CS 229r: Algorithms for Big Data Fall 2015

Lecture 13 — October 15, 2015

Prof. Jelani Nelson *Scribe: Yakir Reshef*

1 Recap and overview

Last time we started by talking about lower bounds for JL that were worst-case. That is, we stated a lower bound on the reduced dimension m such that for any Π mapping to $\mathbb{R}^m$ there would exist a "bad" set T of x's whose distances would not be preserved by Π. But this left open the question of how well we could do on some particular set T. This was where Gordon's theorem came in. It said

Theorem 1 (Gordon)**.** *Suppose $T \subset S^{n-1}$. If $\Pi \in \mathbb{R}^{m\times n}$ has $\Pi_{ij} = g_{ij}/\sqrt{m}$, where the g_{ij} are iid standard normals, and $m \gtrsim \frac{g^2(T)+1}{\varepsilon^2}$, then*

$$P_\Pi\left(\exists x \in T \,:\, |\|\Pi x\| - 1| > \varepsilon\right) < \frac{1}{10}.$$

where $g(T) = E_g \sup_{x\in T}\langle g, x\rangle$ is the mean width of T, with the expectation taken over a gaussian with mean zero and identity covariance.

Recall that we also mentioned the following result last time concerning $g(T)$, with one direction of the inequality showed by Fernique and the other by Talagrand:

Theorem 2. *Let $T \subset \mathbb{R}^n$ bounded, and let $T_0 \subset T_1 \subset \cdots T$ be such that $|T_0| = 1$ and $|T_r| \le 2^{2^r}$. Define*

$$\gamma_2(T, d) := \inf_{\{T_r\}_{r=0}^\infty} \sup_{x\in T} \sum_{r=0}^{\infty} 2^{r/2} d(x, T_r).$$

Then $\gamma_2(T, \ell_2) \simeq g(T)$.

Henceforth, when we say $\gamma_2(T)$ without specifying a metric, ℓ_2 is implied.

Gordon's theorem implies DJL, because in general $g(T)$ is at most $\sqrt{\log|T|}$ in it can be much smaller for some T. Today we'll show that the converse is also true, i.e., that DJL implies Gordon's theorem.

2 Today: DJL ⇒ Gordon's theorem

2.1 Statement of result

The main result is summarized in the following theorem.

Theorem 3 ([1]). *Define* $L = \lceil \log n \rceil$, $\tilde{\varepsilon} = \varepsilon/(c\gamma_2(T))$, $\tilde{\varepsilon}_r = \max\{2^{r/2}\tilde{\varepsilon}, 2^{r/2}\tilde{\varepsilon}^2\}$, $\delta_r = \frac{\delta}{C2^r 8^{2^r}}$. *Let* $T \subset S^{n-1}$. *Then if* D *satisfies* $(\varepsilon_r, \delta_r)$*-DJL for* $r = 0, \ldots, L$, *then*

$$P_{\Pi \sim D}\left(\sum_{x \in T} \left| \|\Pi x\|_2^2 - 1 \right| > \varepsilon\right) < \delta$$

To see why this implies Gordon's theorem, we consider the random sign matrix, e.g., $\Pi_{ij} = \frac{\sigma_{ij}}{\sqrt{m}}$. We know that this matrix satisfies $(\tilde{\varepsilon}, \tilde{\delta})$-DJL for $m \gtrsim \frac{\log(1/\tilde{\delta})}{\tilde{\varepsilon}^2}$, which equals $\frac{2^r \log(1/\tilde{\delta})}{(2^{r/2}\tilde{\varepsilon})^2} \geq \frac{\log(1/\delta_r)}{\varepsilon_r^2}$ for all r. The theorem therefore applies and so we see that we get an (ε, δ) guarantee with $m \gtrsim \log(1/\delta)/\tilde{\varepsilon}^2 \asymp \frac{\gamma_2^2(T)}{\varepsilon^2}\log(1/\delta)$. And since $\gamma_2(T) \simeq g(T)$, this is approximately $\frac{g^2(T)\log(1/\delta)}{\varepsilon^2}$, which gives Gordon's theorem. As mentioned last time, different proofs yield that $\frac{g^2(T)+\log(1/\delta)}{\varepsilon^2}$ actually suffices.

2.2 Proof of result

To prove the above theorem, the lemma below suffices.

Lemma 1. *For a given set* T, *let* T_r *be the sequence that achieves the infimum in the definition of* γ_2. *To achieve* $\sup_{x \in T} \left| \|\Pi x\|_2^2 - 1 \right| < \varepsilon$, *it suffices that for all* $r = 0, \ldots, L$, *the following hold simultaneously for all* $r \in [L]$.

- *For all* $v \in T_{r-1} \cup T_r \cup (T_{r-1} - T_r)$,
$$\|\Pi v\| \leq (1 + 2^{r/2}\tilde{\varepsilon})\|v\| \tag{1}$$
- *For all* $v \in T_{r-1} \cup T_r \cup (T_{r-1} - T_r)$,
$$|\|\Pi v\|^2 - \|v\|^2| \leq \max\{2^{r/2}\tilde{\varepsilon}, 2^r\tilde{\varepsilon}^2\} \cdot \|v\|^2 \tag{2}$$
- *For all* $u \in T_{r-1}$ *and* $v \in T_r - \{u\}$,
$$|\langle \Pi u, \Pi v \rangle - \langle u, v \rangle| \leq \max\{2^{r/2}\tilde{\varepsilon}, 2^r\tilde{\varepsilon}^2\} \cdot \|u\| \cdot \|v\| \tag{3}$$
- *We also have*
$$\|\Pi\| \leq 1 + (1/4)2^{L/2}\tilde{\varepsilon} \tag{4}$$

We note that it is not too bad to show that the first three conditions hold with high probability since they are all JL-type conditions. The third one is a bit less obvious since it's about dot products instead of norms. But notice that $\|u+v\|^2 - \|u-v\|^2 = 4\langle u, v\rangle$. So if $\|u\| = \|v\| = 1$, then Π preserving $u+v$ and $u-v$ means that $\langle \Pi u, \Pi v\rangle = \frac{1}{4}(\|\Pi u + \Pi v\|^2 - \|\Pi u + \Pi v\|^2) = \langle u, v\rangle \pm O(\varepsilon)$. If u and v don't have unit norm you can scale them to achieve the above condition. So the third condition also follows from the DJL assumption. (We neglect the fourth condition above for now, since it's based on a standard argument which is based on a constant-sized net of S^{n-1} that we'll see later in the course.)

We now argue that the lemma suffices to prove our theorem.

Claim 1. *Lemma 1 implies Theorem 3.*

Proof. Define $\tilde{L} = \lceil \log(1/\tilde{\varepsilon}^2) \rceil \leq L$ (Note that if $\tilde{L} > L$ then we're not interested because then we're not reducing dimensionality!) Fix $x \in T$. We will show

$$\left| \|\Pi x\|^2 - \|x\|^2 \right| < \varepsilon$$

Define $e_r(T) = d(x, T_r)$, and define

$$\tilde{\gamma}_2(T) = \sum_{r=1}^{L} 2^{r/2} \cdot e_r(T).$$

Clearly $\tilde{\gamma}_2(T) \leq \gamma_2(T)$.

Also define

$$z_r = \text{argmin}_{y \in T_r} \|x - y\|_2$$

$$\begin{aligned}
|\|\Pi x\|^2 - \|x\|^2| &\leq |\|\Pi z_{\tilde{L}}\|^2 - \|z_{\tilde{L}}\|^2| + |\|\Pi x\|^2 - \|\Pi z_{\tilde{L}}\|^2| + |\|x\|^2 - \|z_{\tilde{L}}\|^2| \\
&\leq \underbrace{|\|\Pi z_0\|^2 - \|z_0\|^2|}_{\alpha} + \underbrace{|\|\Pi x\|^2 - \|\Pi z_{\tilde{L}}\|^2|}_{\beta} + \underbrace{|\|x\|^2 - \|z_{\tilde{L}}\|^2|}_{\Gamma} \\
&\quad + \underbrace{\sum_{r=1}^{\tilde{L}} (|\|\Pi z_r\|^2 - \|z_r\|^2| - |\|\Pi z_{r-1}\|^2 - \|z_{r-1}\|^2|)}_{\Delta} \qquad (5)
\end{aligned}$$

In class we bounded α and Γ, as the bound on α is simple and the bound on Γ sufficiently captures the proof idea. However, in these notes we bound all of $\alpha, \beta, \Gamma, \Delta$.

Bounding α: We have $\alpha \leq \max\{\tilde{\varepsilon}, \tilde{\varepsilon}^2\} \leq \tilde{\varepsilon}$ by Eq. (2).

Bounding β: We have

$$\begin{aligned}
|\|\Pi x\|^2 - \|\Pi z_{\tilde{L}}\|^2| &= |\|\Pi x\| - \|\Pi z_{\tilde{L}}\|| \cdot |\|\Pi x\| + \|\Pi z_{\tilde{L}}\|| \\
&\leq |\|\Pi x\| - \|\Pi z_{\tilde{L}}\|| \cdot \left(|\|\Pi x\| - \|\Pi z_{\tilde{L}}\|| + 2 \cdot \|\Pi z_{\tilde{L}}\| \right) \\
&= |\|\Pi x\| - \|\Pi z_{\tilde{L}}\||^2 + 2|\|\Pi x\| - \|\Pi z_{\tilde{L}}\|| \cdot \|\Pi z_{\tilde{L}}\| \qquad (6)
\end{aligned}$$

We thus need to bound $|\|\Pi x\| - \|\Pi z_{\tilde{L}}\||$ and $\|\Pi z_{\tilde{L}}\|$. By Eq. (1) we have $\|\Pi z_{\tilde{L}}\| \leq (1 + 2^{\tilde{L}/2}\tilde{\varepsilon}) \leq 2$.

Next, we have

$$\begin{aligned}
|\|\Pi x\| - \|\Pi z_{\tilde{L}}\|| &= |\|\Pi x\| - \|\Pi z_L\| + \|\Pi z_L\| - \|\Pi z_{\tilde{L}}\|| \\
&\leq \|\Pi(x - z_L)\| + \|\Pi(z_L - z_{\tilde{L}})\| \\
&\leq \|\Pi\| \cdot \|x - z_L\| + \| \sum_{r=\tilde{L}+1}^{L} \Pi(z_r - z_{r-1})\|
\end{aligned}$$

$$\leq \|\Pi\| \cdot e_L(T) + \sum_{r=\tilde{L}+1}^{L} \|\Pi(z_r - z_{r-1})\| \tag{7}$$

By Eq. (4), $\|\Pi\| \leq \frac{1}{4}2^{L/2}\tilde{\varepsilon} + 1$. Also by Eq. (1), $\|\Pi(z_r - z_{r-1})\| \leq (1 + 2^{r/2}\tilde{\varepsilon})\|z_r - z_{r-1}\|$. Thus, using $2^{r/2}\tilde{\varepsilon} \geq 1$ for $r > \tilde{L}$,

$$\begin{aligned}
(7) &\leq (\frac{1}{4}2^{L/2}\tilde{\varepsilon} + 1)e_L(T) + \sum_{r=\tilde{L}+1}^{L} (1 + 2^{r/2}\tilde{\varepsilon})\|z_r - z_{r-1}\| \\
&\leq (\frac{1}{4}2^{L/2}\tilde{\varepsilon} + 1)e_L(T) + \sum_{r=\tilde{L}+1}^{L} (1 + 2^{r/2}\tilde{\varepsilon})\|z_r - z_{r-1}\| \\
&\leq \frac{5}{4}2^{L/2}\tilde{\varepsilon}e_L(T) + \sum_{r=\tilde{L}+1}^{L} 2^{r/2+1}\tilde{\varepsilon}\|z_r - z_{r-1}\| \\
&\leq \frac{5}{4}2^{L/2}\tilde{\varepsilon}e_L(T) + 4\sqrt{2}\tilde{\varepsilon} \sum_{r=\tilde{L}+1}^{L} 2^{(r-1)/2} \cdot e_{r-1}(T) \\
&\leq 4\sqrt{2}\tilde{\varepsilon} \sum_{r=\tilde{L}}^{L} 2^{r/2} \cdot e_r(T) \\
&\leq 4\sqrt{2}\tilde{\varepsilon} \cdot \tilde{\gamma}_2(T)
\end{aligned} \tag{8}$$

Thus in summary,

$$\beta \leq (6) \leq 32\tilde{\varepsilon}^2\tilde{\gamma}_2^2(T) + 16\sqrt{2}\tilde{\varepsilon}\tilde{\gamma}_2(T)$$

Bounding Γ: Note $2^{r/2}\tilde{\varepsilon} \geq 1/\sqrt{2}$ for $r \geq \tilde{L}$. Thus

$$\begin{aligned}
|\|x\| - \|z_{\tilde{L}}\|| &\leq e_{\tilde{L}}(T) \\
&\leq \sqrt{2} \cdot 2^{\tilde{L}/2}\tilde{\varepsilon}e_{\tilde{L}}(T) \\
&\leq \sqrt{2}\tilde{\varepsilon} \cdot \tilde{\gamma}_2(T).
\end{aligned}$$

Thus

$$\begin{aligned}
\Gamma &= |\|x\|^2 - \|z_{\tilde{L}}\|^2| \\
&= |\|x\| - \|z_{\tilde{L}}\|| \cdot |\|x\| + \|z_{\tilde{L}}\|| \\
&\leq |\|x\| - \|z_{\tilde{L}}\||^2 + 2|\|x\| - \|z_{\tilde{L}}\|| \cdot \|z_{\tilde{L}}\|| \\
&\leq 2\tilde{\varepsilon}^2 \cdot \tilde{\gamma}_2^2(T) + 2\sqrt{2}\tilde{\varepsilon} \cdot \tilde{\gamma}_2(T)
\end{aligned}$$

Bounding Δ: By the triangle inequality, for any $r \geq 1$

$$\begin{aligned}
|\|\Pi z_r\|^2 - \|z_r\|^2| &= |\|\Pi(z_r - z_{r-1}) + \Pi z_{r-1}\|^2 - \|(z_r - z_{r-1}) + z_{r-1}\|^2| \\
&= |\|\Pi(z_r - z_{r-1})\|^2 + \|\Pi z_{r-1}\|^2 + 2\left\langle \Pi(z_r - z_{r-1}), \Pi z_{r-1} \right\rangle \\
&\quad - \|z_r - z_{r-1}\|^2 - \|z_{r-1}\|^2 - 2\left\langle z_r - z_{r-1}, z_{r-1} \right\rangle|
\end{aligned} \tag{9}$$

$$\leq |\|\Pi(z_r - z_{r-1})\|^2 - \|z_r - z_{r-1}\|^2| + |\|\Pi z_{r-1}\|^2 - \|z_{r-1}\|^2| + 2|\langle \Pi(z_r - z_{r-1}), \Pi z_{r-1}\rangle - \langle z_r - z_{r-1}, z_{r-1}\rangle|. \quad (10)$$

By Eq. (2) we have

$$|\|\Pi(z_r - z_{r-1})\|^2 - \|z_r - z_{r-1}\|^2| \leq \max\{2^{r/2}\tilde{\varepsilon}, 2^r\tilde{\varepsilon}^2\} \cdot 2e_{r-1}^2(T) \leq 2^{r/2+2}\tilde{\varepsilon}e_{r-1}^2(T),$$

with the second inequality holding since $2^{r/2}\tilde{\varepsilon} \leq 1$ for $\leq \tilde{L}$.

By Eq. (3) we also have

$$|\langle \Pi(z_r - z_{r-1}), \Pi z_{r-1}\rangle - \langle z_r - z_{r-1}, z_{r-1}\rangle| \leq 2^{r/2+1}\tilde{\varepsilon}e_{r-1}.$$

Therefore

$$|\|\Pi z_r\|^2 - \|z_r\|^2| - |\|\Pi z_{r-1}\|^2 - \|z_{r-1}\|^2| \leq \tilde{\varepsilon}(2e_{r-1}(T) + 4e_{r-1}^2(T))2^{r/2}$$

Noting $e_r(T) \leq 1$ for all r,

$$\begin{aligned}\Delta &\leq 10\tilde{\varepsilon}\left(\sum_{r=1}^{\tilde{L}} 2^{r/2}e_{r-1}(T)\right)\\ &= 10\sqrt{2}\tilde{\varepsilon}\left(\sum_{r=0}^{\tilde{L}-1} 2^{r/2}e_r(T)\right)\\ &\leq 10\sqrt{2}\tilde{\varepsilon}\tilde{\gamma}_2(T).\end{aligned}$$

Finishing up: We have thus established

$$\begin{aligned}|\|\Pi x\|^2 - \|x\|^2| &\leq \tilde{\varepsilon} + 32\tilde{\varepsilon}^2\tilde{\gamma}_2^2(T) + 16\sqrt{2}\tilde{\varepsilon}\tilde{\gamma}_2(T) + 8\tilde{\varepsilon}^2\tilde{\gamma}_2^2(T) + 2\sqrt{2}\tilde{\varepsilon}\tilde{\gamma}_2(T) + 10\sqrt{2}\tilde{\varepsilon}\tilde{\gamma}_2(T)\\ &= \tilde{\varepsilon} + 28\sqrt{2}\tilde{\varepsilon}\tilde{\gamma}_2(T) + 40\tilde{\varepsilon}^2\tilde{\gamma}_2^2,\end{aligned}$$

which is at most ε for $\tilde{\varepsilon} \leq \varepsilon/(84\sqrt{2}\tilde{\gamma}_2(T))$.

□

3 Doing JL fast

Typically we have some high-dimensional computational geometry problem, and we use JL to speed up our algorithm in two steps: (1) apply a JL map Π to reduce the problem to low dimension m, then (2) solve the lower-dimensional problem. As m is made smaller, typically (2) becomes faster. However, ideally we would also like step (1) to be as fast as possible. So far the dimensionality reduction has been a dense matrix-vector multiplication. So we can ask: can we do better in terms of runtime?

There are two possible ways of doing this: one is to make Π sparse. We saw in pset 1 that this sometimes works: we replaced the AMS sketch with a matrix each of whose columns has exactly 1 non-zero entry. The other way is to make Π structured, i.e., it's still dense but has some structure that allows us to multiply faster. We'll start talking today about sparse JL.

3.1 Sparse JL

One natural way to speed up JL is to make Π sparse. If Π has s non-zero entries per column, then Πx can be multiplied in time $O(s \cdot \|x\|_0)$, where $\|x\|_0 = |\{i : x_i \neq 0\}|$. The goal is then to make s, m as small as possible.

First some history: [2] showed that you can set

$$\Pi_{ij} = \begin{cases} +/\sqrt{m/3} & \text{w.p. } \frac{1}{6} \\ -\sqrt{m/3} & \text{w.p. } \frac{1}{6} \\ 0 & \text{w.p. } \frac{2}{3} \end{cases}$$

and that this gives DJL, even including constant factors. But it provides a factor-3 speedup since in expectation only one third of the entries in Π are non-zero. On the other hand, [5] proved that if Π has independent entries then you can't speed things up by more than a constant factor.

The first people to exhibit a Π without independent entries and therefore to break this lower bound were [3], who got $m = O(\frac{1}{\varepsilon^2}\log(1/\delta))$, $s = \tilde{O}(\frac{1}{\varepsilon}\log^3(1/\delta))$ nonzeros per column of Π. So depending on the parameters this could either be an improvement or not.

Today we'll see [4], which showed that you can take $m = O(\frac{1}{\varepsilon^2}\log(1/\delta))$ and $s = O(\frac{1}{\varepsilon}\log(1/\delta))$, a strict improvement. You do this by choosing exactly s entries in each column of Π to have non-zero entries and then choosing the signs of those entries at random and normalizing appropriately. Alternatively, you can break each column of Π up into s blocks of size m/s, and choose exactly 1 non-zero entry in each block. The resulting matrix is exactly the count sketch matrix.

The analysis uses Hanson-Wright. Previously, for dense Π, we observed that $\Pi x = A_x\sigma$ where A_x was an $m \times mn$ matrix whose i-th row had $x^T/\sqrt{m}$ in the i-th block of size n and zeros elsewhere. Then we said $\|\Pi x\|^2 = \sigma^T A_x^T A_x \sigma$, which was a quadratic form, which allowed us to appeal to HW. We'll do something similar here.

First some notation: we'll write $\Pi_{ij} = \frac{\sigma_{ij}\delta_{ij}}{\sqrt{s}}$ where $\delta_{ij} \in \{0,1\}$ is a random variable indicating whether the corresponding entry of Π was chosen to be non-zero. (So the δ_{ij} are not independent.) For every $r \in [m]$, define $x(r)$ by $(x(r))_i = \delta_{ri}x_i$. The claim is now that $\Pi x = A_x\sigma$ where A_x is an $m \times mn$ matrix whose i-th row contains $x(r)^T/\sqrt{s}$ in the i-th block of size n and zeros elsewhere. This allows us to again use the HW trick: specifically, we observe that $A_x^T A_x$ is a block-diagonal matrix as before. And since we're bounding the difference between $\sigma^T A_x^T A_x \sigma$ and its expectation, it is equivalent to bound $\sigma^T B_x \sigma$ where B_x is $A_x^T A_x$ with its diagonals zeroed out.

Now condition on B_x and recall that HW says that for all $p \geq 1$, $\|\sigma^T B_x\sigma\|_p \leq p\|B_x\| + \sqrt{p}\|B_x\|_F$. Then, taking p-norms with respect to the δ_{ij} and using the triangle inequality, we obtain the bound

$$\|\sigma^T B_x\sigma\|_p \leq p\|\|B_x\|\|_p + \sqrt{p}\|\|B_x\|_F\|_p$$

If we can bound the right-hand-side, we'll obtain required DJL result by application of Markov's inequality, since $\sigma^T B_x\sigma$ is positive. Therefore, it suffices to bound the p-norms with respect to the δ_{ij} of the operator and Frobenius norms of B_x.

We start with the operator norm: since B_x is block-diagonal and its i-th block is $x(r)x(r)^T - \Lambda(r)$ where $\Lambda(r)$ is the diagonal of $x(r)x(r)^T$, we have $\|B_x\| = \frac{1}{s}\max_{1\leq r\leq m}\|x(r)x(r)^T - \Lambda(r)\|$. But the operator norm of the difference of positive-definite matrices is at most the max of either operator norm. Since both matrices have operator norm at most 1, this gives us tht $\|B_x\| \leq 1/s$ always.

So it remains only to bound the Frobenius norm. This is where we'll pick up next time.

References

[1] Samet Oymak, Benjamin Recht, Mahdi Soltanolkotabi. Isometric sketching of any set via the Restricted Isometry Property. *CoRR* abs/1506.03521, 2015.

[2] Dimitris Achlioptas. Database-friendly random projections. *J. Comput. Syst. Sci.* 66(4): 671–687, 2003.

[3] Anirban Dasgupta, Ravi Kumar, Tamás Sarlós. A sparse Johnson: Lindenstrauss transform. *STOC*, pages 341–350, 2010.

[4] Daniel M. Kane, Jelani Nelson. Sparser Johnson-Lindenstrauss Transforms. *Journal of the ACM*, 61(1): 4:1–4:23, 2014.

[5] Jirí Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2): 142–156, 2008.