COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2023.

Lecture 4

Logistics

- Problem Set 1 due next Friday 9/22, at 11:59pm.
- Second quiz will be released today after class, due Monday 8:00pm.
- Change to challenge problems grading.

  ✓+ = 3 ✓ = 2 ✓- = 1

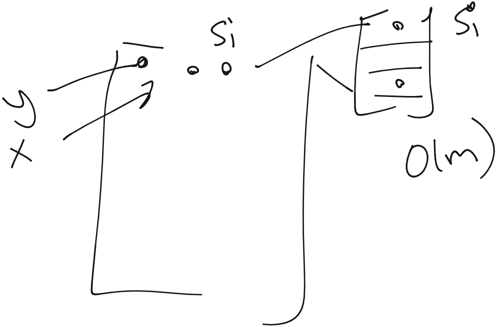
  Full score: 15 points

Last Time

Last Class:

- 2-level hashing and its analysis via linearity of expectation.
  Gives optimal $O(1)$ query time and $O(m)$ expected space usage.
- Practical random hash functions: 2-universal and pairwise independent hashing.

low-collision prob.



Last Time

Last Class:

- 2-level hashing and its analysis via linearity of expectation. Gives optimal $O(1)$ query time and $O(m)$ expected space usage.
- Practical random hash functions: 2-universal and pairwise independent hashing.

This Time:

- Hashing for load balancing in distributed systems. 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.

Markov

- Some of the pset questions use Chebyshev's inequality. After today you will know enough to solve everything on the pset.

Efficiently Computable Hash Functions

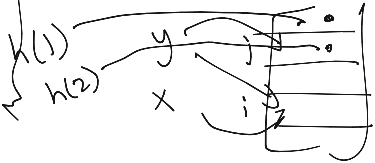
2-Universal Hash Function (low collision probability). A random hash function from $\mathbf{h} : U \to [n]$ is two universal if:

$$\Pr[\mathbf{h}(x) = \mathbf{h}(y)] \leq \frac{1}{n}.$$

Pairwise Independent Hash Function. A random hash function from $\mathbf{h} : U \to [n]$ is pairwise independent if for all $i, j \in [n]$:

$x \neq y$

$$\Pr[\mathbf{h}(x) = i \cap \mathbf{h}(y) = j] = \frac{1}{n^2}.$$



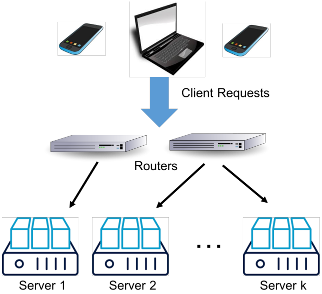
$h(1) = \text{rand}(n)$

$h(2) = h(1) + 1 \mod n$

$h(x) = \text{rand}(n)$

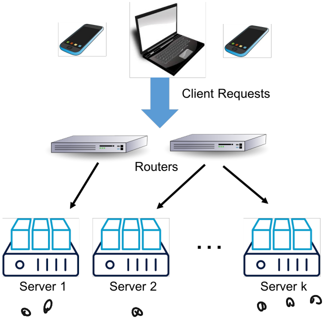
Another Application

Randomized Load Balancing:



Another Application

Randomized Load Balancing:



Simple Model: n requests randomly assigned to k servers. How many requests must each server handle?

- Often assignment is done via a random hash function. Why?

- what if servers go down   consitenly
- more secure.

Weakness of Markov's

R_i = # request assigned to server i.

$$\mathbb{E}[\mathrm{R}_i] = \frac{n}{k}$$

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

Weakness of Markov's

$$\mathbb{E}[\mathsf{R}_i] = \sum_{j=1}^{n} \mathbb{E}[\mathbb{I}_{\text{request } j \text{ assigned to } i}] = \sum_{j=1}^{n} \Pr\left[\underbrace{j \text{ assigned to } i}_{\frac{1}{k}}\right] = \frac{n}{k}.$$

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

Weakness of Markov's

$$\mathbb{E}[\mathsf{R}_i] = \sum_{j=1}^{n} \mathbb{E}[\mathbb{I}_{\text{request } j \text{ assigned to } i}] = \sum_{j=1}^{n} \Pr[j \text{ assigned to } i] = \frac{n}{k}.$$

If we provision each server be able to handle twice the expected load, what is the probability that a server is overloaded?

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

Weakness of Markov's

$$\mathbb{E}[\mathsf{R}_i] = \sum_{j=1}^{n} \mathbb{E}[\mathbb{I}_{\text{request } j \text{ assigned to } i}] = \sum_{j=1}^{n} \Pr[j \text{ assigned to } i] = \frac{n}{k}.$$

If we provision each server be able to handle twice the expected load, what is the probability that a server is overloaded?

Applying Markov's Inequality

$$\Pr\left[\mathsf{R}_i \geq 2\mathbb{E}[\mathsf{R}_i]\right] \leq \frac{\mathbb{E}[\mathsf{R}_i]}{2\mathbb{E}[\mathsf{R}_i]} = \frac{1}{2}.$$

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

Weakness of Markov's

$$\mathbb{E}[\mathsf{R}_i] = \sum_{j=1}^{n} \mathbb{E}[\mathbb{I}_{\text{request } j \text{ assigned to } i}] = \sum_{j=1}^{n} \Pr[j \text{ assigned to } i] = \frac{n}{k}.$$

If we provision each server be able to handle twice the expected load, what is the probability that a server is overloaded?

Applying Markov's Inequality

$$\Pr[\mathsf{R}_i \geq 2\mathbb{E}[\mathsf{R}_i]] \leq \frac{\mathbb{E}[\mathsf{R}_i]}{2\mathbb{E}[\mathsf{R}_i]} = \frac{1}{2}.$$

Not great...half the servers may be overloaded.

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

Chebyshev's inequality

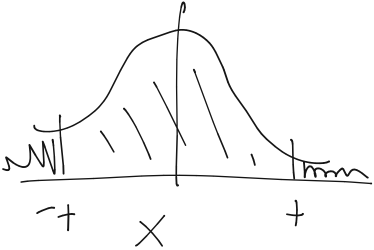
With a very simple twist, Markov's inequality can be made much more powerful.

Chebyshev's inequality

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

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

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



Chebyshev'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:

Chebyshev'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:

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

Chebyshev'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:

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

Chebyshev'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:

Chebyshev's inequality: $Y = X - \mathbb{E}X$

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

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

Chebyshev's inequality

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

$|a| < b$

$\underline{a^2 < b^2}$

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 $X - \mathbb{E}[X]$)

Chebyshev's inequality

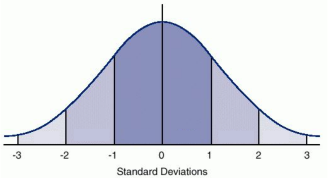
$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}$$

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

Chebyshev's inequality

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

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

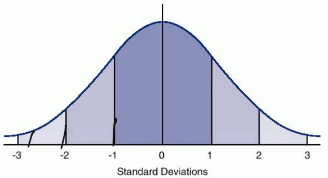


X: any random variable, t, s: any fixed numbers.

Chebyshev's inequality

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

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



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

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

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.

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?

$$\mathbb{E}[S] \stackrel{?}{=} \mu = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$\Pr(|S - \mu| > t) = \Pr(|S - \mathbb{E}S| > t)$$

$$\mathrm{Var}(S)$$

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?

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

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

$$= \boxed{\frac{\sigma^2}{n}}$$

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?

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

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?

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

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?

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

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 $\epsilon > 0$,

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

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

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

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

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

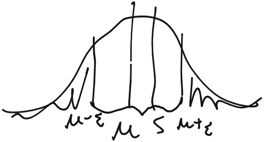
$\sigma = 1$

$\varepsilon = .01$

$$\Pr(|S - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}[S]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}. = .05$$

$$\lim_{n\to\infty} \frac{\sigma^2}{\varepsilon n} = 0$$

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



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.

- Cannot show from vanilla Markov's inequality.

Load Balancing Variance

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

$\mathbb{E}R_i = \frac{n}{k}$

$$R_i = \sum_{j=1}^{n} R_{i,j}$$

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, $\mathbf{R}_i$ as:

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

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

$$\mathbb{E}[R_{ij}] = \frac{1}{k} \qquad \mathbb{E}[R_{ij}^2] = \mathbb{E}R_{ij} = \frac{1}{k}$$

$= 1$ w.p. $\frac{1}{k}$

0 o.w.

$$\mathrm{Var}(R_{ij}) = \mathbb{E}[R_{ij}^2] - (\mathbb{E}[R_{ij}])^2$$

$$= \frac{1}{k} - \frac{1}{k^2} \leq \frac{1}{k}$$

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

Load Balancing Variance

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

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

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

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

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:

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

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

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

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, R_i as:

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

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

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

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

Load Balancing Variance

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

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

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

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

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, 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.

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

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

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 $\text{Var}[R_i] \leq \frac{n}{k}$.

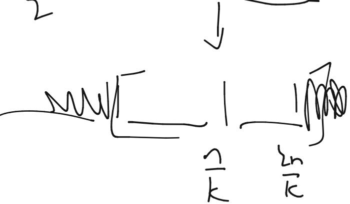
$\mathbb{E} R_i = \frac{n}{k}$

Applying Chebyshev's:

A B

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

$\frac{1}{2}$



if $R_i \geq \frac{2n}{k}$ then

$|R_i - \mathbb{E} R_i| \geq \frac{n}{k}$

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 $\text{Var}[R_i] \leq \frac{n}{k}$.

$$\leq \frac{\text{Var}(R_i)}{(n/k)^2}$$

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.

Bounding the Load via Chebyshevs

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

Applying Chebyshev's:

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

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

Applying Chebyshev's:

$$\Pr\left(\mathsf{R}_i \ge \frac{2n}{k}\right) \le \Pr\left(|\mathsf{R}_i - \mathbb{E}[\mathsf{R}_i]| \ge \frac{n}{k}\right) \le \frac{n/k}{n^2/k^2} = \frac{k}{n}. << \frac{1}{2}$$

- 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.

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 $\text{Var}[R_i] \le \frac{n}{k}$.

$\mathbb{E}R_i = n$ $\quad k = n \quad \frac{1}{n} \quad 0$

Applying Chebyshev's:

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

$\Pr(R_i > 2)$

$n/10$

- 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.

Maximum Server Load

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. $\mathbb{E}[R_i] = \frac{n}{k}$. $\mathrm{Var}[R_i] = \frac{n}{k}$.

Maximum Server Load

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?

$$\Pr\left(\underline{\max_{\substack{i=1\\k}}(R_i) \geq \frac{2n}{k}}\right)$$

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

Maximum Server Load

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. $\mathbb{E}[R_i] = \frac{n}{k}$. $\text{Var}[R_i] = \frac{n}{k}$.

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?

$$\Pr\left(\max_i(\mathrm{R}_i) \geq \frac{2n}{k}\right) = \Pr\left(\left[\mathrm{R}_1 \geq \frac{2n}{k}\right] \text{ or } \left[\mathrm{R}_2 \geq \frac{2n}{k}\right] \text{ or } \ldots \text{ or } \left[\mathrm{R}_k \geq \frac{2n}{k}\right]\right)$$

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

Maximum Server Load

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?

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

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

Maximum Server Load

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?

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

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

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

Maximum Server Load

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?

$P(A \cap B)$
$= P(A)P(B)$

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

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. $\mathbb{E}[R_i] = \frac{n}{k}$. $\mathrm{Var}[R_i] = \frac{n}{k}$.

The Union Bound

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

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

The Union Bound

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

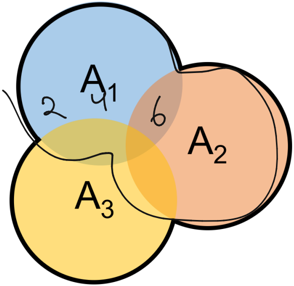
$$\Pr(A_1 \cup A_2 \cup \ldots \cup A_k) \le \Pr(A_1) + \Pr(A_2) + \ldots + Pr(A_k).$$

3 dice

A_1 = dice hits 6

A_2 : dice hits 6

:



$A_1 = \{2, 4, 6\}$

$A_2 = \{6\}$

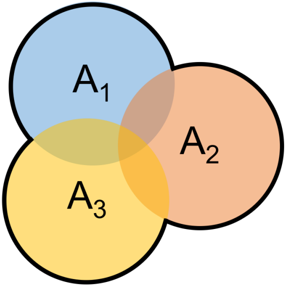
$P(A_1 \cup A_2 \cup A_3) = Pr(\text{at least one hits } 6)$

$\le \frac{1}{2}$

The Union Bound

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

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

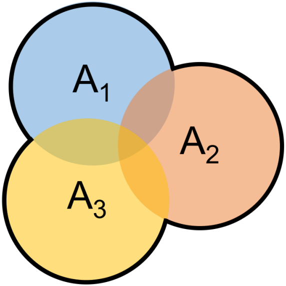


When is the union bound tight?

The Union Bound

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

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



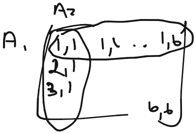
When is the union bound tight? When $A_1, ..., A_k$ are all disjoint.

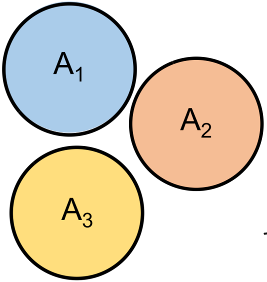
Union Bound: For any random events $A_1, A_2, \ldots, A_k$,

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

$\Pr(\text{roll } 1 \text{ or roll } 2)$

$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$





2 dice

A_1 = die 1 rolls 1

A_2 = die 2 rolls 1

36 outcomes

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

Applying the Union Bound

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?

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

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

Applying the Union Bound

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?

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

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

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

Applying the Union Bound

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?

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

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

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

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

Applying the Union Bound

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?

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

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

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

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

Applying the Union Bound

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?

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

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

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

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