Lecture 1 Introduction

What this course is about:

How does one optimally extract information from data?

$S_n = \{z_1, z_2, \ldots, z_n\} \overset{iid}{\sim} P$

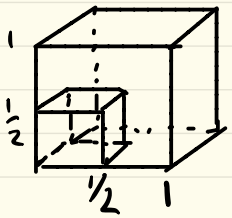
(P: probability measure on $\mathbb{R}^d$)

What makes this hard?

Curse of dimensionality

- n is big $\Rightarrow$ computation becomes harder
- d is big $\Rightarrow$ statistical accuracy becomes worse

Problems become harder in high dimensions



$Vol(2Q) = 2^d Vol(Q)$

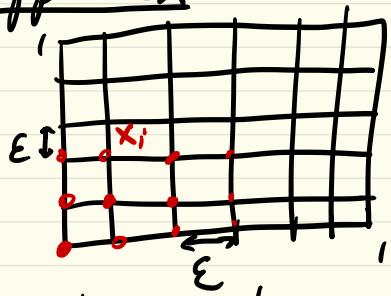
$\Rightarrow$ need exponentially many small balls to cover a set in HD.

Ex: (Monte-Carlo Simulation)

How do we compute an integral

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

Naive approach: Discretize



$(\frac{1}{\varepsilon})^d$ points

If f is Lipschitz, then

$$\left| \varepsilon^d \cdot \sum_i f(x_i) - \int_{[0,1]^d} f(x)\,dx \right| \leq O(\varepsilon)$$

So you need $\left(\frac{1}{\varepsilon}\right)^d$ function eval to estimate the integral to ε accuracy.

Can do better with a randomized algo:

$$E \triangleq \int_{[0,1]} f(x) = \mathbb{E} f(X)$$

where $X \sim \mathrm{Unif}([0,1]^d)$

Draw $x_1, \ldots, x_n \overset{iid}{\sim} \mathrm{Unif}([0,1]^d)$

and set $I = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$

Then

$$\mathbb{E}(I - E)^2 = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} f(x_i)\right)$$

$$= \frac{1}{n} \underbrace{\mathrm{Var}(f(X))}_{\text{bounded, if } f \text{ is bounded.}} =$$

(independence)

So expected error is

$$\boxed{\mathbb{E}|I - E| \leq \frac{1}{\sqrt{n}}}$$

Ex: (Machine Learning)

Suppose we see pairs

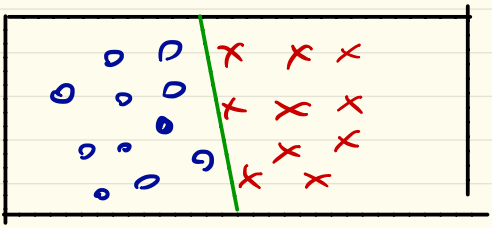
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$$

- x_i are called feature vectors
- y_i are called labels

Eg: x_i encodes properties of individual (incidence of smoking, distance of household to factory, genetics)

$$y_i = \begin{cases} 1 & \text{if cancer} \\ -1 & \text{if no cancer} \end{cases}$$

Goal: Find a predictive relationship between x and y.



Typical Approach:

- Fix a hypothesis class $\mathcal{H}$

Eg: $\mathcal{H} = \{h(x) = \langle v, x \rangle : v \in \mathbb{R}^d\}$

$\mathcal{H} = \{$quadratics in $x\}$

$\mathcal{H} = \{$neural networks$\}$

- Fix a loss function $\ell(t, y)$. (t: observed, y: prediction)

Eg: $\ell(t,y) = \frac{1}{2}(t-y)^2$ MSE

$\ell(t,y) = \max\{0, 1-ty\}$ Hinge

$\ell(t,y) = \log(1+e^{-ty})$ Logistic

. Solve Empirical Risk Minimization

(*) $\min_{h \in H} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)}_{Z_h}$ ← Optimization

Is this any good?

If $(x_i, y_i) \overset{i.i.d.}{\sim} P$ we would like (*) to approximate

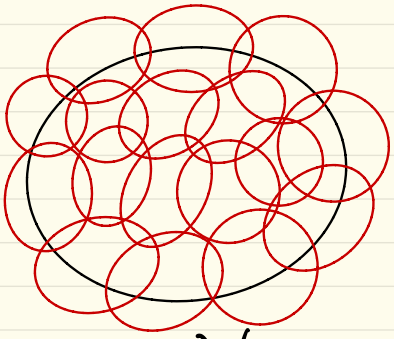
$$\min_{h \in H} \underbrace{\mathbb{E}_{(x,y)\sim P} \ell(h(x), y)}_{\mathbb{E} Z_h}$$

So we need to control

$$\sup_{h \in H} |Z_h - \mathbb{E} Z_h|$$

How to bound the supremum of infinitely many random variables?

One idea: discretize $\mathcal{H}$



Cover $\mathcal{H}$ with $\bigcup_{i=1}^{k} B_\varepsilon(h_i)$

- Control $|Z_{h_i} - \mathbb{E}Z_{h_i}| \le \delta$ with probability $1-p$
- Control $\sup_{i=1,\dots,k} |Z_{h_i} - \mathbb{E}Z_i| \le \delta$ with probability $1-kp$ [union bound]
- Extend by regularity of the loss $\sup_{h \in \mathcal{H}} |Z_h - \mathbb{E}Z_h| \le \delta + L(\varepsilon)$ w.p. $1-kp$

But: $k \approx \left(\frac{1}{\varepsilon}\right)^d$. So p needs to be much smaller than ε^d

$\Longrightarrow$ Concentration!