CSE 312

Foundations of Computing II

Lecture 21: Chernoff Bound & Union Bound

Review Tail Bounds

Putting a limit on the probability that a random variable is in the "tails" of the distribution (e.g., not near the middle).

Usually statements in the form of

$$P(X \geq a) \leq b$$

or

$$P(|X - \mathbb{E}[X]| \geq a) \leq b$$

Review Markov's and Chebyshev's Inequalities

Theorem (Markov's Inequality). Let X be a random variable taking only non-negative values. Then, for any $t > 0$,

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Theorem (Chebyshev's Inequality). Let X be a random variable. Then, for any $t > 0$,

$$P(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Agenda

- Chernoff Bound
  - Example: Server Load
  - The Union Bound
- Probability vs statistics
  - Estimation

Chernoff-Hoeffding Bound

Theorem. Let $X = X_1 + \cdots + X_n$ be a sum of independent RVs, each taking values in $[0,1]$, such that $\mathbb{E}[X] = \mu$. Then…

for every $\delta \in [0,1]$, $\quad P(|X - \mu| \geq \delta \cdot \mu) \leq e^{-\frac{\delta^2 \mu}{4}}$ $\quad$ both tails

for every $\delta \geq 0$, $\quad P(X - \mu \geq \delta \cdot \mu) \leq e^{-\frac{\delta^2 \mu}{4}}$ $\quad$ right/upper tail

Herman Chernoff, Herman Rubin, Wassily Hoeffding

Example: If $X \sim \mathrm{Bin}(n, p)$, then $X = X_1 + \cdots + X_n$ is a sum of independent $\{0,1\}$-Bernoulli variables, and $\mu = np$

Note: More accurate versions are possible, but with more cumbersome right-hand side (e.g., see textbook)

Review Chernoff-Hoeffding Bound – Binomial Distribution

Theorem. (CH bound, binomial case) Let $X \sim \mathrm{Bin}(n,p)$. Let $\mu = np = \mathbb{E}[X]$. Then, for any $\delta \in [0,1]$,

$$P(|X-\mu| \geq \delta \cdot \mu) \leq e^{-\frac{\delta^2 np}{4}}.$$

Example:

$p = 0.5$
$\delta = 0.1$

n	Chebyshev $\frac{1}{\delta^2}\cdot\frac{1}{n}\cdot\frac{1-p}{p}$	Chernoff $e^{-\frac{\delta^2 np}{4}}$
800	0.125	0.3679
2600	0.03846	0.03877
8000	0.0125	0.00005
80000	0.00125	3.72×10^{-44}

Review Chernoff Bound – Example

$$\mathbb{P}(|X-\mu| \geq \delta \cdot \mu) \leq e^{-\frac{\delta^2 \mu}{4}}.$$

Alice tosses a fair coin n times, what is an upper bound for the probability that she sees heads at least $0.75 \times n$ times?

$p = 1/2$

$\mu = np = n/2$

Target $\frac{3n}{4} = \frac{n}{2} + \frac{n}{4} = \mu + \frac{1}{2}\mu$

Apply Chernoff bound with $\delta = \frac{1}{2}$

Bound is $e^{-\frac{\delta^2 \mu}{4}} = e^{-\frac{\left(\frac{1}{2}\right)^2\left(\frac{n}{2}\right)}{4}} = e^{-\frac{n}{32}}$

a. $e^{-n/64}$
b. $e^{-n/32}$
c. $e^{-n/16}$
d. $e^{-n/8}$

Chernoff vs Chebyshev – Summary

$$\frac{1}{\delta^2} \cdot \frac{1}{n} \cdot \frac{1-p}{p}$$

Chebyshev, linear decrease in n

vs

Chernoff, exponential decrease in n

$$e^{-\frac{\delta^2 np}{4}}$$

Why is the Chernoff Bound True?

Proof strategy (upper tail): For any $s > 0$:

- $P(X \geq (1+\delta) \cdot \mu) = P\left(e^{tX} \geq e^{t(1+\delta) \cdot \mu}\right)$
- Then, apply Markov + independence:

$$P\left(e^{tX} \geq e^{t(1+\delta) \cdot \mu}\right) \leq \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\delta)\mu}} = \frac{\mathbb{E}[e^{tX_1}] \cdots \mathbb{E}[e^{tX_n}]}{e^{t(1+\delta)\mu}}$$

- Find t minimizing the right-hand-side.

Agenda

- Chernoff Bound
  - Example: Server Load
  - The Union Bound
- Probability vs statistics
  - Estimation

Application – Distributed Load Balancing

We have k processors, and $n \gg k$ jobs.
We want to distribute jobs evenly across processors.

Strategy: Each job assigned to a randomly chosen processor!

X_i = load of processor i $\quad X_i \sim \mathrm{Binomial}(n, 1/k) \quad \mathbb{E}[X_i] = n/k$

$X = \max\{X_1, \dots, X_k\}$ = max load of a processor

Question: How close is X to n/k?

Distributed Load Balancing

Claim. (Load of single server)

$$P\left(X_i > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) \leq 1/k^4.$$

Example:

- $n = 10^6 \gg k = 1000$
- Perfect load balancing would give load $\frac{n}{k} = 1000$ per server
- $\frac{n}{k} + 4\sqrt{n \ln k \,/k} \approx 1332$
- *"The probability that server* i *processes more than* 1332 *jobs is at most 1-over-one-trillion!"*

Distributed Load Balancing

Claim. (Load of single server)

$$P\left(X_i > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) = P\left(X_i > \frac{n}{k}\left(1 + 4\sqrt{\frac{k \ln k}{n}}\right)\right) \le 1/k^4.$$

Proof. Set $\mu = \mathbb{E}[X_i] = \frac{n}{k}$ and $\delta = 4\sqrt{\frac{k}{n}\ln k}$

$$P\left(X_i > \mu\left(1 + 4\sqrt{\frac{k \ln k}{n}}\right)\right) = P\big(X_i > \mu(1+\delta)\big)$$

$$= P(X_i - \mu > \delta\mu) \qquad \text{Upper tail}$$

$$\le e^{-\frac{\delta^2 \mu}{4}} = e^{-4 \ln k} = \frac{1}{k^4}$$

$\delta^2 = 4^2 \cdot \frac{k \ln k}{n}$

so $\delta^2 \mu = 4^2 \ln k$

What about the maximum load?

Claim. (Load of single server)

$$P\left(X_i > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) \leq 1/k^4.$$

What about $X = \max\{X_1, \dots, X_k\}$?

Note: $X_1, \dots, X_k$ are <u>not</u> (mutually) independent!

In particular: $X_1 + \cdots + X_k = n$

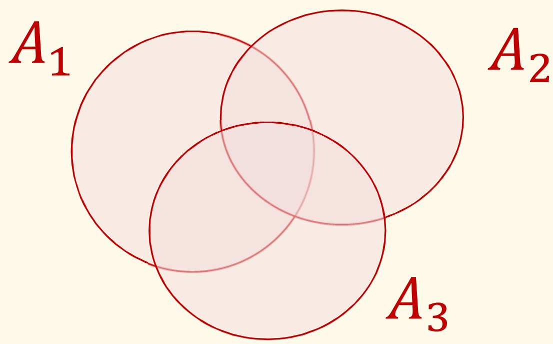
When non-trivial outcome of one RV can be derived from other RVs, they are non-independent.

Detour – Union Bound – A nice name for something you already know

Theorem (Union Bound). Let $A_1, \ldots, A_n$ be arbitrary events. Then,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Intuition (3 evts.):



Detour – Union Bound - Example

Suppose we have $N = 200$ computers, where each one fails with probability 0.001.

What is the probability that at least one server fails?

Let A_i be the event that server i fails.

Then event that at least one server fails is $\bigcup_{i=1}^{n} A_i$

$$P\left(\bigcup_{i=1}^{N} A_i\right) \leq \sum_{i=1}^{N} P(A_i) = 0.001N = 0.2$$

What about the maximum load?

Claim. (Load of single server)

$$P\left(X_i > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) \le 1/k^4.$$

What about $X = \max\{X_1, \dots, X_k\}$?

$$P\left(X > \frac{n}{k} + 4\sqrt{n \ln k\,/k}\right) = P\left(\left\{X_1 > \frac{n}{k} + 4\sqrt{n \ln k\,/k}\right\} \cup \cdots \cup \left\{X_k > \frac{n}{k} + 4\sqrt{n \ln k\,/k}\right\}\right)$$

$$\le P\left(X_1 > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) + \cdots + P\left(X_k > \frac{n}{k} + 4\sqrt{n \ln k\,/k}\right)$$

Union bound

$$\le \frac{1}{k^4} + \cdots + \frac{1}{k^4} = k \times \frac{1}{k^4} = \frac{1}{k^3}$$

What about the maximum load?

Claim. (Load of single server)

$$P\left(X_i > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) \leq 1/k^4.$$

Claim. (Max load) Let $X = \max\{X_1, \dots, X_k\}$.

$$P\left(X > \frac{n}{k} + 4\sqrt{\frac{n \ln k}{k}}\right) \leq 1/k^3.$$

Example:

- $n = 10^6 \gg k = 1000$
- $\frac{n}{k} + 4\sqrt{n \ln k / k} \approx 1332$
- *"The probability that some server processes more than* 1332 *jobs is at most 1-over-one-billion!"*

Using tail bounds

- Tail bounds are *guarantees*, unlike our use of CLT
- Often, we actually start with a target upper bound on failure probability
  - In the load-balancing example, the value of δ in terms of n and k was worked out in order to get failure probability $\leq 1/k^4$
    - We didn't start out with this weird value
  - See example in section and on homework
- We use these bounds to design (randomized) algorithms or analyze their guaranteed level of success.

Agenda

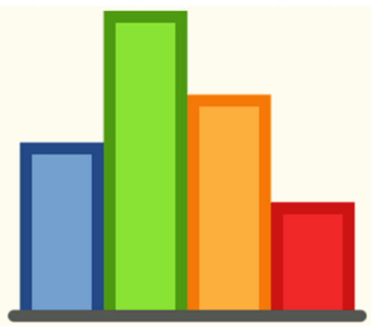
- Chernoff Bound
  - Example: Server Load
  - The Union Bound
- Probability vs statistics
  - Estimation

Probability vs Statistics

$\text{Ber}(p = 0.5)$

Probability
Given model, predict data

$P(THHTHH)$



$\text{Ber}(p = ??)$

Statistics
Given data, predict model

$THHTHH$

Recall Formalizing Polls

Population size N, true fraction of voting in favor p, sample size n.

Problem: We don't know p

What type of r.v. is X_i?

		$\mathbb{E}[X_i]$	$\mathrm{Var}(X_i)$
a.	Bernoulli	p	$p(1-p)$

Polling Procedure

for $i = 1, \ldots, n$:

1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote

$$X_i = \begin{cases} 1, & \text{voting in favor} \\ 0, & \text{otherwise} \end{cases}$$

Report our estimate of p: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Recall Formalizing Polls

We assume that poll answers $X_1, \dots, X_n \sim \mathrm{Ber}(p)$ i.i.d. for unknown p

Goal: Estimate p

We did this by computing $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Why is that a good estimate for p?

More generally …

In estimation we… .

- Assume: we know the type of the random variable that we are observing samples from
  - We just don't know the parameters, e.g.
    - the bias p of a random coin $\mathrm{Bernoulli}(p)$
    - The arrival rate λ for the $\mathrm{Poisson}(\lambda)$ or $\mathrm{Exponential}(\lambda)$
    - The mean μ and variance σ of a normal $\mathcal{N}(\mu, \sigma)$

- Goal: find the "best" parameters to fit the data
    - Next time: "best" = parameters that would be "most likely" to generate the observed samples