Week 5: Cheroff Bounds and Sampling

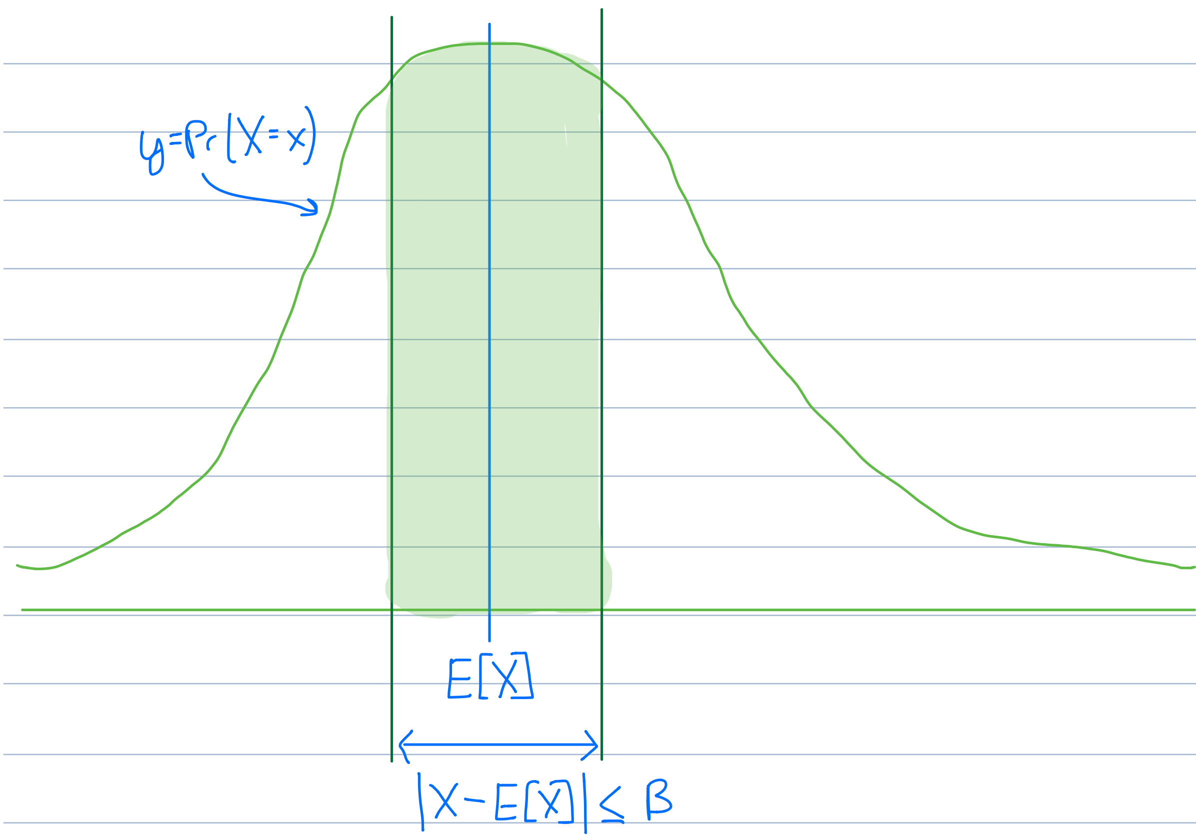
February 12, 2019

Last week: Hashing
(1) Chaining
(2) Linear Probing
(3) Cuckoo Hashing

Today
1) Chernoff-Hoeffding Bounds
2) Examples: Coin flipping
load balancing
Sampling
approximate counting

Deviation from the Mean

X = random variable
$E[X]$ = expected value



Goal: Show $\Pr(|X-E[X]| > B)$ is small.

Ex. Flipping coins

→ flip n coins

→ $Pr(\text{heads}) = 1/2$

→ $X = \#$ heads

1) $E[X] = \frac{n}{2} \leftarrow \mu$

→ $X_j = 1$ if j^{th} coin is heads

$= 0$ otherwise

→ $E[X_j] = Pr(X_j = 1) = 1/2$

→ $E[X] = E[\sum X_j] = \sum E[X_j] = n \cdot \frac{1}{2}$

2) Markov's Inequality: $Pr(X \geq \frac{3}{4}n) \leq \frac{n/2}{3/4\, n} = \frac{2}{3}$

→ not so useful here!

3) Chebychev

→ $Var(X_j) = p(1-p) = \frac{1}{4}$

→ $Var(X) = \frac{n}{4}$

→ $Pr(X \geq \frac{3}{4}n) \leq Pr(|X-\mu| \geq \frac{n}{4}) \leq \frac{Var(X)}{\left(\frac{n}{4}\right)^2}$

$$\leq \frac{\frac{n}{4}}{\left(\frac{n}{4}\right)^2} \leq 4/n$$

Better... What about $Pr(|X-\mu| \geq \varepsilon n)$??

Chernoff Bound: $X_1, X_2, \ldots, X_n$ independent

$X_j \in \{0,1\}$

$\Pr(X_j) = p$

$\mu = E[X] = np$

1) General Version:

$\delta > 0$

$$\Pr\left[X \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

$0 < \delta < 1$:

$$\Pr\left(X \leq (1-\delta)\mu\right) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$

Ex: $\Pr(X \geq e\mu) \leq \left(\frac{e^{e-1}}{e^{e}}\right)^{\mu} \leq e^{-\mu}$

2) Simplified versions

$\underline{\delta > 0}$

$$\Pr(X \geq (1+\delta)\mu) \leq e^{-\left(\delta^2/2+\delta\right)\mu}$$

$$\Pr(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2} \leq e^{-\mu\delta^2/3}$$

$\underline{\delta \leq 1}$

$$\Pr(X \leq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}$$

These are what I memorize!

Coin flipping.

$$\Pr(X \geq \tfrac{3}{4}n) \leq \Pr\left(X \geq \left(1+\tfrac{1}{2}\right)\tfrac{n}{2}\right)$$
$$\leq \Pr(X \geq (1+1/2)\,\mu) \leq e^{-\frac{n}{2}\cdot\left(\frac{1}{2}\right)^2\cdot\frac{1}{3}}$$
$$\leq \frac{1}{e^{-n/24}}$$

Decreases exponentially in n!! Very unlikely...

Tighter?

$$\Pr(X \geq \mu + \lambda) = \Pr\left(X \geq \left(1 + \frac{\lambda}{\mu}\right)\mu\right)$$

$$\leq e^{-\frac{\mu}{3} \cdot \left(\frac{\lambda}{\mu}\right)^2}$$

$\lambda \leq \mu$:

$$\leq e^{\frac{-\lambda^2}{3\mu}}$$

Choose $\lambda = \sqrt{3c\mu \ln(n)} = \sqrt{\frac{3}{2} c\, n \ln(n)}$

$$\leq e^{-\frac{3c\mu \ln(n)}{3\mu}}$$

$$\leq e^{-c\ln(n)}$$

$$\leq \left(\frac{1}{n}\right)^c$$

Same idea: $\Pr(X \leq \mu - \lambda) \leq \left(\frac{1}{n}\right)^c$

Conclusion: With high probability (w.h.p.)
# of heads is $\frac{n}{2} \pm \Theta\left(\sqrt{n \log n}\right)$

$$\Pr(\text{fail}) \leq \Pr(X \geq \mu + \lambda) + \Pr(X \leq \mu - \lambda)$$
$$\leq \frac{1}{n^c} + \frac{1}{n^c} = 2/n^c$$

Proof of Chernoff Bound

① $$\Pr[X \geq (1+\delta)\mu] = \Pr[tX \geq t(1+\delta)\mu]$$

Assume $t > 0$

$$= \Pr\left[e^{tX} \geq e^{t(1+\delta)\mu}\right]$$

② Markov's Inequality

$$\Pr\left[e^{tX} \geq e^{t(1+\delta)X}\right] \leq \frac{E\left[e^{tX}\right]}{e^{t(1+\delta)X}}$$

③ Independence

$$E\left[e^{tX}\right] = E\left[e^{t\sum X_j}\right] = E\left[\prod_{j=1}^{n} e^{tX_j}\right]$$

independence $\longrightarrow$ $$= \prod_{j=1}^{n} E\left[e^{tX_j}\right]$$

④ Approximation

$$E[e^{tX_j}] = pe^{t} + (1-p)e^{0}$$

$$= 1 - p(1-e^{t}) \qquad X = p(1-e^{t})$$

$$\leq e^{-p(1-e^{t})} \qquad e^{-X} \geq 1 - X$$

⑤ Conclusion:

$$\prod_{j=1}^{n} E[e^{tX_j}] \leq \left(e^{-p(1-e^{t})}\right)^{n} \leq e^{-pn(1-e^{t})}$$

$$\leq e^{-\mu(1-e^{t})}$$

Choose $t = \ln(1+\delta)$ $\qquad \leq e^{\mu\delta}$

$$\Pr(X \geq (1+\delta)\mu) \leq \frac{e^{\mu\delta}}{e^{t(1+\delta)\mu}} \leq \frac{e^{\mu\delta}}{(1+\delta)^{(1+\delta)\mu}}$$

$$\leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

Variations:

① X_j not identical: $\Pr(X_j = 1) = p_j$

$\mu = E[X] = \sum p_j$

$$\prod_{j=1}^{n} E[e^{tX_j}] \leq \left(e^{-p(1-e^t)}\right)^n \leq e^{-pn(1-e^t)}$$

$$\leq \prod_{j=1}^{n} e^{-p_j(1-e^t)}$$

$$= e^{-\sum p_j(1-e^t)}$$

$$= e^{-\mu(1-e^t)}$$

OK ✓

② X_j not indicator: $X_j \in [0,1]$ ← bounded!

e.g. $X_j = 0.75$

$$E[e^{tX_j}] = pe^t + (1-p)e^0$$

Hoeffding's Lemma shows:

$$E[e^{ty}] \leq e^{tE[y]} e^{t^2/8}$$

So it still works...

③ X_j not in $[0,1]$: $X_j \in [0, B]$ ← $B \geq 1$

Define $y_j = \frac{X_j}{B}$, $Y = \sum y_j = \frac{1}{B} \sum X_j$

$$\mu_y = E[y_j] = \frac{1}{B} E[X_j] = \frac{1}{B} \mu_x$$

$$\Pr\left(\sum X_j \geq (1+\delta)\mu_x\right) \leq \Pr\left(B \sum y_j \geq (1+\delta) B \mu_y\right)$$

$$\leq \Pr\left(\sum y_j \leq (1+\delta) \mu_y\right)$$

$$\leq e^{-\mu_y \delta^2/3}$$

$$\leq e^{\frac{-\mu_x \delta^2}{3B^2}}$$ ← reduces exponent by B^2

OK if B is constant or small compared to μ_x

④ Two-sided error:

$$\Pr\left[|X-\mu| \geq \delta\mu\right] \leq \Pr(X \geq (1+\delta)\mu)$$
$$+ \quad \text{← union bound}$$
$$\Pr(X \leq (1-\delta)\mu)$$

$$\leq 2e^{-\mu\delta^2/3}$$

④ Absolute bound:

$$\Pr\left[\,|X-\mu| \ge \delta\,\right] \le 2e^{\frac{-2\delta^2}{n}}$$

poly factor of n
ok if $\delta^2 > n$

if $B>1$: $\Pr\left[\,|X-\mu| \ge \delta\,\right] \le 2e^{\frac{-2\delta^2}{nB^2}}$

Hoeffding Bound

Ex: Coin flipping
$\delta = \Theta(\sqrt{n \log n})$

⑤ Negative Correlation / Negative Association

if $E\left[\prod e^{tX_j}\right] \le \prod E\left[e^{tX_j}\right]$

then... ok!

Works for balls in bins...

Ex Polling / Sampling

What fraction of Singaporeans like basketball?

Assume population size n, pn like basketball
$0 \le p \le 1$

Goal: find E such that:

$$\Pr\left[\,|E - pn| \le \delta\,\right] \ge 1 - \varepsilon$$

"Confidence interval $\pm\delta$ w.p. $\ge 1-\varepsilon$"

"if $\delta = 0.02$ and $\varepsilon = 0.1$: $\pm 2\%$ with 90% confidence"

$[p-\delta,\ p+\delta]$ = confidence interval

Algo: Choose sample of S people.
Let $X_j = 1$ if person j likes basketball.
Return:

$$E = \left(\frac{1}{s}\right) \sum X_j$$

Analysis

$$\Pr\left[|E-p| \geq \delta\right] = \Pr\left[\,\left|\sum X_j - ps\right| \geq \delta s\right]$$

$$X = \sum X_j, \quad \mu = E[X] = ps$$

$$= \Pr\left[|X-\mu| \geq \left(\frac{\delta}{p}\right)\mu\right]$$

Chernoff:

$$(1) \leq 2e^{-\frac{(\delta/p)^2}{2+\delta/p}\cdot \mu}$$

$$\text{if } \delta < p$$

$$\leq 2e^{-\frac{\delta^2 s}{3p}}$$

$$\text{Choose } s \geq \frac{3p}{\delta^2} \ln\left(\frac{2}{\varepsilon}\right)$$

$$\leq \varepsilon$$

Need to know p??

$p \leq 1$

$$\Rightarrow s \geq \frac{3}{\delta^2}\ln\left(\frac{2}{\varepsilon}\right)$$

Sample size does not depend on n!!

→ Good for Singapore OR China!!

Hoeffding: ② $\leq 2e^{\frac{-2\delta^2 s^2}{s}}$

$\leq 2e^{-2\delta^2 s}$

Choose $s \geq \frac{1}{2\delta^2} \cdot \ln\left(\frac{2}{\varepsilon}\right)$

$\leq \varepsilon$

(OK if $\delta > p$?)

Load balancing: n jobs

K servers

Algo: assign jobs randomly to servers

[Already analyzed $n = K$]

Fix a server. $X_j = 1$ if job j is on server.

$$E[X_j] = Pr[X_j] = \frac{1}{K} \qquad E[X] = \frac{n}{K}$$

$\boxed{3 \text{ Cases}}$

$n \gg K$

$n > \Omega(K \log n)$

$n = \Theta(K)$

or: $Pr(X \geq (1 + \frac{\lambda K}{n}) \frac{n}{K})$

$\leq e^{-\frac{\lambda^2 K^2}{n^2} \cdot \frac{n}{K} \cdot \frac{1}{3}}$

$\leq e^{\frac{-\lambda^2 K}{3n}}$

$\lambda = \sqrt{3c \frac{n}{K} \log n}$

① $$Pr\left[X \geq \frac{n}{K} + \lambda\right] \leq 2e^{\frac{-2\lambda^2}{n}}$$

$$\lambda \leq \sqrt{\frac{cn \log n}{2}}$$

$$\leq \frac{2}{n^c}$$

With Prob $\geq 1 - \frac{2}{n^{c-1}}$, no server has more than

$\frac{n}{K} + \sqrt{\frac{c}{2} n \log n}$ tasks

② $\Pr\left[X \geq e\frac{n}{K}\right] \leq e^{-\frac{n}{K}}$

if $n > cK\log n$

$$\leq e^{-c\log n} \leq \frac{1}{n^c}$$

With Prob $\geq 1 - \frac{1}{n^{c-1}}$, no server has more than $\frac{en}{K}$ jobs

③ $\Pr\left(X \geq \log(n)\frac{n}{K}\right) \leq e^{-\frac{(\log n - 1)^2 \cdot \frac{n}{K}}{1+\log n}}$

$$\leq e^{-\frac{\frac{1}{2}\log^2 n \cdot \frac{n}{K}}{2\log n}}$$

$$\leq e^{-\frac{n\log n}{4K}}$$

$$\leq \left(\frac{1}{n}\right)^{-\frac{n}{4K}}$$

If $n > 4cK$, no server has more than $\log(n)$ jobs w.p. $\geq 1 - \frac{1}{n^{c-1}}$

Coloring a Bipartite Graph

$G = (U, V, E)$, bipartite

Color each node in V red or blue to

minimize: $\forall u \in U$, $|\underset{\substack{\uparrow \\ \text{red neighbors}}}{R(u)} - \underset{\substack{\uparrow \\ \text{blue neighbors}}}{B(u)}|$

<u>OR</u>: $\geq \left(\frac{1}{2} - \delta\right)$ nbrs of u blue

$\geq \left(\frac{1}{2} - \delta\right)$ nbrs of u red

Fix u: let $K = \deg(u)$

$Y_j = 1$ if j^{th} neighbor is blue

$R(u) = K - B(u)$

$$|R(u) - B(u)| = 2\left|\frac{K}{2} - \sum Y_j\right|$$

$$\Pr\left(\left|\sum Y_j - \frac{K}{2}\right| \geq \lambda\right) \leq 2e^{-\frac{2\lambda^2}{K}}$$

$$\lambda = \sqrt{\frac{c}{2} K \ln(n)}$$

$$\leq 2 \cdot \frac{1}{n^c}$$

$\Rightarrow$ (Union bound over U) $\Pr(\text{any } u \text{ fails}) \leq \frac{2}{n^{c-1}} \leq \frac{1}{n^{c-2}}$

Eg, if $\frac{K}{4} \geq \sqrt{\frac{c}{2} K \ln(n)} \Rightarrow K^2 \geq 8 c K \ln(n)$

$K \geq 8 c \ln(n)$

then $\geq \frac{1}{4}$ nbrs red

$\geq \frac{1}{4}$ nbrs blue

W.h.p: $\Theta(K \log n)$ gap

$\geq \left(\frac{1}{2} - \delta\right)$ red/blue nbrs if $K \geq \Theta(\log n)$

Other problems ① Find median of array
② Find average of array ??
} Similar ??

Random Graphs

Build $G=(V,E)$ as follows:

① $V=$ set of nodes, $E=\emptyset$

② For each $u \in V$: add $2c\log(n)$ random edges

$\longrightarrow (u,?)$

Claim: G is connected [Similar to pset]

Claim: G has diameter $O(\log n)$ w.h.p.

Analysis:

→ Principle of deferred decision

→ start at u → choose edges $(u,?)$

→ find nodes in N_1

→ Choose edges for N_1 ⟹ find nodes N_2

etc.

Notation:

Let $N_0 = \{u\}$ ⟵ start with <u>any</u> u

Let $N_1 =$ nbrs on edges chosen by N_0

Let $N_{j+1} =$ nbrs on edges chosen by $N_j \setminus \bigcup_{i=1}^{j} N_i$

How big is N_1? $2C\log(n)$

Now big is N_2?

→ Choose $|N_1| C\log n$ edges

→ Let $X_j = 1$ if edge hits new node

$= 0$ otherwise

→ $\Pr(X_j) = \frac{|N_1| + |N_0|}{n} \leq \frac{2c\log n}{n} \leq \frac{1}{2}$

$E[|N_2|] = E[\Sigma X_j] = |N_1| c\log n \cdot \frac{1}{2} = \mu \geq C\log n$

$\boxed{C > 16}$

$\Pr(|N_2| \leq (\frac{1}{2})\mu) \leq e^{-\mu(\frac{1}{2})^2 \cdot \frac{1}{2}} \leq e^{-2\log n}$

$\leq \frac{1}{n^2}$

$\Rightarrow$ w.p. $\geq 1 - \frac{1}{n^2}$, $|N_2| \geq \log(n) \cdot |N_1| \geq 2|N_2|$

For each node: divide edges selected into 2 parts:

Phase 1) $c\log n$

Phase 2) $c\log n$

$\Rightarrow$ Use Phase 1 edges to grow set

$\Rightarrow$ Use Phase 2 edges to reach the rest

Let $T_j = N_0 \cup N_1 \cup \cdots \cup N_{j-1}$ ← all nodes found in first $j-1$ steps

Phase 1) Expansion: $T_j \leq \frac{n}{2}$

Consider N_j : $T_j \leq \frac{n}{2}$

→ Choose edges for N_{j-1}

→ $\geq |N_{j-1}| c \cdot \log n$ edges

→ $X_i = 1$ if edge i hits new node (not in T_j)

→ $Pr(X_i) \geq 1 - \frac{T_j}{n} \geq \frac{1}{2}$

$$M = E[|N_j|] \geq E[\Sigma X_i] \geq |N_{j-1}| \frac{c}{2} \log n \geq 16 \log n$$

$$Pr\left[|N_j| \leq \frac{1}{2} M\right] \leq e^{-M(\frac{1}{4})(\frac{1}{2})} \leq e^{-2\log n} \leq \frac{1}{n^2}$$

$\Rightarrow$ Wp $\geq 1 - \frac{1}{n^2}$, $|N_j| \geq 2|N_{j-1}|$

[Why don't we condition on $|N_{j-1}|$??

$Pr[\text{any } N_j \text{ fails}] \leq \frac{\log n}{n^2}$

$\Rightarrow$ for $j^* = \log(n)$, either $T_{j^*} \geq \frac{n}{2}$ or $|N_{j^*}| \geq 2^{\log n} |N_0|$
$\geq n$

End of phase 1) $T_{j^*} \geq \frac{n}{2}$

Phase 2) Finish

Let E' be the set of edges chosen by nodes in T_{j^*}

$|E'| \geq \frac{n}{2} \cdot C \log n$

What is $Pr(V \text{ not hit by } E') \leq \left(1 - \frac{1}{n}\right)^{\frac{n}{2} C \log n}$
$\leq \frac{1}{n^3}$

$Pr(\text{any node in } V \text{ not hit}) \leq n \cdot \frac{1}{n^3} \leq \frac{1}{n^2}$

$Pr(\text{Phase 1 fail or Phase 2 fail}) \leq \frac{\log n}{n^2} + \frac{1}{n^2} \leq 1/n$

Question: Bound max degree of graph.