12. More frequency moment sketching

Thursday, October 7, 2021 3:06 PM

Last time: Idealized F_0 sketch and constant approximation

Today: ε-approximation F_0 sketch and F_2 sketch

Non-idealized Flajolet-Martin counting [1985]
(Geometric sampling)

1. Pick h from a 2-wise family $[n] \to [n]$ for n a power of 2.
2. Maintain $X = \max_{a_i \in \text{stream}} \text{lsb}(h(a_i))$, where lsb is the least-significant 1-bit of a number
3. Output $\tilde{d} = 2^X$

$\text{lsb}(1011011100) = 3$

$\text{lsb}(10110000) = 5$

For fixed j, let $Z_j = \left|\{b_i \in \text{stream} \mid \text{lsb}(h(b_i)) = j\}\right|$ ← unique $\{a_1, \dots, a_n\}$

Let $Z_{>j} = \left|\{b_i \in \text{stream} \mid \text{lsb}(h(b_i)) > j\}\right|$

Let $Y_i = \begin{cases} 1 & \text{if } \text{lsb}(h(b_i)) = j \\ 0 & \text{otherwise} \end{cases}$ $\qquad \mathbb{E} Y_i = \dfrac{1}{2^{j+1}}$

$$\text{Var}(Y_i) = \mathbb{E} Y_i^2 - (\mathbb{E} Y_i)^2 = \frac{1}{2^{j+1}} - \frac{1}{2^{2j+2}} < \frac{1}{2^{j+1}}$$

Then $Z_j = \sum_{b_i \in \text{str}} Y_i$. $\quad \mathbb{E} Z_j = \dfrac{d}{2^{j+1}}$, where $d = |\text{uniq}(\text{stream})|$

$$\mathbb{E} Z_{>j} = d\left(\frac{1}{2^{j+2}} + \frac{1}{2^{j+3}} + \cdots\right) = \frac{d}{2^{j+1}}$$

$$\text{Var}(Z_j) = \text{Var}\left(\sum Y_i\right) = \mathbb{E}\left(\sum Y_i\right)^2 - \left(\mathbb{E} \sum Y_i\right)^2$$

$$\mathbb{E}\left[(Y_1 + \cdots + Y_d)(Y_1 + \cdots + Y_d)\right] = \mathbb{E}\left[Y_1^2 + \cdots + Y_d^2 + \overbrace{2\sum_{i_1 \neq i_2} Y_{i_1} Y_{i_2}}^{\text{By 2-ind, } (\mathbb{E} Y_{i_1})(\mathbb{E} Y_{i_2}) = \mathbb{E}(Y_{i_1} Y_{i_2})}\right]$$

$$= \sum\left[\mathbb{E} Y_i^2 - (\mathbb{E} Y_i)^2\right] = \sum \text{Var}(Y_i) < \frac{d}{2^{j+1}} \implies \text{Var}(Z_j) < \frac{d}{2^{j+1}}.$$

Now for $j^* = \lfloor \lg d - 5 \rfloor$, we have

$$\lg d - 6 \leq j^* \leq \lg d - 5$$

Thus
$$\frac{d}{2^{\lg d - 4}} \leq \mathbb{E} Z_{j^*} \leq \frac{d}{2^{\lg d - 5}}$$

$$\implies 16 \leq \mathbb{E} Z_{j^*} \leq 32$$

$$\Rightarrow \quad 16 \le \mathbb{E} Z_{j^*} \le 32$$

$$\text{Prob}(Z_{j^*} = 0) \le \text{Prob}(|Z_{j^*} - \mathbb{E} Z_{j^*}| \ge 16) \le \frac{\text{Var}(Z_{j^*})}{256} < \frac{\frac{d}{2^{\lg d - 5}}}{256} = \frac{1}{8}$$

Chebyshev.

For $j = \lfloor \lg d + 5 \rfloor$, we have $j > \lg d + 4$

$$\mathbb{E} Z_{>j} \le \frac{d}{2^{\lg d + 5}} = \frac{1}{32} \Rightarrow \text{Prob}(Z_{>j} \ge 1) < \frac{1}{32} \text{ by Markov}$$

Thus w.p. $1 - \frac{1}{8} - \frac{1}{32}$, the max lsb will be b/t j^* and j, in a constant range,

i.e. we get a $O(1)$-approximation with high probability   64-approx

Our estimate $\tilde{d}$ satisfies $\frac{d}{C} \le \tilde{d} \le Cd$ for some constant C.

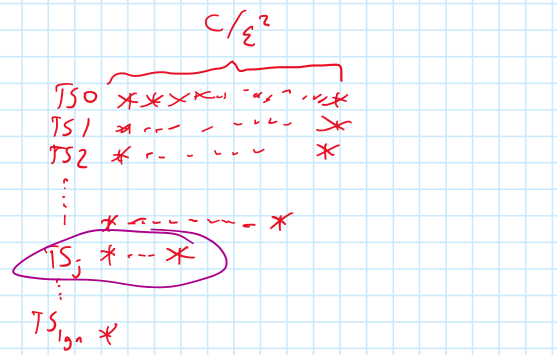
with prob. $1 - \frac{8}{C}$ using $O(\log \log n)$ bits space for the max lsb

and $O(\log n)$ bits for the hash function

Refine to $1+\varepsilon$ solution

Trivial solution TS stores the first $\frac{C}{\varepsilon^2}$ distinct elements.

TS is a $(1+\varepsilon)$-solution if $d \le \frac{C}{\varepsilon^2}$.



Algorithm

1. Instantiate $TS_0, \dots, TS_{\lg n}$
2. Pick $g : [n] \to [n]$ from 2-wise family
3. Feed a_i to $TS_{lsb(g(a_i))}$
4. Output $B_j 2^{j+1}$ where $B_j = |TS_j| \approx \frac{1}{\varepsilon^2}$

More precisely, use $j = \left\lceil \lg \frac{\varepsilon^2 \hat{d}}{32} \right\rceil$ where $\hat{d}$ is a constant approx.

Let r.v. B_j be the # distinct elem hashed by g to TS_j.

Let $\mathbb{E} B_j = \frac{d}{2^{j+1}} = Q_j$.

By Chebyshev, $B_j = Q_j \pm O(\sqrt{Q_j})$ with good probability

$$= (1 \pm O(\varepsilon)) Q_j \quad \text{if } Q_j \approx \frac{1}{\varepsilon^2}$$

$$= (1 \pm O(\varepsilon)) Q_j \quad \text{if} \quad Q_j \approx \frac{1}{\varepsilon^2}$$

$$\left[1 \pm \frac{O(\sqrt{Q_j})}{Q_j} \right] Q_j, \qquad \frac{O\left(\frac{1}{\varepsilon}\right)}{\frac{1}{\varepsilon^2}} = O(\varepsilon)$$

Notice that Q_j's decrease by factors of 2, and with good probability, one of the Q_j's is close to $\frac{1}{\varepsilon^2}$

Final space : $\underbrace{\frac{C}{\varepsilon^2}}_{\#\text{ unique}} \cdot \underbrace{(\lg n)}_{\text{storing entries + hash}} \cdot \underbrace{(\lg n)}_{\text{copies of } \pi s} \cdot \underbrace{(\lg \frac{1}{\delta})}_{\text{medians to obtain high success prob.}} = O\left(\varepsilon^2 (\lg^2 n)(\lg \frac{1}{\delta}) \right)$ bits

---

Can do better e.g. $O\left(\frac{1}{\varepsilon^2} \log \log n + \underbrace{\log n}_{\text{hash}} \right)$ for HyperLogLog

[Kane, Nelson, Woodruff, 2010, PODS]: $O\left(\frac{1}{\varepsilon^2} + \log n \right)$ optimal

---

F_2 second moment freq sketch

Consider a stream $S = \{ a_1, \dots, a_n \}$, $a_i \in [m]$

Let $f_s = | \{ a_i \mid a_i = s \} |$, the frequency of s in the stream.

We want to estimate $F_2 = \sum_{s=1}^{m} f_s^2$.

Algorithm:

1. Let $h : [m] \to \{-1, 1\}$ be 4-wise independent.
2. Let $x_s = h(s)$, Bernoulli r.v. with equal prob. of going to -1 or 1.
3. Output $a = \left(\sum_{i=1}^{n} h(a_i) \right)^2$ i.e. each time we see s, add x_s to a variable, and square at the end.

Notice: $a = \left(\sum_{i=1}^{m} x_s f_s \right)^2$

Claim: $\mathbb{E}(a^2) \le 3 (\mathbb{E} a)^2$

proof. $\mathbb{E}(a^2) = \mathbb{E}\left(\sum^{m} x_s f_s \right)^4 = \mathbb{E}\left(\sum x_s x_t x_u x_v f_s f_t f_u f_v \right)$

Claim: $\mathbb{E}(a^2) \le 3(\mathbb{E}a)^2$

proof. $$\mathbb{E}(a^2) = \mathbb{E}\left(\sum_{s=1}^{m} x_s f_s\right)^4 = \mathbb{E}\left(\sum_{1\le s,t,u,v\le m} x_s x_t x_u x_v f_s f_t f_u f_v\right)$$

Notice, if any s, t, u, v are distinct, by 4-independence, the expectation is 0.

$\Rightarrow$ only cases are either all 4 variables the same or 2 pairs

$$\mathbb{E}(a^2) = \binom{4}{2}\mathbb{E}\left(\sum_{s=1}^{m}\sum_{t=s+1}^{m} x_s^2 x_t^2 f_s^2 f_t^2\right) + \mathbb{E}\left(\sum_{s=1}^{m} x_s^4 f_s^4\right)$$

$$= 6\sum_{s=1}^{m}\sum_{t=s+1}^{m} f_s^2 f_t^2 + \sum_{s=1}^{m} f_s^4$$

$$\le 3\left(\sum_{s=1}^{m} f_s^2\right)^2 = 3\,\mathbb{E}^2 a.$$ ∎

Thm 6.3 The average x of $r = \frac{2}{\varepsilon^2\delta}$ estimates $a_1, \dots, a_r$ using independent sets of 4-way ind. hash functions is

$$\text{Prob}\left(|x - \mathbb{E}x| > \varepsilon\,\mathbb{E}x\right) < \frac{\text{Var}(x)}{\varepsilon^2\,\mathbb{E}^2 x} \le \delta$$

proof. $\text{Var}(x) \le \delta\,\varepsilon^2\,\mathbb{E}^2 x$, and the rest follows by Chebyshev. ∎