COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2022.

Lecture 13

Logistics

Markovs , Chebyshev's

- The exam is this Thursday in class.
- Closed book, no calculator (will be designed so neither are needed).
- My office hours are **today at 2:30pm in LGRC A215** and **tomorrow at 4:00pm in CS 140**.
- Suggested studying approach: Review the study guide to get a sense of what you need to know, and then mostly focus on doing practice questions. Review slides only as needed.
- The very last topic on the study guide, high dimensional geometry, will not be covered.

– No SimHash , HyperLogLog

Midterm Format

Rough Outline: (subject to changes)

- Question 1: 4-5 True/False questions.
- Question 2: 4-5 short answers, sort of like quiz questions.
- Question 3: 4-5 part question on analyzing an algorithm. Similar in style to but easier than a homework question.
- Question 4: Challenging 4-5 part question on analyzing an algorithm – more similar to a homework question.
- Question 5: Extra Credit. 4-5 part question with limited proofs on the Johnson-Lindenstrauss lemma.

Midterm Format

Rough Outline: (subject to changes)

- Question 1: 4-5 True/False questions.
- Question 2: 4-5 short answers, sort of like quiz questions.
- Question 3: 4-5 part question on analyzing an algorithm. Similar in style to but easier than a homework question.
- Question 4: Challenging 4-5 part question on analyzing an algorithm – more similar to a homework question.
- Question 5: Extra Credit. 4-5 part question with limited proofs on the Johnson-Lindenstrauss lemma.

I encourage you to review the JL material as Q5 should not be too difficult if you know the outline of the JL lemma proof. Do not need to know details.

Questions

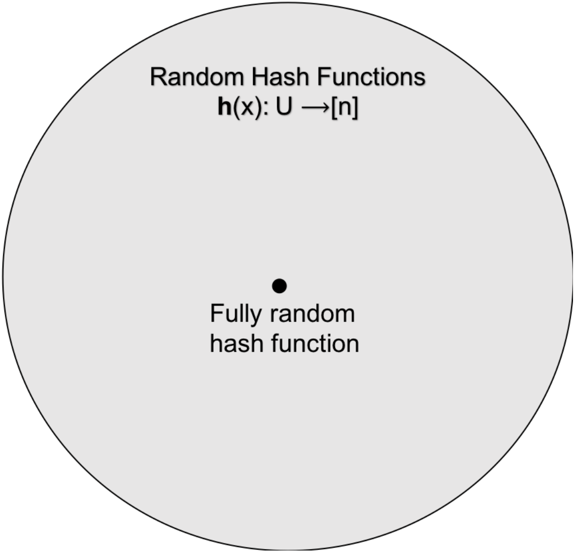
Content or Format Questions?

Bloom Filters: optimal # hash functions $\frac{\ln 2 \cdot m}{n}$

Questions

Questions

Random Hash Functions



Concentration Bounds

Concentration Bound Requirements

Markov's	Chebyshev's	Chernoff	Bernstein
$\Pr(X > t) \le \frac{\mathbb{E}X}{t}$ X non-negative $\mathbb{E}X$	$\mathbb{E}X$, $\mathrm{Var}(X)$ $\Pr(\lvert X - \mathbb{E}X\rvert \ge t) \le \frac{\mathrm{Var}(X)}{t^2}$	sum of independent $\{0,1\}$ n^2	sum of ind. vars $X_1, \dots, X_n$ $X_i \in [-m, m]$ $\mathrm{Var}(\sum X_i)$

$m = \frac{\log(\binom{n}{2}/\delta)}{\varepsilon^2} = O\left(\frac{\log(n/\delta)}{\varepsilon^2}\right)$

$= \sum \mathrm{Var}(X_i) \le 4m^2 = O(m^2)$

For JL Lemma:

χ-squared bound

for $m = \frac{\log(1/\delta)}{\varepsilon^2}$

$m = \frac{1}{\delta \varepsilon^2}$

$\|\Pi y\| \approx \|y\|$

union bound over $\binom{n}{2}$ y vectors

$\delta / \binom{n}{2}$

Sampling For Mean Estimation

Say I have n numbers $\underbrace{x_1, \ldots, x_n}$ all lying in $[-M, M]$ with mean $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. How can I estimate μ without reading all the numbers?

$Y_1 \ldots Y_t$ $\quad$ $Y_i = X_j$ w.p. $\frac{1}{n}$

independent

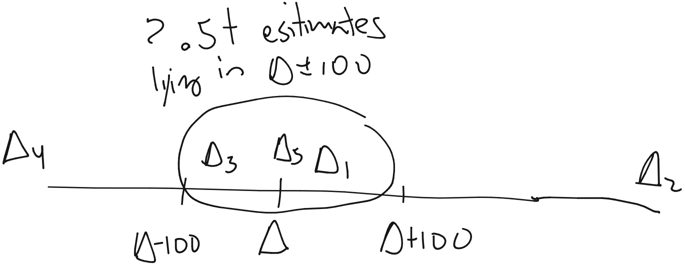
$t = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ $\quad$ $\pm \varepsilon M$

track s_i, x_i $\quad$ where $x_i = \arg\min_{j=1 \ldots n} h(x_j)$

$s_i = h(x_i)$

.21	5
.15	20
.20	

Johnson-Lindenstrauss Proof Outline



Median Trick

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 $O(T(\mathcal{A}))$.

run A t times independently $\rightarrow \Delta_1, \Delta_2 \ldots \Delta_t$

output $\tilde{\Delta} = \text{median}(\Delta_1, \ldots \Delta_t)$

A $\Pr(X > .5t) \leq$

B $\Pr(|\Delta - \tilde{\Delta}| \leq 100)$

1-B $\Pr(|\tilde{\Delta} - \Delta| \geq 100) \leq .01$

$X =$ # successful trials. $\mathbb{E}X = .6t$

If $X > .5t$ then median lies in $\Delta \pm 100$

1-A $\Pr(X \leq .5t) \leq .01$

$\Pr(X \leq \frac{5}{6}\mathbb{E}X) \leq \Pr(|X - \mathbb{E}X| \geq \frac{1}{6}\mathbb{E}X)$

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 $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

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 $O(T(\mathcal{A}))$.

$X = \#$ successful trials

$$\Pr\left(|X - \mathbb{E}X| \geq \frac{1}{6}\mathbb{E}X\right) \leq .01$$

$X = X_1 + \dots X_t$

$$\leq 2\exp\left(\frac{-1/6^2 \cdot .6t}{2+1/6}\right)$$

$$= 2\exp(-ct) \qquad t >> c \qquad t = \frac{1}{c}\cdot\log(2/\theta)$$

$$\leq .01 \qquad 2\exp(-c\cdot t) = 2\exp(-\log(2/\theta))$$

$$\mu = \mathbb{E}\left[\sum_{i=1}^{t} X_i\right] = \mathbb{E}[X] = .6t \qquad = 2\cdot\frac{.6}{2} = \theta$$

$\delta = 1/6$

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

$X = \sum X_i$

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

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$. $= 500$

   (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?

$\Pr\left(|X - \mathbb{E}X| \geq \frac{1}{6}\mathbb{E}X\right) \leq \frac{Var(X)}{\left(\frac{1}{6}\mathbb{E}X\right)^2} \leq \frac{.3t}{\frac{1}{36}\cdot .6^2 \cdot t^2} = \frac{c}{t} \leq \frac{1}{100}$

$Var(X)$ $\quad .6(1-.6) \leq .3t \quad t = 100c$

$X = \sum X_i \quad Var(X) = \sum_{i=1}^{t} Var(X_i)$

$\theta \quad t = O\left(\frac{1}{\theta}\right)$

$Var(X) = \sum Var(X_i) = \sum_{i=1}^{1000} p_i(1-p_i) \leq 250 \leq 500$

$\leq .25$

$t = O(\log(1/\theta))$

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 $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

$p_i - p_i^2$ $\quad 1 - 2p_i = 0$, $p_i = .5$

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 $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2 \mu}{2+\delta}\right)$.

Example Problems

ALWAYS, SOMETIMES, or NEVER:

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

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