CSE 521: Design and Analysis of Algorithms I **Spring 2016**

Lecture 6: Curse of Dimensionality, Dimension Reduction

Lecturer: Shayan Oveis Gharan *April 13th* *Scribe: Hessam Bagherinezhad*

High dimensional vectors appear frequently in recent development in CS, examples of which are user-movie ratings of netflix, DNA strings of patients, and images pixel values. In this lecture we study higher dimensions geometry and see how randomization help us to design algorithm. We first start with a few definitions, then we discuss several properties of high dimensional geometry, and finally we investigate a way to map points from high dimensions to a low dimensional space preserving the ℓ_2 distances.

6.1 Introduction

Definition 6.1. *For a vector $v \in \mathbb{R}^d$, ℓ_2 norm of v is defined $\|v\|_2 = \sqrt{\sum_i v_i}$, and ℓ_∞ of v is defined $\|v\|_\infty = \max_i v_i$.*

Definition 6.2. *A d dimensional ball is defined as $B_d = \{(x_1, \cdots, x_d) : \|x\|_2 \leq 1\}$, and similarly a d dimensional cube is defined as $H_d = \{(x_1, \cdots, x_d) : \|x\|_\infty \leq 1\}$.*

In $\mathbb{R}^2$, a ball is the circle with unit radius around the origin, and a cube is the square whose side is 2 around the origin. In $\mathbb{R}^3$, they become sphere and cube. It is hard to have intuition for ball and cube in $\mathbb{R}^4$, but one way to think about them is that each side of a cube in $\mathbb{R}^4$ is a cube in $\mathbb{R}^3$, and if we cut a ball with a hyperplane in $\mathbb{R}^4$, the set of points from the ball on the hyperplane would form a sphere.

In geometry, diameter of a circle is any line segment that passes through the center of the circle. And for a square it's the line segment that connects two opposite vertices. In more general form, the diameter of a shape (a set of points) is the supremum of distances between every two points in that set, $\sup_{x,y} d(x, y)$. The diameter of a d dimensional ball is always 2, and the diameter of a d dimensional cube is $2\sqrt{d}$, which is the distance between $(1, 1, \cdots, 1)$ and $(-1, -1, \cdots, -1)$.

Volume is a notion for the size of a shape D, and is formally defined as $\int_D 1 dD$ where the integration is with respect to the Lebesgue measure. Volume of a d dimensional ball is $\frac{\pi^{\frac{d}{2}}}{(\frac{d}{2})!}$, and volume of a d dimensional cube is 2^d. It is interesting to observe even though a 2-dimensional ball covers most of the area of a 2-dimensional cube (see Figure 6.1), the ratio of the volumes of the in d dimensions is exponentially small in d, $\frac{\text{Var}(B_d)}{\text{Var}(H_d)} = \frac{1}{d^{\Omega(d)}}$. In other words, as d grows, the ball gets exponentially smaller than the cube.

6.2 Near Orthogonal Vectors

In $\mathbb{R}^d$, the maximum number of vectors that you can find such that *each* pair of them are orthogonal to each other is exactly d. One example of such vectors set is the standard basis vectors, $e^1, \ldots, e^d$ where for each i, e^i is the vector

$$e^i_j = \begin{cases} 1 & \text{if j=i} \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

In this section we want to study the maximum number of vectors one can choose in $\mathbb{R}^d$ such that the angle between each pair of them is close to 90° (say a number between 88° and 92°)? For $d = 2$, the answer is still 2, but we are going to show that in $\mathbb{R}^d$ we can choose exponential in d many vectors such that each pair of them are almost orthogonal.

Theorem 6.3. *$\exists v^1, \cdots, v^m \in \mathbb{R}^d$ where $m \geq e^{\Omega(d)}$, such that $\forall i, j : \angle v^i, v^j \approx 90°$.*

Proof. Choose each vector v^i randomly where each coordinate of v^i is picked independently and uniformly from $\{\frac{+1}{\sqrt{d}}, \frac{-1}{\sqrt{d}}\}$. The factor $1/\sqrt{d}$ is chosen to make sure that each v^i has norm exactly 1. This implies that $\langle v^i, v^j \rangle = \cos(\angle v^i, v^j)$. So, to show that $\angle v^i, v^j \approx 90°$, it is enough to show that $\langle v^i, v^j \rangle$ is very close to 0. We show that with high probability the angle between any two vectors sampled this way is close to 90°, and then we use a union bound to finish the proof.

Fix a unit vector u, and let v be a random vector chosen as described in the previous paragraph. We can easily calculate the expected value of their inner product. We have

$$\mathbb{E}\left[\langle v, u \rangle\right] = \mathbb{E}\left[\sum_i v_i \cdot u_i\right] = \sum_{i=1}^{d} u_i \cdot \mathbb{E}\left[v_i\right] = 0.$$

Now, for each i, let $X_i = v_i \cdot u_i$. Given that $\mathbb{E}\left[\sum_i X_i\right] = 0$, we use Hoeffding inequality to find a bound for the inner product. Observe that for each i, $\frac{-u_i}{\sqrt{d}} \leq X_i \leq \frac{+u_i}{\sqrt{d}}$. So,

$$\mathbb{P}\left[\left|\sum v_i \cdot u_i\right| > \epsilon\right] = \mathbb{P}\left[\left|\sum X_i\right| > \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{d}(2\frac{v_i}{\sqrt{d}})^2}\right) = \exp\left(\frac{-2\epsilon^2}{4/d}\right) = e^{\frac{-\epsilon^2 d}{2}}$$

We showed a probability bound for the inner product of two vectors, now we use union bound to prove a bound for the inner product of all pairs of vector.

$$\begin{aligned}
\mathbb{P}\left[\forall i, j : |\langle v^i, v^j \rangle| \leq \epsilon\right] &= 1 - \mathbb{P}\left[\forall i, j |\langle v^i, v^j \rangle| > \epsilon\right] \\
&\geq 1 - \sum_{i<j} \mathbb{P}\left[|\langle v^i, v^j \rangle| > \epsilon\right] \\
&\geq 1 - \binom{m}{2} e^{\frac{-\epsilon^2 d}{2}} \geq 1 - m^2 e^{\frac{-\epsilon^2 d}{2}}
\end{aligned}$$

If we set $\epsilon = \sqrt{\frac{5 \lg m}{d}}$, we have:

$$\mathbb{P}\left[\forall i, j : |\langle v^i, v^j \rangle| \leq \sqrt{\frac{5 \lg m}{d}}\right] \geq 1 - \frac{1}{\sqrt{m}}$$

Now if we set $m = e^{\frac{d}{10000}}$, then $\epsilon = \sqrt{\frac{5 \frac{d}{10000}}{d}} \leq \frac{1}{40}$. This means that for every pair v^i, v^j of $e^{\frac{d}{10000}}$ vectors we have $-\frac{1}{40} \leq \cos(\angle v^i, v^j) \leq \frac{1}{40}$, so $85° \leq \angle v^i, v^j \leq 95°$. □

6.3 Sample Vectors with Uniform Direction

Let's say we want to sample a d dimensional vector v with a uniformly random direction. In other words, we would like to choose a uniformly random point on a d-dimensional ball. One simple way is to randomly

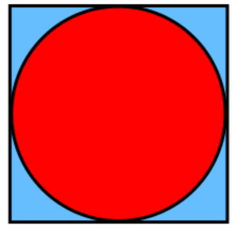

Figure 6.1: Illustration of why sampling coordinates uniformly random doesn't give a rotationally uniform vector.

sample the coordinates of a vector $v = \langle v_1, \cdots, v_d \rangle$ from range $[-1, 1]$. However, with this approach of sampling, the direction of the vectors would not be uniformly distributed. For example in $\mathbb{R}^2$, if we sample a vector by sampling each coordinate uniformly and independently, we're actually choosing a uniform random point from H_2 (square of with side of 2 around the origin). The point that gets picked is either inside B_2 (the red area of Figure 6.1) or outside of it. If the point is inside the ball, then the angle is uniformly distributed. If it's outside the ball (the blue area of Figure 6.1), it's angle is more biased towards 45°, 135°, 225°, and 315°.

We say a vector $g \in \mathbb{R}^d$ is a Guassian vector, if each coordinate of g is a uniformly and independently chosen $\mathcal{N}(0, 1)$ random variable. To choose a random direction in $\mathbb{R}^d$ it is enough to choose a Guassian vector. That vector points to a random direction. Similarly, to choose a uniformly random point on the surface of the d dimensional ball we can use $g/\|g\|$. Next, we show that a Gaussian vector g is rotationally invariant which implies the aforementioned properties.

First, let us state a nice fact about Gaussians.

Fact 6.4. *Given two Gaussian random variables $g_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $g_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, any linear combination of them, $ag_1 + bg_2$, has the same distribution as a Gaussian random variable with mean $a\mu_1 + b\mu_2$ and variance $a^2\sigma_1^2 + b^2\sigma_2^2$.*

Knowing the fact above, we can make the following claim.

Claim 6.5. *Given a d dimensional unit vector $\|v\| = 1$, and a gaussian vector $g = \langle g_1, \cdots, g_d \rangle$ whose coordinates are drawn iid from $\mathcal{N}(0, 1)$, the inner product of v and g is a standard normal random variable, i.e. $\langle v, g \rangle \sim \mathcal{N}(0, 1)$.*

Note that the claim is that the distribution of $\langle v, g \rangle$ doesn't depend on the direction of v. This means that the Gaussian vector g is equally directed towards any direction, meaning that it's rotationally uniform.

To prove the claim, we use Fact 6.4. In particular, we can write

$$\langle v, g \rangle = \sum_i v_i \cdot g_i \stackrel{\text{D}}{=} \mathcal{N}(0, v_1^2 + \cdots + v_d^2) = \mathcal{N}(0, 1).$$

where we used that $\|v\| = 1$.

6.4 Concentration of Measure for Gaussians

Let $X \sim \mathcal{N}(0, 1)$. Observe that $\mathbb{E}\left[X^2\right] = 1$. Concentration of measure for gaussians states that as we get more samples $X_1, \ldots, X_k \sim \mathcal{N}(0, 1)$, the mean of the square of the samples $\frac{1}{k}\sum_i X_i^2$ gets exponentially closer

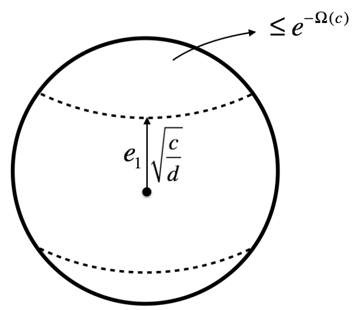

Figure 6.2: Most of the area of a d dimensional ball is on the belt around its equator.

to 1.

Theorem 6.6. *For k standard normal random variables $X_1, \cdots, X_k \sim \mathcal{N}(0,1)$, and $0 < \epsilon < 1$, we have*

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_i X_i^2 - 1\right| \ge \epsilon\right] \le 2e^{\frac{-k\epsilon^2}{8}}$$

We can use this bound to show that most of the area of a d dimensional ball is on the belt around its equator.

Theorem 6.7. *If we slice the ball B_d with the hyperplane $xe^1 = \sqrt{\frac{c}{d}}$, where e^1 is as defined in (6.1), the volume of the ball above (or below) the hyperplane is less than $e^{-\Omega(c)}$.*

Proof. To prove the claim it is enough to show that a uniformly random point $g/\|g\|$ on the surface of unit d-dimensional ball has a small inner product with e^1 vector. To be precise we show that

$$\mathbb{P}\left[|\langle \frac{g}{\|g\|}, e^1\rangle| \ge \sqrt{\frac{c}{d}}\right] \le e^{-\Omega(c)}.$$

First, observe that

$$\mathbb{P}\left[|\langle \frac{g}{\|g\|}, e^1\rangle| \ge \sqrt{\frac{c}{d}}\right] = \mathbb{P}\left[\langle \frac{g}{\|g\|}, e^1\rangle^2 \ge \frac{c}{d}\right] = \mathbb{P}\left[\langle g, e^1\rangle^2 \ge \|g\|^2 \frac{c}{d}\right] = \mathbb{P}\left[g_1^2 \ge \|g\|^2 \frac{c}{d}\right]$$

We use the concentration of measure of independent Gaussians, to show that $\|g\|^2$ is highly around d.

$$\mathbb{P}\left[|\,\|g\|^2 - d| > \epsilon d\right] = \mathbb{P}\left[\left|\sum_i g_i^2 - d\right| > \epsilon d\right] = \mathbb{P}\left[\left|\frac{1}{d}\sum g_i^2 - 1\right| > \epsilon\right] \le e^{\frac{-d\epsilon^2}{8}},$$

where in the last inequality we used Theorem 6.6. So, for $\epsilon = \frac{1}{2}$ we have:

$$\mathbb{P}\left[\|g\|^2 < \frac{d}{2}\right] \le e^{-\Omega(d)}$$

On the other hand, g_1 is just a $\mathcal{N}(0,1)$ random variable, so

$$\mathbb{P}\left[g_1^2 \ge 2c\right] = \mathbb{P}\left[|g_i| \ge \sqrt{2c}\right] \le e^{-\Omega(c)},$$

where the latter simply follows from the Gaussian probability density function.

Now, we can finish the proof by a union bound argument. If $g_1^2 \geq \frac{c}{d} \|g\|^2$, then we must either have $\|g\|^2 > d/2$ or $g_1^2 > 2c$. So, by union bound,

$$\mathbb{P}\left[g_1^2 \geq \|g\|^2 \frac{c}{d}\right] \leq \mathbb{P}\left[g_1^2 \geq 2c\right] + \mathbb{P}\left[\|g\|^2 \geq d/2\right] \leq e^{-\Omega(c)} + e^{-\Omega(d)} \leq 2e^{\Omega(c)}.$$

□

6.5 Dimension Reduction

Now we describe a central result of high dimensional geometry. Given n points $x^1, x^2, \cdots, x^n \in \mathbb{R}^d$. We would like to find n points $y^1, y^2, \cdots, y^n \in \mathbb{R}^k$, where $k \ll d$ such that for all i, j

$$(1-\epsilon)\left\|x^i - x^j\right\| \leq \left\|y^i - y^j\right\| \leq (1+\epsilon)\left\|x^i - x^j\right\|. \tag{6.2}$$

Ideally, we would like to have ϵ very close to zero and $k \ll d$.

Theorem 6.8 (Johnson-Lindenstrauss)**.** . *Given n points $x^1, x^2, \cdots, x^n \in \mathbb{R}^d$, there exists a linear mapping $Gx_i = y_i \in \mathbb{R}^k$ such that each i, j satisfy (6.2) and $k = O(\frac{\lg n}{\epsilon^2})$.*

The remarkable fact about this theorem is that k does not depend on d; it only depends on the number of points and ϵ.

One simple idea to prove the above theorem is to let the rows of G be k standard basis vectors chosen uniformly at random. However, this doesn't work very well in the worst case. A bad example is where all x_i's are zero on all coordinates except the last k. Such a mapping G with high probability distorts most pair of points.

The right mapping is to let G be a $k \times n$ matrix whose element are drawn iid from $\mathcal{N}(0, \frac{1}{k})$, i.e., each row of G is a d-dimensional Gaussian vector scaled by $1/\sqrt{k}$,

$$G = \begin{bmatrix} \sqrt{1/k} & g^1 \\ \sqrt{1/k} & g^2 \\ \cdots & \\ \sqrt{1/k} & g^k. \end{bmatrix}$$

Firstly, observe that for any vector $v \in \mathbb{R}^d$, $\mathbb{E}\left[\|Gv\|^2\right] = 1$. This is because

$$\mathbb{E}\left[\|Gv\|^2\right] = \mathbb{E}\sum_{i=1}^{k} \frac{1}{k}\langle g^i, v\rangle^2 = \sum_{i=1}^{k} \mathbb{E}\left[\langle g^i, v\rangle^2\right] = 1,$$

where the last equality uses the rotation invariance property of Gaussians. In particular, let $X \sim \mathcal{N}(0,1)$. Then, since $\langle g^i, v\rangle \overset{\mathrm{D}}{=} \mathcal{N}(0,1)$, we can write

$$\mathbb{E}\left[\langle g^i, v\rangle^2\right] = \mathbb{E}X^2 = \mathrm{Var}(X) - \mathbb{E}\left[X\right]^2 = \mathrm{Var}(X) = 1.$$

Claim 6.9. *For any unit vector v,*

$$\mathbb{P}\left[\left|\|Gv\|^2 - 1\right| > \epsilon\right] \leq e^{-\epsilon^2 k/8}$$

Proof. By the rotation invariance property, Gv has the same distribution as Ge^1. We can write

$$Ge^1 = \begin{pmatrix} g_1^1/\sqrt{k} \\ g_1^2/\sqrt{k} \\ \vdots \\ g_1^k/\sqrt{k} \end{pmatrix}.$$

So,

$$\left\| Ge^1 \right\|^2 = \sum_{i=1}^{k} \frac{1}{k}(g_1^i)^2.$$

The claim follows by Theorem 6.6. □

Now, we are ready to finish the proof of Theorem 6.8. Fix a pair $1 \le i < j \le n$. In the following claim we show that $\left\|y^i - y^j\right\| \approx \left\|x^i - x^j\right\|$ with high probability. Then, Theorem 6.8 follows by a union bound argument.

Claim 6.10. *For any* $1 \le i < j \le n$*,*

$$\mathbb{P}\left[(1-\epsilon)\left\|x^i - x^j\right\| \le \left\|y^i - y^j\right\| \le (1+\epsilon)\left\|x^i - x^j\right\|\right] \ge 1 - e^{-\epsilon^2 k/8}.$$

Proof. The claim essentially follows from Claim 6.9 after doing some algebraic manipulations. The main nontrivial step is to use the linearity of the projection operator, i.e., $y^i - y^j = Gx^i - Gx^j = G(x^i - x^j)$.

$$\begin{aligned}
\mathbb{P}\left[(1-\epsilon)\left\|x^i - x^j\right\| \le \left\|y^i - y^j\right\| \le (1+\epsilon)\left\|x^i - x^j\right\|\right] &= \mathbb{P}\left[(1-\epsilon)^2\left\|x^i - x^j\right\|^2 \le \left\|y^i - y^j\right\|^2 \le (1+\epsilon)^2\left\|x^i - x^j\right\|^2\right] \\
&\ge \mathbb{P}\left[1-\epsilon \le \frac{\left\|y^i - y^j\right\|^2}{\left\|x^i - x^j\right\|^2} \le 1+\epsilon\right] \\
&= \mathbb{P}\left[\left|\frac{\left\|Gx^i - Gx^j\right\|^2}{\left\|x^i - x^j\right\|^2}\right| \le \epsilon\right] \\
&= \mathbb{P}\left[\left|\left\|G\frac{x^i - x^j}{\left\|x^i - x^j\right\|}\right\|^2\right| \le \epsilon\right] \le e^{-k\epsilon^2/8},
\end{aligned}$$

where the last inequality follows by Claim 6.9. □

To prove Theorem 6.6 it is enough to let $k = 24/\epsilon^2$. Then, we get

$$\mathbb{P}\left[\forall i,j : (1-\epsilon)\left\|x^i - x^j\right\| \le \left\|y^i - y^j\right\| \le (1+\epsilon)\left\|x^i - x^j\right\|\right] \ge 1 - n^2 e^{-\epsilon^2 k/8} \ge 1 - 1/n.$$

This completes the proof of Theorem 6.8.

See the course website for several recent works related to dimension reduction and Johnson-Lindenstrauss theorem.