COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2023.

Lecture 13 (Midterm Review)

Summary

Last Class:

$x-y$

- Introduced the idea of low-distortion embeddings and the JL Lemma. $\|x-y\|_2$ $\|y\|$
- Reduction of JL Lemma to the Distributional JL Lemma via union bound.
- We will finish the proof of the JL Lemma after the midterm. Ignore any practice questions on this topic.

This Class:

- Midterm review.

Midterm Format

Rough Outline: (subject to small changes)

1 pt for answer | 1 pt for explanation

25/30

- Question 1: 4-5 always, sometimes, nevers or true falses.
- Question 2: 3-4 short answers, sort of like quiz questions.
- Question 3-4: Multipart questions, similar to core competency problems.

6

- Question 5: Extra credit question. Similar to a harder core competency problem.

31/30

Content, Format, or Logistics Questions?

- Bernstein
- Chernoff
- bloom filter FPR

midtermconcepts.pdf

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

$= p - p^2$

$\{0,1\} - 0 : 1-p$

$1 : p$

$\binom{m}{2} \le \frac{m^2}{2}$

m^2

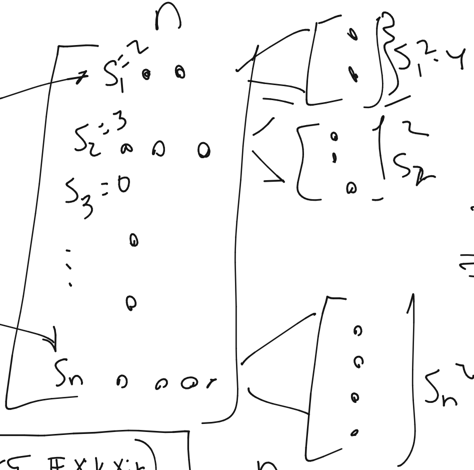
$\boxed{(1+x) < e^x}$

$\exp(\ldots k) \le \frac{1}{k}$

Questions

$\sum_{i=1}^{n} \mathbb{E} S_i^2$

X_1 ... X_n



no collision

$S_i^2 = \left(\sum X_{ij}\right)^2$

1 if item j hashes to bucket i

S_1^2

S_2

$\mathbb{E} \sum_k \sum_j X_{ij} \cdot X_{ik}$

$\sum_k \sum_j \mathbb{E} X_{ij} X_{ik}$

pairwise ind

$\sum_i \left(\sum_k \mathbb{E} X_{ik} + \sum_{i k \neq j} \mathbb{E} X_{ik} X_{jk} \right)$

$\sum_i \sum_k \mathbb{E} X_{ik}$

$\Pr(h(x_j) = h(x_k))$

2-universal

$= n^2 \frac{1}{n}$

$\sum_{i=1}^{n} \mathbb{E}\left[S_i^2\right]$ collisions

$\frac{1}{n^2} = \Pr\left(h(x_j) = i \text{ AND } h(x_k) = i\right)$

$\Pr(h(x) = h(y)) \le \frac{1}{n}$

$\Pr(h(x) = h(y) = 1) = 0$

$\Pr(h(x) = h(y) = 2) = \frac{2}{n}$

$h(2) = h(1) + 1 \mod (n)$

LSH

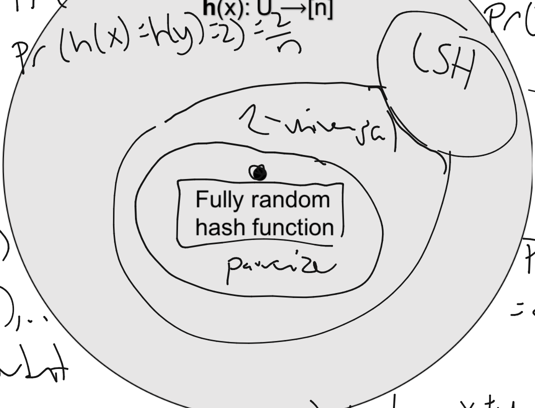
$\Pr(h(1) = 1 + h(2) = 1)$

$= 0$

hash tables

rarely

cms



$h(x) = \text{rand}(n)$

$h(x), h(y), h(z), \ldots$

are independent

$\Pr(h(x) = h(y))$

$= \sum_{i=1}^{n} \Pr(h(x) = i \text{ and } h(y) = i)$

$\sum_{i=1}^{n} \frac{1}{n^2} = \frac{1}{n}$

2-universal $\Pr(h(x) = h(y)) \le \frac{1}{n}$ $x \ne y$

pair-wise independent $\Pr(h(x) = i \text{ AND } h(y) = j) = \frac{1}{n^2}$

Concentration Bounds

$\Pr(X \geq t) \leq \frac{\mathbb{E}X}{t}$

$\mathrm{Var}(X_i) = p - p^2$

$p(1-p)$

$\leq 1/4$

Concentration Bound Requirements

$m=1$ $m=2$

Markov's	Chebyshev's	Chernoff	Bernstein
$\mathbb{E}[X]$	$\mathrm{Var}(X)$	$X = \sum X_i$ ind.	$X = \sum X_i$ ind.
$X \geq 0$	$\mathbb{E}(X)$	$X_i \in \{0,1\}$	$\lvert X_i \rvert \leq M$
		$\mathbb{E}[X] = \mu$	$\sigma^2 = \mathrm{Var}(X)$
		$\Pr(\lvert X - \mathbb{E}X \rvert \geq \delta\mu)$	t

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

$\frac{1}{10}$

$$\exp\left(\frac{-1/2^2 \cdot \mu}{2 + 1/2}\right)$$

3

$\delta = \frac{10}{\mu}$

Example Problems

3. Consider an algorithm $\mathcal{A}$ running in time $T(\mathcal{A})$, that with probability .6 outputs an estimate of the number of triangles in an input graph up to error ± 100, and with probability .4 outputs some bad estimate with worse error. Describe an algorithm that outputs an estimate of the number of triangles in an input graph up to error ± 100 with probability $\geq .99$ and runs in time $O(T(\mathcal{A}))$.

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$, $\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

Example Problems

3. Consider an algorithm $\mathcal{A}$ running in time $T(\mathcal{A})$, that with probability .6 outputs an estimate of the number of triangles in an input graph up to error ± 100, and with probability .4 outputs some bad estimate with worse error. Describe an algorithm that outputs an estimate of the number of triangles in an input graph up to error ± 100 with probability $\geq .99$ and runs in time $O(T(\mathcal{A}))$.

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^n X_i\right]$, for any $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2 \mu}{2+\delta}\right)$.

Example Problems

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.

   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$.

   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2 \mu}{2+\delta}\right)$.

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.

   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$. $1/5$

   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?

a) $X = \sum X_i \rightarrow \mathbb{E}X = \sum_{i=1}^{1000} \mathbb{E}X_i = \sum_{i=1}^{1000} p_i = 500$

b) $\Pr(|\sum X_i - 500| \geq \delta \cdot 500) \leq 2\exp\left(\frac{\delta^2 \mu}{2\delta}\right)$ $\boxed{\delta = 1/5}$

$\Pr(|\sum X_i - 500| \geq 100) \leq 2\exp\left(\frac{1/5^2 \cdot 500}{2 + 1/5}\right)$

$2\exp\left(\frac{20}{4}\right) \leq 2\exp(-5) \leq .01$

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,

$$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right).$$

ALWAYS, SOMETIMES, or NEVER:

$(1-\delta)^n = 1 - \delta n + \delta^2 \ldots + \delta \ldots$

2. $\Pr[\max(X_1, \ldots X_n) \geq t] \leq \sum_{i=1}^{n} \Pr[X_i \geq t]$ for any random variables $X_1, \ldots, X_n$.

$= \Pr(X_1 \geq t \text{ OR } X_2 \geq t \text{ OR} \ldots X_n \geq t) \leq \sum \Pr(X_i \geq t)$

union bound

Alg which answers middle ?s with prob $1-\delta$

Use this to answer n questions

(c) $\Pr[\mathbf{X} = s \cap \mathbf{Y} = t] = \Pr[\mathbf{X} = s] \cdot \Pr[\mathbf{Y} = t]$.

✓ independent

JL

$\Pr(\text{all questions get correct}) \geq 1 - n\delta$

$1 - \Pr(\text{fails at least one})$

union bound $\leq n \cdot \delta = \sum_{i=1}^{n} \Pr(\text{fails question } i)$