Plan

Logistics

Review

Distinct Elements

Games Wednesday 6pm

write up by yourself

Problem set due Friday at 5pm

Solutions → self-grade

Forms (after class)

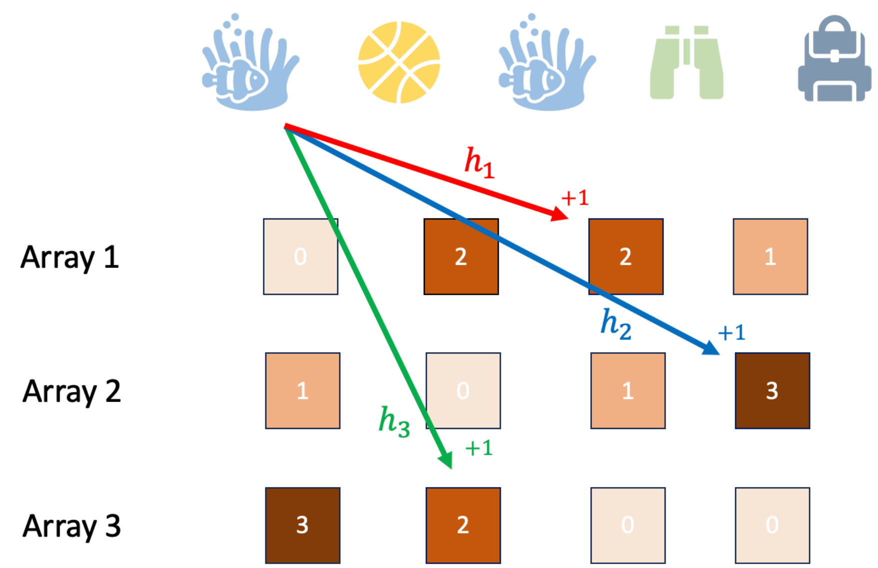
☐ Read written notes before and after

☐ Come talk to me!

e.g., Please explain X

☐ Work in groups, talk to people

Frequent Items



$$\tilde{f}(v) = \min_j A_j[h_j(v)]$$

$$f(v) \le \hat{f}(v) \le f(v) + \frac{2n}{\cancel{m}} \quad \frac{2c}{m}$$

wp $\cancel{1-\delta}$ 9/10

$$A_j[h_j(v)] = f(v) + \underbrace{\sum_{y \in U \setminus v} f(y)\,\mathbb{1}[h_j(v) = h_j(y)]}_{\text{error}}$$

$$\Pr(\text{error}_1 + \text{error}_2 \le \tfrac{2c}{m}) \ge \text{const}$$

$$= f(v) + \text{error}_1 + \text{error}_2$$

$$\Pr(\text{error}_1 + \text{error}_2 \ge \tfrac{2c}{m}) \le 1 - \text{const}$$

$$\Pr(\text{error for } j \ge \tfrac{2n}{m}) \le 1/2$$

(1) $\mathbb{E}[\text{error}] \le \frac{n}{m}$

(2) Apply markov's with $= \frac{2n}{m}$

$$\Pr(\text{error}_1 + \text{error}_2 \ge \tfrac{2c}{m} \text{ for all})$$

$$= \Pr(\text{error} \ge \tfrac{2c}{m} \text{ for } j)^t$$

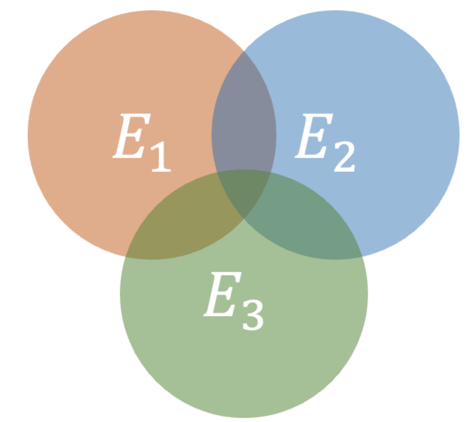
$$\le (1 - \text{const})^t = 1/10$$

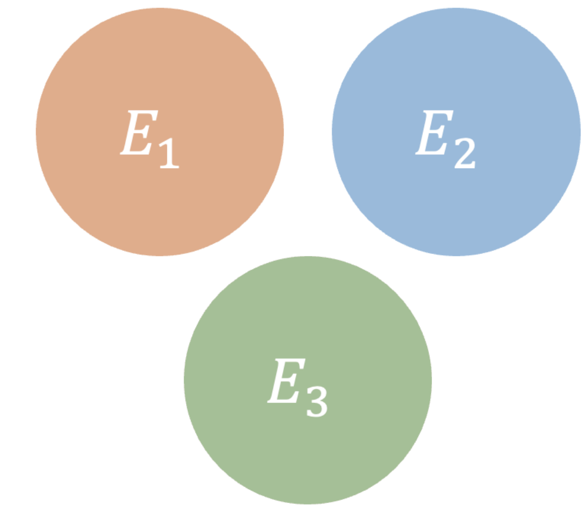
$O(m\cdot b) = O\left(\frac{2k}{\epsilon}\cdot \log(1/\delta)\right)$

Proved

$f(v) \le \hat{f}(v) \le f(v) + \frac{2n}{m} \iff$ "error is small"

wp $1-\delta$ for one item v





$\Pr(E_1 \cup \dots \cup E_n) \le \Pr(E_1) + \dots + \Pr(E_n)$

$\Pr($ error is small for all $v)$

$= 1 - \Pr($error is not small for some $v)$

$= 1 - \Pr($error $\ge \frac{2n}{m}$ for $v_1 \cup \dots$ error $\ge \frac{2n}{m}$ for $v_n)$

$\ge 1 - 1/10 = 9/10$

$\Pr($error $\ge\ \cup$ error $\ge \dots \cup$ error $\ge)$

$\le \Pr($error $\ge) + \Pr($error $\ge) + \dots +$

$\le \delta + \dots + \delta \le \delta \cdot n = 1/10$

$\delta = 1/10n$

Tools

- Markov's Inequality
- Linearity of Expectation
- Union Bound

+ Chebyshev's Inequality
+ Linearity of Variance

Markov's Inequality

X non-negative

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

Chebyshev's Inequality

all X, $\sigma^2 = \mathrm{Var}(X)$

$$\Pr(|X - \mathbb{E}[X]| \geq k \cdot \sigma) \leq \frac{1}{k^2}$$

- Markov's only for non-negative
- Chebyshev's requires variance
- two-sided bound from Chebyshev's

Chebyshev's : $\Pr(|X - \mathbb{E}[X]| \geq k \cdot \sigma) \leq \frac{1}{k^2}$

$$S = (X - \mathbb{E}[X])^2$$

$$\Pr(S \geq t) \leq \frac{\mathbb{E}[S]}{t}$$

$$\Pr((X - \mathbb{E}[X])^2 \geq t) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t} = \frac{Var(X)}{t} = \frac{\sigma^2}{t}$$

$t = k^2 \cdot \sigma^2$

$$\Pr((X - \mathbb{E}[X])^2 \geq k^2 \cdot \sigma^2) \leq \frac{\sigma^2}{k^2 \cdot \sigma^2} = \frac{1}{k^2}$$

$$\Pr(|X - \mathbb{E}[X]| \geq k \cdot \sigma) \leq \frac{1}{k^2}$$

Linearity of Variance

X_i indep X_j

For any pairwise independent r.v.s $X_1, \dots, X_n$

$$Var(X_1 + X_2 + \dots + X_n) = Var(X_1) + Var(X_2) + \dots + Var(X_n)$$

$$Var(X_1 + X_2 + X_3) = Var(X_1) + Var(X_2 + X_3) + 2\,Cov(X_1, X_2 + X_3)$$

$\mathbb{E}[X_1]$

$$\begin{aligned}
Cov(X_1, X_2 + X_3) &= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2 + X_3 - \mu_3)] \\
&= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2) + (X_1 - \mu_1)(X_3 - \mu_3)] \\
&= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] + \mathbb{E}[(X_1 - \mu_1)(X_3 - \mu_3)] \\
&= Cov(X_1, X_2) + Cov(X_1, X_3)
\end{aligned}$$

$C_1, C_2, \ldots, C_{100}$ $\qquad C_i = \begin{cases} 1 & \text{wp } 1/2 \\ 0 & \text{wp } 1/2 \end{cases}$

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

$$E[H] = E\sum_{i=1}^{100} C_i = \sum_{i=1}^{100} E[C_i] = 100 \cdot 1/2 = 50$$

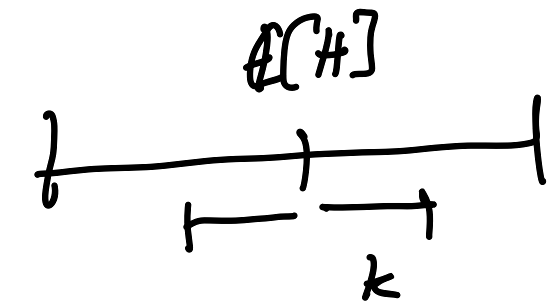
$$Var(H) = Var \sum_{i=1}^{100} C_i = \sum_{i=1}^{100} Var(C_i) = 100 \cdot 1/4 = 25$$
$$= \sigma^2$$
$$\sigma = 5$$

$$Var(c_i) = E[C_i^2] - E[C_i]^2$$
$$= (1^2 \cdot 1/2 + 0^2 \cdot 1/2) - (1/2)^2$$
$$= \frac{1}{2} - 1/4 = 1/4$$

Markov's

$$Pr(H \geq 70) \leq \frac{E[H]}{70} = \frac{50}{70} = 5/7$$

Chebyshov's



$$Pr(|H - E[H]| \geq k \cdot \sigma) \leq \frac{1}{k^2}$$

$k = 4$

$$Pr(|H - 50| \geq 4 \cdot 5) \leq \frac{1}{4^2} = 1/16$$

$$Pr(|H - 50| \geq 20) \leq 1/16$$

$$Pr(H \geq 70 \cup H \leq 30) \leq 1/16$$

Distinct Elements

$x_1, \ldots, x_n$

e.g., 1, 10, 2, 4, 9, 10, 2, 4

$D = \#\text{ distinct} = 5$

Distinct

↳ users

↳ values

↳ queries

↳ DNA motifs

Naive dictionary uses $O(D)$

Our goal return

estimate $\tilde{D} \approx D$

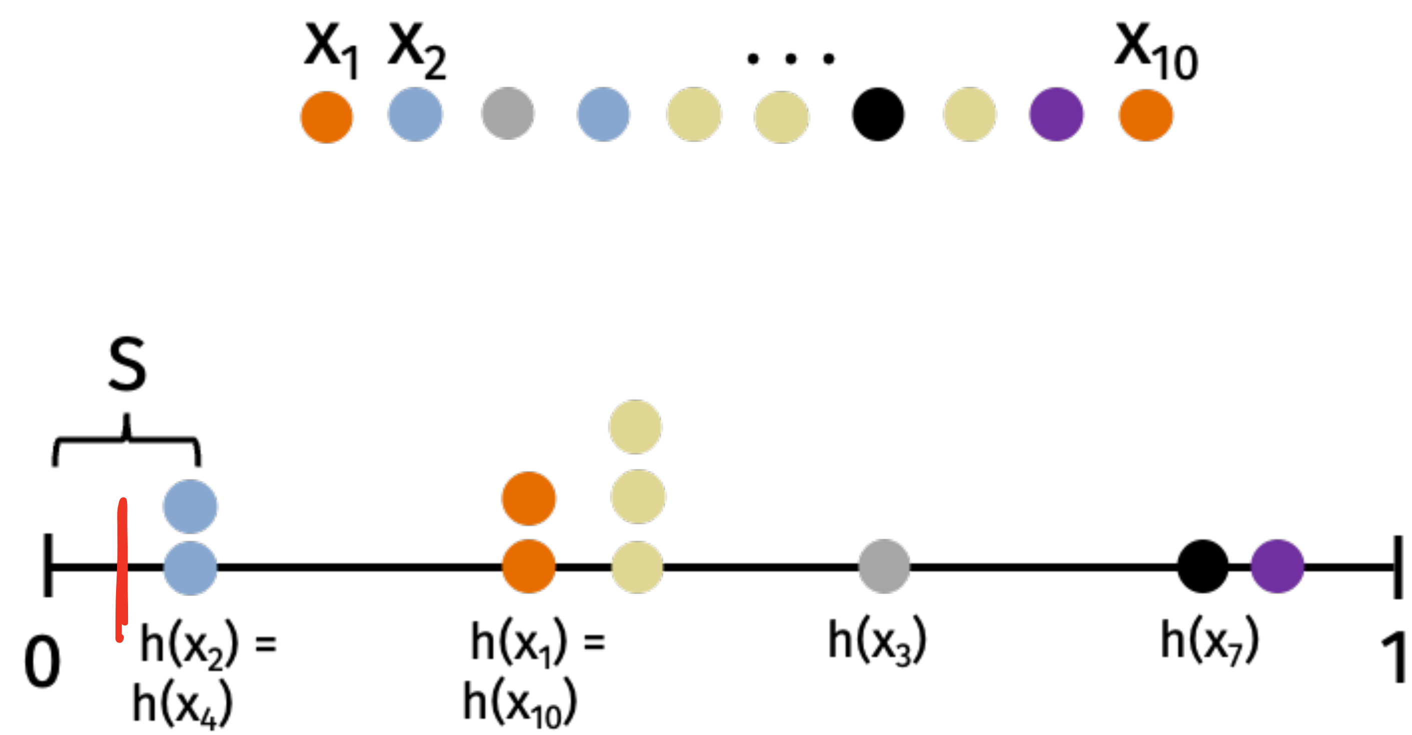
$1 > \epsilon > 0$

$(1-\epsilon) D \leq \tilde{D} \leq (1+\epsilon) D$ wp $1-\delta$

using $O\left(\frac{1}{\epsilon^2 \cdot \delta} \cdot \log D\right)$

Min Hash

Choose $h: \mathcal{U} \to [0, 1]$



$S = 1$

For every $i \in \{1, \dots, n\}$

$$S = \min(S, h(x_i))$$

Return $\tilde{D} = \frac{1}{S} - 1$

$$\Pr(S \geq \Delta) = (1 - \Delta)^D$$

Intuition: More distinct items, S is smaller

(1) Show $\mathbb{E}[X] = \sum_{x=1}^{\infty} Pr(X \geq x)$

X is integer valued, non-negative r.v.

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \cdot Pr(X = x)$$

(2) $\mathbb{E}[X] = \int_0^{\infty} Pr(X \geq x)\, dx$

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \cdot \Pr(X=x)$$

$$= 0 \cdot \Pr(X=0) + 1 \cdot \Pr(X=1) + 2 \cdot \Pr(X=2) + 3 \cdot \Pr(X=3) + \dots$$

$$\begin{array}{llll} = \Pr(X=1) + & \Pr(X=2) & + \Pr(X=3) & = \Pr(X \geq 1) \\ & + \Pr(X=2) & + \Pr(X=3) & + \Pr(X \geq 2) \\ & & + \Pr(X=3) & + \Pr(X \geq 3 \\ & & & + \\ & & & \vdots \end{array}$$

$$= \sum_{x=1}^{\infty} \Pr(X \geq x)$$

$$X \overset{\swarrow \text{continuous}}{=} X - 0 = x\Big]_{x=0}^{X}$$

$$= \int_0^X dx = \int_0^\infty \mathbb{1}[X \geq x]\, dx$$

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{1}[X \geq x]\, dx\right]$$

$$= \int_0^\infty \mathbb{E}[\mathbb{1}[X \geq x]]\, dx$$

$$= \int_0^\infty \Pr(X \geq x)\, dx$$

expectation of indicator is probability

$$\mathbb{E}[S] = \int_{s=0}^{1} \Pr(S \geq s)\, ds$$

$$= \int_{s=0}^{1} (1-s)^D\, ds$$

$$= \frac{-(1-s)^{D+1}}{D+1}\Bigg]_{s=0}^{1} = \frac{1}{D+1}$$

$$\mathbb{E}[S^2] = \int_{s=0}^{1} \Pr(S^2 \geq s)\, ds$$

$$= \int_{s=0}^{1} \Pr(S \geq \sqrt{s})\, ds$$

$$= \int_{s=0}^{1} (1 - \sqrt{s})^D\, ds$$

$$\overset{\text{wolfram}}{=} \frac{2}{(D+1)(D+2)}$$

$$\mathbb{E}[S] = \frac{1}{D+1} \overset{*}{=} \mu$$

$$0 < \epsilon < 1$$

$$Var(S) = \mathbb{E}[S^2] - \mathbb{E}[S]^2 = \sigma^2$$

$$Var(S) \overset{\Delta}{=} \sigma^2$$

$$\mathbb{E}[S] \overset{\Delta}{=} \mu$$

$$= \frac{2}{(D+1)(D+2)} - \frac{1}{(D+1)^2}$$

$$\leq \frac{2}{(D+1)(D+1)} - \frac{1}{(D+1)^2} = \frac{1}{(D+1)^2} \overset{*}{=} \mu^2$$

$$\mu^2 = \left(\frac{1}{D+1}\right)^2 = \frac{1}{(D+1)^2} \geq \sigma^2$$

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

$$k = \epsilon$$

$$Pr(|S-\mu| \geq \epsilon\mu) \leq \frac{1}{\epsilon^2}$$

$$1/\epsilon^2 \geq 1$$

$$1 \geq \epsilon^2 \iff 1 \geq \epsilon$$

Variance Reduction

Repeat core subroutine

Choose k hash functions

$h_1, \ldots, h_k : \mathcal{U} \to [0,1]$

For every i

For every j

$$S_j = \min(S_j, h_j(x_i))$$

$$S = \frac{S_1 + \ldots + S_k}{k}$$

$$\tilde{D} = \frac{1}{S} - 1$$

$$\mathbb{E}[S] = \mathbb{E}\left[\frac{1}{k} \sum_{j=1}^{k} S_j\right]$$

$$= \frac{1}{k} \sum_{j=1}^{k} \mathbb{E}[S_j] = \frac{1}{k} \cdot k \cdot \frac{1}{D+1}$$

$$= \frac{1}{D+1} = \mu$$

$$\mathrm{Var}(S) = \mathrm{Var}\left(\frac{1}{k} \sum_{j=1}^{k} S_j\right)$$

$$= \frac{1}{k^2} \mathrm{Var}\left(\sum_{j=1}^{k} S_j\right)$$

$$= \frac{1}{k^2} \sum_{j=1}^{k} \mathrm{Var}(S_j)$$

$$\leq \frac{1}{k^2} \cdot k \cdot \mu^2 = \frac{\mu^2}{k}$$