MAT-280: Mathematical Foundations for Big Data Spring 2016

Lecture 1 — April 1, 2016

Scribes: Sam Fleischer and Lily Silverstein

1 Probability Basics

Expectation is denoted by $\mathbb{E}$. Probability is denoted by $\mathbb{P}$.

The following two inequalities are independent of dimension.

Theorem 1 (Markov's Inequality)**.** *Let $X \geq 0$ be a random variable (RV) with $\mathbb{E}(X) < \infty$. Then*

$$\mathbb{P}(X > t) < \frac{\mathbb{E}(X)}{t}.$$

Theorem 2 (Chebyshev's Inequality)**.** *Let X be a RV with $\mathbb{E}(X^2) < \infty$. Then*

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) \leq \frac{var(X)}{t^2}.$$

Proof. We apply Markov's Inequality to the RV $(X - \mathbb{E}(X))^2$ (which we know is ≥ 0). Then

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}(X)| > t) &= \mathbb{P}\left((X - \mathbb{E}(X))^2 > t^2\right) \\ &\leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{t^2} \\ &= \frac{\text{var}(X)}{t^2}\end{aligned}$$

□

The result of Markov's Inequality converges linearly to 0 as $t \to \infty$, while the result of Chebyshev's Inequality converges quadratically to 0. Chebyshev's Inequality converges faster, but there is a tradeoff - we need knowledge of the variance of the RV X. In big data applications, we often want to make statements about large subsets of the data. The above inequalities are most useful when the bounds on the right hand side are close to 0.

1.1 Concentration Inequalities

Concentration Inequalities can be thought of as "blessings" of high dimensionality. In big data applications we typically look at sums of many random variables $X_1, X_2, \ldots, X_n$ as opposed to individual random variables. We often ask the question "How large is the sum $X_1 + X_2 + \cdots + X_n$?" We limit our discussion to independent and identically distributed (IID) random variables.

Theorem 3 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be IID RVs with $|X_i| \leq a$ and $\mathbb{E}(X_i) = 0$ for all i. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > t\right) \leq 2\exp\left[\frac{-t^2}{2na^2}\right].$$

When n is large, Hoeffding's Inequality shows there is little variance around the expected value. In other words, the probability measure "concentrates" around $O(\sqrt{n})$. More precisely,

$$\left|\sum_{i=1}^{n} X_i\right| \sim a\sqrt{n \log n}.$$

For example, if $t = a\sqrt{2n \log n}$, then by Hoeffding's Inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \sqrt{2n \log n}\right) &\leq 2\exp\left[\frac{-(a\sqrt{2n \log n})^2}{2na^2}\right] \\ &= 2\exp[-\log n] \\ &= \frac{2}{n}\end{aligned}$$

If $t = \dfrac{na}{2}$, the by Hoeffding's Inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \frac{na}{2}\right) &\leq 2\exp\left[\frac{-\left(\frac{na}{2}\right)^2}{2na^2}\right] \\ &= 2\exp\left[-\frac{n}{8}\right]\end{aligned}$$

Hoeffding's Inequality shows that choosing a larger t gives greater confidence that the value will lie below t, whereas choosing a smaller t gives the likelihood of highly concentrated data. This is the tradeoff between high concentration and high probability.

Theorem 4 (Bernstein's Inequality)**.**

Theorem 5 (Chernoff's Inequality)**.**

2 Spheres and Cubes in High Dimensions

2.1 Geometry of the d-dimensional Sphere

Consider the unit sphere is d dimensions. Its volume is given by

$$V(d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\,\Gamma\left(\frac{d}{2}\right)}$$

where Γ is the Gamma function. Recall that for positive integers n, $\Gamma(n) = (n-1)!$. Using Stirling's Formula,

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}}\Big(\frac{n}{e}\Big)^n$$

we can see $\gamma\big(\frac{d}{2}\big)$ grows much faster than $\pi^{\frac{d}{2}}$, and hence

$$V(d) \to 0 \qquad \text{as} \qquad d \to \infty.$$

In words, the volume of the d-dimensional sphere with radius 1 goes to 0 as the dimension d increases to infinity, i.e. unit sphere in high dimensions have almost no volume (compare this to the unit cube, which always has volume 1).

Also notice that "most" of the volume of the d-dimensional sphere is contained near the boundary of the sphere. That is, for a d-dimensional sphere of radius r, most of the volume is contained in an annulus of width proportional to $\frac{r}{d}$.

2.2 Geometry of the d-dimensional Cube

Most of the volume of the high-dimensional cube is located in its corners.

Proof (probabilistic argument). Pick a point at random in the box $[-1,1]^d$. We want to calculate the probability that the point is also in the sphere.

Let $x = [x_1, \ldots, x_d] \in \mathbb{R}^d$ and each $x_i \in [-1,1]$ chosed uniformly at random. The event that x also lies in the sphere means

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \leq 1.$$

Let $z_i = x_i^2$ and note that

$$\mathbb{E}(z_i) = \frac{1}{2}\int_{-1}^{1} t^2 \mathrm{d}t = \frac{1}{3} \implies \mathbb{E}\left(\|x\|_2^2\right) = \frac{d}{3}$$

and

$$\mathrm{var}(z_i) = \frac{1}{2}\int_{-1}^{1} t^4 \mathrm{d}t - \left(\frac{1}{3}\right)^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45} \leq \frac{1}{10}$$

Using Chernoff's Inequality,

$$\begin{aligned}
\mathbb{P}\left(\|x\|_2^2 \leq 1\right) &= \mathbb{P}\left(\sum_{i=1}^{d} x_i^2 \leq 1\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{d} (z_i - \mathbb{E}(z_i)) \leq 1 - \frac{d}{3}\right) \\
&\leq \exp\left[\frac{-(\frac{d}{3})^2}{\frac{4d}{10}}\right] \\
&\leq \exp\left[-\frac{d}{4}\right].
\end{aligned}$$

Since this value converges to 0 as the dimension d goes to infinity, this shows random points in high cubes are most likely outside the sphere. In other words, almost all the volume of hypercubes lie in their corners. □

2.3 Comparisons between the d-dimensional Sphere and Cube

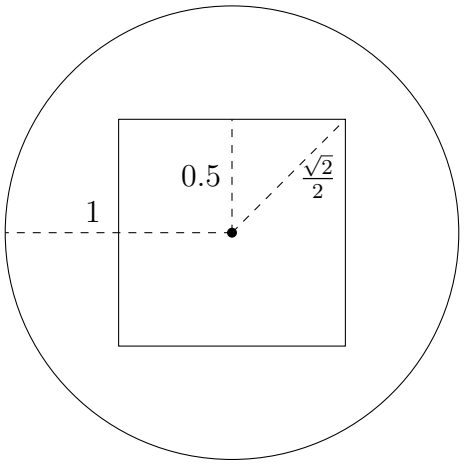

Figure 1: 2-dimensional unit sphere and unit cube, centered at the origin.

In two dimensions (Fig. 1), the unit square is completely contained in the unit sphere. The distance from the center to a vertex (radius of the circumscribed sphere) is $\frac{\sqrt{2}}{2}$ and the apothem (radius of the inscribed sphere) is $\frac{1}{2}$. In four dimensions (Fig. 2), the distance from the center to a vertex is 1, so the vertices of the cube touch the surface of the sphere. However, the apothem is still $\frac{1}{2}$. The result, when projected in two dimensions no longer appears convex, however all hypercubes are convex. This is part of the strangeness of higher dimensions - hypercubes are both convex and "pointy." In dimensions greater than 4 the distance from the center to a vertex is $\frac{\sqrt{d}}{2} > 1$, and thus the vertices of the hypercube extend far outside the sphere.

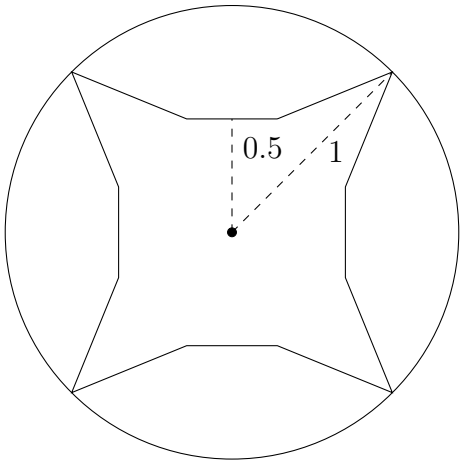

Figure 2: Projections of the 4-dimensional unit sphere and unit cube, centered at the origin.

3 Distances in High Dimensions

Theorem 6 (homework)**.** *Two random variables in high dimensions are almost orthogonal.*

3.1 Behavior of the ℓ^p norm in High Dimensions

Assume we are given n points $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ in $\mathbb{R}^d$. Define the ℓ^p norm of $x \in \mathbb{R}^d$ as

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

Define the *relative contrast* of the data set as

$$\frac{D_{\max} - D_{\min}}{D_{\min}}$$

where

$$D_{\max} = \max_i \left\{ \left\| x^{(i)} \right\|_p \right\}, \qquad \text{and}$$
$$D_{\min} = \min_i \left\{ \left\| x^{(i)} \right\|_p \right\}$$

Relative contrast is useful in machine learning by identifying when a nearest-neighbor calculation is significant.