Boosting the probability Cont'd

Analysis

Define

$$Y_i = \begin{cases} 1, & \text{if } |\widehat{f}_a - f_a| \geq \epsilon \|f\|_2; \\ 0, & \text{otherwise.} \end{cases}$$

- For $k = O(1/\epsilon^2)$, we have $P(Y_i = 1) < \frac{1}{3}$.
- Note that $\mu = E(\sum_i Y_i) \leq \frac{t}{3}$. Then by the Chernoff bound,

t次中估计失败的次数

$\leq \frac{t}{3}$

P (media fail) ≤ $$P(\sum_i Y_i > \frac{t}{2}) \leq P(\sum_i Y_i > (1 + \frac{1}{2})\mu)$$ $(1+\frac{1}{2})\mu$ $\frac{t}{3}$

有一半以上是failure的概率

$$\leq \exp\left(-\mu(1/2)^2/4\right) < \exp\left(-t/48\right) < \delta,$$

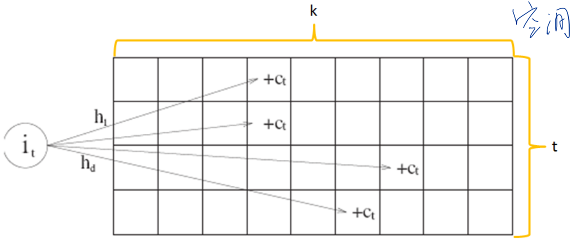
或 或 或

thus, we have $t = O(\log 1/\delta)$.

- Finally, we can get an $(\epsilon, \delta)-$approximation in space complexity $O(\frac{\log 1/\delta}{\epsilon^2})$ counters.

Count min or Cormode-Muthukrishnan sketch

空间大小 $M = t \times k$



Algorithm

1: $C[1...t][1...k] \leftarrow \overleftarrow{0}$, where $k = \frac{2}{\epsilon}$ and $t = \lceil \log(1/\delta) \rceil$;

2: Choose t independent hash functions $h_1, h_2, \cdots, h_t : [n] \rightarrow [k]$;

没有 $g(x)$，值只会变大

发生 collision

Process item (j, c), where $c = 1$:

3: for $i = 1$ to t do $C[i][h_i(j)] \leftarrow C[i][h_i(j)] + c$;

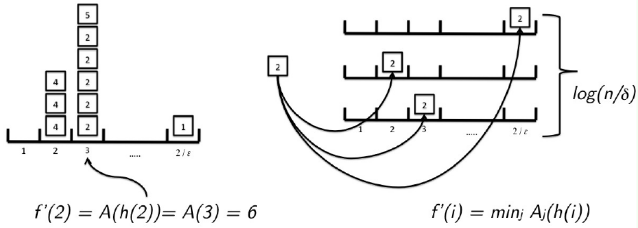
Output:

所以 不是无偏估计

4: On query a, report $\widehat{f_a} = \min_{1 \le i \le t} C[i][h_i(a)]$;

CM sketch analysis

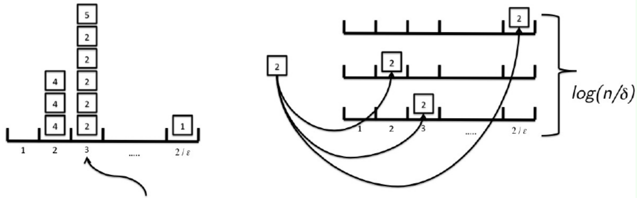
Analysis



- Clearly, for each i, we immediately have $f(a) \leq count[i, h_i(a)]$. However, the bound may be poor.

CM sketch analysis

Analysis



$f'(2) = A(h(2)) = A(3) = 6$ $\qquad f'(i) = min_j\, A_j(h(i))$

- Clearly, for each i, we immediately have $f(a) \leq count[i, h_i(a)]$. However, the bound may be poor.
- To get a better estimator, we will take the minimum over all the rows in count.

CM sketch analysis cont.

Analysis

- For a fixed a, we now analyze the collision in one such counter, say in $count[i, h_i(a)]$. Let r.v. X_i denote this collision.
- For $j \in [n] \setminus \{a\}$, let

$$Y_{i,j} = \begin{cases} 1, & \text{if } h_i(j) = h_i(a); \\ 0, & \text{otherwise.} \end{cases}$$

  be the indicator of the event $h_i(j) = h_i(a)$. Notice that j makes a contribution to the counter iff $Y_{i,j} = 1$ (Note that $E(Y_{i,j}) = \frac{1}{k}$).
- Thus, we have $X_i = \sum_{j \in [n] \setminus \{a\}} f_j Y_{i,j}$. By linearity of expectation,

高估的

所有元素除了a的计数

与f_a没有关系

$$E[X_i] = X_i = \sum_{j \in [n] \setminus \{a\}} \frac{f_j}{k} = \frac{\|f\|_1 - f_a}{k} = \frac{\|f_{-a}\|_1}{k}.$$

不是无偏估计，无法用切比雪夫.

- Since each $f_j \geq 0$, we have $X_i \geq 0$, and we can apply Markov's inequality to get (by choosing the value of k) Markov.

每一行:

$$P[X_i \geq \epsilon \|f\|_1] \leq P[X_i \geq \epsilon \|f_{-a}\|_1] \leq \frac{\|f_{-a}\|_1}{k\epsilon \|f_{-a}\|_1} = \frac{1}{2}.$$

一取合适的k使…$=\frac{1}{2}$

$k = 2 \cdot \frac{1}{\varepsilon}$

CM sketch analysis cont.

Analysis

- The above probability is for one counter. We have t such counters, mutually independent. The excess in the output $\widehat{f}_a - f_a$, is the minimum of excesses X_i over all $i \in [t]$. Thus

$$P[\widehat{f}_a - f_a \geq \epsilon \|f\|_1] \leq P[\widehat{f}_a - f_a \geq \epsilon \|f_{-a}\|_1]$$

$\Downarrow$ f_i 相互独立

$$= P[\min\{X_1, \cdots, X_t\} \geq \epsilon \|f_{-a}\|_1] = \Pi_{i=1}^{t} P[X_i \geq \epsilon \|f_{-a}\|_1] \leq \frac{1}{2^t}. < \delta$$

$\Rightarrow t > \log_2 \frac{1}{\delta}$

- Using our choice of t, this probability is at most δ. Thus, we have shown that, with high probability,

$$f_a \leq \widehat{f}_a \leq f_a + \epsilon \|f_{-a}\|_1.$$

只增不减，不用 $f_a - \epsilon\|\ \|_1$

$k \leq \frac{2}{\varepsilon}$, $t > \log_b \frac{1}{\delta}$ $(b>1)$

- Thus the space requirement is therefore $M = O(\frac{\log 1/\delta}{\epsilon})$ counters.

Ω O

Take-home messages

- Data streaming
- Deterministic algorithm
- Randomized algorithm
  - Naive sampling
  - Count sketch
  - Count min sketch