1 Overview

In the last lecture we looked at $(\epsilon, \delta)$-DJL: There exists a distribution $\mathcal{D}_{\epsilon, \delta}$ over $\mathbb{R}^{m \times n}$ with $m = O(\epsilon^{-2} \log(1/\delta))$ such that $\forall x, \|x\| = 1$ we have that $P_{\Pi \sim \mathcal{D}_{\epsilon, \delta}}[|\|\Pi x\|_2^2 - 1| < max\{\epsilon, \epsilon^2\}] < \delta$. Moreover, we have seen metric JL which tells us that for all $X \subset l^2_2$ there exist a matrix $\Pi \in \mathbb{R}^{m \times n}$ with $m = \Theta(\epsilon^{-2} \log |X|)$ such that for all $x \in T$ we have that $|\|\Pi x\|_2^2 - 1| < \epsilon$, where $T = \{\frac{x-y}{\|x-y\|}, x, y \in X\}.$

2 Today

Today we are going to look at three main things:

- Johnson-Lindenstrauss lower bound
- Beyond worst-case analysis and more specifically what is called Gordon’s Theorem
- Bounding supremum of gaussian processes.

3 Main Section

3.1 Lower bounds on Dimensionality Reduction

3.1.1 Lower bounds on Distributional Johnson Lindenstrauss Lemma

The first lower bound was proved by Jayram and Woodruff [1] and then by Kane,Meka,Nelson [2]. The lower bound tells that any $(\epsilon, \delta)$-DJL for $\epsilon, \delta \in (0, 1/2)$ must have $m = \Omega(min\{n, \epsilon^{-2} \log(\frac{1}{\delta})\})$. The second proof builds on the following idea: Since for all $x$ we have the probabilistic guarantee $P_{\Pi \sim \mathcal{D}_{\epsilon, \delta}}[|\|\Pi x\|_2^2 - 1| < max\{\epsilon, \epsilon^2\}] < \delta$, then it is true also for any distribution over $x$. We are going to pick $x$ according to the uniform distribution over the sphere. Then this implies that there exists a matrix $\Pi$ such that $P_x[|\|\Pi x\|_2^2 - 1| < \epsilon] < \delta$. It is shown that this cannot happen for any fixed matrix $\Pi \in \mathbb{R}^{m \times n}$ unless $m$ satisfies the lower bound.

3.1.2 Johnson-Lindenstrauss lower bound for the linear mapping case

If we insist on linear maps, then a lower bound of Kasper and Nelson [3] tells us that $m$ must be at least $\Omega(min\{n, \epsilon^{-2} \log |X|\})$. The hard set is shown to exist via the probabilistic method,
and is constructed by taking the union of \( \{0, e_1, \ldots, e_n\} \) in \( \mathbb{R}^n \) together with plus sufficiently many random vectors. We just need to take a fine finite net that approximates all possible linear maps that work for the simplex (i.e. \( \{e_1, \ldots, e_n\} \)) and then argue that for each linear map in the net then with some good probability there exists a point from the random ones such that its length is not preserved by the map. So, approximating well enough the set of linear maps and adjusting the parameteres then we can take a union bound over all matrices in the net.

### 3.1.3 Johnson-Lindenstrauss lower bound for the general case

This lower bound has been found by Alon [4]. It tells that \( m \) must be at least \( \Omega(\min\{n, \epsilon^{-2} \log n / \log(1/\epsilon)\}) \)
to preserve distances between the set of points \( X = \{0, e_1, \ldots, e_n\} \). The hard set is the simplex \( X = \{0, e_1, \ldots, e_n\} \). Let \( f \) be the mapping. Translate the embedding so that \( f(0) = 0 \) (i.e. translate the embedding so that each point \( x \in X \) is actually mapped to \( f(x) - f(0) \); this does not affect any distances). Now write \( f(e_i) = v_i \). Then we have that \( \|v_i\| = 1 \pm \epsilon \) since \( f \) preserves the distance from \( e_i \) to 0. We also have for \( i \neq j \) that \( \|v_i - v_j\| = \sqrt{2}(1 \pm \epsilon) \). This implies since \( \|v_i - v_j\| = \|v_i\|^2 + \|v_j\|^2 - 2\langle v_i, v_j \rangle \) we get that \( \langle v_i, v_j \rangle = O(\epsilon) \). Setting \( w_i = \frac{v_i}{\|v_i\|} \) we have that \( \|w_i\| = 1 \) and \( \langle w_i, w_j \rangle = O(\epsilon) \). Let \( \Pi \) be the matrix that has \( w_i \) as columns. Then observe that \( A = \Pi^T\Pi \) is a matrix with 1 on the diagonal and elements with absolute value at most \( \epsilon \) everywhere.

We have that \( \text{rank}(A) = \text{rank}(\Pi^T\Pi) = \text{rank}(\Pi) \leq m \). We are going to use the following lemma which solves the problem for \( \epsilon = \frac{1}{\sqrt{n}} \) and then bootstrap it to work for all values of \( \epsilon \).

**Lemma 1.** Any real symmetric matrix that is \( \frac{1}{\sqrt{n}} \)-near to the identity matrix, i.e. its diagonals are 1 and off-diagonals are in \([-1/\sqrt{n}, 1/\sqrt{n}]\), must have \( \text{rank}(A) \geq \Omega(n) \).

**Proof.** Let \( \lambda_1, \ldots, \lambda_r \) be the non-zero eigenvalues of \( A \), where \( r = \text{rank}(A) \). By Cauchy-Schwarz, we have \( r \geq \left( \sum_{i=1}^{r} \lambda_i \right)^2 / n \). By standard linear algebra the numerator is the trace of \( A \) squared and the denominator is just the Frobenius norm of \( A \) squared. We have that \( \text{tr}(A) = n \) and \( \|A\|_F^2 \leq n + n(n - 1)\epsilon^2 \). Plugging everything into the inequality along with the fact that \( \epsilon = \frac{1}{\sqrt{n}} \) we get the desired result.

**Theorem 2.** Any real symmetric matrix that is \( \epsilon \)-near to the identity matrix must have \( \text{rank}(A) \leq \min\{n, \epsilon^{-2} \log n / \log \epsilon\} \).

**Proof.** Define the matrix \( A^{(k)} \) such that \( (A^{(k)})_{ij} = a_{ij}^k \). We will build our proof on the following claim: It holds that \( \text{rank}(A^{(k)}) \leq \left( \frac{r + k - 1}{k} \right) \) where \( r = \text{rank}(A) \). Assume that the claim is true we pick \( k \) to be the closest integer to \( n \epsilon^{-1} \sqrt{n} \). Thus \( k \leq \frac{1}{\sqrt{n}} \), so we have that \( \Omega(n) \leq r(A^{(k)}) \leq \left( \frac{r + k - 1}{k} \right) \). Using the fact that \( \binom{r}{k} \leq (ek^{-1})^k \) and walking thought the calculations we can get the desired result.

What remains is to prove the claim. Let \( t_1, \ldots, t_r \) be the row-space of \( A \). This means that \( \forall i \exists \beta \in \text{mathbbR}^r \) such that \( a_i = \sum_{q=1}^{r} \beta_q t_q \). Then observe that \( (A^{(k)})_{ij} = a_{ij}^k = (\sum_{q=1}^{r} \beta_q (t_q)_{ij})^k = \sum_{q_1, \ldots, q_k} \Pi_{z=1}^{k} \beta_{q_z} \Pi_{z=1}^{k} t_{q_z} \). It is easy to see that each vector of this form is a linear combination of
vectors of the form \((\Pi_{z=1}^{y=1}(t_q z_1), \Pi_{z=1}^{y=1}(t_q z_2), \ldots)\). where \(\sum d_z = k\). This is a standard combinatorics problem of putting \(r\) balls into bins \(k\) bins with repetition, so the answer is \(\binom{r+k-1}{k}\).

\[\]  

3.2 Beyond Worst Case Analysis

Given a subset \(T\) of the unit sphere- for example \(\{T = \frac{x-y}{\|x-y\|}, x, y \in X\}\) ideally we would like that \(\forall x \in T, ||\Pi x||^2 - 1| < \epsilon\). We want that \(\mathbb{P}_\Pi(\sup_{x \in T} ||\Pi x||^2 - 1 > \epsilon) < \delta\).

We are moving with Gordon’s Theorem [10] which was several times [5] [7] [9]. First of all we are going to define the gaussian mean width of the set.

**Definition 3.** *The Gaussian mean width of a set \(T\) is defined as* \(g(T) = \mathbb{E}_g \sup_{x \in T} \langle g, x \rangle\).*

Back to Gordon’s Theorem. Suppose that \(\Pi_{ij} = \frac{\delta_{ij}}{\sqrt{m}}\) for random signs, with \(m \geq \Omega(\epsilon^{-2}(g^2(T) + \log \frac{1}{\delta}))\). Then we have that \(\mathbb{P}_\Pi(\sup_{x \in T} ||\Pi x||^2 - 1 > \epsilon) < \delta\). Actually, we just need a distribution that decays as fast as a gaussian, has variance one and zero mean.

Let us give a simple example of the gaussian mean width. For example, if \(T\) is the simplex then we have that \(g(T) = \|g\|_\infty\) which is roughly equal to \(\log n\) by standard computations on gaussians. Actually, what Gordon’s theorem tells us is that if the vectors of \(T\) have a nice geometry then one can improve upon Johnson-Lindenstrauss: the more well-clustered the vectors are, the lower dimension you can achieve.

We continue with the following claim: \(\forall T\) which is a subset of the unit sphere, \(g(T) \leq O(\sqrt{\log N})\), where \(N = |T|\).

**Proof.** Define \(Z_i = |\langle g_i, x_i \rangle|\), where \(T = \{x_1, \ldots, x_N\}\). Then

\[
g(T) \leq \mathbb{E}_g \max\{Z_1, \ldots, Z_N\} = \int_0^\infty \mathbb{P}_g(\max\{Z_1, \ldots, Z_N\} > u) du =
\]

\[
= \int_0^{2\sqrt{\log n}} \mathbb{P}_g(\max\{Z_1, \ldots, Z_N\} > u) du + \int_{2\sqrt{\log n}}^{2\sqrt{\log n}} \mathbb{P}_g(\max\{Z_1, \ldots, Z_N\} > u) du \leq
\]

\[
\leq 2\sqrt{\log n} + \int_{2\sqrt{\log n}}^{\infty} \mathbb{P}_g(\exists Z_i \geq u) du \leq 2\sqrt{\log n} + \int_{2\sqrt{\log n}}^{\infty} \sum_i \mathbb{P}_g(Z_i \geq u) du \leq
\]

\[
2\sqrt{\log N} + \int_{2\sqrt{\log n}}^{\infty} Ne^{-u^2/2} du \leq 2\sqrt{\log n} + O(1)
\]

\[\]
3.2.1 How to bound $g(T)$

- $g(T) \leq \sqrt{\log |T|}$, as we just showed. In fact if every vector in $T$ has norm at most $\alpha$, then one gets $g(T) \preceq \alpha \sqrt{\log |T|}$.

- Let $T' \subset T$ be such that $\forall x \in T$, $\exists x' \in T'$ such that $\|x - x'\| \leq \varepsilon$. That is, $T'$ is an $\varepsilon$-net of $T$. This implies that $\langle g, x \rangle = \langle g, x' \rangle + (g, x - x') \leq \|g\|_2 \varepsilon$, which implies that $g(T) \leq g(T') + \varepsilon \|g\|_2^2 \leq g(T') + \varepsilon \sqrt{n} \leq O(\sqrt{\log |T'|}) + \varepsilon \sqrt{n}$. Thus if $T$ is covered well by a small net, one can get a better bound.

- Let $T \subset T_1 \subset T_2 \subset \ldots \subset T$, such that $T_r$ is a $(2^{-r})$-net of $T$ (we are assuming every vector in $T$ has at most unit norm). Then $x = x^{(0)} + (x^{(1)} - x^{(0)}) + (x^{(2)} - x^{(1)}) + \ldots$. Then we have

$$
\mathbb{E} \sup_{x \in T} \langle g, x \rangle \leq \mathbb{E} \sup_{x \in T} \langle g, x^{(0)} \rangle + \sum_{r=1}^{\infty} \mathbb{E} \sup_{x \in T} \langle g, x^{(r)} - x^{(r-1)} \rangle
\lesssim \log^{1/2} |T_0| + \sum_{r=1}^{\infty} (\sup_{x \in T} \|x^{(r)} - x^{(r-1)}\|) \cdot \log^{1/2}(|T_r| \cdot |T_{r-1}|)
\lesssim \log^{1/2} |T_0| + \sum_{r=1}^{\infty} \frac{1}{2^r} \cdot \log^{1/2} |T_r|.
$$

The last inequality holds since by the triangle inequality, $\|x^{(r)} - x^{(r-1)}\| \leq \|x^{(r)} - x\| + \|x^{(r-1)} - x\| \leq 3/2^{r-1}$. Furthermore, $|T_{r-1}| \leq |T_r|$, so $\log(|T_r| \cdot |T_{r-1}|) \leq \log(|T_r|^2) = 2 \log |T_r|$.

Thus $g(T) \leq \sum_{r=0}^{\infty} 2^{-r} \log^{1/2} |T_r| = \sum_{r=0}^{\infty} 2^{-r} N(T_2, \|\cdot\|_2, 2^{-r})$, where $N(T, d, \varepsilon)$ is the size of the best $\varepsilon$-net of $T$ under metric $d$. Bounding this sum by an integral, we have that $g(T)$ is at most a constant factor times $\int_0^{\infty} \log^{1/2} N(T, \|\cdot\|_2, u) du$. This inequality is called Dudley’s inequality.

- Write $S_0 \subset S_1 \subset \ldots \subset S$, such that $|S_0| = 1$ and $\|S_s\| \leq 2^{2s}$. One can show that the Dudley bound is in fact

$$
\inf_{\{S_s\}_{s=0}^{\infty}} \sum_{s=0}^{\infty} 2^{s/2} \sup_{x \in T} d_{\|\cdot\|_2}(x, S_s).
$$

Write

$$
\gamma_2(T) = \inf_{\{S_s\}_{s=0}^{\infty}} \sup_{x \in T} \sum_{s=0}^{\infty} 2^{s/2} d_{\|\cdot\|_2}(x, S_s).
$$

It was shown by Fernique [8] that $g(T) \preceq \gamma_2(T)$ for all $T$. Talagrand later showed that in [11] the lower bound is also true, and hence $g(T) = \Theta(\gamma_2(T))$; this is known as the “majorizing measures” theorem, which we will not prove in this class.

3.2.2 Johnson Lindenstrauss implies Gordon’s Theorem

What we already saw is that Gordon’s theorem implies the Johnson Lindenstrauss lemma. In fact this summer by Oymak, Recht and Soltanolkotabi [6] it was proved that with right parameteres the Distributional Johnson Lindenstrauss lemma implies Gordon’s theorem. Basically take a DJL $\varepsilon' = \frac{\varepsilon}{\gamma_2(T)}$ then for $m \geq \varepsilon^{-2}(\gamma_2(T) \log \frac{1}{\delta})$ (where we hide constants in the inequalities) we take
the guarantee for Gordon’s Theorem. Actually, their proof works by preserving the sets $S_s$ (plus differences and sums of vectors in these sets) at different scales. The result is not exactly optimal because it is known $m = O(e^{-2(\gamma_2^2(T) + \log(1/\delta))})$ suffices (see for example [9, 7]), but it provides a nice reduction from Gordon’s theorem to DJL.

References


