Count-min & Count-Sketch

Goal: given turnstile stream of items
find "Heavy Hitter":
most frequent items

$x \in \mathbb{R}^n$, x_i = # times i appears
Find $\{(i, x_i) \mid x_i$ "large"$\}$

$\hat{x} \in \mathbb{R}^n$ with $\|x - \hat{x}\|_\infty <$ bound

How small can error be?
Simplest: $\varepsilon \cdot \|x\|_1$ ← "Heavy Hitters"
Doable w/ $O(\frac{\log n}{\varepsilon})$ words

ℓ_∞/ℓ_1 Guarantee
Fancier: $\frac{1}{k}\|x - x_k\|_1$
where $x_k \in \mathbb{R}^n$ has k largest entries of x

Not $\varepsilon \cdot$ # items seen, but
$\varepsilon \cdot$ # less frequent items seen.

Still $O(k \log n)$ words
Count-min sketch, Cormode-Muthukrishnan

ℓ_∞/ℓ_2 guarantee
Even fancier: $\frac{1}{\sqrt{k}} \cdot \|x - x_k\|_2$

$O(k \log n)$, Count-Sketch,
Charikar, Chen, Farach-Colton
HW will show: $\frac{1}{\sqrt{k}}\|x - x_{2k}\|_2 \le \frac{1}{k}\|x - x_k\|_1$

How do these behave?

Zipf's law: i^{th} most common word in English appears $\sim \frac{1}{i}$ times

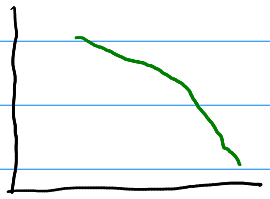
Power law: i^{th} largest $x_i \sim \frac{1}{i^\alpha}$ for some $\alpha > 0$.

Seen in URLs in web
Frequencies in music
Population in cities
:

[approximately] when it comes to largest entries

Log log plot: line

actual data, often



Similar behavior: lognormal, ——

$$\|x\|_1 = \sum_{i=1}^{n} \frac{1}{i^\alpha} \sim \begin{cases} 1 & \text{if } \alpha > 1 \\ \log n & \text{if } \alpha = 1 \\ n^{1-\alpha} & \text{if } \alpha < 1 \end{cases}$$

$$\|x\|_2 = \sqrt{\sum \frac{1}{i^{2\alpha}}} \sim \begin{cases} 1 & \text{if } \alpha > \frac{1}{2} \\ \sqrt{\log n} & \text{if } \alpha = \frac{1}{2} \\ n^{\frac{1}{2}-\alpha} & \text{if } \alpha < \frac{1}{2} \end{cases}$$

What effect does $\|x - x_k\|$ have?

$$\|x - x_k\|_1 = \sum_{i=k+1}^{n} \frac{1}{i^\alpha} = \begin{cases} k^{1-\alpha} & \text{if } \alpha > 1 \\ \log \frac{n}{k} & \text{if } \alpha = 1 \\ n^{1-\alpha} & \text{if } \alpha < 1 \end{cases}$$

Suppose $\alpha > 1$. $\quad \frac{1}{2} < \alpha < 1$

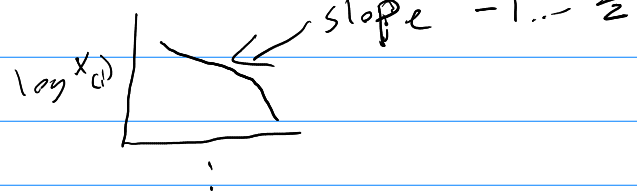
Heavy Hitters: $\frac{1}{k}\|x\|_1 \approx \frac{1}{k} \frac{n^{1-\alpha}}{k}$

Count-min: $\frac{1}{k}\|x - x_k\|_1 \approx \frac{1}{k^\alpha} \frac{n^{1-\alpha}}{k}$

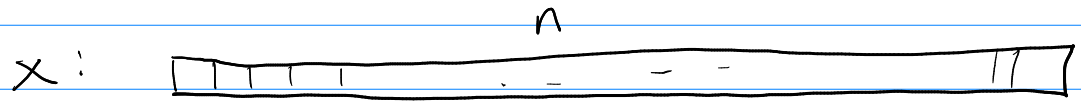
Count-sketch: $\frac{1}{\sqrt{k}}\|x - x_k\|_2 \approx \frac{1}{k^\alpha} \frac{1}{k^\alpha}$

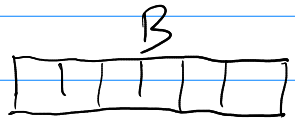
Count-sketch much better if $\frac{1}{2} < \alpha < 1$.

most common regime.



How they work



Hash to 

$B = O(k)$

$h: [n] \to [B]$

How to deal with collisions?

$$Y_j = \sum_{\substack{u \in [n] \\ h(u)=j}} x_u$$

Estimate $\hat{x}_u$ as $Y_{h(u)}$

In strict turnstile ($x_u \geq 0 \ \forall u$):

$$\hat{x}_u \geq Y_{h(u)}$$

In general,

$$E[\hat{x}_u - x_u] = E\left[\sum_{u' \neq u} x_{u'} \cdot 1_{h(u')=h(u)}\right]$$

$$= \sum_{u' \neq u} x_{u'} \cdot Pr[h(u') = h(u)]$$

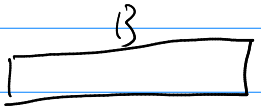
$$\leq \|x\|_1 \cdot \frac{1}{B}$$ w/ Pairwise independence

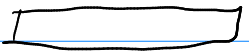
So if $B = 4k$, $|\hat{x}_u - x_u| \leq \frac{\|x\|_1}{k}$ w. P. $3/4$

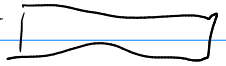
$|\hat{x}_u - x_u| \le \frac{\|x\|_1}{k}$ w.p. $\frac{3}{4}$ for fixed u.

Want $\|\hat{x}_u - x_u\| \le$ " , i.e. all u.

Repeat R times:

h_1: $y^{(1)} =$ 

h_2: $y^{(2)} =$ 

h_R: $y^{(R)} =$ 

Each $\hat{x}_u^{(r)}$ is close to x_u w/ $\frac{3}{4}$ prob.
Different r are independent.
How to combine?

Strict turnstile: $\hat{x}_u^{(r)} \ge x_u \quad \forall r$

Set $\hat{x}_u = \min_r \hat{x}_u^{(r)}$

works if any of R work. Failure prob 4^{-R}
$R = \log n \Rightarrow$ all n $\hat{x}_u$ work.

Each r uses $O(k)$ space
$\Rightarrow O(k \log n)$ total

to solve heavy hitters

Extensions

- What about nonstrict turnstile?
  - still $|\hat{x}_u^{(r)} - x_u| \le \frac{\|x\|_1}{K}$ w. p. $3/4$.
  - but $\min \hat{x}_u^{(r)}$ is bad.
    Replace min w/ median.
    Same asymptotics, worse constant.

- bound w/ $\frac{\|x - x_K\|_1}{K}$

$B = O(K)$

Y_j: 

Split x into x_K: largest K terms
$x - x_K$: rest

Most likely:
- $h(u)$ will not collide w/ any of top K [pr. $1 - \frac{K}{B}$]
- error from rest $\le \frac{\|x - x_K\|_1}{K}$ [as before]

- bound w/ $\frac{\|x - x_K\|_2}{\sqrt{K}}$

problem: what if all $x_u = 1$?

$\forall j$, $Y_j \approx \frac{n}{B} \Rightarrow \hat{x}_u \approx \frac{n}{B} = \Theta\left(\frac{\|x\|_1}{K}\right)$

$|\hat{x}_u - 1| >> \frac{\|x\|_2}{\sqrt{K}} = \sqrt{\frac{n}{K}}$

Change to

$$y_i^{(r)} = \sum_{u:\, h(u)=i} x_u \cdot s_r(u)$$

for $s_r : [n] \to \{\pm 1\}$

$$\hat{x}_u^{(r)} = y_{h(u)}^{(r)} \cdot s_r(u)$$

$$\hat{x}_u = \underset{r}{\text{median}}\ \hat{x}_u^{(r)}.$$

Analysis:

$$\mathbb{E}\left[(\hat{x}_u^{(r)} - x_u)^2\right] = \mathbb{E}\left[\left(\sum_{u' \neq u} \mathbb{1}_{h(u')=h(u)} \cdot x_{u'} \cdot s(u) \cdot s(u')\right)^2\right]$$

w/ Pairwise indep. of s, cross terms cancel!

$$= \sum_{u' \neq u} x_{u'}^2 \cdot \Pr[h(u')=h(u)]$$

$$= \frac{\|x\|_2^2}{B}$$

Hence $|\hat{x}_u^{(r)} - x_u| \leq \frac{\|x\|_2}{\sqrt{B}}$ w. p. $3/4$.

Rest, same as before.