Count-Min Sketch

Counting (frequent) elements from stream $e_1, \dots, e_N \in [n]$ $\quad e_j = (i_t, \Delta_t)$ $\quad \Delta_t \in \mathbb{N}$

add Δ_t to f_{i_t}

Algorithm

Table C [k rows] [w columns]

we specify w later

specified later

Let $h_1, \dots, h_k$ be 2-universal independent hash functions from $[n] \to [w]$

For each item $e_j = (i_t, \Delta_t)$ in the stream do

  For $\ell = 1$ to k do

$$C[\ell, h_\ell(i_t)] \leftarrow C[\ell, h_\ell(i_t)] + \Delta_t$$

For each $i \in [n]$ set

$$\tilde{f}_i = \min_{\ell \in [k]} C[\ell, h_\ell(i)]$$

15 ? 10

1	2	3	4	5	6
10	20	1	2	5	8
4	7	9	1	7	1
8	1	7	2	3	4

Theorem

If always $f \geq 0$ (frequency vector) and $k = \Omega(\log \frac{1}{\delta})$ and $w > \frac{2}{\varepsilon}$, then for any fixed $i \in [n]$: $\Pr[\tilde{f}_i \geq f_i + \varepsilon \|f\|_1] \leq \delta$

$\|f\|_1 = \sum_{i \in [n]} |f_i|$ 1-norm

Proof: Fix $i \in [n]$

Random variable $Z_\ell = C[h_\ell(i)]$: value of counter in row ℓ that i is hashed to, $\forall \ell \in [k]$

$$E[Z_\ell] = f_i + \sum_{i \neq j} f_j \cdot \Pr[h_\ell(j) = h_\ell(i)] = f_i + \sum_{j \neq i} \frac{1}{w} f_j$$

($\uparrow$ 2-universal h_ℓ)

$$= f_i + \underbrace{\frac{1}{w}}_{< \varepsilon/2} \underbrace{\sum_{j \neq i} f_j}_{\leq \|f\|_1} \leq f_i + \frac{\varepsilon}{2} \cdot \|f_1\|$$

$$E[Z_\ell - f_i] \leq \frac{\varepsilon}{2} \|f\|_1$$

By Markov inequality, $\Pr[Z_\ell - f_i \geq \varepsilon \|f\|_1] \leq \frac{1}{2}$

$$\Pr\left[\min_{\ell \in [k]} Z_\ell \geq f_i + \varepsilon \|f\|_1\right] = \Pr\left[\bigwedge_{\ell \in [k]} Z_\ell \geq f_i + \varepsilon \|f\|_1\right]$$

$$\leq \left(\frac{1}{2}\right)^k \leq \delta$$

□

Typical: set $\delta = \frac{1}{n^{c+1}}$ (for some constant $c \geq 1$)

$\rightarrow \Pr[\tilde{f}_i > f_i + \varepsilon \|f\|_1] \leq \frac{1}{n^{c+1}}$ for each i

$\Rightarrow$ For all i simultaneously

$\Pr[\forall i : \tilde{f}_i \leq f_i + \varepsilon \|f\|_1]$

$= 1 - \Pr[\exists i : \tilde{f}_i > f_i + \varepsilon \|f\|_1]$

$\leq n \cdot \frac{1}{n^{c+1}} = \frac{1}{n^c}$ Union Bound

$= 1 - \frac{1}{n^c}$ ("with high probability")

Space: $O(\underbrace{k \cdot w \cdot \log \|f\|_1}_{\text{counters}} + \underbrace{k \cdot \log n}_{\text{hash functions}}) =$

$= O(\frac{1}{\varepsilon} \log \frac{1}{\delta} \cdot (\log n + \log \|f\|_1))$

Sketch

sketch is a summary of n-dimensional data D s.t.

$$\text{sketch}(D_1 \circ D_2) = \text{sketch}(D_1) + \text{sketch}(D_2)$$

composability property

In particular, Linear sketches are of the form $\text{sketch}(D) = \Pi \cdot D$

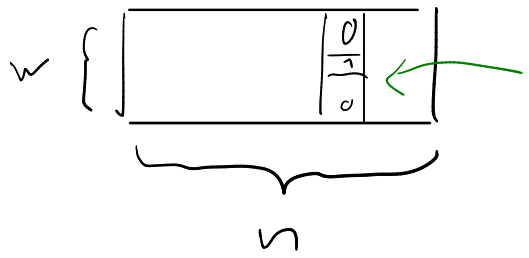
(Π: $t \times n$ matrix)

CountMin as a Linear Sketch

Need to show that sketch vector produced by CountMin sketch (for fixed hash functions) can be obtained by multiplying a fixed sketch matrix by the stream's frequency vector. (But we do not do this explicitly in the algorithm, only in the analysis)

Specifically, for a fixed hash function $h_\ell : [n] \to [w]$

we can define $w \times n$ matrix Π_ℓ



for every column $i \in [n]$ a 1-entry exactly at the position $h_\ell(i)$

$\rightsquigarrow \Pi_\ell f \rightsquigarrow$ w-dimensional vector representing the counters of the hashed items

$$\Pi = \begin{pmatrix} \Pi_1 \\ \Pi_2 \\ \vdots \\ \Pi_k \end{pmatrix}$$