Sketching Algorithms for Big Data Fall 2017

Lecture 5 — September 14, 2017

Prof. Jelani Nelson *Scribe: Rafail Ketsetsidis*

1 Overview

In the past two lectures we discussed algorithms for estimating the L_2. Specifically, the Alon-Matias-Szeged (AMS) algorithm [2] and the Johnson-Lindenstrauss (JL) lemma [1], as well as a generalization of the JL lemma to the L_p norm for $p \in (0, 2]$, and for $p > 2$ in the case of strictly positive incremental updates.

In this lecture we discuss various lower bounds for the both the Distributional JL (DJL) lemma and the JL lemma. Particularly, the history of the lower bound computation for these two versions, as well as proofs of the original lower bound by Johnson-Lindenstraus [1], a better lower bound by Alon [3], the optimal lower bound for DJL by Kane, Meka, and Nelson [7], and an optimal lower bound for JL under a condition by Larsen and Nelson [8], which we will finish on the next lecture.

2 Restating the problem

Two lectures ago professor Indyk talked about this lemma at MIT when he talked about L_2 norm estimation in data streams. Using the AMS sketch we could provide a $1 + \epsilon$ approximation of the L_2, which worked by essentially giving an embedding into L_2. Professor Indyk showed that if we look at the squared dot product between any row and x, its expectation will be the squared norm of x. So if we average $\Theta(\frac{1}{\epsilon^2})$ estimators, as in previous lectures, then we get a $1 + \epsilon$-approximation. Therefore we can see that AMS essentially maps x into Πx, where Π is a random-sign matrix with n columns and $\Theta(\frac{1}{\epsilon^2})$ rows, with each cell equal to ± 1. We then normalize the matrix by $\frac{1}{\sqrt{m}}$ (m is the number of estimators, so the number of rows in Π), so now we can we estimate the norm of x by estimating the norm of Πx.

2.1 Distributional JL Lemma

The Distributional JL lemma [1], as we saw it last time, is an alternative statement of the original statement of the JL lemma. it states that $\forall\ \epsilon, \delta \in (0, \frac{1}{2})\ \exists\ D_{\epsilon,\delta}$ on $\mathbb{R}^{m \times d}$ such that $\forall u \in \mathbb{R}^d$ we have that

$$\mathbb{P}_{\Pi \sim D_{\epsilon,\delta}}(\|\Pi u\|_2 \notin [1 - \epsilon, 1 + \epsilon] \cdot \|u\|_2) < \delta$$

where $m = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$.

An example of a distribution $D_{\epsilon,\delta}$ is matrices comprised of i.i.d. Gaussian entries.

2.2 JL Lemma

The JL lemma [1] states that $\forall\ n, d > 1,\ \forall\ \epsilon \in (0, \frac{1}{2}),\ \forall\ X \subseteq \mathbb{R}^d, |X| = n\ \exists\ f : X \to \mathbb{R}^m, m = O(\frac{1}{\epsilon^2} \log n)$ such that

$$\forall x, y \in X,\ (1-\epsilon)\|x-y\|_2 \leq \|f(x)-f(y)\|_2 \leq (1+\epsilon)\|x-y\|_2$$

This means that f is a near-isometry on X.

The JL lemma can be obtained by the DJL by setting $\delta < \frac{1}{n^2}$ and $u = x - y$ We can prove that JL follows directly from DJL because we can set $\delta < \frac{1}{n^2}, u = x - y$, so by choosing $f(x) = \Pi x$, then with high probability $\|\Pi u\|_2 \in [1-\epsilon, 1+\epsilon] \cdot \|u\|_2$. Then by union bound over all such pairs (which are $\binom{n}{2} < n^2$ in number) we get the desired result.

Can we achieve the same guarantee with a lower m?

3 History of lower bounds

In 2011 it was resolved that the DJL lower bound is optimal [7]. Observe that the derivation of JL from DJL does not imply that the lower bound of m will be the same for JL, as we can choose the function f, and for a non-linear map we might be able to achieve a lower bound. Only within the last year was it proved that the JL lower bound is also optimal. It turns out that the way to achieve the lower bound for JL is to actually choose a random linear map (i.i.d from a Gaussian distribution), so being able to "look" into the point set does not change anything.

- In the original JL paper by Johnson and Lindenstrauss [1], they show that there exist $n+1$ points in $\mathbb{R}^n$ such that any JL map needs $m \gtrsim \frac{\log n}{\log(2c+1)}$. This lower bound is achievable for large C, as shown by Indyk and Naor [4].
- Alon [3] showed that the for same point set that JL analyzed in their paper, we can achieve a better lower bound: $m \gtrsim \min\{n, \frac{1}{\epsilon^2} \frac{\log n}{\log \frac{1}{\epsilon}}\}$ Obviously we cannot get a lower bound better than n, because the points live in an n-dimensional space. This was the best lower bound known until within the last year.
- In 2011 there were two works, one by Jayram and Woodruff [6] and one by Kane, Meka and Nelson [7] which show that for DJL $m \gtrsim \min\{d, \frac{1}{\epsilon^2} \log \frac{1}{\delta}\}$. This is optimal, because one distribution that always works is the distribution that always returns the identity matrix, so that gives $m = d$, and the Gaussian distribution, which gives $\frac{1}{\epsilon^2} \frac{\log n}{\log \frac{1}{\epsilon}}$, so their minimum is the optimal lower bound (as both are achievable).
- After that, a paper by Larsen and Nelson [8] showed that $\forall \epsilon, n$, there exist $O(n^3)$ points in $\mathbb{R}^n$ such that any **linear** map needs $m \gtrsim \min\{n, \frac{1}{\epsilon^2} \log n\}$.

  Note that there are non-linear maps for the point sets used in the above paper that do better than JL.
- Larsen and Nelson in 2016 [9] showed that $\forall \epsilon \in (0, \frac{1}{2}), n, d$, with $\epsilon > \frac{\log^{0.5} n}{\min(n,d)}$, there exists a hard point set in $\mathbb{R}^d$ such that $m \gtrsim \frac{1}{\epsilon^2} \log n$, even if the map is non-linear.

  In this paper it was conjectured that $\forall\ n, d, \epsilon$ the optimal bound is $m = \Theta(\min\{n, d, \frac{1}{\epsilon^2} \log(\epsilon^2 n)\})$

- Alon and Klartag in 2017 [10] showed the lower bound of the conjecture above.

Regarding results in L_p, there is a theorem by Johnson and Naor [5] which shows that any norm space that has this kind of guarantee is very close to L2 in some precise sense. Specifically for every finite dimensional subspace of that norm space there is a low-distortion embedding to L_2.

3.1 Proof of the original JL lower bound [1]

First we decide on the hard point set $X = \{0, e_1, \ldots, e_n\} \subseteq \mathbb{R}^n$

Claim 1. *If we embed these points into m-dimensional space and you preserve distances up to a factor of c, then our target dimension has to be a factor of $\frac{\log n}{\log(2c+1)}$.*

Proof. We will assume that the embedding maps zero to zero, as we can translate without changing any istances. We want to preserve pairwise distances, and particularly distances of points to zero. They used to have distance 1 to zero and distance $\sqrt{2}$ to each other. Now the distance to zero is in $[1, c]$, and the distance to each other will be in $[\sqrt{2}, c\sqrt{2}]$. Let $\widetilde{e_i}$ be the images of the e_is under the current embedding. Now, for each one of these points we are going to consider a ball around it, with radius $\frac{1}{2}$.

Observe that the balls around the $\widetilde{e_i}$ are disjoint. Suppose that two of them overlap, the points will have distance at most one (as the radii of the balls are $\frac{1}{2}$, which is a contradiction since they need to have distance of at least $\sqrt{2} > 1$.

Also all balls lie in a radius $(C + \frac{1}{2})$-ball around zero.

Since the balls are disjoint, the sum of their volumes is bounded by the volume of the big ball. This means that $n \cdot \mathrm{vol}_m(B(\frac{1}{2})) \leq \mathrm{vol}_m(B(C + \frac{1}{2}))$. This implies that $n \leq \frac{\mathrm{vol}_m(B(c+\frac{1}{2}))}{\mathrm{vol}_m(B(\frac{1}{2}))} = (2c+1)^m$.

By taking log and dividing by $\log(2c+1)$, we trivially get that $m \geq \frac{\log n}{\log(2c+1)}$. □

3.2 Proof of the lower bound by Alon [3]

We will be using the same point set $X = \{0, e_1, \ldots, e_n\} \subseteq \mathbb{R}^n$. Also, as before, we will be assuming that zero gets mapped to zero.

Let B be an $n \times m$ matrix whose columns are equal to $f(e_i) \in \mathbb{R}^m$. Also, the columns have norm $1 \pm \epsilon$ (as we want to preserve the distance to zero).

We get that:

$$\|e_i - e_j\|_2^2 = \|e_i\|_2^2 + \|e_j\|_2^2 - 2\langle e_1, e_2\rangle = 2 - 2\langle e_1, e_2\rangle$$

and:

$$\begin{aligned}
\|f(e_i) - f(e_j)\|_2^2 &= \|f(e_i)\|_2^2 + \|f(e_j)\|_2^2 - 2\langle f(e_i), f(e_j) \\
&= (1 \pm \epsilon) + (1 \pm \epsilon) - 2\langle f(e_i), f(e_j)\rangle \Rightarrow \\
\|e_i - e_j\|_2^2 + O(\epsilon) &= 2 - 2\langle f(e_i), f(e_j)\rangle + O(\epsilon) \Rightarrow \\
\langle e_i, e_j\rangle &= \langle f(e_i), f(e_j)\rangle \pm O(\epsilon)
\end{aligned}$$

If f preserves distances then it also approximately preserves dot products, as it is an L_2 embedding. Therefore all the dot products between the columns of B are approximately ϵ.

B is ϵ-incoherent matrix (after dividing the columns by their norms). This means that the columns have unit norm, and for all columns u, v we have that $|\langle u, v\rangle| \leq \epsilon$.

Now define $A = B^T B \in \mathbb{R}^{n \times n}$. All of its diagonal entries equal to 1 and all non-diagonal elements are close to zero ($\pm\epsilon$). So A is a near-identity matrix.

Claim 2. $\epsilon < \frac{1}{\sqrt{n}} \Rightarrow rank(A) = \Omega(n)$.

Proof. (outline) The trace of A is n. The Fubini norm squared of A (which is equal to the norm of the vector that we get if we transform A to a vector of length n^2) is at most $n + \epsilon n^2$. A is a real symmetric matrix, so by the spectral theorem it has r real eigenvalues, where $r = \text{rank}(A)$, so the trace will be equal to n. The sum of the squares of the eigenvalues is given by the Fubini norm, so we can relate those two sums using Cauchy-Schwartz to get a lower bound.

Observe that if you show a lower bound on the rank of A, this gives us a lower bound on m because $\text{rank}(A) = \text{rank}(B) \leq m$. □

To get the lower bound for any ϵ (not necessarily less than $\frac{1}{\sqrt{n}}$), we define $A(k)$ with $A(k)_{i,j} = A^k_{i,j}$, and we take k big enough so that ϵ^k will be arbitrarily small. It turns out that the rank of $A(k)$ does not become much larger, and it is lower bounded, so at the end we get a lower bound for any ϵ.

3.3 KMN 2011 – Optimal DJL lower bound proof [7]

D is a distribution over $\mathbb{R}^{m \times n}$ such that for all u we have that $\mathbb{P}_{\Pi \sim D}(\Pi \text{ fails to preserve } u) < \delta$. The goal is to do this with m as small as possible.

As it holds for all u, it will also hold for a random u, so we have that $\mathbb{P}_{u \sim U}\mathbb{P}_{\Pi \sim D}(\Pi \text{ fails to preserve } u) < \delta \Rightarrow \mathbb{P}_{\Pi \sim D}\mathbb{P}_{u \sim U}(\Pi \text{ fails to preserve } u) < \delta$.

We get that $\exists\ \Pi$ such that $\mathbb{P}_{u \sim U}(\Pi \text{ fails to preserve } u) < \delta$, because otherwise every Π would be above average (i.e. we wouldn't get the probabilty above).

It remains to show that if we choose U to be uniform on a sphere (so if u is a random point on the sphere), because there is no matrix that can preserve a random point on the sphere with probability δ unless the number of rows of that matrix are at least equal to the proven lower bound. This part will not be shown in class.

3.4 LN 2016 – JL Optimal lower bound proof [9]

3.4.1 Overview

The technique that we will follow will be to construct an encoding (injection) argument and use the pigeonhole principle. In particular this is going to be an existence proof, in which we will show that

there exists a hard point-set X with the desired properties (unlike proofs so far which explicitly constructed it).

Let $\mathcal{X}$ be a collection of n-size subsets of $\mathbb{R}^d$. If for every $X \in \mathcal{X}$ there exists a $(1+\epsilon)$-distortion embedding $f_X : X \to \mathbb{R}^m$, we will show that there exists an encoding (injection) function Enc: $\mathcal{X} \to \{0,1\}^{cnm}$. This implies that $2^{cnm} \geq |\mathcal{X}| \Rightarrow m \gtrsim \frac{\log|\mathcal{X}|}{n}$.

3.4.2 Outline

We want to construct an $\mathcal{X}$ that is as big as possible (as this gets us a better lower bound), and show how the encoding works (prove that it is an injection). First, we will construct a collection $\mathcal{X}$ of ordered multisets of size n, so every point set is a tuple of points, which means that the order matters. *Note that here we are slightly changing the definition given above.*

We will proceed by contradiction. Suppose that every point-set is easy, i.e. it has an embedding into dimension m. In particular every point-set in $\mathcal{X}$ has an embedding into dimension m, say f_X. We will use the collection of embeddings, to define an encoder taking a point set to a bit-stream of length cnm. We will show that this encoder is an injection, which would mean that the right side would be bigger than the left side, which would mean that $cnm \geq |\mathcal{X}| \Rightarrow m \geq \frac{\log|\mathcal{X}|}{n}$.

Set $k = \frac{c^2}{\epsilon^2}$, define for $S \subseteq [d], |S| = k$, $y_S = \frac{1}{\sqrt{k}} \sum_{i \in S} e_i$, so y_S is the indicator vector of the set S, normalized so that it has unit norm. We are going to consider point sets of the form $X = (0, e_1, e_2, \ldots, e_d, y_{S_1}, \ldots, y_{S_{n-d-1}})$, so $\mathcal{X}$ is the collection of sets of this form.

For brevity we will solve this for a specific value of d. Set $d = \frac{n}{\log \frac{1}{\epsilon}}$.

Then $|\mathcal{X}| = \binom{d}{k}^{n-d-1} \Rightarrow m \gtrsim \frac{(n-d-1)\log\binom{d}{k}}{n} \gtrsim k \log(\frac{d}{k}) \gtrsim \frac{1}{\epsilon^2} \log(\frac{\epsilon^2 n}{\log \frac{1}{\epsilon}})$ (by Stirling's approximation).

It remains to show the existence of the encoding.

The rest of the proof will be shown next Tuesday.

References

[1] William Johnson, Joram Lindenstrauss Extensions of Lipschitz mappings into a Hilbert space *Contemporary Mathematics*, 189–206, 1984.

[2] Noga Alon, Yossi Matias, Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[3] Noga Alon. Problems and results in extremal combinatoricsI. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[4] Piotr Indyk, Assaf Naor. Nearest-neighbor-preserving embeddings *ACM Trans. Algorithms* , 3(3):31, 2007

[5] William Johnson, Assaf Naor The Johnson-Lindenstrauss lemma almost characterizes Hilbert space, but not quite *SODA*, 885–891, 2009.

[6] T.S. Jayram, David Woodruff. Optimal Bounds for Johnson-Lindenstrauss Transforms and Streaming Problems with Sub-Constant Error *SODA*, 1–10, 2011.

[7] David Kane, Raghu Meka, Jelani Nelson Almost Optimal Explicit Johnson-Lindenstrauss Families. *APPROX – RANDOM*, 628–639, 2011.

[8] Kasper Larsen, Jelani Nelson. The Johnson-Lindenstrauss Lemma Is Optimal for Linear Dimensionality Reduction. *CoRR*, abs/1411.2404, 2014.

[9] Kasper Larsen, Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma *CoRR*, abs/1609.02094, 2016.

[10] Noga Alon, Bo'az Klartag Optimal compression of approximate inner products and dimension reduction *FOCS* (to appear) 2017.