COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Spring 2026.

Lecture 5

Logistics

- Problem Set 1 is due this Friday at 11:59pm.
- Make sure to start early and attend office hours if you feel stuck on any of the questions.
- No class on Thursday (Monday schedule)
- Quiz 3 will be posted immediately after class today and due next Monday at 8pm.
- Only 38/55 students took Quiz 2. Make sure you are completing the weekly quizzes.

Last Time

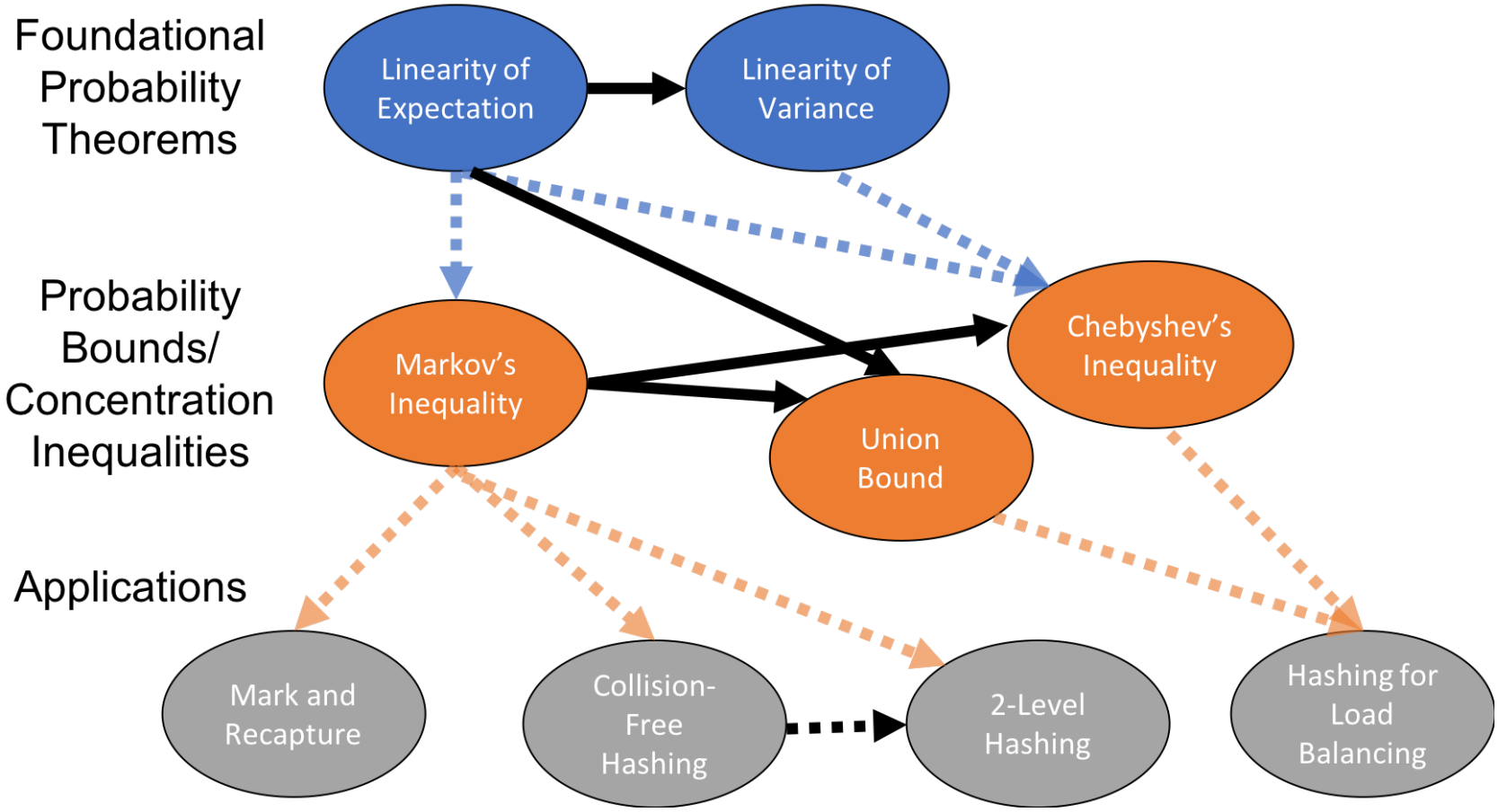
Last Class:

- Chebyshev's inequality and the law of large numbers.
- The union bound.
- Application to hashing for load balancing.
- Start on exploring higher moment bounds.

This Time:

- Higher moment bounds $\rightarrow$ exponential concentration bounds and the central limit theorem.

Concept Map


Foundational Probability Theorems
Linearity of Expectation
Linearity of Variance
Probability Bounds/ Concentration Inequalities
Markov's Inequality
Chebyshev's Inequality
Union Bound
Applications
Mark and Recapture
Collision-Free Hashing
2-Level Hashing
Hashing for Load Balancing

Quiz Question 1

12 1 point

My (not very popular) photo hosting service receives 2 download requests per day. Each download request is completed successfully with probability 0.95. Give an upper bound on the probability that my service fails to complete at least one request successfully. Hint: do not assume independence of the request completions.

Type your answer...

If the failures were independent: $1 - .95^2 = 0.0975$. Only a bit smaller than the upper bound of 0.1.

Quiz Question 2

10 Formula 1 point

The expected temperature on Saturday is $\mu = 75$ degrees. The variance of the temperature is $\sigma^2 = 5.1$ degrees. Give an upper bound on the probability that the temperature does not lie between 65 and 85 degrees. Give you answer to three decimal places.

Answer

Flipping Coins

We flip $n = 100$ independent coins, each are heads with probability 1/2 and tails with probability 1/2. Let H be the number of heads.

$$\mathbb{E}[\mathsf{H}] = \frac{n}{2} = 50 \text{ and } \mathsf{Var}[\mathsf{H}] = \frac{n}{4} = 25$$

Markov's:	Chebyshev's:	In Reality:
$\Pr(\mathsf{H} \geq 60) \leq .833$	$\Pr(\mathsf{H} \geq 60) \leq .25$	$\Pr(\mathsf{H} \geq 60) = 0.0284$
$\Pr(\mathsf{H} \geq 70) \leq .714$	$\Pr(\mathsf{H} \geq 70) \leq .0625$	$\Pr(\mathsf{H} \geq 70) = .000039$
$\Pr(\mathsf{H} \geq 80) \leq .625$	$\Pr(\mathsf{H} \geq 80) \leq .0278$	$\Pr(\mathsf{H} \geq 80) < 10^{-9}$

- H has a simple Binomial distribution, so can compute these probabilities exactly.
- Markov and Chebyshev's inequalities are extremely general and so can be loose – we would like to incorporate more information about the underlying distribution to get tighter bounds.

Tighter Bounds

Fourth Moment Bound:

$$\Pr(|X - \mathbb{E}[X]| \geq t) = \Pr\left((X - \mathbb{E}[X])^4 \geq t^4\right) \leq \frac{\mathbb{E}\left[(X-\mathbb{E}[X])^4\right]}{t^4}.$$

Chebyshev's:	4^{th} Moment:	In Reality:
$\Pr(H \geq 60) \leq .25$	$\Pr(H \geq 60) \leq .186$	$\Pr(H \geq 60) = 0.0284$
$\Pr(H \geq 70) \leq .0625$	$\Pr(H \geq 70) \leq .0116$	$\Pr(H \geq 70) = .000039$
$\Pr(H \geq 80) \leq .04$	$\Pr(H \geq 80) \leq .0023$	$\Pr(H \geq 80) < 10^{-9}$

- Can apply Markov's to $f(|X - \mathbb{E}[X]|)$ for any monotonically increasing function f and potentially generate new and tighter bounds.
- Why monotonic? $\Pr(|X - \mathbb{E}[X]| > t) = \Pr(f(|X - \mathbb{E}[X]|) > f(t))$.

H: total number heads in 100 random coin flips. $\mathbb{E}[H] = 50$.

Exponential Concentration Bounds

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

$$M_t(X) = e^{t\cdot(X-\mathbb{E}[X])} = \sum_{k=0}^{\infty} \frac{t^k (X - \mathbb{E}[X])^k}{k!}$$

- $M_t(X)$ is monotonic for any $t > 0$.
- Weighted sum of all moments, with t controlling how slowly the weights fall off (larger t = slower falloff).
- Choosing t appropriately lets one prove a number of very powerful exponential concentration bounds (exponential tail bounds).
- Chernoff bound, Bernstein inequalities, Hoeffding's inequality, Azuma's inequality, Berry-Esseen theorem, etc.
- We will not cover the proofs in this class – just the statements.

Bernstein Inequality

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ all falling in $[-M, M]$[-1,1]. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$ and $\sigma^2 = \mathrm{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathrm{Var}[X_i]$. For any $t \geq 0 s \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq s\sigma\right) \leq 2\exp\left(-\frac{s^2}{4}\right).$$

Assume that $M = 1$ and plug in $t = s \cdot \sigma$ for $s \leq \sigma$.

Compare to Chebyshev's: $\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq s\sigma\right) \leq \frac{1}{s^2}$.

- An exponentially stronger dependence on s!

Comparision to Chebyshev's

Consider again bounding the number of heads H in $n = 100$ independent coin flips.

Chebyshev's:	Bernstein:	In Reality:
$\Pr(\mathsf{H} \geq 60) \leq .25$	$\Pr(\mathsf{H} \geq 60) \leq .21$	$\Pr(\mathsf{H} \geq 60) = 0.0284$
$\Pr(\mathsf{H} \geq 70) \leq .0625$	$\Pr(\mathsf{H} \geq 70) \leq .005$	$\Pr(\mathsf{H} \geq 70) = .000039$
$\Pr(\mathsf{H} \geq 80) \leq .04$	$\Pr(\mathsf{H} \geq 80) \leq 4^{-5}$	$\Pr(\mathsf{H} \geq 80) < 10^{-9}$

Getting much closer to the true probability.

H: total number heads in 100 random coin flips. $\mathbb{E}[\mathsf{H}] = 50$.

The Chernoff Bound

A useful variation of the Bernstein inequality for binary (indicator) random variables is:

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

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

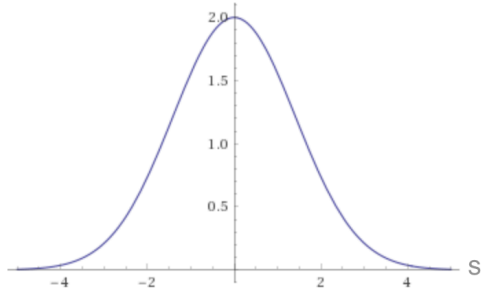
- As δ gets larger and larger, the bound falls of exponentially fast.
- How does this set up differ from a Binomial distribution?

Interpretation as a Central Limit Theorem

Bernstein Inequality (Simplified): Consider independent random variables $X_1, \ldots, X_n$ falling in [-1,1]. Let $\mu = \mathbb{E}[\sum X_i]$, $\sigma^2 = \mathrm{Var}[\sum X_i]$, and $s \leq \sigma$. Then:

$$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| \geq s\sigma\right) \leq 2\exp\left(-\frac{s^2}{4}\right).$$

Can plot this bound for different s:



Looks a lot like a Gaussian (normal) distribution.

$\mathcal{N}(0, \sigma^2)$ has density $p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}$.

Gaussian Tails

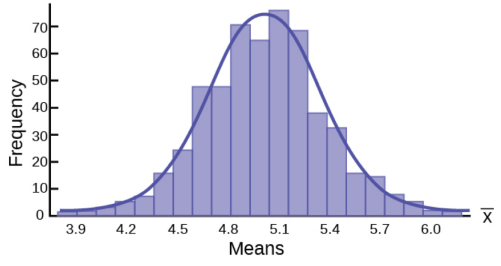
$$\mathcal{N}(0, \sigma^2) \text{ has density } p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}.$$

Exercise: Using this can show that for $X \sim \mathcal{N}(0, \sigma^2)$: for any $s \geq 0$,

$$\Pr\left(|X| \geq s \cdot \sigma\right) \leq 2e^{-\frac{s^2}{2}}.$$

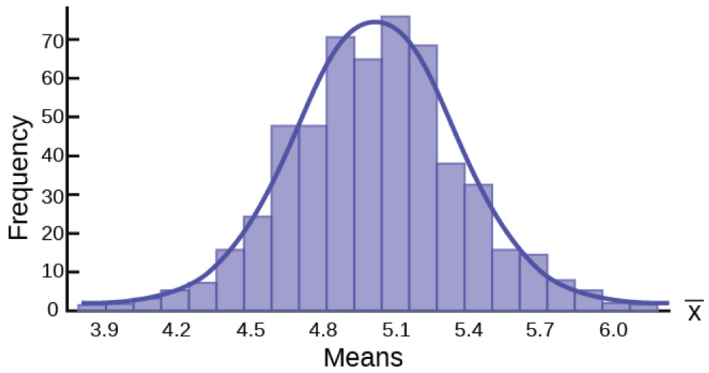
Essentially the same bound that Bernstein's inequality gives!

Central Limit Theorem Interpretation: Bernstein's inequality gives a quantitative version of the CLT. The distribution of the sum of *bounded* independent random variables can be upper bounded with a Gaussian (normal) distribution.



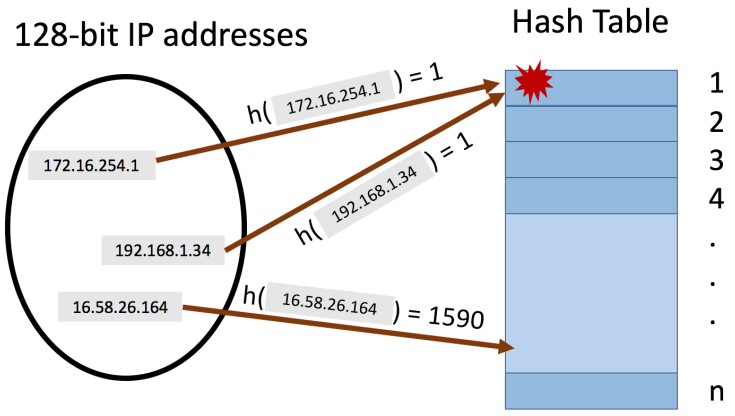
Central Limit Theorem

Stronger Central Limit Theorem: The distribution of the sum of *n bounded* independent random variables converges to a Gaussian (normal) distribution as *n* goes to infinity.



- Why is the Gaussian distribution is so important in statistics, science, ML, etc.?
- Many random variables can be approximated as the sum of a large number of small and roughly independent random effects. Thus, their distribution looks Gaussian by CLT.

Return to Random Hashing



We hash m values $x_1, \ldots, x_m$ using a random hash function into a table with $n = m$ entries.

- I.e., for all $j \in [m]$ and $i \in [m]$, $\Pr(\mathbf{h}(x_j) = i) = \frac{1}{m}$ and hash values are chosen independently.

What will be the maximum number of items hashed into the same location, with probability $\geq .99$?

$O(m)$ $O(\sqrt{m})$ $O(\log m)$ $O(1)$

Maximum Load in Randomized Hashing

Let S_i be the number of items hashed into position i and $\mathsf{S}_{i,j}$ be 1 if x_j is hashed into bucket i ($\mathsf{h}(x_j) = i$) and 0 otherwise.

$$\mathbb{E}[\mathsf{S}_i] = \sum_{j=1}^{m} \mathbb{E}[\mathsf{S}_{i,j}] = m \cdot \frac{1}{m} = 1 = \mu.$$

By the Chernoff Bound: for any $\delta \geq 0$,

$$\Pr(\mathsf{S}_i \geq 1 + \delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta \cdot \mu\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right)$$

m: total number of items hashed and size of hash table. $x_1, \ldots, x_m$: the items.
h: random hash function mapping $x_1, \ldots, x_m \to [m]$.

Maximum Load in Randomized Hashing

$$\Pr(\mathsf{S}_i \geq 1+\delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

For large δ, $\frac{\delta^2}{2+\delta} \approx \delta$. E.g., set $\delta = 20\log m$. Gives:

$$\Pr(\mathsf{S}_i \geq 20\log m + 1) \leq 2\exp\left(-\frac{(20\log m)^2}{2+20\log m}\right) \leq \exp(-18\log m) \leq \frac{2}{m^{18}}.$$

Apply Union Bound:

$$\Pr(\max_{i\in[m]} \mathsf{S}_i \geq 20\log m + 1) = \Pr\left(\bigcup_{i=1}^{m}(\mathsf{S}_i \geq 20\log m + 1)\right)$$

$$\leq \sum_{i=1}^{m} \Pr(\mathsf{S}_i \geq 20\log m + 1) \leq m \cdot \frac{2}{m^{18}} = \frac{2}{m^{17}}.$$

m: total number of items hashed and size of hash table. S_i: number of items hashed to bucket i. $\mathsf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

Upshot: If we randomly hash m items into a hash table with m entries the maximum load per bucket is $O(\log m)$ with very high probability.

- So, even with a simple linked list to store the items in each bucket, worst case query time is $O(\log m)$.
- Using Chebyshev's inequality could only show the maximum load is bounded by $O(\sqrt{m})$ with good probability (good exercise).
- The Chebyshev bound holds even with a pairwise independent hash function. The stronger Chernoff-based bound can be shown to hold with a *k-wise independent hash function* for $k = O(\log m)$.

Questions on Exponential Concentration Bounds?

This concludes the probability foundations part of the course – on to algorithms.