Lecture 03: Balls & Bins (Birthday Paradox, Max-Load)
There are \( m \) balls and \( n \) bins

Each ball is independently thrown into a bin that is chosen uniformly at random

\((T_1, \ldots, T_m)\) be the joint distribution such that \( T_i \) represents the bin into which the \( i \)-th ball is thrown
For $1 \leq i < j \leq m$, let $X_{i,j}$ be the indicator variable for the event that the $i$-th ball and the $j$-th ball fall into the same variable, i.e., indicator variable for the event $T_i = T_j$

We are interested in computing the “expectation of the random variable $X_{i,j}$”

$$E[X_{i,j}] = P[X_{i,j} = 1] = P[T_i = T_j]$$

$$= \sum_{t=1}^{n} P[T_i = T_j = t] = n \cdot 1/n^2 = 1/n$$
Expected Number of Collisions

- Let \( X = \sum_{1 \leq i < j \leq m} X_{i,j} \)
- The random variable \( X \) counts the number of collisions that occur.
- We are interested in the expected number of collisions

\[
E[X] = E \left[ \sum_{1 \leq i < j \leq m} X_{i,j} \right] \\
= \sum_{1 \leq i < j \leq m} E[X_{i,j}] \quad \text{By the Linearity of Expectation} \\
= \binom{m}{2} \frac{1}{n}
\]

- Note that for \( m = \sqrt{2n} \), we have \( E[X] = 1 \)
- The “average number of collisions” at \( m = \sqrt{2n} \) is 1, but how does the probability of this event behave like?
There are $m$ people in a room. Assume that the birthday of people are distributed uniformly at random over the 365 days in the year. What is the number of people $m$ that ensures that two people share a birthday with probability 0.9?
We want to find the value of $m$ such that throwing $m$ balls in $n = 365$ bins ensures a collision with probability 0.9.

Let $\text{NoColl}_{\leq t}$ represent the probability that $T_1, \ldots, T_t$ are all distinct.

Note that

$$\mathbb{P}[\text{NoColl}_{\leq t}] = 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{t-1}{n}\right)$$

$$= \prod_{t=0}^{t-1} \left(1 - \frac{t}{n}\right)$$
We are interested in finding $m$ such that the probability of collision is high

Alternately, we are interested in showing the probability of $\text{NoColl} \leq m$ is small

$$\mathbb{P}[\text{NoColl} \leq m] = \prod_{t=1}^{m-1} \left(1 - \frac{t}{n}\right)$$

$$\leq \prod_{t=1}^{m-1} \exp\left(-\frac{t}{n}\right) = \exp\left(-\sum_{t=0}^{m-1} \frac{t}{n}\right)$$

$$= \exp\left(-(m - 1)m/2n\right)$$

Substituting $m = c\sqrt{n}$, for a suitable constant $c > 0$, ensures that $\exp\left(-(m - 1)m/2n\right) \leq 0.1$
We are interested in finding out \( m \) such that we can throw \( m \) balls without getting a collision, with high probability.

That is, we are interested in showing that the probability of \( \text{NoColl} \leq m \) is high

\[
P[\text{NoColl} \leq m] = \prod_{t=1}^{m-1} \left(1 - \frac{t}{n}\right) \geq \prod_{t=0}^{m-1} \exp \left(-\frac{t}{n} - \frac{t^2}{n^2}\right)
\]

\[
= \exp \left(- \sum_{t=0}^{m-1} \frac{t}{n} - \sum_{t=0}^{m-1} \frac{t^2}{n^2}\right)
\]

\[
= \exp \left(- \frac{m(m-1)}{2n} - \frac{m(m-0.5)(m-1)}{3n^2}\right)
\]

For \( m = d\sqrt{n} \), the first term in the exponent dominates and the second term is \( o(1) \).
For a constant $d > 0$ we can ensure that the final probability term is $\geq 0.9$.

Conclusion: As $m$ increases from $d\sqrt{n}$ to $c\sqrt{n}$ the probability of no-collisions transitions from 0.9 to 0.1.

Recommended: Plot the probability of no-collisions for $m = 1$ to $m = n$, for large values of $n$. How quickly does the probability transition from “high” to “low” as $n$ increases?
Let \( X_i \) represent the indicator variable for the \( i \)-th bin being empty, i.e., the indicator of the event: \( T_j \neq i \), for all \( i \in \{1, \ldots, m\} \)

Note that

\[
E[X_i] = P[X_i = 1] = \left(1 - \frac{1}{n}\right)^m
\]

Let \( X = \sum_{t=0}^{n} X_i \) represent the number of empty bins

\[
E[X] = E\left[\sum_{t=1}^{n} X_i\right] = \sum_{t=1}^{n} E[X_i] = n \left(1 - \frac{1}{n}\right)^m \approx n \exp\left(-\frac{m}{n}\right)
\]

For \( m = n \), we expect (roughly) \( n/e \) empty bins! For \( m = n \log n \), we expect (roughly) 1 empty bin.
There are \( \binom{m}{k} \) ways of choose the balls (indexed by)

\[ 1 \leq i_1 < i_2 < \cdots < i_k \leq m \] that fall in the bin.

The probability that these balls fall into the bin is \( \frac{1}{n^k} \).

The probability that other balls fall outside is \( \left(1 - \frac{1}{n}\right)^{m-k} \).

Let \( X_{i,=k} \) represent the indicator variable that bin \( i \) contains exactly \( k \) balls.

Note that, we have

\[
P [ X_{i,=k} ] = \binom{m}{k} \frac{1}{n^k} \left(1 - \frac{1}{n}\right)^{m-k}
\] (1)
Max Load

- Let $L_i$ represent the load of the $i$-th bin, i.e., the number of balls in the $i$-th bin
- Note that $L_i$ is the random variable $\{|k: T_k = i|\}$
- Let $M$ be the maximum load of the bins
- That is, $M$ is the random variable $\max\{L_1, \ldots, L_n\}$
- We are interested in understanding how $\mathbb{E}[M]$ behaves like

**Theorem (Max Load)**

*For $m = n$, we have*

$$\mathbb{E}[M] = \Theta\left(\frac{\log n}{\log \log n}\right)$$
First, we want to show that $\mathbb{M}$ is $\leq c \frac{\log n}{\log \log n}$ with $\approx 1$ probability.

This will imply that $\mathbb{E}[\mathbb{M}]$ is upper bounded by (roughly) $c \frac{\log n}{\log \log n}$.

This is known as the “First Moment Technique”.
Let $X_{i,\geq k}$ be the indicator variable for bin $i$ getting $\geq k$ balls

Note that

$$P[X_{i,\geq k} = 1] \leq \binom{m}{k} \frac{1}{n^k}$$

Think: Why is this true?

We will upper bound this probability further

$$P[X_{i,\geq k} = 1] \leq \binom{m}{k} \frac{1}{n^k} \leq \left(\frac{m}{n}\right)^k \frac{1}{k!}$$

By union bound:

$$P[\exists i \in [n]: X_{i,\geq k} = 1] \leq \left(\frac{m}{n}\right)^k \frac{n}{k!}$$
Analysis of Max Load (Upper Bound)

- Let $k = k^* = c \frac{\log n}{\log \log n}$ such that $k! \geq n^2$ and $m = n$
- We get

$$\mathbb{P} \left[ \exists i \in [n]: X_{i, \geq k^*} = 1 \right] \leq \frac{1}{n}$$

- Negating, we get:

$$\mathbb{P} \left[ \forall i \in [n]: X_{i, \geq k^*} = 0 \right] \geq 1 - \frac{1}{n}$$

- Note that the event "$\forall i \in [n]: X_{i, \geq k^*} = 0$" implies the event "$M < k^*$"
- So, we have $\mathbb{P} [M < k^*] \geq 1 - \frac{1}{n}$
- This implies that $\mathbb{E} [M] \leq \left(1 - \frac{1}{n}\right) (k^* - 1) + \frac{1}{n} \cdot n \leq k^*$
We are interested in showing that

\[ E[M] \geq d \frac{\log n}{\log \log n} \]

There are multiple ways to show this. In particular, we can use a “Second Moment Technique” to prove this result. One of the reading materials proves the result using this technique. We will, instead, use a more general technique that shows a close connection between the balls-and-bins problem and its approximation using independent Poisson Distributions.
Recall Equation 1. The probability of a bin to have $k$ balls is

$$P[X_i = k] = \binom{m}{k} \frac{1}{n^k} \left(1 - \frac{1}{n}\right)^{m-k}$$

$$= \binom{m}{k} \left(\frac{1}{n(1 - \frac{1}{n})}\right)^k \left(1 - \frac{1}{n}\right)^m$$

$$\approx \frac{1}{k!} \left(\frac{m}{n-1}\right)^k \left(1 - \frac{1}{n}\right)^m$$

$$\approx \exp \left(-\frac{m}{n}\right) \cdot \frac{(m/n)^k}{k!}$$
Let $Y$ be the distribution over the sample space $\{0, 1, 2, \ldots\}$ such that
\[
P[Y = k] = \exp(-\mu) \frac{\mu^k}{k!}
\]
Prove: This is a probability distribution with mean $\mu$
This distribution is the Poisson Distribution with mean $\mu$
Intuition of the Approximation

- **Reality**: The load distribution of $n$ bins when $m$ balls are thrown is represented by the joint random variables $(L_1, \ldots, L_n)$
- **Approximation**: Consider the distribution $(Y^{(1)}, \ldots, Y^{(n)})$, where each $Y^{(i)}$ is an independent Poisson distribution with mean $m/n$

**Theorem (Intuitive: Poisson Approximation)**

If $f$ is a “well-behaved” function, then

$$\mathbb{E}[f(L_1, \ldots, L_n)] \preceq \mathbb{E}[f(Y^{(1)}, \ldots, Y^{(n)})]$$

**Example**: We want to show that the max-load is $\geq d \frac{\log n}{\log \log n}$ with high probability. So, we choose $f$ as the indicator variable for the event that the maximum of the inputs is $< d \frac{\log n}{\log \log n}$. Then we show that the $\mathbb{E} \left[ f(Y^{(1)}, \ldots, Y^{(n)}) \right]$ is “small.” So, we have

$$\mathbb{E} \left[ f(L_1, \ldots, L_n) \right]$$ is also “small.”

Coupon Collector’s Problem

Think:

- How many balls need to be thrown so that every bin has at least one ball?
- How many balls need to be thrown so that every bin has at least \( r \) balls?