Artificial Intelligence

Instructor: Jie Shen

Dept. of Computer Science

September 3, 2024

AI applications...

- ChatGPT
- content generation
- auto driving, robotics
- ...

2018 : V1
late 19 : V2 → open-source from OpenAI
late 20 : V3 → ~~Closed~~ private
01/24 : 3.5
07/24 : V4

V2 : small : 120 M
XL : 4-5 B
Llama 3.1 : 8 B
70 B
405 B

≈ 2,000 GPU cards

48 GB VRAM

A100 : $16,000
RTX 6000 $6,000 × 8

Time : 6 months
> 30
Couple of years

performance

f (# params of model.

samples.

compute resource)

OpenAI 2020:

Scaling laws.

Conclusion: always go with larger models

if $ is fixed

$f \propto$ # param$^{0.73}$

samples$^{0.5}$

Deepmind 2022:

in 2020 is wrong.

param ≃ # samples

Meta 2023:

"You selected wrong obj func to optimize"

obj → inference cost

⇒ smaller & model

2024:

"I'm hot"

But...

Theoretical Foundation of AI

Syllabus

- Review of calculus, probability, linear algebra
- random projection
- singular value decomposition, principal component analysis
- dictionary learning and sparse coding
- low-rank matrix estimation, with applications to recommender systems
- Large language models: the Transformer
- GPT, Bert, DistillBert
- Scaling laws, chain of thought
- Contrastive learning
- Deep learning in bioinformatics

(20 min → proof, 15 min → paper review)

↓
mid-term
Oct 25

Course Staff

Instructor: Jie Shen (jie.shen@stevens.edu)

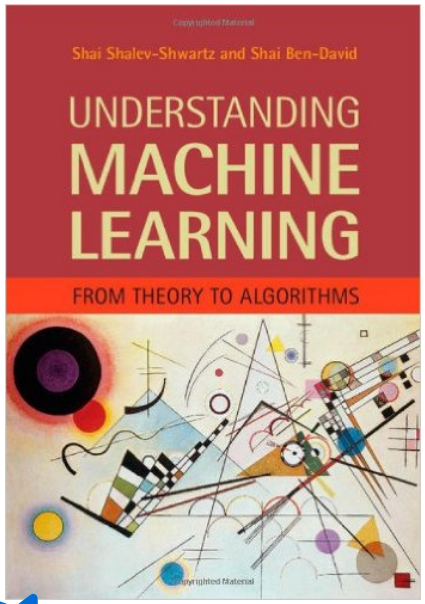
TA - Section A: Ziruo (Rosie) Zhao (zzhao83@stevens.edu)

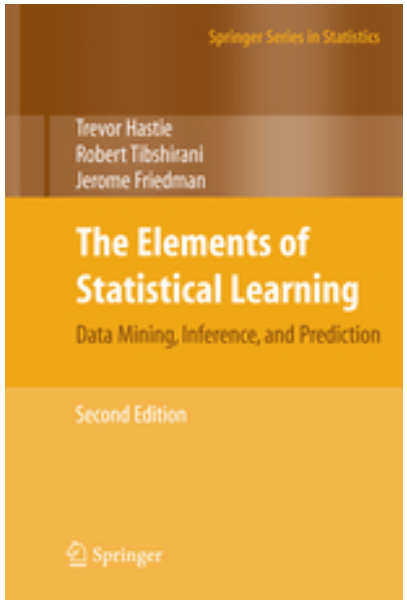
TA - Section B: Krishna Deb (kdeb@stevens.edu)

Office Hours: 1:00 - 2:00 Friday at GS 351

Textbook & Reference

- No Required Textbook
- Recommended (available online)





2014: 1000

2024: 10000+

- Research Papers in NeurIPS, ICML COLT pure theory

ICLR

oral presentation → spotlight

~100

Grading

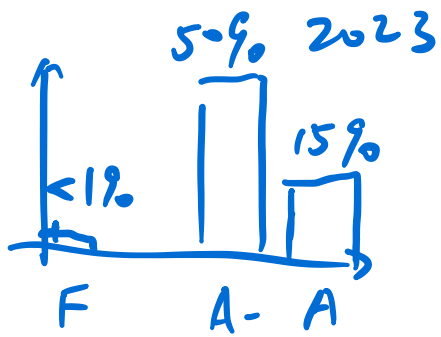
5 Problems 20 pt / problem

20 pt guaranteed.

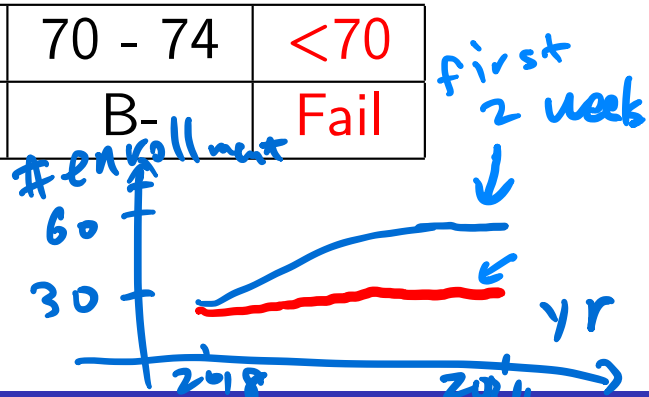
20 pt 99%

→ 60 pt : open-ended

- Midterm Exam (60%)
  - closed-book
  - **Section A: Oct 22 - mark it in your calendar!**
  - **Section B: Oct 25 - mark it in your calendar!**
- Final Paper Presentation (40%)
  - **Section A: Dec 10, or Dec 3 + Dec 10**
  - **Section B: Dec 13, or Dec 6 + Dec 13**
  - more details will be announce in October
- Final Grade



90 - 100	85 - 89	80 - 84	75 - 79	70 - 74	<70
A	A-	B+	B	B-	Fail



About the Course

Tough

- Not for introductory purpose
- Research oriented
  - Emphasize on both theoretical and application aspects
  - Analyze computational cost
  - Understand statistical accuracy testing loss v.s. #samples
- Strong background in calculus, linear algebra, and probability
  - If cannot do Quiz 0, consider dropping the course

About the Course

Overarching goal: Students can do independent research

- read research articles
- implement algorithms
- push the frontier of AI

General paradigm

- paper reading is assigned every week
- Not graded, but you are welcome to discuss during office hours

Quiz 0 (20 min)

4:10 → break

4:15 → continue.

~ 5:00 → dismiss

1. Let $\boldsymbol{x} = (1\ 2\ 3)$, $\boldsymbol{y} = (1\ 1\ 1)$. Calculate $\boldsymbol{x}\boldsymbol{y}^\top$ and $\boldsymbol{x}^\top\boldsymbol{y}$.

2. Show that for all $x > 0$, $\log(1+x) \leq x - \frac{x^2}{2} + \frac{x^3}{3}$.

3. Show that $\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}$ for all $x \in \mathbb{R}$, where e is the base of the natural logarithm.

apply Taylor wiki on LHS RHS

$f(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \log(1+x)$

$f(x) \geq 0$ when $x > 0$

$f(0) = 0$

$f(x) \geq f(0)$ $x > 0$

$f \uparrow \iff f' \geq 0$

$f : \mathbb{R}^d \to \mathbb{R}$

$f'(x)$ ~~$= 0$~~

$= \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots \frac{\partial f}{\partial x_d} \right)$

$x = (x_1, \cdots x_d)$

$x_i \in \mathbb{R}$

$f(x) = x_1 + x_2 + x_3 + \cdots x_d$

$\nabla f = (1, 1, 1, \cdots 1)$

$f(x) = x_1^2 x_2 + x_3 x_4$

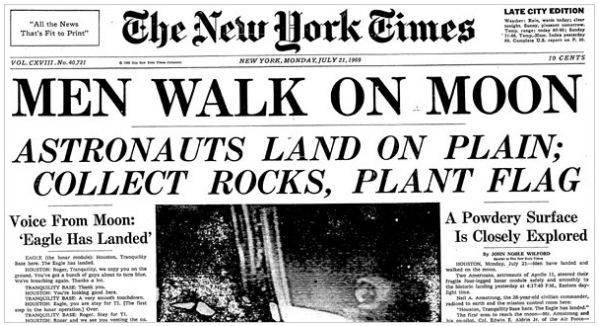
$\nabla f(x) = (2x_2 \cdot x_1, x_1^2, x_4, x_3)$

Linear Algebra Overview

A d-dimensional column vector $\boldsymbol{x}$ is a set of d numbers

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$$

- Bold lowercase letters for vectors
- Almost all the data is vector



Vector Operations

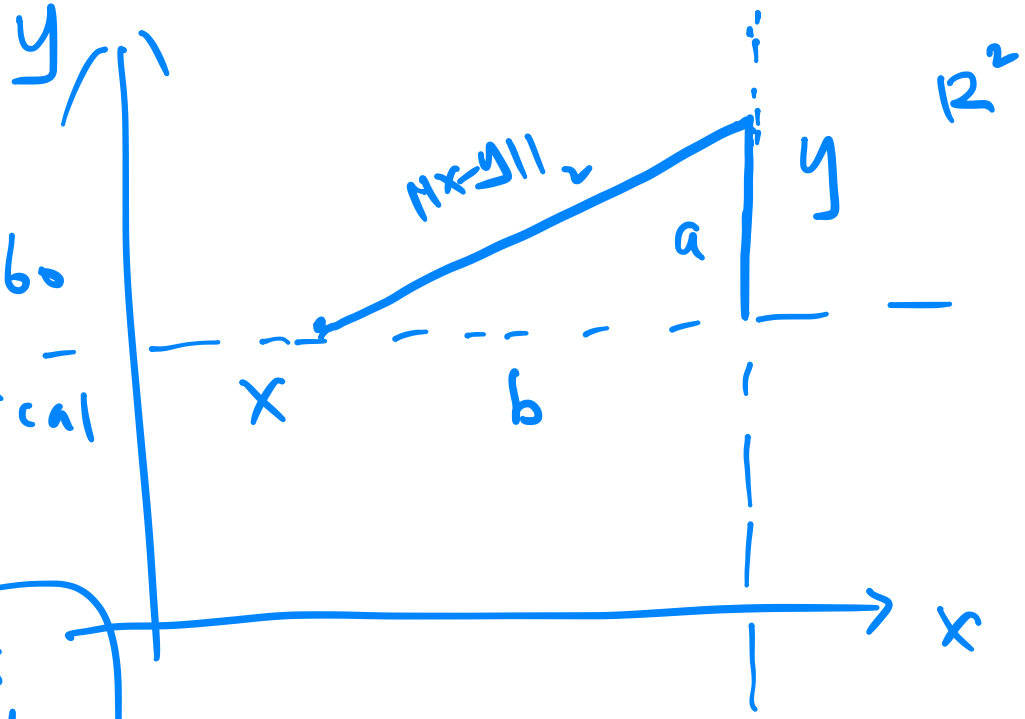
Suppose $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ are column vectors, $a, b \in \mathbb{R}$

- $\boldsymbol{x}^\top \stackrel{\text{def}}{=} \begin{pmatrix} x_1 & x_2 & \dots & x_d \end{pmatrix}$
- $a\boldsymbol{x} \stackrel{\text{def}}{=} \begin{pmatrix} ax_1 & ax_2 & \dots & ax_d \end{pmatrix}^\top$
- $\boldsymbol{x} + \boldsymbol{y} \stackrel{\text{def}}{=} \begin{pmatrix} x_1 + y_1 & x_2 + y_2 & \dots & x_d + y_d \end{pmatrix}$
- $a\boldsymbol{x} + b\boldsymbol{y}$
- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \stackrel{\text{def}}{=} \sum_{i=1}^d x_i y_i \in \mathbb{R}$
  - Sometimes use $\boldsymbol{x}^\top \boldsymbol{y}$, $\boldsymbol{x} \cdot \boldsymbol{y}$

Vector Norms

Linear reg: min w s.t. $\|w\|_2 \le \square$ → CS 560 Statistical ML.

- $\|\boldsymbol{x}\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{d} x_i^2}$
  - Broadly used
  - $\|\boldsymbol{x} - \boldsymbol{y}\|_2$
- $\|\boldsymbol{x}\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{d} |x_i|$ → saves samples
- $\|\boldsymbol{x}\|_\infty \stackrel{\text{def}}{=} \max_{1 \le i \le d} |x_i|$



$\|x-y\|_1 = a+b$

$\|\ \|_\infty = \max\{a,b\}$

over-fitting

train loss ≈ 0

test loss → +∞

test performance = f(# samples during training)

In some applications, # samples is capped.

Matrix

Vector: a set of numbers
Matrix: a set of vectors

- Bold capital letters $\boldsymbol{X} \in \mathbb{R}^{d \times n}$
- $\boldsymbol{X} = (x_{ij})_{1 \le i \le d, 1 \le j \le n} = (\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \dots \ \boldsymbol{x}_n)$
- $a\boldsymbol{X}$ for $a \in \mathbb{R}$
- $a\boldsymbol{X} + b\boldsymbol{Y}$ when $\boldsymbol{X}, \boldsymbol{Y}$ have the same size
- Multiplication: $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, $\boldsymbol{Y} \in \mathbb{R}^{p \times m}$
  - Can do $\boldsymbol{X}\boldsymbol{Y}$ only when $n = p$
  - $\boldsymbol{X}\boldsymbol{Y} \in \mathbb{R}^{d \times m}$
  - For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{x}^\top \boldsymbol{y} \in \mathbb{R}$, $\boldsymbol{x}\boldsymbol{y}^\top \in \mathbb{R}^{d \times d}$
- Transpose
- Symmetric matrix, diagonal
- Inverse of a square matrix

Probability Overview

Probability: measure of likelihood that an event will occur.

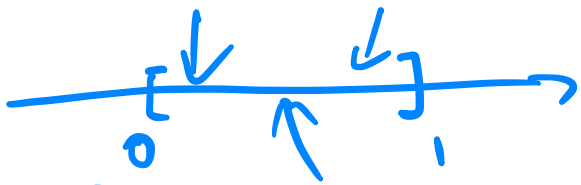
- From 0 to 1
- Coin tossing (heads or tails)

- Random variable X
- Events $= \{0, 1\}$
- X has distribution $\mathcal{D}$

Probability Overview

- If X is discrete, probability mass function $p(x) = P(X = x)$
  - Takes value from a countable set
  - $\{0, 1\}$
  - $\{1, 2, 3, \dots\}$
- If X is continuous, probability density function (PDF) $p(x)$

$$P(X \leq x) = \int_{-\infty}^{x} p(z)dz$$



  - Uniform distribution
  - Normal distribution

Gaussian: $N(0, 1)$

$$p(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

- $P(X \leq x)$: cumulative density function

Expected Value

$X_1 \quad X_2 \quad \cdots \quad X_N \qquad S = \sum_{i=1}^{N} X_i \geq 100$

$E[X_i] = 0.2 \qquad E[S] = 0.2N \geq 100$

$N \geq 500$

Expected value

- Discrete: $\mathbb{E}[X] = \sum xp(x)$
- Continuous: $\mathbb{E}[X] = \int xp(x)dx$
- Practice: Play a game for money. Each time

$$\Pr(X = 1) = 0.6, \quad \Pr(X = -1) = 0.4.$$

$N_1 = 100$ (Jie)

$N_2 = 10{,}000$ (")

$N_3 = 510$ ( )

300

When can we win 100 dollars?

$N = 500$

$\Pr(100) = 0.6^{100} \approx 0$

40 $N_1 \quad \cdots \quad N_{40}$

$\frac{1}{t} \sum_{j=1}^{t} N_j \approx 500$

$t \to \infty$

Expectation

- Average of multiple outcomes
- Not quite useful in practice
  - gambling
  - weather forecasting (rainy, sunny, dry)
  - in expectation = I guess
- But, $\mathbb{E}[X]$ implies $P(X)$ w.p. 299, if play 2000 time

You must win $100

Markov's Inequality

$\mathbb{E}[X] = 10$ $\quad \Pr(X > t) \leq \frac{10}{t}$

$\Leftrightarrow \Pr(X < t) \geq 1 - 10/t$

Theorem. If $X > 0$, $P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$ for all $t > 0$.

- Proof of correctness
- Proof of tightness
- Negative random variables
  - moment-generating function

pick $t = 100$

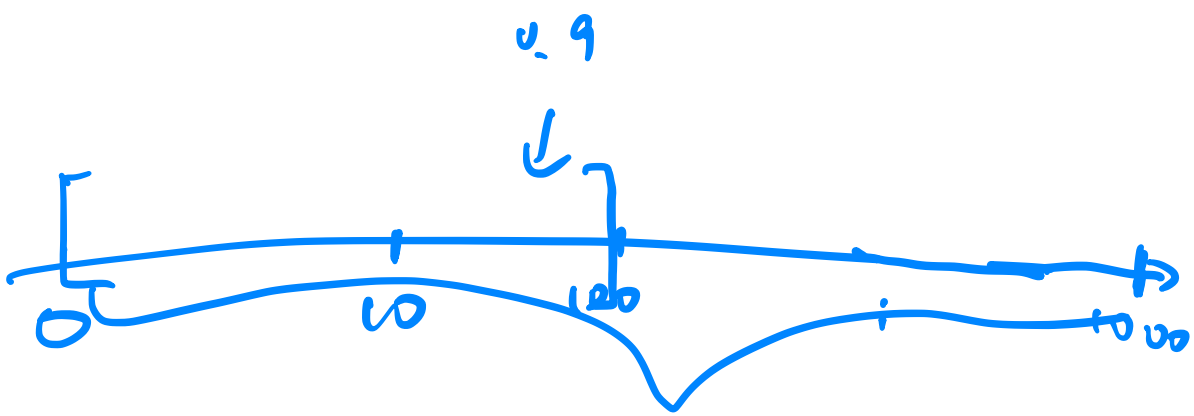
$\Rightarrow \Pr(X < 100) \geq 1 - 0.1 = 0.9$

w.p. 0.9, $X < 100$

$t = 1000$

$\Pr(X < 1000) \geq 0.99$

w.p. 0.99 $X < 1000$



Variance

- $\mathrm{Var}(X) \stackrel{\text{def}}{=} \mathbb{E}[X - \mathbb{E}X]^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$
- $X_1, X_2, \ldots X_n$ are independent, then $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$

→ large var

→ Today : small var

Chebyshev's Inequality

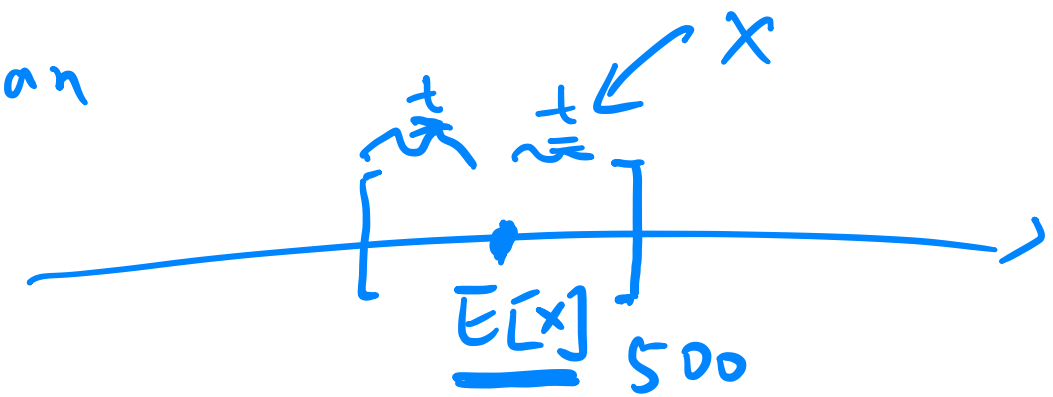
$$\Pr(|X - E[X]| > t) \leq \frac{Var(X)}{t^2}$$

$\forall X. \quad \forall t > 0$ ⇕

w.p. $1 - \frac{Var(X)}{t^2}$.

$|X - E[X]| \leq t$

X concentrates around its mean



w.p. $1 - \frac{Var(X)}{t^2}$

$|X - E[X]| \leq t$

$S = \sum_{i=1}^{N} X_i \geq 100$

$E[S] = 0.2N$

$Var[S] = Var(\sum X_i)$

$= \sum Var(X_i) = 0.96N$

$Var(X_i) = EX_i^2 - (EX_i)^2$

$= 1 - 0.2^2 = 0.96$

$\Rightarrow$ w.p. $1 - \frac{0.96N}{t^2} = 0.99$

$= 100$ ~~$|S - 0.2N| \leq t$~~

$0.2N - t \leq S \leq 0.2N + t$

$t > 0$

$t = 0.2N - 100 > 0 \Rightarrow N > 500$

$$\begin{cases} 0.2N - t = 100 \\ 1 - \dfrac{0.96N}{t^2} = 0.99 \end{cases} \Rightarrow \begin{matrix} N_1 = < 500 \\ N_2 = 3000 \end{matrix}$$

Solve over t, N

$\Rightarrow N = \square$

HW① Solve $N = \begin{matrix} \approx 565 \approx 500 \\ \approx 3000 \approx 3000 \\ \approx 3325 \end{matrix}$

HW② write a python program to see empirical result

Theory: $N = 3000$ $\hat{N} \approx 500$ 450 [450 600]

Emp: $N < 600$ correct

Hoeffding's Inequality

Symmetric Bernoulli distribution: $P(X = 1) = P(X = -1) = 1/2$

Theorem. Let $X_1, X_2, \ldots, X_n$ be independent symmetric Bernoulli random variables. Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$. Then, for any $t \geq 0$,

$$P\left(\sum_{i=1}^{n} a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|\boldsymbol{a}\|_2^2}\right)$$

Hoeffding's

- Proof of correctness
- Generalize to non-symmetric distribution

$$P(X = 1) = p \in [0, 1], \quad P(X = -1) = 1 - p$$

→ Hoeffding's on bounded R.V.

Assume $X_1, \dots X_n$ independent

$X_i \in [a, b] \quad \forall 1 \le i \le n$

Then.

$$Pr\left(\left| \sum X_i - E \sum X_i \right| > t \right)$$

$$\le e^{-\frac{t^2}{4(b-a)^2 n}}$$

In our case.

$X_i \in [-1, 1]$

are independent.

$$S = \sum X_i$$

$$Pr\left(|S - 0.2N| > t \right) \le e^{-\frac{t^2}{16n}}$$

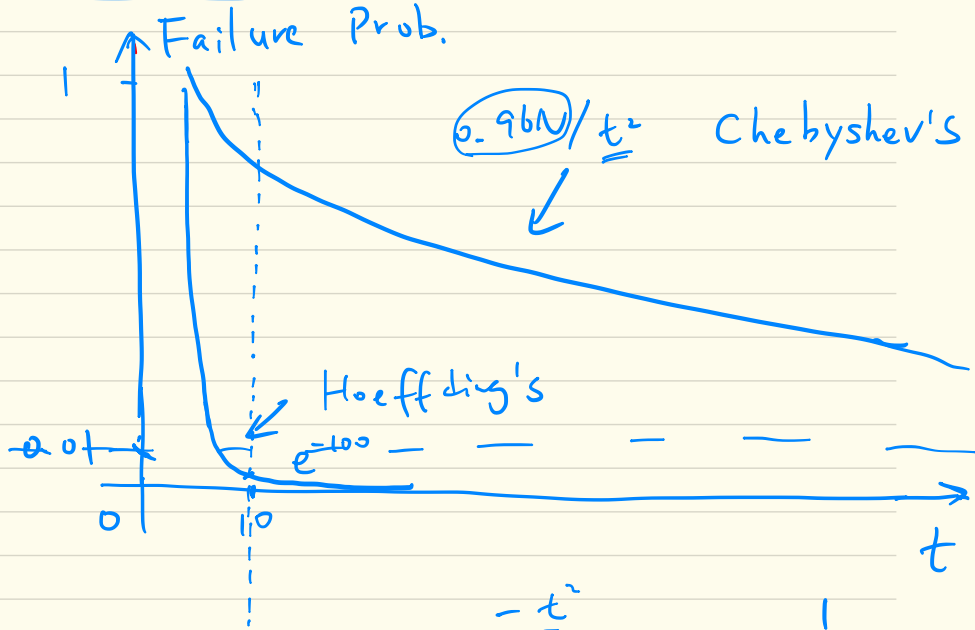
$\Leftrightarrow$ w.p. $1 - e^{-t^2/16}$

$$|S - 0.2N| < t$$

Hoeffding's v.s. Chebyshev's

w.p. $1 - e^{-t^2/16N}$ $\quad$ $1 - \frac{0.96N}{t^2}$

$|S - 0.2N| \le t = 10$ $\quad$ $|S - 0.2N| \le t = 10$



$t \uparrow$ $\quad e^{-t^2} \downarrow$

$t = 10$

w.p. $0.99\ldots9$ (>100)

e^{-t^2} $\quad e^{-100}$ $\quad 0.000\ldots01$ (>100)

$\frac{1}{t^2}$ $\quad \frac{1}{100} = 0.01$ $\quad$ w.p. 0.99

Hoeffding's

C.

@. w.p. $0.99\cdots9$ (>100 nines)

$|S - 0.2N| < 10$

w.p. 0.99

$|S - 0.2N| < 10$

$e^{-t^2} = 0.01$

$t = \sqrt{\ln 100}$

$= \sqrt{2\ln 10}$ (≈ 2.4)

$< \sqrt{8} \approx 2.5$

w.p. 0.99

$|S - 0.2N| < 2.5$

$\frac{1}{t^2} = 0.01$

$t = 10$

w.p. 0.99

$|S - 0.2N| < 10$

Today: ~~For~~ Try Hoeffding's to solve same problem.

$N = \square$

stronger →

	M.	C.	H.
w.p.	$1-\frac{EX}{t}$	$1-\frac{Var(X)}{t^2}$	$1-e^{-t^2/n}$
	$X \le t$	$\lvert X-EX\rvert \le t$	$\lvert \sum X_i - E\sum X_i\rvert \le t$
			~~$X-EX$~~
Condition			
①	$EX < \infty$	both $EX < \infty$ and EX^2 ($Var X$)	$E\lvert X\rvert^k < \infty$ for *all* $k = 1, \dots \infty$
		$Var(X) = EX^2 - (EX)^2$, $< \infty$, $< \infty$	
②	$X > 0$	any X	$X = \sum X_i$, X_i are indep.

estimate: stronger →

condition: stronger →

To show: $\boxed{\begin{aligned} Pr(X_i = +1) &= 0.5 \\ Pr(X_i = -1) &= 0.5 \end{aligned}}$

$$Pr\left(\sum X_i - E\sum X_i > t\right) \le e^{-\frac{t^2}{2n}}$$

Proof: LHS

$\forall \lambda > 0$ $= Pr\left(\lambda\left(\sum X_i - E\sum X_i\right) > \lambda t\right)$

e^x $= Pr\left(e^{\lambda\left(\sum X_i - E\sum X_i\right)} > e^{\lambda t}\right)$

($e^{\lambda\left(\sum X_i - E\sum X_i\right)} > 0$, $e^{\lambda t} > 0$)

Markov's $\le \dfrac{E\left[e^{\lambda\left(\sum X_i - E\sum X_i\right)}\right]}{e^{\lambda t}}$

$= \dfrac{E\left[e^{\lambda \cdot \sum X_i}\right]}{e^{\lambda t}}$ Holds $\forall \lambda > 0$

$\le \dfrac{e^{\frac{n}{2}\lambda^2}}{e^{\lambda t}} = e^{\frac{n}{2}\lambda^2 - \lambda t}$

$e^{a+b} = e^a \cdot e^b$

$$E\left[e^{\lambda \sum X_i}\right] = E\left[\prod_{i=1}^{n} e^{\lambda X_i}\right]$$

since $X_1, \ldots X_n$ are indep.

$\Rightarrow e^{\lambda X_1}, \ldots e^{\lambda X_n}$ are indep.

A, B are indep. $\Rightarrow E[A \cdot B] = E[A] \times E[B]$

$$= \prod_{i=1}^{n} E\left[e^{\lambda X_i}\right] \rightarrow \text{MGF of } X_i$$

$$= \prod_{i=1}^{n} \left(e^{\lambda \cdot 1} \cdot \frac{1}{2} + e^{\lambda \cdot (-1)} \cdot \frac{1}{2}\right)$$

$$= \prod_{i=1}^{n} \frac{1}{2}\left(e^{\lambda} + e^{-\lambda}\right)$$

$$\leq \prod_{i=1}^{n} e^{\frac{\lambda^2}{2}} = e^{\frac{n}{2}\lambda^2}$$

$e^{-t^2/2n}$

$\Rightarrow \quad \forall \lambda > 0$

$$Pr\left(\sum X_i - E\sum X_i > t\right) \leq e^{\frac{n}{2}\lambda^2 - \lambda t}$$

minimize

Find "optimal" λ w.r.t λ.

$\lambda = 1.$

~~$\lambda = 10$~~

$\lambda = 7.777$

RHS = 0.3

(RHS) = 0.1

RHS = 0.4

$\lambda = 1$ w.p. 0.7.

$\sum X_i - E < t$

($\lambda = 10$) w.p. 0.9

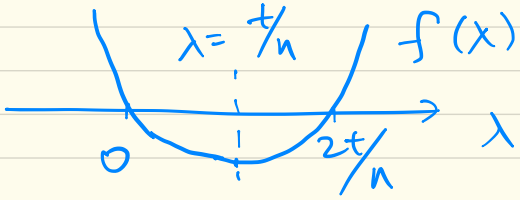
$\sum X_i - E < t$

$\lambda = 7.7777$ w.p. 0.6

$\sum X_i - E < t$

$$\min f(\lambda) = \frac{n}{2}\lambda^2 - t\lambda$$

s.t. $\lambda > 0$



$$f(t/n) = -\frac{t^2}{2n}$$

4 : 20

$$Pr(|X - EX| > t) < f(t)$$

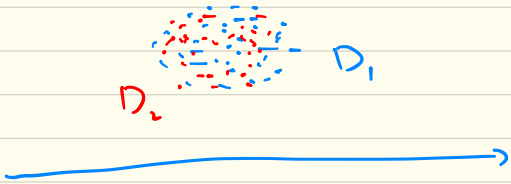
↑ Failure prob.

w.p. $1 - f(t)$

$$|X - EX| \leq t$$

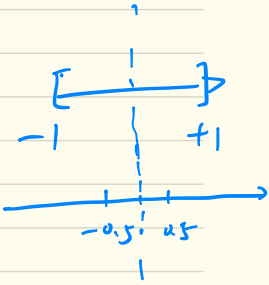
Hoeffding's needs all info of distribution / $\mathbb{E}X^k < \infty$.

$D_1 = D_2$



$\mathbb{E}X = \mathbb{E}X$

$X \sim D_1$ $\quad X \sim D_2$



$\mathrm{Var}(X) = \mathrm{Var}(X)$

$D_1 \quad D_2$

$\mathbb{E}X^2 = \mathbb{E}X^2$

$D_1 \quad D_2$

$\mathbb{E}X^3 \quad \mathbb{E}$

$D_1 \quad D_2$

$\vdots$

$\mathbb{E}X^\infty \quad \mathbb{E}X^\infty$

if $\underset{D_1}{E} X^k = \underset{D_2}{E} X^k$

for all $k = 1, \ldots \infty$

$\Rightarrow D_1 = D_2$

$\Rightarrow$ any distribution is uniquely determined by the moments $\{E X^k\}_{k=1}^{\infty}$

Moment Generating func. of X:

$\triangleq \underset{X \sim D}{E}[e^{\lambda X}] = g(\lambda)$

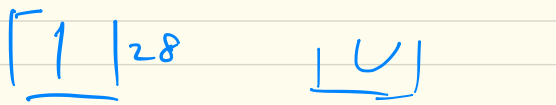
$e^y = 1 + y + \frac{1}{2} y^2 + \frac{1}{3} \cdot y^3 + \frac{1}{4} \cdot y^4 + \ldots$

$D \Leftrightarrow g(\lambda)$

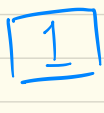
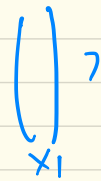
Random Projection

How to effective reduce dim ?

- - - dataset



784 - dim

Query  → 

Retrieve all imges with "1"

metric on x_1 . x_2

$d \to 0$ $x_1 \approx x_2$ $d(x_1, x_2)$

$\to \infty$ $x_1 \neq x_2$ $= \|x_1 - x_2\|_2$

n: # img in database.

d: dim

$O(nd)$

① reduce n → hashing (not in this course)

② reduce d → dim reduction ✓

$d = 784$ $k = 1$

10

100

Given k

How to evaluate performance of the new data?

X_i in DB.

$$\| q - X_i \|_2$$

$q', x_1' \cdots x_n' \in R^k$

Goal: on new data,

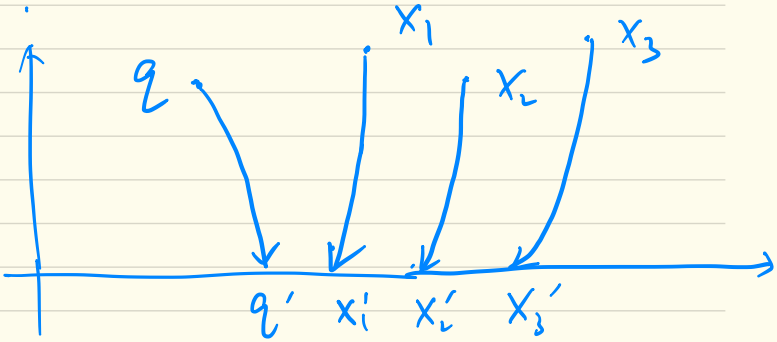
Top 100 are same as

Top 100 R^d

$$f: \quad q \to q' = f(q) \quad R^d \to R^k.$$

$$X_i \to X_i' = f(X_i)$$

$$\| q' - X_i' \| = \| f(q) - f(X_i) \|$$



$$\| f(q) - f(X_i) \|_2 \approx \| q - X_i \|_2$$

$f: X \to X'$

$x \to M \cdot x$.

M: $k \times d$ matrix

$M_{ij} \sim N(0, \frac{1}{=})$

↓

random matrix

random projection.

$\forall q, x$

$\| M \cdot q - M \cdot x \| \approx \| q - x \|$

$\Downarrow$

$\| M \cdot (q - x) \| \approx \| q - x \|$

Goal: w.p. 0.99 $\| M \cdot x \| \approx \| x \|$

$\| M \cdot x \|$: R.V.

$\| x \|$: $=$

$\| M \cdot x \| \approx E[\| M \cdot x \|]$

next week.

key msg from concentration inequalities:

w. h.p. $X \approx \mathbb{E}[X]$