Today's Goal

Today is all bout three killer applications of the balls and bins problem.

- Computing the probability of collisions: How many balls do we need to throw before we two balls land in the same bin?
- The coupon collector problem: How many balls do we need to throw before every bin has at least one ball?
- Load balancing: If we throw m balls in n bins, what is the smallest k such that there's at least a 50% chance that no bin has more than k items?
    - Example: We mail 350,000,000 pieces of junk mail to 350,000,000 addresses randomly. With probability > 50%, nobody will get more than 12 pieces of junk mail.

We'll cover each topic in decreasing order of detail.

- Please read the notes again afterwards, especially on load balancing. It's an important skill to be able to read the notes.

Collisions: Motivation and Simulation

Lecture 18, CS70 Summer 2025

Balls and Bins: Collisions

- Motivation and Simulation
- Theoretical Analysis
  - Union Bound
  - Exact Solution (Counting)
  - Exact Solution (Product Rule)
  - Explicit Formula for Critical m

Balls and Bins: Coupon Collecting

- Simulation and Union Bound
- $m_{e^{-1}} \approx n \ln n + n$

Balls and Bins: Load Balancing (Extra)

Background Information: Hashing

A hash function maps items from a set (which may be infinite) to an integer from a finite range.

Examples:

- When you create a git commit and get back an identifier like "8a35c5ea4e042702144366cfd4225496fa384c01", this is a 160-bit hash of the information in your commit, using hashing algorithm SHA-1.
- In a hash table (see CS61B), the hash function tells you in which bucket to place the data.
- The SHA256 hash function maps a sequence of bits (e.g., the contents of a file) to a 256-bit integer. All entries on the Bitcoin blockchain are identified by a SHA256 hash (with certain properties!).

Note that cryptographic hashes (like SHA-1 and SHA-256) have additional requirements that "hash table hashes" don't have... security is complicated...

Hash Collisions

In many applications of hashing, we want collisions to be extremely unlikely.

Examples:

- Git commit hashes.
- File integrity hashes.

  - ISO
    - PGP signature
    - **PGP fingerprint:** 0x54449A5C
    - **SHA256:** b72dd6ffef7507f8b7cddd7c69966841650ba0f82c29a318cb2d182eb3fcb1db

Note: For hash tables (see CS61B), collisions are totally fine and are expected, though we want to avoid degenerate cases where tons of items end up in the same bucket.

Other Types of Collision Avoidance

There are other cases in computer science where we select random numbers and wish to avoid "collisions".

Examples:

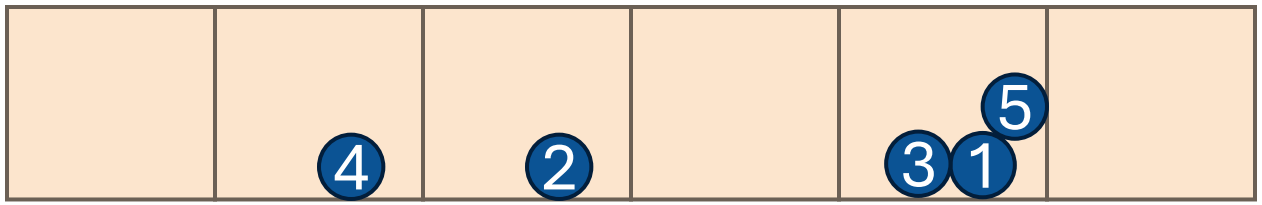
- Assigning unique identifiers to resources. For example, RFID tags use a 96-bit identifier. Chance of two tags from two different manufacturers having the same identifier is very small.
- "Nonces" in cryptographic protocols. Recall use of random numbers in passkey example from RSA lecture – avoid "replay attacks!"

Collisions and Balls and Bins

We can model the problem of collisions as the Balls and Bins problem.

Example, suppose we have a hash function with 6 possible outputs, and we hash 5 objects.

- One possible outcome is given below.
- We see that balls 1, 3, and 5 have “collided”.



Definition of A, Simulation

Define A as the event where there is at least one collision, i.e., at least two balls land in the same bin.

- As the number of bins m grows, the chance of a collision decreases.
- We want A to be false in our applications.

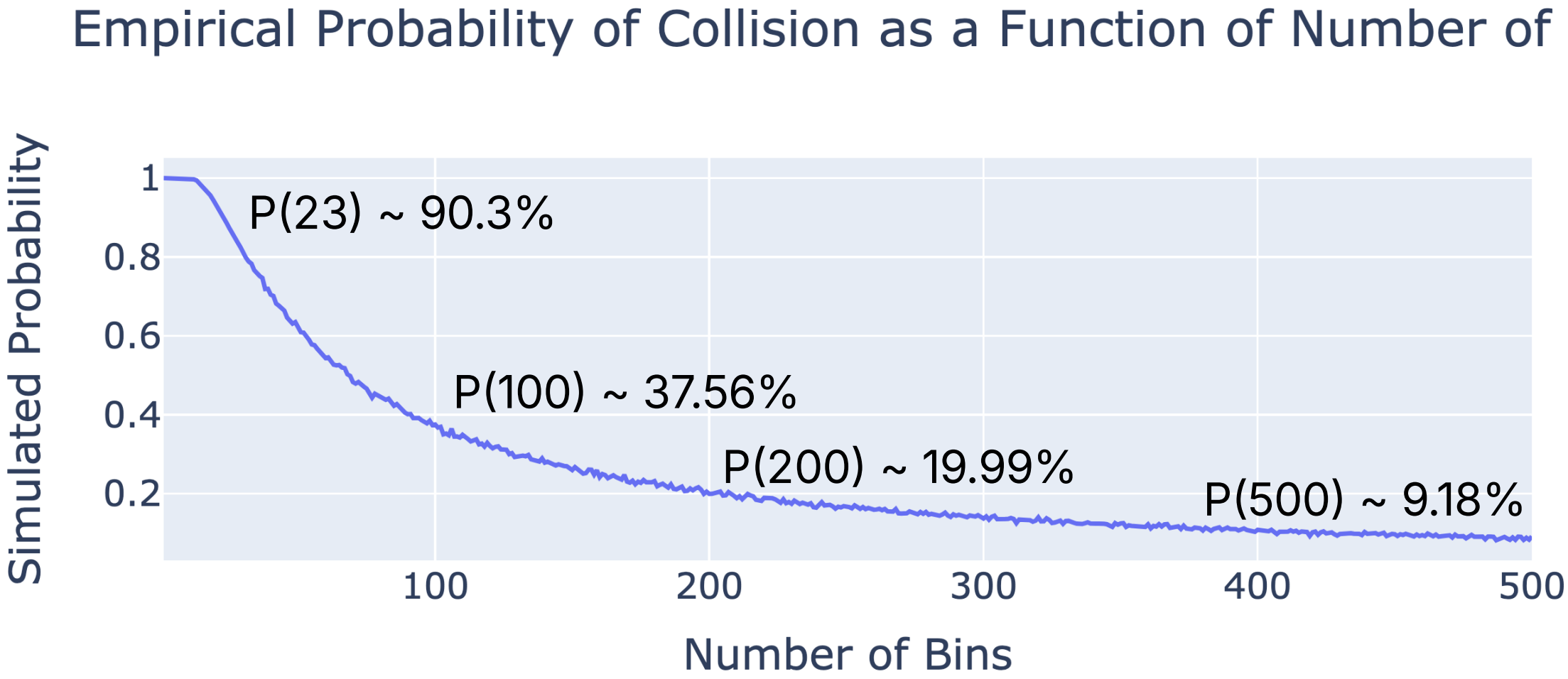
Let's do a quick simulation to understand the probability of a collision as a function of m.

- https://joshh.ug/cs70/collision_simulator.html

Note: The notes define A in reverse, i.e., in the notes, A is the probability that there is no collision.

Probability as a Function of M

Below, we see the empirical probability of a collision as a function of the number of bins (for $n = 10$ balls).



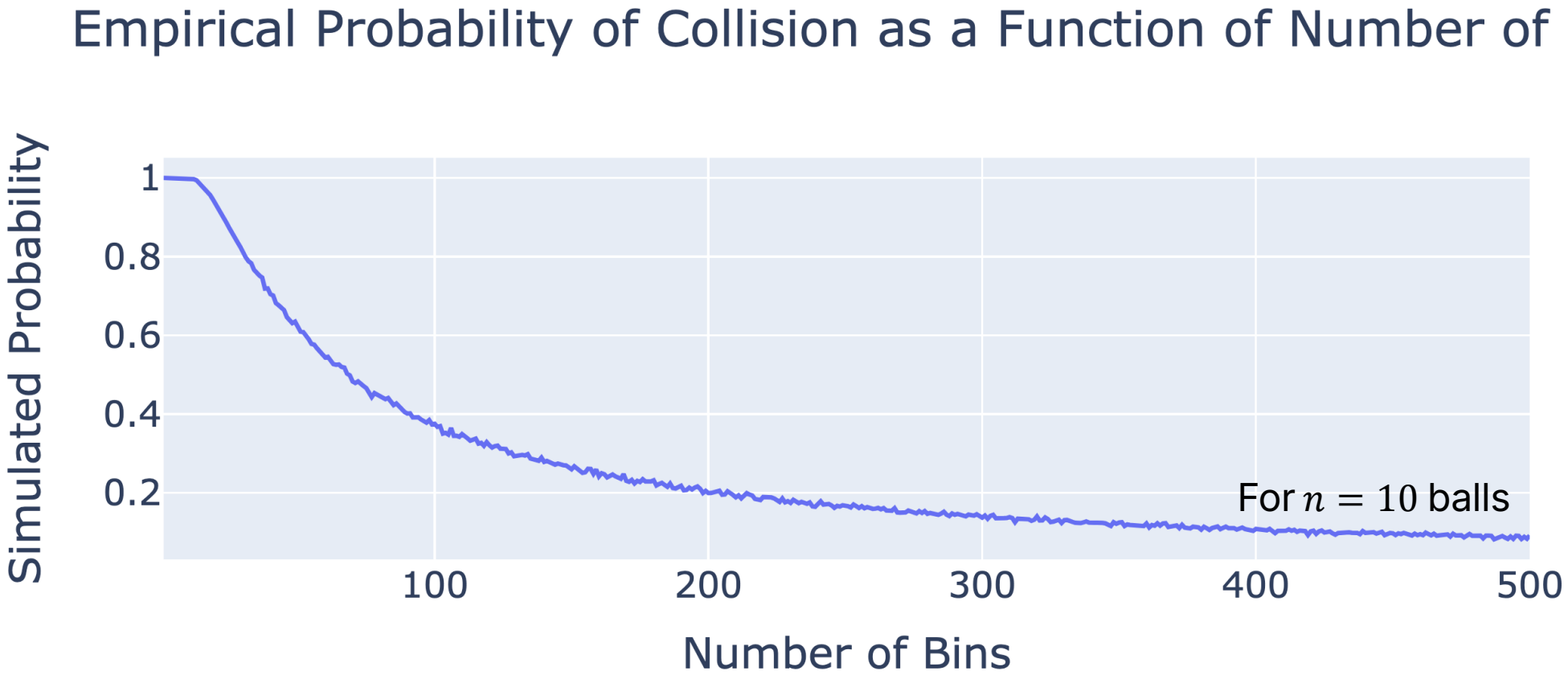
Theoretical Analysis for Collisions: Union Bound

Lecture 18, CS70 Summer 2025

Goal

In this part of the lecture, we'll work to approximate this function with a union bound.

- Example usage: Suppose we want to pick a number of bins so that if we throw 10,000 balls, the chance of a collision is less than 1%.



Balls and Bins: Ball Pairs and Collisions

Our first analysis will center around the idea of pairs of balls.

In the balls and bins problem with m balls, there are $\binom{m}{2} = \frac{m(m-1)}{2}$ pairs of balls, ignoring order.

For example, if m = 5, we have $\binom{5}{2} = \frac{5(5-1)}{2} = 10$ pairs of balls.

- 12, 13, 14, 15, 23, 24, 25, 34, 35, 45

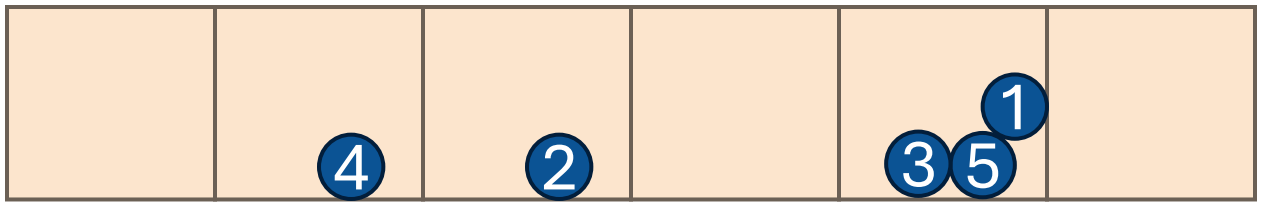
Let C_i be the event that the ith pair of balls collide.

Balls and Bins: Testing Your Understanding

For example, if m = 5, we have $\binom{5}{2} = \frac{5(5-1)}{2} = 10$ pairs of balls.

- 12, 13, 14, 15, 23, 24, 25, 34, 35, 45

Assuming the pairs of balls above are numbered 1 through 10, and we have the experimental outcome below, which C_i are true, i.e., which collisions occur?



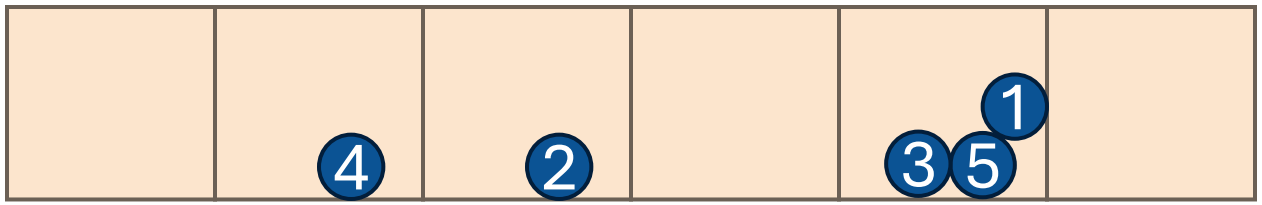
Balls and Bins: Testing Your Understanding

For example, if m = 5, we have $\binom{5}{2} = \frac{5(5-1)}{2} = 10$ pairs of balls.

- 12, 13, 14, 15, 23, 24, 25, 34, 35, 45

Assuming the pairs of balls above are numbered 1 through 10, and we have the experimental outcome below, which C_i are true, i.e. which collisions occur?

- C_2, C_4, C_9



Probability of a Collision

If A is the probability that a collision occurs, how should we write A in terms of events $C_1, C_2, \ldots$?

$$A = \overset{??}{\bigcup_{i=1}} C_i$$

$$A = \overset{??}{\bigcap_{i=1}} C_i$$

Probability of a Collision

Let A be the event that some collision occurs. That is:

$$A = \bigcup_{i=1}^{\binom{m}{2}} C_i$$

Why?

- There are $\binom{m}{2}$ possible collision events, and A is true if any of them are true.

We want to know $\mathrm{P}(A)$. Then we'll pick m such that this probability is below our desired threshold.

Collision Independence

Question: Are these events independent?

$$A = \bigcup_{i=1}^{\binom{m}{2}} C_i$$

Probability of a Collision

Question: Are these events independent?

- No. Example: if 1 and 2 collide, and 2 and 3 collide, then 1 and 3 also collide!

$$A = \bigcup_{i=1}^{\binom{m}{2}} C_i$$

Probability of a Collision

Question: Practically speaking, can we just use the principle of inclusion and exclusion to compute the probability $P(A)$ as a function of m?

$$A = \bigcup_{i=1}^{\binom{m}{2}} C_i$$

Probability of a Collision

Question: Practically speaking, can we just use the principle of inclusion and exclusion to compute the probability $P(A)$ as a function of m?

- No, using the PIE would require calculating roughly 2^{m^2} terms.

$$A = \bigcup_{i=1}^{\binom{m}{2}} C_i$$

Probability of a Collision

One approach is to simply truncate the PIE sum:

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) = P(C_1) + P(C_2) + P(C_3) + \cdots + P\left(C_{\binom{m}{2}}\right) - P(C_1 \cap C_2) - P(C_1 \cap C_2) - \cdots$$

As we discussed in lecture 17, as we increase n_s and allow combinations involving more terms, we get increasing accuracy.

$$P(|A_1 \cup \cdots \cup A_n|) = \sum_{k=1}^{n_s} (-1)^{k-1} \sum_{S \subseteq \{1,\ldots,n\}:|S|=k} P\left(\left|\bigcap_{i \in S} A_i\right|\right)$$

Probability of a Collision

In our case, let's just truncate to the probabilities of the individual events, i.e., choose $n_s = 1$.

- Recall, we called this the union bound.

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq P(C_1) + P(C_2) + P(C_3) + \cdots + P\left(C_{\binom{m}{2}}\right)$$

Question: What is $P(C_i)$?

- In other words, what's the chance that any particular pair of balls collides?

Probability of a Collision

In our case, let's just truncate to the probabilities of the individual events.

- Recall, we called this the union bound.

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq P(C_1) + P(C_2) + P(C_3) + \cdots + P\left(C_{\binom{m}{2}}\right)$$

Question: What is $P(C_i)$?

- In other words, what's the chance that any particular pair of balls collides?
- Let the position of the first ball be bin k. The chance that the second ball also lands in bin k is $1/n$.
- Thus, the answer is $1/n$.

Probability of a Collision

In our case, let's just truncate to the probabilities of the individual events.

- Recall, we called this the union bound.

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq P(C_1) + P(C_2) + P(C_3) + \cdots + P\left(C_{\binom{m}{2}}\right)$$

Question: What is $P(C_i)$?

- Thus, the answer is $1/n$.

So what is the union bound?

Probability of a Collision

In our case, let's just truncate to the probabilities of the individual events.

- Recall, we called this the union bound (a.k.a. Boole's inequality).

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq \frac{1}{n} \times \binom{m}{2}$$

Probability of a Collision

In our case, let's just truncate to the probabilities of the individual events.

- Recall, we called this the union bound (a.k.a. Boole's inequality).

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq \frac{1}{n} \times \binom{m}{2} = \frac{m(m-1)}{2n}$$

Probability of a Collision

In our case, let's just truncate to the probabilities of the individual events.

- Recall, we called this the union bound (a.k.a. Boole's inequality).

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq \frac{1}{n} \times \binom{m}{2} = \frac{m(m-1)}{2n} \leq \frac{m^2}{2n}$$

Thus, the probability that we get a collision is less than roughly $\frac{m^2}{2n}$.

Example for m = 10

In our case, let's just truncate to the probabilities of the individual events.

- Recall, we called this the union bound (a.k.a. Boole's inequality).

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) \leq \frac{1}{n} \times \binom{m}{2} = \frac{m(m-1)}{2n} \leq \frac{m^2}{2n}$$

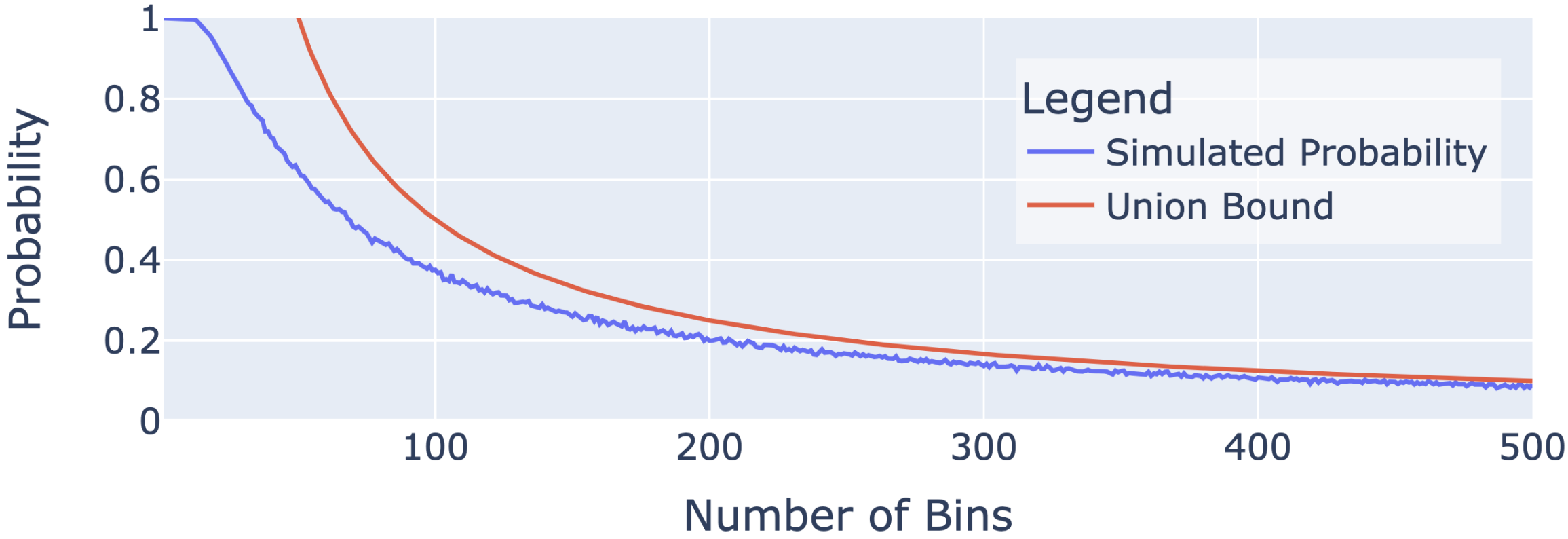
Thus, the probability that we get a collision is less than roughly $\frac{m^2}{2n}$.

So if there are 10 balls, the probability of a collision is less than $\frac{100}{2n} = \frac{50}{n}$

Union Bound as a Function of the Number of Bins

Below, I've plotted $\frac{50}{n}$ alongside the simulated probability.

- Note: The union bound is greater than 1 for small n. This is fine. It's still an upper bound on the probability.
  - Technically: The graph only shows it's an upper bound for our *experimental estimate* of the probability… but Boole's inequality shows it's a real upper bound.

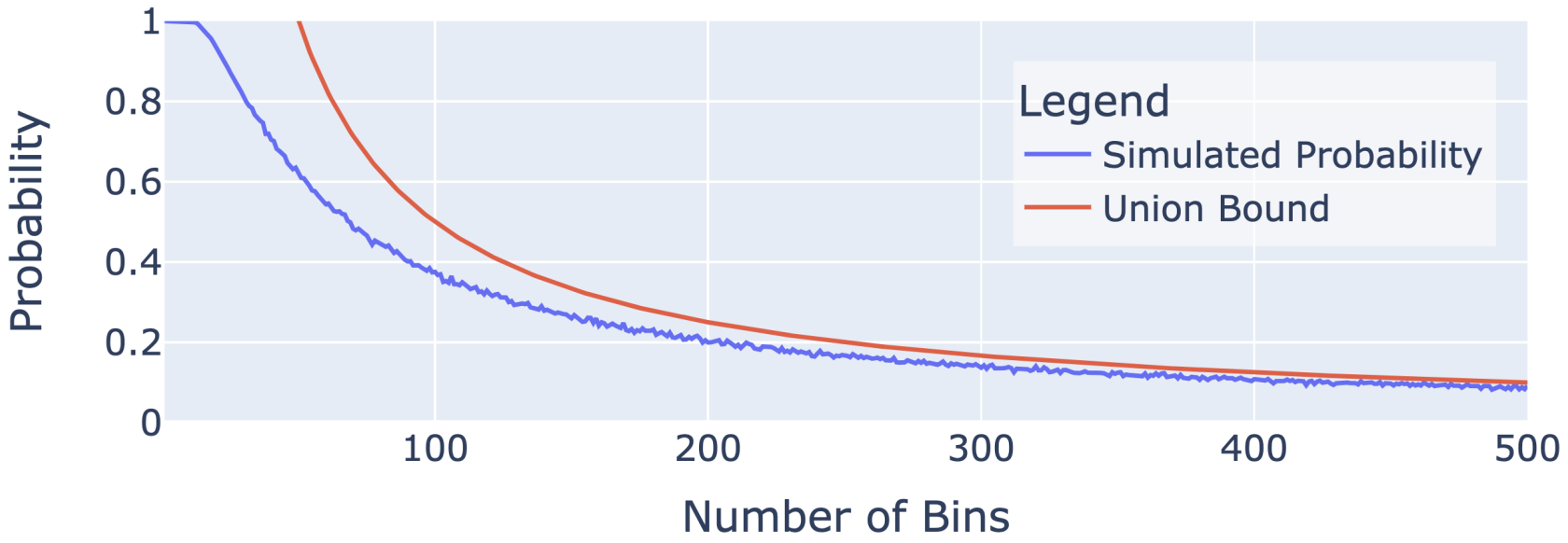


Test Your Understanding

Using the union bound: If we have 365 balls, roughly how many bins do we need to avoid a collision with probability 50%.

Recall the union bound says the probability is less than approximately $\frac{m^2}{2n}$.

Thoughts?



Test Your Understanding

Using the union bound: If we have 365 bins, roughly how many balls can you throw before you have a 50% chance of a collision.

Recall the union bound says the probability is less than approximately $\frac{m^2}{2n}$.

- We can compute the m that gets us there with:

$$\frac{m^2}{2n} = \frac{1}{2}$$

$$n = m^2$$

$$m = \sqrt{n}$$

$$m = \sqrt{365} \approx 19.1$$

Birthday Paradox Connection

This problem we just solved is just the birthday paradox.

- Instead of trying many different m's, we computed an upper bound.
- Union bound: The real m can be no smaller than approximately* $\sqrt{365} = 19.1$.
- In lecture 16, we smallest m that has a probability of collision above 50% is 23.

*: The reason it is approximate is because we replaced $m(m-1)$ by m^2. We get a slightly tighter bound by avoiding this replacement.

Theoretical Analysis for Collisions: Exact Solution (Counting)

Lecture 18, CS70 Summer 2025

Balls and Bins: Collisions

- Motivation and Simulation
- Theoretical Analysis
  - Union Bound
  - Exact Solution (Counting)
  - Exact Solution (Product Rule)
  - Explicit Formula for Critical m

Balls and Bins: Coupon Collecting

- Simulation and Union Bound
- $m_{e^{-1}} \approx n \ln n + n$

Balls and Bins: Load Balancing (Extra)

An Exact Solution

Earlier, we saw that the probability of a collision could be computed using an exponentially long sum with $\binom{m}{1}$ singleton terms, $\binom{m}{2}$ pairwise terms, etc.

$$\mathrm{P}\left(\bigcup_{i=1}^{\binom{m}{2}} C_i\right) = P(C_1) + P(C_2) + P(C_3) + \cdots + P\left(C_{\binom{m}{2}}\right) - P(C_1 \cap C_2) - P(C_1 \cap C_2) - \cdots$$

This challenging-to-compute expression was a consequence of our decision to frame our argument around pairs of balls.

- But we already know how to solve this problem exactly! We did it two lectures ago. Let's review the idea from those two slides in lecture 16.

Birthdays (50 case)

Suppose we have 50 people in a room who all have a birthday between day 1 and day 365. What is the chance that none of those 50 people share the same birthday?

- Total number of sequences of birthdays: 365^{50}
- Number of sequences of birthdays w/no repeats:

$$|\bar{A}| = 365 \times 364 \times \cdots \times 316$$

Probability of no repeated birthdays:

$$P(|\bar{A}|) = \frac{365 \times 364 \times \cdots \times 316}{365^{50}}$$

Birthdays (m case)

Suppose we have m people in a room who all have a birthday between day 1 and day 365. What is the chance that none of those m people share the same birthday?

- Total number of sequences of birthdays: 365^m
- Number of sequences of birthdays w/no repeats:

$$|\bar{A}| = 365 \times 364 \times \cdots \times (365 - m + 1)$$

Probability of no repeated birthdays:

$$P(|\bar{A}|) = \frac{365 \times 364 \times \cdots \times (316 - m + 1)}{365^m}$$

N bins, M Balls

Suppose we have m balls tossed in bins number 1 to n. What is the chance that none of those m balls land in the same bin?

- Total number of sequences of ball locations: n^m
- Number of sequences w/no repeats:

$$|\bar{A}| = n \times (n-1) \times \cdots \times (n-m+1)$$

Probability of no collisions:

$$P(|\bar{A}|) = \frac{n \times (n-1) \times \cdots \times (n-m+1)}{n^m}$$

N bins, M Balls

Suppose we have m balls tossed in bins number 1 to n. What is the chance that none of those m balls land in the same bin?

Probability of no collisions. Let's rewrite in a slightly different form:

$$P(|\bar{A}|) = \frac{n \times (n-1) \times \cdots \times (n-m+1)}{n^m}$$

$$= \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-m+1}{n}$$

$$= \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$$

Theoretical Analysis for Collisions: Exact Solution (Product Rule)

Lecture 18, CS70 Summer 2025

Balls and Bins: Collisions

- Motivation and Simulation
- Theoretical Analysis
  - Union Bound
  - Exact Solution (Counting)
  - Exact Solution (Product Rule)
  - Explicit Formula for Critical m

Balls and Bins: Coupon Collecting

- Simulation and Union Bound
- $m_{e^{-1}} \approx n \ln n + n$

Balls and Bins: Load Balancing (Extra)

Using the Product Rule

Another approach is to use the product rule.

- For $1 \leq i \leq m$, define A_i as the event where the ith ball lands in a different bin than the previous $i - 1$ balls.

Goal: Use the product rule to write $P(\bar{A})$ in terms of the events A_i.

Using the Product Rule

Another approach is to use the product rule.

- For $1 \leq i \leq m$, define A_i as the event where the ith ball lands in a different bin than the previous $i-1$ balls.

$$P(\bar{A}) = P\left(\bigcap_{i=1}^{n} A_i\right)$$

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_m \middle| \cap_{i=1}^{m-1} A_i\right)$$

Questions, what are:

- $P(A_1) = ?$
- $P(A_2|A_1) = ??$
- $P(A_3|A_1 \cap A_2) = ??$
- $P\left(A_m \middle| \cap_{i=1}^{m-1} A_i\right) = ??$

No special link to answer this one.

Note: $P(\bar{A}) = 1 - \mathrm{P}\left(\mathrm{U}_{i=1}^{\binom{m}{2}} C_i\right)$

Using the Product Rule

Another approach is to use the product rule.

- For $1 \leq i \leq m$, define A_i as the event where the ith ball lands in a different bin than the previous $i-1$ balls.

$$P(\bar{A}) = P\left(\bigcap_{i=1}^{n} A_i\right)$$

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_m \middle| \cap_{i=1}^{m-1} A_i\right)$$

Questions, what are:

- $P(A_1) = 1$
- $P(A_2|A_1) = \frac{n-1}{n}$
- $P(A_3|A_1 \cap A_2) = ??$
- $P\left(A_m \middle| \cap_{i=1}^{m-1} A_i\right) = ??$

Using the Product Rule

Another approach is to use the product rule.

- For $1 \leq i \leq m$, define A_i as the event where the ith ball lands in a different bin than the previous $i-1$ balls.

$$P(\bar{A}) = P\left(\bigcap_{i=1}^{n} A_i\right)$$

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_m \middle| \cap_{i=1}^{m-1} A_i\right)$$

Questions, what are:

- $P(A_1) = 1$
- $P(A_2|A_1) = \frac{n-1}{n}$
- $P(A_3|A_1 \cap A_2) = \frac{n-2}{n}$
- $P\left(A_m \middle| \cap_{i=1}^{m-1} A_i\right) = \frac{n-m+1}{n}$

Using the Product Rule

Another approach is to use the product rule.

- For $1 \leq i \leq m$, define A_i as the event where the ith ball lands in a different bin than the previous $i-1$ balls.

$$P(\bar{A}) = 1 \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-m+1}{n}$$

$$= \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$$

This is, of course, the same answer we derived using the direct counting method:

$$\frac{n \times (n-1) \times \cdots \times (n-m+1)}{n^m}$$

$$= \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$$

Summary of Approaches So Far

Summary of approaches so far:

- Define A as the union of collision events (exponentially complex, can union bound).
- Solve for $P(\bar{A})$ directly through counting.
- Define $\bar{A}$ as the intersection of non-collision events (straightforward and simple product rule).

In real problem solving, using unions, intersections, or direct counting may be appropriate.

Theoretical Analysis for Collisions: Explicit Formula for Critical m

Lecture 18, CS70 Summer 2025

Balls and Bins: Collisions

- Motivation and Simulation
- Theoretical Analysis
  - Union Bound
  - Exact Solution (Counting)
  - Exact Solution (Product Rule)
  - Explicit Formula for Critical m

Balls and Bins: Coupon Collecting

- Simulation and Union Bound
- $m_{e^{-1}} \approx n \ln n + n$

Balls and Bins: Load Balancing (Extra)

Computing Desired m

Suppose we want to find the n and m values that satisfy a specific probability.

- Example: There are n bins. How many balls can we throw before there is a 5% chance of getting a collision?

$$\left(1-\frac{1}{n}\right)\times\left(1-\frac{2}{n}\right)\times\cdots\times\left(1-\frac{m-1}{n}\right)$$

Computing Desired m

Suppose we want to find the n and m values that satisfy a specific probability.

- Example: There are $n = 100$ bins. How many balls can we throw before there is a 5% chance of getting a collision?

$$\left(1 - \frac{1}{100}\right) \times \left(1 - \frac{2}{100}\right) \times \cdots \times \left(1 - \frac{5}{100}\right)$$

- Plug in $m = 1$, $m = 2$, ..., until we get to a value that is below 0.95.
  - Formula above for $m = ?$

Computing Desired m

Suppose we want to find the n and m values that satisfy a specific probability.

- Example: There are $n = 100$ bins. How many balls can we throw before there is a 5% chance of getting a collision?

$$\left(1 - \frac{1}{100}\right) \times \left(1 - \frac{2}{100}\right) \times \cdots \times \left(1 - \frac{5}{100}\right)$$

- Plug in $m = 1$, $m = 2$, …, until we get to a value that is below 0.95.
  - Formula above for $m = 6$ - remember that final term is $1 - \frac{m-1}{n}$
  - Largest m that gives "confidence" above 0.95 is $m = 5$.

This brute force approach isn't ideal. We'd rather have an expression that gives the "critical" m directly.

Computing Desired m

First, we'll take the log of the function to turn it into a sum (natural log for later...):

$$P(\bar{A}) = \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$$

$$\ln\left(P(\bar{A})\right) = \ln\left(1 - \frac{1}{n}\right) + \ln\left(1 - \frac{2}{n}\right) + \cdots + \ln\left(1 - \frac{m-1}{n}\right)$$

Computing Desired m

First, we'll take the log of the function to turn it into a sum (natural log for later...):

$$P(\bar{A}) = \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$$

$$\ln\left(P(\bar{A})\right) = \ln\left(1 - \frac{1}{n}\right) + \ln\left(1 - \frac{2}{n}\right) + \cdots + \ln\left(1 - \frac{m-1}{n}\right)$$

Then, let's make use of a handy fact (derived from the Taylor series expansion):

$$\ln(1-x) \approx 0 - x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots$$

So for small x, $\ln(1-x) \approx -x$ and $\ln\left(P(\bar{A})\right) \approx -\frac{1}{n} - \frac{2}{n} - \frac{3}{n} - \cdots - \frac{m-1}{n}$

Computing Desired m

$$\ln\left(P(\bar{A})\right) \approx -\frac{1}{n} - \frac{2}{n} - \frac{3}{n} - \cdots - \frac{m-1}{n}$$

$$= -\frac{1}{n}\sum_{i=1}^{m-1} i$$

$$= -\frac{1}{n} \times \frac{m(m-1)}{2}$$

$$= -\frac{m(m-1)}{2n}$$

$$\approx -\frac{m^2}{2n}$$

- Error is small as long as $\frac{1}{\mathrm{n}}, \ldots, \frac{m-1}{n}$ are small enough.

Computing Desired m for a Specific Critical Probability

$$\ln\left(P(\bar{A})\right) \approx -\frac{m^2}{2n}$$

Exponentiating both sides we have:

$$P(\bar{A}) \approx e^{-m^2/2n}$$

Suppose we want to find the critical m so that the probability of a collision is 50%.

$$e^{-m^2/2n} = 0.5$$

$$-\frac{m^2}{2n} = \ln(0.5)$$

$$m^2 = -2n\ln(0.5)$$

$$m = \sqrt{-2\ln(0.5)}\sqrt{n} \approx 1.177\sqrt{n}$$

Computing Desired m for a Specific Critical Probability

$$\ln\left(P(\bar{A})\right) \approx -\frac{m^2}{2n}$$

Exponentiating both sides we have:

$$P(\bar{A}) \approx e^{-m^2/2n}$$

Suppose we want to find the critical m so that the probability of a collision is 50%.

$$m \approx 1.177\sqrt{n}$$

- So any m greater than this should have a collision probability of less than 50%.
- Note: Assumes that $\frac{1}{\mathrm{n}}, \ldots, \frac{m-1}{n}$ is small.

Evaluating the Critical M Formula

Now we have two ways to compute the critical m:

- Exact: Try out $m = 1$, $m = 2$, etc. for $\left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \cdots \times \left(1 - \frac{m-1}{n}\right)$
- Approximate: Compute $1.177\sqrt{n}$ directly.

Let's compare for a specific n. Let m_o be the largest m such that the probability of collision is less than 50%.

Example, for $n = 365$, we had the table to the right:

- Exact: $m_o = 22$
- Approximate: $1.177 \times \sqrt{365} = 22.48$

m	$P(A)$	%
1	0	0%
2	0.0027	0.27%
3	0.008	0.8%
4	0.016	1.6%
10	0.117	11.7%
20	0.411	41.1%
22	0.493	49.3%
23	**0.507**	**50.7%**

Exact vs. Approximate (Other n)

We can compare the exact and approximate solution for other choices of n.

n	10	20	50	100	200	365	500	1000	10^4	10^5	10^6
$1.177\sqrt{n}$	3.7	5.3	8.3	11.8	16.6	22.5	26.3	37.3	118	372	1177
Exact m_o	4	5	8	12	16	22	26	37	118	372	1177

Our approximation is very good even for small n. When n is large, the error observed in the error table above is negligible.

Other Critical Values

$$P(\bar{A}) \approx e^{-m^2/2n}$$

Naturally, we can use this to find critical m for other desired levels of confidence:

$$e^{-m^2/2n} = 0.95$$

$$-\frac{m^2}{2n} = \ln(0.95)$$

$$m^2 = -2n\ln(0.95)$$

$$m^2 = -2n\ln(0.95)$$

$$m = \sqrt{-2\ln(0.95)}\sqrt{n} = 0.32\sqrt{n}$$

Confidence and Critical Values

$$P(\bar{A}) \approx e^{-m^2/2n}$$

No matter what confidence level λ we specify, our critical value is always of the form $c\sqrt{n}$, where c is some constant.

$$e^{-m^2/2n} = \lambda$$

$$m = \sqrt{-2\ln(\lambda)}\sqrt{n}$$

Not discussed, but might be interesting for you to explore: How does the gap between this approximation and the exact m_o change as we vary λ and n?

- Example question: What does the error table look like for $\lambda = 0.999999$? How big did n have to get before the approximation was good?

Hash functions with large ranges (optional)

Recall: Git commits are identified by a 160-bit hash value.

- Collisions would cause real problems (confusion between commits)
- Define event A_m as "there is a collision in m commits"

Question: What is $P(A_m)$ when m is a billion?

Using previous approximation: $P(A_m) = 1 - e^{-m^2/2n}$

What is n?

Hash functions with large ranges (optional)

Recall: Git commits are identified by a 160-bit hash value.

- Collisions would cause real problems (confusion between commits)
- Define event A_m as "there is a collision in m commits"

Question: What is $P(A_m)$ when m is a billion?

Using previous approximation: $P(A_m) = 1 - e^{-m^2/2n}$

What is n? $n = 2^{160}$

Can we express m as a power of 2?

Hash functions with large ranges (optional)

Recall: Git commits are identified by a 160-bit hash value.

- Collisions would cause real problems (confusion between commits)
- Define event A_m as "there is a collision in m commits"

Question: What is $P(A_m)$ when m is a billion?

Using previous approximation: $P(A_m) = 1 - e^{-m^2/2n}$

What is n? $n = 2^{160}$

Can we express m as a power of 2? $m \approx 2^{30}$

So $\frac{m^2}{2n} \approx \frac{(2^{30})^2}{2 \cdot 2^{160}} = \frac{2^{60}}{2^{161}} = 2^{-101}$ and $P(A_m) \approx 1 - e^{-2^{-101}}$

More approximation magic: $e^{-x} \approx 1 - x$ ... so $P(A_m) \approx \frac{1}{2^{101}}$

Note: Probability of being hit by lightning in a year: about $\frac{1}{2^{20}}$ - much higher!

Coupon Collecting: Simulation and Union Bound

Lecture 18, CS70 Summer 2025

Balls and Bins: Collisions

- Motivation and Simulation
- Theoretical Analysis
  - Union Bound
  - Exact Solution (Counting)
  - Exact Solution (Product Rule)
  - Explicit Formula for Critical m

Balls and Bins: Coupon Collecting

- Simulation and Union Bound
- $m_{e^{-1}} \approx n \ln n + n$

Balls and Bins: Load Balancing (Extra)

Coupon Collecting

In the coupon collecting problem, we imagine a contest where every time you buy a box of cereal, there is a coupon in the box.

- n different coupons.
- Once you collect all n coupons, you can redeem them all for a discount on your next cereal.

Let m be the number of boxes of cereal you buy. How many boxes of cereal must we buy, i.e., how big must m get before we have a 50% chance of getting all of the coupons?

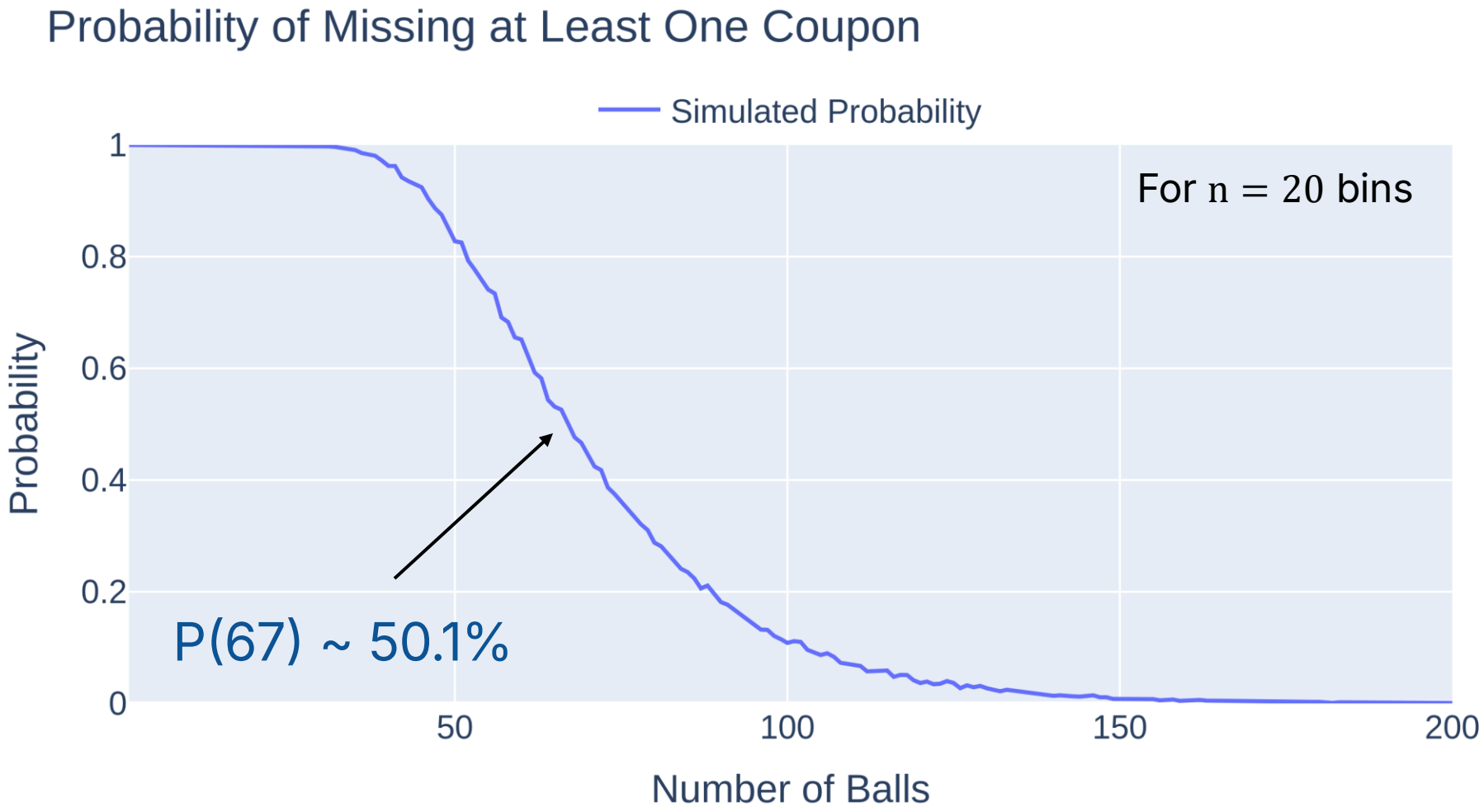
Let's try another simulation:

- https://joshh.ug/cs70/coupon_collector_simulator.html

Simulation Visualization

We can visualize the results of our simulation.

- We want to be able to compute the m where P crosses 50%.



Coupon Collecting

Coupon collecting problem in terms of balls and bins: Given n bins, how many balls m do we need to throw before there is at least one ball in every bin?

- Let A be the event where any bin is empty.
- Let A_i be the event where the ith bin is empty.

Naturally, we have that $P(A) = P(\cup_{i=1}^{n} A_i)$

Coupon Collecting

What is the probability that bin i is empty*?

- We've done this exact problem before: $\left(1-\frac{1}{n}\right)^m$
- Approach in lecture 16 was counting # outcomes with and without bin i.
- Can also compute as follows: Chance of first ball missing bin i is $\frac{n-1}{n}$ which can also be written as $\left(1-\frac{1}{n}\right)$. The chance of missing m independent throws is $\left(1-\frac{1}{n}\right)^m$

*If it helps make the problem more concrete, you can pick an arbitrary i, e.g., "what is the probability that bin 1 is empty?"

Coupon Collecting

Coupon collecting problem: Given n bins, how many balls m do we need to throw before there is at least one ball in every bin?

- Let A be the event where any bin is empty.
- Let A_i be the event where the ith bin is empty.

Naturally, we have that $P(A) = P(\cup_{i=1}^{n} A_i)$

- $P(A_i) = \left(1 - \frac{1}{n}\right)^m$
- Can apply the union bound (ignoring all pairwise, three-way, etc. interactions) and get:

$$P(A) \leq n\left(1 - \frac{1}{n}\right)^m$$

Example for n=20

Suppose there are n different coupons. We have that the probability of missing at least one coupon is given by:

$$P(A) \leq n\left(1-\frac{1}{n}\right)^m$$

For $n = 20$, we have:

$$P(A) \leq 20\left(\frac{19}{20}\right)^m$$

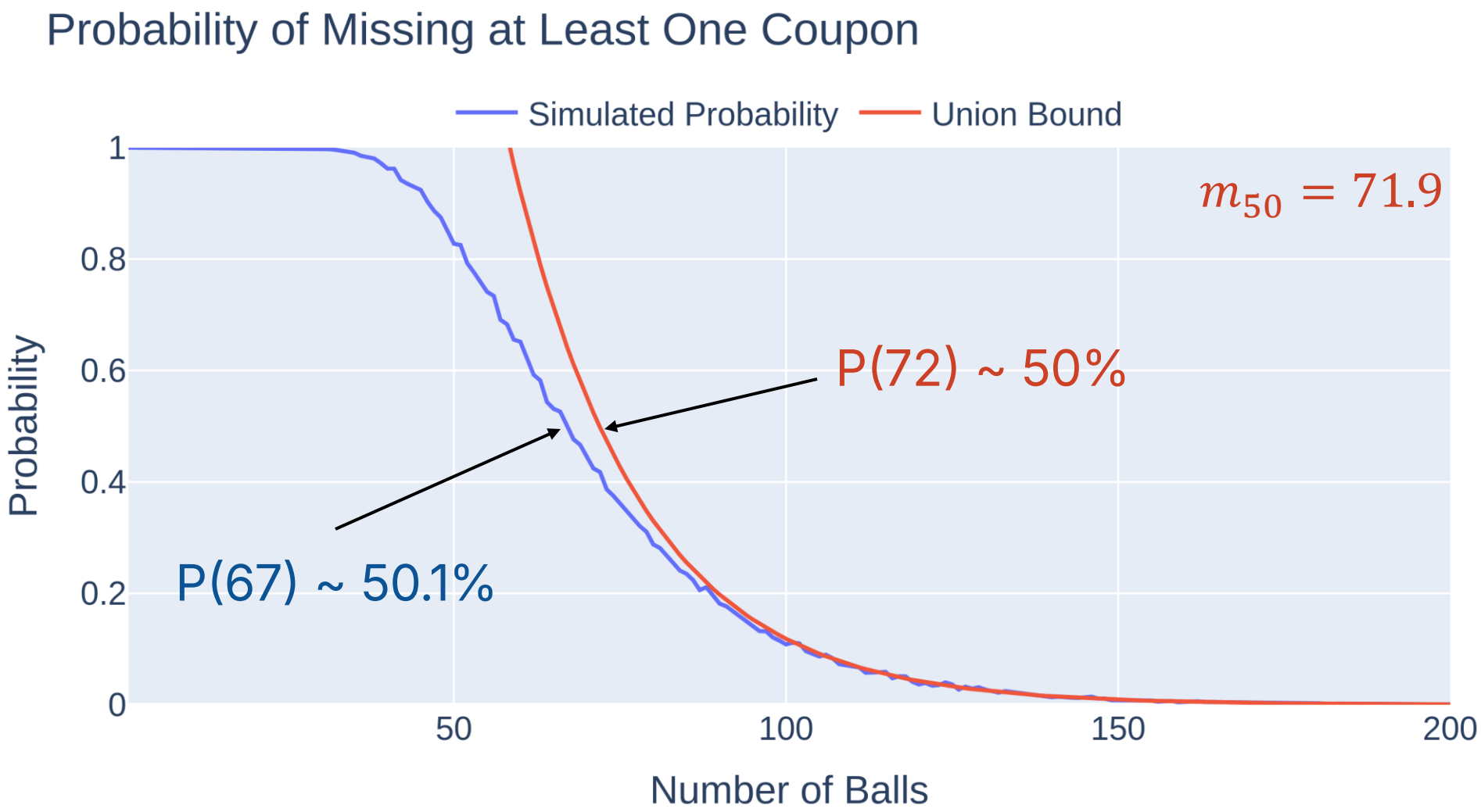
To get a 50 chance of having all the coupons, we can compute the critical m_{50}.

$$\frac{0.5}{20} = \left(\frac{19}{20}\right)^{m_{50}}$$

$$\ln\left(\frac{0.5}{20}\right) / \ln\left(\frac{19}{20}\right) = m_{50} \qquad 71.9 = m_{50}$$

Union Bound Visualization

We can visualize the union bound for the $n = 20$ case as shown below:



General Case

Suppose there are n different coupons. We have that the probability of missing at least one coupon is given by:

$$P(A) \leq n\left(1 - \frac{1}{n}\right)^m$$

For general n and desired probability of 50%, we have:

$$\frac{0.5}{n} = \left(\frac{n-1}{n}\right)^{m_{50}}$$

$$\ln\left(\frac{0.5}{n}\right) / \ln\left(\frac{n-1}{n}\right) = m_{50}$$

$$\frac{\ln(0.5) - \ln(n)}{\ln\left(1 - \frac{1}{n}\right)} = m_{50}$$

Coupon Collecting Summary

Using the union bound, we have that the probability of missing at least one coupon is:

$$P(A) \leq n\left(1 - \frac{1}{n}\right)^m$$

Using basic algebra, the critical m_{50} that yields a union bound of 50% is:

$$\frac{\ln(0.5) - \ln(n)}{\ln\left(1 - \frac{1}{n}\right)} = m_{50}$$

Coupon Collecting: $m_{e^{-1}} \approx n \ln n + n$

Lecture 18, CS70 Summer 2025

Balls and Bins: Collisions

- Motivation and Simulation
- Theoretical Analysis
  - Union Bound
  - Exact Solution (Counting)
  - Exact Solution (Product Rule)
  - Explicit Formula for Critical m

Balls and Bins: Coupon Collecting

- Simulation and Union Bound
- $m_{e^{-1}} \approx n \ln n + n$

Balls and Bins: Load Balancing (Extra)

Coupon Collecting Summary and N Ln N Formula

Using the union bound, we have that the probability of missing at least one coupon is:

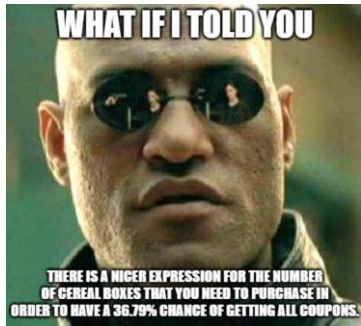
$$P(A) \leq n\left(1 - \frac{1}{n}\right)^m$$

Using basic algebra, the critical m_{50} that yields a union bound of 50% is:

$$\frac{\ln(0.5) - \ln(n)}{\ln\left(1 - \frac{1}{n}\right)} = m_{50}$$

Using another approximation, we can find a nicer formula for a critical m.

- Goal: Show that $m_{36.79} \approx n \ln n + n$, i.e. the m which yields a $e^{-1} \approx 36.79\%$ chance of missing at least one coupon is $n \ln n + n$
- Example: For $n = 20$, $m_{36.79} \approx 20 \ln 20 + 20 = 79.9$



N Ln N Formula Proof

Our union bound is:

$$P(A) \leq n\left(1-\frac{1}{n}\right)^m$$

If we approximate $\left(1-\frac{1}{n}\right)^n \approx e^{-1}$ (not hard to show), then $\left(1-\frac{1}{n}\right)^m \approx e^{-m/n}$, and:

$$P(A) \leq n\left(1-\frac{1}{n}\right)^m \approx ne^{-m/n}$$

Suppose we select $m = n\ln n + n$, then this expression becomes:

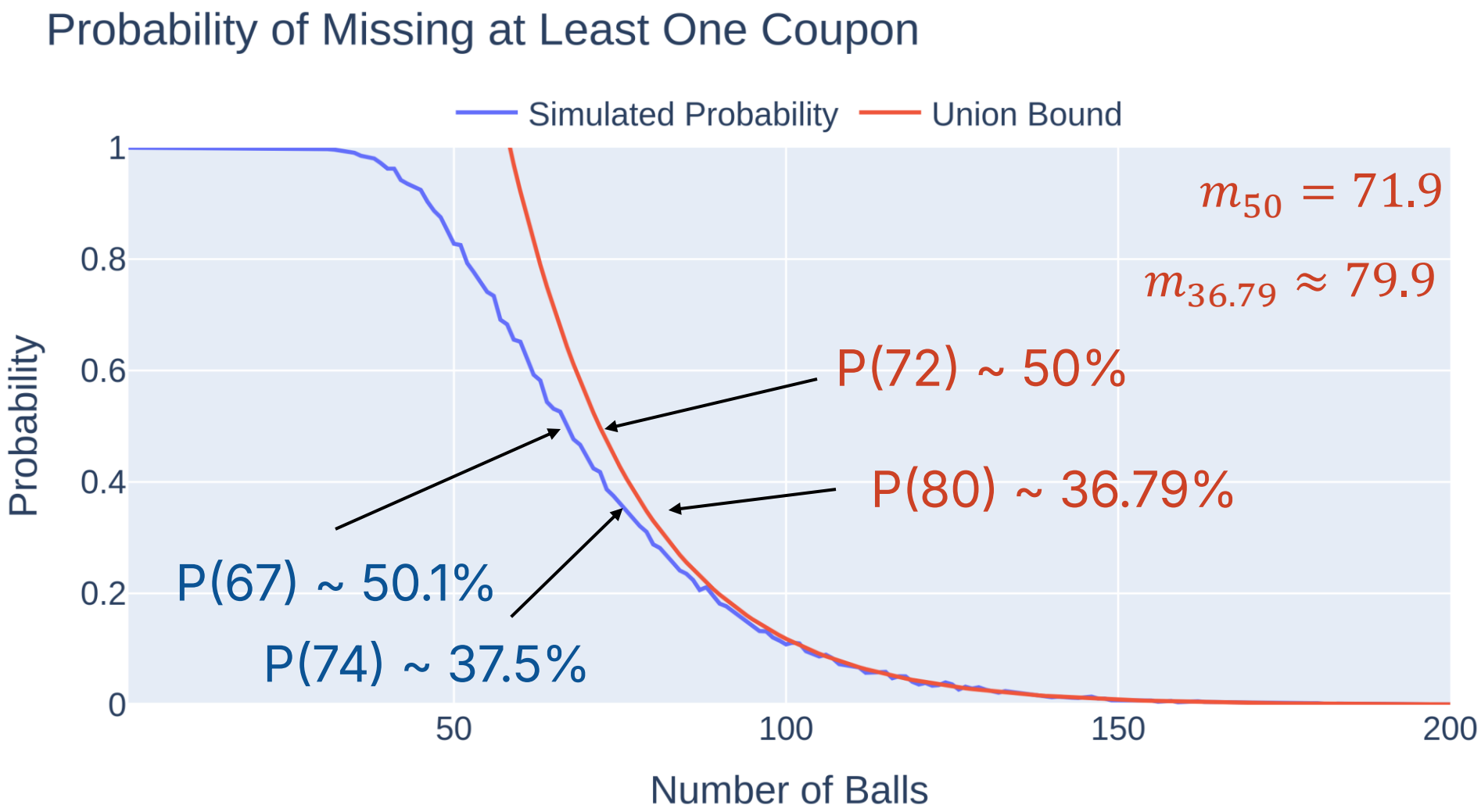
$$ne^{-m/n}\Big|_{m=n\ln n+n} = ne^{-\ln n-1}$$

$$= ne^{-\ln n} \times e^{-1}$$

$$= e^{-1}$$

Union Bound Visualization

We can visualize the union bound for the $n = 20$ case as shown below:



Coupon Collecting Summary

Using the union bound, we have that the probability of missing at least one coupon is:

$$P(A) \leq n\left(1 - \frac{1}{n}\right)^m$$

Using basic algebra, the critical m_{50} that yields a union bound of 50% is:

$$\frac{\ln(0.5) - \ln(n)}{\ln\left(1 - \frac{1}{n}\right)} = m_{50}$$

Using another approximation, the critical $m_{36.79}$ that yields a union bound of e^{-1} can be approximated by:

$$n \ln n + n \approx m_{36.79}$$

A better name for this: $m_{e^{-1}}$

Summary

Today we saw two different applications of balls and bins:

- Collisions
- Coupon Collecting

In the notes, they also consider an additional problem: Load Balancing. See extra slides.

In all three cases, we leveraged the union bound to produce concise formulas to compute quantities of interest.

Load Balancing (Extra)

Lecture 18, CS70 Summer 2025

Load Balancing

Suppose we send computational loads to a variety of different servers.
Suppose we send them randomly, with no regard for scheduling.

We can model this as throwing m balls into n bins.

In the context of load scheduling, we might wonder how heavily loaded the busiest server will be.

- More precisely: What is the smallest k such that the server with the highest load has to handle k loads with probability 50%.

Let's again start with a simulation.

- joshh.ug/cs70/load_balancing.html

Results

After running 50 simulations with $n = m = 201$, we have:

- k appears to be (based on simulation) 5.

Maximum Load Observed Per Simulation:

```
 1:   50/50  (100.0%)
 2:   50/50  (100.0%)
 3:   50/50  (100.0%)
 4:   49/50  ( 98.0%)
 5:   31/50  ( 62.0%)
 6:    6/50  ( 12.0%)
 7:    1/50  (  2.0%)
 8:    0/50  (  0.0%)
 9:    0/50  (  0.0%)
10:    0/50  (  0.0%)
```

Analysis

Define A_k as the event that the load in any bin has at least k balls.

- We want to know the smallest $P(A_k)$ such that $P(A_k) \leq 1/2$
- It is extremely difficulty to approach this problem directly (try!)

Analysis

Define A_k as the event that the load in any bin has at least k balls.

- We want to know the smallest $P(A_k)$ such that $P(A_k) \leq 1/2$
- It is extremely difficulty to approach this problem directly (try!)

The big idea in the notes is how we define the problem. We do something very clever:

- Define $A_k(i)$ as the event that the load in bin i is at least k.
- Define k_c as the smallest k such that $P\big(A_k(1)\big) \leq \frac{1}{2n}$
- Then it turns out that k_c is also the smallest k such that $P(A_k) \leq 1/2$.

In other words, we transform a problem across all bins into a problem of just one bin.

Analysis

The big idea in the notes is how we define the problem. We do something very clever:

- Define $A_k(i)$ as the event that the load in bin i is at least k.
- Define k_c as the smallest k such that $P\big(A_k(1)\big) \leq \frac{1}{2n}$
- Then it turns out that k_c is also the smallest k such that $P(A_k) \leq 1/2$.

Why does this work?

- Since $P(A_k) = P(\mathrm{U}_{i=1}^{n} A_k(i))$, we have that:

$$P(A_k) \leq \sum_{i=1}^{n} P\big(A_k(i)\big) \leq n \times \frac{1}{n} = \frac{1}{2}$$

Analysis

The big idea in the notes is how we define the problem. We do something very clever:

- Define $A_k(i)$ as the event that the load in bin i is at least k.
- Define k_c as the smallest k such that $P\big(A_k(1)\big) \leq \frac{1}{2n}$
- Then it turns out that k_c is also the smallest k such that $P(A_k) \leq 1/2$.

$$P(A_k) \leq \sum_{i=1}^{n} P\big(A_k(i)\big) \leq n \times \frac{1}{n} = \frac{1}{2}$$

Last step: Find k_c by reasoning about a single bin (see notes). Result turns out to be:

$$k \approx \frac{\ln n}{\ln \ln n}$$