DS-563/CS-543 2025-02-05

Today:
- Amplifying the probability of success of the AMS algorithm for F_2
- The Chernoff bound

Review of what we know already:

We will refer to copies of this algorithm as Z_i

- Algorithm that outputs a multiplicative $(1+\varepsilon)$-approximation to F_2 w.p. $\geq \frac{8}{9}$
- How: - Maintain $k = \lceil 18/\varepsilon^2 \rceil$ sketches

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

- Estimate: $Z = \frac{\sum_{i=1}^{k} Y_i^2}{k}$
- Space: $O(1/\varepsilon^2)$

Challange from last time:

Improve probability of outputting multiplicative $(1+\varepsilon)$-approximation to $1-\delta$ for $\delta \in (0, 1/9)$.

Attempt 1 (last time):
- Take the mean of more estimators Y_i^2

- Outcome via Chebyshev's inequality: needed $k = \Theta\left(\frac{1}{\delta \varepsilon^2}\right)$
- So total space $O\left(\frac{1}{\delta \varepsilon^2}\right)$

Bigger overhead than Count-Min Sketch, which required $O(\log(1/\delta))$

Attempt 2 (now): Run k' copies of Z_i & return the median of estimates

to be set later

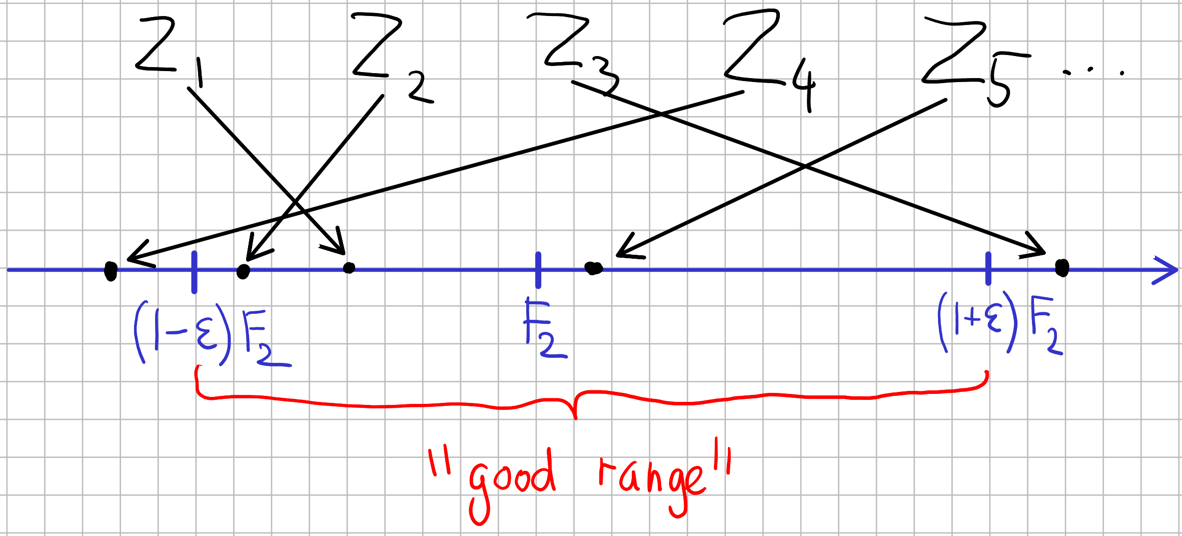
Pictorially:

median of means ← median of Z_j

means of estimates ← Z_1 = mean of $Y_{1,i}$ Z_2 = mean of $Y_{2,i}$ ... $Z_{k'}$ = mean of $Y_{k',i}$

estimates ← $Y_{1,1}^2 \; Y_{1,2}^2 \ldots Y_{1,k}^2$ $Y_{2,1}^2 \; Y_{2,2}^2 \ldots Y_{2,k}^2$ ... $Y_{k',1}^2 \; Y_{k',2}^2 \ldots Y_{k',k}^2$

sketches ← $Y_{1,1} \; Y_{1,2} \ldots Y_{1,k}$ $Y_{2,1} \; Y_{2,2} \ldots Y_{2,k}$ ... $Y_{k',1} \; Y_{k',2} \ldots Y_{k',k}$



each Z_i in this range w.p. $\geq \frac{8}{9}$

The median of Z_i's outside of the "good range" only if

① at least half of Z_i's $> (1+\varepsilon)F_2$

or

at least half of Z_i's $< (1-\varepsilon)F_2$

We'll show that probability of

$\circledast \equiv$ (at least half of Z_i's outside of the "good range")

is small. If $\circledast$ does not occur, then neither do ① and ②, and the median of Z_i's is a multiplicative $(1+\varepsilon)$-approximation.

The Chernoff bound

Setup: $X_1, X_2, \ldots, X_n$ – independent random variables that take on values in $[0,1]$

$$X \overset{\text{def}}{=} \sum_{i=1}^{n} X_i \quad \& \quad \mu \overset{\text{def}}{=} \mathbb{E}[X]$$

Claim: For $\varepsilon \in [0,1]$,

$$\Pr[X \leq (1-\varepsilon)\mu] \leq e^{-\varepsilon^2 \mu/2}$$

and

$$\Pr[X \geq (1+\varepsilon)\mu] \leq e^{-\varepsilon^2 \mu/3}.$$

For $\varepsilon \geq 1$,

$$\Pr[X \geq (1+\varepsilon)\mu] \leq e^{-\varepsilon \mu/3}.$$

Bounding the probability of ⊛:

For $i \in [k']$, $X_i = \begin{cases} 1 & \text{if } z_i \text{ in "good range"} \\ 0 & \text{otherwise} \end{cases}$

independent random variables

Define $X \overset{\text{def}}{=} \sum_{i=1}^{k'} X_i$. Then $\mathbb{E}[X] \geq \frac{8}{9}k'$.

$$\Pr[⊛] = \Pr\left[X \leq \frac{k'}{2}\right] = \Pr\left[X \leq \frac{9}{16} \cdot \frac{8}{9}k'\right]$$
$$\leq \Pr\left[X \leq \left(1 - \frac{7}{16}\right)\mathbb{E}[X]\right]$$

$$\overset{\text{Chernoff bound}}{\leq} e^{-\left(\frac{7}{16}\right)^2 \mathbb{E}[X]/3} \leq e^{-\left(\frac{7}{16}\right)^2 \cdot \frac{8}{9} k'/3} = e^{-\frac{49}{864} k'}$$

By setting $k' = \left\lceil \frac{864}{49} \ln(1/\delta) \right\rceil$, we get
$\Pr[\circledast] \leq e^{-\ln(1/\delta)} = \delta.$

Total space: $O(k' \cdot k) = O\left(\frac{1}{\varepsilon^2} \log(1/\delta)\right)$

overhead of $O(\log(1/\delta))$
(asymptotically better than $O(1/\delta)$)

Note: Nothing specific about F_2 here. The same strategy, taking the median of several estimators, would work for any estimator as long as it gives a "good" estimate with probability p, where p is a constant greater than $1/2$. The number of estimators needed is $O(\log(1/\delta))$.