Plan

Logistics

Hashing Around the Clock

Concentration Inequalities

Load Balancing (Review-ish)

Thanks for coming to gamos!

Problem set tomorrow

All but one said they liked pace,

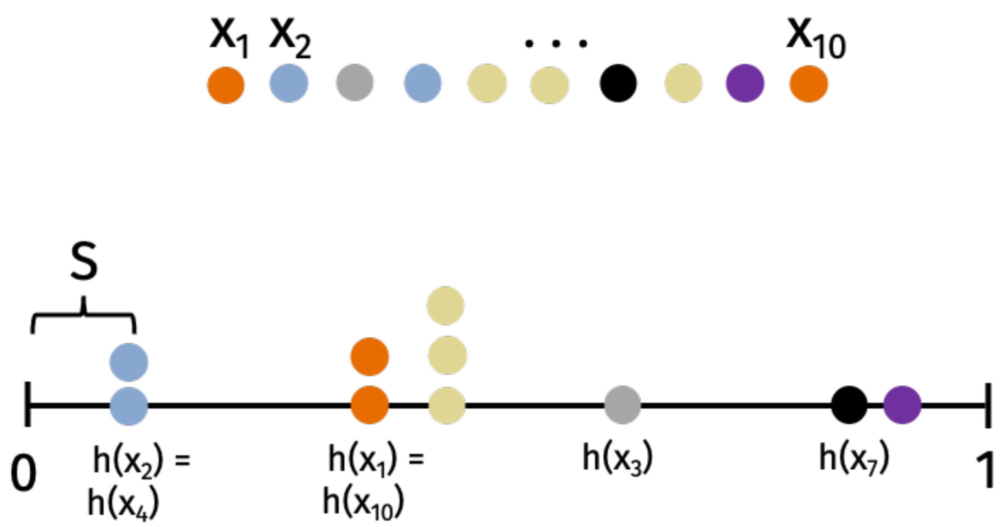
↳ I'll (try to) slow down

↳ more group activities

Not available tomorrow,

ask me today!!

Hashing Around the Clock



(1) $\mathbb{E}$[requests to move]

(2) Pr(any server "owns" c fraction) $\leq 1/10$

union bound

want

$\Pr(\text{one server "owns" } c) \leq \frac{1}{10n}$

hints on pset

$\Pr(\text{one server "owns" } c)$

$= (1-c)^{n-1}$ (then a miracle occurs)

$\leq \frac{1}{10n}$

Review in lecture

Concentration Inequalities

Chebyshev gave "disappointing" bound yesterday :(

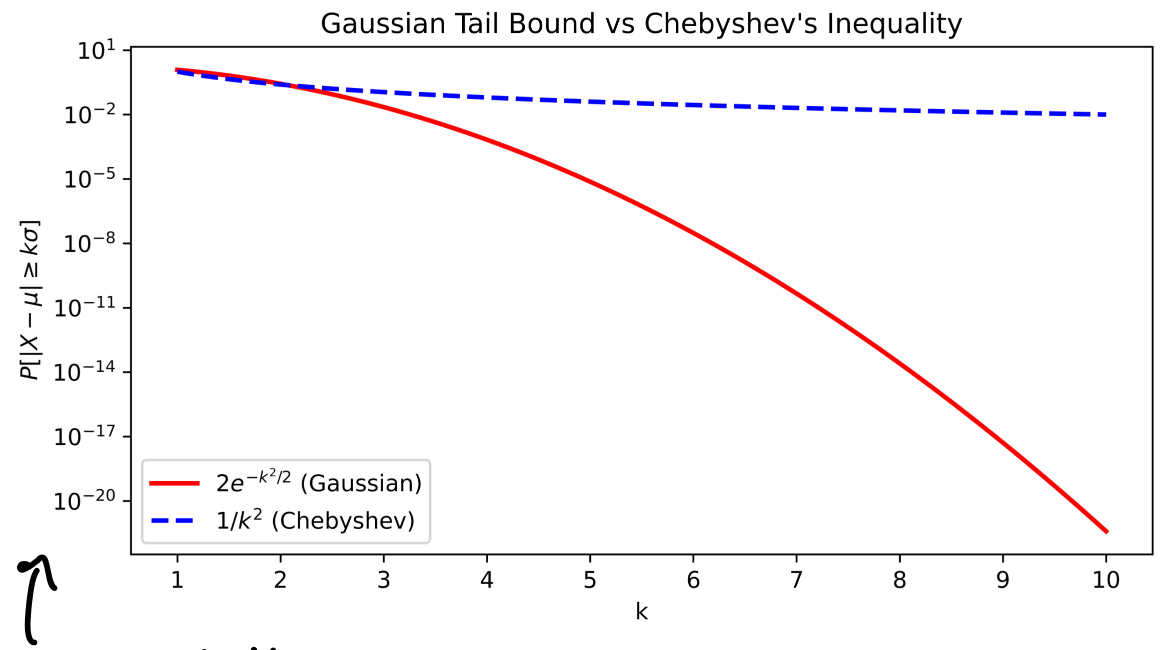
$\mu = \mathbb{E}[X] \qquad \sigma^2 = \mathrm{Var}(X)$

Chebyshev

$$\Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Gaussian X

$$\Pr(|X-\mu| \geq k\sigma) \leq 2e^{-k^2/2}$$



Log scale!! :)

Is Chebyshev just bad?

We need assumptions!

$\Pr(X_1 = x_1, \ldots, X_k = x_k)$

$= \Pr(X_1 = x_1) \ldots \Pr(X_k = x_k)$

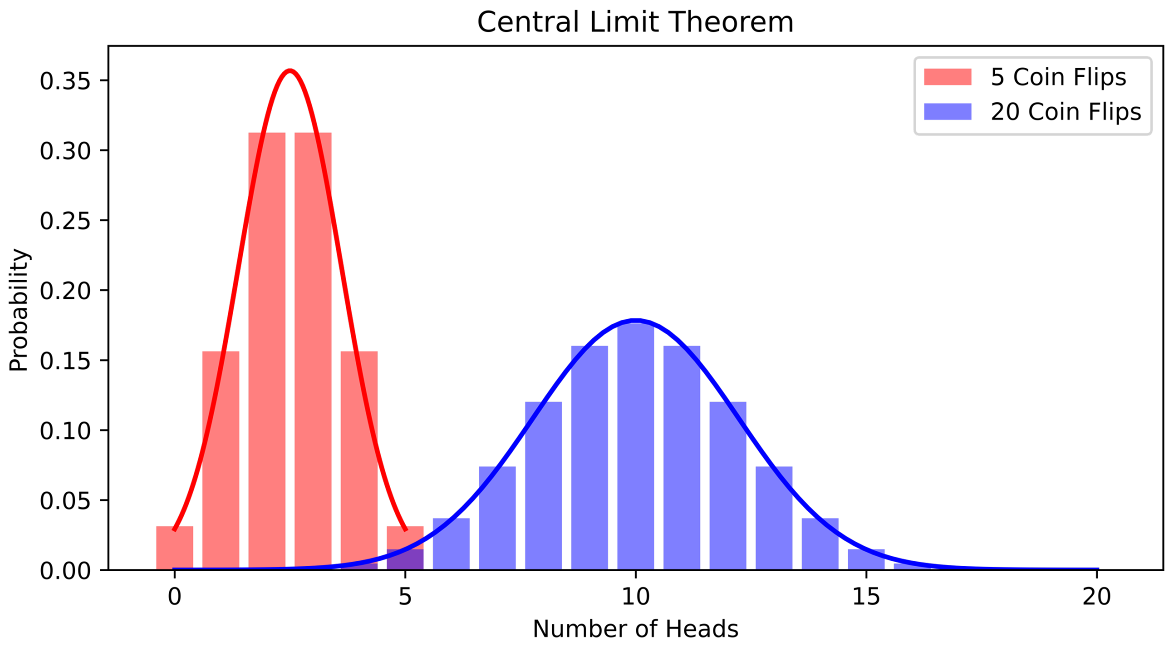
Central Limit Theorem: Any sum of $\boxed{\textit{mutually independent}}$ and identically distributed random variables $X_1, \ldots, X_k$ with mean μ and finite variance σ^2 converges to a Gaussian random variable with mean $k \cdot \mu$ and variance $k \cdot \sigma^2$ as k goes to infinity. Formally,

!!

linearity of variance

$$\lim_{n \to \infty} \sum_{i=1}^{k} X_i \sim \mathcal{N}(k\mu, k\sigma^2).$$

linearity of expectation



better approximation

coin flip

$$H = \sum_{i=1}^{100} c_i$$

$$\mathbb{E}[H] = 50$$

$$Var(H) = 25$$

Chebyshev:

$$Pr(|X-50| \geq 20) \leq .0625$$

If CLT held exactly,

$$Pr(|X-50| \geq k \cdot 5) \leq 2e^{-k^2/2}$$

$$k = 4$$

$$Pr(|X-50| \geq 20) \leq 2\exp(-16/2)$$

$$= .00067$$

Let's be formal! ↳ Different forms

↳ Use typed notes and/or wikipedia

↳ Different assumptions $\Rightarrow$ different bounds

indicator!! :)

Chernoff Bound: Let $X_1, \ldots, X_k$ be independent $\boxed{\text{binary}}$ random variables. That is, $X_i \in \{0,1\}$. Let $p_i = \mathbb{E}[X_i]$ where $0 < p_i < 1$. Choose a parameter $\epsilon > 0$. Then the sum $S = \sum_{i=1}^k X_i$, which has mean $\mu = \sum_{i=1}^k p_i$, satisfies

$$\Pr(S \geq (1+\epsilon)\mu) \leq \exp\left(\frac{-\epsilon^2\mu}{2+\epsilon}\right)$$

and, if $0 < \epsilon < 1$,

$$\Pr(S \leq (1-\epsilon)\mu) \leq \exp\left(\frac{-\epsilon^2\mu}{2}\right).$$

$$\Rightarrow \quad \Pr(|S-\mu| \geq \epsilon\mu)$$

$$\leq \exp\left(-\frac{\epsilon^2\mu}{2+\epsilon}\right) + \exp\left(\frac{-\epsilon^2\mu}{2}\right)$$

$$\overset{*}{\leq} 2\exp\left(-\frac{\epsilon^2\mu}{3}\right)$$

$2+\epsilon \leq 3$ when $\epsilon < 1$

$\frac{1}{3} \leq \frac{1}{2+\epsilon}$

$-\frac{1}{2+\epsilon} \leq -\frac{1}{3} \Rightarrow$

$$\exp\left(\frac{-\epsilon^2\mu}{2+\epsilon}\right) \overset{*}{\leq} \exp\left(\frac{-\epsilon^2\mu}{3}\right)$$

Less restrictive?

between -1 and 1

Bernstein Inequality: Let $X_1, \ldots, X_k$ be independent random variables with each $X_i \in [-1, 1]$. Let $\mu = \sum_{i=1}^k \mathbb{E}[X_i]$ and $\sigma^2 = \sum_{i=1}^k \mathrm{Var}[X_i]$. Then, for any $k \le \frac{\sigma}{2}$, the sum $S = \sum_{i=1}^k X_i$ satisfies

$$\Pr(|S - \mu| > k\sigma) \le 2\exp\left(\frac{-k^2}{4}\right).$$

between a_i and b_i

Hoeffding's Inequality: Let $X_1, \ldots, X_k$ be independent random variables with each $X_i \in [a_i, b_i]$. Let $\mu = \sum_{i=1}^k \mathbb{E}[X_i]$. Then, for any $k > 0$, the sum $S = \sum_{i=1}^k X_i$ satisfies

$$\Pr(|S - \mu| > k) \le 2\exp\left(\frac{-k^2}{\sum_{i=1}^k (b_i - a_i)^2}\right).$$

Coin Flips

$$S = \sum_{i=1}^{k} X_i \qquad X_i = \begin{cases} 1 & \text{wp } b \\ 0 & \text{wp } 1-b \end{cases}$$

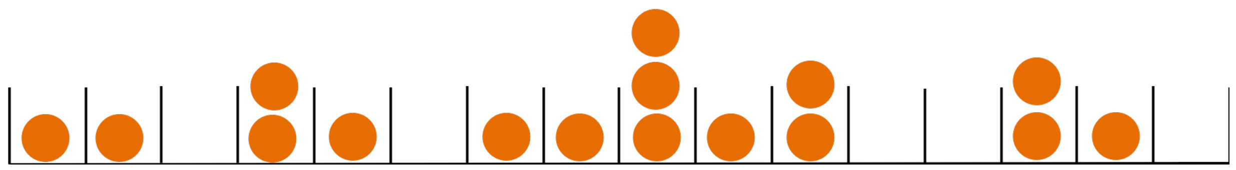
Choose $k \geq \dfrac{3\log(2/\delta)}{\epsilon^2}$

(1) $\mathbb{E}[S] = bk$

(2) $\Pr(\,|S - bk| \geq \epsilon k\,) \leq \delta$

Load Balancing

m requests to n servers



$$\Pr\left(\max_i S_i \ge c\right) \le 1/10$$

$$\Pr\left(S_i \ge c\right) \le \frac{1}{10n}$$

$$S_i = \sum_{j=1}^{m} \mathbb{1}[j \text{ goes to } i]$$

binary independent sum!

Chernoff

$$\Pr(S_i \ge 1+\epsilon) \le \exp\left(\frac{-\epsilon^2}{2+\epsilon}\right)$$

$$\exp\left(\frac{-\epsilon^2}{2+\epsilon}\right) \overset{\text{want}}{\le} \frac{1}{10n} \qquad \epsilon \ge 2$$

$$\exp\left(\frac{-\epsilon^2}{2+\epsilon}\right) \le \exp\left(\frac{-\epsilon^2}{2\epsilon}\right) \overset{\text{want}}{=} 1/10n$$

$$\frac{-\epsilon}{2} = \log(1/10n)$$

$$\epsilon = 2\log(10n)$$

$$\Pr\left(S_i \geq 1 + \sqrt{3\log(10n)}\right) \leq \frac{1}{10n}$$

$$\Pr\left(S_i \geq O(\log n)\right) \leq \frac{1}{10n}$$

$$\Rightarrow \quad \Pr\left(\max_i S_i \geq O(\log n)\right) \leq \frac{1}{10n}$$

Practice: Hash to 2 servers and choose least loaded

$O(\log n)$ or $O(\log\log n)$ or $O(1)$ maximum load?

Log logn on desmos! crazy!

atoms in universe $\approx 10^{82}$

$$\log_{10}\log_{10} 10^{82} = \log_{10} 82 \approx 1.91$$