23 Feb 2026 Distinct Elements Problem

Announcements:

① Weds 2/25: Eshan C guest lecturer

② Weds quiz covers 2/11, 2/18, 2/23

Streaming model of computation

An algorithm processes a sequence of

m tokens

from a universe of

N potential tokens $N := 2^b$

token $\in \{0,1\}^b$

using storage space $O(\text{poly}(b + \log m))$

and must afterward report an estimate of some function of the token sequence.

Ex.
- most frequently occurring token
- # distinct tokens
- entropy of token distrib
- ... et cetera

An (ε, δ)-PAC algorithm satisfies the property

$$\Pr\left(\left|\text{ALG's answer} - \text{TRUTH}\right| > \varepsilon \cdot \text{TRUTH}\right) \leq \delta.$$

Estimating distinct elements

Suppose we have stream

$$\theta_1, \theta_2, \ldots, \theta_m \in \{0,1\}^b$$

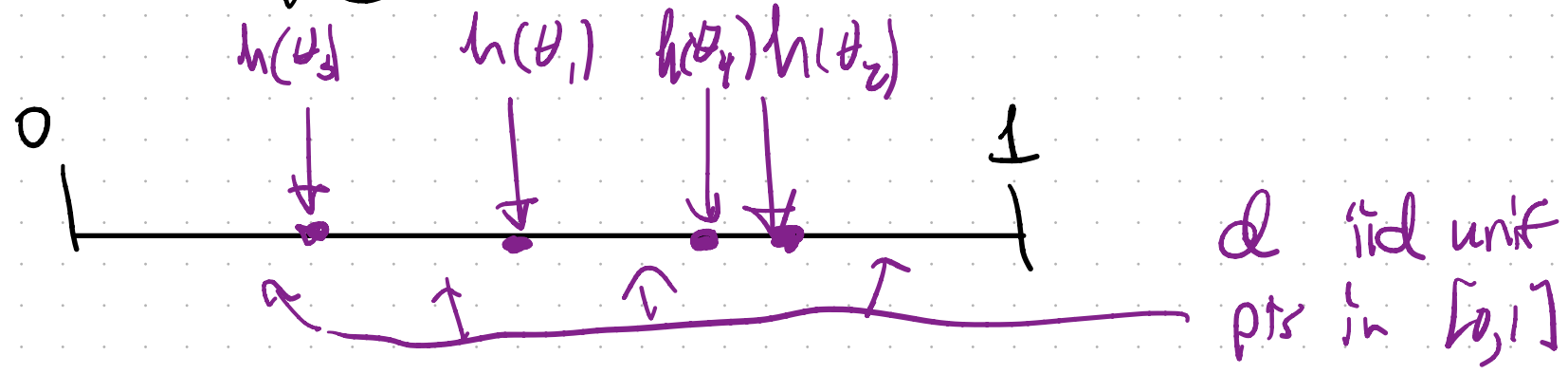
and (unrealistic) we have mutually independent random

$$h(\theta_1), h(\theta_2), \ldots, h(\theta_m) \in [0,1]$$

if $\theta_i == \theta_j$ $h(\theta_i) = h(\theta_j)$

else $h(\theta_i)$ indep of $h(\theta_j)$.

If there are d distinct tokens in the stream, what would the set of distinct hash vals look like?



For d iid unif pts in $[0,1]$

$$\mathbb{E}[\min] = \frac{1}{d+1}.$$

Distinct Elements Estimator (part 1)

Initialize 2-uniform hash $h: [N] \to [p]$

(p large prime) ← $O(\log p)$ bits of storage

for each token θ_i, compute $h(\theta_i)$

and store $Z = \min\{h(\theta_1), \dots, h(\theta_m)\}$.

// memory requirement is only storing Z and the most recent hash $h(\theta_i)$.

// If d distinct tokens,

expect $Z \approx \frac{p}{d+1}$.

After seeing entire stream,

report $\hat{d} = \frac{p}{Z}$.

Claim. If $p \gg d$

$$\Pr\left(\hat{d} \notin \left[\frac{1}{6}d,\ 6d\right]\right) < \frac{1}{3}.$$

Proof. Two bad events

$$\mathcal{E}_1: \quad \hat{d} < \frac{d}{6} \iff Z > \frac{6p}{d}.$$

$$\mathcal{E}_2: \quad \hat{d} > 6d \iff Z < \frac{p}{6d}.$$

$\Pr(\mathcal{E}_2)$: the event $Z < \frac{p}{6d}$ happens

when at least one of

$$\{h(\theta_1), \ldots, h(\theta_m)\} \in \left[1, \ldots, \left\lceil \frac{p}{6d} \right\rceil - 1\right].$$

Assume WLOG $\theta_1, \ldots, \theta_d$ are the d distinct stream elements.

$$\Pr\left(\{h(\theta_1), \ldots, h(\theta_m)\} \text{ intersects } \left\{1, \ldots, \frac{p}{6d} - 1\right\}\right)$$

$$\leq \mathbb{E}\left[\#\, i \in \{1, \ldots, d\} : h(\theta_i) \in \left\{1, \frac{p}{6d} - 1\right\}\right]$$

$$= \sum_{i=1}^{d} \Pr\left(h(\theta_i) < \frac{p}{6d}\right)$$

$$< \quad d \cdot \frac{1}{6d} \quad = \quad \frac{1}{6} .$$

$$\Pr(\mathcal{E}_2) < \frac{1}{6}$$

$\Pr(\mathcal{E}_1)$: $\mathcal{E}_1$ happens when

$\{h(\theta_1), h(\theta_2), \ldots, h(\theta_d)\}$ disjoint

from $\{1, 2, \ldots, \frac{6p}{d}\}$.

Step 1. $\Pr\left(h(\theta_1) \notin \{1, \ldots, \frac{6p}{d}\}\right) = \frac{p - 6p/d}{p}$

$$= 1 - \frac{6}{d}$$

$$\Pr\left(h(\theta_2) \notin \{1, \ldots, \frac{6p}{d}\}\right) = 1 - \frac{6}{d}$$

$$\vdots$$

$$\Pr\left(h(\theta_d) \notin \{1, \ldots, \frac{6p}{d}\}\right) = 1 - \frac{6}{d}$$

Assumes $h(\theta_1), \ldots, h(\theta_d)$ mutually indep.

$$\Pr\left(\{h(\theta_1), \ldots, h(\theta_d)\} \text{ disj from } \{1, \ldots, \frac{6p}{d}\}\right) = \left(1 - \frac{6}{d}\right)^d$$

$$< \left[e^{-6/d}\right] d \quad = \quad e^{-6}.$$

Lem. When $X_1, \dots, X_n$ are pairwise indep't,

$$\mathrm{Var}(X_1 + \dots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Proof. Let $\bar{X}_i = \mathbb{E}(X_i)$.

$$\bar{X} = \bar{X}_1 + \dots + \bar{X}_n$$

$$= \mathbb{E}[X_1 + \dots + X_n].$$

$$\mathrm{Var}(X_1 + \dots + X_n) = \mathbb{E}\left[\left((X_1 + \dots + X_n) - \bar{X}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{1 \le i < j \le n} X_i X_j - 2\sum X_i \bar{X} + \bar{X}^2\right]$$

$$\sum \mathrm{Var}(X_i) = \sum_{i=1}^{n} E\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)$$

Abandoning proof to avoid careless algebra error.

Theme of proof. $\mathrm{Var}(X_1 + \cdots + X_n)$ is $E[\text{quadratic polynomial}]$.

Each monomial depends on ≤ 2 variables, so can't detect whether they are fully indep. or just pairwise indep.

Applying to

$$\Pr(\mathcal{E}_1) = \Pr\left(\{h(\theta_1), \ldots, h(\theta_d)\} \text{ disj from } \{1, \ldots, \frac{6p}{d}\}\right)$$

$i = 1, \ldots, d$:

$$X_i = \begin{cases} 1 & \text{if } h(\theta_i) \in \{1, \ldots, \frac{6p}{d}\} \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbb{E}(X_i) = \frac{6}{d}$$

$$\mathrm{Var}(X_i) = \frac{6}{d}\left(1 - \frac{6}{d}\right) < \mathbb{E}[X_i].$$

$$\mathcal{E}_1 = \{X_1 + \ldots + X_d = 0\}.$$

Have: $E(X_1 + \dots + X_d) = \frac{6}{d} \cdot d = 6$

$$\mathrm{Var}(X_1 + \dots + X_d)$$

using pairwise indep. →

$$= \mathrm{Var}(X_1) + \dots + \mathrm{Var}(X_d)$$

$$\leq E(X_1) + \dots + E(X_d)$$

$$= 6$$

$$\Pr(X_1 + \dots + X_d = 0)$$

Chebyshev →

$$\leq \frac{\mathrm{Var}(X_1 + \dots + X_d)}{6^2}$$

$$\leq \frac{6}{6^2} = \frac{1}{6}.$$

$\Pr(\mathcal{E}_1), \Pr(\mathcal{E}_2) \leq 1/6$

$$\Pr(\mathcal{E}_1 \cup \mathcal{E}_2) \leq 1/3.$$

$$\Pr\left(\hat{d} \in \left[\frac{1}{6}d,\ 6d\right]\right) \geq \frac{2}{3}.$$

Reducing relative error to $1 \pm \varepsilon$.

Let $Z_1, Z_2, \ldots, Z_t$ be the t smallest hash values.

$$t = \left\lceil \frac{12}{\varepsilon^2 \delta} \right\rceil.$$

$$\hat{d} = \frac{t \cdot p}{Z_t}.$$

$Z_1 \approx \frac{p}{d+1}$

$Z_t \approx \frac{tp}{d+1}$

Analysis in §3.2