**15-750: Graduate Algorithms** March 24, 2017

Lecture 23: Dimensionality reduction, Johnson-Lindenstrauss Lemma

*Lecturer: David Witmer* *Scribe: Oleksandr Rudenko*

# 1 Introduction

## 1.1 Intuition and setup

Suppose we are given a set of $n$ points $\{x_1, \dots, x_n\}$ in a large-dimensional space $\mathbb{R}^d$. And let's say we want to perform some operations on this points that depends only on their pairwise distances (like find two closest points etc). Then, the lower the dimension $d$, the less computation we have to do. The goal of this lecture is to find out the way to project given points to a lower-dimensional space such that the distances are preserved.

How small can we make $d$ and still maintain the Euclidean distances between the points? Clearly, we can always make $d = n-1$, since any set of $n$ points lies on a $n-1$-dimensional subspace. And this is (essentially) tight: e.g., the case when $x_2 - x_1, x_3 - x_1, \dots, x_n - x_1$ are all orthogonal vectors.

But what if we were OK with the distances being approximately preserved? In fact, while there could only be $d$ orthogonal unit vectors in $\mathbb{R}^d$, there could be as many as $e^{c\varepsilon^2 d}$ unit vectors which are $\varepsilon$-orthogonal, i.e., whose mutual inner products all lie in $[-\varepsilon, \varepsilon]$. Near-orthogonality allows us to pack exponentially more vectors!

So, if we wanted $n$ points exactly at unit (Euclidean) distance from each other, we would need $n-1$ dimensions (equilateral triangle in $\mathbb{R}^2$, regular tetrahedron in $\mathbb{R}^3$). But if we wanted to pack in $n$ points which were at distance $(1 \pm \varepsilon)$ from each other, we could pack them into

$$O\left(\frac{\log n}{\varepsilon^2}\right)$$

dimensions.

## 1.2 The Johnson Lindenstrauss lemma

**Lemma (Johnson-Lindenstrauss):** Let $\varepsilon \in (0, \frac{1}{2})$. Then for **any** set of points $S = \{x_1, \dots, x_n\}$ in $\mathbb{R}^d$, there exists a mapping $A : \mathbb{R}^d \to \mathbb{R}^k$ with $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ such that

$$\forall x_1, x_2 \in S: \ \|x_1 - x_2\| (1-\varepsilon) \le \|Ax_1 - Ax_2\| \le \|x_1 - x_2\| (1+\varepsilon)$$

Note that the target dimension $k$ is independent of the original dimension $d$, and depends only on the number of points $n$ and the accuracy parameter $\varepsilon$.

It is easy to see that we need at least $\Omega\left(\frac{\log n}{\varepsilon}\right)$ using a packing argument. Noga Alon showed a lower bound of $\Omega\left(\frac{\log n}{\varepsilon^2 \log \frac{1}{\varepsilon}}\right)$. Recently, Larsen and Nelson showed that any *linear* dimensionality reduction scheme must require $\Omega\left(\frac{\log n}{\varepsilon^2}\right)$ dimensions for some data sets. Very recently (2016) they also showed that it is true for *any* map (but with some constraints on $\varepsilon$, which is not very significant).

# 2 Main procedure

## 2.1 Construction of the map

The JL lemma is pretty surprising, but the construction of the map is perhaps even more surprising: it is a super-simple random construction. Let $M$ be a $k \times d$ matrix, such that every entry of $M$ is filled with an i.i.d. draw from a standard normal $N(0, 1)$ distribution. Then our $A$ will be a linear transformation assosiated with a scaled matrix $\frac{1}{\sqrt{k}}M$. That is $A(x) = \frac{1}{\sqrt{k}}Mx$

Thats it. Note that because $A$ assosiated with a matrix multiplication, it is linear.

**Lemma 2:** Let $\varepsilon \in (0, \frac{1}{2})$. If $A$ is constructed as above with $k = c\varepsilon^{-2} \log \delta^{-1}$, and $x \in \mathbb{R}^d$ is a unit vector, then

$$\mathbf{Pr}[\ \|A(x)\|^2 \in [1 - \varepsilon, 1 + \varepsilon]\ ] \geq 1 - \delta$$

If we are able to prove this lemma, then we can prove JL lemma. Indeed, let's set $\delta = \frac{1}{n^2} \Rightarrow k = O\left(\frac{\log n}{\varepsilon^2}\right)$. Now for each $x_i, x_j \in S$ we get that the squared length of $x_i - x_j$ is maintained to within $1 \pm \varepsilon$ with probability at least $1 - \frac{1}{n^2}$. By a union bound we can get that all distances maintained with at least $1 - \binom{n}{2}\frac{1}{n^2} \geq \frac{1}{2}$.

A few comments about this construction:

1. The above proof shows not only the existence of a good map, we also get that a random map as above works with constant probability! In other words, a Monte-Carlo randomized algorithm for dimension reduction.

2. The algorithm does not even look at the set of points $S$: it works for any set with high probability. Hence, we can pick this map $A$ before the points in $S$ arrive.

3. if we pick $k$ to be a bit larger (up to constant), we can decrease our constant probability $\frac{1}{2}$ to be exponentially small (so zero in practice).

## 2.2 Starting point and normal distributions

Well be using basic facts about Gaussians, lets just recall them.

1. The probability density function for the Gaussian $N(\mu, \sigma^2)$ is

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

2. If $G_i \sim N(\mu_i, \sigma_i^2)$, $c$ is a constant, then:

$$cG_1 \sim N(c\mu_1, c^2\sigma_1^2)$$
$$G_1 + G_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Recall that we want to argue about the squared length of $A(x) \in \mathbb{R}^k$. Observe that each coordinate of the vector $Mx$ behaves like

$$Y \sim (G_1, \ldots, G_d) \cdot x$$

where each $G_i \sim N(0, 1)$. Then by the above properties we get $Y \sim N(0, x_1^2 + \ldots + s_d^2) = N(0, 1)$. So each of $k$ coordinates of projection is just a normal random variable.

Then the squared length of $A(x) = \frac{1}{\sqrt{k}}M$ is $Z = \sum_i \frac{1}{k}G_i^2$. Since $\mathbb{E}[G_i^2] = Var(G_i) + \mathbb{E}[G_i]^2 = 1$, we get $\mathbb{E}[Z] = 1$.

### 2.3 Concentration about the Mean

Now to show that $Z$ does not deviate too much from 1. And $Z$ is the sum of a bunch of independent and identical random variables. If only the $G_i$s were all bounded, we could have used a Chernoff bound and be done. But these are not bounded, so this is where well need to do a little work. Basically, because normal distribution has so small tail, everything will work out (In fact, while the Laplace distribution with distribution $f(x) \sim e^{-\lambda|x|}$ for $x \in \mathbb{R}$ also has pretty thin tailsexponential tails, this wont work the same, even if you squint as hard as you like. It turns out you need sub-Gaussian tails).

Okay, let's start:

$$\mathbf{Pr}[\, Z \geq 1+\varepsilon \,] = \mathbf{Pr}[\, e^{tkZ} \leq e^{tk(1+\varepsilon)} \,] \leq |\text{Markov}| \leq \frac{\mathbb{E}[e^{tkZ}]}{e^{tk(1+\varepsilon)}} = \prod_i \frac{\mathbb{E}[e^{tG_i^2}]}{e^{t(1+\varepsilon)}}$$

for any $t > 0$. Also:

$$\mathbb{E}[e^{tG^2}] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tg^2} e^{-\frac{g^2}{2}} dg = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{z^2}{2}} \frac{dz}{\sqrt{1-2t}} = \frac{1}{\sqrt{1-2t}}$$

for $t < \frac{1}{2}$. So our current bound on the upper tail is that for all $0 < t < \frac{1}{2}$ we have

$$\mathbf{Pr}[\, Z \geq 1+\varepsilon \,] \leq \left( \frac{1}{e^{t(1+\varepsilon)}\sqrt{1-2t}} \right)^k$$

Then:

$$\begin{aligned} \left( \frac{1}{e^{t(1+\varepsilon)}\sqrt{1-2t}} \right) &= \exp\left( -t - \frac{1}{2}\log(1-2t) \right) \\ &= \exp\left( \frac{(2t)^2}{4} + \frac{(2t)^3}{6} + \ldots \right) \leq \exp(t^2(1+2t+2t^2 \ldots)) \\ &= \exp\left( \frac{t^2}{1-2t} \right) \end{aligned}$$

Plugging this back, we get

$$\mathbf{Pr}[\, Z \geq 1+\varepsilon \,] \leq e^{\frac{-k\varepsilon^2}{8}}$$

is we set $t = \frac{\varepsilon}{4}$ and use $1 - 2t \geq \frac{1}{2}$ for $\varepsilon \leq \frac{1}{2}$.

Almost done: lets take stock of the situation. We observed that $\|A(x)\|^2$ was distributed like an average of squares of Gaussians, and by a Chernoff-like calculation we proved that

$$\mathbf{Pr}[\, \|A(x)\|^2 > 1+\varepsilon \,] \leq e^{\frac{-k\varepsilon^2}{8}} \leq \frac{\delta}{2}$$

for $k = \frac{8}{\varepsilon^2}\log\frac{2}{\delta}$. A similar calculation bounds the lower tail, and nishes the proof of Main Lemma.

## 3 Using Random Signs instead of Gaussians

In practice it is painfull to draw from a normal distributions. How about populating $A$ with draws from other, simpler distributions? How about setting each entry of $M$ be $\pm 1$, and letting $A$ be a scaled version of $M$. And it will also work!

If we look closely on the prove of lemma and see what properties of normal distribution we used, we will se, that the lemma holds for any sub-gaussian distribution.

**Definition 3.1.** A random variable $V$ is said to be subgaussian with parameter $c$ if for all real $s$, we have $\mathbb{E}[e^{sV}] \leq e^{cs^2}$

The following facts are true:

1. For $G \sim N(0,1)$, $G$ is $\frac{1}{2}$-subgaussian.
2. For $\pm 1$ random variable is $\frac{1}{2}$-subgaussian.