CS-GY 9223 Lecture 1 Supplemental
Load Balancing, Union Bound, Chebyshev's Inequality

NYU Tandon School of Engineering, Prof. Christopher Musco

LAST LECTURE

Showed how to design an $O(m)$ sized hash table that supported $O(1)$ time queries using a 2-level approach.

In doing so, we did not provide a bound on the maximum number of elements in each row of the level one hash table.

For some applications, this would be nice to have!

Load balancing problem:

Suppose Google answers map search queries using servers $A_1, \ldots, A_q$. Given a query like "new york to rhode island", common practice is to choose a random hash function $h \to \{1 \ldots, q\}$ and to route this query to server:

$$A_{h(\text{“new york to rhode island”})}$$

The advantage of this is that duplicate requests always get routed to the same server, saving computation time.

Goal: Ensure that requests are distributed evenly, so no one server gets loaded with too many requests. We want to avoid downtime and slow responses to clients.

LECTURE ROAD MAP

1. Show that Linearity of Expectation + Markov's are too weak to get any interesting theoretical bounds.
2. Introduce two new tools: the Union Bound and Chebyshev Inequality to prove something much more interesting.

These four simple tools combined are surprising powerful and flexible. Along with exponential tail bounds (next class), they form the cornerstone of randomized algorithm design.

Suppose we have n servers and m requests, $x_1, \ldots, x_m$. Let s_i be the number of requests sent to server $i \in \{1, \ldots, n\}$:

$$s_i = \sum_{j=1}^{m} \mathbb{1}[h(x_j) = i].$$

Formally, our goal is to understand the value of maximum load on any server, which can be written as the random variable:

$$S = \max_{i \in \{1,\ldots,n\}} s_i.$$

A good first step in any analysis of random variables is to first think about expectations. If we have n servers and m requests, for any $i \in \{1, \ldots, n\}$:

$$\mathbb{E}[s_i] = \sum_{j=1}^{m} \mathbb{E}\left[\mathbb{1}[h(x_j) = i]\right] = \frac{m}{n}.$$

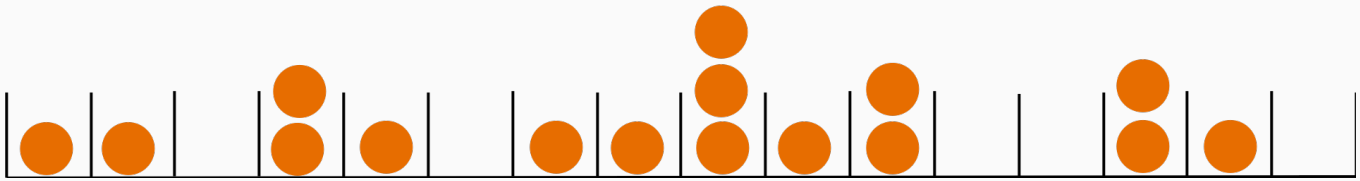
But it's very unclear what the expectation of $S = \max_{i \in \{1,\ldots,n\}} s_i$ is... in particular, $\mathbb{E}[S] \neq \max_{i \in \{1,\ldots,n\}} \mathbb{E}[s_i]$.

Exercise: Convince yourself that for two random variables A and B, $\mathbb{E}[\max(A, B)] \neq \max(\mathbb{E}[A], \mathbb{E}[B])$ even if those random variable are independent.

SIMPLIFYING ASSUMPTIONS

Number of servers: To reduce notation and keep the math simple, let's assume that $m = n$. I.e., we have exactly the same number of servers and requests.

Hash function: Continue to assume a fully (uniformly) random hash function h.



Often called the "balls-into-bins" model.

$\mathbb{E}[s_i]$ = expected number of balls per bin $= \frac{m}{n} = 1$. We would like to prove a bound of the form:

$$\Pr[\max_i s_i \geq C] \leq \frac{1}{10}.$$

for as tight a value of C. I.e., something much better than $C = n$.

BOUNDING A UNION OF EVENTS

Goal: Prove that for some $C \ll n$,

$$\Pr[\max_i s_i \geq C] \leq \frac{1}{10}.$$

$\cup$ means "or". $\cap$ means "and".

Equivalent statement: Prove that for some $C \ll n$,

$$\Pr[(s_1 > C) \cup (s_2 \geq C) \cup \ldots \cup (s_n \geq C)] \leq \frac{1}{10}.$$

Need to bound the probability of a union of different events.

These events are not independent!!

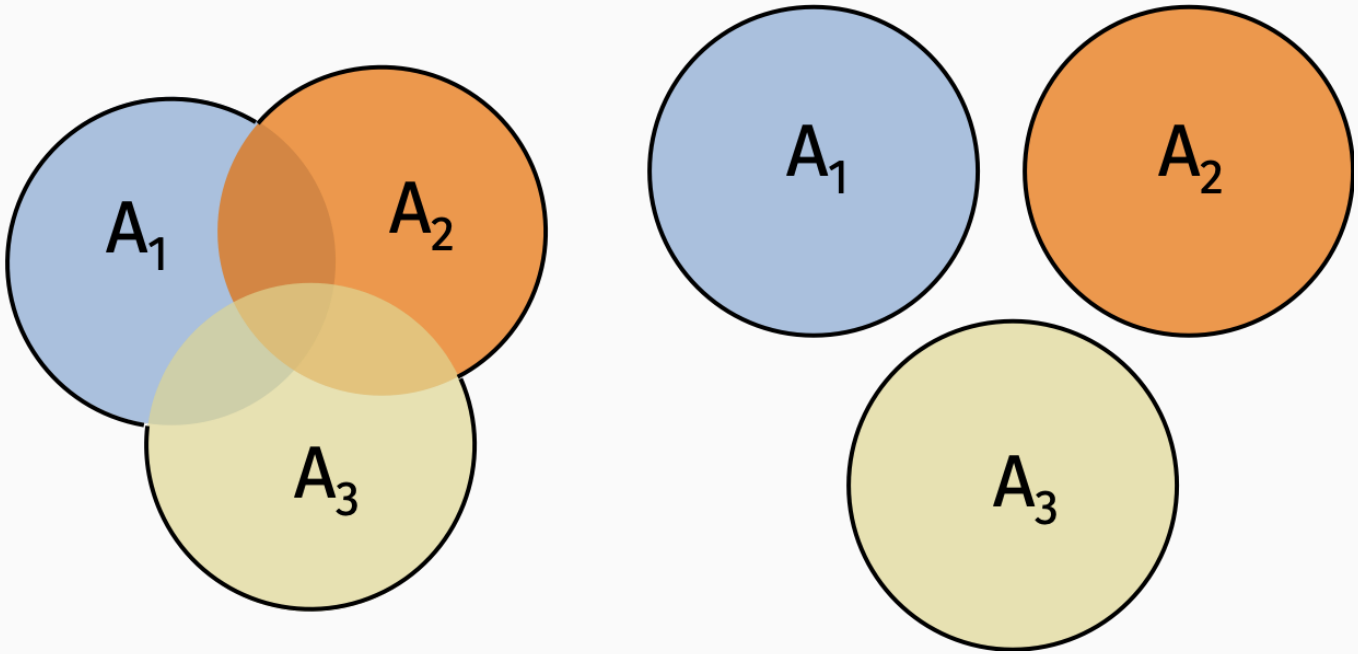
$n =$ number of balls and number of bins. s_i is number of balls in bin i. $C =$ upper bound on maximum number of balls in any bin.

USE A UNION BOUND

Lemma (Union Bound)

For <u>any</u> random events $A_1, \ldots, A_k$:

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_k] \leq \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_k].$$



Proof by picture.

APPLICATION OF UNION BOUND

We want to prove that:

$$\Pr[\max_i s_i \geq C] = \Pr[(s_1 \geq C) \cup (s_2 \geq C) \cup \ldots \cup (s_n \geq C)] \leq \frac{1}{10}.$$

To do so, it suffices to prove that for all i:

$$\Pr[s_i \geq C] \leq \frac{1}{10n}.$$

Why? Because then by the union bound,

$$\Pr[\max_i s_i \geq C] \leq \sum_{i=1}^{n} \Pr[s_i \geq C] \quad \text{(Union bound)}$$

$$\leq \sum_{i=1}^{n} \frac{1}{10n} = \frac{1}{10}. \quad \square$$

NEW GOAL

Prove that for some $C \ll n$,

$$\Pr[s_i \geq C] \leq \frac{1}{10n}.$$

This should look hard! We need to prove that $s_i < C$ (i.e. the i^{th} bin has a small number of balls) with very high probability (specifically $1 - \frac{1}{10n}$.

Markov's inequality is too weak of a bound for this.

$n =$ number of balls and number of bins. s_i is number of balls in bin i. $C =$ upper bound on maximum number of balls in any bin.

ATTEMPT WITH MARKOV'S INEQUALITY

Goal: Prove that $\Pr[s_i \geq C] \leq \frac{1}{10n}$.

- **Step 1.** Verify we can apply Markov's: s_i takes on non-negative values only. Good to go!
- **Step 2.** Apply Markov's: $\Pr[s_i \geq C] \leq \frac{\mathbb{E}[s_i]}{C} = \frac{1}{C}$.

To prove our target statement, need to see $C = 10n$.

Meaningless! There are only n balls, so of course there can't be more than $10n$ in the most overloaded bin.

n = number of balls and number of bins. s_i is number of balls in bin i. $\mathbb{E}[s_i] = 1$. C = upper bound on maximum number of balls in any bin. **Markov's inequality**: for positive r.v. X, $\Pr[X \geq t] \leq \mathbb{E}[X]/t$.

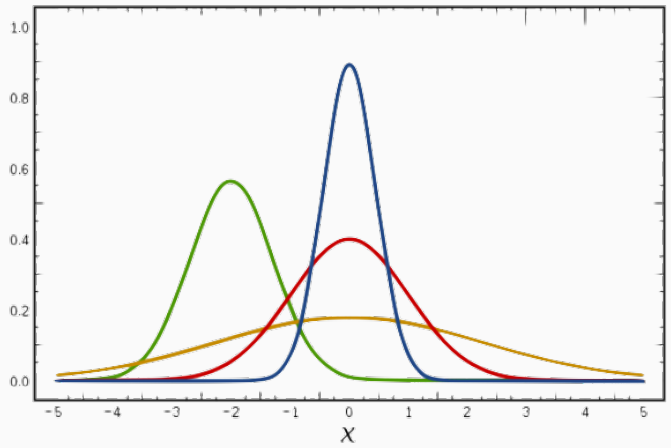
A NEW TOOL: CHEBYSHEV'S INEQUALITY

A new concentration inequality:

Lemma (Chebyshev's Inequality)

Let X be a random variable with expectation $\mathbb{E}[X]$ and variance $\sigma^2 = \mathrm{Var}[X]$. Then for any $k > 0$,

$$\Pr[|X - \mathbb{E}[X]| \geq k \cdot \sigma] \leq \frac{1}{k^2}$$



$\sigma = \sqrt{\mathrm{Var}[X]}$ is called the standard deviation of X. Intuitively this bound makes sense: it is tighter when σ is smaller.

COMPARISON TO MARKOV'S INEQUALITY

Properties of Chebyshev's inequality:

- **Good:** No requirement of non-negativity. X can be anything.
- **Good:** Two-sided. Bounds the probability that $|X - \mathbb{E}X|$ is large, which means that X isn't too far above or below its expectation. Markov's only bounded probability that X exceeds $\mathbb{E}[X]$.
- **Bad/Good:** Requires a bound on the variance of of X.

No hard rule for which to apply! Both Markov's and Chebyshev's are useful in different settings.

Proof of Chebyshev's inequality

Idea: Apply Markov's inequality to the (non-negative) random variable $S = (X - \mathbb{E}[X])^2$.

Lemma (Chebyshev's Inequality)

Let X be a random variable with expectation $\mathbb{E}[X]$ and variance $\sigma^2 = \mathrm{Var}[X]$. Then for any $k > 0$,

$$\Pr[|X - \mathbb{E}[X]| \geq k \cdot \sigma] \leq \frac{1}{k^2}$$

Let $S = (E - \mathbb{E}[X])^2$.

$$\Pr[S \geq k^2\sigma^2] \leq \frac{\mathbb{E}[S]}{k^2\sigma^2} \qquad \text{(Markov inequality)}$$

$$\Pr[\sqrt{S} \geq k\sigma] \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{k^2\sigma^2}$$

$$\Pr[|X - \mathbb{E}[X]| \geq k\sigma] \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}. \quad \square$$

APPLICATION TO BALLS INTO BINS

Goal: Prove that $\Pr[s_i \geq C] \leq \frac{1}{10n}$.

- **Step 1.** To apply Chebyshev's inequality, we need to understand $\sigma^2 = \mathrm{Var}[s_i]$.

Use linearity of variance. Let $s_{i,j}$ be a $\{0,1\}$ indicator random variable for the event that ball j falls in bin i. Clearly:

$$s_i = \sum_{j=1}^{n} s_{i,j}.$$

And $s_{i,1}, \ldots, s_{i,n}$ are (pairwise) independent so:

$$\mathrm{Var}[s_i] = \sum_{j=1}^{n} \mathrm{Var}[s_{i,j}].$$

$n =$ number of balls and number of bins. s_i is number of balls in bin i. $\mathbb{E}[s_i] = 1$. $C =$ upper bound on max number of balls in bin.

Use identity from first class: $\mathrm{Var}[s_{i,j}] = \mathbb{E}[s_{i,j}^2] - \mathbb{E}[s_{i,j}]^2$.

$$s_{i,j} = \begin{cases} 1 \text{ with probability } \frac{1}{n} \\ 0 \text{ otherwise.} \end{cases}$$

$$\mathbb{E}[s_{i,j}] = 1 \cdot \frac{1}{n} + 0 \cdot (1 - \frac{1}{n}) = \frac{1}{n}.$$

$$\mathbb{E}[s_{i,j}^2] = 1^2 \cdot \frac{1}{n} + 0^2 \cdot (1 - \frac{1}{n}) = \frac{1}{n}.$$

So:

$$\mathrm{Var}[s_{i,j}] = \mathbb{E}[s_{i,j}^2] - \mathbb{E}[s_{i,j}]^2 = \frac{1}{n} - \frac{1}{n^2}.$$

$n =$ number of balls and number of bins. $s_{i,j}$ is event ball j lands in bin i.

APPLYING CHEBYSHEV'S

Goal: Prove that $\Pr[s_i \geq C] \leq \frac{1}{10n}$.

Step 1. To apply Chebyshev's inequality, we need to understand $\sigma^2 = \mathrm{Var}[s_i]$.

$$\mathrm{Var}[s_i] = \sum_{j=1}^{n} \mathrm{Var}[s_{i,j}] = \sum_{j=1}^{n} \frac{1}{n} - \frac{1}{n^2} = 1 - \frac{1}{n}.$$

Step 2. Apply Chebyshev's inequality:

$$\Pr\left[|s_i - \mathbb{E}[s_i]| \geq k \cdot \sqrt{1 - 1/n}\right] \leq \frac{1}{k^2}$$

$$\text{which implies } \Pr\left[|s_i - 1| \geq k \cdot 1\right] \leq \frac{1}{k^2}.$$

$n =$ number of balls and number of bins. $s_i =$ number of balls in bin i. $s_{i,j}$ is event ball j lands in bin i. $\mathbb{E}[s_i] = 1$.

APPLYING CHEBYSHEV'S

Goal: Prove that $\Pr[s_i \geq C] \leq \frac{1}{10n}$.

We just proved: $\Pr[|s_i - 1| \geq k] \leq \frac{1}{k^2}$.

Setting $k = \sqrt{10n}$ gives:

$$\Pr[|s_i - 1| \geq \sqrt{10n}] \leq \frac{1}{10n}.$$

So, we have that:

$$\Pr[s_i \geq \sqrt{10n} + 1] \leq \frac{1}{10n}.$$

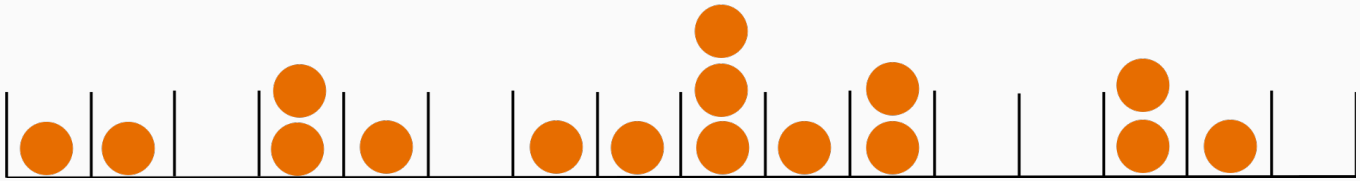
By the union bound argument from earlier, it thus holds that:

$$\Pr[\max_{i \in \{1,\ldots,n\}} s_i \geq \sqrt{10n} + 1] \leq \frac{1}{10}.$$

n = number of balls and number of bins. s_i is number of balls in bin i. C = upper bound on maximum number of balls in any bin.

FINAL RESULT

When hashing n balls into n bins, the maximum bin contains $o(\sqrt{n})$ balls with probability $\frac{9}{10}$.



Much better than the trivial bound of n!

TAKEAWAYS

Techniques used that will appear again:

- Union bound to control the maximum of many random variables.
- Chebyshev's inequality to bound a variable whose variance we can compute.
- To compute the variance, break down random variable into smaller pieces and apply linearity of variance.

Next class: We will use even stronger tools to prove a better bound of $o(\log n)$ for the most loaded bin.