COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2024.
Lecture 13 (Midterm Review)

Midterm Logistics

- Thursday in class 10-11:15 am.
- Closed book, no calculator.
- Study guide posted under course schedule on the midterm line.
- Practice midterm posted in Moodle.

Midterm Format/Content

Rough Format:

- 1. Set of true/false or always/sometimes/never questions
- 2. Set of short answer questions.
- 3-4. Two multi-part 'problem set like' questions.
- 5. One bonus question.

Midterm Format/Content

Rough Format:

- 1. Set of true/false or always/sometimes/never questions
- 2. Set of short answer questions.
- 3-4. Two multi-part 'problem set like' questions.
- 5. One bonus question.

Content:

- See study guide for a detailed list.
- Should know key tools/analysis approaches. But don't need to memorize derivations from class.
- It will help to be familiar with homework problems, but don't need to memorize answers/derivations from them.

Midterm Studying

- I would focus on doing practice problems more than reviewing material.
- Definitely do the practice exam in semi-timed, closed note environment.
- Try to do as many other practice questions from the study guide, by redoing homework questions/problems given in class, etc.

Practice

A parallel computer consists of n processors and n memory modules. During a step, each processor sends a memory request to a random module. A module that receives 1 or 2 requests satisfies its requests; modules that receive more than two requests will not satisfy them. What is the expected number of modules that satisfy their requests?

Practice

A parallel computer consists of n processors and n memory modules. During a step, each processor sends a memory request to a random module. A module that receives 1 or 2 requests satisfies its requests; modules that receive more than two requests will not satisfy them. What is the expected number of modules that satisfy their requests?

X = # modules satisfy requests

$X = X_1 + \dots + X_n$

$X_i = 1$ if module i satisfies, 0 otherwise

$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$

b_i = # balls in bin i

$\Pr(b_i = 1) + \Pr(b_i = 2)$

$= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} + \frac{\binom{n}{2}}{n^2}\left(1 - \frac{1}{n}\right)^{n-2}$

$\mathbb{E}[X] = n\left(1 - \frac{1}{n}\right)^{n-1} + \frac{\binom{n}{2}}{n}\left(1 - \frac{1}{n}\right)^{n-2}$

Practice

You store n items in a hash table with 2n buckets using a fully random hash function. Give an upper bound on the maximum load in any bucket, which holds with probability at least $1 - 1/n$. Does the answer change significantly if you have n instead of $2n$ buckets?

$$\mathbb{E}[X] = n\underbrace{\left(1-\frac{1}{n}\right)^{n-1}}^{1/e} + \frac{\binom{n}{2}}{n}\left(1-\frac{1}{n}\right)^{n-2}$$

$n \to \infty$

$$\approx \frac{n}{e} + \frac{n(n-1)}{2n}\cdot\frac{1}{e} = n\left(\frac{1}{e}+\frac{1}{2e}\right)$$

Practice

$b_1 \ldots b_n$

You store n items in a hash table with 2n buckets using a fully random hash function. Give an upper bound on the maximum load in any bucket, which holds with probability at least $\boxed{1 - 1/n}$. Does the answer change significantly if you have n instead of $2n$ buckets?

$O(\sqrt{n})$

$O(n)$

$\mathbb{E}\, b_i = \frac{1}{2}$

max load $\approx O\left(\frac{\lg n}{\lg \lg n}\right)$

max load $\leq O(\log n)$

$\Pr(\text{max load} \geq t) = \Pr(b_1 \geq t \text{ or } b_2 \geq t \ldots \text{ or } b_n \geq t)$

$\overset{\text{union bound}}{\leq} n \cdot \Pr(b_i \geq t)$

we need to set t s.t. $\boxed{\Pr(b_i \geq t) \leq \frac{1}{n^2}}$

Then $\Pr(\text{max load} \leq t) \geq 1 - n \Pr(b_i \geq t)$

$\geq 1 - \frac{n}{n^2} \geq 1 - \frac{1}{n}$

Consider the scenario above, where you use linear probing instead of chaining. Does the expected look up time change significantly if you use n rather than $2n$ buckets?

$\Pr(b \geq t)$ where $b_j = 1$ if ball j lands in bin 1

$\Pr\left(\sum_{j=1}^{n} b_j \geq t\right)$ $b_j = 0$ otherwise

$$\Pr\left(\sum b_j \geq \delta\mu\right) \leq \exp\left(\frac{-\mu\delta^2}{2+\delta}\right)$$

$\mu = \mathbb{E}\sum b_j$ $\mu = \frac{1}{2}$ $\delta = 2t$

$$\Pr\left(\sum b_j \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}\cdot 2t^2}{2+2t}\right) \sim \exp(-t)$$

$t = c\log n$ for constant c

$\exp(-t) \leq \frac{1}{n^2}$

Practice

You have an algorithm that succeeds with probability $2/3$. You run it t times independently and would like to ensure that the probability that the algorithms fails on a majority (i.e., $> t/2$) of these trials is at most δ. How large should you set t?

Practice

You have an algorithm that succeeds with probability 2/3. You run it t times independently and would like to ensure that the probability that the algorithms fails on a majority (i.e., $> t/2$) of these trials is at most δ. How large should you set t?

Practice

$Y = \#\text{failures} \quad \mathbb{E}[Y] = t/3$

You have an algorithm that succeeds with probability 2/3. You run it t times independently and would like to ensure that the probability that the algorithm fails on a majority (i.e., $> t/2$) of these trials is at most δ. How large should you set t?

$$\Pr\left(|X - \mathbb{E}X| \geq \frac{t}{6}\right) \leq \frac{\mathrm{Var}(X)}{t^2} \leq \frac{t}{t^2} = \frac{1}{t}$$

→ median

→ BPP success prob can boosted

→ Chernoff $\quad t = O(\log(1/\delta))$

→ Bernstein

Chebyshev $\quad t = O(1/\delta)$

Markov $\quad \Pr(Y \geq \frac{t}{2}) = \Pr(Y \geq \mathbb{E}(Y) \cdot \frac{3}{2}) \leq \frac{2}{3}$

Practice

How to improve dependence on δ: use Bernstein

- use median trick $O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$

There is a slot machine that on each pull pays out a random prize between $[0, 1]$. After t plays roughly how good of an estimate of the expected value of playing do you have?

$1/\sqrt{t}$

$\mathbb{E}[X] = 0$ $\quad$ $\mathbb{E}[X] = 1/2$

$\mathrm{Var}[X] \leq 1$

$\bar{X} = \frac{1}{t}\sum_{i=1}^{t} X_i$

$\mathrm{Var}(\bar{X}) \leq \frac{1}{t^2}\sum_{i=1}^{t} \overbrace{\mathrm{Var}}^{\leq 1}(X_i)$

$\leq \frac{1}{t}$ $\quad$ $\mathrm{Var}(\bar{X}) = \mathbb{E}(\bar{X} - \mathbb{E}X)^2 = \frac{1}{t}$

to get ε error in estimating value w.p. $1-\delta$ how big should I set t? $t = O\left(\frac{1}{\delta\varepsilon^2}\right)$

by Chebyshev's: $\Pr(|\bar{X} - \mathbb{E}X| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} \leq \frac{1}{t\varepsilon^2}$

Practice

There is a slot machine that on each pull pays out a random prize between $[0, 1]$. After t plays roughly how good of an estimate of the expected value of playing do you have?

Practice

$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$

You would like to estimate the inner product of a vector $a \in \mathbb{R}^n$ with a vector $x \in \mathbb{R}^n$. You know a completely but don't know x. Describe how to obtain an estimate of the inner product without reading all of x via important sampling.

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \sum_{i=1}^{n} a_i x_i$$

$$M = \sum a_i x_i$$

sample i_j w.p. p_{i_j}

how to choose p_{i_j}?

$$m = \frac{1}{t} \sum_{j=1}^{t} a_{i_j} x_{i_j} \cdot \frac{1}{p_{i_j}}$$

$$Var(m) = \frac{1}{t} Var\left(a_{i_j} x_{i_j} \cdot \frac{1}{p_{i_j}}\right) = \frac{1}{t} \mathbb{E}\left(a_{i_j} x_{i_j} \frac{1}{p_{i_j}} - \mathbb{E}[\ldots]\right)^2$$

$$\leq \frac{1}{t} \mathbb{E}\left(a_{i_j}^2 x_{i_j}^2 \frac{1}{p_{i_j}^2}\right)$$

Practice

You would like to estimate the inner product of a vector $a \in \mathbb{R}^n$ with a vector $x \in \mathbb{R}^n$. You know a completely but don't know x. Describe how to obtain an estimate of the inner product without reading all of x via important sampling.

$$\mathrm{Var}(m) \leq \frac{1}{t} \mathbb{E}\left[a_{i_j}^2 x_{i_j}^2 \frac{1}{p_{i_j}^2}\right]$$

$$= \frac{1}{t} \sum_{j=1}^{n} p_j \cdot a_j^2 x_j^2 \frac{1}{p_j^2}$$

$$\mathrm{Var}(m) \quad = \frac{1}{T} \sum_{j=1}^{n} \frac{a_j^2 x_j^2}{p_j}$$

opt: $p_j \propto \frac{a_j x_j}{\sum a_j x_j}$

$$= \frac{1}{t} \sum_{j=1}^{n} \frac{a_j^2 x_j^2}{\frac{a_j^2}{\|a\|_2^2}} = \frac{1}{t} \cdot \|a\|_2^2 \cdot \|x\|_2^2 \qquad p_j = \frac{a_j^2}{\sum a_j^2} = \frac{a_j^2}{\|a\|_2^2}$$

$\|a\|_\infty, \|x\|_\infty \leq M$

$$\min V = \frac{1}{T}\sum_{j=1}^{n} \frac{a_j^2 x_j^2}{p_j} = \frac{n}{t}\sum a_j^2 x_j^2$$

$$= \frac{n}{t} \cdot n \cdot m^4 = \frac{n^2 \cdot m^4}{t}$$

$$\left(\frac{n\, m^2}{\sqrt{t}}\right)$$

$$\frac{\partial V}{\partial p_1} = \quad = \frac{\partial V}{\partial p_n} \qquad \text{s.t. } p_j \in [0,1]$$

$$\sum p_j = 1$$

$$\simeq \frac{a_j^2 x_j^2}{p_j^2} \qquad \text{so set } p_j^2 \propto a_j^2 x_j^2$$

What could you get w/ uniform sampling?

$$a = \begin{bmatrix} \infty \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad x \begin{bmatrix} \infty \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbb{E}\left[\|AB - C\|_F^2\right]$$