

COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2022.

Lecture 5

- Problem Set 1 is due this Friday at 11:59pm.
- Quiz question on class pacing:
 - Way too fast: 5.
 - A bit too fast: 45.
 - Just right: 53.
 - A bit too slow: 2.
 - Way too slow: 0.

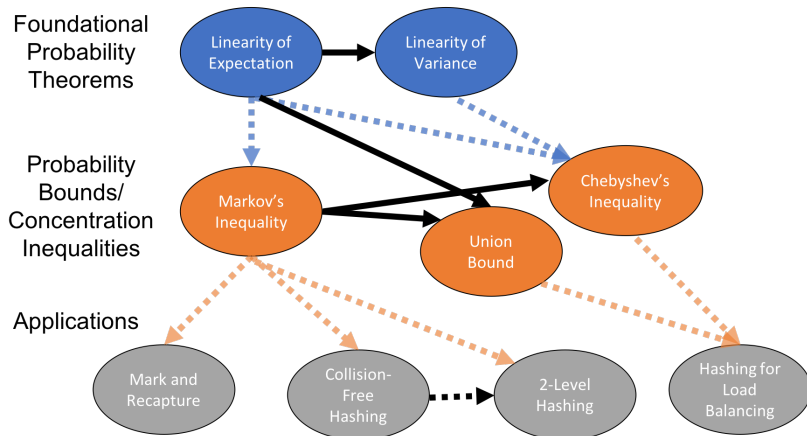
Last Class:

- 2-universal and pairwise independent hash functions.
- Chebyshev's inequality and the **law of large numbers**.
- The union bound.
- Application to hashing for load balancing.

This Time:

- Exponential concentration bounds and the **central limit theorem**.

Concept Map



Quiz Questions

My (not very popular) photo hosting service receives 5 download requests per day. Each download request is completed successfully with probability 0.98. 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.

Answer:

Check

Quiz Questions

The expected temperature on Saturday is $\mu = 75$ degrees. The variance of the temperature is $\sigma^2 = 12$ degrees. Give an upper bound on the probability that the temperature does not lie between 65 and 85 degrees.

Answer:

Check

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.

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

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

H has a simple Binomial distribution, so can compute these probabilities exactly.

Tighter Concentration Bounds

To be fair... Markov and Chebyshev's inequalities apply much more generally than to Binomial random variables like coin flips.

Can we obtain tighter concentration bounds that still apply to very general distributions?

- Markov's: $\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$. First Moment.
- Chebyshev's: $\Pr(|X - \mathbb{E}[X]| \geq t) = \Pr(|X - \mathbb{E}[X]|^2 \geq t^2) \leq \frac{\text{Var}[X]}{t^2}$. Second Moment.
- What if we just apply Markov's inequality to even higher moments?

A Fourth Moment Bound

Consider any random variable X :

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

Application to Coin Flips: Recall: $n = 100$ independent fair coins, H is the number of heads.

- Bound the fourth moment:

$$\mathbb{E}\left[(H - \mathbb{E}[H])^4\right] = \mathbb{E}\left[\left(\sum_{i=1}^{100} H_i - 50\right)^4\right] = \sum_{i,j,k,\ell} c_{ijkl} \mathbb{E}[H_i H_j H_k H_\ell] = 1862.5$$

where $H_i = 1$ if coin flip i is heads and 0 otherwise. Then apply some messy calculations...

- Apply Fourth Moment Bound: $\Pr(|H - \mathbb{E}[H]| \geq t) \leq \frac{1862.5}{t^4}$.

Tighter Bounds

Chebyshev's:

$$\Pr(H \geq 60) \leq .25$$

$$\Pr(H \geq 70) \leq .0625$$

$$\Pr(H \geq 80) \leq .04$$

4th Moment:

$$\Pr(H \geq 60) \leq .186$$

$$\Pr(H \geq 70) \leq .0116$$

$$\Pr(H \geq 80) \leq .0023$$

In Reality:

$$\Pr(H \geq 60) = 0.0284$$

$$\Pr(H \geq 70) = .000039$$

$$\Pr(H \geq 80) < 10^{-9}$$

Can we just keep applying Markov's inequality to higher and higher moments and getting tighter bounds?

- Yes! To a point.
- In fact – don't need to just apply Markov's to $|X - \mathbb{E}[X]|^k$ for some k . Can apply to any monotonic function $f(|X - \mathbb{E}[X]|)$.
- **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 $t > 0$:

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

- $M_t(\mathbf{X})$ 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.

Bernstein Inequality

Bernstein Inequality: Consider **independent** random variables X_1, \dots, X_n all falling in $[-M, M]$ **[-1,1]**. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$ and $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. 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 !

Comparison to Chebyshev's

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

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

Getting much closer to the true probability.

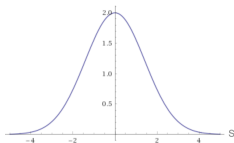
H : total number heads in 100 random coin flips. $\mathbb{E}[H] = 50$.

Interpretation as a Central Limit Theorem

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

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