Lecture 4: Balls & Bins

Rafael Oliveira

University of Waterloo
Cheriton School of Computer Science

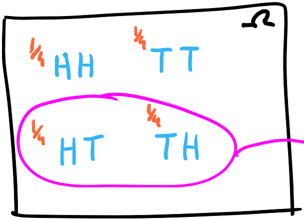
rafael.oliveira.teaching@gmail.com

September 16, 2021

Overview

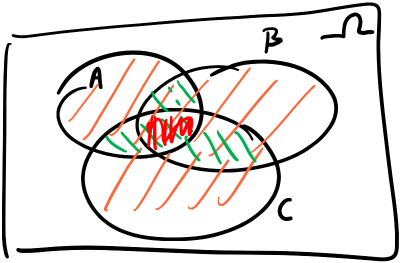
Event Spaces and Inclusion-Exclusion

Ω - event space (outcomes of tossing fair coin twice unbiased/independent)



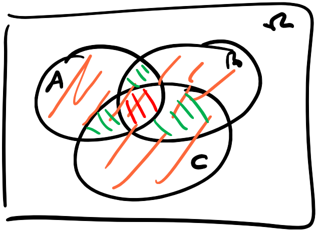
event : exactly 1 H come up

orange region: $A \cup B \cup C$



$$\Pr[A \cup B \cup C] = \Pr[A] + \Pr[B] + \Pr[C] - \Pr[A \cap B] - \Pr[A \cap C] - \Pr[B \cap C] + \Pr[A \cap B \cap C]$$

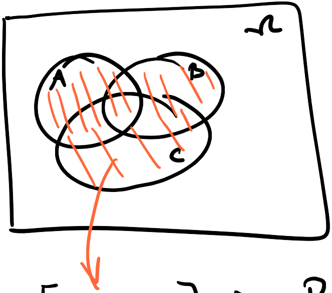
Union Bound and Inclusion-Exclusion

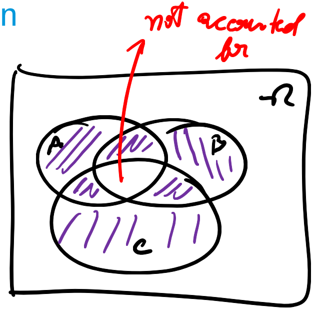


$$\Pr[\text{at least one } A, B \text{ or } C \text{ happens}] \stackrel{\Delta}{=} \Pr[A \cup B \cup C]$$

$$\leq \Pr[A] + \Pr[B] + \Pr[C]$$

Union Bound and Inclusion-Exclusion



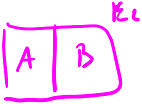


$$\Pr[A \cup B \cup C] \geq \Pr[A] + \Pr[B] + \Pr[C] - \Pr[A \cap B] - \Pr[A \cap C] - \Pr[B \cap C]$$

we can generalize this to any level of intersections

Conditional Probability and Bayes Rule

- The *conditional probability* of E_1 given E_2 is



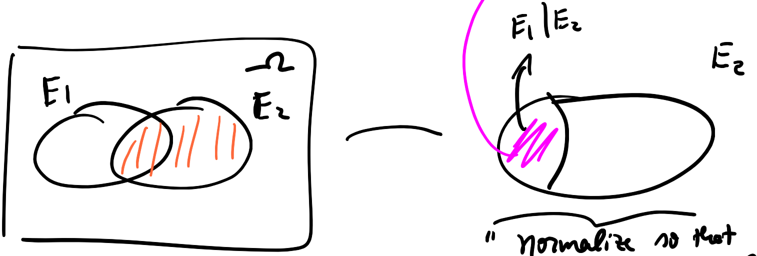
$$\Pr[E_1 \mid E_2] := \frac{\Pr[E_1 \cap E_2]}{\Pr[E_2]}$$

$\Pr[E_2] = \Pr[A] + \Pr[B]$

$\Pr[A \mid E_2] = \frac{\Pr[A]}{\Pr[E_2]}$

Probability of E_1 happens given that event E_2 happened

both E_1 and E_2 happens



" normalize so that E_2 has" Pr 1

Conditional Probability and Bayes Rule

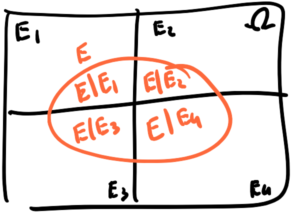
- The *conditional probability* of E_1 given E_2 is

$$\Pr[E_1 \mid E_2] := \frac{\Pr[E_1 \cap E_2]}{\Pr[E_2]}$$

- If $E_1, \ldots, E_k$ partition our sample space, then for any event E

$$\Pr[E] = \sum_{i=1}^{k} \Pr[E \mid E_i] \cdot \Pr[E_i]$$

$\Pr[E \cap E_i]$



Conditional Probability and Bayes Rule

- The *conditional probability* of E_1 given E_2 is

$$\Pr[E_1 \mid E_2] := \frac{\Pr[E_1 \cap E_2]}{\Pr[E_2]}$$

- If $E_1, \ldots, E_k$ partition our sample space, then for any event E

$$\Pr[E] = \sum_{i=1}^{k} \Pr[E \mid E_i] \cdot \Pr[E_i]$$

- **Simple Bayes' rule**:

Pr[E1 ∩ E2]

$$\Pr[E_1 \mid E_2] = \frac{\Pr[E_2 \mid E_1] \cdot \Pr[E_1]}{\Pr[E_2]}$$

Conditional Probability and Bayes Rule

- The *conditional probability* of E_1 given E_2 is

$$\Pr[E_1 \mid E_2] := \frac{\Pr[E_1 \cap E_2]}{\Pr[E_2]}$$

- If $E_1, \ldots, E_k$ partition our sample space, then for any event E

$$\Pr[E] = \sum_{i=1}^{k} \Pr[E \mid E_i] \cdot \Pr[E_i]$$

- **Simple Bayes' rule**:

$$\Pr[E_1 \mid E_2] = \frac{\Pr[E_2 \mid E_1] \cdot \Pr[E_1]}{\Pr[E_2]}$$

- **Bayes' rule**: $E_1, \ldots, E_k$ partition our sample space then for event E

$$\Pr[E_i \mid E] = \frac{\Pr[E \cap E_i]}{\Pr[E]} = \frac{\Pr[E \mid E_i] \cdot \Pr[E_i]}{\sum_{j=1}^{k} \Pr[E \mid E_j] \cdot \Pr[E_j]}$$

Balls and Bins Questions

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

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.

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.

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.

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.

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?

maximum load

Balls and Bins Questions

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

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.

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.

Strategy: devise randomized algorithm
with good expected running time
prove concentration of measure
around expectation

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!

Applications ranging from:

1. data structures
2. routing in parallel computers
3. many more!

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.

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

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.

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

all balls

Expected Number of Balls in a Bin

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

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

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.

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

$$\mathbb{E}[B_{ij}] = 1 \cdot \Pr[\text{ball } i \text{ in bin } j] + 0 \cdot (1 - \ldots)$$

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.

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

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.

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

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. after m throws

Expected number of empty bins

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

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

Expected number of empty bins

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

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

$$= \sum_{i=1}^{n} \mathbb{E}[N_i] \qquad \text{(linearity of expectation)}$$

Expected number of empty bins

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

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

$$\Pr[b \text{ does not fall in } i] = \left(1 - \frac{1}{n}\right)$$

$$\Pr[\text{bin } i \text{ empty}] = \left(1 - 1/n\right)^m$$

Expected number of empty bins

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

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

$\left(1-\frac{1}{n}\right)^n \sim e^{-1}$

Expected number of empty bins

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

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

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"?

(example where expectation ≠ typical)

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"?

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:

third ball

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

never any collision (only 1 ball)

second ball

k balls in bin

$1 - \frac{k}{n}$ $(k+1)^{th}$ ball doesn't cause collision

(avoid k occupied bins)

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:

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

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 $\geq 1/2$.

Thus, expect to see collision (two balls in the same bin) when $m = \Theta(\sqrt{n})$. This appears in several places:

- hashing
- factoring
- many more

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?

union bound

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

$\cup$ (balls in $S \to$ bin 1), $S \subset [n]$, $|S| = k$

S ← set of k balls in bin 1

all possible sets of k balls

$$\Pr[S \to \text{bin } 1] = \prod_{i \in S} \Pr[\text{ball } i \to \text{bin } 1]$$

$$= \prod_{i \in S} \frac{1}{n} = \frac{1}{n^k}$$

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?

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

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

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?

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

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

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

↑ # subsets of size 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?

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

By 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 $m = n$

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

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

all bins have $\leq k$ load

Maximum load in a bin when $m = n$

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

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

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

Maximum load in a bin when $m = n$

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

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

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.

Maximum load in a bin when $m = n$

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

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

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 $O\left(\frac{\ln n}{\ln \ln n}\right)$.

Maximum load in a bin when $m = n$

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

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

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).

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

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

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

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

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

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

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

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

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

fraction of empty bins

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

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

k-th attemp succeed

fail in first k-1 attempts

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

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

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

rearrange

$$\Pr[X_i \geq k] = \sum_{j=k}^{\infty} \Pr[X_i = j]$$

$$= \sum_{k=1}^{\infty} (1-p)^{k-1} = \frac{1}{1-(1-p)} = \frac{1}{p} = \frac{n}{i}$$

$$\mathbb{E}[X_i] = 1 + \Pr[\text{first throw fails}] \cdot \mathbb{E}[X_i]$$

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

$$\Pr[X = k] = (1-p)^{k-1} \cdot p$$

$$1 \cdot p + \sum_{k=2}^{\infty} k \cdot (1-p)^{k-1} \cdot p$$

$$\sum_{j=1}^{\infty} (j+1) \cdot (1-p)^{j} \cdot p$$

$$= (1-p) \sum_{j=1}^{\infty} (j+1)(1-p)^{j-1} p$$

$$= (1-p) \sum_{j=1}^{\infty} (1-p)^{j-1} p + (1-p)\,\mathbb{E}[X]$$

$$\mathbb{E}[X] = \underbrace{1 \cdot p}_{\text{1st success}} + \underbrace{(1-p)}_{\text{1st fail}} \cdot \underbrace{\mathbb{E}[X+1]}_{\sum (k+1)\cdot(1-p)^k p}$$

$$= p + \underbrace{(1-p)\,\mathbb{E}[1]}_{1-p} + \underbrace{(1-p)}_{\text{pr. 1st throw fails}}\mathbb{E}[X]$$

$$= 1 + (1-p)\,\mathbb{E}[X].$$

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
- many more places

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 constant probability.[1]

[1] 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 constant probability.[1]

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*?

[1] 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 constant probability.[1]

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 $O(\ln \ln n)$!

[1] 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 constant probability.[1]

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 $h + 1$).

[1] we'll maybe see lower bound later

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

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

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

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

- 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 $\dfrac{n}{2^{2^{h-3}}}$ 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 $\dfrac{n}{2^{2^{h-3}}}$ bins of height h
- So expect $\log\log n$ maximum height after throwing n balls.

How do we turn this into a proof?
See [Mitzenmacher & Upfal, Chapter 14] and 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.