COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2024.

Lecture 4

Logistics

- Problem Set 1 due next Friday 9/20, at 11:59pm.
- Second quiz will be released today after class, due Monday 8:00pm.

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.

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

$h(x) = ax + b \bmod p \bmod n$

Last Time

Last Class:

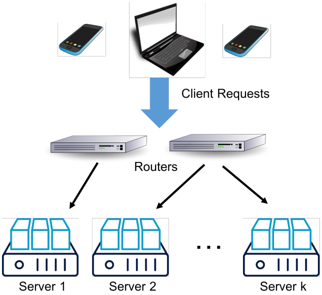
- 2-level hashing and its analysis via linearity of expectation. Gives optimal $O(1)$ query time and $O(m)$ expected space usage.
- Practical random hash functions: 2-universal and pairwise independent hashing.

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.
- Some of the problem set questions use Chebyshev's inequality. After today you will be able to solve them.

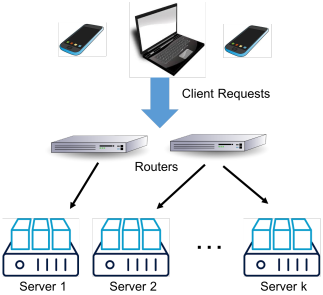
Another Application

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?

L consistency, cache performance

Weakness of Markov's

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

└ #requests assigned to server i

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

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

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

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

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

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

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

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

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

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

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

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:

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

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:

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

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 $t > 0$:

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

r.v.

deviation

$\Pr(Y > s) \leq \frac{\mathbb{E}[Y]}{s}$

$Y = X^2$

$s = t^2$

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:

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

$X - \mathbb{E}X$

Chebyshev's inequality

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

$X = -2 \quad -1 \quad 0 \quad 1 \quad 2$

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

$Pr(|X| \geq 2) = \frac{2}{5}$

$Pr(X^2 \geq 4) = \frac{2}{5}$

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

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

Chebyshev's inequality:

$\mathbb{E}X^4$

$\mathbb{E}[(X-\mathbb{E}X)^2]$

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

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

Chebyshev's inequality

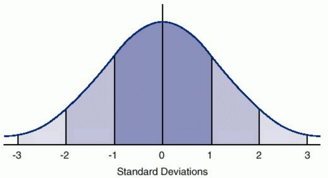
$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}$$

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

Chebyshev's inequality

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

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

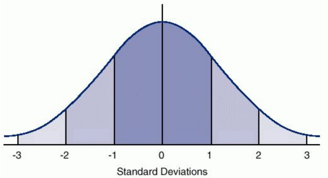


X: any random variable, t, s: any fixed numbers.

Chebyshev's inequality

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

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



Chebyshev: 3σ: $\frac{1}{9} \approx .1$

Gaussian: $\approx .01$

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

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?

Law of Large Numbers

Consider drawing independent identically distributed (i.i.d.) random variables $X_1, \dots, X_n$ 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?

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

$$\mu = \mathbb{E}X_i$$

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

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?

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

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?

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

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?

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

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

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

σ^2 may not be finite

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?

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

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

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

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?

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

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

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

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

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?

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

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:

requests

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

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, 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.

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

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

$p(1-p)$

$p - p^2$

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

$$\mathrm{Var}[R_{i,j}] = \mathbb{E}[R_{i,j}^2] - \mathbb{E}[R_{i,j}]^2 = \frac{1}{k} - \frac{1}{k^2}$$

$\mathbb{E}[R_{i,j}] = \frac{1}{k}$

$\frac{1}{k^2}$

> 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.

$$\begin{aligned} \mathsf{Var}[\mathsf{R}_{i,j}] &= \mathbb{E}[\mathsf{R}_{i,j}^2] - \mathbb{E}[\mathsf{R}_{i,j}]^2 \\ &= \mathbb{E}[\mathsf{R}_{i,j}] - \mathbb{E}[\mathsf{R}_{i,j}]^2 \end{aligned}$$

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:

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

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

$$\begin{aligned}\text{Var}[R_{i,j}] &= \mathbb{E}[R_{i,j}^2] - \mathbb{E}[R_{i,j}]^2 \\ &= \mathbb{E}[R_{i,j}] - \mathbb{E}[R_{i,j}]^2 \\ &= \frac{1}{k} - \frac{1}{k^2} \leq \frac{1}{k}\end{aligned}$$

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.

$$\begin{aligned} \text{Var}[R_{i,j}] &= \mathbb{E}[R_{i,j}^2] - \mathbb{E}[R_{i,j}]^2 \\ &= \mathbb{E}[R_{i,j}] - \mathbb{E}[R_{i,j}]^2 \\ &= \frac{1}{k} - \frac{1}{k^2} \leq \frac{1}{k} \implies \text{Var}[R_i] \leq \frac{n}{k}. \end{aligned}$$

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

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

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

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

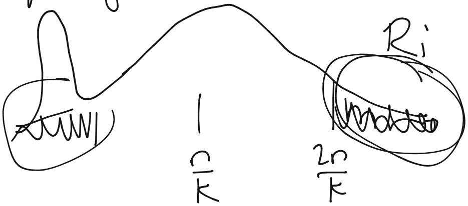
Bounding the Load via Chebyshevs

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

Applying Chebyshev's:

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

load   capacity



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 $\text{Var}[\mathsf{R}_i] \leq \frac{n}{k}$.

Applying Chebyshev's:

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

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 $\mathrm{Var}[\mathsf{R}_i] \leq \frac{n}{k}$.

Applying Chebyshev's:

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

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 $\mathrm{Var}[\mathsf{R}_i] \leq \frac{n}{k}$.

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

Applying Chebyshev's:

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

capacity

- Overload probability is 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 $2 \cdot \mathbb{E}[\mathrm{R}_i] = \frac{2n}{k}$. 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. $\mathbb{E}[\mathrm{R}_i] = \frac{n}{k}$. $\mathrm{Var}[\mathrm{R}_i] = \frac{n}{k}$.

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?

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

$$\Pr\left(R_i \geq \frac{2n}{k}\right)$$

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

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

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?

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

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

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?

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

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

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?

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

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

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?

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

$\Pr\left(\bigcup_{i=1} A_i\right)$

$1 - \Pr\left(\bigcap_{i=1} \bar{A}_i\right)$

$1 - \prod_{i=1} \bar{A}_i$

$\geq \left(1 - \frac{k}{n}\right)$

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

The Union Bound

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

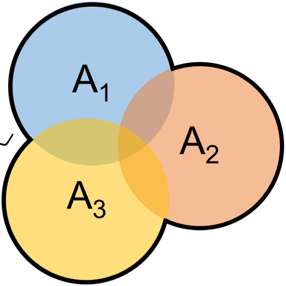
$$\Pr\left(A_1 \cup A_2 \cup \ldots \cup A_k\right) \leq \Pr(A_1) + \Pr(A_2) + \ldots + Pr(A_k).$$

The Union Bound

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

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

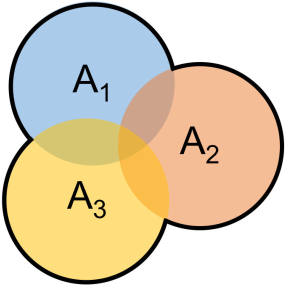
special case of markovs applied to indicator r.v.s. for $A_1 \ldots A_n$



The Union Bound

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

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

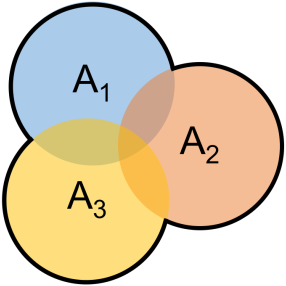


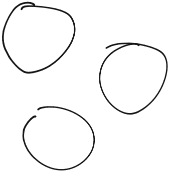
When is the union bound tight?

The Union Bound

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

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



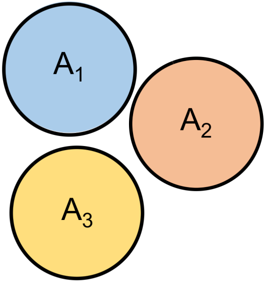


When is the union bound tight? When $A_1, ..., A_k$ are all disjoint.

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.

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?

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

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

bounded by $\frac{k}{n}$ via 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?

$$
\begin{aligned}
\Pr\left(\max_i(R_i) \geq \frac{2n}{k}\right) &= \Pr\left(\bigcup_{i=1}^{k}\left[R_i \geq \frac{2n}{k}\right]\right) \\
&\leq \sum_{i=1}^{k} \Pr\left(\left[R_i \geq \frac{2n}{k}\right]\right) && \text{(Union Bound)} \\
&\leq \sum_{i=1}^{k} \frac{k}{n} && \text{(Bound from Chebyshev's)}
\end{aligned}
$$

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

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?

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

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

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

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

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

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.

$\mathbb{E}[H] = 50$

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.

$$\mathbb{E}[H] = \frac{n}{2} = 50 \text{ and } \mathrm{Var}[H] = \mathrm{Var}\left(\sum_{i=1}^{100} H_i\right)$$

$$= \sum_{i=1}^{100} \mathrm{Var}(H_i)$$

$$\mathbb{E}[H_i^2] - \mathbb{E}[H_i]^2$$

$$\frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$\sum_{i=1}^{100} \frac{1}{4} = 25$$

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.

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

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:

50/60

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

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

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

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.

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

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

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.

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

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

Bin(n,p)
Bin(100,.5)

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.

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?

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.

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.

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

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

A Fourth Moment Bound

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

A Fourth Moment Bound

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