COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2022.

Lecture 4

Logistics

- Problem Set 1 due next Friday 9/23, at 11:59pm.
- Second quiz will be released today after class, due Monday 8:00pm.
- I will hold additional office hours next Tuesday 11am-12pm.

Last Time

Last Class:

- Expected collision analysis for hashing and collision free hashing via Markov's inequality. Gives $O(1)$ query time and $O(m^2)$ space for item look-up problem.
- 2-level hashing and its analysis via linearity of expectation. Gives optimal $O(1)$ query time and $O(m)$ space.

This Time:

- 2-universal and pairwise independent hash functions
- 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.

Efficiently Computable Hash Function

So Far: we have assumed a **fully random hash function** $\mathrm{h}(x)$ with $\Pr[\mathrm{h}(x) = i] = \frac{1}{n}$ for $i \in 1, \ldots, n$ and $\mathrm{h}(x), \mathrm{h}(y)$ independent for $x \neq y$.

- To compute a random hash function we have to store a table of x values and their hash values. Would take at least $O(m)$ space and $O(m)$ query time to look up $\mathrm{h}(x)$ if we hash m values. Making our whole quest for $O(1)$ query time pointless!

x	h(x)
x_1	45
x_2	1004
x_3	10
⋮	⋮
x_m	12

Efficiently Computable Hash Functions

What properties did we use of the randomly chosen hash function?

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

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

Exercise: Rework the two level hashing proof to show that this property is really all that is needed.

When $\mathbf{h}(x)$ and $\mathbf{h}(y)$ are chosen independently at random from $[n]$, $\Pr[\mathbf{h}(x) = \mathbf{h}(y)] = \frac{1}{n}$ (so a fully random hash function is 2-universal)

Efficient Alternative: Let p be a prime with $p \geq |U|$. Choose random $\mathbf{a}, \mathbf{b} \in [p]$ with $\mathbf{a} \neq 0$. Represent x an an integer and let

$$\mathbf{h}(x) = (\mathbf{a}x + \mathbf{b} \mod p) \mod n.$$

Pairwise Independence

Another common requirement for a hash function:

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

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

Pairwise hash functions are 2-universal:

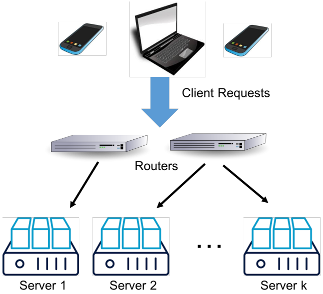
$$\Pr[\mathbf{h}(x) = \mathbf{h}(y)] = \sum_{i=1}^{n} \Pr[\mathbf{h}(x) = i \cap \mathbf{h}(y) = i] = n \cdot \frac{1}{n^2} = \frac{1}{n}.$$

A closely related $(\mathbf{a}x + \mathbf{b}) \mod p$ construction gives pairwise independence on top of 2-universality.

Remember: A fully random hash function is both 2-universal and pairwise independent. But it is not efficiently implementable.

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?

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.

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:

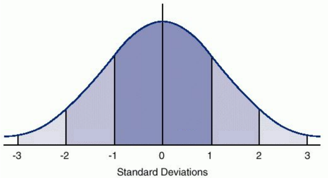
$$\Pr(|X - \mathbb{E}[X]| \geq t) \quad \Pr(|X| \geq t) = \Pr(X^2 \geq t^2) \leq \frac{\mathbb{E}[X^2]}{t^2} \quad \frac{\mathrm{Var}[X]}{t^2}.$$

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

Chebyshev's inequality

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

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



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

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.

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

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

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

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

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:

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

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

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

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

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.

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?

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

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

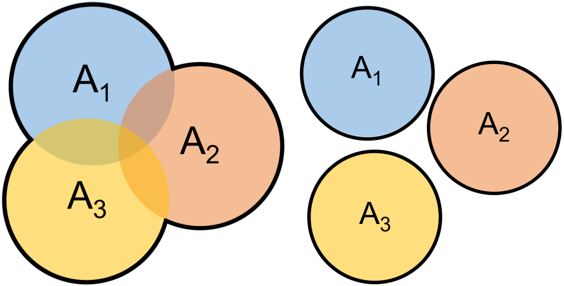
How do we do this? Note that $R_1, \ldots, R_k$ are correlated in a somewhat complex way.

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

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.

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