26 Feb 2025

More Streaming Algorithms

Flajolet – Martin Estimator for # Distinct Elements

Let $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$

Draw $h : \mathbb{F}_p \to \mathbb{F}_p$ unif random from a 2-universal hash family.

Calculate $Z = \min \{ h(x_i) \}$ — This calculation requires only $O(1)$ space!

Output the estimate $\frac{p}{Z+1}$.

<u>Last time:</u> If # distinct elements is d,

$$\Pr(\text{estimate is} > 6d) < \frac{1}{6}. \quad \text{[Markov's]}$$

<u>Still need to show:</u>

$$\Pr(\text{estimate} < \frac{d}{6}) < \frac{1}{6} \quad \text{[Chebyshev's]}.$$

We defined

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

$$Y = X_1 + \cdots + X_d$$

and observed that...

- estimate $< \frac{d}{6}$ when $Y = 0$.
- $\mathbb{E}[X_i] = \frac{6}{d}$ $\quad \mathrm{Var}(X_i) < \mathbb{E}[X_i]$
- $\mathbb{E}[Y] = 6 = \frac{6}{d} \cdot d$

• $\underbrace{\mathrm{Var}(Y) = \sum_{i=1}^{d} \mathrm{Var}(X_i)}_{\text{The variance of the sum of pairwise indep rand vars is the sum of their variances.}} < \sum_{i=1}^{d} E[X_i] = 6.$

$$\Pr(Y=0) \le \Pr(|Y-EY| \ge 6)$$

$$= \Pr\left((Y-EY)^2 \ge 36\right)$$

$$\overset{\text{(Markov's)}}{<} \frac{1}{6}.$$

Chebyshev's Ineq

$\therefore$ so $\Pr\left(\frac{d}{6} \le \text{estimate} \le 6d\right) > \frac{2}{3}$.

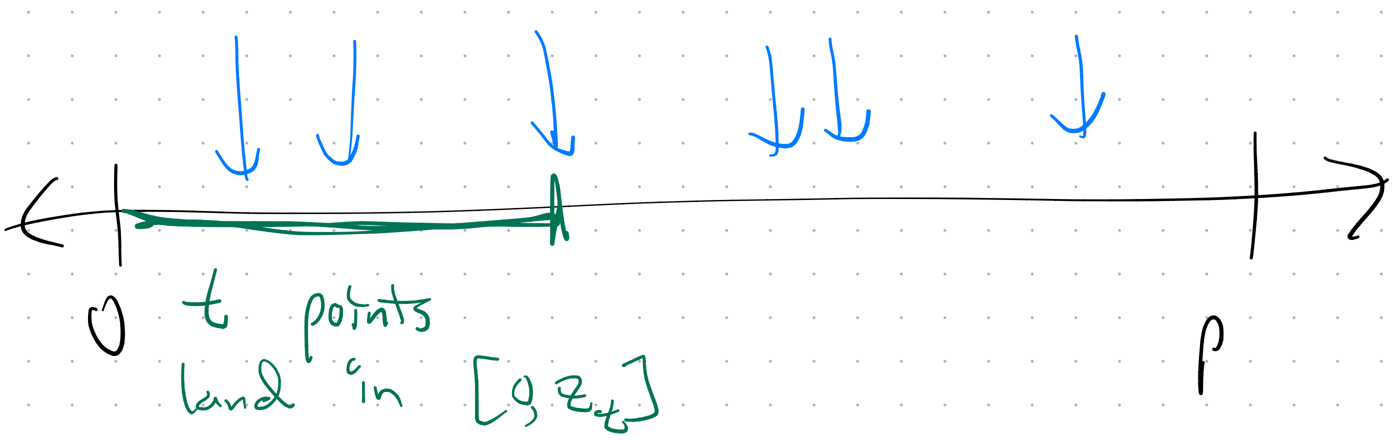
To improve accuracy, let $t > 1$.

Let $z_1 < z_2 < \dots < z_t$ be the t smallest distinct values among $\{h(x_1), \dots, h(x_n)\}$.

Estimate is $\dfrac{p \cdot t}{z_t}$

hash range

estimated density of hash values.



Recap: Assume $x_1, \ldots, x_d$ are the d distinct elements.

Again let d = # distinct elements.

$$X_i = \begin{cases} 1 & \text{if } h(x_i) \leq \frac{(1+\varepsilon)tp}{d} \\ 0 & \text{otherwise} \end{cases}$$

$$Y = X_1 + \cdots + X_d$$

Observe $Y < t$ preceisely when

$$Z_t > \frac{(1+\varepsilon)tp}{d} \iff \frac{pt}{Z_t} < \frac{d}{1+\varepsilon}.$$

$$\mathbb{E}[X_i] = \frac{(1+\varepsilon)t}{d}$$

$$\mathbb{E}[Y] = d \cdot \mathbb{E}[X_i] = (1+\varepsilon)t.$$

$$\mathrm{Var}(Y) = d \cdot \mathrm{Var}(X_i) < d \cdot \mathbb{E}[X_i] = (1+\varepsilon)t.$$

$$\Pr\left((Y - \mathbb{E}Y)^2 > \varepsilon^2 t^2\right) < \frac{(1+\varepsilon)t}{\varepsilon^2 t^2}$$

Can make RHS as large as we want by making t large.

To make $\Pr\left(\text{estimate} < \frac{d}{1+\varepsilon}\right) < \frac{\delta}{2}$,

solve $\frac{1+\varepsilon}{\varepsilon^2} \cdot \frac{1}{t} < \frac{\delta}{2}$

$$t > \boxed{\frac{2(1+\varepsilon)}{\varepsilon^2 \delta}}$$

For this value of t, similar Chebyshev Ineq calculation $\Longrightarrow$

$$\Pr\left(\text{estimate} > \frac{d}{1-\varepsilon}\right) < \frac{\delta}{2}.$$

Misra-Gries Algorithm for finding Frequent Elements

- Processes a stream of n (not necessarily distinct) elements
- Uses $O(k)$ storage
- Reports a list of (at most) k elements that includes every element occurring $> \frac{n}{k+1}$ times in the stream.

... plus false positives.

Maintain list of pairs

(x^i, c^i)

element seen in stream or $\perp$

positive integer

"number of uncancelled occurrences of x^i in the stream"