CS7800: Advanced Algorithms

Class 22: Randomized Algorithms III

- Balls and Bins: Chernoff Bounds
- Universal Hashing

Jonathan Ullman

November 25, 2025

Balls and Bins: Maximum Load

m balls n bins

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

So far:

① Trivial bound: $\mathbb{E}(\max L_i) \leq m$

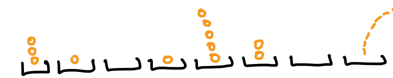
② Markov's inequality: $\mathbb{E}(\max L_i) \leq \infty$

③ Chebyshev's inequality: $\mathbb{E}(\max L_i) \leq O\left(\frac{m}{n} + \sqrt{m}\right)$

Union bound

Today: Tighter analysis with Chernoff Bounds $\mathbb{E}(\max L_i) = O\left(\frac{m}{n} + \frac{\log n}{\log \log n}\right)$

Balls and Bins: Maximum Load



m balls n bins

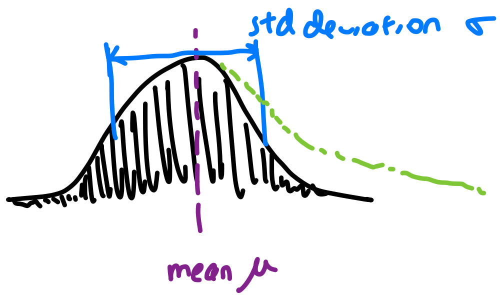
- Let L_i be the number of balls in bin i
- Expected maximum load is $\mathbb{E}(\max_i L_i) = \sum_{k=1}^{\infty} \mathbb{P}(\max_i L_i \geq k)$
- Let $L_{i,j} = \begin{cases} 1 & \text{if ball } j \text{ is in bin } i \\ 0 & \text{otherwise} \end{cases}$ $\Rightarrow L_i = L_{i,1} + \ldots + L_{i,m}$

Want to bound $\mathbb{P}(\max_i L_i \geq k) \leq n \cdot \mathbb{P}(L_i \geq k)$

L_i is a sum of m independent <u>simple</u> random variables

Aside: Central Limit Theorem

Gaussian distribution Z



$$\mathbb{P}(Z \geq \mu + 3\sigma) \leq .003$$

$$\mathbb{P}(Z \geq \mu + 6\sigma) \leq .000001 \ ?$$

$$\mathbb{P}(Z \geq \mu + t\sigma) \leq e^{-t^2}$$

$\sigma^2 = \mathrm{Var}(Z)$

Chebyshev $\mathbb{P}(Z \geq \mu + t\sigma) \leq 1/t^2$

If $Z_1, \dots, Z_m$ are independent random variables with $\mu = \mathbb{E}(Z_i)$, $\sigma^2 = \mathrm{Var}(Z_i)$ then $Z = \dfrac{Z_1 + \dots + Z_m}{\sqrt{m}}$ then

$Z \xrightarrow{m \to \infty}$ Gaussian with mean μ and variance σ^2

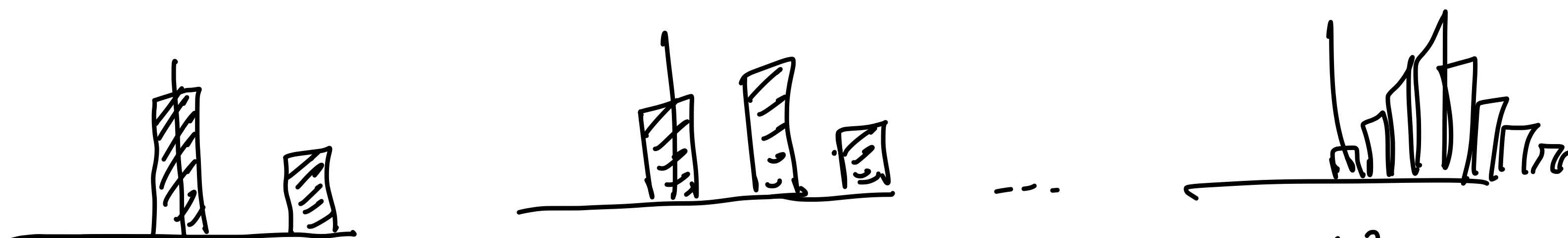
Aside: Central Limit Theorem

CLT:

If $z_1, \dots, z_m$ are independent random variables with $\mu = \mathbb{E}(z_i)$, $\sigma^2 = Var(z_i)$ then $Z = \dfrac{(z_1-\mu)+(z_2-\mu)+\dots+(z_m-\mu)}{\sqrt{\sigma^2 m}}$

$Z \xrightarrow{m\to\infty}$ Gaussian with mean 0 and variance 1

e.g. $z_i = \begin{cases} 1 & \text{wp } 1/3 \\ 0 & \text{wp } 2/3 \end{cases}$



z_i $\qquad$ z_1+z_2 $\qquad$ ... $\qquad$ $z_1+\dots+z_{100}$

Chernoff Bounds

$$Z_i = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases} \qquad Z = Z_1 + \dots + Z_m$$

$Z_1, \dots, Z_m$ independent $\qquad \mu = \mathbb{E}(Z) = pm$

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

$\mathbb{P}(Z - \mu \geq \varepsilon\mu)$

$\varepsilon < 1$

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

$\varepsilon > 1$

$e^{-\mu\varepsilon}$

$$\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \leq \left(\frac{e}{\varepsilon}\right)^{\varepsilon\mu}$$

Chernoff Bounds

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

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

$Z_1, \ldots, Z_m$ independent

$Z = Z_1 + \ldots + Z_m$

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

Proof:

$$\begin{aligned}
\mathbb{P}(Z \geq t\mu) &= \mathbb{P}(e^{sZ} \geq e^{st\mu}) \\
&\leq e^{-st\mu} \cdot \mathbb{E}(e^{sZ}) \\
&= e^{-st\mu} \cdot \mathbb{E}\left(\prod_i e^{sZ_i}\right) \\
&= e^{-st\mu} \cdot \prod_i \mathbb{E}(e^{sZ_i})
\end{aligned}$$

What is this dark magic?

Markov

Independence

$Z_i \in \{0,1\}$ so this is something "simple"

$$\begin{aligned}
\mathbb{E}(e^{sZ_i}) &= pe^{s} + (1-p) \\
&= 1 + p(e^{s} - 1)
\end{aligned}$$

Chernoff Bounds

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

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

$Z = Z_1 + \dots + Z_m$

$Z_1, \dots, Z_m$ independent

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

Proof Cont'd:

Balls and Bins: Maximum Load

m balls n bins

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

Apply Chernoff Bound

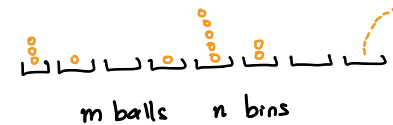
$$L_{i,j} = \begin{cases} 1 & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases} \qquad L_i = L_{i,1} + \ldots + L_{i,m} \qquad \mathbb{E}(L_i) = \frac{m}{n}$$

Chernoff: $\mathbb{P}\left(L_i \geq (1+\varepsilon)\frac{m}{n}\right) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{m/n}$

$\frac{m}{n}$ "big" $\quad \frac{m}{n} \geq \log n$

$$\mathbb{P}\left(L_i \geq (1+\varepsilon)\frac{m}{n}\right) \leq \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \leq e^{-\frac{\varepsilon^2 m/n}{4}} \rightsquigarrow \mathbb{E}(\max_i L_i) = O\left(\frac{m}{n}\right)$$

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 $L_{i,j} = \begin{cases} 1 & \text{if ball } j \text{ is in bin } i \\ 0 & \text{otherwise} \end{cases}$ $\Rightarrow$ $L_i = L_{i,1} + \dots + L_{i,m}$

Apply Chernoff Bound

$L_{i,j} = \begin{cases} 1 & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1-1/n \end{cases}$ $\quad L_i = L_{i,1} + \dots + L_{i,m}$ $\quad \mathbb{E}(L_i) = \frac{m}{n}$

Chernoff: $\mathbb{P}\left(L_i \geq (1+\varepsilon)\frac{m}{n}\right) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{m/n}$

$m/n = 1$ ("$\frac{m}{n}$ small")

$\mathbb{P}(L_i \geq 1+\varepsilon) \leq \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \leq \left(\frac{e}{\varepsilon}\right)^{\varepsilon} \rightsquigarrow \mathbb{E}(\max_i L_i) = O\left(\frac{\log n}{\log \log n}\right)$

Application: Hash Tables

"universe"

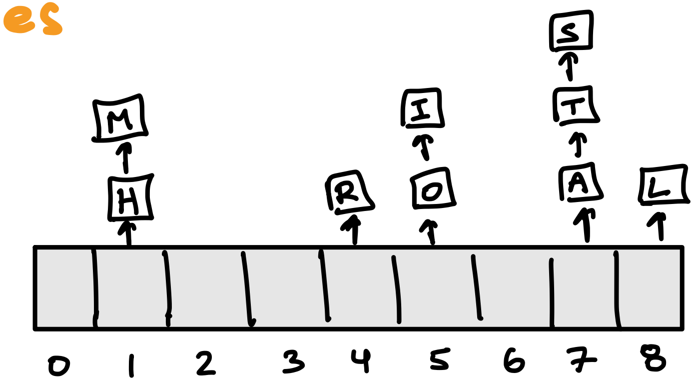
Goal: Store a set of m elements $S \subseteq \mathcal{U}$, such that we can efficiently check if $x \in S$

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

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

Application: Hash Tables

Linear chaining:
a common way to deal with collisions



Looks a lot like balls in bins!

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

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

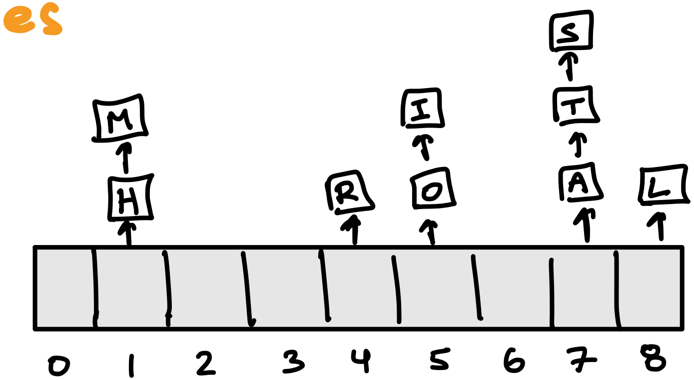
"collisions"

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

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

Application: Hash Tables

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



Looks a lot like balls in bins!

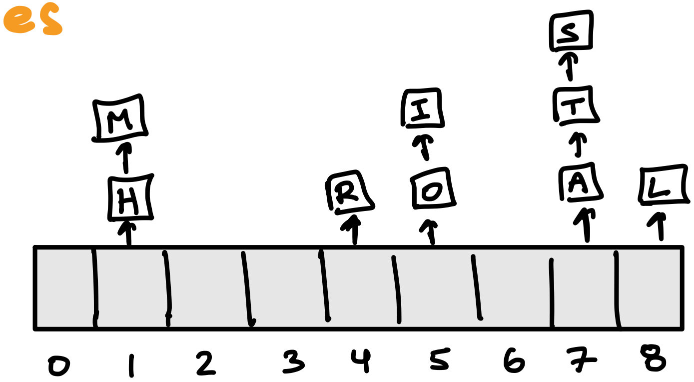
Randomized hash function:

- Model h as a uniformly random function $\mathcal{U} \to \{0,1,\dots,n-1\}$
- Fix the set S and study $\mathbb{E}\left(\max_x \ell(x)\right)$

expectation over random choice of h

Application: Hash Tables

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



Looks a lot like balls in bins!

Uniformly random hash functions: (load factor $\frac{m}{n} = 1$)

- Expected max load is $\Theta\left(\frac{\log n}{\log\log n}\right)$ [Will be true 99.999% of the time]

- For any $x \in \mathcal{U}$, expected lookup time is

$\mathbb{P}(h(x)=h(y)) = 1/n$

$$\mathbb{E}(\ell(x)) = \mathbb{E}(\#y \in S \text{ s.t. } h(x)=h(y)) = \sum_{y \in S} \mathbb{E}(1_{\text{if } h(y)=h(x)}) = n \cdot \left(\frac{1}{n}\right) = 1$$

Universal Hash Families

A <u>hash family</u> is a set of hash functions

$$\mathcal{H} \subseteq \{ h : \mathcal{U} \to \{0, 1, \ldots, n-1\} \}$$

<u>Definition</u>: $\mathcal{H}$ is 2-universal if for every distinct $x, y \in \mathcal{U}$

$$\mathbb{P}_{h}\left(h(x) = h(y) \right) \leq \frac{1}{n}$$

Choose h from $\mathcal{H}$ with equal probability

Behaves like a uniformly random hash fn if we only look at pairs of points

Constructing Universal Hashing

$\mathcal{H}$ is 2-universal if for every distinct $x, y \in \mathcal{U}$

$\mathbb{P}_h(h(x) = h(y)) \leq \frac{1}{n}$

Construction:

- Fix a prime $p \geq |\mathcal{U}|$, bins n
- Let $h_{a,b}(x) \equiv (ax + b \mod p) \mod n$

$$\mathcal{H}_{p,n} = \{ h_{a,b} : a \in \mathbb{Z}_p^{\neq 0}, b \in \mathbb{Z}_p \}$$

$|\mathcal{H}_{all}| = n^{|\mathcal{U}|} \geq n^p$

$|\mathcal{H}_{p,n}| = (p-1)p$

Theorem: $\mathcal{H}_{p,n}$ is a 2-universal hash family

Constructing Universal Hashing

$\mathcal{H}$ is 2-universal if for every distinct $x, y \in \mathcal{U}$

$\mathbb{P}_h(h(x) = h(y)) \leq \frac{1}{n}$

$h_{a,b}(x) = (ax + b \mod p) \mod n$

a,b random
$a \neq 0$

p fixed

Lemma 1: For every prime p and $a \neq 0$ there is a unique $a^{-1} \in \{1, 2, \ldots, p-1\}$ such that $a^{-1} \cdot a = 1 \mod p$

Proof: ① $az = c \mod p$ has at most one solution

if $az = az' \mod p$ for $z, z' \in \{1, 2, \ldots, p-1\}$ $\Rightarrow$ $a(z - z') = 0 \mod p$

$\Rightarrow$ $z - z'$ divisible by p

$\Rightarrow$ $z - z' = 0$

② $az = 0 \mod p$ has no solutions

Constructing Universal Hashing

$\mathcal{H}$ is 2-universal if for every distinct $x, y \in \mathcal{U}$

$$\mathbb{P}_h(h(x) = h(y)) \leq \frac{1}{n}$$

$$h_{a,b}(x) = (ax + b \mod p) \mod n$$

a,b random
$a \neq 0$

p fixed

Lemma 2: If $x \neq y$ and $r \neq s$ then the system

$$ax + b = r \mod p$$
$$ay + b = s \mod p$$

has a unique solution

Proof:

$$a = x^{-1}(r - b)$$

$$x^{-1}(r-b)y + b = s$$

$$b = s - x^{-1}(r-b)y$$

Constructing Universal Hashing

$\mathcal{H}$ is 2-universal if for every distinct $x, y \in \mathcal{U}$

$\mathbb{P}_h(h(x) = h(y)) \leq \frac{1}{n}$

$h_{a,b}(x) = (ax + b \mod p) \mod n$

a,b random
$a \neq 0$

p fixed

Thm: $\mathcal{H}_{p,n}$ is 2-universal

Proof: By Lemma 2,

$$\mathbb{P}_{a,b}(ax+b = r \mod p \wedge ay+b = s \mod p) = \frac{1}{p(p-1)}$$

$$\mathbb{P}_{a,b}(h_{a,b}(x) = h_{a,b}(y)) = \frac{N}{p(p-1)}$$

where N is number of $r \neq s$ in $\mathbb{Z}_p$ so that $r = s \mod n$

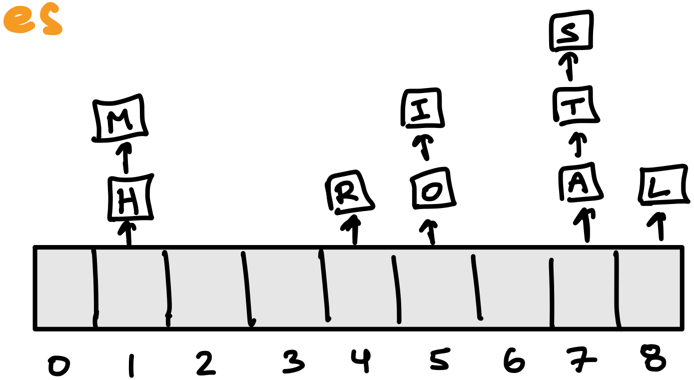
$$\leq \frac{p(p-1)}{n \cdot p(p-1)}$$

$$= \frac{1}{n}$$

$$N \leq \underbrace{p}_{\text{choices of } r} \cdot \underbrace{\frac{p-1}{n}}_{\text{choices of } s \text{ for a given } r}$$

Application: Hash Tables

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



Looks a lot like balls in bins!

Universal ~~Uniformly random~~ hash functions: (load factor $\frac{m}{n} = 1$)

- Expected max load is $\cancel{\Theta\left(\frac{\log n}{\log\log n}\right)}$ $\Theta(\sqrt{n})$
- For any $x \in \mathcal{U}$, expected lookup time is

$$\mathbb{E}(\ell(x)) = \sum_{y \in S} \mathbb{P}_h(h(x) = h(y)) \le n \cdot \frac{1}{n} = 1$$

Still true for universal hash families!