CS 329M: Topics in Artificial Intelligence **Spring 2017**

Lecture 1: April 3

Lecturer: James Zou *Scribe: Tri Dao, Yosuke Tanigawa, Ziyi Yang, Salil Bhate, Horia Margarit*

1.1 Random geometry in high dimension

In this section, we will develop a geometric intuition on high dimensional vectors.

1.1.1 Why do we care about high dimensional vectors?

A lot of data we care about are in very high dimension. In the context of computational genomics, for example, one can think of a vector of 3 billion dimensions, each element of which represents a base pair. In recommendation systems, one can think about a high dimension vector to represent a personal preference, where each element of the vector corresponds to one movie in the library. For natural language processing (NLP), one can represent a semantic meaning of word with word embedding. High dimensional vectors are common in machine learning and proper understanding of their geometry is important.

1.1.2 Three puzzles to test your intuition in high dimensional vector space

High dimensional vectors do not follow our intuition, which is based on the 3D space we live in. Here are a few puzzles to test your intuition.

We first need few definitions.

Definition 1 (d-ball) *A d-dimensional ball (d-*ball*) is the set of points that has 1 or smaller distance from the origin, $B_d^1 := \{x \in \mathbb{R}^d \colon \|x\|^2 \leq 1\}$.*

Definition 2 (d-cube) *A d-*cube *is a rectangular region in d-dimensional space, $\left[-\frac{1}{2}, \frac{1}{2}\right]^d$.*

1. The volume of the cube is 1 for all d. Which of the following is closest to the volume of a d-ball for $d = 100$?

   A. > 1

   B. $(1/2, 1)$

   C. $(1/100, 1/2)$

   D. $(0, 1/100)$.

   The answer is D. The volume is pretty much 0.

2. As $d \to \infty$, pick two points uniformly randomly from the d-cube. What is the average Euclidean distance between the two points?

   A. 0

B. $\sqrt{2}$

C. π

D. ∞

The answer is D.

3. Same as question 2, but for a d-ball?

A. 0

B. $\sqrt{2}$

C. π

D. ∞

The answer is B.

1.1.3 A few results from probability theory

We state here a few results (without proof) from probability theory that we will need.

Theorem 3 (Law of large numbers) *Let $X_1, \ldots, X_n$ be iid samples of a random variable X with $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}(X)$, then the following inequality holds:*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Intuitively, this theorem gives a probabilistic bound on the difference between the empirical average and the actual average.

Theorem 4 (Hoeffding's bound) *Suppose X_i are iid samples of X, with $\mu = \mathbb{E}[X]$, $X \in [a, a+A]$[1], then the following inequality holds:*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 n/A^2}.$$

Note that Hoeffding's bound is much stronger than the one given by the law of large numbers, since it is exponential in n.

Proposition 5 (A variant of Hoeffding's bound for Gaussian variables) *If $X_i \sim N(0,1)$, for $i = 1, \ldots, n$, iid, then the following inequality holds:*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 1\right| > \epsilon\right] \leq 2e^{-\epsilon^2 n/9}.$$

[1] X is bounded in a range in this case. One can see that the range parameter of random variables convey information on their variance.

It is interesting to see that we can still have exponential bound even though Gaussian distribution is not bounded[2]

Note that if we draw $X \sim N(0, I_d)/\sqrt{d}$, then $\|X\|_2^2 = \frac{1}{d}\sum_{i=1}^{d} X_i^2$ holds.[3][4]. Proposition 5 says X has distance almost exactly 1 from the origin (with high probability), i.e., on the surface of d-ball.

These bounds are called *concentration inequalities*.

1.1.4 Explanation of the puzzles

1.1.4.1 The volume of the d-ball

The volume of a ball in d dimensions with radius r is $\text{Vol}(B_d^r) = \frac{\pi^{d/2} r^d}{\Gamma(d/2+1)}$, where $\Gamma(n) = (n-1)!$. As d increases, the denominator increases faster than the numerator, so the volume goes to 0[5].

Moreover, we can see the volume of the ball is concentrated on the surface

$$\frac{\text{Vol}(B_d^{1+\epsilon})}{\text{Vol}(B_d^1)} = (1+\epsilon)^d \to \infty.$$

1.1.4.2 Distance of two points randomly sampled from a d-cube

Pick $X, Y \sim \mathsf{Uniform}[-0.5, 0.5]^d$. This is equivalent to picking each X_i, Y_i from $\mathsf{Uniform}[-0.5, 0.5]$. Let $U_i = X_i - Y_i$ then

$$\|X - Y\|_2^2 = \sum_{i=1}^{d} (X_i - Y_i)^2 = \sum_{i=1}^{d} U_i^2 \to \infty \quad (\text{as } d \to \infty).$$

1.1.4.3 Distance of two points randomly sampled from a d-ball

Picking two points from the ball is pretty much the same as picking two points from the surface, which again is almost the same as picking $X, Y \sim N(0, I_d)/\sqrt{d}$. Therefore

$$\|X - Y\|_2^2 = \frac{1}{d}\sum_{i=1}^{d} (X_i' - Y_i')^2, \quad \text{where } X_i', Y_i' \sim N(0,1).$$

Let $U_i = X_i' - Y_i'$ then $U_i \sim N(0, 2)$, so we obtain $\|X - Y\|_2^2 \approx 2$ by Proposition 5.

---

[2]Intuitively, this is because the tails of the Gaussian distribution decay exponentially so we can control the probability outside of a bounded interval.

[3]Note that drawing d independent samples from a one dimensional Gaussian is equivalent to drawing one sample from a d-dimensional Gaussian with diagonal covariance.

[4]Note that we used the normalized random vector to avoid the sum blowing up

[5]The numerator is exponential in $\frac{d}{2}$ and the denominator is factorial of $\frac{d}{2}$, which grows like d^d.

1.2 Johnson–Lindenstrauss random projections

1.2.1 Motivation

In machine learning applications, we often have to deal with very high dimensional vectors. Any operation on this vector (even storage itself) is expensive. Can we leverage the geometry of high dimensional vectors discussed above to store and compute them efficiently? Here we introduce an idea called *random projections*.

If we have access to the pairwise distances between points, we can solve many machine learning problems, such as clustering, classification and so on. However, computation of pairwise distances in high dimension can be very expensive. Can we project these points to a lower dimensional space, with little distortion in the pairwise distances?

1.2.2 Johnson–Lindenstrauss Theorem

Let $A \in \mathbb{R}^{d\times k}$ be a *random projection matrix* whose entries A_{ij} are iid sampled from the scaled Normal distribution, i.e.

$$A_{ij} \sim N(0,1)/\sqrt{d}.$$

We generally assume $d \ll k$.

Here is the setup: We have a set of n points in $\mathbb{R}^k$: $\mathcal{V} = \{v \in \mathbb{R}^k\}$. Generate a random projection matrix A of size $d \times k$ where $d \ll k$. We project each vector v by computing $Av \in \mathbb{R}^d$. Our hope is that $\forall v, w \in \mathcal{V}$, $\|Av - Aw\|^2 \approx \|v - w\|^2$. If this holds, we can throw away the original data points and work with the low dimension projections.

The question is: what is the smallest d that gives us statistical guarantee?

Theorem 6 (Johnson–Lindenstrauss Theorem) *Suppose that $d = \mathcal{O}\left(\frac{\log n}{\epsilon^2}\right)$, then with high probability, for all $v, w \in \mathcal{V}$,*

$$\|v - w\|(1 - \epsilon) \le \|Av - Aw\| \le \|v - w\|(1 + \epsilon). \tag{1.1}$$

Proof: Fix a pair $v, w \in \mathcal{V}$. The entries of A are sampled from $N(0,1)/\sqrt{d}$, which is rotationally symmetric. Hence without loss of generality (by change of coordinates), we can assume that $v - w = (1, 0, \ldots, 0)$. Thus we only care about the first column of A, which we will denote as $(A_{1,1}, \ldots, A_{d,1})$. Then, we have

$$\|Av - Aw\| = \|v - w\| \sqrt{\sum_{i=1}^{d} A_{i,1}^2} = \|v - w\| \sqrt{\frac{1}{d}\sum_{i=1}^{d} X_i^2}.$$

where $X_i \sim N(0,1)$. We only need to show that $\sqrt{\frac{1}{d}\sum_{i=1}^{d} X_i^2} \approx 1$. Applying Proposition 5, we obtain

$$\Pr\left[\left|\frac{1}{d}\sum_{i=1}^{d} X_i^2 - 1\right| > \epsilon\right] \le 2e^{-\epsilon^2 d/9}.$$

Choose $d = \mathcal{O}\left(\frac{\log n}{\epsilon^2}\right)$ then the right hand side is $\mathcal{O}\left(\frac{1}{n^2}\right)$. Applying union bound over $\frac{n(n-1)}{2} = \mathcal{O}(n^2)$ pairs, we conclude that (1.1) holds for all $v, w \in \mathcal{V}$ with high probability as desired. ■

The key idea in the proof was to make a sum of independent random variables so that we can apply the bounds in theorems we presented before.

1.2.3 Example of an application of Johnson–Lindenstrauss Theorem

A cute puzzle/application to think about: Let's say we have two big matrices S_1, S_2 of size $k \times n_1$ and $k \times n_2$, where k is very large. We want to compute $S_1^T S_2$ (which is quite normal in machine learning computations). We can do random projections with random projection matrices A_1 and A_2, and then compute $(A_1 S_1)^T (A_2 S_2)$, which will be approximately $S_1^T S_2$. Why is this the case?

1.2.4 Projecting using PCA

Instead of using random projections, we could project using PCA (which we'll see in the next lecture). In fact, this does *very* badly at preserving pairwise distances, since it is optimizing a completely different objective. One can construct examples of multivariate distributions where a projection onto even $d - 1$ principal components distorts distances with high probability.