Today:

- A few more comments on the AMS sketch
- Amplifying the probability of a good estimate
- ~~Chernoff bounds~~
- ~~Constructing k-wise independent hash functions~~

} Next time

Review: AMS sketch for $F_2 = \sum_x f^2(x)$

↑ number of occurrences (aka. frequency) of x

Last time:

Output $Z = \frac{\sum_{i=1}^{k} Y_i^2}{k}$

where each $Y_i = \sum_{x \in X} h_i(x) f(x)$

↑ hash function independently assigning each $x \in X$ to $\{-1, 1\}$ uniformly at random

Analysis last time:

If $k = \lceil 8/\varepsilon^2 \rceil$, then

$$\Pr\left[|Z - F_2| \geq \varepsilon F_2\right] \leq \frac{1}{9}$$

$\Updownarrow$ equivalently

$$\underbrace{(1-\varepsilon)F_2 \leq Z \leq (1+\varepsilon)F_2}_{\text{this kind of guarantee: multiplicative } (1+\varepsilon)\text{-approximation}} \quad \text{w.p.} \geq \frac{8}{9}$$

Implementing in small space as a streaming algorithm:

Each Y_i:

- Initially, $Y_i = 0$
- For each insertion $x \in X$

$$Y_i \leftarrow Y_i + h_i(x)$$

($f(x)$ increases by 1)

For each deletion $x \in X$

$$Y_i \leftarrow Y_i - h_i(x)$$

($f(x)$ decreases by 1)

This keeps $Y_i = \sum_{x \in X} h_i(x) f(x)$

Outputting estimate:

Still $Z = \frac{\sum_{i=1}^{k} y_i^2}{k}$

Space usage (ignoring hash functions): $O(1/\varepsilon^2)$

How about hash functions?

4-wise independence suffices to make our analysis of $\mathbb{E}[h(x)h(y)h(z)h(t)]$ valid (and $\mathbb{E}[h(x)h(y)]$)

} we get later to how to construct such functions in small space

---

How to make the probability of mult. $(1+\varepsilon)$-approximation higher?

Say, $\geq 1-\delta$ for $\delta \in (0, \frac{1}{3})$?

Compared to Count-Min Sketch (CMS):

- CMS: never underestimates, so take minimum of several basic estimators and it suffices that one of them is close

- AMS: can both overestimate and underestimate so taking maximum or minimum of several estimators won't work

---

Idea 1: Apply Chebyshev's inequality with higher number k of basic estimates we average

Last time via Chebyshev's: (bound on variance of Z)

$$\Pr\left[|Z - \underbrace{\mathbb{E}[Z]}_{=F_2}| \geq a\sqrt{\frac{2F_2^2}{k}}\right] \leq \frac{1}{a^2}$$

for any $a > 0$

To bound the probability by δ, set $a = \sqrt{\frac{1}{\delta}}$:

$$\Pr\left[|Z - F_2| \geq \sqrt{\frac{2}{\delta k}}F_2\right] \leq \delta$$

Next set $k = \left\lceil \frac{2}{\delta \varepsilon^2} \right\rceil$, to get

$$\Pr[|Z - F_2| \geq \varepsilon F_2] \leq \delta$$

to get multiplicative $(1+\varepsilon)$-approximation

Total space: $O\left(\frac{1}{\delta\varepsilon^2}\right)$

Overhead:

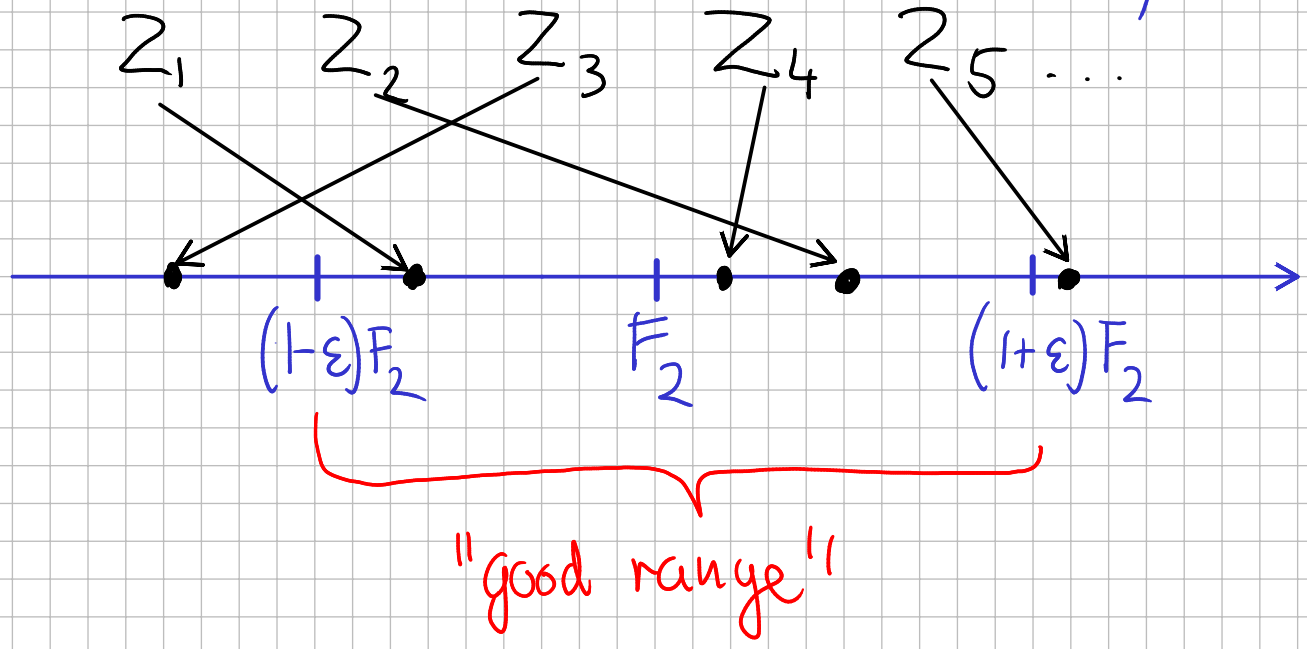
- CMS: $\sim \log(1/\delta)$
- here: $\sim 1/\delta$ ← Can we do better?

Idea 2: Take the median of multiple estimators Z_i for $k = \lceil 18/\varepsilon^2 \rceil$

Algorithm:

- Run k (to be set later) independent Z_i's (each outputs the mean of $k = \lceil 18/\varepsilon^2 \rceil$ independent copies of the basic estimator Y^2)
- Output the median of their estimates

Big picture: use "median of means"
(Idea I was just one big mean
or ≈ "mean of means")



- Z_i's independent & the probability each Z_i out of "good range" is at most $8/9$
- Intuition: For large k', vast majority of Z_i's should be in the good range
- The median not in "good range" only if half of estimates above the range or below it