Data Science Through a Geometric Lens Fall 2019

Lecture 05 — October 14, 2019

Prof. Cyrus Rashtchian *Topics: Dimensionality Reduction and JL*

Overview. In the last lecture we studied the AMS algorithm for estimating the ℓ_2 norm of a stream. We viewed a stream of updates $(i, 1)$ as $x_i = x_i + 1$, i.e. increasing the respective coordinate in a count vector by 1. We wanted to approximate $\|x\|_2^2$.

Today we will talk about the Johnson-Lindenstrauss (JL) sketch. This will serve as a transition between streaming/sketching and other topics, because the real point of the JL algorithm is *dimensonality reduction*. We will only spend two days on this topic, but we will see other applications later as well. Dimensionality reduction is useful for many applications, such as nearest neighbor search, clustering, regression, and convex geometry problems (e.g, minimum enclosing ball).

For historical reasons, it is often called the "Johnson-Lindenstrauss Lemma". At a high-level the JL Lemma maintains a linear sketch Ax, where A is a $k \times m$ random matrix. Then, for a vector x, we will actually look Ax, which is k-dimensional with $k \ll n$. The punchline is that this is a way to use a **low** dimensional ℓ_2 norm $||Ax||_2^2$ to estimate a *high* dimensional ℓ_2 norm $||x||_2^2$ for $x \in \mathbb{R}^n$. In general, we want to map to a lower dimensional ($k \ll n$) space while preserving distances, or angles, or volumes, etc. Moreover, we usually want to preserve approximation guarantees of some algorithm (running over Ax instead of x).

Brief interlude on metric embeddings

One way to measure this how good Ax approximates x is the distortion. In general, we can think of a map $f : X \to \ell_2^k$ from an arbitrary metric space to Euclidean space. Then, the **distortion** is the smallest number $C > 0$ such that

$$d(x, y) \ \leq \ \|f(x) - f(y)\|_2 \ \leq \ C \cdot d(x, y).$$

This is well-studied for non-linear functions f as well, and in fact, many deep learning methods will implicitly learn such a function f. Although, in machine learning, the focus is usually on embedding with small "semantic distortion" and this is much harder to formalize. We will focus mostly on linear maps $f(x) = Ax$.

We will see that a random A will work well for mapping ℓ_2 to a lower dimensional ℓ_2. But it's not always possible to find good maps for different metric spaces. Brinkman and Charikar showed that for ℓ_1, it turns out that dimensionality reduction with distortion C requires dimension $k \geq n^{\Omega(1/C^2)}$. In other words, for any n, there exists a set of n points requiring dimension $k \geq n^{\Omega(1/C^2)}$ to preserve the ℓ_1 distance using the ℓ_2 distance with distortion C. This is true for any time of embedding function f, not just linear maps.

Review of Normal Random Variables

First we recap some basics about the normal distribution. The normal distribution is distributed on the whole real number and the shape is like a bell. It has two parameters: the mean μ and the variance σ^2. We denote it by $\mathcal{N}(\mu, \sigma)$, and the density function for $\mathcal{N}(\mu, \sigma)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}$$

One important feature of the normal distribution is that it is closed under linear transformations. If X and Y are independent normally-distributed random variables, then $X + Y$ also has a normal distribution (and by independence, the mean is $\mu = \mu_1 + \mu_2$ and variance is $\sigma^2 = \sigma_1^2 + \sigma_2^2$). More information about these facts and others can be found on the Wikipedia page[1] (worth looking at).

1 The Johnson-Lindenstrauss Lemma

Theorem 1 (Johnson-Lindenstrauss Lemma)**.** *Let $\varepsilon \in (0, 1/2)$ be an accuracy parameter, and let $X = \{x_1, x_2, \ldots, x_n\}$ in $\mathbb{R}^d$ be a set of of high-dimensional points. There exists a matrix $A \in \mathbb{R}^{k \times d}$ with $k = O(\log(n)/\varepsilon^2)$ such for any $x, y \in X$, we have*

$$(1 - \varepsilon)\|x - y\|_2^2 \leq \|Ax - Ay\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2.$$

To prove this, we will show that a random mapping actually works with high probability, assuming the entries are i.i.d. standard normal (we mention ± 1 variations later).

Theorem 2 (Norm Preservation)**.** *Assume that B is a random matrix in $\mathbb{R}^{k \times d}$ with entries sampled i.i.d. from $\mathcal{N}(0, 1)$ and define $A = \frac{1}{\sqrt{k}} \cdot B$. Then, for any $x \in \mathbb{R}^d$, we have*

$$(1 - \varepsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2.$$

with probability at least $1 - 2e^{-(\varepsilon^2 - \varepsilon^3)k/4}$.

The second theorem above is often called the "Distributional JL Lemma". Notice that $\varepsilon < 1/2$ means that $\varepsilon^2 - \varepsilon^3 > \varepsilon^2/2$. So the probability of failure is at most $e^{-\varepsilon^2 k/8}$.

An important aspect of the second theorem is that we can assume without loss of generality that x has unit norm, that is, $\|x\|_2 = 1$. Why? Because A is a linear map, and therefore, $\|cAx\|_2 = c\|Ax\|_2$. So we can always renormalize x and get the same conclusion (that is, if the theorem holds for unit vectors, it holds for all vectors).

The main use of it is that it immediately implies the JL Lemma. We will set $k = O(\log(n)/\varepsilon^2)$.

Proof of Theorem 1. The proof is constructive and is an example of the probabilistic method. Choose an f which is a random projection. More precisely, set $k = \lceil \frac{24}{\varepsilon^2} \cdot \log n \rceil$ and define

$$f(x) = \frac{1}{\sqrt{k}} \cdot Bx,$$

where B is a random matrix in $\mathbb{R}^{k \times d}$ with entries sampled i.i.d. from $\mathcal{N}(0, 1)$. We define $A = B/\sqrt{k}$, so that $f(x) = Ax$.

Note that there are $\binom{n}{2} < n^2$ pairs of vectors in $x, y \in X$ to handle to prove Theorem 1. Consider one such pair, and notice that $f(x) - f(y) = A(x - y)$ by definition.

By Theorem 2, we *fail* to achieve

$$(1 - \varepsilon)\|x - y\|_2^2 \leq \|A(x - y)\|_2^2 \leq (1 + \varepsilon)\|x - y\|_2^2.$$

with probability at most $e^{-\varepsilon^2 k/8}$. Plugging in $k = \lceil \frac{24}{\varepsilon^2} \cdot \log n \rceil$, we have that this is at most $1/n^3$. We conclude by taking a union bound over the (at most) n^2 possible pairs x, y, so that the probability that we succeed is at least $1 - 1/n$. □

[1] `https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables`

Aside: the above proof technique is known as 'the probabilistic method' because the theorem is a deterministic statement while the proof is via a probabilistic argument. This can be very useful to show the existence of objects that you desire. Our bound is very good though, so we will succeed with high probability (at least $1 - 1/n$) by increasing k by only a constant.

1.1 Proving the norm preservation theorem

Claim 3. *Assume that B is a random matrix in $\mathbb{R}^{k \times d}$ with entries sampled i.i.d. from $\mathcal{N}(0,1)$ and define $A = \frac{1}{\sqrt{k}} \cdot B$. For any $x \in \mathbb{R}^d$,*

$$\mathbb{E} \|Ax\|_2^2 = \|x\|_2^2$$

Proof. The proof considers the expected value of a single entry j, which we denote as

$$[Ax]_j^2 = \frac{1}{k}[Bx]_j^2.$$

We can compute the expectation for coordinate j (where use that $\mathbb{E}\, B_{ij} = 0$ and that $\mathbb{E}\, B_{ij}^2 = 1$):

$$\begin{aligned}
\mathbb{E}[Bx]_j^2 &= \mathbb{E}\left(\sum_{i=1}^{d} B_{ij} x_i\right)^2 \\
&= \mathbb{E} \sum_{i=1}^{d} \sum_{r=1}^{d} B_{ij} B_{rj} x_i x_r \\
&= \mathbb{E} \sum_{i=1}^{d} B_{ii}^2 x_i^2 \\
&= \sum_{i=1}^{d} x_i^2\, \mathbb{E}\, B_{ii}^2 = \sum_{i=1}^{d} x_i^2 = \|x\|_2^2.
\end{aligned}$$

Therefore we have that summing over j gives:

$$\mathbb{E} \|Ax\|_2^2 = \mathbb{E} \sum_{j=1}^{m} [Ax]_j^2 = \frac{1}{k} \sum_{j=1}^{m} \mathbb{E}[Bx]_j^2 = \|x\|_2^2.$$

□

We need a concentration lemma about sums of squares of normal random variables. The proof is very similar to the proof of Chernoff bound. Also, turns out that a sum of k squared normals is called a 'chi-squared' χ^2 distribution with k degrees of freedom.[2]

Lemma 4. *Let $Z = \sum_{j=1}^{m} Z_j^2$ where each Z_j is i.i.d. from $\mathcal{N}(0,1)$. Then,*

$$\Pr(Z \geq (1+\varepsilon)k) \leq e^{-(\varepsilon^2 - \varepsilon^3)k/4}.$$

and

$$\Pr(Z \leq (1-\varepsilon)k) \leq e^{-(\varepsilon^2 - \varepsilon^3)k/4}.$$

[2] `https://en.wikipedia.org/wiki/Chi-squared_distribution`

Proof. By Markov's inequality, and letting $0 < t < 1/2$ be a parameter, we have that

$$\begin{aligned}
\Pr(Z \geq (1+\varepsilon)k) &= \Pr\left(e^{tZ} \geq e^{t(1+\varepsilon)k}\right) \\
&= \Pr\left(e^{t\sum_{j=1}^m Z_j^2} \geq e^{(1+\varepsilon)kt}\right) \\
&\leq \frac{\mathbb{E}\, e^{t\sum_{j=1}^m Z_j^2}}{e^{(1+\varepsilon)kt}} \\
&= \frac{(\mathbb{E}[e^{tZ_1^2}])^m}{e^{(1+\varepsilon)kt}} \\
&= \left(\frac{1}{1-2t}\right)^{k/2} e^{-(1+\varepsilon)kt}
\end{aligned}$$

where the final step evaluates the expectation. To see this, we use that this expectation, by definition, exactly the moment generating function for a squared normal random evaluated at t, and we know that

$$\mathbb{E}[e^{tZ_1^2}] \;=\; \frac{1}{\sqrt{2\pi}} \int e^{tz^2} e^{-z^2/2}\, dz \;=\; \frac{1}{\sqrt{1-2t}}.$$

We will choose $t = \frac{\varepsilon}{2(1+\varepsilon)}$, which will minimize the above expression (and is less than $1/2$ because $\varepsilon > 0$). Plugging t in gives

$$\Pr(Z \geq (1+\varepsilon)k) \leq \left((1+\varepsilon)e^{-\varepsilon}\right)^{k/2} \leq e^{-(\varepsilon^2-\varepsilon^3)k/4},$$

where the final step uses the upper bound

$$\ln(1+\varepsilon) \leq \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{2}$$

which comes from the Taylor expansion. □

We now prove the norm preservation Theorem 2.

Proof of Theorem 2. Recall B is a random matrix in $\mathbb{R}^{k\times d}$ with entries sampled i.i.d. from $\mathcal{N}(0,1)$, and we defined $A = B/\sqrt{k}$. Claim 3 establishes that

$$\mathbb{E}\, \|Ax\|_2^2 = \|x\|_2^2$$

To bound the failure probability, we note that $Z_j = [Bx]_j/\|x\|_2$ is distributed as $\mathcal{N}(0,1)$, and of course the variables $Z_1, \ldots, Z_m$ are independent. We can prove this as follows:

$$\mathbb{E}\, Z_j = \sum_{i=1}^d \mathbb{E}\, B_{ji} x_i = 0.$$

$$\mathrm{Var}(Z_j) = \mathbb{E}(Z_j^2) = \sum_{i=1}^d \mathbb{E}\, B_{ji}^2 x_i^2 = \|x\|_2^2.$$

For ease of notation, we now use that $\|x\|_2 = 1$ without loss of generality. Therefore, we see that

$$\|Ax\|_2^2 > (1+\varepsilon)$$

is equivalent to (using $A = B/\sqrt{k}$)

$$\|Bx\|_2^2 > (1+\varepsilon)k$$

is equivalent to (using $Z_j = [Bx]_j$)

$$\sum_{j=1}^{k} Z_j^2 > (1+\varepsilon)k.$$

Invoking Lemma 4 we get the desired bound on the probability, where the factor of 2 comes from union bounding over the $(1-\varepsilon)$ case as well. □

1.2 Interpreting AMS as dimensionality reduction

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. 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 $\frac{1}{\sqrt{k}}$ (k 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 History of lower bounds

In 2011 it was resolved that the DJL lower bound is optimal [9]. Observe that the derivation of JL from DJL does not imply that the lower bound of k 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 [2], they show that there exist $n+1$ points in $\mathbb{R}^n$ such that any JL map needs $k \gtrsim \frac{\log n}{\log(2c+1)}$. This lower bound is achievable for large C, as shown by Indyk and Naor [6].
- Alon [5] showed that the for same point set that JL analyzed in their paper, we can achieve a better lower bound: $k \gtrsim \min\{n, \frac{1}{\epsilon^2}\frac{\log n}{\log \frac{1}{\epsilon}}\}$ 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 [8] and one by Kane, Meka and Nelson [9] which show that for DJL $k \gtrsim \min\{d, \frac{1}{\epsilon^2}\log\frac{1}{\delta}\}$. 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 [10] showed that $\forall \epsilon, n$, there exist $O(n^3)$ points in $\mathbb{R}^n$ such that any **linear** map needs $k \gtrsim \min\{n, \frac{1}{\epsilon^2}\log n\}$.

  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 [11] 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 $k \gtrsim \frac{1}{\epsilon^2}\log n$, 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 [12] showed the lower bound of the conjecture above.

Regarding results in L_p, there is a theorem by Johnson and Naor [7] 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.

2.1 Proof of the original JL lower bound [2]

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

Claim 5. *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 $m \geq \frac{\log n}{\log(2c+1)}$. □

2.2 Proof of the lower bound by Alon [5]

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

Proof. (outline) The trace of A is n. The Frobenius 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 Frobenius 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 k 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.

References

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

[2] Johnson, William B., and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics.*, 26(1):189–206, 1984.

[3] Indyk, Piotr. Stable distributions, pseudorandom generators, embeddings and data stream computation. *Journal of the ACM (JACM)* :53.3 (2006): 307-323.

[4] Li, Ping. Estimators and tail bounds for dimension reduction in $\ell_\alpha (0 < \alpha \leq 2)$ using stable random projections. *J. Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics*, 2008.

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

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

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

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

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

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

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

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