COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Spring 2020.

Lecture 3

By Thursday:

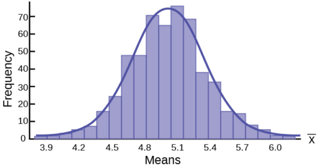
- Sign up for Piazza.
- Sign up for Gradescope (code on class website) and fill out the Gradescope consent poll on Piazza. Contact me via email if you don't consent to use Gradescope.
- First problem set will be **available in the next day or two, due 2/14.**

Last Class We Covered:

- Markov's inequality: the most fundamental concentration bound.
- Random hash functions, collision free hashing, and two-level hashing (analysis with linearity of expectation and Markov's inequality.)
- 2-universal and pairwise independent hash functions.
- Chebyshev's inequality and an elementary proof of the law of large numbers.

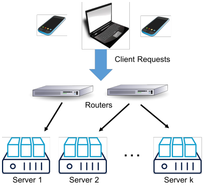
Today: We'll see even stronger concentration bounds than Chebyshev's inequality – exponential tail bounds.

- Will show a version of the central limit theorem.



First: We'll show learn about the union bound and apply it to randomized load balancing.

Randomized Load Balancing:



- n requests randomly assigned to k servers using a random hash function.
- Letting R_i be the number of requests assigned to server i, $\mathbb{E}[R_i] = \frac{n}{k}$ and we provision each server with the capacity to serve twice its expected load: $\frac{2n}{k}$ requests.
- What is the probability that a server exceeds its capacity?
- To apply Chebyshev's inequality, need to bound $\text{Var}[R_i]$.

Recall that we can write $\mathbf{R}_i$ as:

$$\mathbf{R}_i = \sum_{j=1}^{n} \mathbf{R}_{i,j} \quad \text{Var}[\mathbf{R}_i] = \sum_{j=1}^{n} \text{Var}[\mathbf{R}_{i,j}] \qquad \text{(linearity of variance)}$$

where $\mathbf{R}_{i,j}$ is 1 if request j is assigned to server i and 0 o.w.

$$\begin{aligned}\text{Var}[\mathbf{R}_{i,j}] &= \mathbb{E}\left[\left(\mathbf{R}_{i,j} - \mathbb{E}[\mathbf{R}_{i,j}]\right)^2\right] \\ &= \Pr(\mathbf{R}_{i,j} = 1) \cdot \left(1 - \mathbb{E}[\mathbf{R}_{i,j}]\right)^2 + \Pr(\mathbf{R}_{i,j} = 0) \cdot \left(0 - \mathbb{E}[\mathbf{R}_{i,j}]\right)^2 \\ &= \frac{1}{k} \cdot \left(1 - \frac{1}{k}\right)^2 + \left(1 - \frac{1}{k}\right) \cdot \left(0 - \frac{1}{k}\right)^2 \\ &= \frac{1}{k} - \frac{1}{k^2} \le \frac{1}{k} \implies \text{Var}[\mathbf{R}_i] \le \frac{n}{k}.\end{aligned}$$

n: total number of requests, k: number of servers randomly assigned requests, $\mathbf{R}_i$: number of requests assigned to server i.

Letting R_i be the number of requests sent to server i, $\mathbb{E}[R_i] = \frac{n}{k}$ and $\text{Var}[R_i] \leq \frac{n}{k}$.

Applying Chebyshev's:

$$\Pr\left(R_i \geq \frac{2n}{k}\right) \leq \Pr\left(|R_i - \mathbb{E}[R_i]| \geq \frac{n}{k}\right) \leq \frac{n/k}{n^2/k^2} = \frac{k}{n}.$$

- Overload probability is extremely small when $k \ll n$!
- Might seem counterintuitive – bound gets worse as k grows.
- When k is large, the number of requests each server sees in expectation is very small so the law of large numbers doesn't 'kick in'.

n: total number of requests, k: number of servers randomly assigned requests, R_i: number of requests assigned to server i.

What is the probability that the maximum server load exceeds $2 \cdot \mathbb{E}[R_i] = \frac{2n}{k}$. I.e., that some server is overloaded if we give each $\frac{2n}{k}$ capacity?

$$\Pr\left(\max_i (R_i) \geq \frac{2n}{k}\right) = \Pr\left(\left[R_1 \geq \frac{2n}{k}\right] \cup \left[R_2 \geq \frac{2n}{k}\right] \cup \ldots \cup \left[R_k \geq \frac{2n}{k}\right]\right) = \Pr$$

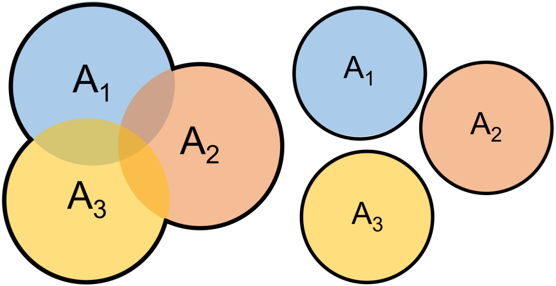
We want to show that $\Pr\left(\bigcup_{i=1}^k \left[R_i \geq \frac{2n}{k}\right]\right)$ is small.

How do we do this? Note that $R_1, \ldots, R_k$ are correlated in a somewhat complex way.

n: total number of requests, k: number of servers randomly assigned requests, R_i: number of requests assigned to server i. $\mathbb{E}[R_i] = \frac{n}{k}$. $\text{Var}[R_i] = \frac{n}{k}$.

Union Bound: For any random events $A_1, A_2, \ldots, A_k$,

$$\Pr\left(A_1 \cup A_2 \cup \ldots \cup A_k\right) \leq \Pr(A_1) + \Pr(A_2) + \ldots + Pr(A_k).$$



When is the union bound tight? When $A_1, \ldots, A_k$ are all disjoint.

On the first problem set, you will prove the union bound, as a consequence of Markov's inquality.

What is the probability that the maximum server load exceeds $2 \cdot \mathbb{E}[\mathbf{R}_i] = \frac{2n}{k}$. I.e., that some server is overloaded if we give each $\frac{2n}{k}$ capacity?

$$\Pr\left(\max_i(\mathbf{R}_i) \geq \frac{2n}{k}\right) = \Pr\left(\bigcup_{i=1}^{k}\left[\mathbf{R}_i \geq \frac{2n}{k}\right]\right)$$

$$\leq \sum_{i=1}^{k} \Pr\left(\left[\mathbf{R}_i \geq \frac{2n}{k}\right]\right) \qquad \text{(Union Bound)}$$

$$\leq \sum_{i=1}^{k} \frac{k}{n} = \frac{k^2}{n} \qquad \text{(Bound from Chebyshev's)}$$

As long as $k \leq O(\sqrt{n})$, with good probability, the maximum server load will be small (compared to the expected load).

n: total number of requests, k: number of servers randomly assigned requests, $\mathbf{R}_i$: number of requests assigned to server i. $\mathbb{E}[\mathbf{R}_i] = \frac{n}{k}$. $\mathrm{Var}[\mathbf{R}_i] = \frac{n}{k}$.

The number of servers must be small compared to the number of requests ($k = O(\sqrt{n})$) for the maximum load to be bounded in comparison to the expected load with good probability.

- There are many requests routed to a relatively small number of servers so the load seen on each server is close to what is expected via law of large numbers.
- **A Useful Exercise:** Given n requests, and assuming all servers have fixed capacity C, how many servers should you provision so that with probability $\geq 99/100$ no server is assigned more than C requests?

n: total number of requests, k: number of servers randomly assigned requests.

Questions on union bound, Chebyshev's inequality, random hashing?

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.

$$\mathbb{E}[\mathbf{H}] = \frac{n}{2} = 50 \text{ and } \mathrm{Var}[\mathbf{H}] = \frac{n}{4} = 25 \rightarrow s.d. = 5$$

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

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

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

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:

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

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

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

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

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 $|\mathbf{X} - \mathbb{E}[\mathbf{X}]|^k$ for some k. Can apply to any monotonic function $f(|\mathbf{X} - \mathbb{E}[\mathbf{X}]|)$.
- Why monotonic? $\Pr(|\mathbf{X} - \mathbb{E}[\mathbf{X}]| > t) = \Pr(f(|\mathbf{X} - \mathbb{E}[\mathbf{X}]|) > f(t))$.

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

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 the this class.

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 $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. 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!

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

Chebyshev's:	Bernstein:	In Reality:
$\Pr(H \geq 60) \leq .25$	$\Pr(H \geq 60) \leq .15$	$\Pr(H \geq 60) = 0.0284$
$\Pr(H \geq 70) \leq .0625$	$\Pr(H \geq 70) \leq .00086$	$\Pr(H \geq 70) = .000039$
$\Pr(H \geq 80) \leq .04$	$\Pr(H \geq 80) \leq 3^{-7}$	$\Pr(H \geq 80) < 10^{-9}$

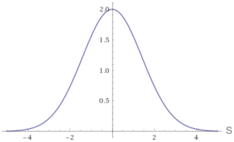
Getting much closer to the true probability.

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

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

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

Can plot this bound for different s:



Looks a lot like a Gaussian (normal) distribution.

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

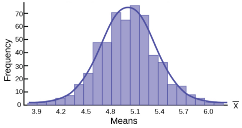
$$\mathcal{N}(0, \sigma^2) \text{ has density } p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{x^2}{2\sigma^2}}.$$

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

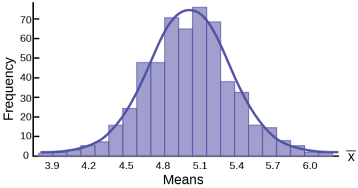
$$\Pr\left(|\mathbf{X}| \geq s \cdot \sigma\right) \leq O(1) \cdot e^{-\frac{s^2}{2}}.$$

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.



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.



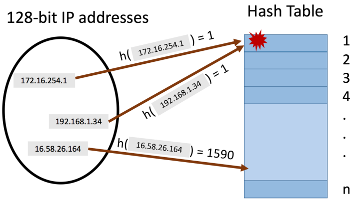
- Why is the Gaussian distribution is so important in statistics, science, ML, etc.?
- Many random variables can be approximated as the sum of a large number of small and roughly independent random effects. Thus, their distribution looks Gaussian by CLT.

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.



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 $i \in [n]$, $\Pr(\mathbf{h}(x) = i) = \frac{1}{m}$ and hash values are chosen independently.

What will be the maximum number of items hashed into the same location?

$O(n)$ $O(\log n)$ $O(\sqrt{n})$ $O(1/n)$

What will be the maximum number of items hashed into the same location? $O(\log m)$

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

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

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

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

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

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

Set $\delta = 20\log m$. Gives:

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

Apply Union Bound:

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

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

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

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.
- 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?

This concludes probability review/concentration bounds.