SKETCHING ALGORITHM

A stream $a_1, a_2 \dots, a_m, \dots$ where $a_i \in U$

$F[a]$ = # times a appeared so far , $a \in U$: COUNTER

$\|F\| = \sum_{a \in U} F[a]$ → ISSUE U is huge, we don't enough memory!

a is the most frequent if $F[a] \geq F[b] \quad \forall b \in U$

GOAL
- small space
- two user-defined parameters
  - δ = error probability
  - ε = we cannot compute $F[]$ exactly so we give an ε-approximation

COUNT-MIN SKETCH

- $|U| = n$
- update $F[i]{+}{+}$ (in general, you can have $F[i] \mathrel{+}= v$ — v can be negative)
- query: return $F[i]$ $\quad i \in U = [n]$

$$\|F\| = \sum_{i \in U} F[i]$$

Given ε, δ we estimate $\tilde{F}[i]$ such that

$$F[i] \le \tilde{F}[i] \le F[i] + \varepsilon \|F\|$$

(ε: absolute error)

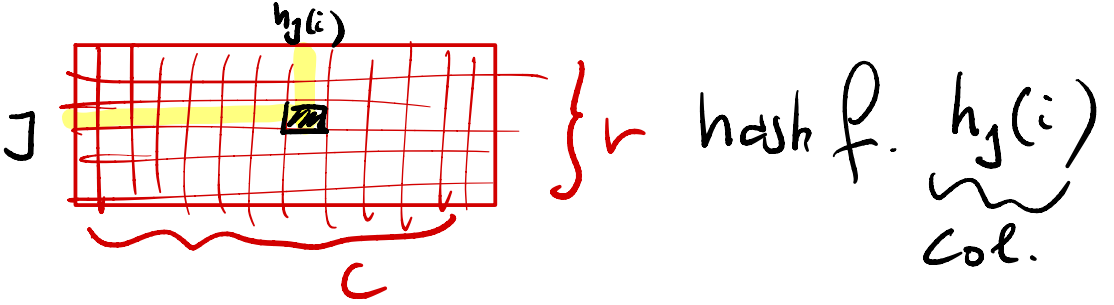
holds with probability $\ge 1 - \delta$

Limitation: if $F[i]$ is small compared to $\varepsilon \|F\|$, the estimation is not useful

CM-sketch is a sort of Bloom filter with counters

Input: ε, δ Output: $r \times c$ table

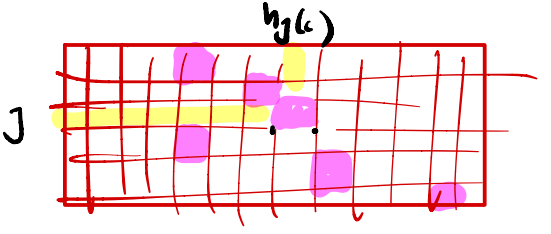
- $r = \#\text{rows} = \ln \delta^{-1}$
- $c = \#\text{columns} = \frac{e}{\varepsilon}$ where e = the base of the natural lg: ln = 2.71828...
- $T = r \times c$ table



space $O(\varepsilon^{-1} \lg \delta^{-1})$ words

- $h_0, h_1, \ldots, h_{r-1} \in \mathcal{H}$ randomly chosen from universal family
- TIME is $O(r) = O(\ln \delta^{-1})$

each item $i \in U \to$ table cells: $h_0(i), h_1(i), \ldots, h_{r-1}(i)$



update : $F[i]$++ increment by 1 the cells at positions $\langle j, h_j(i) \rangle$ for $0 \le j \le r-1$

query $\tilde{F}[i]$ → return $\min_{0 \le j \le r-1} T[j][h_j(i)]$

Issue: $\tilde{F}[i] \ge F[i]$ by construction,
but it could be too large!

Let j be the row that gives the minimum

$\tilde{F}[i] = T[j][h_j(i)] = F[i] + X_{ji}$

"garbage" due to the updates $F[h]$++ of other items $k \neq i$

we want to show that $\Pr[X_{ij} > \varepsilon \|F\|] < \delta$

Indicator variable

$$I_{jik} = \begin{cases} 1 & \text{if } h_j(i) = h_j(k) \\ 0 & \text{o.w.} \end{cases} \qquad k \neq i$$

$$\Rightarrow X_{ji} = \sum_{\substack{k \neq i \\ h_j(i) = h_j(k)}} F[k] = \sum_{k=1}^{n} I_{jik} \cdot F[k] \quad (*)$$

$h = U \to [c]$
hash
2-wise indep

Q. $E[I_{jik}] = \Pr[I_{jik} = 1] = \frac{1}{c} = \frac{\varepsilon}{e}$

$$E[X_{ji}] = \sum_{k=1}^{n} E[I_{jik}] \cdot F[k] = \sum_{k=1}^{n} \frac{\varepsilon}{e} \cdot F[k] = \frac{\varepsilon}{e} \|F\|$$

$(*)$

$F[k]$: not a random var.

$\|F\| = \sum_k F[k]$

$$e \cdot E[X_{ji}] = \varepsilon \|F\|$$

We apply MI $\left(\Pr[X \geq z] \leq \frac{E[X]}{z} : z = \varepsilon \|F\|, \; X = X_{ji}\right)$

$$\Pr[X_{ji} > \varepsilon \|F\|] \overset{MI}{\leq} \frac{E[X_{ji}]}{\varepsilon \|F\|} = \frac{E[X_{ji}]}{e\,E[X_{ji}]} = \frac{1}{e}$$

We are done:

$$\Pr[\tilde{F}[i] > F[i] + \varepsilon \|F\|] =$$

$$\Pr[\forall j : \Pr[X_{ji} > \varepsilon \|F\|]) = \prod_{j=1}^{r} \frac{1}{e} = \frac{1}{e^r} = \delta$$

$h_0, h_1, \ldots, h_{r-1}$ are independently chosen

$r = \ln \frac{1}{\delta}$