Lecture 3: Balls & Bins

Rafael Oliveira

University of Waterloo
Cheriton School of Computer Science

rafael.oliveira.teaching@gmail.com

September 16, 2020

Overview

- Introduction
  - Probability basic notions
  - Balls and Bins
  - Analyses
- Coupon Collector and Power of Two Choices
  - Coupon Collector
  - Power of Two Choices
- Acknowledgements

Basic notions

$[n] = \{1, 2, \ldots, n\}$

Expectation: if random variable X takes values in $[n]$

then $\mathbb{E}[X] = \sum_{i \in [n]} \Pr[X = i] \cdot i$

Conditional probability: if E and A are probability events

$\Pr[E] = \Pr[E \mid A] \cdot \Pr[A] + \Pr[E \mid \text{not } A] \cdot \Pr[\text{not } A]$

Balls and Bins Questions

Setup: we have m balls and we want to put them in n bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

Balls and Bins Questions

Setup: we have m balls and we want to put them in n bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

We are interested in the following questions:

Balls and Bins Questions

Setup: we have *m* balls and we want to put them in *n* bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

We are interested in the following questions:

- What is the *expected* number of balls in a bin?

Balls and Bins Questions

Setup: we have *m* balls and we want to put them in *n* bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

We are interested in the following questions:

- What is the *expected* number of balls in a bin?
- What is the *expected* number of empty bins?

Balls and Bins Questions

Setup: we have m balls and we want to put them in n bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

We are interested in the following questions:

- What is the *expected* number of balls in a bin?
- What is the *expected* number of empty bins?
- What is "typically" the *maximum* number of balls in any bin?

Balls and Bins Questions

Setup: we have *m* balls and we want to put them in *n* bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

We are interested in the following questions:

- What is the *expected* number of balls in a bin?
- What is the *expected* number of empty bins?
- What is "typically" the *maximum* number of balls in any bin?
- What is the *expected* number of bins with k balls in them?

Balls and Bins Questions

Setup: we have m balls and we want to put them in n bins.

As it is NBA playoff season, we will do this by throwing each ball into a *uniformly random* bin *independently*.

We are interested in the following questions:

- What is the *expected* number of balls in a bin?
- What is the *expected* number of empty bins?
- What is "typically" the *maximum* number of balls in any bin?
- What is the *expected* number of bins with k balls in them?
- For what values of m do we expect to have *no empty bins*? (coupon collector)

Why Learn About Balls and Bins?

In next lectures, we are going to learn about and analyse *randomized algorithms*. While we will usually analyse the *expected running times* of the algorithms, we would also like to know if the algorithm runs in time close to its expected running time *most of the time*.

Why Learn About Balls and Bins?

In next lectures, we are going to learn about and analyse *randomized algorithms*. While we will usually analyse the *expected running times* of the algorithms, we would also like to know if the algorithm runs in time close to its expected running time *most of the time*.

Running time *small* with *high probability* *better than* small expected running time.

Why Learn About Balls and Bins?

In next lectures, we are going to learn about and analyse *randomized algorithms*. While we will usually analyse the *expected running times* of the algorithms, we would also like to know if the algorithm runs in time close to its expected running time *most of the time*.

Running time *small* with *high probability* *better than* small expected running time.

In **this lecture**, we will analyse random processes (*balls & bins*) which underlie several randomized algorithms! (ranging from data structures to routing in parallel computers and beyond!)

Why Learn About Balls and Bins?

In next lectures, we are going to learn about and analyse *randomized algorithms*. While we will usually analyse the *expected running times* of the algorithms, we would also like to know if the algorithm runs in time close to its expected running time *most of the time*.

Running time *small* with *high probability* *better than* small expected running time.

In **this lecture**, we will analyse random processes (*balls & bins*) which underlie several randomized algorithms! (ranging from data structures to routing in parallel computers and beyond!)

After we learn about these basic processes, in lecture 7 (concentration inequalities) we will be concerned with statements of the first kind (what is the probability of deviating far from its expectation).

Expected Number of Balls in a Bin

Let us label the m balls $1, \ldots, m$, and the n bins $1, 2, \ldots, n$.
Let B_{ij} be the indicator variable that ball i was thrown into bin j.

Expected Number of Balls in a Bin

Let us label the m balls $1, \ldots, m$, and the n bins $1, 2, \ldots, n$.
Let B_{ij} be the indicator variable that ball i was thrown into bin j.

$$B_{ij} = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[\#\text{ balls in bin } j] = \mathbb{E}\left[\sum_{i=1}^{m} B_{i,j}\right]$$

Expected Number of Balls in a Bin

Let us label the m balls $1, \ldots, m$, and the n bins $1, 2, \ldots, n$.
Let B_{ij} be the indicator variable that ball i was thrown into bin j.

$$\begin{aligned}\mathbb{E}[\#\text{ balls in bin } j] &= \mathbb{E}\left[\sum_{i=1}^{m} B_{i,j}\right] \\ &= \sum_{i=1}^{m} \mathbb{E}\left[B_{i,j}\right] \qquad \text{(linearity of expectation)}\end{aligned}$$

Expected Number of Balls in a Bin

Let us label the m balls $1, \ldots, m$, and the n bins $1, 2, \ldots, n$.
Let B_{ij} be the indicator variable that ball i was thrown into bin j.

$$B_{ij} = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{o.w.} \end{cases}$$

$$\begin{aligned}
\mathbb{E}[\# \text{ balls in bin } j] &= \mathbb{E}\left[\sum_{i=1}^{m} B_{i,j}\right] \\
&= \sum_{i=1}^{m} \mathbb{E}\left[B_{i,j}\right] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{m} \Pr\left[\text{ball } i \text{ in bin } j\right]
\end{aligned}$$

Expected Number of Balls in a Bin

Let us label the m balls $1, \ldots, m$, and the n bins $1, 2, \ldots, n$.
Let B_{ij} be the indicator variable that ball i was thrown into bin j.

$$\begin{aligned}
\mathbb{E}[\#\text{ balls in bin } j] &= \mathbb{E}\left[\sum_{i=1}^{m} B_{i,j}\right] \\
&= \sum_{i=1}^{m} \mathbb{E}\left[B_{i,j}\right] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{m} \Pr\left[\text{ball } i \text{ in bin } j\right] \\
&= \sum_{i=1}^{m} \frac{1}{n} = \frac{m}{n} && \text{(uniformly at random)}
\end{aligned}$$

Expected Number of Balls in a Bin

Let us label the m balls $1, \ldots, m$, and the n bins $1, 2, \ldots, n$.
Let B_{ij} be the indicator variable that ball i was thrown into bin j.

$$
\begin{aligned}
\mathbb{E}[\#\text{ balls in bin } j] &= \mathbb{E}\left[\sum_{i=1}^{m} B_{i,j}\right] \\
&= \sum_{i=1}^{m} \mathbb{E}\left[B_{i,j}\right] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{m} \Pr\left[\text{ball } i \text{ in bin } j\right] \\
&= \sum_{i=1}^{m} \frac{1}{n} = \frac{m}{n} && \text{(uniformly at random)}
\end{aligned}
$$

When $m = n$, expectation of one ball per bin. How often will this actually happen?

Expected number of empty bins

Let N_i be the indicator variable that bin i is empty.

Expected number of empty bins

Let N_i be the indicator variable that bin i is empty.

$$N_i = \begin{cases} 1 & \text{if bin } i \text{ empty} \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbb{E}[\#\text{ empty bins}] = \mathbb{E}\left[\sum_{i=1}^{n} N_i\right]$$

Expected number of empty bins

Let N_i be the indicator variable that bin i is empty.

$$\begin{aligned}\mathbb{E}[\#\text{ empty bins}] &= \mathbb{E}\left[\sum_{i=1}^{n} N_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[N_i\right] && \text{(linearity of expectation)}\end{aligned}$$

Expected number of empty bins

Let N_i be the indicator variable that bin i is empty.

$$\begin{aligned}\mathbb{E}[\#\text{ empty bins}] &= \mathbb{E}\left[\sum_{i=1}^{n} N_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[N_i] && \text{(linearity of expectation)} \\ &= \sum_{i=1}^{n} \Pr[\text{bin } i \text{ is empty}]\end{aligned}$$

Expected number of empty bins

Let N_i be the indicator variable that bin i is empty.

$$
\begin{aligned}
\mathbb{E}[\#\text{ empty bins}] &= \mathbb{E}\left[\sum_{i=1}^{n} N_i\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[N_i\right] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{n} \Pr\left[\text{bin } i \text{ is empty}\right] \\
&= \sum_{i=1}^{n} (1 - 1/n)^m \\
&= n \cdot (1 - 1/n)^m \approx n \cdot e^{-m/n}
\end{aligned}
$$

1st ball not in bin i

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

m balls unif. random

Expected number of empty bins

Let N_i be the indicator variable that bin i is empty.

$$
\begin{aligned}
\mathbb{E}[\#\text{ empty bins}] &= \mathbb{E}\left[\sum_{i=1}^{n} N_i\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[N_i] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{n} \Pr[\text{bin } i \text{ is empty } j] \\
&= \sum_{i=1}^{n} (1 - 1/n)^m \\
&= n \cdot (1 - 1/n)^m \approx n \cdot e^{-m/n}
\end{aligned}
$$

n/e

When $m = n$, expected fraction of empty bins is $\frac{1}{e}$.

Head Scratching Moment

When $m = n$, first calculation had expectation of *one ball per bin*.

Head Scratching Moment

When $m = n$, first calculation had expectation of *one ball per bin*.

When $m = n$, second calculation had expectation of $1/e$ *fraction of empty bins*.

Head Scratching Moment

When $m = n$, first calculation had expectation of *one ball per bin*.

When $m = n$, second calculation had expectation of $1/e$ *fraction of empty bins*.

Which expectation should I actually "expect"?

Head Scratching Moment

When $m = n$, first calculation had expectation of *one ball per bin*.

When $m = n$, second calculation had expectation of *$1/e$ fraction of empty bins*.

with high prob.

Which expectation should I actually "expect"?

As we mentioned earlier, this is where *concentration of probability measure* tries to address. It turns out that the *second random variable* (and thus second calculation) is concentrated around the mean (i.e., expectation).

So we "expect" (or it is "typical") to see around $1/e$-fraction of empty bins when $m = n$

Maximum load in a bin

What is the "typical" maximum number of balls in a bin?

As we saw in the previous slide, "typical" is related to concentration of probability measure.

Maximum load in a bin

What is the “typical” maximum number of balls in a bin?

As we saw in the previous slide, “typical” is related to concentration of probability measure.

Let us first see a simpler problem, which is known as the *birthday paradox*: for what value of m do we expect to see two balls in one bin?

Birthday Paradox

The probability that there are no collisions after we have thrown m balls is:

$$1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdot \dots \cdot \left(1 - \frac{m-1}{n}\right) \leq e^{-1/n} \cdots e^{-\frac{m-1}{n}} \approx e^{\frac{-m^2}{2n}}$$

1st, 2nd, 3rd, mth

Birthday Paradox

The probability that there are no collisions after we have thrown m balls is:

$$1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdot \dots \cdot \left(1 - \frac{m-1}{n}\right) \leq e^{-1/n} \cdots e^{-\frac{m-1}{n}} \approx e^{\frac{-m^2}{2n}}$$

This is $\leq 1/2$ when $m = \sqrt{2n\ln(2)}$. For $n = 365$, this is $m \approx 22.4$ for the probability that two people *(balls)* have birthday on the same date *(bins)* to become $\geq 1/2$.

Birthday Paradox

The probability that there are no collisions after we have thrown m balls is:

$$1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \cdot \left(1 - \frac{m-1}{n}\right) \leq e^{-1/n} \cdots e^{-\frac{m-1}{n}} \approx e^{\frac{-m^2}{2n}}$$

This is $\leq 1/2$ when $m = \sqrt{2n\ln(n)}$. For $n = 365$, this is $m \approx 22.4$ for the probability that two people *(balls)* have birthday on the same date *(bins)* to become $\geq 1/2$.

Thus, we expect to see a collision (two balls in the same bin) when $m = \Theta(\sqrt{n})$. This appears in several places, such as hashing, factoring, etc.

Maximum load in a bin when $m = n$

What is the probability that a particular bin (say bin 1) has $\geq k$ balls in it?

Maximum load in a bin when $m = n$

What is the probability that a particular bin (say bin 1) has $\geq k$ balls in it?

$$\Pr[\text{bin 1 has } \geq k \text{ balls}] \leq \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \Pr[\text{ball } i \text{ in bin 1}]$$

union bound

Pr[balls in S fall in 1]

Maximum load in a bin when $m = n$

What is the probability that a particular bin (say bin 1) has $\geq k$ balls in it?

$$\Pr[\text{bin 1 has } \geq k \text{ balls}] \leq \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \Pr[\text{ball } i \text{ in bin 1}]$$

$$= \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \frac{1}{n}$$

uniformly at random

Maximum load in a bin when $m = n$

What is the probability that a particular bin (say bin 1) has $\geq k$ balls in it?

$$\begin{aligned}
\Pr[\text{bin 1 has } \geq k \text{ balls}] &\leq \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \Pr[\text{ball } i \text{ in bin 1}] \\
&= \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \frac{1}{n} \\
&= \binom{n}{k} \cdot \frac{1}{n^k} \leq \left(\frac{ne}{k}\right)^k \cdot \frac{1}{n^k} = \frac{e^k}{k^k}
\end{aligned}$$

$1/n^k$

Maximum load in a bin when $m = n$

What is the probability that a particular bin (say bin 1) has $\geq k$ balls in it?

$$\Pr[\text{bin 1 has } \geq k \text{ balls}] \leq \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \Pr[\text{ball } i \text{ in bin 1}]$$

$$= \sum_{\substack{S \ subset[n] \\ |S|=k}} \prod_{i \in S} \frac{1}{n}$$

$$= \binom{n}{k} \cdot \frac{1}{n^k} \leq \left(\frac{ne}{k}\right)^k \cdot \frac{1}{n^k} = \frac{e^k}{k^k}$$

By union bound

union bound

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq \sum_{i=1}^{n} \Pr[\text{bin i has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k}$$

Maximum load in a bin when $m = n$

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k \ln k}$$

Maximum load in a bin when $m = n$

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k \ln k}$$

$$\Pr[\text{max load is } \leq k] = 1 - \Pr[\text{some bin has } \geq k \text{ balls}] \geq 1 - e^{\ln n + k - k \ln k}$$

complement

small

large

Maximum load in a bin when $m = n$

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k \ln k}$$

$>$ still true

$$\Pr[\text{max load is } \leq k] = 1 - \Pr[\text{some bin has } \geq k \text{ balls}] \geq 1 - e^{\ln n + k - k \ln k}$$

When will the above probability be large (say $>> 1/2$)?

Maximum load in a bin when $m = n$

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k \ln k}$$

$$\Pr[\text{max load is } \leq k] = 1 - \Pr[\text{some bin has } \geq k \text{ balls}] \geq 1 - e^{\ln n + k - k \ln k}$$

$\sim 3 \ln n$

$\frac{3 \ln n}{\ln(\ln(n))}$

When will the above probability be large (say $>> 1/2$)?

When $k \ln k > \ln n$. Setting $k = 3 \frac{\ln n}{\ln \ln n}$ does it.

Maximum load in a bin when $m = n$

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k \ln k}$$

$$\Pr[\text{max load is } \leq k] = 1 - \Pr[\text{some bin has } \geq k \text{ balls}] \geq 1 - e^{\ln n + k - k \ln k}$$

When will the above probability be large (say $>> 1/2$)?

When $k \ln k > \ln n$. Setting $k = 3\frac{\ln n}{\ln \ln n}$ does it.

With high probability, max load is $O\left(\frac{\ln n}{\ln \ln n}\right)$.

typical behaviour

Maximum load in a bin when $m = n$

$$\Pr[\text{some bin has } \geq k \text{ balls}] \leq n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k \ln k}$$

$$\Pr[\text{max load is } \leq k] = 1 - \Pr[\text{some bin has } \geq k \text{ balls}] \geq 1 - e^{\ln n + k - k \ln k}$$

When will the above probability be large (say $>> 1/2$)?

When $k \ln k > \ln n$. Setting $k = 3\frac{\ln n}{\ln \ln n}$ does it.

With high probability, max load is $O\left(\frac{\ln n}{\ln \ln n}\right)$.

This comes up in hashing and in analysis of approximation algorithms (for instance, best known approximation ratio for congestion minimization).

Coupon Collector

For what value of m do we expect to have no empty bins?

Coupon Collector

For what value of m do we expect to have no empty bins?

Why is this problem called the coupon collector problem?

Because we can formulate it in the following way:

- suppose each bin is a different coupon
- we buy one coupon at random (like kinder eggs/pack action cards)
- what is the number of coupons that we need to buy to collect all of them?

Coupon Collector

For what value of m do we expect to have no empty bins?

Why is this problem called the coupon collector problem?

Because we can formulate it in the following way:

- suppose each bin is a different coupon
- we buy one coupon at random (like kinder eggs/pack action cards)
- what is the number of coupons that we need to buy to collect all of them?

Let X_i be the number of balls thrown to get from $i+1$ empty bins to i empty bins. Let X be the number of balls thrown until we have no empty bins.

$$X = \sum_{i=0}^{n-1} X_i$$

Coupon Collector

- $X_i \leftarrow \#$ balls thrown to get from i empty bins to $i-1$ empty bins
- $X \leftarrow \#$ balls thrown until we have no empty bins

Coupon Collector

- $X_i \leftarrow \#$ balls thrown to get from i empty bins to $i-1$ empty bins
- $X \leftarrow \#$ balls thrown until we have no empty bins

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]$$

Coupon Collector

- $X_i \leftarrow$ # balls thrown to get from i empty bins to $i-1$ empty bins
- $X \leftarrow$ # balls thrown until we have no empty bins

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]$$

What is $\mathbb{E}[X_i]$?

Coupon Collector

- $X_i \leftarrow$ # balls thrown to get from i empty bins to $i-1$ empty bins
- $X \leftarrow$ # balls thrown until we have no empty bins

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

What is $\mathbb{E}[X_i]$?

success prob.

X_i geometric random variable with parameter $p = \frac{i}{n}$.

Number of trials until the first success, where success probability p.

In X_i setting have only i empty bins

Coupon Collector

- $X_i \leftarrow \#$ balls thrown to get from i empty bins to $i-1$ empty bins
- $X \leftarrow \#$ balls thrown until we have no empty bins

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]$$

What is $\mathbb{E}[X_i]$?

X_i geometric random variable with parameter $p = \frac{i}{n}$.

Number of trials until the first success, where success probability p.

failed in attempts 1, .., k-1

$$\Pr[X_i = k] = (1-p)^{k-1} \cdot p$$ ← succeded in k

Threw k balls
k^{th} ball landed on empty bin

Coupon Collector - Computing $\mathbb{E}[X]$

X_i takes values in $\mathbb{N}^*$

$$\mathbb{E}[X_i] = \sum_{k=1}^{\infty} k \cdot \Pr[X_i = k] = \sum_{k=1}^{\infty} \Pr[X_i \geq k] = \sum_{k=1}^{\infty} (1-p)^{k-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$$

rearrange sum

$\geq k-1$, $\geq k-2$, $\vdots$, ≥ 1

geom. sum

geom. sum

$p = \frac{i}{n}$ for X_i

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{n}{i} = n \cdot \underbrace{\sum_{i=1}^{n} \frac{1}{i}}_{H_n} \approx n \ln n .$$

Coupon Collector - Computing $\mathbb{E}[X]$

This $n \ln n$ bound shows up in cover time of random walks in complete graph, number of edges needed in graph sparsification, etc.

Power of Two Choices

We now know that when n balls are thrown into n bins, the maximum load is $\Theta(\ln n / \ln \ln n)$ with high probability (we'll maybe see lower bound later).

Power of Two Choices

We now know that when n balls are thrown into n bins, the maximum load is $\Theta(\ln n / \ln \ln n)$ with high probability (we'll maybe see lower bound later).

Consider following variant: what if when throwing a ball in a bin, *before* we throw the ball we choose *two* bins *uniformly at random* and put the ball in the *bin with fewer balls*?

Power of Two Choices

We now know that when n balls are thrown into n bins, the maximum load is $\Theta(\ln n / \ln \ln n)$ with high probability (we'll maybe see lower bound later).

Consider following variant: what if when throwing a ball in a bin, *before* we throw the ball we choose *two* bins *uniformly at random* and put the ball in the *bin with fewer balls*?

This simple modification reduces maximum load to $O(\ln \ln n)$!

Power of Two Choices

We now know that when n balls are thrown into n bins, the maximum load is $\Theta(\ln n / \ln \ln n)$ with high probability (we'll maybe see lower bound later).

Consider following variant: what if when throwing a ball in a bin, *before* we throw the ball we choose *two* bins *uniformly at random* and put the ball in the *bin with fewer balls*?

This simple modification reduces maximum load to $O(\ln \ln n)$!

Intuition/idea: let the height of a bin be the # balls in that bin. This process tells us that to get one bin with height $h+1$ we must have at least two bins of height h.

We can bound # bins with height at least h (because this will tell us how likely it is to get to height $h+1$).

A bit more intuition

$N_h :=$ number of bins with height at least h

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \underbrace{\left(\frac{N_h}{n}\right)^2}$$

Pr selecting two bins of height h

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \left(\frac{N_h}{n}\right)^2$$

- Say we have only $n/4$ bins with 4 items (i.e. height 4)

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \left(\frac{N_h}{n}\right)^2$$

- Say we have only $n/4$ bins with 4 items (i.e. height 4)
- Probability of selecting two such bins is $1/16$

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \left(\frac{N_h}{n}\right)^2$$

- Say we have only $n/4$ bins with 4 items (i.e. height 4)
- Probability of selecting two such bins is $1/16$
- So we should expect only $n/16$ bins with height 5
- And only $n/256 = n/16^2 = n/2^{2^3}$ bins with height 6

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \left(\frac{N_h}{n}\right)^2$$

- Say we have only $n/4$ bins with 4 items (i.e. height 4)
- Probability of selecting two such bins is $1/16$
- So we should expect only $n/16$ bins with height 5
- And only $n/256 = n/16^2 = n/2^{2^3}$ bins with height 6
- Repeating this, we should expect $\frac{n}{2^{2^{h-3}}}$ bins of height h

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \left(\frac{N_h}{n}\right)^2$$

- Say we have only $n/4$ bins with 4 items (i.e. height 4)
- Probability of selecting two such bins is $1/16$
- So we should expect only $n/16$ bins with height 5
- And only $n/256 = n/16^2 = n/2^{2^3}$ bins with height 6
- Repeating this, we should expect $\frac{n}{2^{2^{h-3}}}$ bins of height h
- So expect $\log \log n$ maximum height after throwing n balls.

A bit more intuition

$N_h :=$ number of bins with height at least h

$$\Pr[\text{at least one bin of height } h+1] \leq \left(\frac{N_h}{n}\right)^2$$

- Say we have only $n/4$ bins with 4 items (i.e. height 4)
- Probability of selecting two such bins is $1/16$
- So we should expect only $n/16$ bins with height 5
- And only $n/256 = n/16^2 = n/2^{2^3}$ bins with height 6
- Repeating this, we should expect $\frac{n}{2^{2^{h-3}}}$ bins of height h
- So expect $\log\log n$ maximum height after throwing n balls.

How do we turn this into a proof?

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

Use following Chernoff bound[1] on binomial random variable $B(n,p)$ with n trials and success probability p.[2]

$$\Pr[B(n,p) \geq 2np] \leq e^{-np/3}$$

[1] we will see Chernoff in lecture 7

[2] That is, $\Pr[B(n,p) = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

Use following Chernoff bound[1] on binomial random variable $B(n,p)$ with n trials and success probability p.[2]

$$\Pr[B(n,p) \geq 2np] \leq e^{-np/3}$$

- $\beta_4 := n/4$ and $\beta_{i+1} = 2\beta_i^2/n$.

[1] we will see Chernoff in lecture 7

[2] That is, $\Pr[B(n,p) = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

Use following Chernoff bound[1] on binomial random variable $B(n, p)$ with n trials and success probability p.[2]

$$\Pr[B(n,p) \geq 2np] \leq e^{-np/3}$$

- $\beta_4 := n/4$ and $\beta_{i+1} = 2\beta_i^2/n$.
- $E(h, t) :=$ event that after all t balls are thrown, $N_h \leq \beta_h$

[1] we will see Chernoff in lecture 7

[2] That is, $\Pr[B(n,p) = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

Use following Chernoff bound[1] on binomial random variable $B(n, p)$ with n trials and success probability p.[2]

$$\Pr[B(n,p) \geq 2np] \leq e^{-np/3}$$

- $\beta_4 := n/4$ and $\beta_{i+1} = 2\beta_i^2/n$.
- $E(h, t) :=$ event that after all t balls are thrown, $N_h \leq \beta_h$
- $\Pr[E(4, n)] = 1$ (why?)

$n \geq N_4 \cdot 4 \Rightarrow N_4 \leq n/4 = \beta_4$

[1] we will see Chernoff in lecture 7

[2] That is, $\Pr[B(n,p) = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

Use following Chernoff bound[1] on binomial random variable $B(n, p)$ with n trials and success probability p.[2]

$$\Pr[B(n, p) \geq 2np] \leq e^{-np/3}$$

- $\beta_4 := n/4$ and $\beta_{i+1} = 2\beta_i^2/n$.
- $E(h, t) :=$ event that after all t balls are thrown, $N_h \leq \beta_h$
- $\Pr[E(4, n)] = 1$
- We will prove that if $E(h, n)$ holds with high probability then so does $E(h+1, n)$ (so long as h is "small enough")

(see what small is later)

[1] we will see Chernoff in lecture 7

[2] That is, $\Pr[B(n, p) = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

- $Y_t(h)$ be the indicator variable that t^{th} ball has height $\geq h+1$ (i.e., was placed in a bin that had height h)
- $\Pr[Y_t(h) = 1 \mid E(h,t)] \leq \left(\frac{N_h}{n}\right)^2 \leq \frac{\beta_h^2}{n^2}$
- If $p_i := \frac{\beta_h^2}{n^2}$ then

$$\Pr\left[\sum_{t=1}^{n} Y_t(h) > k \mid E(h,n)\right] \leq \Pr\left[\sum_{t=1}^{n} B(n,p_i) > k \mid E(h,n)\right]$$

-

$$\Pr[N_{h+1} > k \mid E(h,n)] = \Pr\left[\sum_{t=1}^{n} Y_t(h) > k \mid E(h,n)\right]$$

$$\leq \Pr\left[\sum_{t=1}^{n} B(n,p_i) > k \mid E(h,n)\right]$$

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[N_{h+1} > k \mid E(h, n)] \leq \Pr\left[\sum_{t=1}^{n} B(n, p_i) > k \mid E(h, n)\right]$$

total # bins of height h (β_h)

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[N_{h+1} > k \mid E(h,n)] \leq \Pr\left[\sum_{t=1}^{n} B(n, p_i) > k \mid E(h,n)\right]$$

Setting $k = \beta_{h+1} = 2np_h$ above, we get

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[N_{h+1} > k \mid E(h,n)] \leq \Pr\left[\sum_{t=1}^{n} B(n,p_h) > k \mid E(h,n)\right]$$

Setting $k = \beta_{h+1} = 2np_h$ above, we get

$$\begin{aligned}
\Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] &\leq \Pr\left[\sum_{t=1}^{n} B(n,p_h) > \beta_{h+1} \mid E(h,n)\right] \\
&\leq \frac{\Pr\left[\sum_{t=1}^{n} B(n,p_h) > \beta_{h+1}\right]}{\Pr[E(h,n)]} \\
&\leq \frac{1}{\Pr[E(h,n)] \cdot e^{np_h/3}} \qquad \text{(Chernoff)}
\end{aligned}$$

conditional probability

$= 2np_h$

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[N_{h+1} > k \mid E(h,n)] \leq \Pr\left[\sum_{t=1}^{n} B(n,p_i) > k \mid E(h,n)\right]$$

Setting $k = \beta_{h+1} = 2np_h$ above, we get

$$\begin{aligned}
\Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] &\leq \Pr\left[\sum_{t=1}^{n} B(n,p_i) > \beta_{h+1} \mid E(h,n)\right] \\
&\leq \frac{\Pr\left[\sum_{t=1}^{n} B(n,p_i) > \beta_{h+1}\right]}{\Pr[E(h,n)]} \\
&\leq \frac{1}{\Pr[E(h,n)] \cdot e^{np_h/3}} \qquad \text{(Chernoff)}
\end{aligned}$$

not $E(h+1, n)$

$\approx n^2$

Thus, setting $p_h \cdot n \geq 6 \ln n$ we get

$$\Pr[\text{not } E(h+1,n) \mid E(h,n)] = \Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] \leq \frac{1}{n^2 \Pr[E(h,n)]}$$

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[\text{not } E(h+1,n) \mid E(h,n)] = \Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] \leq \frac{1}{n^2 \Pr[E(h,n)]}$$

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[\text{not } E(h+1,n) \mid E(h,n)] = \Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] \leq \frac{1}{n^2 \Pr[E(h,n)]}$$

Now, to bound the final probability, we have:

- $p_h \cdot n \geq 6 \ln n$ for $h = O(\ln \ln n)$ (easy calculation - we did it)
- Handle the case where $p_h \cdot n < 6 \ln n$. (another Chernoff bound - see Lap Chi's notes)

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[\text{not } E(h+1,n) \mid E(h,n)] = \Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] \le \frac{1}{n^2 \Pr[E(h,n)]}$$

Now, to bound the final probability, we have:

→ conditional probability

$$\Pr[\text{not } E(h+1,n)] = \Pr[\text{not } E(h+1,n) \mid E(h,n)] \cdot \Pr[E(h,n)]$$
$$+ \Pr[\text{not } E(h+1,n) \mid \text{not } E(h,n)] \cdot \Pr[\text{not } E(h,n)]$$

bad event (have many bins with height h+1)

- $p_h \cdot n \ge 6 \ln n$ for $h = O(\ln \ln n)$ (easy calculation - we did it)
- Handle the case where $p_h \cdot n < 6 \ln n$. (another Chernoff bound - see Lap Chi's notes)

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[\text{not } E(h+1,n) \mid E(h,n)] = \Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] \leq \frac{1}{n^2 \Pr[E(h,n)]}$$

Now, to bound the final probability, we have:

$$\begin{aligned}\Pr[\text{not } E(h+1,n)] &= \Pr[\text{not } E(h+1,n) \mid E(h,n)] \cdot \Pr[E(h,n)] \\ &+ \Pr[\text{not } E(h+1,n) \mid E(h,n)] \cdot \Pr[\text{not } E(h,n)] \\ &\leq \frac{1}{n^2} + \Pr[\text{not } E(h,n)] \qquad \text{(so long as } p_h n \geq 6 \ln n\text{)}\end{aligned}$$

- $p_h \cdot n \geq 6 \ln n$ for $h = O(\ln \ln n)$ (easy calculation - we did it)
- Handle the case where $p_h \cdot n < 6 \ln n$. (another Chernoff bound - see Lap Chi's notes)

Proof Sketch (proof [Mitzenmacher & Upfal, Chapter 14])

$$\Pr[\text{not } E(h+1,n) \mid E(h,n)] = \Pr[N_{h+1} > \beta_{h+1} \mid E(h,n)] \leq \frac{1}{n^2 \Pr[E(h,n)]}$$

Now, to bound the final probability, we have: *if E(h, n) happens whp then so does E(h+1, n)*

$$\begin{aligned}\Pr[\text{not } E(h+1,n)] &= \Pr[\text{not } E(h+1,n) \mid E(h,n)] \cdot \Pr[E(h,n)] \\ &\quad + \Pr[\text{not } E(h+1,n) \mid E(h,n)] \cdot \Pr[\text{not } E(h,n)] \\ &\leq \frac{1}{n^2} + \Pr[\text{not } E(h,n)] \quad (\text{so long as } p_h n \geq 6 \ln n)\end{aligned}$$

small

To finish the proof, need to show:

- $p_h \cdot n \geq 6 \ln n$ for $h = O(\ln \ln n)$ (easy calculation - we did it)
- Handle the case where $p_h \cdot n < 6 \ln n$. (another Chernoff bound - see Lap Chi's notes)

Acknowledgement

- Lecture based largely on Lap Chi's notes and on [Motwani & Raghavan 2007, Chapter 3].
- See Lap Chi's notes at `https://cs.uwaterloo.ca/~lapchi/cs466/notes/L04.pdf`

References I

Motwani, Rajeev and Raghavan, Prabhakar (2007)
Randomized Algorithms

Mitzenmacher, Michael, and Eli Upfal (2017)
Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.
Cambridge university press, 2017.