Lecture 1

Today

- Syllabus
- Motivation
- Probability Review

Syllabus

Instructor: Mateo Diaz (mateodd@jhu.edu)

Office Hours: Th 3:00 - 4:30 pm
Wyman S429

TAs:

- Ian McPherson (impher1@jhu.edu)
  OH: F 9:00 - 10:30 am
- Pedro Izquierdo (pizquie1@jh.edu)
  OH: M 9:30 - 10:15 pm

Resources

- Canvas ← HW
- Website (mateodd25.github.io/mds) ← Course materials

- Piazza
- Gradescope ← All submissions

Ask your questions here

Grades

Four Components

- Homework (5 psets) 50%
- Midterm 20%
- Final Project 20%
- Participation 10%

Textbook

We will follow a couple of references (you can see a complete list on the website), but the

main one will be "High-Dimensional Probability" by Roman Vershynin.

Motivation

In this class we will gather tools to answer: statistically / computationally

How to efficiently extract information from data?

often samples

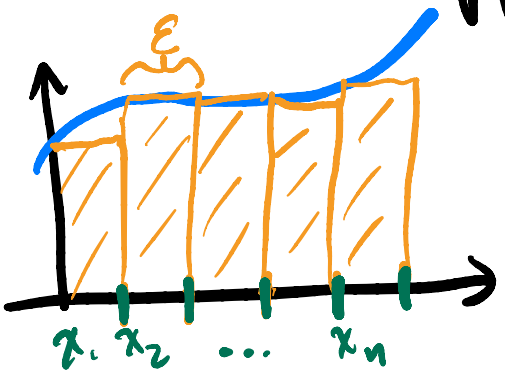
As written this question is ill-defined. Let's consider a couple of examples.

Example 1 (Monte Carlo method): Suppose I want to estimate

$$\int_{[0,1]^d} f(x)\,dx$$

In 1D, I can use a Remann

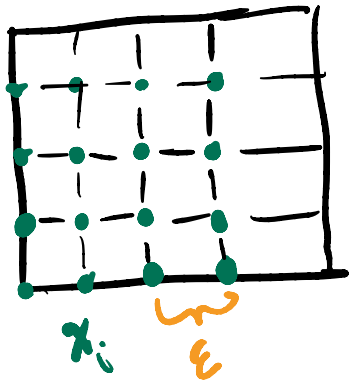
sum to approximate using a grid via



$$\frac{1}{n}\sum_{j=1}^{n} f(x_j) \approx \int_{[0,1]^d} f(x)\,dx.$$

If f is Lipschitz, then $n=\frac{1}{\varepsilon}$ suffices for an error on the order of ε. (why?)

In $\mathbb{R}^d$ I can play the same game with a grid. But in this case I need $n = (\frac{1}{\varepsilon})^d$ to have



Exponentially large

$$\left|\frac{1}{n}\sum f(x_i) - \int_{[0,1]^d} f(x)\,dx\right| \le O(\varepsilon).$$

This is what is known as the curse of dimensionality. Fortunately, we can have a more efficient method using random samples.

Quick probability interlude (Recorderis)

① Expectation

Given a random vector X with density $\rho(x)$ and $f: \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E} f(x) = \int_{\mathbb{R}^d} f(x)\,\rho(x)\,dx$$

② Variance

Given a random variable X,

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

③ Linearity

(a) $\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}X_1 + \dots + \mathbb{E}X_n$

(b) If X_i are independent, then

$$\mathrm{Var}(X_1 + \dots + X_n) = \mathrm{Var}(X_1) + \dots + \mathrm{Var}(X_n)$$

④ Jensen's Inequality

If X r.v. and $\varphi: \mathbb{R} \to \mathbb{R}$ convex

$$\varphi(\mathbb{E}X) \le \mathbb{E}\,\varphi(X).$$

④ (strong) Law of large Numbers

If $X_1, X_2, \dots$ are independent identically distributed r.v., then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}X \quad \text{almost surely (a.s.)}$$

with probability = 1.

Back to the example, take $x_1, \dots, x_n \sim \text{Unif}([0,1]^d)$. Then,

$$\int_{[0,1]^d} f(x)\,dx = \mathop{\mathbb{E}}_{X \sim \text{Unif}([0,1]^d)} f(X)$$

and by the Strong Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to \mathbb{E} f(X) \quad \text{a.s.}$$

How fast does this happen?

$$\mathbb{E}\left(\frac{1}{n}\sum f(x_i) - \mathbb{E} f(X)\right)^2 = \mathbb{E}\left(\frac{1}{n}\sum \left(f(x_i) - \mathbb{E} f(X)\right)\right)^2$$

$$\text{Var}\left(\sum Y_i\right) = \sum \text{Var}(Y_i) \text{ for } Y_i \text{ iid} = \frac{1}{n}\,\text{Var}(f(x_i))$$

Bounded for f over $[0,1]^d$

Thus, taking square roots

$$\mathbb{E}\ \text{error} \le \frac{c}{\sqrt{n}}$$

$\varphi(\mathbb{E}x) \le \mathbb{E}\varphi(x)$ for φ convex

This is independent of the dimension!

Example 2: Now suppose we are given

access to a coin and they ask us to determine whether it is fair or not. Due to the SLLN, we could just try to check if

$$\left| \frac{1}{n} \sum c_i - 0 \right| \text{ is large.}$$

(c_i: coin toss ± 1)

But how large? How many samples do we need? A classical result that we learn in Stats 101 is the Central Limit Theorem.

Theorem (CLT):

Let $X_1, \dots, X_n$ be iid with

$$\mathbb{E} X_i = \mu, \quad \text{Var } X_i = \sigma^2.$$

Define, $S_n = \sum_{i=1}^{n} X_i$

$$Z_n = \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}\, S_n}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

Then,

$$Z_n \xrightarrow{d} N(0,1).$$

↑ in distribution.

Thus, letting $Z \sim N(0,1)$, we have

$$\mathbb{P}[Z_n > t] \leq \underbrace{|\mathbb{P}[Z_n > t] - \mathbb{P}[Z > t]|}_{\text{Error } (t)} + \mathbb{P}[Z > t]$$

The second term is super nice

Fact (Exercise): $\mathbb{P}[Z > t] \leq e^{-t^2/2}$.

The second term is not so nice.

Fact (Berry Essen):

$\sup\{\text{Error}(t)\} = O\left(\frac{1}{\sqrt{n}}\right)$. ✝

Thus, using these results we can only conclude that if the losses are fair

$$P\left(\frac{1}{n}\sum^{n} c_i \geq 3/4\right) = P\left(Z_n \geq \frac{\sqrt{n}}{2}\right)$$

CLT ↘

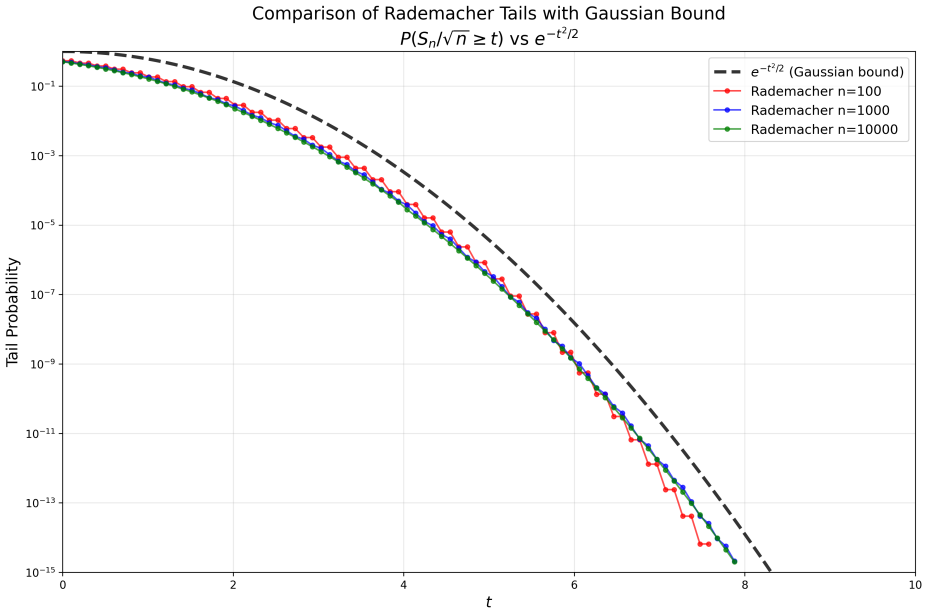
$$\leq \frac{C}{\sqrt{n}} + e^{-n/8}$$

↗ Giant compare to

Overview

We will cover:

- Concentration Inequalities
- Dimension Reduction
- Clustering
- Foundations of Statistical Inference
- " " " Learning
- Minimax Lower Bounds.


Comparison of Rademacher Tails with Gaussian Bound
$P(S_n/\sqrt{n} \geq t)$ vs $e^{-t^2/2}$
$e^{-t^2/2}$ (Gaussian bound)
Rademacher n=100
Rademacher n=1000
Rademacher n=10000
Tail Probability
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}
0
2
4
6
8
10
t