COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2021.

Lecture 4

- Problem set one posted this morning. Due next Friday 9/24, at 8pm.
- Class pacing: Just right – 53/87, A bit too fast – 29/87.
- If things feel way too fast, come to my office hours or send me an email and we can discuss how to make things more manageable.
- **Reminder:** My office hours are now Thursdays 9am on Zoom. Pratheba's in person office hours are Mondays 12pm-1pm in CS 207.

reminder: correct quiz points

Last Class:

- 2-Level Hashing Analysis (linearity of expectation and Markov's inequality)
- 2-universal and pairwise independent hash functions
- Start on hashing for load balancing.

2-universal hash functions are:

✓ a) more efficient to compute than fully independent hash functions.

✓ b) low collision prob.

~~c) are deterministic~~

~~d) zero collisions~~

h(x) = ax+b mod...

h(x)

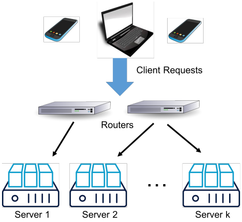
Last Class:

- 2-Level Hashing Analysis (linearity of expectation and Markov's inequality)
- 2-universal and pairwise independent hash functions
- Start on hashing for load balancing.

This Time:

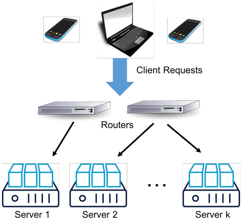
- Finish hashing for load balancing. Motivating:
  - Stronger concentration inequalities: Chebyshev's inequality, exponential tail bounds, and their connections to the law of large numbers and central limit theorem.
  - The union bound to bound the probability that one of multiple possible correlated events happens.

Randomized Load Balancing:



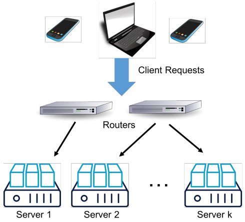
- n requests randomly assigned to k servers.
- Expected load on server i is $\mathbb{E}[R_i] = \quad$.

Randomized Load Balancing:



- n requests randomly assigned to k servers.
- Expected load on server i is $\mathbb{E}[\mathsf{R}_i] = \frac{n}{k}$.

Randomized Load Balancing:



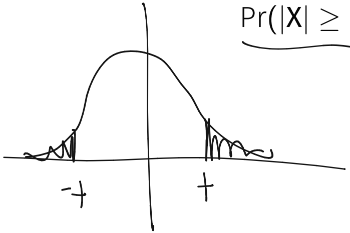
- n requests randomly assigned to k servers.
- Expected load on server i is $\mathbb{E}[\mathbf{R}_i] = \frac{n}{k}$.
- By Markov's inequality, if we provision each server to handle twice this expected load (so $\frac{2n}{k}$ requests), it will be overloaded with probability $\leq 1/2$.

With a very simple twist, Markov's inequality can be made much more powerful.

With a very simple twist, Markov's inequality can be made much more powerful.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$



With a very simple twist, Markov's inequality can be made much more powerful.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$

X^2 is a nonnegative random variable. So can apply Markov's inequality:

With a very simple twist, Markov's inequality can be made much more powerful.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$

X^2 is a nonnegative random variable. So can apply Markov's inequality:

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

With a very simple twist, Markov's inequality can be made much more powerful.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$

X^2 is a nonnegative random variable. So can apply Markov's inequality:

$|X - \mathbb{E}X|$

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

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

With a very simple twist, Markov's inequality can be made much more powerful.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$

X^2 is a nonnegative random variable. So can apply Markov's inequality:

Chebyshev's inequality:

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

With a very simple twist, Markov's inequality can be made much more powerful.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$

X^2 is a nonnegative random variable. So can apply Markov's inequality:

Chebyshev's inequality:

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

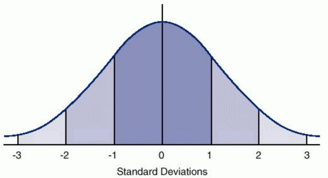
(by plugging in the random variable $X - \mathbb{E}[X]$)

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

X: any random variable, t, s: any fixed numbers.

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

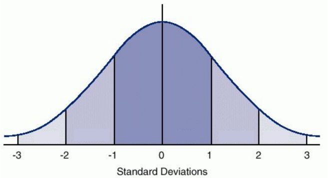
What is the probability that X falls s standard deviations from it's mean?



X: any random variable, t, s: any fixed numbers.

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

What is the probability that X falls s standard deviations from it's mean?



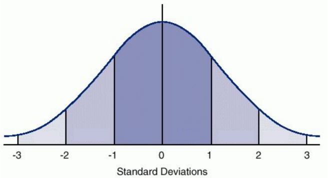
$$\Pr(|X - \mathbb{E}[X]| \geq s \cdot \sqrt{\mathrm{Var}[X]}) \leq \frac{\mathrm{Var}[X]}{s^2 \cdot \mathrm{Var}[X]} = \frac{1}{s^2}.$$

t

X: any random variable, t, s: any fixed numbers.

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

What is the probability that X falls s standard deviations from it's mean?



$$\Pr(|X - \mathbb{E}[X]| \geq s \cdot \sqrt{\mathrm{Var}[X]}) \leq \frac{\mathrm{Var}[X]}{s^2 \cdot \mathrm{Var}[X]} = \frac{1}{s^2}.$$

Why is this so powerful?

X: any random variable, t, s: any fixed numbers.

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\mathbb{E}(S) = \frac{1}{n} \cdot \sum \mathbb{E} X_i = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$\mathrm{Var}[S] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum \mathrm{Var}(X_i)$$

$$= \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\mathrm{Var}[S] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[X_i\right]$$

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\mathrm{Var}[S] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n^2}\cdot n \cdot \sigma^2$$

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\mathrm{Var}[S] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

$$\mathbb{E}[S] = \mu$$

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\mathrm{Var}[S] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

By Chebyshev's Inequality: for any fixed value $\epsilon > 0$,

$$\Pr(|S - \mathbb{E}[S]| \geq \epsilon) \leq \frac{\mathrm{Var}[S]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\text{Var}[S] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

By Chebyshev's Inequality: for any fixed value $\epsilon > 0$,

$$\Pr(|S - \mu| \geq \epsilon) \leq \frac{\text{Var}[S]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

LAW OF LARGE NUMBERS

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\mathrm{Var}[S] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

By Chebyshev's Inequality: for any fixed value $\epsilon > 0$,

$$\Pr(|S - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}[S]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Law of Large Numbers: with enough samples n, the sample average will always concentrate to the mean.

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2.

How well does the sample average $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate the true mean μ?

$$\text{Var}[S] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}[X_i] = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

By Chebyshev's Inequality: for any fixed value $\epsilon > 0$,

$$\Pr(|S - \mu| \geq \epsilon) \leq \frac{\text{Var}[S]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Law of Large Numbers: with enough samples n, the sample average will always concentrate to the mean.

- Cannot show from vanilla Markov's inequality.

LOAD BALANCING VARIANCE

We can write the number of requests assigned to server i, R_i as:

$$\mathsf{R}_i = \sum_{j=1}^{n} \mathsf{R}_{i,j}$$

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

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

We can write the number of requests assigned to server i, R_i as:

$$\text{Var}[R_i] = \sum_{j=1}^{n} \text{Var}[R_{i,j}] \qquad \text{(linearity of variance)}$$

where $R_{i,j}$ is 1 if request j is assigned to server i and 0 otherwise.

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

LOAD BALANCING VARIANCE

We can write the number of requests assigned to server i, R_i as:

$$\text{Var}[R_i] = \sum_{j=1}^{n} \text{Var}[R_{i,j}] \qquad \text{(linearity of variance)}$$

where $R_{i,j}$ is 1 if request j is assigned to server i and 0 otherwise.

$$\text{Var}[R_{i,j}] = \mathbb{E}\left[\left(R_{i,j} - \mathbb{E}[R_{i,j}]\right)^2\right]$$

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

LOAD BALANCING VARIANCE

We can write the number of requests assigned to server i, R_i as:

$$\text{Var}[R_i] = \sum_{j=1}^{n} \text{Var}[R_{i,j}] \qquad \text{(linearity of variance)}$$

where $R_{i,j}$ is 1 if request j is assigned to server i and 0 otherwise.

$$\begin{aligned}\text{Var}[R_{i,j}] &= \mathbb{E}\left[\left(R_{i,j} - \mathbb{E}[R_{i,j}]\right)^2\right] \\ &= \Pr(R_{i,j} = 1) \cdot \left(1 - \mathbb{E}[R_{i,j}]\right)^2 + \Pr(R_{i,j} = 0) \cdot \left(0 - \mathbb{E}[R_{i,j}]\right)^2\end{aligned}$$

$$\frac{1}{k} \cdot \left(1 - \frac{1}{k}\right)^2 + \left(1 - \frac{1}{k}\right) + \left(0 - \frac{1}{k}\right)^2$$

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

We can write the number of requests assigned to server i, R_i as:

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

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

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

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

We can write the number of requests assigned to server i, R_i as:

$$\mathrm{Var}[R_i] = \sum_{j=1}^{n} \mathrm{Var}[R_{i,j}] \qquad \text{(linearity of variance)}$$

where $R_{i,j}$ is 1 if request j is assigned to server i and 0 otherwise.

$$\begin{aligned}
\mathrm{Var}[R_{i,j}] &= \mathbb{E}\left[\left(R_{i,j} - \mathbb{E}[R_{i,j}]\right)^2\right] \\
&= \Pr(R_{i,j} = 1) \cdot \left(1 - \mathbb{E}[R_{i,j}]\right)^2 + \Pr(R_{i,j} = 0) \cdot \left(0 - \mathbb{E}[R_{i,j}]\right)^2 \\
&= \frac{1}{k} \cdot \left(1 - \frac{1}{k}\right)^2 + \left(1 - \frac{1}{k}\right) \cdot \left(0 - \frac{1}{k}\right)^2 \\
&= \frac{1}{k} - \frac{1}{k^2} \leq \frac{1}{k}
\end{aligned}$$

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

LOAD BALANCING VARIANCE

We can write the number of requests assigned to server i, $\mathbf{R}_i$ as:

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

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

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

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, $\mathbb{E}[\mathsf{R}_i] = \frac{n}{k}$ and $\text{Var}[\mathsf{R}_i] \le \frac{n}{k}$.

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

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

Applying Chebyshev's:

$$\Pr\left(\mathsf{R}_i \geq \frac{2n}{k}\right) \leq \Pr\left(|\mathsf{R}_i - \mathbb{E}[\mathsf{R}_i]| \geq \frac{n}{k}\right) \leq \frac{\text{Var}(R_i)}{n^2/k^2}$$

$$\Pr\left((R_i - \mathbb{E}R_i) \geq \frac{n}{k} \text{ or } (R_i - \mathbb{E}R_i) \leq -\frac{n}{k}\right)$$

$$\leq \frac{n/k}{n^2/k^2} = \frac{k}{n}$$

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

BOUNDING THE LOAD VIA CHEBYSHEVS

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:

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

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

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

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

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

Chebyshevs inequality

Applying Chebyshev's:

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

$\frac{1}{2}$

- Overload probability is extremely small when $k \ll n$!

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

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

Applying Chebyshev's:

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

- 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?

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

MAXIMUM SERVER LOAD

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

$$\Pr\left(\max_i(\mathrm{R}_i) \geq \frac{2n}{k}\right)$$

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

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?

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

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

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?

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

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

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?

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

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

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

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

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

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

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

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

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

How do we do this? Note that $\mathrm{R}_1, \ldots, \mathrm{R}_k$ are correlated in a somewhat complex way.

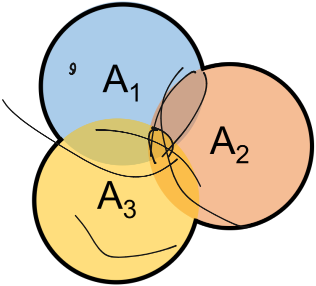
n: total number of requests, k: number of servers randomly assigned requests, R_i: number of requests assigned to server i. $\mathbb{E}[\mathrm{R}_i] = \frac{n}{k}$. $\mathrm{Var}[\mathrm{R}_i] = \frac{n}{k}$.

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

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

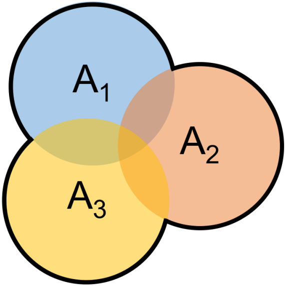
Union Bound: For any random events $A_1, A_2, \ldots, A_k$,

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



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

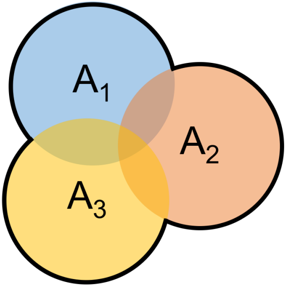
$$\Pr\left(A_1 \cup A_2 \cup \ldots \cup A_k\right) \leq \Pr(A_1) + \Pr(A_2) + \ldots + Pr(A_k).$$



When is the union bound tight?

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

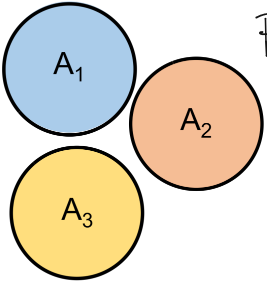
$$\Pr(A_1 \cup A_2 \cup \ldots \cup A_k) \leq \Pr(A_1) + \Pr(A_2) + \ldots + Pr(A_k).$$



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

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

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

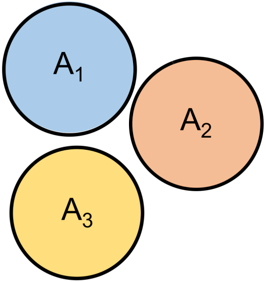


$$P(D=1 \text{ or } D=2 \text{ or } D=3) = \frac{1}{2} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

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

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

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



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

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

APPLYING THE UNION BOUND

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?

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

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

APPLYING THE UNION BOUND

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?

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

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

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

APPLYING THE UNION BOUND

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?

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

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

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

$$k \cdot \frac{k}{n} = \frac{k^2}{n}$$

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

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?

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

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

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

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

APPLYING THE UNION BOUND

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. $\mathbb{E}[\mathbf{R}_i] = \frac{n}{k}$. $\text{Var}[\mathbf{R}_i] = \frac{n}{k}$.

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.

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

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.

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

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 More Natural Variant:** 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 $\mathbf{H}$ be the number of heads.

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.

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

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.

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

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.

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

Markov's:

$\Pr(\mathsf{H} \geq 60) \leq .833$

$\Pr(\mathsf{H} \geq 70) \leq .714$

$\Pr(\mathsf{H} \geq 80) \leq .625$

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.

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

Markov's:	Chebyshev's:
$\Pr(\mathsf{H} \geq 60) \leq .833$	$\Pr(\mathsf{H} \geq 60) \leq .25$
$\Pr(\mathsf{H} \geq 70) \leq .714$	$\Pr(\mathsf{H} \geq 70) \leq .0625$
$\Pr(\mathsf{H} \geq 80) \leq .625$	$\Pr(\mathsf{H} \geq 80) \leq .0278$

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.

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

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.

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

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?