4 March 2026

Finishing Count Sketch + Reservoir Sampling

COUNT SKETCH

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

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

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

output $\hat{f}_x = \text{median}(v_1, \ldots, v_t)$.

Analysis focuses on one estimator V_l, $l \in [t]$.

Fix x (token whose frequency is queried).

$$X_{yl} = \begin{cases} 1 & \text{if } h_l(y) = h_l(x) \\ 0 & \text{otherwise} \end{cases}$$

$$Z_{yl} = g_l(x) g_l(y) X_{yl} f_y$$

↑ the "noise term" in V_l coming from token y.

Fact: $V_l = \sum_{y \in [N]} Z_{yl}$

$$\mathbb{E}[v_\ell] = \sum_y \mathbb{E}[z_{y\ell}]$$

$$= \sum_y \mathbb{E}\left[g_\ell(x) g_\ell(y) X_{y\ell} f_y \right]$$

$$= f_x + \sum_{y \neq x} \mathbb{E}\left[g_\ell(x) g_\ell(y) X_{y\ell} \right] f_y$$

PW IND $\mathbb{E}[g_\ell(x) g_\ell(y)] \cdot \mathbb{E}[X_{y\ell}]$

$\mathbb{E}[g_\ell(x)] \cdot \mathbb{E}[g_\ell(y)]$ only depends on g_ℓ

only depends on h_ℓ

1 0

$$\boxed{\mathbb{E}[v_\ell] = f_x}$$

Now focus on $\mathrm{Var}(v_\ell)$, with an eye toward using Chebyshev.

$$\mathrm{Var}(v_\ell) = \mathbb{E}\left[(v_\ell - F_x)^2 \right]$$

$$= \mathbb{E}\left[\left(\sum_{y \neq x} z_{y\ell} \right)^2 \right]$$

$$= \sum_{y \neq x} \mathbb{E}\left[z_{y\ell}^2 \right] + \sum_{y \neq x} \sum_{w \notin \{x,y\}} \mathbb{E}\left(z_{y\ell} z_{w\ell} \right)$$

$$z_{y\ell}^2 = g_\ell(x)^2 \, g_\ell(y)^2 \, X_{y\ell}^2 \, f_y^2$$

(+1, +1, $X_{y\ell}$)

$$= X_{y\ell} \, f_y^2$$

$$\mathbb{E}\left(z_{y\ell}^2 \right) = \mathbb{E}\left[X_{y\ell} \right] \cdot f_y^2 = \frac{1}{B} f_y^2 .$$

$$z_{y\ell} z_{w\ell} = g_\ell(x) g_\ell(y) X_{y\ell} f_y \, g_\ell(x) g_\ell(w) X_{w\ell} f_w$$

(+1)

$$\mathbb{E}\left[z_{y\ell} z_{w\ell} \right] = \mathbb{E}\left[g_\ell(y) g_\ell(w) X_{y\ell} X_{w\ell} \right] f_y f_w$$

dep on g — dep on h

$$= \mathbb{E}\left(g_\ell(y) g_\ell(w)\right) \cdot \mathbb{E}\left(\chi_{y\ell}\chi_{w\ell}\right) f_y f_w$$

$$= \mathbb{E}\left[g_\ell(y)\right] \mathbb{E}\left[g_\ell(w)\right] \mathbb{E}\left(\chi_{y\ell}\chi_{w\ell}\right) f_y f_w$$

$0 \qquad 0$

$$= 0$$

$$\mathbb{E}\left[v_\ell\right] = f_x$$

$$\mathrm{Var}\left(v_\ell\right) = \frac{1}{B} \sum_{y \neq x} f_y^2 \leq \frac{1}{B} \|f\|_2^2 .$$

$$\Pr\left(\left|v_\ell - f_x\right| > \varepsilon \|f\|_2\right) < \frac{\mathrm{Var}\left(v_\ell\right)}{\varepsilon^2 \|f\|_2^2}$$

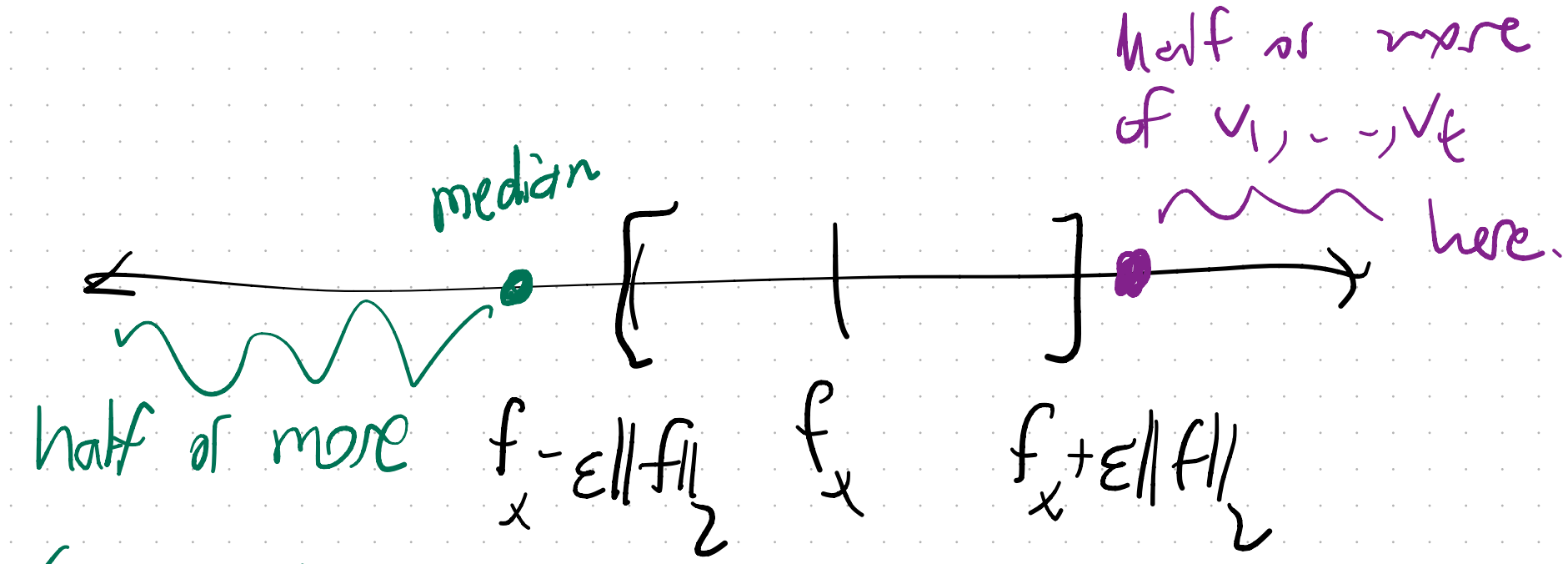
$$\leq \frac{1}{\varepsilon^2 B} < \frac{1}{3}$$

... by setting $B > 3/\varepsilon^2$.

If median $(v_1, \ldots, v_t)$ differs from f_x by more than $\varepsilon \|f\|_2$

it means

$$\#\{\ell \mid |v_\ell - f_x| > \varepsilon \|f\|_2\} \geq \frac{t}{2}.$$



$$\mathbb{E}\left[\#\{\ell \mid |v_\ell - f_x| > \varepsilon \|f\|_2\}\right] < \frac{t}{3}.$$

Hoeffding: $\Pr(\text{this}) < e^{-\frac{2\cdot(t/6)^2}{t}}$

$= e^{-t/18}$.

To make $e^{-t/18} < \delta$ set

$$t > 18 \ln(1/\delta).$$

Conclusion: Count sketch

with $B = \Theta\left(\frac{1}{\varepsilon^2}\right)$

$$t = \Theta\left(\ln \frac{1}{\delta}\right)$$

satisfies $\varepsilon \cdot \|f\|_2$ with

prob $\geq 1 - \delta$.

$N = 2^b$

Space requirement:

$$O\left(\frac{1}{\varepsilon^2} \ln\left(\frac{1}{\delta}\right) \underbrace{\ln(mN)}_{\ln|m| + b}\right) \text{ bits.}$$

2-universal

Storing one hash function from

$$[N] \longrightarrow [B]$$

takes how much space?

$B = p$ prime

$[N] \subseteq \mathbb{F}_p^d$ $\quad d = \lceil \log_p(N) \rceil$.

Use inner product $h_{a,b}(x) = \langle a, x \rangle + b \mod p$.

Description of h: $d+1$ values mod p.

coords of a $\quad$ b

$(d+1) \cdot \log_2(p)$ bits

$\leq (\log_p(N) + 2) \cdot \log_2(p)$ bits

$= O(\log_2(N))$

$B = 2^k$ $\quad k \in \mathbb{N}$,

use k indep hash functs with $\{0,1\}^k$ o/p

$k \cdot \log_2(N)$ bits.