CS 331, Fall 2024
Lecture 20 (11/6)

Today: – Basic boosting
– Chebyshev's inequality
– Mean/median boosting
– Morris counter

Basic boosting (Part VII, Section 3)

Today: improving the failure probability

$$\Pr[A \text{ does} \ldots] \leq \delta$$

Basic setting: you have A, succeeds @ outputting "good" x w.p. $\geq p$

e.g. Find Pivot $p = \frac{1}{2}$

Contention resolution $p = \frac{1}{n}$

What if p small?

If we can check whether x good...

- Run A T times
- Check all of them, return any good

We'll get good x except w.p.

$$(1-p)^T$$

Key fact: $\forall\, p \in (0, \frac{1}{2})$,

$$\frac{1}{4} \le (1-p)^{1/p} \le \frac{1}{e}$$

Thus, $T \gtrsim \frac{1}{p} \log\left(\frac{1}{\delta}\right)$

$$\Longrightarrow (1-p)^T \le \delta$$

What if we can't verify??

Motivation: running experiment to estimate x^*

design algo to output x, $\mathbb{E}[x] = x^*$

but...
- we care about $|x - x^*|$
- can't check (don't know x^*)

Today: how to prove things like

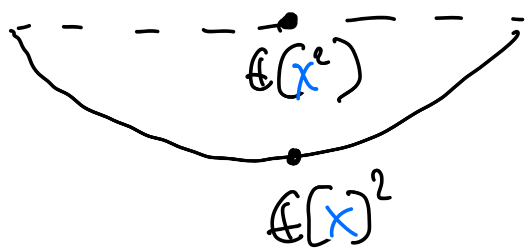
$$\Pr\Big[x \notin \underbrace{[x^* - R,\ x^* + R]}_{\text{"confidence interval"}}\Big] \le \delta$$

- Mean boosting: improve R
- Median boosting: improve δ

Chebyshev's inequality (Part VII, Section 3.1)

Key players:

$$\mathrm{Var}[X] := \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$$



(convexity)

$$\mathrm{Stdev}[X] := \sqrt{\mathrm{Var}[X]}$$ "spread"

$\boxed{\text{Aside}}$ Variance crash course

Reminder about $\mathbb{E}$: (no caveats!!!)

① $\mathbb{E}[cX] = c\,\mathbb{E}[X]$

② $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

$$\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] = \mathbb{E}\left[X^2\right]$$

$$- 2\mathbb{E}\left[X \cdot \underbrace{\mathbb{E}[X]}_{\text{constant, use ①}}\right] + \mathbb{E}[X]^2$$

$$= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathrm{Var}[X]$$

Also, we have

① $\mathrm{Var}[cX] = c^2 \mathrm{Var}[X]$

② $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

only if X, Y independent!

(Compare to $\mathbb{E}$, ① & ② both always true)

Our intuition: for reasonable δ,
pick $R \approx \text{stdev}(X)$

$\boxed{\text{Markov's inequality}}$

Let r.v. $X \geq 0$, then $\forall\, \delta \in (0,1)$,

$$\Pr\left[X \geq \frac{\mathbb{E}(X)}{\delta}\right] \leq \delta$$

e.g. $X < 10\,\mathbb{E}(X)$ except w.p. $\frac{1}{10}$

Proof: if not, let $\tau = \frac{\mathbb{E}(X)}{\delta}$,

$$\mathbb{E}(X) = \Pr(X \geq \tau) \cdot \tau + \Pr(X < \tau) \cdot 0$$
$$> \delta \cdot \tau > \mathbb{E}(X)$$

Chebyshev's inequality

Apply Markov's with $X \leftarrow (X - \mathbb{E}[X])^2$

$$\Pr\left[(X - \mathbb{E}[X])^2 \geq \frac{\mathrm{Var}[X]}{\delta}\right] \leq \delta$$

$$\Leftrightarrow |X - \mathbb{E}[X]| \geq \frac{\mathrm{stdev}[X]}{\sqrt{\delta}}$$

Thus we have shown except w.p. δ,

$$X \in \left[\mathbb{E}[X] - \frac{\mathrm{stdev}[X]}{\sqrt{\delta}}, \mathbb{E}[X] + \frac{\mathrm{stdev}[X]}{\sqrt{\delta}}\right]$$

e.g. $\delta = \frac{1}{4}$, confidence interval

$$\left[\mathbb{E}[X] - \underbrace{2\,\mathrm{stdev}[X]}_{\text{radius } R}, \mathbb{E}[X] + 2\,\mathrm{stdev}[X]\right]$$

Mean / median boosting (Part VII, Section 3.2)

Idea 1: Improve R (mean boosting)

Recall that $R \propto \mathrm{Stdev}(X)$

How to halve R? Decrease Var!

Basic fact:

Let $X_1, X_2, \ldots, X_k$ <u>independent</u> copies of X,

$$\overline{X} = \frac{1}{k} \sum_{i \in [k]} X_i$$

$$\mathrm{Var}(\overline{X}) = \frac{1}{k^2} \mathrm{Var}\left(\sum_{i \in [k]} X_i\right)$$

$$= \frac{k}{k^2} \mathrm{Var}(X) = \frac{1}{k} \mathrm{Var}(X)$$

Takeaway: if $k \approx \frac{1}{\varepsilon^2}$,

then our confidence gets εx better

$\mathrm{Var}(\overline{X}) = \varepsilon^2 \mathrm{Var}(X)$, $\sqrt{\varepsilon^2}$ x for stdev.

Idea 2: Improve δ (median boosting)

Suppose r.v. X has

$\Pr(X \in I) \geq \frac{3}{4}$ for interval $I \subset \mathbb{R}$

Claim: let $\hat{X} = \mathrm{median}(X_1, X_2, \ldots, X_k)$

$k \geq 12 \log\left(\frac{1}{\delta}\right)$

Then, $\hat{X} \in I$ except w.p. δ

How to prove?

Step 1: it's enough to show $\geq \frac{k}{2}$ copies $\in I$

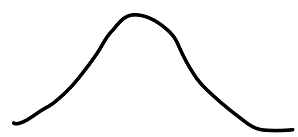
e.g. if $\hat{x} \notin I$ to the right, clearly $< \frac{k}{2} \in I$

$\hat{x} < \frac{1}{2}$ elements

Our goal: we have coin that H w.p. $\geq \frac{3}{4}$

toss k coins, $\geq \frac{k}{2}$ are H.

Step 2: "binomial concentration"



Suppose $i \geq \frac{k}{2}$ copies miss I

$$\Pr\left[\geq \frac{k}{2} \text{ are } T\right] = \sum_{i=k/2}^{k} \binom{k}{i} (1-p)^i p^{k-i}$$

$$\leq \sum_{i=k/2}^{k} \binom{k}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{k-i}$$

Observe if $i \leftarrow i+1$, $\underbrace{\binom{k}{i}}_{\text{decreases}} \underbrace{\left(\frac{1}{4}\right)^{i} \left(\frac{3}{4}\right)^{k-i}}_{\text{decreases } 3x}$

Thus geom sequence. Enough to get first term

$$\binom{k}{k/2}\left(\frac{1}{4}\right)^{k/2}\left(\frac{3}{4}\right)^{k/2} = O(\delta)$$

$$\leq 2^k \cdot \left(\frac{3}{16}\right)^{k/2} = \left(\frac{3}{4}\right)^{k/2}$$

Good enough if $k = O\left(\log\left(\frac{1}{\delta}\right)\right)$!

3-stage plan:

- Basic interval via $\mathrm{Var}[X]$
  $R = 2\,\mathrm{stdev}(X)$, $\delta = 1/4$
- Mean boost $\quad \times \frac{4}{\varepsilon^2}$
  $R = \varepsilon\,\mathrm{stdev}[X]$, $\delta = 1/4$
- Median boost $\quad \times 12\log\left(\frac{1}{\delta}\right)$
  $R = \varepsilon\,\mathrm{stdev}(X)$, $\delta = \delta$

Morris counter (Part VII, Section 3.3)

Goal: build data structure to store counter $i \leftarrow 0$

- Count(): $i \leftarrow i+1$
- Report(): return i

How much space needed to Count() n times?

Trivial: $O(\log(n))$ space (store i)

Today: $O(\log\log(n))$ space (randomized)

Motivation:
- Count many large things
- Constant factors help a lot!
- Web crawler, search engine, etc.
- Morris: spellchecker, needed 26^3 trigram counters

Algo: $X \leftarrow 0$,

Count(): w.p. 2^{-X}, $X \leftarrow X+1$

Report(): return $2^{X}-1$

It works ????

Intuition: $X \approx \log(i+1)$

Increase once in each bucket

$i \in (3,4]$, $[5,8]$, $[9,16]$, ...

Let X_k = value of X after $i=k$

Claim: $\mathbb{E}[2^{X_k}] = k+1$ $\quad \forall k \in [n]$

$\mathrm{Var}[2^{X_k}] \leq 2k^2$

3-step plan: gives estimate in

$$[(1-\varepsilon)n, (1+\varepsilon)n] \text{ w.p. } \geq 1-\delta$$

(NY22]: space $O\left(\log\log\frac{n}{\delta} + \log\frac{1}{\varepsilon}\right)$

Proof of $\mathbb{E}$: true when $k=0$. Induct!

$$\mathbb{E}\left(2^{X_{k+1}}\right)$$

$$= \sum_{j=0}^{\infty} \Pr[X_k = j] \cdot \overbrace{\left(2^j\left(1-\frac{1}{2^j}\right) + 2^{j+1}\left(\frac{1}{2^j}\right)\right)}^{=2^j+1}$$

$$= \sum_{j=0}^{\infty} \left(\Pr[X_k=j]\,2^j + \Pr[X_k=j]\right)$$

$$= \mathbb{E}\left(2^{X_k}\right) + 1. \blacksquare$$ (Var: see notes)