# COMPSCI 690RA: Randomized Algorithms and Probabilistic Data Analysis

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2024

Lecture 4

# Logistics

- Problem Set 2 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

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.

# Quiz Questions

**Question 4**

Not complete

Points out of 1.00

Flag question

Edit question

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 $1/2 - \epsilon$.

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

# Quiz Questions

**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?

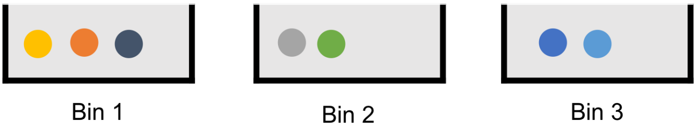
- a. $\Theta(n)$
- b. $\Theta(\sqrt{n})$
- c. $\Theta(\log n)$
- d. $\Theta\left(\frac{\log n}{\log \log n}\right)$

Check

# 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.
- **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 $\frac{1}{n^{c-1-o(1)}}$.

# Balls Into Bins Via Chebyshev's Inequality

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

**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 $\mathbb{E}[\mathbf{b}_i^2]$.

# 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.

**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

# The Chernoff Bound

**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 $\delta \geq 0$

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

**Chernoff Bound (alternate version):** Consider independent random variables $X_1, \ldots, X_n$ 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 $\delta \geq 0$

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

As $\delta$ gets larger and larger, the bound falls off exponentially fast.