27 Apr 2022 Sketching: Count-Min & Count Sketches

Stream of tokens $a_1, \ldots, a_n$.
- Each $a_i \in [m]$

Algorithm maintains data structure of size s bits.

Goal: $s = O(\mathrm{poly}(\log n, \log m))$.

Frequency vector of the stream: $\vec{f} \in \mathbb{R}^m$

f_j = # of times token j appears in stream $\in \{0, 1, \ldots, n\}$.

$\|\vec{f}\|_1 = n$.

Two algorithms today.

Name	Space Required	Approximation Error	Probability
CountMin	$O\left(\frac{\log(mn)\log(1/\delta)}{\varepsilon}\right)$	$\le \varepsilon n = \varepsilon \|f\|_1$ one-sided	$1-\delta$
CountSketch	$O\left(\frac{\log(mn)\log(1/\delta)}{\varepsilon^2}\right)$	$\le \varepsilon \|f\|_2$ two-sided	$1-\delta$

Count Min

Given positive integers B, t. (# of hash buckets: B; # of hash functions: t)

Sample t independent hash functions $h_1, \ldots, h_t : [m] \to [B]$ from a 2-universal hash distribution.

h is 2-universal if $\forall x \neq y \in [m]$, $\forall i, j \in [B]$, $\Pr(h(x) = i \wedge h(y) = j) = \frac{1}{B^2}$.

Initialize a 2-D array C, size $B \times t$, $C[k, l] = 0 \quad \forall k, l$

// $C[k,\ell]$ counts # stream elements a_i s.t. $h_\ell(a_i) = k$.

for each a_i in the stream:
    for $\ell \in [t]$:
        let $k = h_\ell(a_i)$
        increament $C[k,\ell]$.
    endfor
endfor
// Done processing stream.

To answer FreqQuery(x):
    return $\min_{\ell \in [t]} \{ C[h_\ell(x), \ell] \}$.

Analysis. Space Requirement

Store t hash function descriptions each requiring $O(\log m)$ bits.

$$O(t \log m)$$

Store $B \times t$ array of counters in $\{0, \ldots, n\}$.

$$O(Bt \log n)$$

Approximate Correctness

For any given ℓ,
$$\mathbb{E}\left[C[h_\ell(x), \ell]\right] = f_x + \sum_{y \neq x} f_y \cdot \Pr\left(h_\ell(y) = h_\ell(x)\right)$$
$$= f_x + \frac{1}{B} \sum_{y \neq x} f_y$$
$$= f_x + \frac{1}{B}(n - f_x).$$

The random variable $C[h_\ell(x), \ell] - f_x$ is ≥ 0
and has exp val. $\leq \frac{n}{B}$.

Markov: $\Pr\left(C[h_\ell(x), \ell] - f_x \geq \frac{2n}{B}\right) \leq \frac{1}{2}$.

Independence of $h_1, \ldots, h_t$: $\Pr\left(\min_\ell \{C[h_\ell(x), \ell]\} - f_x \geq \frac{2n}{B}\right) \leq \frac{1}{2^t}$.

To get error $\leq \varepsilon n$ w/ prob $\geq 1-\delta$
set

$$\frac{2n}{B} \leq \varepsilon n \qquad \frac{1}{2^t} \leq \delta$$

$$\Downarrow \qquad \Downarrow$$

$$B = \left\lceil \frac{2}{\varepsilon} \right\rceil \qquad t \geq \log\left(\frac{1}{\delta}\right)$$

space required is $O(t \log m + Bt \log n)$

$$= O\left(\log(m) \log\left(\tfrac{1}{\delta}\right) + \tfrac{1}{\varepsilon} \log\left(\tfrac{1}{\delta}\right) \log(n)\right).$$

Count sketch: similar to Count Min but each token is added to the counter with a random +/− sign.

Consistently use same sign for x whenever it appears.

sign(x), sign(y) indep. when $x \neq y$.

This could help because if many other tokens collide in same hash bucket as x, their counters may cancel each other out.

CountSketch

$\left(B = \left\lceil \frac{3}{\varepsilon^2} \right\rceil, \quad t = 18 \ln\left(\frac{1}{\delta}\right) \right)$

Given B, t pos. integers.

Sample indep 2-univ hash functions

$$h_1, \ldots, h_t : [n] \to [B]$$

$$g_1, \ldots, g_t : [n] \to \{\pm 1\}.$$

Initialize $C[k, \ell] = 0 \quad \forall\, k \in [B],\ \ell \in [t]$

for each element a_i in stream:

  for each $\ell \in [t]$:

    $C[h_\ell(a_i), \ell] \leftarrow C[h_\ell(a_i), \ell] + g_\ell(a_i)$

  endfor

endfor

To answer FreqQuery(x):

  return median element of $\left\{ g_\ell(x) \cdot C[h_\ell(x), \ell] \;\middle|\; \ell \in [t] \right\}$

Analysis: Space is $O(t \log(m) + Bt \log(n))$

$$= O\left(\log\left(\tfrac{1}{\delta}\right)\log(m) + \tfrac{1}{\varepsilon^2}\log\left(\tfrac{1}{\delta}\right)\log(n)\right)$$

Correctness:

$$\mathbb{E}\left[g_\ell(x) \cdot C[h_\ell(x), \ell]\right]$$

$$= f_x + \sum_{y \neq x} f_y \cdot \underbrace{\mathbb{E}\left[g_\ell(x) \cdot g_\ell(y) \cdot \mathbb{1}\{h_\ell(y) = h_\ell(x)\}\right]}_{\text{“}0\text{”}}$$

$\mathbb{1}\{h_\ell(y) = h_\ell(x)\} = \begin{cases} 1 & \text{if } h_\ell(y) = h_\ell(x) \\ 0 & \text{if not} \end{cases}$

$$= f_x.$$

$$\mathrm{Var}\left(g_\ell(x) \cdot C[h_\ell(x), \ell] \right) \le \frac{1}{B} \|f\|_2^2$$

To finish up, Chebyshev $\Rightarrow$

$$\Pr\left(\left| g_\ell(x) \cdot C[h_\ell(x), \ell] - f_x \right| > \varepsilon \|f\|_2 \right)$$

$$\le \frac{1}{\varepsilon^2 B}. \qquad \left(B \ge \left\lceil \frac{3}{\varepsilon^2} \right\rceil \right)$$

$$\le \frac{1}{3}.$$

Finish up using Hoeffding.