2 Mar 2026

Sketching element frequencies.

Recap: $N = 2^b$ $\quad [N] \cong \{0,1\}^b$ "space of tokens"

$x_1, x_2, \ldots, x_m$ stream of tokens

$\vec{f} \in \mathbb{R}^N$ frequency vector

$$f_x = \#\{ i \mid x_i = x \}$$

$=$ number of times x occurs in stream

(a number between $\emptyset$ and m)

Misra-Gries outputs a list of k tokens guaranteed to include every token x s.t. $f_x > \frac{m}{k+1}$.

Sketching $\vec{f}$ means maintaining a data struct C that occupies small space (poly$(b, \log m)$)

and enables approximately answering queries

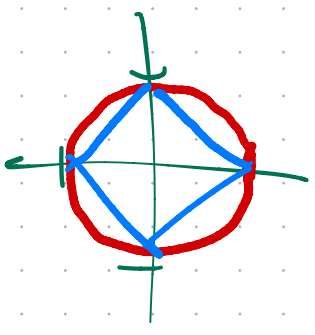
$x \mapsto \hat{f}_x$ (estimate of f_x).

(ε, δ)-PAC

$\Pr(\text{normalized error} > \varepsilon) < \delta$.

Count-Min Sketch: $\Pr\left(|\hat{f}_x - f_x| > \varepsilon \|\vec{f}\|_1\right)$

Count Sketch: $\Pr\left(|\hat{f}_x - f_x| > \varepsilon \|\vec{f}\|_2\right)$ ← (tougher to satisfy)



2-D. $\left\| \begin{matrix} x \\ y \end{matrix} \right\|_1^2 = (|x| + |y|)^2$ $\quad \left\| \begin{matrix} x \\ y \end{matrix} \right\|_2^2 = x^2 + y^2$

$= |x|^2 + 2k|x||y| + |y|^2 \geq = |x|^2 + |y|^2$

In $\mathbb{R}^N$, $\|x\|_1 \geq \|x\|_2$ always

$\|x\|_1$ can exceed $\|x\|_2$ by factor $\sqrt{N}$.

E.g. $x = \vec{1}$.

Count-Min Sketch

Sample random $h_1, \ldots, h_t : [N] \to [B]$

from 2-universal family, where $t = \lceil \log_2 (\frac{1}{\delta}) \rceil$.

initialize $C \equiv 0$ (dimension $B \times t$)

for $i = 1, 2, \ldots, m$:

  let $(b_1, \ldots, b_t) = (h_1(x), \ldots, h_t(x))$

  $\forall j \; C[b_j, j] \leftarrow C[b_j, j] + 1$ // $C[b]$ counts how many tokens hashed to bucket b.

end

Query(x): return $\hat{f}_x = \min_j C[h_j(x), j]$

obs. $\hat{f}_x$ never underestimates f_x.

$\hat{f}_x \geq f_x$ with prob. 1.

Q: $\Pr(\hat{f}_x - f_x > \varepsilon)$?

$\hat{f}_x - f_x$ = # tokens $x_i \neq x$ s.t. $h(x_i) = h(x)$.

$$\mathbb{E}\left[\hat{f}_x - f_x\right] = \;??$$

Suppose $x_i \neq x$. What is

$$\Pr\left(h(x_i) \neq h(x)\right) \overset{?}{=} \frac{B-1}{B}$$

$$\Pr\left(h(x_i) = h(x)\right) = \frac{1}{B}.$$

$$\mathbb{E}\left[\hat{f}_x - f_x\right] = \sum_{i : x_i \neq x} \Pr\left(h(x_i) = h(x)\right)$$

$$= \frac{1}{B} \cdot \#\{i \mid x_i \neq x\}.$$

$$= \frac{1}{B} \cdot \left(m - f_x\right) \leq \frac{1}{B} \cdot m = \frac{1}{B} \cdot \|f\|_1.$$

Idea #1. Set $B = \frac{1}{\varepsilon\delta}$.

$$\mathbb{E}\left[\hat{f}_x - f_x\right] \leq \varepsilon\delta \|f\|_1$$

$$\Pr\left(\hat{f}_x - f_x > \varepsilon \|f\|_1\right) \leq \delta \quad \text{by Markov.}$$

Space requirement:

- C stores B counters in range $0 \ldots m$

  $B \cdot \log(m) = \frac{1}{\varepsilon\delta} \log(m)$ bits.

- description of h: $O(\log B \cdot \log N) = O\left(\log\left(\frac{1}{\varepsilon\delta}\right) \cdot b\right)$.

Now with $\beta = \lceil \frac{2}{\varepsilon} \rceil$, $t = \lceil \log_2(\frac{1}{\delta}) \rceil$

we have

- Never underestimate $\Pr(\hat{f}_x \geq f_x) = 1$
- Probably overshoot f_x by at most $\varepsilon \cdot \|f\|_1$

$$\Pr\left(\hat{f}_x - f_x > \varepsilon \|f\|_1\right) < \delta$$

- Space requirement:

C needs $\underbrace{\frac{2}{\varepsilon}}_{\text{rows}} \times \underbrace{\log(\frac{1}{\delta})}_{\text{cols}} \times \underbrace{\log m}_{\text{bits per entry}}$

$h_1, \ldots, h_t$ needs $O\left(\log(\frac{1}{\varepsilon}) \cdot b \cdot \log(\frac{1}{\delta})\right)$

Count-Sketch. Use random signs to bring about cancellations in the "noise" without cancelling "signal."

Algorithm:

$t = ??$

$B = ??$

initialize array C of dimension $B \times t$

$C \equiv 0$.

sample random hash functions (2-univ)

$h_1, h_2, \ldots, h_t : [N] \to [B]$

$g_1, g_2, \ldots, g_t : [N] \to \{\pm 1\}$.

for $i = 1, \ldots, n$:

$(b_1, \ldots, b_t) = (h_1(x_i), \ldots, h_t(x_i))$

$\forall j \quad C[b_j, j] = C[b_j, j] + g_j(x_i)$

// could increment or decrement

end for

Query(a) :

for $j = 1, \ldots, t$ let

$a_j = C[h_j(x), j] \cdot g_j(x)$

// in absence of hash collisions this would be exactly f_x.

output $\hat{f}_x = \text{median}(a_1, \ldots, a_t)$.

What can we prove about this?

- Error is two-sided: $\hat{f}_x - f_x$ could be positive or negative.
- Error likely to be small:

$$\Pr\left(|\hat{f}_x - f_x| > \varepsilon \|f\|_2 \right) <$$

small when $B >> \frac{1}{\varepsilon^2}$ and $t >> \log_2\left(\frac{1}{\delta}\right)$.

- Space

C requires $B \times t \times \log(2m)$ bits

$h_1, \ldots, h_t, g_1, \ldots, g_t$

$O(\log B \cdot t \cdot b)$.

Total space = $O(\text{Greek} \cdot (b + \log m))$.

Analysis of count sketch

Focus on column j with hash functions $h = h_j$, $g = g_j$,

estimate $a_j = C[h(x), j] \cdot g(x)$.

Claim 1. $\mathbb{E}[q_j] = f_x$.

Proof: Let $X_{by} = \begin{cases} 1 & \text{if } h(y) = b \\ 0 & \text{if } h(y) \neq b \end{cases}$

$$Z_y = g(y)$$

$$C[b,j] = \sum_{y \in [N]} f_y \cdot X_{by} \cdot Z_y$$

$$q_j = \sum_{b \in [B]} X_{bx} \, C[b,j] \cdot Z_x$$

$$\mathbb{E}[q_j] = \sum_b \sum_y f_y \cdot \mathbb{E}[X_{bx} X_{by} Z_x Z_y]$$

$\hookrightarrow 0$ except if $x = y$

$$= \sum_b f_x \, \mathbb{E}[X_{bx}]$$

$$= f_x$$