).

- **Scaling:** If $X \sim N(\mu, \sigma^2)$, then $cX \sim N(c\mu, c^2\sigma^2)$.
- **Sums:** If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$.

Recall some fundamental properties of Gaussian (Normal) distributions.

- **Scaling:** If $X \sim N(\mu, \sigma^2)$, then $cX \sim N(c\mu, c^2\sigma^2)$.
- **Sums:** If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, then $\sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$.

A crucial property we will use extensively relates to projections of multivariate Gaussians.

Fact 8.3. Let $G_i \sim N(0,1)$ be independent standard Gaussians, and let $\vec{G} = (G_1, G_2, \ldots, G_n) \in \mathbb{R}^n$. For any fixed vector $x \in \mathbb{R}^n$, the inner product is distributed as:

$$\langle x, G \rangle \sim N\left(0, \sum x_i^2\right) = N(0, \|x\|^2).$$

If $\|x\| = 1$, then $\langle x, G \rangle \sim N(0,1)$.

8.2.1 Gaussians in Subspaces

We can also define Gaussian distributions constrained to a subspace.

Definition 8.4 (Gaussian from a Subspace)**.** Suppose V is a subspace of $\mathbb{R}^n$ of dimension $d \leq n$. Pick an orthonormal basis $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_d$ for V. We define a **Gaussian from subspace V**, denoted $g \sim N(V)$, as:

$$g = \sum_{i=1}^{d} g_i \vec{v}_i,$$

where $g_i \sim N(0,1)$ are independent.

Fact 8.5. Let $g \sim N(V)$. For any vector $x \in \mathbb{R}^n$:

$$\langle x, g \rangle \sim N(0, \sigma^2)$$

where $\sigma^2 = \|\text{Proj}_V(x)\|^2$. If $\|x\| = 1$, then $\sigma^2 \leq 1$.

8.3 Recap of Martingale Concentration

We recall the Azuma-Hoeffding inequality and its extensions.

Let $Z_1, Z_2, \ldots$ be independent r.v.s, and suppose X_i is a function of $Z_1, \ldots, Z_i$. If the sequence of differences behaves nicely, we have concentration. For example, if $X_i | Z_1, \ldots, Z_{i-1}$ behaves like a Rademacher variable (taking values ± 1 with probability $1/2$), Azuma-Hoeffding gives:

$$\Pr\left(\left|\sum_{i=1}^{T} X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2T}\right).$$

8.3.1 Gaussian Concentration for Martingales

This concentration extends naturally to Gaussian random variables.

Theorem 8.6 (Gaussian Concentration for Martingales)**.** *Suppose we have a martingale difference sequence such that $X_i | Z_1, \ldots, Z_{i-1}$ is Gaussian with mean 0.*

1. *If the conditional variance is 1, then:*

$$\Pr\left(\left|\sum_{i=1}^{T} X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2T}\right).$$

2. *More generally, if $X_i | Z_1, \ldots, Z_{i-1} \sim N(0, \sigma_i^2)$, then:*

$$\Pr\left(\left|\sum_{i=1}^{T} X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2\sum \sigma_i^2}\right).$$

Crucially, these concentration bounds also hold when T is a **stopping time** (not a fixed quantity).

Definition 8.7 (Stopping Time)**.** T is a stopping time with respect to a sequence $X_1, X_2, \ldots$ if the event $\{T = t\}$ depends only on the values of $X_1, \ldots, X_t$.

8.4 Discrepancy Minimization

We now turn to the main application: discrepancy minimization.

Definition 8.8 (Discrepancy)**.** Given a set system $\mathcal{S} = (S_1, S_2, \ldots, S_m)$ where each $S_i \subseteq [n] = \{1, \ldots, n\}$. A 2-coloring of $[n]$ is a map $\chi : [n] \to \{-1, 1\}$. The **discrepancy** of this coloring is

$$\operatorname{disc}(\chi) = \max_{i \in [m]} \left|\sum_{j \in S_i} \chi(j)\right|.$$

We want to find a coloring χ that minimizes the discrepancy (achieves good balance).

8.4.1 Randomized Coloring

Fact 8.9. Consider a simple randomized approach: set $\chi(j) \in \{-1, 1\}$ uniformly and independently at random. For any set S_i, $E\left[\sum_{j \in S_i} \chi(j)\right] = 0$. By Chernoff-Hoeffding bounds:

$$\Pr\left(\left|\sum_{j \in S_i} \chi(j)\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2n}\right).$$

By setting $\lambda = O(\sqrt{n \log m})$ and taking a union bound over all m sets, we find that the discrepancy is $\leq \lambda$ with high probability (w.h.p. $1 - 1/\text{poly}(m)$).

However, we can achieve tighter bounds.

Theorem 8.10 (Spencer's Theorem)**.** *There exists a coloring χ such that*

$$disc(\chi) \leq O\left(\sqrt{n \log(m/n)}\right).$$

If $m = n$, this guarantees $O(\sqrt{n})$ discrepancy. (This is often summarized as "six standard deviations suffice").

Spencer's original proof was non-constructive (using the entropy/pigeonhole principle on an exponentially large family). Bansal (2010) provided the first algorithmic proof using semidefinite programming and rounding. We will present a proof due to Lovett and Meka, which uses Linear Algebra, Gaussians, and Martingales.

8.4.2 Step 1: Relaxation to Partial Colorings

Instead of requiring $\chi : [n] \to \{-1, 1\}$, we consider a "convex" fractional relaxation, allowing $\chi : [n] \to [-1, 1]$.

Bad news: If we only minimize the fractional discrepancy, the problem is trivial: set $\chi(j) = 0$ for all j, achieving "zero fractional discrepancy".

Fix: We require that most variables are "close" to ± 1, allowing only a small fraction of variables to be far from $\{-1, 1\}$.

Lemma 8.11 (Partial Coloring Lemma (Lemma 1))**.** *Let $x_0 = 0$ be the starting point in $[-1, 1]^n$. We can find $x \in [-1, 1]^n$ such that:*

1. $\left|\sum_{j \in S_i} x_j\right| \leq \sqrt{|S_i|} \cdot \Delta + 1/poly(n)$ *for all i.*
2. $\#\{j \text{ s.t. } x_j \notin \{-1, 1\}\} \leq n/2$.

Here $\Delta = c\sqrt{\log(m/n)}$ for some constant c.

We will actually prove a more general statement involving arbitrary vectors, which implies Lemma 8.11.

Lemma 8.12 (Generalized Partial Coloring Lemma (Lemma 2)). *Given any vectors $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$, any starting point $x_0 \in [-1,1]^n$, and a small $\delta > 0$ (e.g., $1/poly(n)$). We can find $x \in [-1,1]^n$ such that:*

1. $|\langle a_i, x - x_0 \rangle| \leq \|a_i\|_2 \cdot \Delta$ *for all* i.
2. $\#\{j \text{ s.t. } x_j \in (-(1-\delta), 1-\delta)\} \leq n/2$.

Here $\Delta = c\sqrt{\log(m/n)}$.

To see that Lemma 8.12 implies Lemma 8.11, set a_i to be the indicator vector of set S_i (so $\|a_i\|_2 = \sqrt{|S_i|}$) and set $x_0 = 0$. We then take the solution given by Lemma 8.12 and round the variables close to ± 1 to exactly ± 1.

8.4.3 From Partial to Total Coloring

Before proving Lemma 8.12, let's see how Lemma 8.11 implies Spencer's Theorem. We use an iterative approach.

Start at $X_0 = 0$. Apply Lemma 8.11. This yields a set J of $\geq n/2$ variables colored $\{\pm 1\}$. The remaining $\leq n/2$ variables are fractional.

Freeze the variables in J. Consider the remaining variables $[n] \setminus J$. Start where the previous run stopped (using that configuration as the new X_0) and run Lemma 8.11 again on this smaller instance of size $\leq n/2$.

This finds a new set J' of $\geq \frac{1}{2}|[n] \setminus J|$ variables at ± 1. Repeat.

The net discrepancy may add up over the iterations. The total discrepancy is bounded by:

$$\begin{aligned}\chi(\mathcal{S}) &\leq c\sqrt{n\log(m/n)} + c\sqrt{\frac{n}{2}\log(m/(n/2))} + c\sqrt{\frac{n}{4}\log(m/(n/4))} + \ldots \\ &= c \cdot \sum_{i \geq 0} \sqrt{\frac{n}{2^i}\log\left(\frac{2^i m}{n}\right)} \\ &= O(\sqrt{n\log(m/n)}).\end{aligned}$$

The sum converges, dominated by the first term. This shows how to get Spencer's Theorem from Lemma 1 (and hence from Lemma 2).

8.5 Proof of Lemma 8.12: Gaussian Random Walks

How to prove Lemma 8.12? We use a beautiful algorithm utilizing ideas from:

- Random Walks
- Gaussians
- Martingales

Recall the goal (Lemma 8.12). WLOG, assume a_i are unit vectors, so we want $|\langle a_i, x - x_0 \rangle| \leq \Delta$. For convenience, we will prove that $\#\{j \text{ s.t. } x_j \in (-(1-\delta), 1-\delta)\} \leq 7n/10$ (instead of $\leq n/2$).

8.5.1 *The Algorithm Idea*

The idea is to start at x_0 and take tiny Gaussian steps. ($X^{t+1} = X^t +$ gaussian).

If we just do this, it is not great. We need to maintain constraints:

1. Coordinate bounds: $x_j \in [-1, 1]$ for all j.

2. Discrepancy bounds: $\langle a_i, x - x_0 \rangle \in [-\Delta, \Delta]$ for all i.

The Key Idea: If we get close to violating some constraint, we "freeze" the solution, forcing subsequent steps to lie in a subspace orthogonal to that constraint.

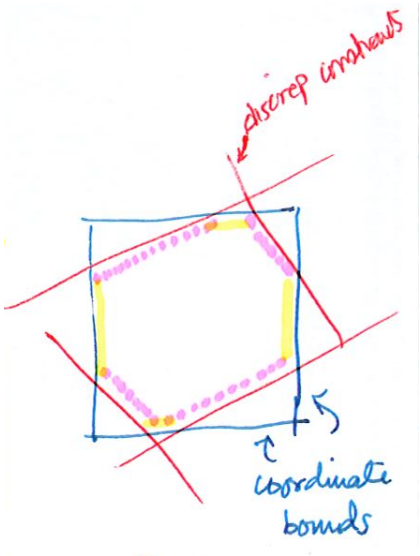

Figure 8.1: Feasible Region for the Random Walk.

We define when a variable or a constraint is close to being violated.

Definition 8.13. A variable j is **frozen** if $|x_j| > 1 - \delta$. A constraint a_i is **dangerous** if $|\langle a_i, x - x_0 \rangle| > \Delta - \delta$.

Intuition: The variable bounds are often much closer to the origin than the discrepancy bounds. So we expect the random walk to hit a variable bound earlier, meaning we will freeze many more variables than constraints become dangerous.

8.5.2 *The Algorithm Details*

At time t, we have a solution x^t. Let $F^t = \{j : |x_j^t| \geq 1 - \delta\}$ be the set of frozen variables. Let $D^t = \{i : |\langle a_i, x^t - x^0 \rangle| \geq \Delta - \delta\}$ be the set of dangerous constraints.

We define the subspace V_t where the next step must lie.

Definition 8.14. Let V_t be the subspace orthogonal to all frozen variables and all dangerous constraints.

$$V_t = \left(\text{span}\left(\{e_j : j \in F^t\} \cup \{a_i : i \in D^t\}\right)\right)^{\perp}.$$

The algorithm proceeds as follows. We set the step size $\epsilon \leq \frac{\delta}{10\sqrt{\log(mn)}}$.

Algorithm 3: Lovett-Meka Algorithm

2.1 **while** $\dim(V_t) \geq n/2$ **do**
2.2 (This means we have few frozen/dangerous constraints/variables.)
2.3 Pick $g_t \sim N(V_t)$ (a Gaussian from that subspace).
2.4 $x^{t+1} \leftarrow x^t + \epsilon \cdot g_t$.

The algorithm stops when $\dim(V_t) < n/2$. This means we have at least $n/2$ frozen variables or dangerous constraints in total (or more precisely, the dimension spanned by them is $> n/2$). We need to show that most of these are frozen variables.

8.5.3 *Analysis of the Algorithm*

What could go wrong with the algorithm?

1. The solution x^t "jumps" outside the feasible region ($[-1, 1]^n$ or the discrepancy bounds).
2. The algorithm stops (when $\dim(V_t) < n/2$), but very few variables are frozen (mostly dangerous constraints).

Let's address these concerns.

1. Staying Feasible. We know that x^{t-1} was "good". For non-frozen variables, $|x_j^{t-1}| \leq 1 - \delta$. For non-dangerous constraints, $|\langle a_i, x^{t-1} \rangle| \leq \Delta - \delta$.

For x^t to go outside the feasible region, the step must be large: $|\epsilon g^t| \geq \delta$.

We analyze the probability of a large Gaussian step. Since the components of g^t have variance ≤ 1:

$$\Pr(|\epsilon g^t| \geq \delta) \leq 2\exp\left(-\frac{(\delta/\epsilon)^2}{2}\right).$$

We chose ϵ such that $\delta/\epsilon \geq \sqrt{\log(mn)}$.

$$\Pr(|\epsilon g^t| \geq \delta) \leq 2\exp\left(-\frac{100\log(mn)}{2}\right) = \frac{2}{(mn)^{50}}.$$

We can take a union bound over all $m+n$ constraints and over all time steps, provided the total number of steps T is less than $(mn)^{49}$. We will actually prove that $T = O(1/\epsilon^2) = O(\text{poly}(m,n))$ many steps.

2. How many steps? We analyze the expected progress of the algorithm using the ℓ_2 norm.

Fact 8.15. $E\left[\|x^T - x^0\|^2\right] = \sum_{t<T} \epsilon^2 \cdot E[\dim(V_t)]$.

Proof.

$$\begin{aligned} E\left[\|x^{t+1} - x^0\|^2\right] &= E\left[\|x^t - x^0 + \epsilon g_t\|^2\right] \\ &= E\left[\|x^t - x^0\|^2\right] + 2\epsilon E\left[\langle g_t, x^t - x^0\rangle\right] + \epsilon^2 E\left[\|g_t\|^2\right]. \end{aligned}$$

The middle term is o because g_t is independent of x^t (conditioned on V_t) and $E[g_t] = 0$ (by symmetry). The last term $E\left[\|g_t\|^2 \mid V_t\right]$ is the dimension of the subspace V_t. So, $E\left[\|x^{t+1} - x^0\|^2\right] = E\left[\|x^t - x^0\|^2\right] + \epsilon^2 \cdot E[\dim(V_t)]$. The result follows by induction. □

Since the algorithm runs as long as $\dim(V_t) \geq n/2$, we have:

$$E\left[\|x^T - x^0\|^2\right] \geq \sum_t \epsilon^2 \cdot \frac{n}{2} = T \cdot \frac{n}{2}\epsilon^2.$$

On the other hand, since x^T must remain in $[-1,1]^n$, we must have $\|x^T - x^0\|^2 \leq O(n)$.

So, $T\frac{n}{2}\epsilon^2 \leq O(n)$. This implies $T \leq O(1/\epsilon^2)$. The algorithm is very likely to stop after $O(1/\epsilon^2) = O(\text{poly}(n,m))$ steps.

3. What happens at the stopping time? We need to analyze how many frozen vs dangerous constraints we have. We want to bound the number of dangerous constraints.

Let's see what the probability is that a specific constraint i (say a_i) becomes dangerous. Let $Y_t = \langle a_i, x^t - x^0\rangle$. We want $\Pr(|Y_T| > \Delta - \delta)$.

Y_T is a martingale (it is the sum of the noise contributions $\epsilon\langle a_i, g_t\rangle$). Let $Z_t = \epsilon\langle a_i, g_t\rangle$. Z_t is Gaussian $N(0, \sigma_t^2)$ where $\sigma_t^2 \leq \epsilon^2$ (since a_i is a unit vector and g_t is a Gaussian in a subspace).

We use the Gaussian concentration for martingales, noting that T is a stopping time.

$$\begin{aligned} \Pr(|Y_T| > \Delta - \delta) &\leq 2\exp\left(-\frac{(\Delta-\delta)^2}{2\sum E[\sigma_t^2]}\right) \\ &\leq 2\exp\left(-\frac{(\Delta-\delta)^2}{2T\epsilon^2}\right). \end{aligned}$$

We know $T\epsilon^2 = O(1)$. We set $\Delta = O(\sqrt{\log(m/n)})$.

$$\Pr(\text{constraint } i \text{ dangerous}) \leq 2\exp(-O(\Delta^2)) = 2\exp(-O(\log(m/n))) \leq \frac{n}{10m}.$$

(By choosing the constant c in Δ large enough).

Now we can calculate the expected number of dangerous constraints:

$$E[\#\text{dangerous constraints}] = \sum_{i=1}^{m} \Pr(\text{constraint } i \text{ dangerous}) \leq m \cdot \frac{n}{10m} = n/10.$$

By Markov's inequality,

$$\Pr(\#\text{dangerous} \geq n/5) \leq 1/2.$$

When the algorithm stops, we have $\dim(V_t) < n/2$. This implies that the span of frozen variables and dangerous constraints has dimension $> n/2$. With probability $\geq 1/2$, we have #Dangerous $< n/5$. In this case, the number of frozen variables must be large enough to account for the remaining dimension (e.g., #Frozen $\geq n/2 - n/5 = 3n/10$).

This successfully proves Lemma 8.12 (except that we proved $\#\{j \text{ s.t. } x_j \in (-(1-\delta), 1-\delta)\} \leq 7n/10$ instead of $\leq n/2$.)).

8.6 To wrap up

Today we saw Gaussian RVs and their use for:

- Dimension reduction (for distance preservation).
- Discrepancy minimization.

Along the way, we needed:

- Concentration bounds for sums of squares of (independent) Gaussians (Chi-squared distribution).
- Concentration bounds for Gaussians (but using martingale techniques).
- Random walks with Gaussians.

These ideas are simple but very powerful! The dimension reduction techniques (approximate) are useful in various surprising contexts, such as Compressive Sensing (also the "single-pixel camera").

8.7 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 $k = O(\log n/\varepsilon^2)$.

8.8 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 8.16. *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 $\mathrm{poly}(n, D, 1/\varepsilon)$ 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 $\Omega\left(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon}\right)$, and then Kasper Green Larson and

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)

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.

Larson and Nelson (2017)

8.9 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 8.17. *[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 8.16, 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

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

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.

8.10 Intuition for the Distributional JL Lemma

Let us recall some basic facts about Gaussian 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 8.18. *If $G_1 \sim N(\mu_1, \sigma_1^2)$ and $G_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then for $c \in \mathbb{R}$,*

$$c\,G_1 \sim N(c\mu_1, c^2\,\sigma_1^2) \tag{8.1}$$

$$G_1 + G_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{8.2}$$

Now, here's the main idea in the proof of Lemma 8.17. 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

$$1/\sqrt{k} \cdot N(0,1) = N(0, 1/k) \qquad \text{(by (8.1))}.$$

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

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

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

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.

1. First we give a proof via a direct calculation: it has several steps, but each step is elementary, and you are mostly following your nose.

2. The second proof uses the notion of sub-Gaussian random variables from , and builds some general machinery for concentration bounds.

8.11 A Direct Proof of Lemma 8.17

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. First, let's understand what the expected length of $A(x)$ is, and then we will show concentration about the mean.

Lemma 8.19. *Suppose the entries of M are independent random variables, with mean zero and unit variance. Then for unit vector $x \in \mathbb{R}^D$,*

$$\mathbb{E}[\|A(x)\|^2] = \|x\|^2.$$

Proof. Each entry of the vector Mx is the inner product of x with a vector with independent zero mean and unit variance random variables, and so is itself a random variable with zero mean and variance $\sum_i x_i^2 = 1$. This means that for any entry $i \in [k]$,

$$\mathbb{E}[(Mx)_i^2] = \mathrm{Var}(Mx) + \mathbb{E}[(Mx)_i]^2 = 1.$$

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

Observe that did not use the fact that the matrix entries were Gaussians. We will use it for the concentration bound, which we show next.

8.11.1 Concentration about the Mean

Using that each entry of M is an independent $N(0,1)$ r.v., we can use Proposition 8.18 to infer that $(Mx)_i \sim N(0, x_1^2 + x_2^2 + \ldots + x_D^2) = N(0,1)$. So, each of the k coordinates of Mx behaves just like an independent Gaussian! For brevity, define

$$Z := \|A(z)\|^2 = \sum_{i=1}^{k} \frac{1}{k} (Mx)_i^2,$$

so Z is the *average* of the *squares* of a collection of k independent $N(0,1)$ r.v.s.

Next we show that Z does not deviate too much from 1. Since 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:

$$\Pr[Z \geq 1+\varepsilon] \leq \Pr[e^{tkZ} \geq e^{tk(1+\varepsilon)}] \leq \mathbb{E}[e^{tkZ}]/e^{tk(1+\varepsilon)} \tag{8.3}$$

$$= \prod_i \left(\mathbb{E}[e^{tG^2}]/e^{t(1+\varepsilon)} \right) \tag{8.4}$$

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 even if you don't recall these facts, and don't have internet connectivity and cannot check Wikipedia, it is not that difficult to prove from scratch.

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

$$\frac{1}{\sqrt{2\pi}} \int_{g \in \mathbb{R}} e^{tg^2} e^{-g^2/2} dg = \frac{1}{\sqrt{2\pi}} \int_{z \in \mathbb{R}} e^{-z^2/2} \frac{dz}{\sqrt{1-2t}} = \frac{1}{\sqrt{1-2t}}. \tag{8.5}$$

Plugging back into (8.4), the bound on the upper tail shows that for all $t \in (0, 1/2)$,

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

Let's just focus on part of this expression:

$$\left(\frac{1}{e^t \sqrt{1-2t}} \right) = \exp\left(-t - \frac{1}{2} \log(1-2t) \right) \tag{8.6}$$

$$= \exp\left((2t)^2/4 + (2t)^3/6 + \cdots \right) \tag{8.7}$$

$$\leq \exp\left(t^2(1 + 2t + 2t^2 + \cdots) \right) \tag{8.8}$$

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

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 8.17.

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)

8.12 *Introducing Subgaussian Random Variables*

It turns out that the proof of Lemma 8.17 is a bit cleaner (with fewer calculations) if we use the abstraction provided by the generic Chernoff bound from last lecture, and the notion of subGaussian random variables which we introduce next. This abstraction will also allow us to extend the result to JL matrices having i.i.d. entries from other distributions, e.g., where each $M_{ij} \in_R \{-1, +1\}$.

8.12.1 *Subgaussian Random Variables*

Recall the definitions of the log-MGF $\psi(t)$ and its Legendre-Fenchel dual $\psi^*(\lambda)$ from §**??**.

Definition 8.20. A random variable V with mean 0 is *subgaussian with parameter* σ if its log-MGF $\psi(t)$ satisfies

$$\psi(t) \leq \frac{\sigma^2 t^2}{2}.$$

for all $t \geq 0$. It is *subgaussian with parameter* σ *up to* t_0 if the above inequality holds for all $|t| \leq t_0$.

In other words, the log-MGF of a subgaussian r.v. is bounded above by that of a Gaussian! At this point, it's useful to recall a fact we asked as an exercise in §**??**:

Fact 8.21. If $\psi_1(t) \geq \psi_2(t)$ for all $t \geq 0$, then $\psi_1^*(\lambda) \leq \psi_2^*(\lambda)$ for all λ.

Using this, the dual function of a subgaussian random variable with parameter σ is bounded *below* by that of a Gaussian $N(0, \sigma^2)$,

which means we have a tighter upper tail bound! Indeed, combining with (**??**), we immediately get:

Theorem 8.22 (Subgaussian Tail Bounds)**.** *If V is zero-mean and subgaussian with parameter σ, then*

$$\Pr[V \geq \lambda] \leq e^{-\lambda^2/(2\sigma^2)}.$$

Most tail bounds you will prove using the subgaussian perspective will come down to showing that some random variable is subgaussian with parameter σ, whereupon you can use Theorem 8.22. Given that you will often reason about sums of subgaussians, you may use the next fact, which is an analog of Proposition 8.18.

Lemma 8.23. *If $V_1, V_2, \ldots$ are independent, zero-mean 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. Using independence and the definition of subgaussian-ness:

$$\mathbb{E}[e^{tV}] = \mathbb{E}[e^{t\sum_i x_i V_i}] = \prod_i \mathbb{E}[e^{tx_i V_i}] \leq \prod_i e^{(tx_i)^2\sigma_i^2/2}.$$

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

8.12.2 *A Couple of Examples*

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

$$\begin{aligned}\mathbb{E}[e^{t(V-\mu)}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{x\in\mathbb{R}} e^{tx} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{t^2\sigma^2/2} \int_{x\in\mathbb{R}} e^{-\frac{(x-t\sigma^2)^2}{2\sigma^2}} dx = e^{t^2\sigma^2/2}. \qquad (8.9)\end{aligned}$$

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,

$$\Pr[V - \mu \geq \lambda] \leq e^{-\frac{\lambda^2}{2\sigma^2}}. \qquad (8.10)$$

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.

8.13 A Proof of Lemma 8.17 using Subgaussian r.v.s

Suppose we choose each M_{ij} to be an independent copy of a subgaussian r.v. with zero mean and unit variance, and let $A(x) = \frac{1}{\sqrt{k}}Mx$ again? We want to show that

$$Z := \|A(x)\|^2 = \frac{1}{k}\sum_{i=1}^{k}(Mx)_i^2 \tag{8.11}$$

has mean $\|x\|^2$, and is concentrated sharply around that value. Conveniently, we had only used the mean and variance of the entries of M in proving Lemma 8.19, so we can still infer that

$$\mathbb{E}[Z] = \mathbb{E}[\|A(x)\|^2 = \|x\| = 1.$$

It just remains to show the concentration.

8.13.1 Sums of Squares of Subgaussians

To add in. Until then see the explanaton in Matousek's paper "On Variants of the Johnson-Lindenstrauss Lemma".

8.13.2 Relating Subgaussian to Gaussians

If you have done the proof for the Gaussian case, and just want to extend the JL Lemma to other subgaussian random variables, you need not do all the work in §8.13.1. Instead you can relate subgaussian concentration to good old Gaussian concentration.

Indeed, the direct proof from §8.11 showed the $(Mx)_i$s were themselves Gaussian with variance $\|x\|^2$. Since the Rademachers are 1-subgaussian, Lemma 8.23 shows that $(Mx)_i$ is subgaussian with parameter $\|x\|^2$. Next, we need to consider Z, which is the average of *squares of* k independent $(Mx)_i$s. The following lemma shows that the MGF of squares of *symmetric* σ-subgaussians are bounded above by the corresponding 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 8.24. *If V is symmetric mean-zero σ-subgaussian r.v., and $W \sim N(0,\sigma^2)$, then $\mathbb{E}[e^{tV^2}] \le \mathbb{E}[e^{tW^2}]$ for $t > 0$.*

Proof. Using the calculation in (8.9) 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!) 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

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

Excellent. Now the bound on the upper tail for sums of squares of symmetric mean-zero σ-subgaussians follows from that of Gaussians. The lower tail (which requires us to bound $\mathbb{E}[e^{tV^2}]$ for $t < 0$) needs one more idea: suppose V is a mean-zero σ-subgaussian with parameter $\sigma^2 = 1$, and suppose $|t| < 1$. A Taylor expansion shows that

$$\mathbb{E}[e^{tV^2}] \leq 1 + t\mathbb{E}[V^2] + t^2 \sum_{i \geq 2} \mathbb{E}[V^{2i}/i!].$$

Since $E[V^2] = 1$ and $|t| < 1$, this is at most $1 + t + t^2\mathbb{E}[e^{V^2}]$. Now use the above bound $\mathbb{E}[e^{V^2}] \leq \mathbb{E}[e^{W^2}]$ to get that $E[e^{tV^2}] \leq 1 + t + t^2/\sqrt{1-2t}$, and the proof proceeds as for the Gaussian case.

In summary, we get the same tail bounds as in §8.11.1, 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 §8.11.1 since the linear terms in (8.6) don't cancel any more, see works by Indyk and Naor or Matousek for details.

Indyk and Naor (2008)
Matoušek (2008)

8.13.3 *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 randomness, 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)

8.14 *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 s-sparse signal for a vector with at most s nonzero entries, i.e., with $|\operatorname{supp}(x)| \leq s$.

It is common to use the notation $\|x\|_0 := |\operatorname{supp}(x)|$, even though this is not a norm.

Formally, x is a n-dimensional s-sparse vector, and a measurement of x with respect to a vector a is a real number given by $\langle a, x \rangle$. If we ask k questions, this gives us a $k \times n$ *sensing matrix* A (whose rows are the measurements), and a k-dimensional vector b of results. We want to reconstruct x with s nonzero entries satisfying $Ax = b$. This is often written as

$$\min \left\{ \|x\|_0 \mid Ax = b \right\}. \tag{8.12}$$

8.14.1 *Sparse Recovery: A First Attempt*

What properties would we like from our sensing matrix A? The first would be some form of consistency: that the problem should be solvable.

Definition 8.25 (Kruskal Rank)**.** An $m \times n$ matrix A has Kruskal rank r if every subset of r of its columns are linearly independent.

Lemma 8.26 (Unique Decoding)**.** *If A has Kruskal rank $\geq 2s$, then for any b we have $Ax = b$ for at most one s-sparse x.*

Proof. Suppose $Ax = Ax'$ for two s-sparse vectors x, x'. Then $A(x - x') = 0$ for the $2s$-sparse vector $x - x'$. The Kruskal rank being $2s$ means this vector $x - x' = 0$, and hence $x = x'$. □

So we can just find some sensing matrix with large Kruskal rank Give examples here and ensure our results will be unique. The next question is: how fast can we find x? (We should also be worried about noise in the measurements.) A generic construction of matrices with large Kruskal rank may not give us efficient solutions to (8.12). Indeed, it turns out that the problem as formulated is NP-hard, assuming A and b are contrived by an adversary.

Of course, asking to solve (8.12) for general A, b is a more difficult problem than we need to solve. In our setting, we can choose A as we like and then are given $b = Ax$, so we can ask whether there are matrices A for which this decoding process is indeed efficient. This is precisely what we do next.

8.14.2 The Basis Pursuit Algorithm

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

$$\min \Big\{ \|x\|_1 \mid Ax = b \Big\}. \tag{8.13}$$

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

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

Definition 8.27. We call a matrix A as *BP-exact for sparsity s* if for all vectors b such that the non-convex program (8.12) has a unique solution x^*, this vector x^* is also the unique optimal solution to the basis pursuit LP (8.13).

In other words, we want a matrix A for which the two programs return the same optimal solution. But do BP-exact matrices exist? If so, how do we efficiently construct them? Our next ingredient will be crucial to show their existence and construction.

Definition 8.28 (Restricted Isometry Property (RIP))**.** A matrix A is (t, ε)-RIP if for all unit vectors x with $\|x\|_0 \leq t$, we have

$$\|Ax\|_2^2 \in [1 \pm \varepsilon].$$

Lemma 8.29 (RIP $\implies$ BP-exact)**.** *If a matrix A is $(3s, \varepsilon)$-RIP for some $\varepsilon \leq 1/9$, then A is BP-exact for sparsity s.*

Proof. Suppose x^* is the unique solution to (8.12) and x the solution to (8.13), so that

$$\|x\|_1 \leq \|x^*\|_1. \tag{8.14}$$

Suppose $x - x^* = \Delta \neq 0$; hence $A\Delta = A(x - x^*) = 0$. If we could somehow show that $\text{supp}(\Delta) \leq 3s$, then using the RIP property for A, we would get

$$0 = \|A\Delta\|_2 \geq (1 - \varepsilon)\|\Delta\|_2 > 0,$$

a contradiction. But of course, Δ could have large support, so we need to work harder. The actual proof breaks up Δ into small pieces (so that the RIP matrix A maintains their length), and argues that there is one large piece that the other pieces cannot cancel out.

Let $S := \text{supp}(x^*)$ be the support of x^*, and $\overline{S}$ be the remaining coordinates. Let's sort these coordinates in decreasing order of their *absolute value*, and group them into buckets of $2s$ consecutive coordinates. Call these buckets $B_1, B_2, \ldots$.

For vector $v \in \mathbb{R}^n$ and subset $T \subseteq [n]$, define vector $v_T \in \mathbb{R}^n$ which agrees with v on the coordinates in S, and which has zeroes elsewhere.

Claim 8.30. $\sum_{j\geq 2} \|\Delta_{B_j}\|_2 \leq \|\Delta_S\|_2/\sqrt{2}$.

Before we prove the claim, let's see how to use it. The claim says that total Euclidean length of the vectors $\{v_{B_j}\}_{j\geq 2}$ is a constant factor smaller than that of $v_{S\cup B_1}$. So even after the near-isometric mapping A, the lengths of the former would not be able to cancel the length of the latter. Formally:

$$\begin{aligned} 0 = \|A\Delta\|_2 &\geq \|A\Delta_{S\cup B_1}\|_2 - \sum_{j\geq 2} \|A\Delta_{B_j}\|_2 \\ &\geq (1-\varepsilon)\,\|\Delta_{S\cup B_1}\|_2 - (1+\varepsilon)\sum_{j\geq 2}\|\Delta_{B_j}\|_2 \\ &\geq (1-\varepsilon)\,\|\Delta_S\|_2 - \frac{1+\varepsilon}{\sqrt{2}}\,\|\Delta_S\|_2 \ , \end{aligned}$$

where the first step uses the triangle inequality for norms, the second uses that each $\Delta_{S\cup B_1}$ and Δ_{B_j} are $3s$-sparse, and the last step uses $\|\Delta_{S\cup B_1}\|_2 \geq \|\Delta_S\|_2$ and also Claim 8.30. Finally, since $\varepsilon \leq 1/9$, we have $1-\varepsilon > \frac{1+\varepsilon}{\sqrt{2}}$, so the only remaining possibility is that $\Delta_S = 0$. The next claim implies that $\Delta_S = 0$ implies that $\Delta = 0$, giving a contradiction and hence the proof of Lemma 8.29.

Claim 8.31. $\|\Delta_S\|_1 \geq \|\Delta_{\overline{S}}\|_1$.

Proof. We finally use that $x = x^* + \Delta$ is the optimizer for the LP, which means

$$\begin{aligned} \|x^*\|_1 > \|x^*+\Delta\|_1 &= \|x^*_S + \Delta_S\|_1 + \|\Delta_{\overline{S}}\|_1 \\ &\geq \|x^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{\overline{S}}\|_1. \end{aligned}$$

(The last step uses the triangle inequality.) Since $\|x^*\|_1 = \|x^*_S\|_1$, we get Claim 8.31. □

The final piece of the argument is to prove Claim 8.30:

Proof of Claim 8.30. Take any bucket B_j for $j \geq 2$. Each entry of Δ in this bucket is smaller than the smallest entry of B_{j-1}, and hence smaller than the average entry of B_{j-1}. And there are $2s$ entries in this bucket B_j, so the Euclidean length of the bucket is

$$\|\Delta_{B_j}\|_2 \leq \sqrt{2s}\cdot\frac{\|\Delta_{B_{j-1}}\|_1}{2s} = \frac{\|\Delta_{B_{j-1}}\|_1}{\sqrt{2s}}.$$

Summing this over all $j \geq 2$, we get

$$\sum_{j\geq 2}\|\Delta_{B_j}\|_2 \leq \sum_{j\geq 2}\frac{\|\Delta_{B_{j-1}}\|_1}{\sqrt{2s}} = \frac{\|\Delta_{\overline{S}}\|_1}{\sqrt{2s}}.$$

Now $\|\Delta_{\overline{S}}\|_1 \leq \|\Delta_S\|_1$ by Claim 8.31. And finally, since the support of Δ_S is of size s, we can bound its ℓ_1 length by $\sqrt{s}$ times its ℓ_2 length, finishing the claim. (Since we wanted that factor of $\sqrt{2}$ in the denominator, we made the buckets slightly larger than the size of S.) □

Exercise:: for any vector $v \in \mathbb{R}^d$, show that $\|v\|_1 \leq \sqrt{\mathrm{supp}(v)}\cdot\|v\|_2$.

This completes the proof for Lemma 8.29. □

Finally, how do we construct RIP matrices? Call a distribution $\mathcal{D}$ over $k \times n$ matrices a *distributional JL family* if Lemma 8.17 is true when A is drawn from $\mathcal{D}$. The following theorem was proved by David Donoho, and by Emanuel Candes and Terry Tao, and by Mark Rudelson and Roman Vershynin. (The connection of their constuction to the distributional JL was made explicit by Baraniuk et al.)

Theorem 8.32 (JL $\implies$ RIP)**.** *If we pick $A \in \mathbb{R}^{k \times n}$ from a distributional JL family with $k \geq \Omega(s \log n/s)$, then with high probability A is BP-exact.*

Proof. The proof is simple, but uses some fairly general ideas worth emphasizing. First, focus on some s-dimensional subspace of $\mathbb{R}^n$ (obtained by restricting to some subset of coordinates). For notational simplicity, we just identify this subspace with $\mathbb{R}^s$.

1. For $\delta = \varepsilon/3$, pick an δ-net N of the sphere S^{s-1} (under Euclidean distances). This can be done by a greedy algorithm: if some point x does not satisfy the covering property at any time, it can be added to the net. We claim the size of the net is $|N| := (4/\delta)^s$. Indeed, define balls of radius $\delta/2$ around the points in N; these are disjoint by the packing property of nets, and are all contained in a ball of radius $1 + \delta$ around the origin. Since the volume of balls of radius r scales as r^s, we have

   $$|N| \leq \left(\frac{1+\delta}{\delta/2}\right)^s = (4/\delta)^s.$$

   Given a metric space (X, d), a *δ-net* is a subset $N \subseteq X$ such that (i) $d(x, y) \geq \delta$ for all $x, y \in N$, and (ii) for each $x \in X$ there exists $y \in N$ such that $d(x, y) \leq \delta$. The former is call the *packing property* and the latter the *covering property* of nets.

2. If A is an δ-isometry on the δ-net $N \subseteq S^{s-1}$, we claim it is a 3δ-isometry on all of S^{s-1}. Indeed, consider the point x that achives the maximum stretch $\arg\max\{\|Ax\|_2 \mid x \in S^{s-1}\}$, and let this stretch be M. Let y be the closest point in N to x; by the packing property $\|x - y\| \leq \delta$. Then $M = \|Ax\| \leq \|Ay\| + \|A(x - y)\| \leq (1 + \delta) + M\delta$. Rearranging, $M \leq \frac{1+\delta}{1-\delta} \leq (1 + 3\delta)$ for $\delta \leq 1/3$, say. For the contraction, consider any $x \in S^{s-1}$, with closest net point y. Then $\|Ax\| \geq \|Ay\| - \|A(x - y)\| \geq 1 - \delta - (1 + 3\delta)\delta \geq 1 - 3\delta$, again as long as $\delta \leq 1/3$.

3. By Lemma 8.17, the random matrix A with m rows is an δ-isometry on each point in the net N, except with probability $\exp(-c\delta^2 m)$ for some constant c.

4. Now apply the above argument to each of the $\binom{n}{s}$ subspaces obtained by restricting to some subset S of coordinates. By a union bound over all subsets S, and over all points in the net for that

subspace, the matrix A is an 3δ-isometry on all points with support in S except with probability

$$\binom{n}{s} \cdot (4/\delta)^s \cdot \exp(-c\delta^2 m) \leq \exp(-\Theta(m)),$$

as long as m is $\Omega(s \log n/s)$. Since $\varepsilon = 3\delta$, we have the proof.

□

This presentation is based on notes by Jirka Matoušek. Also see Chapter 4 of Ankur Moitra's book for more on compressed sensing, sparse recovery and basis pursuit.

8.15 *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 8.33 (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 8.34 (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 least

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

If we define $\sigma^2 = \frac{1}{d-1}$, 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 (8.10). So setting that tail probability to be ε gives

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

□

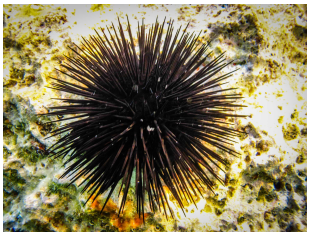

Figure 8.2: Sea Urchin (from uncommoncaribbean.com)

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:

Corollary 8.35 (Near-orthogonality)**.** *Two random vectors from the surface of the unit ball in $\mathbb{R}^d$ (i.e., from the sphere S^{d-1}) 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 one of the vectors u. Then for dot-product $|u \cdot v|$ to be at most ε, the other vector v must fall in the slab of width ε around the hyperplane $\{x \cdot u = 0\}$. Now Theorem 8.34 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.