COMPSCI ~~690RA~~ 614: Randomized Algorithms ~~and~~ w/ Apps to ~~Probabilistic~~ Data Analysis

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2024

Lecture 4

Logistics

- Problem Set ~~2~~ 1 is due next Wednesday 2/21 at 11:59pm.
- Most people think the lectures are ’just right’ or ’a bit too fast’. I’ll try to slow down a bit. If you feel that you are really falling behind, let me know.
- If you are confused on something please ask about it – certainly you are not the only one!

Summary

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

Last Time:

- Concentration bounds – Markov's and Chebyshev's inequalities.
- The union bound. $Pr(A_1 \cup \ldots \cup A_n) \leq \sum P(A_i)$
- Coupon collecting, statistical estimation.
- Randomized load balancing and ball-into-bins

Summary

Last Time:

- Concentration bounds – Markov's and Chebyshev's inequalities.
- The union bound.
- Coupon collecting, statistical estimation.
- Randomized load balancing and ball-into-bins

Today:

- Stronger concentration bounds for sums of independent random variables. I.e., exponential concentration bounds.
- Applications to balls-into-bins and linear probing analysis.

central limit theorem

Quiz Questions

Question 4

Not complete

Points out of 1.00

Flag question

Edit question

error $\approx \frac{1}{\sqrt{n}}$

Let's say I have two biased coins -- one hits heads with probability $1/2 + \epsilon$ and tails with probability $1/2 - \epsilon$. The other hits tails with probability $1/2 + \epsilon$ and heads with probability $1/2 - \epsilon$. \ multiarm bandit, hypothesis testing

How many independent flips of the coins must I perform to distinguish them from each other with probability at least $2/3$.

- a. $O(\log(1/\epsilon))$
- b. $O(1/\epsilon)$
- c. $O(1/\epsilon^2)$
- d. $O(1/\epsilon^4)$

Check

C_H = # tles heads hits heads

C_T = # tails coin

flip n times

$\mathbb{E}[C_H - C_T] = 2\epsilon n > 0$

$\mathbb{E}[C_H] - \mathbb{E}[C_T] = \frac{n}{2} + \epsilon n - \left(\frac{n}{2} - \epsilon n\right)$

$\Pr(C_H > C_T)$

$\Pr(C_H - C_T > 0)$

$\boxed{\mathrm{Var}(C_H - C_T) \leq \frac{n}{2}}$

$\mathrm{Var}[C_H - C_T] = \mathrm{Var}[C_H] + \mathrm{Var}[C_T]$

$\mathrm{Var}\left[\sum_{i=1}^{n} C_{H,i}\right]$

$= n \cdot \mathrm{Var}(C_{H,i}) \leq \frac{1}{4}$

$p(1-p) = \left(\frac{1}{2}+\epsilon\right)\left(\frac{1}{2}-\epsilon\right) = \frac{1}{4} - \epsilon^2$

$\leq \frac{n}{4}$

$\Pr(C_H - C_T < 0) \leq$

$\Pr\left(\left|(C_H - C_T) - \mathbb{E}(C_H - C_T)\right| \geq 2\epsilon n\right) \leq \frac{n/2}{4\epsilon^2 n^2} = \frac{1}{8\epsilon^2 n}$

Quiz Questions

$\mathrm{Var}(X+Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y)$

$X = 0$ w.p. $1-p$, either n or $-n$ w.p. $2p$

$\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}X)^2] = (1-p)\cdot 0 + 2p\cdot n^2$

Question 5

Not complete

Points out of 1.00

Flag question

Edit question

You roll a fair 6-sided die n times independently. You look at the difference between the number of times you rolled a "1" the number of times you rolled a "2". Roughly, how big do we expect this difference to be in magnitude? **Hint:** What is the variance of this difference?

- a. $\Theta(n)$
- b. $\Theta(\sqrt{n})$
- c. $\Theta(\log n)$
- d. $\Theta\left(\frac{\log n}{\log \log n}\right)$

Check

$\mathrm{Var}(X) = p\cdot n^2$

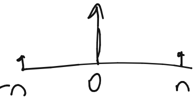
s.d: $\sqrt{p}\cdot n$

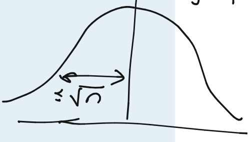
$\mathbb{E}[|X|] = 2p\cdot n$

$X_1 = \#$ ones

$X_2 = \#$ twos

$\mathrm{Var}(X_1 - X_2)$





$X_1 - X_2 = \sum_{i=1}^{n} D_i$

$D_i = 1$ w.p. $1/6$

$= -1$ w.p. $1/6$

0 w.p. $2/3$

$\mathrm{Var}(X_1 - X_2) \le \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$?

$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}\left(\sum D_i\right) \le n = \mathbb{E}[(X_1 - X_2)^2]$

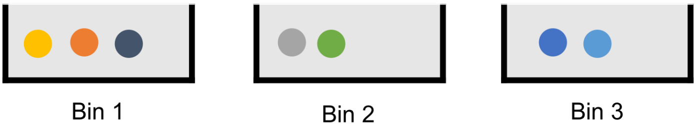
$\mathrm{Var}(D_i) \le 1$

$\mathbb{E}[|X_1 - X_2|]$

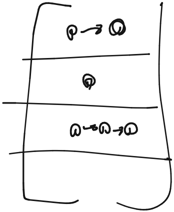
Balls Into Bins

Balls Into Bins

I throw m balls independently and uniformly at random into n bins. What is the maximum number of balls any bin?

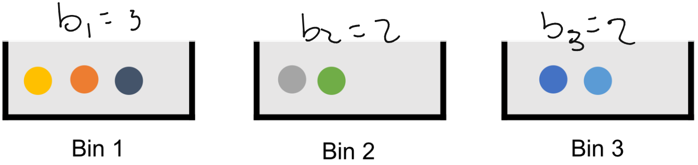


- Applications to randomized load balancing
- Analysis of hash tables using chaining.



Balls Into Bins

I throw m balls independently and uniformly at random into n bins. What is the maximum number of balls any bin?



- Applications to randomized load balancing
- Analysis of hash tables using chaining.
- **Direct Proof:** For any bin i, $\Pr[\mathbf{b}_i \geq \frac{c \ln n}{\ln \ln n}] \leq \frac{1}{n^{c-o(1)}}$. Thus, via union bound, the maximum load is exceeds $\frac{c \ln n}{\ln \ln n}$ with probability at most $\frac{1}{n^{c-1-o(1)}}$.

$n \cdot \frac{1}{n^{c-o(1)}}$

Balls Into Bins Via Chebyshev's Inequality

In our balls into bins analysis we directly bound
$\Pr[\mathbf{b}_i \geq k] \leq \left(\frac{e}{k}\right)^k \cdot \frac{1}{1-e/k}$.

$n = m$

Think Pair Share: Give an upper bound on this probability using Chebyshev's inequality. Hint: write $\mathbf{b}_i$ as a sum of n indicator random variables and compute $\mathrm{Var}[\mathbf{b}_i]$ and/or $\mathbb{E}[\mathbf{b}_i^2]$.

$Pr(b_i \geq k) \leq O(\frac{1}{k^2})$

$b_i = \sum_{j=1}^{n} X_j \qquad Var(b_i) = \sum Var(X_j) = n \cdot Var(X_j)$

$\downarrow$

$\frac{1}{n}\left(1 - \frac{1}{n}\right) \leq \frac{1}{n}$

$Var(b_i) \leq 1 \qquad \overset{\leq 1}{Var(b_i)} = \mathbb{E} b_i^2 - \underset{1}{\underset{\shortparallel}{[\mathbb{E} b_i]}}^2$

$\mathbb{E}(b_i) = 1 \qquad \mathbb{E} b_i^2 \leq 1 + 1 \leq 2$

$Pr[b_i \geq k] \leq Pr[|b_i - \mathbb{E} b_i| \geq k-1] \leq \overset{\leq 1}{\frac{Var(b_i)}{(k-1)^2}} \leq \frac{1}{(k-1)^2}$

$Pr[b_i \geq k] = Pr(b_i^2 \geq k^2] \leq \frac{2}{k^2}$

Balls Into Bins Via Chebyshev's Inequality

By Chebyshev's Inequality: $\Pr[\mathbf{b}_i \geq k] \leq \frac{2}{k^2}$.

Setting $k = c\sqrt{n}$, $\Pr\left[\mathbf{b}_i \geq c\sqrt{n}\right] \leq \frac{2}{c^2 n}$. So via a union bound:

$$\Pr\left[\max_{i=1,\ldots,n} \mathbf{b}_i \geq c\sqrt{n}\right] \leq n \cdot \frac{2}{c^2 n} \leq \frac{2}{c^2}.$$

Balls Into Bins Via Chebyshev's Inequality

By Chebyshev's Inequality: $\Pr[\mathbf{b}_i \geq k] \leq \frac{2}{k^2}$.

Setting $k = c\sqrt{n}$, $\Pr\left[\mathbf{b}_i \geq c\sqrt{n}\right] \leq \frac{2}{c^2 n}$. So via a union bound:

$$\Pr\left[\max_{i=1,\ldots,n} \mathbf{b}_i \geq c\sqrt{n}\right] \leq n \cdot \frac{2}{c^2 n} \leq \frac{2}{c^2}.$$

Upshot: Chebyshev's inequality bounds the maximum load by $O(\sqrt{n})$ with good probability, as compared to $O\left(\frac{\log n}{\log\log n}\right)$ for the direct proof. It is quite loose here.

$Var(X+Y) = Var(X) + Var(Y) + 2\,\mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)$

$\leq 2\sqrt{Var(X)Var(Y)}$ AMGM

$\leq Var(X) + Var(Y)$

$Y=X$

$Var(X+Y) \leq 2Var(X) + 2Var(Y)$

$Var(2X) = 4Var(X)$

Can't give lower bound b/c we could have $X=-Y$

Balls Into Bins Via Chebyshev's Inequality

By Chebyshev's Inequality: $\Pr[\mathbf{b}_i \geq k] \leq \frac{2}{k^2}$.

Setting $k = c\sqrt{n}$, $\Pr\left[\mathbf{b}_i \geq c\sqrt{n}\right] \leq \frac{2}{c^2 n}$. So via a union bound:

$$\Pr\left[\max_{i=1,\ldots,n} \mathbf{b}_i \geq c\sqrt{n}\right] \leq n \cdot \frac{2}{c^2 n} \leq \frac{2}{c^2}.$$

Upshot: Chebyshev's inequality bounds the maximum load by $O(\sqrt{n})$ with good probability, as compared to $O\left(\frac{\log n}{\log \log n}\right)$ for the direct proof. It is quite loose here.

Chebyshev's and Markov's inequalities are extremely valuable because they are very general – require few assumptions on the underlying random variable. But by using assumptions, we can often get tighter analysis.

Exponential Concentration Bounds

Higher Moments

Markov's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$. First moment.

$t = \sqrt{n}\, c$

Chebyshev's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X^2]}{t^2}$. Second moment.

$$\Pr(X \geq t) = \Pr(X^2 \geq t^2) \leq \frac{\mathbb{E}[X^2]}{t^2} = \frac{n}{t^2} = \frac{1}{c^2}$$

$$= \Pr(X^4 \geq t^4) \leq \frac{\mathbb{E}[X^4]}{t^4} \leq \frac{n^2}{t^4} = \frac{1}{c^4}$$

$X = \sum_{i=1}^{n} X_i$

$X_i = 1$ w.p. $\frac{1}{2}$

$X_i = -1$ w.p. $1/2$

$$\frac{\mathbb{E}X^6}{t^6} \leq \frac{n^3}{t^6}$$

$\mathbb{E}X = 0$

$\mathbb{E}X^2 = \mathrm{Var}(X) = \sum \mathrm{Var}(X_i) = n \cdot 1$

$$\mathbb{E}X^4 = \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^4\right] = n \cdot \underbrace{\sum \mathbb{E}X_i^4}_{n} + \underbrace{\binom{n}{2} \sum \mathbb{E}X_i^2\, \mathbb{E}X_j^2}_{n^2} + \cancel{\text{other stuff } X_i^2 X_j X_k \; X_j^3 X_i}$$

other stuff $\to 0$

$= \mathbb{E}[(X_1 + \cdots X_n)^4]$

$= \underline{\underline{n^2}}$

Higher Moments

Markov's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$. First moment.

Chebyshev's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X^2]}{t^2}$. Second moment.

Often (not always!) we can obtain tighter bounds by looking to higher moments of the random variable.

Higher Moments

Markov's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$. First moment.

Chebyshev's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X^2]}{t^2}$. Second moment.

Often (not always!) we can obtain tighter bounds by looking to higher moments of the random variable.

Moment Generating Function: Consider for any $z > 0$:

$$M_z(X) = e^{z \cdot X} = \sum_{k=0}^{\infty} \frac{z^k X^k}{k!}$$

by

Higher Moments

Markov's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$. First moment.

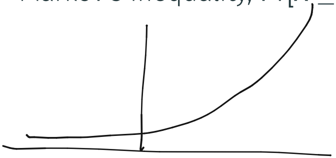
Chebyshev's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X^2]}{t^2}$. Second moment.

Often (not always!) we can obtain tighter bounds by looking to higher moments of the random variable.

Moment Generating Function: Consider for any $z > 0$:

$$M_z(X) = e^{z \cdot X} = \sum_{k=0}^{\infty} \frac{z^k X^k}{k!}$$

$e^{z \cdot t}$ is non-negative, and monotonic for any $z > 0$. So can bound via Markov's inequality, $\Pr[X \geq t] = \Pr[M_z(X) \geq e^{zt}] \leq \frac{\mathbb{E}[M_z(X)]}{e^{zt}}$.



Higher Moments

Markov's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$. First moment.

Chebyshev's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X^2]}{t^2}$. Second moment.

Often (not always!) we can obtain tighter bounds by looking to higher moments of the random variable.

Moment Generating Function: Consider for any $z > 0$:

$$M_z(X) = e^{z \cdot X} = \sum_{k=0}^{\infty} \frac{z^k X^k}{k!}$$

$e^{z \cdot t}$ is non-negative, and monotonic for any $z > 0$. So can bound via Markov's inequality, $\Pr[X \geq t] = \Pr[M_z(X) \geq e^{zt}] \leq \frac{\mathbb{E}[M_z(X)]}{e^{zt}}$.

By appropriately picking z and bounding $\mathbb{E}[M_z(X)]$, we can obtain a variety of exponential tail bounds. Typically require that X is a sum of bounded and independent random variables

Hoeffding, Chernoff, Bernstein, Azuma, ...
Berry Esseen...

Chernoff Bound (simplified version): Consider independent random variables $X_1, \ldots, X_n$ taking values in $\{0,1\}$ and let $X = \sum_{i=1}^{n} X_i$. Let $\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} X_i]$. For any $\delta \geq 0$

$\delta \geq 5$

$$\Pr(X \geq (1+\delta)\mu) \leq \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}}$$

not necessarily binomial

$$\leq \frac{e^{5\mu}}{6^{6\mu}} \qquad \delta = 5$$

$$\leq \frac{6^{5\mu}}{6^{6\mu}} \leq \frac{1}{6^{\mu}}$$

The Chernoff Bound

Chernoff Bound (simplified version): Consider independent random variables $X_1, \dots, X_n$ taking values in $\{0, 1\}$ and let $X = \sum_{i=1}^{n} X_i$. Let $\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} X_i]$. For any $\delta \geq 0$

$$\Pr\left(X \geq (1+\delta)\mu\right) \leq \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}}$$

Chernoff Bound (alternate version): Consider independent random variables $X_1, \dots, X_n$ taking values in $\{0, 1\}$ and let $X = \sum_{i=1}^{n} X_i$. Let $\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} X_i]$. For any $\delta \geq 0$

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

As δ gets larger and larger, the bound falls off exponentially fast.

Balls Into Bins Via Chernoff Bound

Recall that $\mathbf{b}_i$ is the number of balls landing in bin i, when we randomly throw n balls into n bins.

- $\mathbf{b}_i = \sum_{i=1}^{n} \mathbf{I}_{i,j}$ where $\mathbf{I}_{i,j} = 1$ with probability $1/n$ and 0 otherwise. $\mathbf{I}_{i,1}, \ldots \mathbf{I}_{i,n}$ are independent.

Balls Into Bins Via Chernoff Bound

Recall that $\mathbf{b}_i$ is the number of balls landing in bin i, when we randomly throw n balls into n bins.

- $\mathbf{b}_i = \sum_{i=1}^{n} \mathbf{I}_{i,j}$ where $\mathbf{I}_{i,j} = 1$ with probability $1/n$ and 0 otherwise. $\mathbf{I}_{i,1}, \ldots \mathbf{I}_{i,n}$ are independent.
- Apply Chernoff bound with $\mu = \mathbb{E}[\mathbf{b}_i] = 1$:

$$\Pr[\mathbf{b}_i \geq k] \leq \frac{e^k}{(1+k)^{(1+k)}}.$$

Balls Into Bins Via Chernoff Bound

Recall that $\mathbf{b}_i$ is the number of balls landing in bin i, when we randomly throw n balls into n bins.

- $\mathbf{b}_i = \sum_{i=1}^{n} \mathbf{I}_{i,j}$ where $\mathbf{I}_{i,j} = 1$ with probability $1/n$ and 0 otherwise. $\mathbf{I}_{i,1}, \ldots \mathbf{I}_{i,n}$ are independent.
- Apply Chernoff bound with $\mu = \mathbb{E}[\mathbf{b}_i] = 1$:

$$\Pr[\mathbf{b}_i \geq k] \leq \frac{e^k}{(1+k)^{(1+k)}}.$$

- For $k \geq \frac{c \log n}{\log \log n}$ we have:

$$\Pr[\mathbf{b}_i \geq k] \leq \frac{e^{\frac{c \log n}{\log \log n}}}{\left(\frac{c \log n}{\log \log n}\right)^{\frac{c \log n}{\log \log n}}} =$$