Advanced Algorithms

Lecture 17: The Johnson-Lindenstrauss Transform

Lecturer: Nathan Klein

1 Volume of the Ball and the Cube

Before talking about dimension reduction, we'll discuss one more fact about high dimensional geometry: that the volume of the ball and the cube diverge as $d \to \infty$. In particular remember the d-dimensional ball B_d (setting $r = 1$) is the set of points x with $\|x\|_2 \leq 1$. The d-dimensional cube C_d is the set of points x with $\|x\|_\infty = \max_i\{x_i\} \leq 1$.

In particular, we will show that the ball becomes tiny compared to the cube as $d \to \infty$.

Lemma 1.1. $\frac{Vol(B_d)}{Vol(C_d)} = e^{-\Omega(d)}$.

Proof. We will compute the probability a randomly chosen point in C_d lies in B_d. To choose a point x in the cube at random, it suffices to pick $x_1, \ldots, x_d \in [-1, 1]$ uniformly at random. It will lie in the ball exactly when $\|x\|_2 \leq 1$. So, let's apply Hoeffding to the independently sampled random variables $x_i^2 \in [0, 1]$ and let $X = \sum_{i=1}^d x_i^2$, we have

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq t\right] \leq 2e^{-\frac{2t^2}{d}}$$

Now, $\mathbb{E}\left[X\right] = d \cdot \mathbb{E}\left[x_i^2\right] = d \cdot \int_{-1}^1 \frac{x^2}{2} dx = d \cdot \frac{x^3}{6}\big|_{-1}^1 = \frac{d}{3}$. So,

$$\mathbb{P}\left[X \leq 1\right] \leq \mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq d/3 - 1\right] \leq 2e^{-\frac{2(d/3-1)^2}{d}} = e^{-\Omega(d)}$$ □

This adds some nice intuition to our thoughts about the curse of dimensionality with the following corollary: fix any point x in the cube $[-1, 1]^d$. Then, the probability another randomly sampled point in the cube lies within distance 1 of x is vanishingly small: $e^{-\Omega(d)}$.

2 Dimension Reduction

In this class, we will show the following, often called the "Johnson-Lindenstrauss Lemma."

Theorem 2.1 (JL Lemma)**.** *Given n points $x_1, \ldots, x_n \in \mathbb{R}^d$ and an approximation factor $\epsilon > 0$, we can choose $k \in O(\frac{\log n}{\epsilon^2})$ and a matrix $P \in \mathbb{R}^{k \times d}$ so that if $y_i = Px_i$ for all $1 \leq i \leq n$, the points $y_1, \ldots, y_n \in \mathbb{R}^k$ approximately preserve the ℓ_2 distances of $x_1, \ldots, x_n$. In other words, it's the case that for all i, j we have:*

$$(1 - \epsilon)\|x_i - x_j\|_2 \leq \|y_i - y_j\|_2 \leq (1 + \epsilon)\|x_i - x_j\|_2$$

This is a very useful theorem. Suppose you have a collection of n vectors $S \subseteq \mathbb{R}^d$ for some very large d, and given any $x, y \in S$ you want to be able to compute their ℓ_2 distance $d(x, y) = \|x - y\|_2$. This result says that you can reduce the storage from nd to $O(n \log(n))$ as long as you are OK with a small loss in precision. Furthermore, we can compute the ℓ_2 distance between any pair in

only $O(k) \approx O(\log n)$ time instead of $O(d)$ time. This has applications in, for example, speeding up nearest neighbor search over ℓ_2.

A first attempt might be to subsample the dimensions, taking each one with probability $\frac{k}{d}$. However this clearly fails when comparing two vectors $(1, 0, \ldots, 0)$ and $(0, 1, 0, \ldots, 0)$, as you will almost always map them to the same points. So what does work?

2.1 Proof of JL

We will let P be a matrix in $\mathbb{R}^{k \times d}$ where each entry $P_{i,j}$ is an independent standard Gaussian, $\mathcal{N}(0, 1)$, multiplied by $\frac{1}{\sqrt{k}}$. We will prove the following:

Lemma 2.2. *For any $\epsilon > 0, \delta > 0$, if we pick $k = O(\log(1/\delta)/\epsilon^2)$, for any vector x with $\|x\|_2^2 = 1$ we have:*

$$1 - \epsilon \leq \|Px\|_2^2 \leq 1 + \epsilon$$

with probability $1 - \delta$.

At first glance, this lemma does not look so interesting. For example, to achieve this, couldn't we just map every vector to its length? However, using the fact that this mapping is *linear* (setting $y_i = Px_i$ for some matrix P), it becomes enough to prove the whole theorem. Before proving it, let's show that it implies the JL lemma.

Proof of JL. For a pair x_i, x_j, apply Lemma 2.2 to the vector $\frac{x_i - x_j}{\|x_i - x_j\|_2}$. This implies that with probability $1 - \delta$, we have:

$$1 - \epsilon \leq \|P \frac{x_i - x_j}{\|x_i - x_j\|_2}\|_2^2 \leq 1 + \epsilon$$

Pulling out the constant term $\frac{1}{\|x_i - x_j\|_2}$, it gets squared, and we can multiply by it on both sides, yielding:

$$(1 - \epsilon)\|x_i - x_j\|_2^2 \leq \|P(x_i - x_j)\|_2^2 = \|Px_i - Px_j\|_2^2 \leq (1 + \epsilon)\|x_i - x_j\|_2^2$$

but after taking square roots this is exactly what we want for two points x_i, x_j, noting that $|\sqrt{1 - \epsilon}| < 1 - \epsilon$.

$$(1 - \epsilon)\|x_i - x_j\|_2 \leq \|Px_i - Px_j\|_2 \leq (1 + \epsilon)\|x_i - x_j\|_2$$

since $\|y_i - y_j\|_2 = \|Px_i - Px_j\|_2$. So we know that for any particular pair x_i, x_j, the probability the distance is not distorted by more than $1 \pm \epsilon$ is at least $1 - \delta$.

Therefore, we need to choose δ small enough so that we can union bound over all pairs. It's enough to choose $\delta = \frac{1}{n^2}$ to prove the theorem since then after union bounding our probability of success is at least $1 - \binom{n}{2} \cdot \frac{1}{n^2} \geq \frac{1}{2}$, guaranteeing existence of a good matrix (one can be found with high probability by running the same procedure many times). But by Lemma 2.2, to achieve these it suffices to pick $k = O(\log(n^2)/\epsilon^2) = O(\log(n)/\epsilon^2)$, as desired. □

So, let's prove Lemma 2.2.

Proof. For our input x, let's compute the expected ℓ_2 norm of $y = Px$.

$$\mathbb{E}\left[\|Px\|_2^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[y_i^2\right] = \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\left[\left(\sum_{j=1}^{d} x_j g_{ij}\right)^2\right]$$

where the g_{ij} are independent standard Gaussians. But recall that any scaling of a Gaussian with mean μ and variance σ^2 by α is also Gaussian with mean $\alpha\mu$ and variance $\alpha^2\sigma^2$. So, we are summing Gaussians $1, \ldots, d$ with mean 0 and variance x_j^2. And now, recall our second fact about Gaussians: that the sum of two independent ones with mean μ_1, μ_2 and variance σ_1^2, σ_2^2 is Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Therefore this whole inner term $\sum_{j=1}^{d} x_j g_j$ is just a Gaussian with mean 0 and variance $\|x\|_2^2 = 1$. So, the expected value of its square is its variance, 1, and we have:

$$\mathbb{E}\left[\|Px\|_2^2\right] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\left(\sum_{j=1}^{d} x_j g_j\right)^2\right] = \frac{1}{k}\sum_{i=1}^{k} 1 = 1$$

Therefore, the expected value is 1. So now we just need to prove that we don't deviate from this expected value too much, which is something we already have experience with using Hoeffding's inequality. In this case, we have something very specific: the sum of the squares of k standard normals is called a chi-squared (χ^2) random variable with k degrees of freedom. We won't dive into this in this class, but it's good to at least know what a chi-squared distribution is because it is one of the most widely used distributions in statistics and, more specifically, hypothesis testing.

So, this distribution is quite well understood. A standard tail bound for the chi-squared distribution with k degrees of freedom is this:

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] \leq 2e^{-k\epsilon^2/8}$$

This is exactly the probability we do not have $\|Px\|_2^2 \in [1-\epsilon, 1+\epsilon]$. Setting this to $\delta = 2e^{-k\epsilon^2/8}$ and taking logs gives us $k = O(\ln(1/\delta)/\epsilon^2)$ as desired. □

It turns out this is optimal, and the dependence on ϵ and n cannot be improved.