Week 2 Tuesday

- Self-grade due Friday
- Thoughts on "working smart not hard"
  - ↳ focus is a muscle
    (fast dopamine weakens it)
  - ↳ even 15min is enough to start a hard task

Plan:

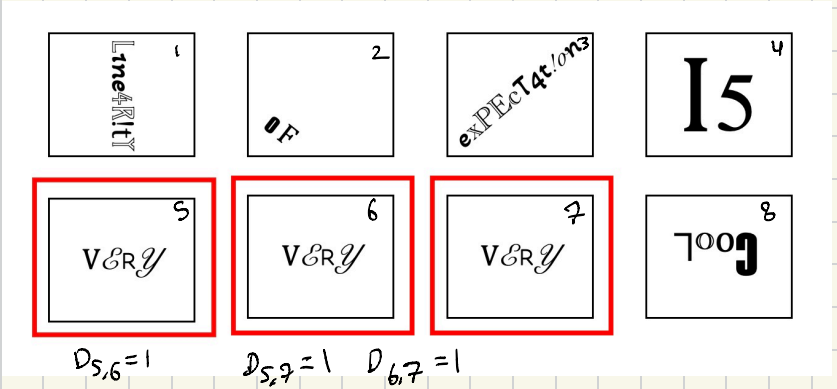
- Finish up set size estimation
  (Formalize "unlikely" event?)
- Start frequent items

Problem: Given query access, how big is a set?

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

$$D = \sum_{i,j=1 \,;\, i<j}^{m} D_{i,j}$$

$$\mathbb{E}[D] = \sum_{i<j} \mathbb{E}[D_{i,j}] = \binom{m}{2} \frac{1}{n} = \frac{m(m-1)}{2n}$$



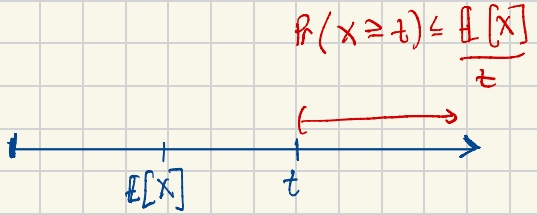
Suppose we made $m = 1{,}000$ queries and saw $D = 10$ duplicates.

How does this compare to what we expect?

Markov's Inequality

Theorem: For any non-negative rv X and $t > 0$,

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



Proof:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x \Pr(X = x) \\
&= \sum_{x:\, x \geq t} x \Pr(X = x) + \sum_{x:\, x < t} x \Pr(X = x) \\
&\geq \sum_{x:\, x \geq t} t \Pr(X = x) + 0 = t \Pr(X \geq t)
\end{aligned}$$

Answer to duplicate question:

Frequent Items

Problem: Most common elements in a stream?

Examples:
- most popular products on Amazon
- most watched videos on YouTube
- most searched queries on Google

A stream of n items $x_1, \dots, x_n$

U = set of all items

K positive integer and $\epsilon > 0$ small constant

Return:
- every item that appears at least $\frac{n}{K}$ times
- only items that appear at least $(1-\epsilon)\frac{n}{K}$ times

Naive Attempt: Store each item with its frequency

Is this a good solution?

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

↑ frequency of item v

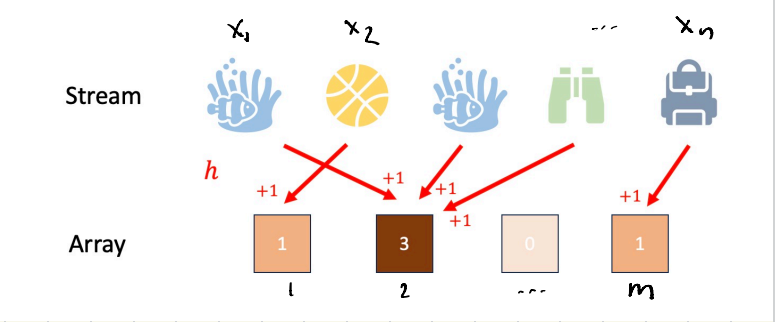
Our strategy:

Maintain estimates $\hat{f}(v)$ s.t.

$$f(v) \le \hat{f}(v) \le f(v) + \frac{\epsilon}{k} n$$

Question: How do we use $\vec{P}$ to solve frequent items problem?

Counting



For x_i in $x_1, \ldots, x_n$:

$$A[h(x_i)] \leftarrow A[h(x_i)] + 1$$

Return $\hat{f}(v) = A[h(v)]$

Week 2 Thursday

- self-grade due Friday
- Problem 2 and 3 due Monday

Plan

- Count-min for frequent items
- Union Bound

Markov's Inequality

For non-negative rv X, $t > 0$

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

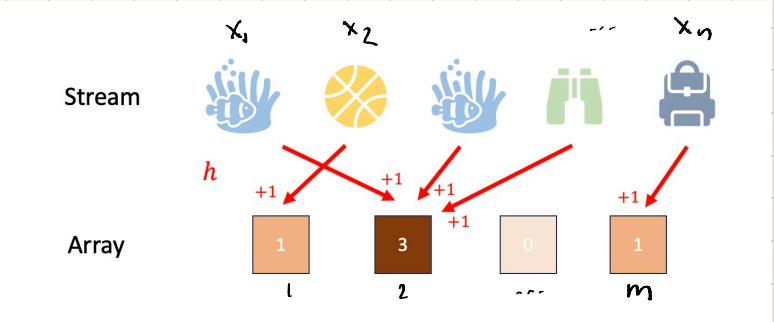
Frequent Items

Problem: Most common elements in a stream?

Examples:

- most popular products on Amazon
- most watched videos on YouTube
- most searched queries on Google

Counting



For x_i in $x_1, \ldots, x_n$:

$$A[h(x_i)] \leftarrow A[h(x_i)] + 1$$

Return $\hat{f}(v) = A[h(v)]$

Key fact: $\Pr(h(v) = h(v')) \leq \frac{1}{m}$

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

↑ true

↑ error

Show that:

$$\Pr\left(\sum_{y \in U \setminus \{v\}} f(y)\, \mathbb{1}[h(y) = h(v)] \geq \frac{2n}{m} \right) \leq \frac{1}{2}$$

With one array,

$$f(v) \leq A[h(v)] \leq f(v) + \frac{2n}{m}$$

wp $\frac{1}{2}$ for any v

Setting $m = \frac{2k}{\epsilon}$,

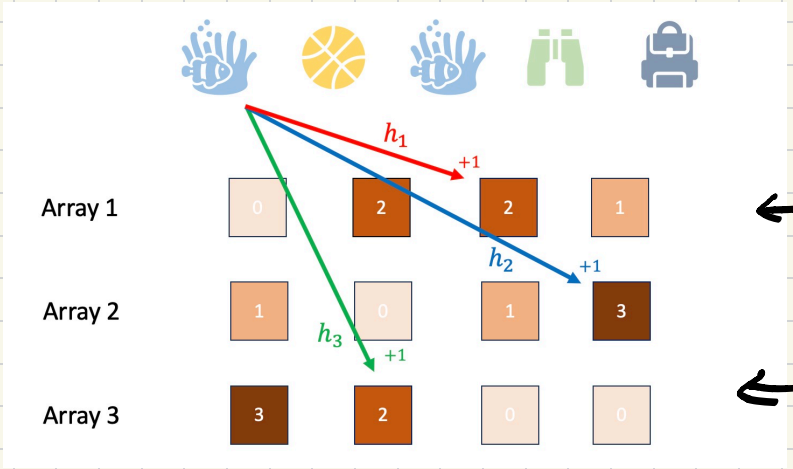
$$f(v) \leq A[h(v)] \leq f(v) + \frac{n}{k}\epsilon$$

wp $1/2$ for any v

↑ a little low...

Strategy: Boost success by repeating subroutine!

Count-Min



$\leftarrow f(v) \le A_1[h_1(v)] \le f(v) + \epsilon\frac{n}{k}$ w.p. $1/2$

$\vdots$

$\leftarrow f(v) \le A_t[h_t(v)] \le f(v) + \epsilon\frac{n}{k}$ w.p. $1/2$

For x_i in $x_1, \ldots, x_n$:

For A_j in $A_1, \ldots, A_t$:

$$A_j[h_j(x_i)] \leftarrow A[h_j(x_i)] + 1$$

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

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

$$\le \prod_{i=1}^{t} Pr\left(A_i[h_i(v)] \ge f(v) + \frac{\epsilon n}{k}\right)$$

$$= \left(\frac{1}{2}\right)^t$$

Count-Min Guarantee

n - length stream

$\epsilon > 0$ error

k integer

$m = \frac{2k}{\epsilon}$ size arrays

Space: $O(mt) = O(\frac{k}{\epsilon} \log_2 1/\delta)$

$\Pr(\hat{f}(v) < f(v)) = 0$

$\Pr(\hat{f}(v) > f(v) + \frac{n}{k}\epsilon) = \frac{1}{2}^t$

... but this only holds for one v

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

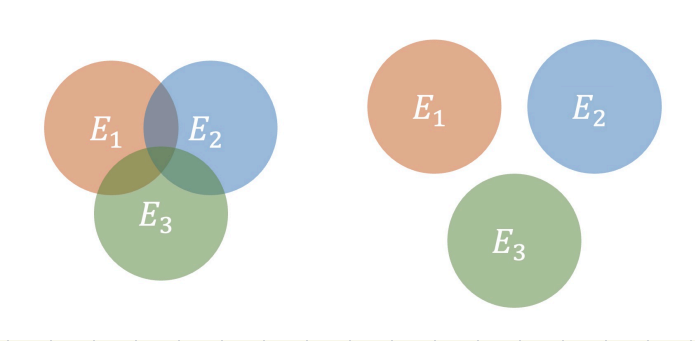
$2^t = 1/\delta$

$t = \log_2 2^t = \log_2 1/\delta$

want $= \delta$ ← failure probability

Union Bound

$$\Pr(E_1 \cup \ldots \cup E_n) \leq \Pr(E_1) + \ldots + \Pr(E_n)$$



Apply union bound to

$$\Pr\left(\hat{f}(v_1) > f(v_1) + \epsilon\frac{n}{k} \cup \ldots \hat{f}(v_n) > f(v_n) + \epsilon\frac{n}{k}\right) \leq$$