RAD (6-7) Basic count sketch

basic count sketch

vi vil estimere frekvens af elemener i en strøm:
$(x_0, \Delta_0), \ldots, (x_{n-1}, \Delta n - 1)$
frekvens af x: $f_x := \sum\{\Delta_i | x_i = x\}$
E.g, with (7, 20), (3, −5), (7, −3), (9, 100), $f_7 = 20 - 3 = 17$

vi bruger 2 hash funktioner og for et given ε > 0 giver vi et estimat $\hat{f}_x$, af frekvensen, f_x, for en given nøgle x så at:
$Pr[|\hat{f}_x - f_x| \leq \epsilon ||f_{-x}||_2] \geq \frac{3}{4}$.

her er Euclidean norm:

$$||f_{-x}||_2 = \sqrt{||f_{-x}||_2^2} = \sqrt{\sum_{y \neq x} f_y^2}$$

$\mathbf{f} = (f_0, \cdots, f_{u-1})$: $\mathbf{f_{-x}} = (f_0, \cdots, f_{x-1}, f_{x+1}, \cdots, f_{u-1})$.

Pseudokode

k er antal elementer vi kan counte (antal spande).

$s(x)$ hasher til -1 eller +1

h(x) bruges til at bestemme hvad vi tæller, det giver sig selv at der kan opstå kollisioner, derved kan vi risikere at en tæller faktisk tæller flere ting.

k afhænger af hvad vi ønsker af fejl, ε

kaldet til proces kan ses som at putte x i en af k spande sammen med et tegn $s(x)$

så en nøgle x bidrager til spanden værdi C[h(x)] med $s(x)f_x$
hvor f_x er frekvensen af x

C er array der fyldes med 0.

1: **function BCS-INITIALIZE** (ε)
2: $k \leftarrow \left\lceil \frac{4}{\varepsilon^2} \right\rceil$ ▷ Trivial version had $k = 1$
3: Pick strongly universal $s : [u] \to \{-1, +1\}$
4: Pick universal $h : [u] \to [k]$ ▷ h independent of s
5: $C[0, \ldots, k-1] \leftarrow 0$

6: **function BCS-PROCESS** (x, Δ)
7: $C[h(x)] \leftarrow C[h(x)] + s(x) \cdot \Delta$ ▷ trivial: $C+ = s(x) \cdot \Delta$

8: **function BCS-QUERY** (x)
9: return $\hat{f}_x = s(x) \cdot C[h(x)]$ ▷ trivial: $\hat{f}_x = s(x) \cdot C$
s(x) bruges til at give det korrekte sign.

Constant time per update/query and $\mathcal{O}(k) = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ space. Vi skal have plads til alle vores tællere.

en hash funktion $h : U \to [k]$ er q-universal hvis der for et bestemt q gælder: **q-wise independence $h(x_1), \ldots, h(x_q)$ are independent and uniformity h(x_i) uniform in [k]. strong universality = 2-universality

den triviale version giver unbiased estimator: $\mathrm{E}\left[\hat{f}_x\right] = f_x$, men stor varians, derfor bruger vi basic count sketch istedet.

$k = \left\lceil \frac{4}{\varepsilon^2} \right\rceil$ så vi får en estmations fejl $\epsilon||f_{-x}||_2$, det kræver middelværdi og varians, dertil kan man bruge chebyshev.

$$\mathrm{E}\left[\hat{f}_x\right] = f_x \quad \mathrm{Var}\left[\hat{f}_x\right] \leq \frac{\|\mathbf{f}_{-x}\|_2^2}{k}$$

$$\Pr\left[\left|\hat{f}_x - f_x\right| \geq \varepsilon \cdot \|\| \mathbf{f}_{-x}\|_2\right] \leq \frac{\operatorname{Var}\left[\hat{f}_x\right]}{\left(\varepsilon \cdot \|\mathbf{f}_{-x}\|_2\right)^2} = \frac{1}{k\varepsilon^2} \leq \frac{1}{4}$$

$$\Pr\left[|X - \mu_X| \geq t\right] \leq \frac{\sigma_X^2}{t^2} = \frac{\operatorname{Var}[X]}{t^2}$$

analyse

Lemma: $\mathrm{E}\left[\hat{f}_x\right] = f_x$ so $\hat{f}_x$ unbiased estimator of f_x.

Proof $E[\hat{f}_x] = E[s(x) \cdot C[h(x)]] = f_x$

her er $C[h(x)]$ hvor mange gange vi har countet h(x).
Det er det samme som at vi løber igennem universet og

et element tælles kun når $B_{xy} = 1$, da B_{xy} er indikator

$$= s(x) \cdot \sum_{y \in [u]} f_y s(y) B_{xy} \text{ where } B_{xy} = [h(y) = h(x)]$$

ligesom at man må trække en konstant ud af en sum så må man også gerne putte den ind i en sum.

$$= \sum_{y \in [u]} f_y s(x) s(y) B_{xy}$$

$s(x)^2 B_{xx} = s(y)^2 B_{yy} = 1$ fordi $s(x) = \{-1, 1\}$, $s(x) = s(y)$ så $s(x)^2 = 1$, $B_{xx} = [h(x) = h(x)] = 1$. Derfor splitter vi vores sum i to: alle de cases fra summen hvor x og y er ens, det er f_x:

$$\hat{f}_x = f_x + \sum_{y \neq x} f_y s(x) s(y) B_{xy}$$

$$\mathrm{E}\left[\hat{f}_x\right] = \mathrm{E}[f_x] + \mathrm{E}[\sum_{y \neq x} f_y s(x) s(y) B_{xy}]$$

Lineraty of exp

$$\mathrm{E}\left[\hat{f}_x\right] = \mathrm{E}[f_x] + \mathrm{E}[\sum_{y \neq x} f_y s(x) s(y) B_{xy}]$$

$\mathbb{E}[f_x] = f_x$, f_x er en konstant

$$\mathrm{E}\left[\hat{f}_x\right] = f_x + \sum_{y \neq x} \mathrm{E}[f_y] \mathrm{E}\left[s(x) s(y) B_{xy}\right] \text{ (linearity of expectation)}$$

$\mathbb{E}[f_y] = f_y$

$$= f_x + \sum_{y \neq x} f_y \mathrm{E}[s(x)] \mathrm{E}\left[s(y) B_{xy}\right] \text{ (independence)}$$

$E[s(x)] = 0$ fordi $E[s(x)] = 1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{2} = 0$, husk s(x) er strækt universal derfor gælder der uniformitet: $\forall x \in U \rightarrow m = \{-1, 1\}, s(x)$ uniform i [m], dvs sandsynligheden for at s(x) hasher til 1 eller -1 er $\frac{1}{2}$. derfor forsvinder summen da alle led ganges med 0. når $x \neq y$ er s(x), s(y) og B_{xy} uafhængige. Derfor gælder der
$\mathrm{E}[s(x) s(y) B_{xy}] = \mathrm{E}[s(x)] \mathrm{E}[s(y)] \mathrm{E}[B_{xy}] = 0 \cdot 0 \cdot B_{xy}$

$$\mathrm{E}\left[\hat{f}_x\right] = f_x$$

Median trick

ved at lave m uafhængige estimater og give deres median kan vi begrænse sandsynligheden for at estimatet er "dårligt" til at være exponentialt småt i m. Det er median tricket. Meget mindre fejl, kigger på sandsynlighed for noget går galt, mere specifikt median er forkert.

Theorem: Uafhængig $X_1, \ldots, X_m$ så at $\Pr\left[X_i \notin [x_{\min}, x_{\max}]\right] \leq 1/4$. så $\Pr\left[X_{(\lceil m/2 \rceil)} \notin [x_{\min}, x_{\max}]\right] \leq e^{(-m/12)}$.

x_max og x_min er en eller anden grænse som vi gerne vil vise at median ikke bliver større end. det er sandsynligheden for at man ryger over eller under den her grænse.

De X_i der ikke er i intervallet $[x_{\min}, x_{max}]$, dvs. $X_i < x_{\min} \vee X_i > x_{max}$

Desto flere eksperimenter desto større er sandsynligheden for at vi ikke rammer forkert på medianen. medianen går galt hvis halvdelen af eksperimenterne fejler i den ene retning, så vi vil gerne være sikker på at sandsynligheden for at eksperimenter fejler er mindre end en 1/2, vi har her at det er mindre end 1/4 så det er ok.

$$X_i \notin [x_{\min}, x_{max}] = Pr\{X_i \leq x_{\min} \vee X_i \geq x_{max}\} \leq \frac{1}{4}$$

"bevis"

- Definer $Z_i = [X_i \notin [x_{\min}, x_{\max}]]$ og $Z = \sum_{i=1}^{m} Z_i$. Z_i er en indikator.
- $Z_{(\lceil m/2 \rceil)} = 1 \Longrightarrow Z \geq m/2$.
- $\mathrm{E}[Z] = \sum_{i=1}^{m} \mathrm{E}\left[Z_i\right] = \sum_{i=1}^{m} \Pr\left[Z_i\right] \leq m/4 \equiv \mu$.
- Bad event implies $Z \geq m/2 = (1+\delta)\mu$ with $\delta = 1$.
- The Z_i are independent, so by Chernoff

  $$\Pr[Z \geq m/2] \leq \exp\left(-\mu\delta^2/3\right) = \exp(-m/12)$$

  Derved

  $$\Pr\left[X_{(\lceil m/2 \rceil)} \notin [x_{\min}, x_{\max}]\right] \leq \Pr[Z \geq m/2] \leq \exp(-m/12)$$

Full count sketch overordnet

i full count sketch bruger vi median tricket til at begrænse fejlsandsynlighed for estimatet vi får på f_x, vi kan bruge median tricket til at få 1 − δ sandsynlighed istedet for $\frac{3}{4}$.

For fejl sandsynlighed $\delta > 0$, gentager vi algorithmen $m = \lceil 12 \log(1/\delta) \rceil$ gange i parallel, så vi får estimater $\hat{f}_x^{(1)}, \ldots, \hat{f}_x^{(m)}$, $\hat{f}_x = \text{median}\left\{\hat{f}_x^{(1)}, \ldots, \hat{f}_x^{(m)}\right\}$. der gælder så ved brug af median tricket at

$$\Pr\left[\left|\hat{f}_x - f_x\right| \leq \varepsilon \|\mathbf{f}_{-x}\|_2\right] \geq 1 - e^{(-m/12)} \geq 1 - e^{(-\log(1/\delta))} = 1 - \delta$$

$$\Pr\left[\left|\hat{f}_x - f_x\right| \geq \varepsilon \|\mathbf{f}_{-x}\|_2\right] \leq \delta$$

Pladsforbrug:
$O(mk \log n) = O(\frac{4 \cdot 12 log(1/\delta)}{\epsilon^2}) = O\left(\frac{\log(1/\delta)}{\varepsilon^2} \log n\right)$ tid til at process et element i streamen er $O(\log(1/\delta))$.

2nd moment estimation (ekstra)

Det bruger count sketch kan vi så bruge til at estimere andet momentet, det kan vi dernæst bruge til at regne en varians ud.

vi vil estimere andet momentet

$$F_2 = \sum_{x \in [u]} f_x^2 = ||f||_2^2$$

$$||f||_p = (\sum_{x \in [u]} |f_x|^p)^{\frac{1}{p}}$$

Men vi skal bruge stærkere hashfunktioner:
en hash funktion $h : U \to [k]$ er q-universal hvis der for et bestemt q gælder: **q-wise independence $h(x_1), \ldots, h(x_q)$ are independent and uniformity h(x_i) uniform in [k].**

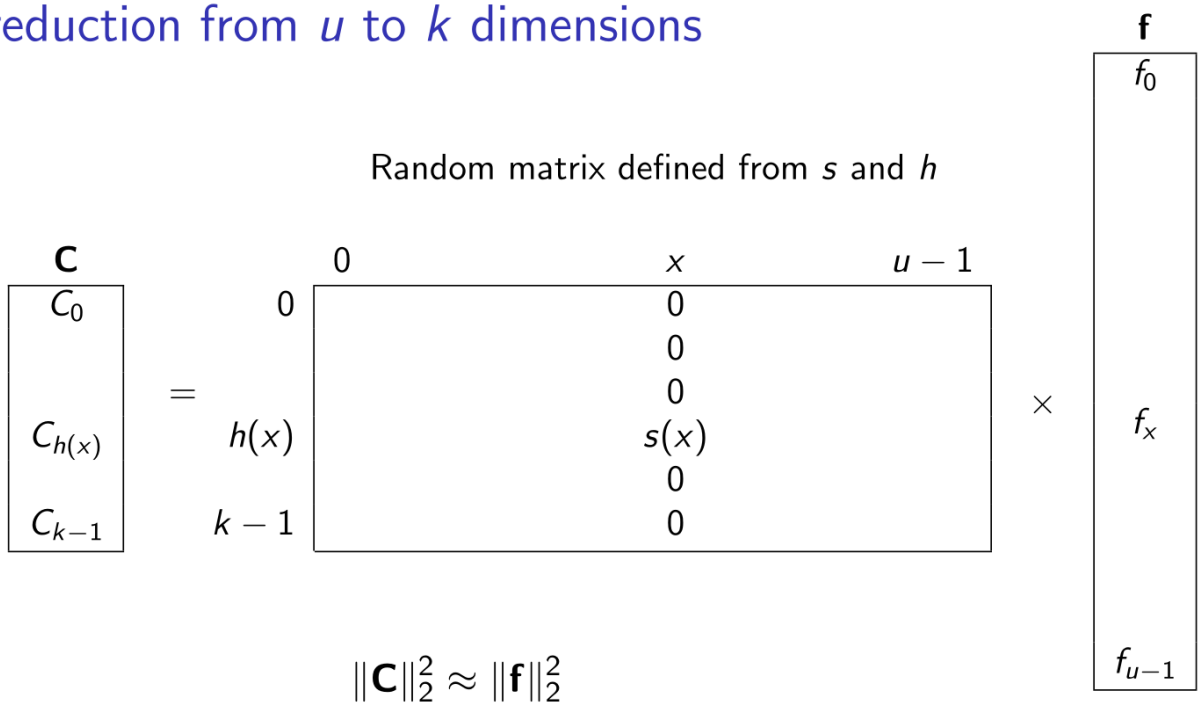
strong universality = 2-universality

for at estimere andet moment:

$$F_2 = \sum_{x \in [u]} f_x^2 = ||f||_2^2$$

s(x) er nu 4 universal istedet for 2 universal.
$k = \frac{8}{\epsilon^2}$
query er nu $\sum_{i \in [k]} C[i]^2$ som giver andet moment.

Linear map and distance preserving dimensionality reduction from u to k dimensions



$$\|\mathbf{C}\|_2^2 \approx \|\mathbf{f}\|_2^2$$