**CSCI 1951-W Sublinear Algorithms for Big Data** **Fall 2020**

# Lecture 10: JL Lemma and Kirszbraun's Extension Theorem

*Lecturer: Jasper Lee* *Scribe: Tyler DeFroscia*

## 1 The Johnson-Lindenstrauss Lemma

Consider the following setting: we have some set

$$S = \{v_1, \ldots, v_n\} \subset \mathbb{R}^d$$

where $d$ is very large, say $n \ll d$. We want to find some function $f : \mathbb{R}^d \to \mathbb{R}^k$ with $k \ll d$, so that $f$ maps $S$ into some lower dimensional space. Moreover, we want $f$ to approximately preserve pairwise distances, so that $\|f(v_i) - f(v_j)\| \approx \|v_i - v_j\|$.

The immediate first question is: how small can we make $k$? The lower the dimension we can project down to the greater the possible benefits, but there must be some lower limit at which we can't preserve pairwise distances sufficiently.

A naive first answer is $n$. Consider $\text{span}\{v_i\} \subset \mathbb{R}^d$. This is a linear subspace of $\mathbb{R}^d$ with dimension $\leq n$. We can then take the orthogonal projection of $S$ onto this subspace, which projects down to $\leq n$ dimensions while preserving pairwise distances.

Even though this is just a first attempt, it gives us an upper bound that's independent of $d$, which suggests that our minimum answer might also be independent of $d$. The Johnson-Lindenstrauss Lemma gives us a better answer to this minimum dimension question.

**Theorem 10.1** (Johnson-Lindenstrauss Lemma) *Fix $S$ as before and take $\epsilon, \delta > 0$. It suffices to take*

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

*so that there exists a random linear map $A : \mathbb{R}^d \to \mathbb{R}^k$ such that with probability $\geq 1 - \delta$ we have $\forall v_i, v_j \in S$ we have*

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

A couple of notes here. First, while in our initial setup we cared about the pairwise distances and here we are using the squared distances, this doesn't matter much, as for small $\epsilon$ we have $\sqrt{1 + \epsilon} \approx 1 + \epsilon/2$. Second, the introduction of the $\delta$ term here isn't necessary if we only care about the existence of such a random map $A$. Indeed, if we only care about existence, then it suffices to take

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

as it can be shown that with this $k$ there is a non-zero probability of finding such an $A$, and so by the probabilistic method $A$'s existence is guaranteed. The $\delta$ term allows us to control the probability of finding such an $A$ and is therefore more of an algorithmic concern.

**Remark 10.2** Larsen and Nelson showed that the bound given by the JL Lemma is tight, ie. the number of dimensions we need is

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

even for non-linear maps.

Before diving into the proof, it's worth quickly noting that by linearity $\|Av_i - Av_j\| = \|A(v_i - v_j)\|$, and so we can work with the norm of the pairwise differences instead of the pairwise distances. We begin with a lemma.

**Lemma 10.3** (Norm Preservation Lemma) *Fix $v \in \mathbb{R}^d$. Consider a random $k \times d$ Gaussian matrix whose entries are all iid $\mathcal{N}(0,1)$. Then*

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

In less formal language, the lemma says that $G$ *approximately preserves the length of $v$ with high probability.* We first see how to prove the JL Lemma using this lemma, and then return to its proof.

*Proof of the JL Lemma.* Apply Lemma 10.3 to the pairwise differences $v_i - v_j$. There are $O(n^2)$ such distances. If we take $k$ as in the JL Lemma, then by Lemma 10.3 we are guaranteed that

$$1 - P\left((1-\epsilon)\|v_i - v_j\|^2 \leq \left\|\frac{1}{\sqrt{k}}G(v_i - v_j)\right\|^2 \leq (1+\epsilon)\|v_i - v_j\|^2)\right) \leq O\left(\frac{\delta}{n^2}\right)$$

and so by applying a union bound, we have that *all* the pairwise distances are within this multiplicative bound with probability $\leq O(\delta)$, which proves the lemma. □

We now return to prove Lemma 10.3.

*Proof of Lemma 10.3.* Since $G$ is a $k \times d$ random matrix, we can write $Gx$ in terms of a matrix product. Using the usual rules of matrix multiplication and the fact that each entry $G_{ij}$ is $\mathcal{N}(0,1)$ and iid, we have that

$$\begin{aligned}
(Gx)_i &= \sum_j G_{ij}x_j && (1)\\
&= \sum_j \mathcal{N}(0,1)x_j && (2)\\
&= \sum_j \mathcal{N}(0,x_j)^2 && (3)\\
&= \mathcal{N}(0,\|x\|^2) && (4)
\end{aligned}$$

where $(Gx)_i$ denotes the $i$-th entry of the resulting vector. We note also that since the entries are iid, $(Gx)_i$ and $(Gx)_j$ are independent for $i \neq j$. Finally, note also that $\|Gx\|^2 = \sum_i (Gx)_i^2 = \sum_i \mathcal{N}(0,\|x\|^2)^2$.

Recall that $\mathcal{N}(0,1)^2 = \chi^2$, the Chi-squared distribution, and that $E\chi^2 = 1$. Thus, we have that

$$\begin{aligned}
E\left(\left\|\frac{1}{\sqrt{k}}Gx\right\|^2\right) &= \frac{1}{k}E\left(\sum_i (Gx)_i^2\right) && (5)\\
&= \|x\|^2 && (6)
\end{aligned}$$

so the expected value of the projected square norm is $\|x\|^2$. This should inspire some confidence, as this suggets we can apply concentration inequalities to control the probability

that the square norm varies by some amount. Indeed, we will apply a Chernoff bound here to bound the probability. We have that

$$P\left(\left\|\frac{1}{\sqrt{k}}Gx\right\|^2 > (1+\epsilon)\|x\|^2\right) = P\left(\sum_{i=1}^{k} Z_i^2 > (1+\epsilon)k\right) \tag{7}$$

where the $Z_i \sim \mathcal{N}(0,1)$. Following the usual steps for computing a Chernoff bound we proceed:

$$= P\left(\exp\left(t\sum Z_i^2\right) > \exp\left(t(1+\epsilon)k\right)\right) \text{ , for } t > 0 \tag{8}$$

$$\leq \frac{(M_{\chi^2}(t))^k}{\exp\left(t(1+\epsilon)k\right)} \tag{9}$$

$$= \left(\frac{M_{\chi^2}(t)}{\exp\left(t(1+\epsilon)\right)}\right)^k \tag{10}$$

Plugging in for the MGF gives us

$$= \left(\exp\left(-t(1+\epsilon)\right)\left(\frac{1}{1-2t}\right)^{1/2}\right)^k \text{ for } t < 1/2 \tag{11}$$

One can compute via usual calculus techniques that the maximum is at $t = \frac{\epsilon}{2(1+\epsilon)} < 1/2$, which gives us the bound

$$\leq \left((1+\epsilon)e^{-\epsilon}\right)^{k/2} \tag{12}$$

We simplify this by using the Taylor expansion of $e^x$, from which we can see that

$$\log(1+\epsilon) \leq \epsilon - \frac{\epsilon^2 - \epsilon^3}{2} \tag{13}$$

and so taking the exponential and plugging back into the bound gives us the new bound of

$$\leq \exp\left(\frac{-k(\epsilon^2-\epsilon^3)}{4}\right) \tag{14}$$

which is what we wanted. Performing the same analysis with the other side of the inequality and using the bound $\log(1-\epsilon) \leq -\epsilon - \epsilon^2/2 + \epsilon^3/2$ gives us the desired bound. □

One might wonder why we had to use Gaussians in the previous lemma. A nice variant of this shows that other simpler distributions will also work:

**Lemma 10.4** *Lemma 10.3 also holds if we consider a random matrix $A$ with* Unif$\{\pm 1\}$ *entries.*

We also get a corollary that essentially tells us that these random maps also preserve inner products for unit vectors:

**Corollary 10.5** *Fix unit vectors $u, v \in \mathbb{R}^d$, and consider a random $k \times d$ Gaussian matrix $G$. Then*

$$P\left(\left|u \cdot v - \left(\frac{1}{\sqrt{k}}Gu\right) \cdot \left(\frac{1}{\sqrt{k}}Gv\right)\right| > \epsilon\right) \leq 4\exp(-k(\epsilon^2-\epsilon^3)/4) \tag{15}$$

*Proof.* To ease notation, we denote the scaled random Gaussian matrix as $f$. Apply Lemma 10.3 to $u+v$ and $u-v$ to get with probability $\geq 1-4\exp(-k(\epsilon^2-\epsilon^3)/4)$ that

$$(1-\epsilon)\|u\pm v\|^2 \leq \|f(u\pm v)\|^2 \leq (1+\epsilon)\|u\pm v\|^2 \tag{16}$$

By expanding the inner product $f(u)\cdot f(v)$, we get

$$\begin{aligned} 4f(u)\cdot f(v) &= \|f(u+v)\|^2 - \|f(u-v)\|^2 && (17)\\ &\geq 4u\cdot v - 2\epsilon\left(\|u\|^2+\|v\|^2\right) && (18)\\ &\geq 4u\cdot v - 4\epsilon && (19)\end{aligned}$$

which implies that $f(u)\cdot f(v)\geq u\cdot v-\epsilon$ with the above probability. Performing the same process for the other side and combining yields the desired inequality. □

## 2 Kirszbraun's Extension Theorem

Consider the $k$-medians problem in $\mathbb{R}^d$. We have some set of $n$ points $S\subset\mathbb{R}^d$, and our goal is to partition $S$ into $k$ subsets $\mathcal{C}=\{C_1,\ldots,C_k\}$ so as to minimize the cost of the clustering, defined as

$$\text{cost}(\mathcal{C}) = \sum_{i=1}^{k}\min_{c_i\in\mathbb{R}^d}\sum_{x\in C_i}\|x-c_i\|_2 \tag{20}$$

We refer to the $c_i$ as the centers of the clusters.

We'd like to apply some sort of dimesionality reduction to solve this $k$-medians problem. It seems natural to apply the JL Lemma here, but the question is on which set to apply it. If we consider the set $S\cup\{c_i\}_i$ of the points along with the cluster centers, we can apply JL to project into a lower dimension while preserving the relevant pairwise distances. Indeed, if we let $m=O(\log n/\epsilon^2)$ and consider the random Gaussian map $\pi=\frac{1}{\sqrt{m}}G$, then we know that all the pairwise distances will be preserved up to a $(1+\epsilon)$ factor, so that

$$\text{cost}_{\pi S}(\pi\mathcal{C}^*) = \sum_{i=1}^{k}\sum_{x\in C_i}\|\pi x-\pi c_i\|_2 \leq (1+\epsilon)\text{cost}_S(\mathcal{C}^*) \tag{21}$$

In other words, the JL Lemma guarantees that any high-dimesional solution is still a low-dimensional solution, up to this $(1+\epsilon)$ factor. However, we still have another concern: is it possible that $\pi$ can create "fake" solutions in $\pi S$ that are significantly better than $\mathcal{C}^*$ on $S$?

Fortunately, the answer is no, due to Kirszbraun's Extension Theorem. First, recall the definition of an $L$-Lipschitz function.

**Definition 10.6** Given $X\subseteq\mathbb{R}^m$, $Y\subseteq\mathbb{R}^d$ and a function $f:X\to Y$, we say $f$ is $L$-Lipschitz if

$$\forall x_1,x_2\in X, \|f(x_1)-f(x_2)\|\leq L\|x_1-x_2\| \tag{22}$$

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

**Theorem 10.7** *(Kirszbraun's Extension Theorem) For any subset $U\subseteq\mathbb{R}^m$ and an $L$-Lipschitz function $\varphi:U\to\mathbb{R}^d$, there exists an extension $\widetilde{\varphi}:\mathbb{R}^m\to\mathbb{R}^d$ such that*

1. $\widetilde{\varphi}_{|U} = \varphi$ *(i.e. $\widetilde{\varphi}$ is an extension of $\varphi$)*

2. $\widetilde{\varphi}$ *is L-Lipschitz*

We now have everything we need to solve $k$-medians in lower dimensions.

**Theorem 10.8** *Given a set $S \subseteq \mathbb{R}^d$, consider an optimal clustering $\mathcal{C}^*$ with centers $\{c_i\}$. Suppose $\pi$ satisfies the JL Lemma conditions for $S \cup \{c_i\}$. Then*

$$(1 - O(\epsilon))\text{cost}_S(\mathcal{C}^*) \leq \min_{\mathcal{C}} \text{cost}_{\pi S}(\mathcal{C}) \leq (1 + O(\epsilon))\text{cost}_S(\mathcal{C}^*) \tag{23}$$

*Proof.* The inequality

$$\min_{\mathcal{C}} \text{cost}_{\pi S}(\mathcal{C}) \leq (1 + O(\epsilon))\text{cost}_S(\mathcal{C}^*) \tag{24}$$

follows from a relatively straightforward application of JL as follows: applying the lemma to $S \cup \{c_i\}$ gives us that

$$\text{cost}_{\pi S}(\pi\mathcal{C}^*) \leq (\sqrt{1+\epsilon})\text{cost}_S(\mathcal{C}^*) \tag{25}$$

and furthemore we have

$$\min_{\mathcal{C}} \text{cost}_{\pi S}(\mathcal{C}) \leq (\sqrt{1+\epsilon})\text{cost}_S(\mathcal{C}^*) \tag{26}$$

which gives us the right hand side of the inequality.

For the other inequality, we use Kirszbraun's Extension Theorem. Let $\mathcal{C}^*_\pi$ with centers $\{c_{\pi i}\}$ be an optimal solution for $\pi S$. Let $U = \pi(S \cup \{c_i\})$ and define $\varphi = \pi^{-1}$, restricting the domain to the image of $U$ to prevent well-definedness issues. The JL Lemma gives us that $\varphi$ is $1/\sqrt{1-\epsilon}$-Lipschitz, and so we invoke Kirszbraun's to extend $\varphi$ to $\widetilde{\varphi}$ on all of $\mathbb{R}^m$. This is $1/\sqrt{1-\epsilon}$-Lipschitz as well, and so

$$\text{cost}_S(\mathcal{C}^*) \leq \text{cost}_S(\widetilde{\varphi}\mathcal{C}^*_\pi) \leq \frac{1}{\sqrt{1-\epsilon}}\text{cost}_{\pi S}(\mathcal{C}^*_\pi) \tag{27}$$

which gives us the right hand side of the inequality. □

To summarize, the idea is that the JL Lemma lets us move from high-dimensional solutions to low-dimensional ones, and Kirszbraun's lets us move from low-dimensional ones to high-dimensional ones.