11. Count distinct sketch

Wednesday, October 6, 2021 2:33 PM

Last time: How to construct random hash families

Today: Frequency moments & estimating F_0.

Consider a sequence $a_1, \dots, a_n \in [m]$, with n, m both large (stream)

$\forall s \in [m]$, let $f_s = |\{i \mid a_i = s\}|$, the frequency of s in stream.

Def. For $p \in \mathbb{N}$, the pth freq moment of the stream is

$$F_p = \sum_{s=1}^{m} f_s^p.$$

($0^0 = 0$)

Aside: $F_0 = |\text{uniq}\{a_1, \dots, a_n\}|$, number of distinct symbols

Aside: $F_1 = n$

Aside: F_2 is useful in computing variance in occurences of elements.

$$\frac{1}{m}\sum_{s=1}^{m}\left(f_s - \frac{n}{m}\right)^2 = \frac{1}{m}\sum_{s=1}^{m}\left(f_s^2 - 2\frac{n}{m}f_s + \left(\frac{n}{m}\right)^2\right) = \left(\frac{1}{m}\sum_{s=1}^{m} f_s^2\right) - \frac{2n}{m^2}\underbrace{\sum_{s=1}^{m} f_s}_{=n} + \frac{n^2}{m^2}$$

$$= \frac{F_2}{m} - \frac{n^2}{m^2},$$

Aside: $\lim_{p\to\infty} F_p^{1/p} = \lim_{p\to\infty}\left(\sum_{s=1}^{m} f_s^p\right)^{1/p}$ is the freq of the most frequent element(s).

F_0 - sketching (distinct elements)

$O(m)$ space solution — bit vector

$O(n \log m)$ space sol — store a list of items seen

Goal: $O(\log n \cdot \log m)$ space

Thm (lower bound) Any exact deterministic algorithm must use at least m bits of memory on some seq. of length $m+1$.

proof. Assume that an alg uses $< m$ bits of memory on all such seq.

Recall that the power set $|\mathcal{P}([m])| = 2^m$, and uniq$\{a_1, \dots, a_m\}$

Recall that the power set $|\mathcal{P}([m])| = 2^m$, and uniq $\{a_1, \dots, a_m\}$ can be any subset except $\emptyset$, so $2^m - 1$ possibilities.

But, we only have 2^{m-1} possible memory states.

Thus, two diff subsets $S_1, S_2 \in \mathcal{P}([m]) \setminus \emptyset$ must have the same state.

$\Rightarrow |S_1| = |S_2|$ because otherwise our alg would be wrong.

Let $b \in S_1$ but $b \notin S_2$. Then $S_1 \cup \{b\} = S_1$ so $|S_1 \cup \{b\}| = |S_1|$

and $S_2 \cup \{b\}$, so $|S_2 \cup \{b\}| = |S_2| + 1$.

But the alg will have same memory state after adding b, so it must be wrong on one of them.

Consider an idealized streaming alg (ISA) [Jelani Nelson's book on sketching]

1. Pick random hash function $h : [n] \to [0, 1]$
2. Calculate $z = \min_{i \in \text{stream}} h(a_i)$
3. Output $\frac{1}{z} - 1$.

Let $\{a_1, \dots, a_n\} = \{b_1, \dots, b_d\}$

$h(b_1), \dots, h(b_d) = X_1, \dots, X_d$ i.i.d. Unif. $[0, 1]$.

$z = \min \{X_i\}_{i=1}^{d}$

Lemma: If $X : \Omega \to [0, +\infty)$ is a nonnegative r.v., then

$$\mathbb{E} X = \int_{[0, +\infty)} \text{Prob}(X > x)\, dx.$$

Claim: $\mathbb{E}\, z = \frac{1}{d+1}$

proof. $\mathbb{E}\, z = \int_0^\infty \text{Prob}(z > \lambda)\, d\lambda = \int_0^1 \text{Prob}(\forall i,\ X_i > \lambda)\, d\lambda$

$$= \int_0^1 \prod_{i=1}^{t} \text{Prob}(X_i > \lambda)\, d\lambda = \int_0^1 (1-\lambda)^t\, d\lambda = \left[\frac{-(1-\lambda)^{d+1}}{d+1} \right]_0^1 = \frac{1}{t+1}$$

Claim: $\mathbb{E}\, z^2 = \frac{2}{(d+1)(d+2)}$

Claim: $\mathbb{E}\, z^2 = \frac{2}{(d+1)(d+2)}$

proof. $\mathbb{E}\, z^2 = \int_0^1 \text{Prob}(z^2 > \lambda)\, d\lambda = \int_0^1 \text{Prob}(z > \sqrt{\lambda})\, d\lambda$

$= \int_0^1 (1-\sqrt{\lambda})^d\, d\lambda = 2\int_0^1 u^d (u-1)\, du = 2\int_0^1 u^d (1-u)\, du$

Let $u = 1-\sqrt{\lambda}$

$= \frac{2}{(d+1)(d+2)}$ ☑

Then, $\text{Var}(z) = \mathbb{E}\, z^2 - (\mathbb{E}\, z)^2 = \frac{d}{(d+1)^2(d+2)} < \frac{1}{(d+1)^2}$.

Averaging algorithm (AA)

1. Run $q = \frac{1}{\varepsilon^2 \eta}$ ISAs in parallel
2. $\bar{z} = \frac{1}{q}\sum_{i=1}^{q} z_i$, where z_i comes from ISAs
3. Output $\frac{1}{\bar{z}} - 1$.

Then $\mathbb{E}(\bar{z}) = \frac{1}{d+1}$. $\text{Var}(\bar{z}) = \frac{1}{q}\cdot\frac{d}{(d+1)^2(d+2)} < \frac{1}{q(d+1)^2}$

By Chebyshev, $\text{Prob}\left(\left|\bar{z} - \frac{1}{d+1}\right| > \frac{\varepsilon}{d+1}\right) < \frac{(d+1)^2}{\varepsilon^2}\cdot\frac{1}{q(d+1)^2} = \eta$

Claim: $\text{Prob}\left(\left|\left(\frac{1}{\bar{z}} - 1\right) - d\right| > O(\varepsilon) d\right) < O(\eta)$.

proof.

$\text{Prob}\left(\left|\bar{z} - \frac{1}{d+1}\right| > \frac{\varepsilon}{d+1}\right) < \eta$

$\text{Prob}\left(\left|\bar{z}d + \bar{z} - 1\right| > \varepsilon\right) < \eta$

$\text{Prob}\left(\left|\frac{1}{\bar{z}} - d - 1\right| > \frac{\varepsilon}{|\bar{z}|}\right) < \eta$

w.p. $1-\eta$, $|\bar{z}| \le \frac{1+\varepsilon}{d+1}$

$\Rightarrow \text{Prob}\left(\left|\frac{1}{\bar{z}} - d - 1\right| > \frac{(d+1)\varepsilon}{1+\varepsilon}\right) < \eta$

$(d+1)(1+d)\varepsilon\left(1 - \varepsilon + \frac{\varepsilon}{2} - O(\varepsilon^2)\right)$

$= (d+1)O(\varepsilon) + d\varepsilon + \varepsilon$

$= O(\varepsilon)$

$$\Rightarrow \text{Prob}\left(\left| \frac{1}{z} - d - 1 \right| > \underbrace{\frac{\varepsilon(d+1)}{1+\varepsilon}}_{O(\varepsilon)d} \right) < \underbrace{\eta}_{O(\eta)}$$

With high prob our estimator D within a factor $1+O(\varepsilon)$ of $1+d$.

Space-complexity $O\left(\frac{1}{\varepsilon^2 \eta} \cdot \text{space}[0,1] \right)$

Median of averages

1. Instantiate $s = \lceil 36 \ln\left(\frac{2}{\delta}\right) \rceil$ ind. copies of averaging alg with $\eta = \frac{1}{3}$
2. Output the median $\hat{d}$ of $\left\{ \frac{1}{z_j} - 1 \right\}_{j=1}^{s}$ where z_j is the j^{th} output of AA.

Claim: $\text{Prob}\left(|\hat{d} - d| > \varepsilon d \right) < \delta$

proof. Let $Y_j = \begin{cases} 1 & \text{if } \left| \frac{1}{z_j} - d \right| > \varepsilon d \\ 0 & \text{otherwise} \end{cases}$

The median fails if at least half of the $Y_j = 1$, i.e. $\sum_{j=1}^{s} Y_j > \frac{s}{2}$.

Note $\text{Prob}\left(\sum Y_j > \frac{s}{2} \right) = \text{Prob}\left(\sum Y_j - \frac{s}{3} > \frac{s}{6} \right)$

Simplify by using the stronger assumption that $\mathbb{E}\, Y_j = \frac{1}{3}$ (i.e. higher failure rate)

$$= \text{Prob}\left(\sum Y_j - \mathbb{E} \sum Y_j > \frac{1}{2} \mathbb{E}\, Y_j \right) < \exp\left(\frac{-(0.5)^2 \cdot s/3}{3} \right) < \delta$$

↑ Chernoff bound

Space-complexity $O\left(\frac{1}{\varepsilon^2} \lg \frac{1}{\delta} \cdot \text{space}[0,1] \right)$

Consider the following algorithm [from book]

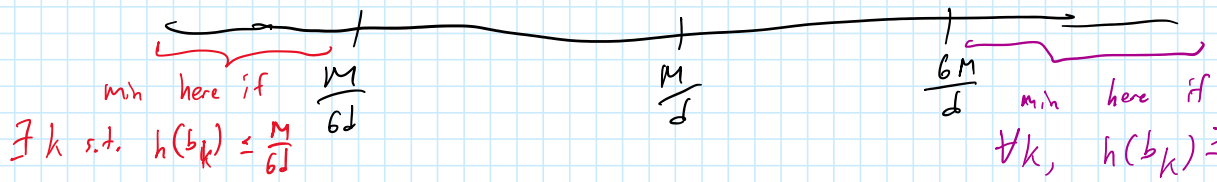
- Same notation as before
- Pick 2-ind hash function $h : [n] \to [M]$, where $M > n$.
- Output $z = \dfrac{M}{\min_{i \in \text{stream}} h(a_i)}$, an estimate for d, # distinct items.

$= \min$

Note: $S = \{ h(b_1), \ldots, h(b_d) \}$ is a set of d random and pairwise ind. uniform values from $[M]$.

Claim: With prob at least $\frac{2}{3} - \frac{d}{M}$, $\frac{d}{6} \le t \le 6d$.

proof. Equivalent to show $\frac{M}{6d} \le \min \le \frac{6M}{d}$.



- $\text{Prob}\left(\min \le \frac{M}{6d}\right) = \text{Prob}\left(\exists k,\ h(b_k) \le \frac{M}{6d}\right)$

$$\le \sum_{i=1}^{d} \text{Prob}\left(h(b_i) \le \frac{M}{6d}\right) \le d\left(\frac{\lceil \frac{M}{6d} \rceil}{M}\right)$$

$$\le d\left(\frac{1}{6d} + \frac{1}{M}\right) \le \frac{1}{6} + \frac{d}{M}.$$

- $\text{Prob}\left(\min \ge \frac{6M}{d}\right) = \text{Prob}\left(\forall k,\ h(b_k) \ge \frac{6M}{d}\right)$

Let $Y_i = \begin{cases} 0, & \text{if } h(b_i) \ge \frac{6M}{d} \\ 1, & \text{otherwise} \end{cases}$, $Y = \sum_{i=1}^{d} Y_i$.

$$= \text{Prob}(Y = 0)$$

Chebyshev $\Big\{$

$$\le \text{Prob}\left(|Y - \mathbb{E}Y| \ge \mathbb{E}Y\right)$$

$$\le \frac{\text{Var}(Y)}{(\mathbb{E}Y)^2} \le \frac{1}{\mathbb{E}Y} \le \frac{1}{6}.$$

Note: $\text{Prob}(Y_i = 1) = \text{Prob}\left(h(b_i) < \frac{6M}{d}\right) \ge \frac{\frac{6M}{d}}{M} = \frac{6}{d}$

$\Rightarrow \mathbb{E}Y_i \ge \frac{6}{d}$, $\mathbb{E}Y \ge 6$

Also, $\text{Var}(Y) = d\,\text{Var}(Y_i)$ by 2-independence.

$\text{Var}(Y_i) = \mathbb{E}(Y_i^2) - \mathbb{E}^2 Y_i = \mathbb{E}Y_i - \mathbb{E}^2 Y_i \le \mathbb{E}Y_i$

$\Rightarrow \text{Var}(Y) \le \mathbb{E}Y$

Thus, with prob. at least $\frac{2}{3} - \frac{d}{M}$, $\frac{d}{6} \le t \le 6d$. ∎