CS7800: Advanced Algorithms

Class 21: Randomized Algorithms II

- Balls and Bins: maximum vs expected load
- Universal Hashing

Jonathan Ullman

November 21, 2025

Probability Case Study: Balls and Bins

Throw <u>m balls</u> into
<u>n bins</u> <u>independently</u>

$$\Omega = \{1, \dots, n\}^m$$

$$\mathbb{P}(\omega_1, \omega_2, \dots, \omega_m) = \frac{1}{n^m}$$

$$\omega = (6, 8, 11, 2, 37, \dots)$$

Ball 1
Bin 6

Ball 2
Bin 8

Questions:

1. How long until bin 1 gets a ball?
2. How long until no bin is empty?
3. What is the maximum number of balls in any bin?

Application: Hash Tables

"universe"

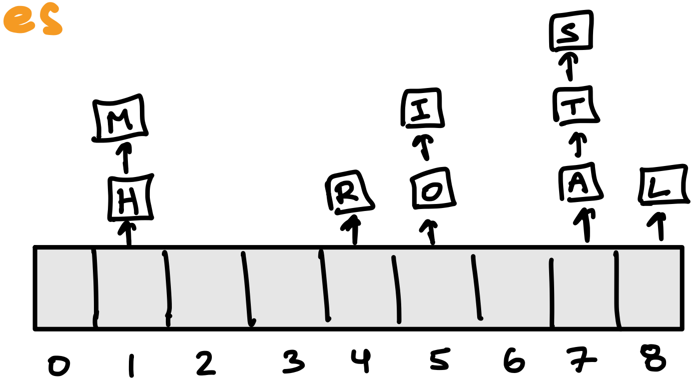
<u>Goal:</u> Store a set of <u>m elements</u> $S \subseteq \mathcal{U}$,

such that we can efficiently check if $x \in S$

↳ A "dictionary" also lets us associate a value with each key x

- A <u>hash table</u> $T[1:n]$ stores the elements in n <u>bins</u>
- A <u>hash function</u> $h: \mathcal{U} \to \{0, 1, \ldots, n-1\}$ maps elements to bins $x \mapsto T[h(x)]$

Application: Hash Tables



Looks a lot like balls in bins!

Linear chaining: a common way to deal with collisions

- Load factor $= \frac{m}{n}$

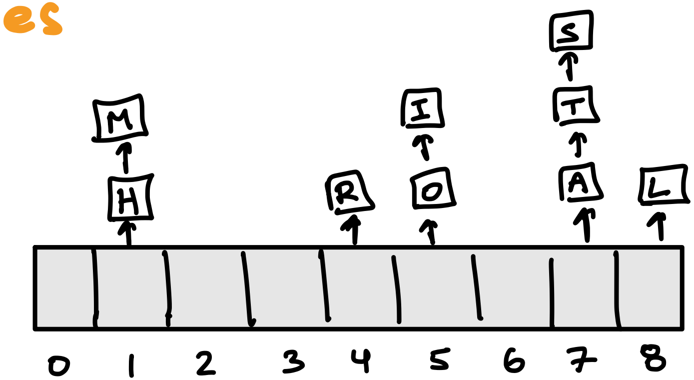
- Let $\ell(x) =$ # of elements in the same bin as x
  $\#\{y \in S : h(y) = h(x)\}$

- Time to lookup $x \in \mathcal{U}$ is $O(\ell(x))$

Worst-case lookup time $= \max_{x \in S} \ell(x)$

Application: Hash Tables

How should we choose the hash function $h: \mathcal{U} \to \{0, 1, \ldots, b-1\}$ to have small maximum load?



Looks a lot like balls in bins!

Balls and Bins: Maximum Load

m balls $\quad n$ bins

- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}\left(\max_i L_i\right) = \sum_{k=1}^{\infty} \underbrace{\mathbb{P}\left(\max_i L_i \geq k\right)}_{\text{want to bound this probability}}$

Balls and Bins: Maximum Load

m balls n bins

- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \underbrace{\mathbb{P}(\max_i L_i \geq k)}_{\text{want to bound this probability}}$

Step 1: "Union Bound"

Markov's Inequality

Thm: For any non-negative random variable X and every k, $\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}(X)}{k}$

Proof:

Balls and Bins: Maximum Load

m balls   n bins

- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \underbrace{\mathbb{P}(\max_i L_i \geq k)}_{\text{want to bound this probability}}$

$$\mathbb{P}(\max_i L_i \geq k) \leq n \cdot \mathbb{P}(L_1 \geq k)$$

$$\leq n \cdot \frac{\mathbb{E}(L_1)}{k}$$

Chebyshev's Inequality

Thm: For any random variable X with $\mu = \mathbb{E}(X)$ and every t,

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\mathbb{E}((X-\mu)^2)}{t^2}$$

$\mathbb{E}((X-\mu)^2) = \mathrm{Var}(X)$

Balls and Bins: Maximum Load

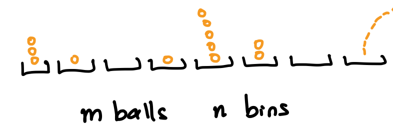
m balls $\quad n$ bins

- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \underbrace{\mathbb{P}(\max_i L_i \geq k)}_{\text{want to bound this probability}}$

$$\mathbb{E}(L_1) = \frac{m}{n}$$

$$\begin{aligned}
\mathbb{P}(\max_i L_i \geq k) &\leq n \cdot \mathbb{P}(L_1 \geq k) \\
&= n \cdot \mathbb{P}(L_1 - \mu \geq k - \mu) \\
&\leq n \cdot \mathbb{P}(|L_1 - \mu| \geq k - \mu) \\
&\overset{\text{Chebyshev}}{\leq} \frac{n \cdot \mathrm{Var}(L_1)}{(k-\mu)^2} \leq \frac{n \cdot \mathrm{Var}(L_1)}{k^2}
\end{aligned}$$

Balls and Bins: Maximum Load



- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \underbrace{\mathbb{P}(\max_i L_i \geq k)}_{\text{want to bound this probability}}$

$\mathrm{Var}(L_i) = ???$

Balls and Bins: Maximum Load

m balls n bins

- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \mathbb{P}(\max_i L_i \geq k)$
- Let $L_{i,j} = \begin{cases} 1 & \text{if ball } j \text{ is in bin } i \\ 0 & \text{otherwise} \end{cases}$ $\Rightarrow$ $L_i = L_{i,1} + \ldots + L_{i,m}$

How can we reason about sums of independent random variables?

Central Limit Theorem

Chernoff Bounds

$$Z_i = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases}$$

$$Z = Z_1 + \dots + Z_m$$

$Z_1, \dots, Z_m$ independent

$$\mu = \mathbb{E}(Z) = pm$$

$$\boxed{\textbf{Thm:} \quad \mathbb{P}\big(Z \geq (1+\varepsilon)\mu\big) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu}}$$

$\varepsilon < 1$

$$e^{-\frac{\mu\varepsilon^2}{4}}$$

ε

$$e^{-\mu\varepsilon}$$

Chernoff Bounds

Thm: $\mathbb{P}(Z \geq (1+\varepsilon)\mu) \leq \left(\dfrac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu}$

$Z_i = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases}$

$Z = Z_1 + \ldots + Z_m$

$Z_1, \ldots, Z_m$ independent

$\mu = \mathbb{E}(Z) = pm$

Proof:

$$\mathbb{P}(Z \geq t\mu) = \mathbb{P}(e^{sZ} \geq e^{st\mu})$$

What is this dark magic?

$$\leq e^{-st\mu} \cdot \mathbb{E}(e^{sZ})$$

Markov

$$= e^{-st\mu} \cdot \mathbb{E}\left(\prod_i e^{sZ_i}\right)$$

Independence

$$= e^{-st\mu} \cdot \prod_i \mathbb{E}(e^{sZ_i})$$

Chernoff Bounds

Thm: $\mathbb{P}\left(Z \geq (1+\varepsilon)\mu\right) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu}$

$Z_i = \begin{cases} 1 \text{ with prob } p \\ 0 \text{ with prob } 1-p \end{cases}$

$Z = Z_1 + \dots + Z_m$

$Z_1, \dots, Z_m$ independent

$\mu = \mathbb{E}(Z) = pm$

Proof Cont'd:

Balls and Bins: Maximum Load

- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \mathbb{P}(\max_i L_i \geq k)$
- Let $L_{i,j} = \begin{cases} 1 & \text{if ball } j \text{ is in bin } i \\ 0 & \text{otherwise} \end{cases}$ $\Rightarrow$ $\underbrace{L_i = L_{i,1} + \ldots + L_{i,m}}_{\text{Apply Chernoff Bound}}$