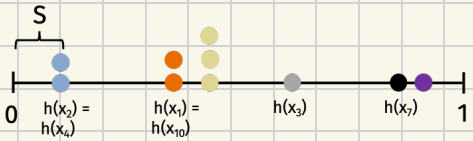
Tuesday, February 10

- Thank you for coming to GEMS/my talk!
- CCMS Applied Math seminar Mondays: 4:15 pm Estella 1021
- CCMS Colloquium Fridays: 11am Davidson Lecture Hall
- No OH Thursday, DM to meet Friday!

Plan

- Clean up discrete elements
- Load balancing

Distinct Elements



$$\bar{S} \leftarrow \sum_{j=1}^{k} \frac{S_j}{k}$$

$$D \leftarrow \frac{1}{\bar{S}} - 1$$

wp $1-\delta$ in $O\left(\frac{\log D}{\epsilon^2 \delta}\right)$ space:

$$(1-\epsilon)\mu \le \bar{S} \le (1+\epsilon)\mu$$

$$\Leftrightarrow \frac{1}{(1+\epsilon)\mu} \le \frac{1}{\bar{S}} \le \frac{1}{(1-\epsilon)\mu}$$

$1-2\epsilon \le \frac{1}{1+\epsilon}$; $\frac{1}{1-\epsilon} \le 1+2\epsilon$ by Desmos

$$\Leftrightarrow (1-2\epsilon)\frac{1}{\mu} - 1 \le \frac{1}{\bar{S}} - 1 \le (1+2\epsilon)\frac{1}{\mu} - 1$$

$\mu \le \frac{1}{2}$

$$\Leftrightarrow (1-4\epsilon)\left(\frac{1}{\mu} - 1\right) \le \hat{D} \le (1+4\epsilon)\left(\frac{1}{\mu} - 1\right)$$

$$\Leftrightarrow (1-4\epsilon) D \le \hat{D} \le (1+4\epsilon) D$$

$$(1-\epsilon) D \le \hat{D} \le (1+\epsilon) D$$

where we hide 4 in big-O

Load Balancing

Goal: Distribute load evenly between servers using hash function

e.g., Google maps routing

Advantage of hash: cache redundant queries

n servers and m (unique) queries

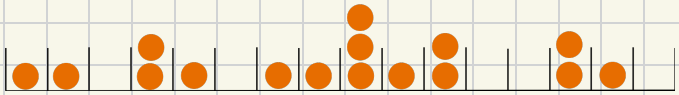
S_i = # queries sent to server i

$$= \sum_{j=1}^{m} \mathbb{1}[h(x_j) = i]$$

Bound $S = \max_{i \in [n]} S_i$

- $\mathbb{E}[S_i] =$
- $\mathbb{E}[S] \overset{2.}{=} \max_{i \in [n]} \mathbb{E}[S_i]$

Let's assume $m = n$, so $\mathbb{E}[S_i] = 1$



Goal: Prove

$$\Pr\left(\max_i S_i \geq C\right) \leq 1/10$$

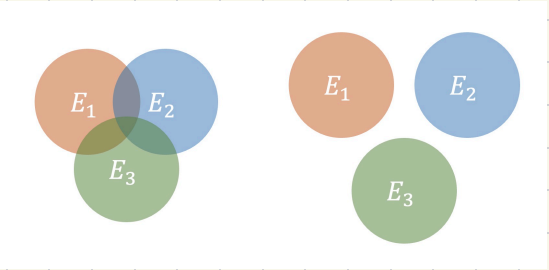
$$\Pr(S_1 \geq C \cup S_2 \geq C \cup \dots \cup S_n \geq C) \leq 1/10$$

By union bound,

$$\Pr(S_1 \geq C \cup \dots \cup S_n \geq C) \leq \sum_{i=1}^{n} \Pr(S_i \geq C) = n \Pr(S_i \geq C) \overset{\text{want}}{\leq} n \cdot \frac{1}{10n}$$

Union Bound

Proof by picture:



Proof by Markov's:

$$\Pr(E_1 \cup \dots \cup E_n) = \Pr\left(\sum_{i=1}^{n} \mathbb{1}[E_i] \geq 1\right)$$

$$\leq \mathbb{E}\sum_{i=1}^{n} \mathbb{1}[E_i] = \sum_{i=1}^{n} \Pr(E_i)$$

Prove that

$\Pr(s_i \geq c) \leq \frac{1}{10n}$

- Markov's ?
- Chebyshev's ?
- what c?

Thursday, February 12

Plan

Concentration Inequalities

Revisit Coins + load balancing

Chebyshev's Inequality

Let X rv with $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}(X)$

For $k > 0$,

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

On load balancing, showed max load $\geq \Theta(\sqrt{n})$ whp.

Using tighter inequality, we can show max load $\geq \Theta(\log n)$ whp.

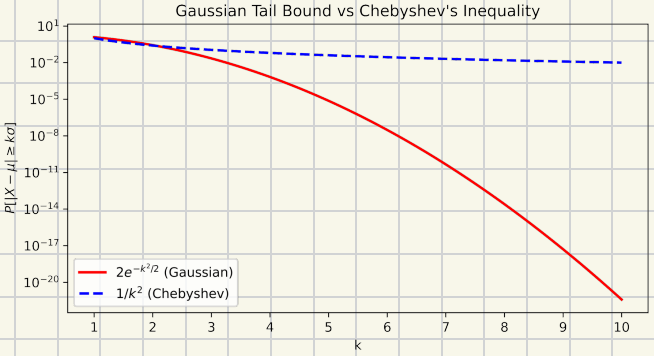
Q: When is Chebyshev loose?

Gaussian Tail Bound

Let $X \sim \mathcal{N}(\mu, \sigma^2)$

For $k > 0$,

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



Q: Is Chebyshev always loose?

Q: For which RV is chebyshev loose?

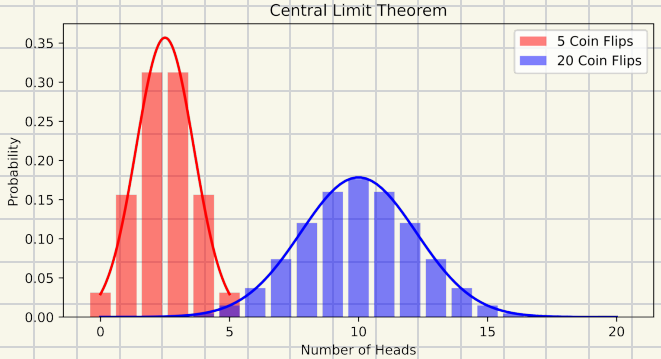
Central Limit Theorem:

The sum of mutually independent $X_1, \ldots, X_n$ with $\mu = \mathbb{E}X_i$, $\sigma^2 = \mathrm{Var}(X_i)$ converges to $\mathcal{N}(n\mu, n\sigma^2)$ i.e.,

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

↑ also holds for rv w/ different mean, var

↑ Today: When does CLT "kick in"?



$n = 100$ $\mu = 50$ $\sigma^2 = 25$ $k = 4$

Chebyshev: $\Pr(|X-\mu| \geq k\sigma) \leq 1/16 = 0.0625$

Exponential Tail Bound: $\Pr(|X-\mu| \geq k\sigma) \leq 2e^{-k^2/2} \approx .0006$

Exponential Concentration Inequalities

$X_1, \dots, X_n$ indep

$S = \sum_{i=1}^{n} X_i \qquad \mu = \mathbb{E}[S_i]$

↳ Make assumptions on X_i:
stronger assumption = stronger bound

↳ Proved using clever applications of Markov's

Chernoff Bound

Binary $X_i \in \{0,1\}$. For $0 < \epsilon < 1$,

$$\Pr(|S-\mu| \geq \epsilon\mu) \leq 2\exp\left(\frac{-\epsilon^2\mu}{3}\right)$$

For $\epsilon > 0$,

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

$\mathrm{Var}\, X_i \leq \mathbb{E}[X_i]$

Bernstein Inequality

$X_i \in [-1,1]. \qquad \sigma^2 = \sum_{i=1}^{n} \mathrm{Var}(X_i).$

For $k \leq \sigma/2$,

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

$$\mathrm{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq (b_i - a_i)^2$$

Hoeffding's Inequality

$X_i \in [a_i, b_i]$

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

Revisiting Coin Flips

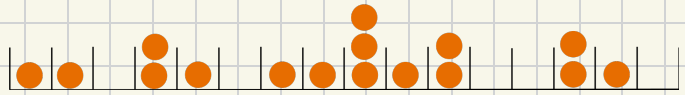
$X_i \in \{0,1\}$ $\quad \mathbb{E}[X_i] = b \ \forall\, i$

Choose $\delta > 0$, $1 > \epsilon > 0$, then $n \geq \frac{3\log(2/\delta)}{\epsilon^2}$.

$$\Pr(|S - bn| \geq \epsilon n) \leq \delta$$

Proof:

Revisiting Load Balancing



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

$S = \max_i S_i$

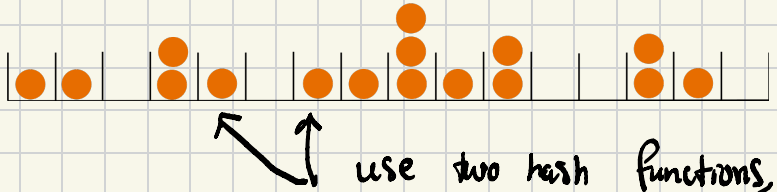
$$\Pr(S \geq c) \leq \frac{1}{10}$$

$\Leftarrow \quad \Pr(S_i \geq c) \leq 1/10n$

with Chebyshev's, $c = O(\sqrt{n})$

Q: Can we do better?

Power of Two Choices



use two hash functions, choose least occupied

Then $Pr(S \geq c) \leq 1/10$ for $c = \log n$

$c = \log \log n$

$c = \log \log \log n$

How about power of three choices?