Plan

Logistics

Review

Frequent Items

Wednesday 6pm Games!

Warner 210

Projects: 1 or 2 ppl,

Check out topics on home page

Problem set due Friday 5pm

↳ overleaf

↳ colab

→ combine into 1 pdf

Advice: aim to finish day assigned

check in form after problem solving session

Recommend: read written notes day before class

Review

X random variable

$$X = \begin{cases} 1 & \text{wp } 1/6 \\ 2 & \text{wp } 1/6 \\ 3 & \vdots \\ 4 & \\ 5 & \\ 6 & \end{cases}$$

$$Pr(X = x) \overset{\text{dice}}{=} 1/6$$

$$\mathbb{E}[X] = \sum_x x \cdot Pr(X = x)$$

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

X, Y r.v. independent iff

$\forall\, x, y \quad Pr(X = x \mid Y = y) = Pr(X = x)$

$\iff Pr(X = x \cap Y = y) = Pr(X = x)\, Pr(Y = y)$

uniform samples

$$D_{i,j} = \begin{cases} 1 & \text{if } i^{th}, j^{th} \text{ same} \\ 0 & \text{else} \end{cases}$$

$D_{1,2}$, $D_{3,4}$ independant !

$D_{1,2}$, $D_{2,3}$ independent !

$D_{1,2}$, $D_{2,1}$ not independent

Facts

constant

$$\mathbb{E}[cX] = c\,\mathbb{E}[X]$$

$$\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$$

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$ always

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$ iff uncorrelated

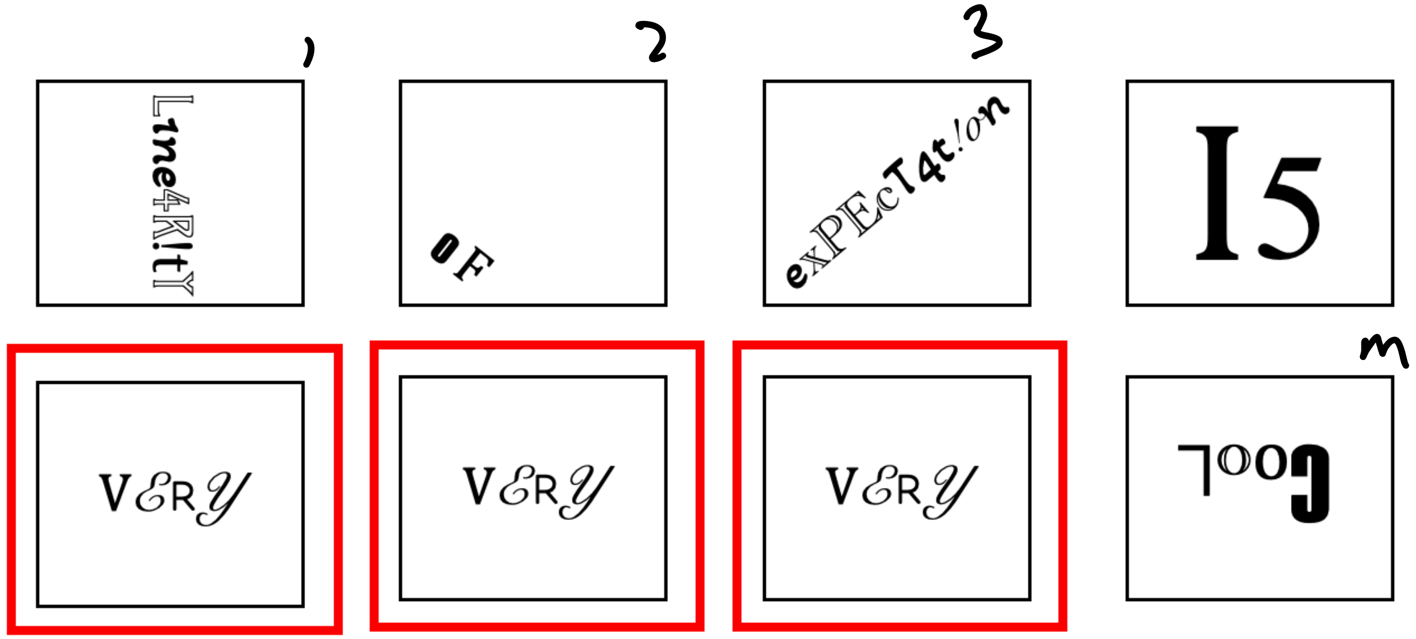
independent

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$ always

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$ iff uncorrelated

independence

Set Size Estimation



$$D = \sum_{i=1}^{m} \sum_{j=i+1}^{m} D_{i,j}$$ ← # samples

$$\mathbb{E}[D] = \frac{m(m-1)}{2 \cdot n}$$ ← set size

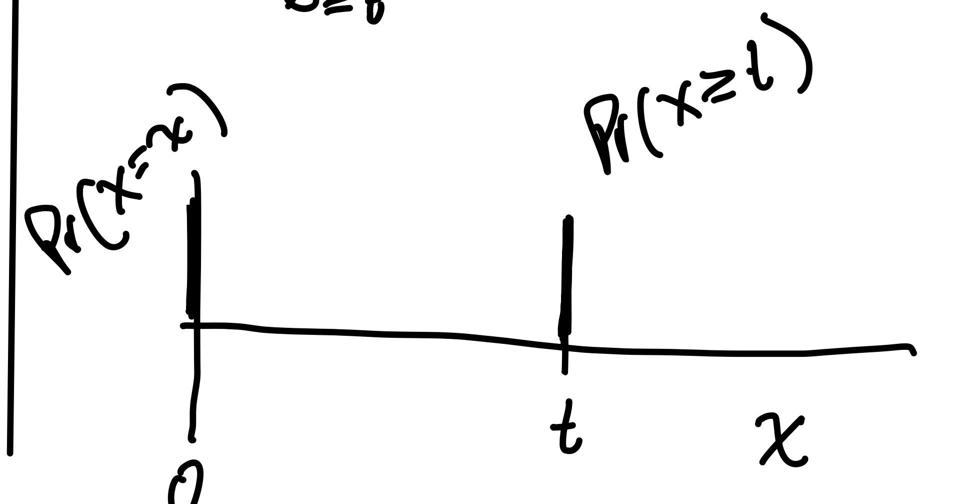
$$\boxed{\text{If we } \hat{D} \simeq \mathbb{E}[D], \text{ then } n \approx \frac{m(m-1)}{2\hat{D}}}$$

$$\boxed{\text{Markov's} \quad X \text{ non-negative}, \; t>0 \quad \Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}}$$

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

$$= \sum_{\substack{x \\ x \geq t}} x \Pr(X=x) + \sum_{\substack{x \\ x<t}} x \Pr(X=x)$$

$$\geq t \sum_{\substack{x \\ x \geq t}} \Pr(X=x) + 0 = t \Pr(X \geq t)$$



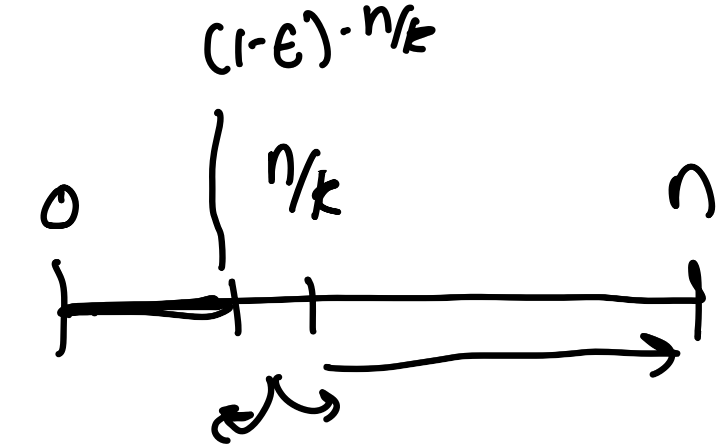
Frequent Items Estimation

$x_1, x_2, x_3, \ldots, x_n$ ← products, searches, videos

Which are most frequent?

Storing counts takes too much space (esp. if items are pairs)

Params: k integer, $1 > \epsilon > 0$ error



Return:

(1) every item that appears $\frac{n}{k}$ times

(2) only items that appear at least $(1-\epsilon)\frac{n}{k}$

We'll estimate frequency:

how often item appears

$$f(v) = \sum_{i=1}^{n} \mathbb{1}[x_i = v]$$

$\begin{cases} 1 \text{ if } x_i = v \\ 0 \text{ else} \end{cases}$

We'll return estimate $\hat{f}(v)$

$$f(v) \leq \tilde{f}(v) \leq f(v) + \frac{\epsilon}{k} \cdot n$$

one-sided error

with probability 9/10

Then return v s.t. $\tilde{f}(v) \geq \frac{n}{k}$

such that

(1) If $f(v) \geq \frac{n}{k}$,

$$\tilde{f}(v) \geq f(v) \geq \frac{n}{k}$$

(2) $\frac{n}{k} \leq \tilde{f}(v) \leq f(v) + \frac{\epsilon}{k} \cdot n$

$\Rightarrow \frac{n}{k} \leq f(v) + \frac{\epsilon}{k} \cdot n$

$\frac{n}{k} - \frac{\epsilon}{k} \cdot n \leq f(v)$

$\Rightarrow (1-\epsilon)\frac{n}{k} \leq f(v)$

Hash Functions!

$h: \mathcal{U} \to \{1, \dots, m\}$ ← consistently maps to a random value

- $\Pr(h(x) = i) = \frac{1}{m}$

- $h(x), h(y)$ independent r.v.

$\Rightarrow \Pr(h(x) = h(y)) \leq \frac{1}{m}$

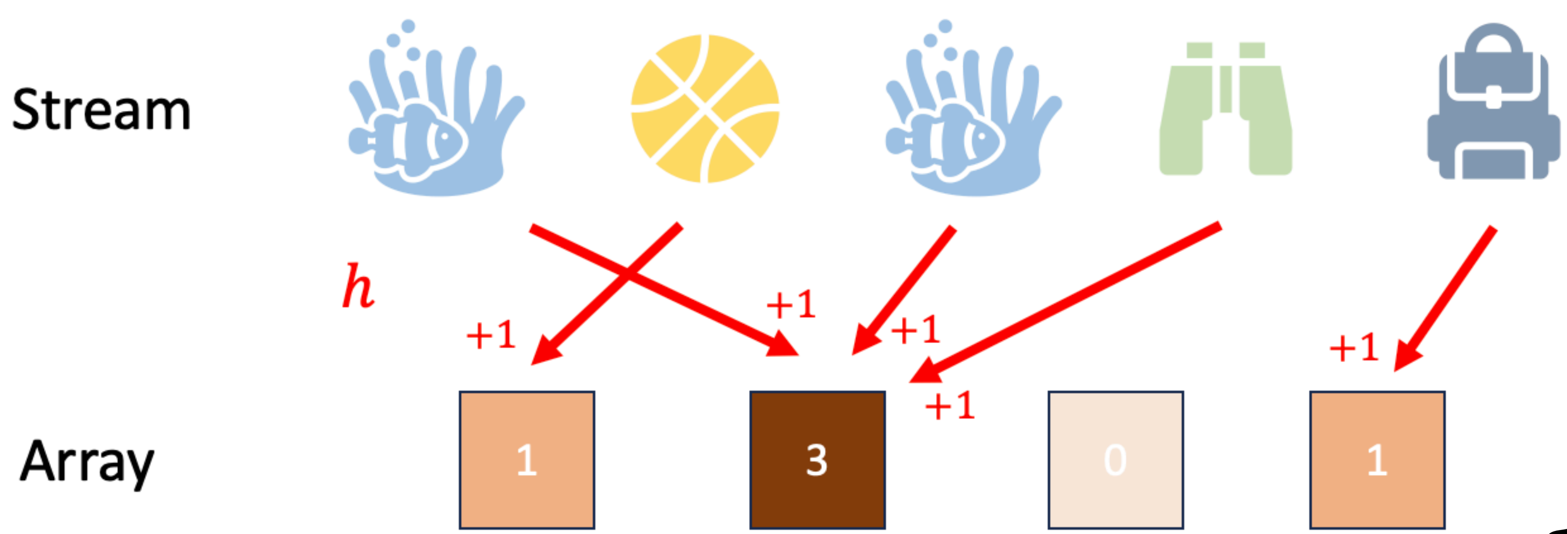
Count-Min Sketch

Choose h hash-function

Initialize m-length array A

For every x_i,

$$A[h(x_i)] = A[h(x_i)] + 1$$



1 3 0 1

$m = 4$

$$\tilde{f}(v) = A[h(v)] \geq f(v)$$

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

all other items

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

(2) $\Pr(\text{error} \geq t) \leq 1/2$

what is t?

$$\mathbb{E}\left[\sum_{y \in U \setminus v} f(y) \mathbb{1}[h(y)=h(v)] \right]$$

$$= \sum_{y \in U \setminus v} \mathbb{E}\left[f(y) \mathbb{1}[h(y)=h(v)] \right]$$ by linearity of expectation

$$= \sum_{y \in U \setminus v} f(y)\, \mathbb{E}\left[\mathbb{1}[h(y)=h(v)] \right]$$

$$1 \cdot Pr(h(y)=h(v))$$
$$+ 0\ (1 - Pr(h(y)=h(v)))$$
$$= Pr(h(y)=h(v))$$
$$\leq \frac{1}{m}$$

$$= \sum_{y \in U \setminus v} f(y) \cdot \frac{1}{m}$$

$$= \frac{1}{m} \sum_{y \in U \setminus v} f(y)$$

$$\leq \frac{1}{m} \cdot n = \frac{n}{m}$$

Want: $\Pr(\text{error} \geq t) \leq 1/2$

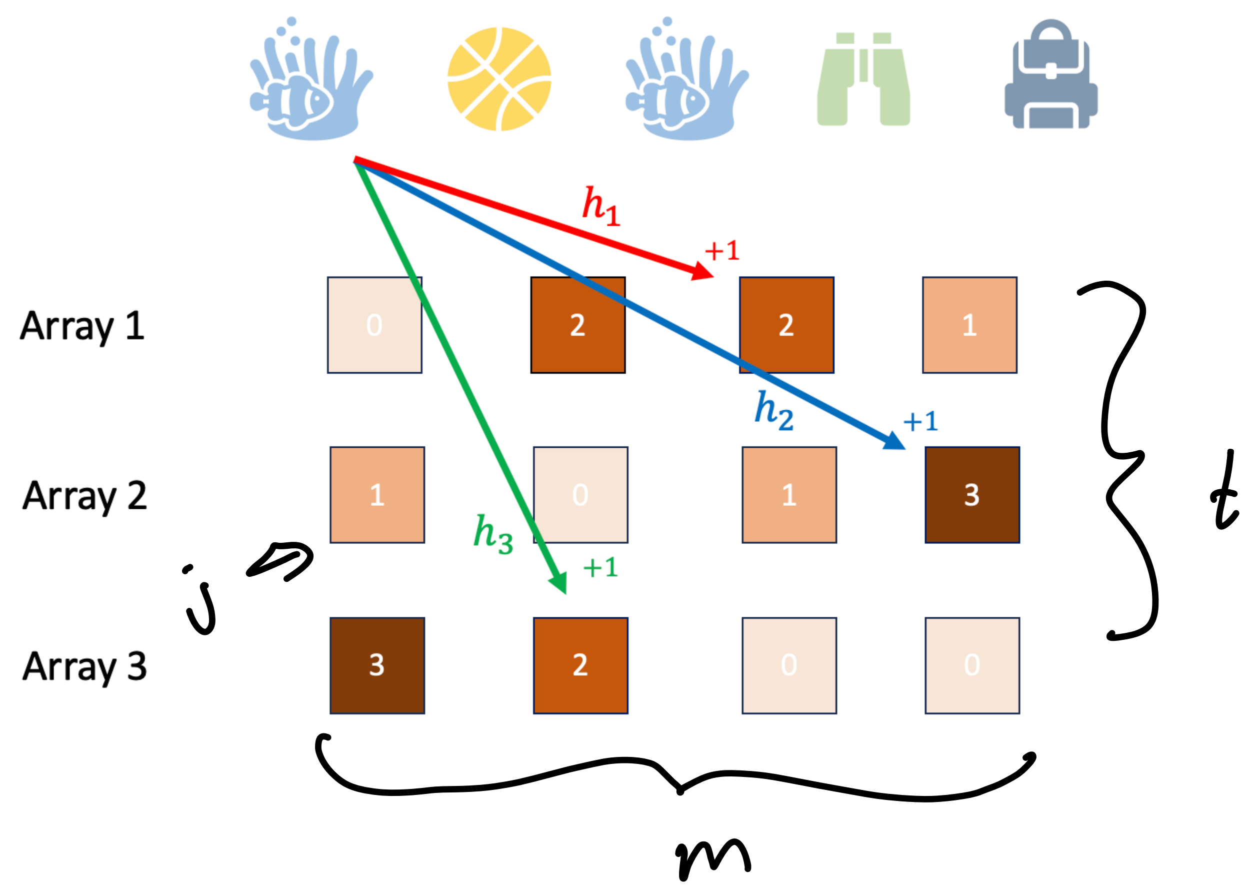
what is t?

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

$$\Pr(\text{error} \geq t) \leq \frac{\mathbb{E}[\text{error}]}{t} \leq \frac{n}{m} \cdot \frac{1}{t} = 1/2$$

$$t = \frac{2n}{m}$$

Boost with repetition!



$$\hat{f}(v) = \min \{ A[h_j(v)] : j \in \{1, \dots, t\} \}$$

$$\geq f(v)$$

$$\Pr\left(\text{error} \geq \frac{2n}{m}\right) \leq 1/2$$

$$m = \frac{2k}{\epsilon} \Rightarrow \Pr\left(\text{error} \geq \frac{n\epsilon}{k}\right) \leq 1/2$$

For every j wp $1/2$

$$f(v) \leq A[h_j(v)] \leq f(v) + \frac{\epsilon n}{k}$$

$$\Pr\left(\hat{f}(v) \geq f(v) + \frac{\epsilon n}{k}\right)$$

indep

$$= \prod_{j=1}^{t} \Pr\left(A[h_j(v)] \geq f(v) + \frac{\epsilon n}{k}\right)$$

$$\leq (1/2)^t = \delta$$

$$\frac{1}{2^t} = \delta \qquad \frac{1}{\delta} = 2^t$$

$$\log_2(1/\delta) = t$$

So we have $f(v) \leq \hat{f}(v) \leq f(v) + \frac{\epsilon n}{k}$ wp $1-\delta$

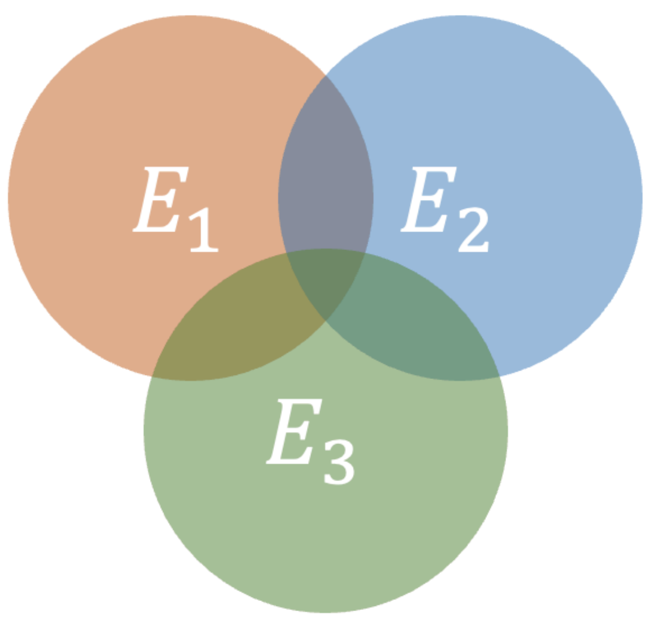
Space $O(mt) = O\left(\frac{k}{\epsilon} \cdot \log(1/\delta)\right)$

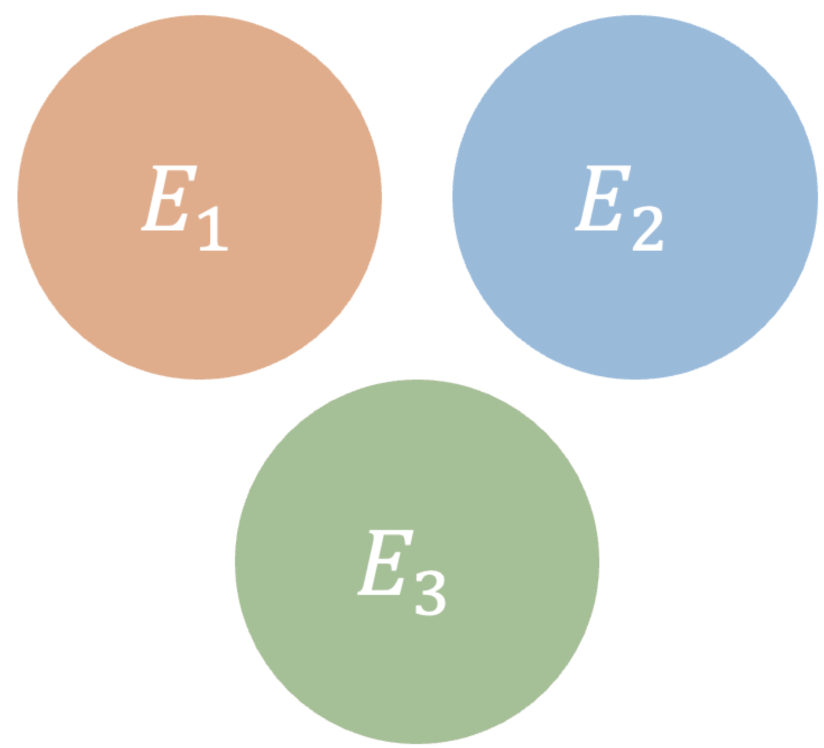
but only holds for one item v

Union Bound

Events $E_1, \dots, E_n$

$$\Pr(E_1 \cup E_2 \cup E_3 \dots \cup E_n) \le \Pr(E_1) + \Pr(E_2) + \dots + \Pr(E_n)$$





$$\Pr(\text{fail for } v_1 \cup \text{fail for } v_2 \cup \dots \cup \text{fail for } v_{|U|})$$

$$\overset{\text{union}}{\leq} \Pr(\text{fail for } v_1) + \Pr(\text{fail for } v_2) + \dots + \Pr(\text{fail for } v_{|U|})$$

$$\leq \delta + \delta + \dots + \delta = \delta \cdot |U| \leq \delta \cdot n \overset{\text{want}}{=} 1/10$$

$$\delta \overset{*}{=} 1/10n$$

$$O\left(\frac{k}{\epsilon} \log(1/\delta)\right) \overset{*}{=} O\left(\frac{k}{\epsilon} \log 10n\right)$$