ECS 189A Sublinear Algorithms for Big Data **Fall 2024**

Lecture 8: JL Lemma + Kirszbraun's Extension Theorem

Lecturer: Jasper Lee *Scribe: Andrew Yeow*

1 Johnson-Lindenstrauss Lemma

Consider the following setting where we have a set:

$$S = \{v_1, ..., v_n\} \subseteq \mathbb{R}^d$$

where d (the dimension of the points) is very large, $n \ll d$. We want to find a function f: $\mathbb{R}^d \to \mathbb{R}^k$ for $k \ll d$. This would turn our dataset into $f(S) = \{f(v_1), ..., f(v_n\} \subseteq \mathbb{R}^k$. The goal for the function f is to preserve pairwise distances (approximately). The big question is, "How small can k be?".

Proposition 8.1. *It suffices to take $k = n$.*

This is a trivial result because:

$$span(S) \leq n$$

Rotate $\mathbb{R}^d$ such that $span(S)$ occupies the first $k = n$ coordinates and the rest are 0.

Note that the minimum k feasible is independent of d.

Theorem 8.2 (Johnson-Lindenstrauss Lemma)**.** *Fix S and some $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$. It suffices to take*

$$k = O\left(\frac{\log(n) + \log(\frac{1}{\delta})}{\epsilon^2}\right)$$

so that there exists a random matrix A: $\mathbb{R}^d \to \mathbb{R}^k$ such that with probability $\geq 1 - \delta$, for all $v_i, v_j \in S$, the squared distance between Av_i and Av_j should be within a $1 \pm \epsilon$ factor of the true squared distance $\|v_i - v_j\|^2$. That is,

$$(1 - \epsilon)||v_i - v_j||_2^2 \leq ||Av_i - Av_j||_2^2 \leq (1 + \epsilon)||v_i - v_j||_2^2$$

If we care about Euclidean distance one important note is that taking the square root of both sides results in an expression that is roughly equivalent because:

$$\sqrt{1 - \epsilon} \approx 1 + \frac{\epsilon}{2}$$

.

Remark 8.3.

$$k = \Omega\left(\frac{\log(n)}{\epsilon^2}\right)$$

dimensions are necessary even for non-linear maps. (Larsen, Nelson 2017)

We make the first observation in order to prove the JL lemma. Note that by linearity, $||Av_i - Av_j|| = ||A(v_i - v_j)||$. Therefore we can equivalently consider the preservation of the norm of the pairwise vector differences, instead of preserving pairwise distances.

Lemma 8.4 (Norm Preservation)**.** *Fix $v \in \mathbb{R}^d$, consider a random $k \times d$ matrix G whose entries are i.i.d. $\mathcal{N}(0,1)$. Then*

$$P\left((1-\epsilon)||v||_2^2 \leq ||\frac{1}{\sqrt{k}}Gv||_2^2 \leq (1+\epsilon)||v||_2^2\right) \geq 1 - 2\exp\left(\frac{-k(\epsilon^2-\epsilon^3)}{4}\right)$$

In other words, the norm of a vector v is preserved by G with a high probability. We tune k such that the failure probability (the exponential on the right hand side) is sufficiently small.

A notable remark is that the k expression is special, because we have an additive $(\log(\frac{1}{\delta}))$ term instead of a multiplicative term.

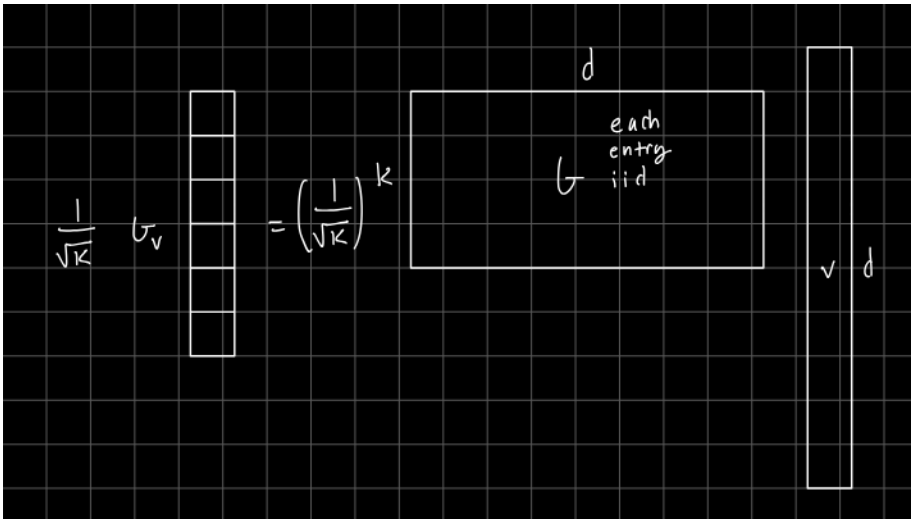
Proof of JL Lemma based on Lemma 8.4. Apply Lemma 8.4 to all the pairwise vector differences $v_i, v_j : v_i, v_j \in S$. This means that there are $O(n^2)$ such vectors. By Lemma 8.4, each pair fails with probability $\leq O(\frac{\delta}{n^2})$. We will set k:

$$k = O\left(\frac{\log(n) + \log(\frac{1}{\delta})}{\epsilon^2 - \epsilon^3}\right)$$

The denominator term $\epsilon^2 - \epsilon^3 = \Theta(\epsilon^2)$. If we substitute k in the JL Lemma and apply the union bound we get that all the pairwise distances are within this multiplicative bound with probability $\geq 1 - \delta$.

□

Proof of Lemma 8.4. Fix $v \in \mathbb{R}^d$.



Consider the norm of the vector Gv. This is a random Gaussian matrix G of dimension $k \times d$ multiplied by a fixed vector, v. Each entry in G is $\mathcal{N}(0,1)$ and iid. The question is, "What ends up still being independent?"

Some observations are that $(Gv)_i$ is the dot product of the i-th row of G with v. This is equivalent to $\sum \mathcal{N}(0,1)v_j$. If you scale a Gaussian by a constant, v_j, the variance is

v_j^2 giving $\sum \mathcal{N}(0, v_j^2)$. Now we are summing a bunch of independent Gaussian random variables. This is really just the Gaussian where we sum the variances together, which is equivalent to $||v||_2^2 \mathcal{N}(0,1)$. Therefore, each entry of the random projection is distributed as something that tells us something about the norm of v.

Another observation is that any two entries $(Gv)_i$ and $(Gv)_j$ in the random projection are independent, because all the rows of G are independent from each other, and so their dot products with a fixed v are independent.

We care about the norm $||\frac{1}{\sqrt{k}} Gv||_2^2$, which is equal to $\frac{1}{k}\sum (Gv)_i^2$. This is an average of k i.i.d. terms, so we expect concentration behavior. Using the previous observations, we get that the norm $||\frac{1}{\sqrt{k}} Gv||_2^2$ is equal to $||v||_2^2 \frac{1}{k} \sum (\mathcal{N}(0,1))^2$. So, it suffices to show that the expectation of the square of a Gaussian is exactly equal to 1. If that were the case, then the expected squared norm of $\frac{1}{\sqrt{k}} Gv$ would be $\|v\|^2$.

Squaring a Gaussian is called a Chi-Squared distribution, $(\mathcal{N}(0,1))^2 = \mathcal{X}^2$ distribution and the expectation of that distribution is the variance of $\mathcal{N}(0,1) = 1$, because of mean 0.

We want to upper bound

$$\mathbb{P}\left(||\frac{1}{\sqrt{k}} G_V||_2^2 > (1+\epsilon)||v||_2^2 \right)$$

Using the observations from above, this expression is equivalent to:

$$\mathbb{P}\left(\sum_{m=1}^{k} Z_m^2 > (1+\epsilon)k \right)$$

where Z_m is i.i.d. $\mathcal{N}(0,1)$. By applying the standard Chernoff trick, this is bounded by:

$$\left(\frac{(M_{\mathcal{X}^2}(t))}{\exp\left(t(1+\epsilon)\right)} \right)^k$$

for any $t > 0$.

The MGF for a Chi-Square distribution is $\sqrt{\frac{1}{1-2t}}$. Plugging in the MGF we get:

$$\left(\frac{\sqrt{\frac{1}{1-2t}}}{\exp\left(t(1+\epsilon)\right)} \right)^k$$

for $t \in (0, \frac{1}{2})$. Pick $t = \frac{\epsilon}{2(1+\epsilon)}$ gives $\left((1+\epsilon)\exp\left(-\epsilon\right)\right)^{\frac{k}{2}}$.

Using the inequality $\ln(1+\epsilon) \leq \epsilon - \frac{\epsilon^2 - \epsilon^3}{2}$, we can claim that:

$$\left((1+\epsilon)\exp -\epsilon\right)^{\frac{k}{2}} \leq \exp\left(\frac{-k}{4}(\epsilon^2 - \epsilon^3)\right)$$

The lower bound would use $\ln(1-\epsilon) \leq -\epsilon - \frac{\epsilon^2}{2}$. (not proven).

□

Lemma 8.5. *Lemma 8.4 also holds for random matrix A with* $\text{Unif}(\pm 1)$ *entries.*

The proof of this is shown by some analogous concentration inequality.

Corollary 8.6 (Inner or Dot Product Preservation)**.** *One implication of JL is that a random projection also preserves inner products. Fix unit vectors u,v* $\in \mathbb{R}^d$*. Take a random Gaussian matrixG. Then:*

$$\mathbb{P}(|u \cdot v - (\frac{1}{\sqrt{k}}Gu)(\frac{1}{\sqrt{k}}Gv)| > \epsilon) \leq 4 \exp \frac{-k(\epsilon^2 - \epsilon^3)}{4}$$

This is saying that for unit vectors or constant length vectors, the dot products are basically preserved.

Proof of Inner or Dot Product Preservation. Apply Lemma 8.4 to the vectors, $u + v$ and $u - v$. We get:

$$(1 - \epsilon)||u + v||_2^2 \leq ||\frac{1}{\sqrt{k}}G(u + v)||_2^2 \leq (1 + \epsilon)||u + v||_2^2$$

We have an equivalent statement for $u - v$:

$$(1 - \epsilon)||u - v||_2^2 \leq ||\frac{1}{\sqrt{k}}G(u - v)||_2^2 \leq (1 + \epsilon)||u - v||_2^2$$

Using the union bound, the failure probability is upper bounded by:

$$4 \exp \frac{-k(\epsilon^2 - \epsilon^3)}{4}$$

Finally, observe that

$$4(\frac{1}{\sqrt{k}}Gu)(\frac{1}{\sqrt{k}}Gv)$$

is equal to

$$||\frac{1}{\sqrt{k}}G(u + v)||_2^2 - ||\frac{1}{\sqrt{k}}G(u - v)||_2^2 \geq 4u \cdot v - 2\epsilon(||u||^2 + ||v||^2)$$

which is upper bounded by

$$4u \cdot v - 4\epsilon$$

and because we assumed u and v have unit norm, meaning $(||u||^2 + ||v||^2)$ is equal to 2. □

2 Statement of Kirszbraun's Extension Theorem

Problem 8.7. Consider the following example problem, k-medians. We have some set of n points $S \subset \mathbb{R}^d$ and our goal is to partition S into k subsets $C = \{C_1, ..., C_k\}$, such that we minimize the cost of the clustering. We want to find the centers $\{c_1, ..., c_k\}$. The cost is defined as:

$$cost(C) = \sum_{i=1}^{k} \min_{C_i \in \mathbb{R}^d} \sum_{x \in C_i} ||x - c_i||_2$$

We want dimensionality reduction so that we can reduce computation. One simple strategy would be to take the set of inputs and apply JL. We'll do something slightly more sophisticated. If we consider optimal clusters C^* with centers $\{c_i^*\} \in \mathbb{R}^d$. The set of points we are considering is the set of input points and the set of optimal centers (that exist and are definable from the input points, but we haven't computed what they are yet), and there are $O(n)$ of these points. Apply JL with $O(\frac{\log n}{\epsilon^2})$ dimensions and will get all pairwise distances preserved within a $1 \pm \epsilon$ factor.

$$cost(\pi C^*) = \sum_{i=1}^{k} \sum_{x \in C_i} ||\pi x - \pi c_i||_2 \approx (1+\epsilon) cost_S(C^*)$$

This means that if there is a good solution in the original high dimensional space, we will be able to find a good solution in the lower dimensional space. However, there is an issue. The issue is that we need to guarantee that we have not created some fake solutions in the lower dimensional space.

Definition 8.8 (L-Lipschitz Function)**.** Given $X \subseteq \mathbb{R}^k, Y \subseteq \mathbb{R}^d, f : X \to Y$ is L-Lipschitz if:

$$\forall x, y \in X, ||f(x) - f(y)|| \leq L||x - y||$$

Kirszbrauns's Extension Theorem shows that we can extend L-Lipschitz functions.

Theorem 8.9 (Kirszbraun's Theorem)**.** *For any subset $U \subseteq \mathbb{R}^k$, L-Lipschitz function $\varphi : U \to \mathbb{R}^d$, there is an extension $\widetilde{\varphi} : \mathbb{R}^k \to \mathbb{R}^d$, such that:*

1. *$\widetilde{\varphi}(u) = \varphi(u), u \in U$*
2. *$\widetilde{\varphi}(u)$ is also L-Lipschitz.*

Theorem 8.10 (Dimensionality Reduction for k-median)**.** *Given set $S \subseteq \mathbb{R}^d$ with optimal solution C^* with centers $\{c_i^*\}$. Suppose π satisfies the JL guarantees for $S \cup \{c_i^*\}$. This projection is taken from $\mathbb{R}^d \to \mathbb{R}^k$. Then*

$$(1 - O(\epsilon)) cost_S(C^*) \leq \min_C cost_{\pi S}(C) \leq (1 + O(\epsilon)) cost_S(C^*)$$

Using JL we can prove that:

$$\min_C cost_{\pi S}(C) \leq cost_{\pi S}(\pi C^*) \leq (1 + O(\epsilon)) cost_S(C^*)$$

Now we want to show that there are no fake solutions in low dimension:

$$(1 - O(\epsilon)) cost_S(C^*) \leq \min_C cost_{\pi S}(C)$$

Let C_π^* with centers $\{c_{\pi,i}^*\}$ be an optimal solution in low dimension. Let $U = \pi(S \cup \{c_i^*\})$ and let $\varphi = \pi^{-1}$. This is just taking the projected point back to the unprojected point and is well-defined (with probability 1 over π). The JL Lemma gives us that:

$$\varphi \text{is} \frac{1}{\sqrt{1-\epsilon}} \text{–Lipschitz}$$

Now we use Kirszbraums's to extend φ to $\widetilde{\varphi}$ on all of $\mathbb{R}^m$. This is $\frac{1}{\sqrt{1-\epsilon}}$–*Lipschitz* as well, giving us:

$$cost_S(C^*) \leq cost_S(\widetilde{\varphi} C_\pi^*) \leq \frac{1}{\sqrt{1-\epsilon}} cost_{\pi S}(C_\pi^*)$$

The main idea is that the JL Lemma allows us to map from a high-dimensional space to a low-dimensional one, while Kirzbraun's allows us to move from a low-dimensional space to a high-dimensional one.