1 Overview

In the last lecture we discussed how distributional JL implies Gordon’s theorem, and began our discussion of sparse JL. We wrote $\|\Pi x\|^2 = \sigma^T A^T_x A_x \sigma$ and bounded the expression using Hanson-Wright in terms of the Frobenius norm.

In this lecture we’ll bound that Frobenius norm and then discuss applications to fast nearest neighbors.

2 Sparse JL from last time

Note that we defined $B_x = A^T_x A_x$ as the center of the product from before, but with the diagonals zeroed out. $B_x$ is a block-diagonal matrix with $m$ blocks $B_{x,1}, \ldots, B_{x,m}$ with

$$(B_{x,r})_{i,j} = \begin{cases} 0, & i = j \\ \delta_{r,i} \delta_{r,j} x_i x_j, & i \neq j. \end{cases}$$

2.1 Frobenius norm

We can write the Frobenius norm as

$$\|B_x\|_F^2 = \frac{1}{s^2} \sum_{r=1}^{m} \sum_{i \neq j} \delta_{r,i} \delta_{r,j} x_i^2 x_j^2$$

$$= \frac{1}{s^2} \sum_{i \neq j} (\sum_{r=1}^{m} \delta_{r,i} \delta_{r,j}) x_i^2 x_j^2$$

where we define the expression in the parentheses to be $Q_{ij}$.

Claim 1.

$$\|Q_{ij}\|_p \lesssim p$$

Let’s assume the claim and show that the Frobenius norm is correct.
\[ \|\|B_x\|_F\|_p = (\mathbb{E}[\|B_x\|_F]^p)^{1/p} \]
\[ = (((\mathbb{E}[\|B_x\|_F^2])^{2/p})^{2/p})^{1/2} \]
\[ = \|\|B_x\|_F^2\|_p^{1/2} \]
\[ \leq \|\|B_x\|_F^2\|_p^{1/2} \]
\[ = \left( \frac{1}{s^2} \sum_{i \neq j} x_i^2 x_j^2 Q_{ij} \right)^{1/2} \]
\[ \leq \frac{1}{s} \sum_{i \neq j} x_i^2 x_j^2 Q_{ij}^{1/2} \]
\[ \leq \sqrt{p} s \left( \sum_{i \neq j} x_i^2 x_j^2 \right)^{1/2} \]
\[ \leq \sqrt{p} s \]
\[ \approx \frac{\epsilon}{\sqrt{\ln 1/\delta}} \approx \frac{1}{\sqrt{m}} \]

Now, we can do the following:

\[ \|\|\Pi x\|_2^2 - 1\|_p = \|\sigma^T B_x \sigma\|_p \leq \sqrt{\frac{p}{m}} + \frac{p}{s} \]

(Markov) \[ \implies \quad \mathbb{P}(\|\|\Pi x\|_2^2 - 1\|_p > \epsilon) \leq \frac{\|\|\Pi x\|_2^2 - 1\|_p^p \epsilon}{\epsilon^p} \]
\[ \leq 2^p \cdot \left( \frac{\max\left( \frac{\sqrt{\frac{p}{m}}, \frac{p}{s}}{\epsilon} \right)}{\epsilon} \right)^p < \delta \]

Now we can prove the claim

**Proof.** Let’s just fix column \( i \). It has \( s \) nonzero elements somewhere. There’s another column \( j \), and the question is how many of the nonzero locations of \( i \) match with nonzero elements of \( j \). Let’s have \( Y_t \) be an indicator random variable for column \( j \) having a nonzero element in the \( t \)th nonzero row of \( i \) (note: this is not the \( t \)th row of all the elements). Then \( Q_{ij} = \sum_{t=1}^s Y_t \). If we had independence across the entries, this would just be a Chernoff bound. But we don’t, so it isn’t.

However, the moments are dominated by the independent case.

\[ \mathbb{E}[\sum_i Y_i]^p = \sum_{s=1}^{\min(p,s)} \sum_{d_1, d_2, \ldots, d_t} \sum_{d_j=p}^{i_1 < i_2 < \ldots < i_t} \mathbb{E}[\prod_{q=1}^s Y_{i_q}] \]

Remember that the expected value of any \( Y_t \) is \( s/n \). The product at the end is just \( (s/n)^t \) in the
independent case. In our case, it’s a conditional product:

$$\mathbb{E} \left[ \prod_{q=1}^{t} Y_q \right] = \mathbb{P}(Y_{i_1} = 1) \cdot \mathbb{P}(Y_{i_2} = 1 | Y_{i_1} = 1) \cdots$$

$$= \frac{s}{m} \cdot \frac{s-1}{m-1} \cdots \frac{s-l+1}{m-l+1}$$

$$\leq \left( \frac{s}{m} \right)^l$$

So the sum is actually dominated by the independent case, which can be handled via Bernstein’s inequality.

Note the runtime to apply the sparse JL map is $O(s \times \text{supp}(x))$

## 3 Fast JL Transform (FJLT)

Now we’ll use a different approach that’ll give $O(n \lg n)$ time, which is better in cases where $x$ is dense. This is due to Ailon & Chazelle ’09 [AC09]. It is called, as the section title suggests, the FJLT.

Here’s their definition of $\Pi$:

$$\Pi = \frac{1}{\sqrt{m}} \cdot PHD$$

where $P$ is an $m \times n$ sampling matrix (note that differs slightly from the paper to make the analysis easier). $H$ is $\sqrt{n}$ times an orthogonal $n \times n$ matrix, i.e. $H^T H = n \cdot I$. Also max $|H_{ij}| = O(1)$, and computing $Hx$ should be fast for any $x$. $D$ is an $n \times n$ diagonal matrix with random signs $\alpha_1, \ldots, \alpha_n$ along the diagonal.

Today we’ll let $P = S_\eta$ be an $n \times n$ diagonal matrix where the $i$th diagonal entry $\eta_i$ equals 1 with probability $m/n$ and 0 otherwise, and the $\eta_i$ are independent across $i$. Note that an example of $H$ could be the unnormalized discrete Fourier transform. Another possibility for $H$ is the unnormalized Hadamard matrix where $H_{i,j} = (-1)^{\langle i, j \rangle}$. Here $n$ is a power of 2 and we are interpreting $i, j$ as elements of $\mathbb{F}_{2^{\log_2 n}}$. Both of these matrices allow $Hx$ to be computed in time $O(n \log n)$. In general, $n \times n$ matrices $F$ which are orthogonal with max $|F_{i,j}| = O(1/\sqrt{n})$ are called bounded orthonormal systems.

Motivation: what if we just sampled coordinates from $x$? That would be $Px$; let $y = (1/\sqrt{m})Px$. Then

$$\mathbb{E} y_i^2 = \frac{\|x\|^2}{m} = \frac{1}{m} \implies \mathbb{E} \|y\|^2 = 1$$

Note that the expected value is good, but the variance is pretty bad: what if all the mass of $x$ is at a single index? We can take intuition from the Heisenberg uncertainty principle, which says that both $x$ and $Hx$ cannot have their mass concentrated on few coordinates.

In [AC09] the following is shown via the Khintchine inequality:
Claim 2.
\[ \forall x, \|x\|_2 = 1, \mathbb{P}_\alpha \left( \|HDx\|_\infty > c \cdot \sqrt{\frac{\log(n/\delta)}{n}} \right) < \delta/2 \]

If we condition on \( \alpha \) so that the event of the above claim holds, then Bernstein implies that for
\[ m \geq \frac{\log(1/\delta) \log(n/\delta)}{\epsilon^2}, \]
we will have \( \| (1/\sqrt{m}) PHDx \|_2^2 = (1 \pm \epsilon)\|x\|_2^2 \) with probability \( 1 - \delta/2 \). Thus by a union bound, the overall failure probability is \( \delta \).

If we actually want to have \( O(\epsilon^{-2} \log(1/\delta)) \) rows, one way to achieve this is to set use the matrix \( \Pi' \cdot (1/\sqrt{m}) PHD \), where \( \Pi' \) is for example a dense random sign matrix with \( m' = O(\epsilon^{-2} \log(1/\delta)) \) rows.

The total time to apply \( \Pi' \cdot \Pi \) is then \( O(n \log n + m' \cdot m) \).

A slightly different analysis can improve the \( \log(n/\delta) \) dependence in \( m \) to actually be \( \log(m/\delta) \) as follows.

Theorem 3. Let \( x \in \mathbb{R}^n \) be an arbitrary unit norm vector, and suppose \( 0 < \epsilon, \delta < 1/2 \). Also let \( \Pi = S_HD \) as described above with a number of rows equal to \( m \gtrsim \epsilon^{-2} \log(1/\delta) \). Then
\[ \mathbb{P}_{\Pi}(\|\Pi x\|_2^2 - 1 > \epsilon) < \delta. \]

Proof. We use the moment method. Let \( \eta' \) be an independent copy of \( \eta \), and let \( \sigma \in \{-1, 1\}^n \) be uniformly random. Write \( z = HDx \) so that \( \|\Pi x\|_2^2 = \sum_i \eta_i z_i^2 \). Then
\[ \| \frac{1}{m} \sum_{i=1}^n \eta_i z_i^2 - 1 \|_p = \| \frac{1}{m} \sum_{i=1}^n \eta_i z_i^2 - 1 \|_{L^p(\eta)} \|_{L^p(\sigma)} \]
\[ = \| \frac{1}{m} \sum_{i=1}^n \eta_i z_i^2 - \frac{1}{m} \mathbb{E}_{\eta'} \sum_{i=1}^n \eta_i' z_i^2 \|_{L^p(\eta)} \|_{L^p(\sigma)} \]
\[ \leq \| \frac{1}{m} \sum_{i=1}^n z_i^2(\eta_i - \eta_i') \|_{L^p(\eta, \eta')} \|_{L^p(\sigma)} \text{ (Jensen)} \]
\[ = \| \frac{1}{m} \sum_{i=1}^n \sigma_i z_i^2(\eta_i - \eta_i') \|_{L^p(\eta, \eta')} \|_{L^p(\sigma)} \text{ (equal in distribution)} \]
\[ \leq \frac{2}{m} \cdot \| \sum_{i=1}^n \sigma_i \eta_i z_i^2 \|_{L^p(\eta)} \|_{L^p(\sigma)} \text{ (triangle inequality)} \]
\[ \leq \frac{2}{m} \cdot \| \sum_{i=1}^n \sigma_i \eta_i z_i^2 \|_p \]
\[ \lesssim \sqrt{p} \cdot \| (\sum_{i=1}^n \eta_i z_i^2)^{1/2} \|_p \text{ (Khinchine)} \]
\[ \leq \sqrt{p} \cdot \| \max_i \eta_i |z_i| : (\sum_i \eta_i z_i^2)^{1/2} \|_p \]
\[ \leq \sqrt{p} \cdot \| \max_i \eta_i z_i^2 \|_p^{1/2} : \| \sum_i \eta_i z_i^2 \|_p^{1/2} \text{ (Cauchy-Schwarz)} \]
\[
\leq \sqrt{\frac{p}{m}} \cdot \| \max_i \eta_i z_i^2 \|_{p/2}^{1/2} \cdot (\| \frac{1}{m} \sum_i \eta_i z_i^2 - 1 \|_{p/2}^{1/2} + 1) \tag{triangle inequality} \quad (2)
\]

We will now bound \( \| \max_i \eta_i z_i^2 \|_{p/2} \). Define \( q = \max \{ p, \log m \} \) and note \( \| \cdot \|_p \leq \| \cdot \|_q \). Then

\[
\| \max_i \eta_i z_i^2 \|_q = \left( \mathbb{E} \max_{\alpha, \eta} \eta_i z_i^{2q} \right)^{1/q} \leq \left( \mathbb{E} \sum_i \eta_i z_i^{2q} \right)^{1/q} = \left( \sum_i \mathbb{E} \eta_i z_i^{2q} \right)^{1/q} \leq \left( n \cdot \max_i \mathbb{E} \eta_i z_i^{2q} \right)^{1/q} = \left( n \cdot \max_i (\mathbb{E} \eta_i) \cdot (\mathbb{E} \eta_i z_i^{2q}) \right)^{1/q} \quad (\alpha, \eta \text{ independent})
\]

\[
= \left( m \cdot \max_i \mathbb{E} z_i^{2q} \right)^{1/q} \leq 2 \cdot \max_i \| z_i \|_q \quad (m^{1/q} \leq 2 \text{ by choice of } q)
\]

\[
= 2 \cdot \max_i \| z_i \|_{2q} \lesssim q \quad \text{(Khintchine)} \tag{3}
\]

Eq. (3) uses that \( H \) is an unnormalized bounded orthonormal system.

Defining \( E = \| \frac{1}{m} \sum_i \eta_i z_i^2 - 1 \|_{p/2}^{1/2} \) and combining (1), (2), (3), we find that for some constant \( C > 0 \)

\[
E^2 - C \sqrt{\frac{pq}{m}} E - C \sqrt{\frac{pq}{m}} \leq 0,
\]

implying \( E^2 \lesssim \max \{ \sqrt{pq/m}, pq/m \} \). By the Markov inequality

\[
\mathbb{P}(\| \Pi x \|_2^2 - 1 > \varepsilon) \leq \varepsilon^{-p} \cdot E^{2p},
\]

and thus to achieve the theorem statement it suffices to set \( p = \log(1/\delta) \) then choose \( m \gtrsim \varepsilon^{-2 \log(1/\delta) \log(1/\delta)} \).

**Remark 4.** Note that the FJLT as analyzed above provides suboptimal \( m \). If one desired optimal \( m \), one can instead use the embedding matrix \( \Pi' \Pi \), where \( \Pi \) is the FJLT and \( \Pi' \) is, say, a dense matrix with Rademacher entries having the optimal \( m' = O(\varepsilon^{-2 \log(1/\delta)}) \) rows. The downside is that the runtime to apply our embedding worsens by an additive \( m \cdot m' \). [AC09] slightly improved this additive term (by an \( \varepsilon^2 \) multiplicative factor) by replacing the matrix \( S \) with a random sparse matrix \( P \).

**Remark 5.** The usual analysis for the FJLT, such as the approach in [AC09], would achieve a bound on \( m \) of \( O(\varepsilon^{-2 \log(1/\delta) \log(n/\delta)}) \). Such analyses operate by, using the notation of the proof...
of Theorem 3, first conditioning on \( \|z\|_\infty \lesssim \sqrt{\log(n/\delta)} \) (which happens with probability at least \( 1 - \delta/2 \) by the Khintchine inequality), then finishing the proof using Bernstein’s inequality. In our proof above, we improved this dependence on \( n \) to a dependence on the smaller quantity \( m \) by avoiding any such conditioning.

### 3.1 Application: High-dimensional approximate nearest neighbors search (ANN)

Let’s assume that we’re working with \( L_2 \) distances in \( \mathbb{R}^d \). Let’s define the exact nearest neighbors problem as follows: we’re given \( n \) points \( P = \{p_1, p_2 \ldots p_n\} \), \( p_i \in \mathbb{R}^d \). We need to create a data structure such that a query on point \( q \in \mathbb{R}^d \) returns a point \( p \in P \) such that the distance \( \|p - q\|_2 \) is minimized. An example application might be image retrieval (similar images). In the approximate case, we want to return a point \( p \) such that \( \|p - q\|_2 \leq c \cdot \min_{p' \in P} \|p' - q\|_2 \). Note that the simple solution is to store all the points in a list and just check them all on query, but that requires \( nd \) time to calculate.

#### 3.1.1 Voronoi diagrams

One way to solve this problem is to construct the Voronoi diagram for the points in the space, which is the division of the space into areas \( A_i \) such that all points \( x \in A_i \) are closest to \( p_i \). Then on a query we do planar point location to find the correct Voronoi cell for a point. However, when \( d \neq 2 \), the curse of dimensionality strikes. In \( d \) dimensions, the Voronoi diagram requires \( n^{\Theta(d)} \) space to store. Note that this is a lower bound!

#### 3.1.2 Approximate Nearest Neighbor (ANN)

This reduces to the problem \( c \)-NN ([HPI12]).

\((c, r)\)-NN: If there exists \( p \in P \) such that \( \|p - q\| \leq r \), then return \( p' \in P \) such that \( \|p' - q\| \leq cr \). If there doesn’t exist such a \( p \), then FAIL.

The easiest reduction is just binary search on \( r \), but the above reference avoids some problems.

<table>
<thead>
<tr>
<th>Space</th>
<th>Time</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( dn + n^{O(1/c^2)} )</td>
<td>( (d + \frac{\log n}{c})O(1) )</td>
<td>[KOR98][IM98]</td>
</tr>
<tr>
<td>( dn + n^{1+\rho(c)} )</td>
<td>( n^{\rho(c)} )</td>
<td></td>
</tr>
<tr>
<td>( \rho(c) = )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( 1/c )</td>
<td></td>
<td>[GIM+99]</td>
</tr>
<tr>
<td>( \frac{1}{c^2} + o(1) )</td>
<td></td>
<td>[AI06]</td>
</tr>
<tr>
<td>( (7/8)c^2 + o(1/c^2) )</td>
<td></td>
<td>[AINR14]</td>
</tr>
<tr>
<td>( (2c^2-1)^{1/2} + o(1) )</td>
<td></td>
<td>[AR15]</td>
</tr>
<tr>
<td>( O(dn) )</td>
<td>( \frac{2^{16}}{c} )</td>
<td>[MNP07]</td>
</tr>
</tbody>
</table>

Today we just show the following result:

1. ANN with \( n^{O(\log(1/c)/c^2)} \) space.

6
2. First, $dn + n \cdot O(c/\epsilon^d)$ space

3. Pretend $r = 1$. Impose uniform $\epsilon/\sqrt{d}$ grid on $\mathbb{R}^d$

4. for each $p_i \in P$, let $B_i = B_{l_2}(p_i, 1)$

5. let $B_i' = \text{set of the grid cells that } B_i \text{ intersects}$

6. Store $B' = \cup B_i'$ in a hash table (key = grid cell ID, value = i).

7. # of grid cell intersected $\leq Vol(B_{l_2}(1+\epsilon)/Vol(\text{grid cell}))$

8. The volume of the ball is $R_d 2^{O(d)}/d^{d/2}$

9. so we have # of grid cell intersected $\leq (c/\epsilon)^d$

Now note $d$ can be reduced to $O(\epsilon^{-2} \log n)$ using the JL lemma, giving the desired space bound.

References


