$\boxed{\text{5 Mar 2025}}$ Count Sketch

Recall: Count-min sketch stores data about a stream of $n$ tokens in $\{0,1\}^b$.

Uses space $O\left(\left(b + \frac{1}{\varepsilon}\right) \log\left(\frac{n}{\varepsilon\delta}\right)\right)$.

Answers frequency queries with additive error $\leq \varepsilon n$ with prob $\geq 1-\delta$.

Count sketch improves the additive error to $\varepsilon \|f\|_2$ rather than $\varepsilon n = \varepsilon \|f\|_1$.

$$\|f\|_2 = \sqrt{\sum_{x \in \{0,1\}^b} f[x]^2}$$

$$\|f\|_1 = \sum_x |f[x]|$$

$\|f\|_2 \leq \|f\|_1$ always.

They are equal when $f$ has only one nonzero coordinate. (i.e. the stream is one token repeatedly)

$\|f\|_2 \ll \|f\|_1$ when many tokens occur rarely.

E.g. 1000 tokens each occurring once.

$\|f\|_1 = 1000$ $\|f\|_2 = \sqrt{1000} \approx 31.6$

Idea: use random signs to cancel hash collisions.

Preprocessing phase

will be roughly $\frac{1}{\varepsilon^2}$.

Initialize $h: \{0,1\}^b \longrightarrow [B]$

$g: \{0,1\}^b \longrightarrow \{\pm 1\}$

each sampled (independently) from a 2-universal hash family.

Initialize $C[j] = 0$ for all $j \in [B]$.

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

$$C[h(x_i)] \longleftarrow C[h(x_i)] + g(x_i)$$

Query(x): return $C[h(x)] \cdot g(x)$.

Improved Count sketch:

Use $h_1, \ldots, h_t : \{0,1\}^b \longrightarrow [B]$

$g_1, \ldots, g_t : \{0,1\}^b \longrightarrow \{\pm 1\}$

all sampled indep'tly from 2UHF.

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

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

$$C[h_j(x_i), j] = C[h_j(x_i), j] + g_j(x_i)$$

QUERY (x): return median of $\{C[h_j(x), j] \cdot g_j(x)\}$.

## Analysis:

Fix a token $x$. To estimate error in QUERY($x$) we define some rand vars

$$X_{y\ell} = \begin{cases} 1 & \text{if } h_\ell(y) = h_\ell(x) \\ \phi & \text{o/w.} \end{cases}$$

"$y$ has a hash collision with $x$"

sign of noise in preproc

sign at query time

does $y$ influence QUERY($x$)?

$$Z_{y\ell} = g_\ell(y)\, g_\ell(x)\, X_{y\ell}$$

"the amt of noise $y$ contributes to the $\ell^{th}$ estimate of QUERY($x$)." $1 \le \ell \le t$.

Sum over every token $y$ in stream $\sigma$.

$$C[h_\ell(x), \ell] = \sum_{y \in \sigma} g_\ell(y) X_{y\ell}$$

(by induction on length of stream)

$$C[h_\ell(x), \ell] \cdot g_\ell(x) = \sum_{y \in \sigma} g_\ell(y)\, g_\ell(x)\, X_{y\ell}$$

$$= \sum_{y \in \sigma} Z_{y\ell}.$$

$$\mathbb{E}\left[ C[h_\ell(x), \ell] \cdot g_\ell(x) \right] = \mathbb{E}\left[ \sum_{y \in \sigma} z_{y\ell} \right]$$

$$= \sum_{y \in \sigma} \mathbb{E}[z_{y\ell}]$$

$$= \sum_{y \in \sigma} \mathbb{E}[g_\ell(y) g_\ell(x)] \cdot \mathbb{E}[X_{y\ell}]$$

$$= f[x] + \sum_{y \in \sigma,\, y \neq x} \cancelto{0}{\mathbb{E}[g_\ell(y) g_\ell(x)]}\, \mathbb{E}[X_{y\ell}]$$

$$= f[x].$$

$$\mathrm{Var}\left( \sum_{y \in \sigma} z_{y\ell} \right) = \mathbb{E}\left[ \left( \sum_{y \in \sigma} z_{y\ell} - f[x] \right)^2 \right]$$

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

$$= \sum_{y \in \sigma,\, y \neq x} \mathbb{E}[z_{y\ell}^2] + \sum_{\substack{y \in \sigma,\, w \in \sigma \\ y \neq x,\, w \neq x}} \mathbb{E}[z_{y\ell} z_{w\ell}]$$

$y, w$ occur in different locations in stream

$$\mathbb{E}\left[ Z_{y\ell}^2 \right] = \mathbb{E}\left[ g_\ell(y)^2 \, g_\ell(x)^2 \, X_{y\ell}^2 \right]$$

$$= \mathbb{E}\left[ X_{y\ell} \right] = \frac{1}{B}.$$

$$\mathbb{E}\left[ Z_{y\ell} Z_{w\ell} \right] = \mathbb{E}\left[ g_\ell(y) \, g_\ell(w) \, \cancelto{1}{g_\ell(x)^2} \, X_{y\ell} X_{w\ell} \right]$$

$$= \mathbb{1}[y=w] \cdot \frac{1}{B}.$$

$$\mathrm{Var}\left( \sum_{y \in \sigma} Z_{y\ell} \right)$$

$$= \sum_{y \in \{0,1\}^b} \frac{1}{B} \cdot f[y]^2$$

$$= \frac{1}{B} \|f\|_2^2.$$