Lecture 03: Balls & Bins (Birthday Paradox, Max-Load)

- There are m balls and n bins
- Each balls is independently thrown into a bin that is chosen uniformly at random
- $(\mathbb{T}_1, \ldots, \mathbb{T}_m)$ be the joint distribution such that $\mathbb{T}_i$ represents the bin into which the i-th ball is thrown

Expected Number of Collisions

- For $1 \leqslant i < j \leqslant m$, let $\mathbb{X}_{i,j}$ 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 $\mathbb{T}_i = \mathbb{T}_j$
- We are interested in computing the "expectation of the random variable $\mathbb{X}_{i,j}$"

$$\begin{aligned}\mathbb{E}\left[\mathbb{X}_{i,j}\right] = \mathbb{P}\left[\mathbb{X}_{i,j} = 1\right] &= \mathbb{P}\left[\mathbb{T}_i = \mathbb{T}_j\right] \\ &= \sum_{t=1}^{n} \mathbb{P}\left[\mathbb{T}_i = \mathbb{T}_j = t\right] = n \cdot 1/n^2 = 1/n\end{aligned}$$

Expected Number of Collisions II

- Let $\mathbb{X} = \sum_{1 \leqslant i < j \leqslant m} \mathbb{X}_{i,j}$
- The random variable $\mathbb{X}$ counts the number of collisions that occur
- We are interested in the expected number of collisions

$$\begin{aligned}
\mathbb{E}\left[\mathbb{X}\right] &= \mathbb{E}\left[\sum_{1 \leqslant i < j \leqslant m} \mathbb{X}_{i,j}\right] \\
&= \sum_{1 \leqslant i < j \leqslant m} \mathbb{E}\left[\mathbb{X}_{i,j}\right] \qquad \text{By the Linearity of Expectation} \\
&= \binom{m}{2} \frac{1}{n}
\end{aligned}$$

- Note that for $m = \sqrt{2n}$, we have $\mathbb{E}\left[\mathbb{X}\right] = 1$
- 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 $\mathrm{NoColl}_{\leqslant t}$ represent the probability that $\mathbb{T}_1, \ldots, \mathbb{T}_t$ are all distinct
- Note that

$$\begin{aligned}\mathbb{P}\left[\mathrm{NoColl}_{\leqslant t}\right] &= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{t-1}{n}\right) \\ &= \prod_{t=0}^{t-1} \left(1 - \frac{t}{n}\right)\end{aligned}$$

Birthday Paradox – Balls and Bins Versions II

- We are interested in finding m such that the probability of collision is high
- Alternately, we are interested in showing the probability of $\mathrm{NoColl}_{\leqslant m}$ is small

$$\begin{aligned}\mathbb{P}\left[\mathrm{NoColl}_{\leqslant m}\right] &= \prod_{t=1}^{m-1}\left(1-\frac{t}{n}\right)\\ &\leqslant \prod_{t=1}^{m-1}\exp\left(-\frac{t}{n}\right) = \exp\left(-\sum_{t=0}^{m-1} t/n\right)\\ &= \exp\left(-(m-1)m/2n\right)\end{aligned}$$

- Substituting $m = c\sqrt{n}$, for a suitable constant $c > 0$, ensures that $\exp\left(-(m-1)m/2n\right) \leqslant 0.1$

- 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 $\mathrm{NoColl}_{\leqslant m}$ is high

$$\begin{aligned}\mathbb{P}\left[\mathrm{NoColl}_{\leqslant m}\right] &= \prod_{t=1}^{m-1}\left(1-\frac{t}{n}\right) \geqslant \prod_{t=0}^{m-1}\exp\left(-\frac{t}{n}-\frac{t^2}{n^2}\right)\\ &= \exp\left(-\sum_{t=0}^{m-1} t/n - \sum_{t=0}^{m-1} t^2/n^2\right)\\ &= \exp\left(-\frac{m(m-1)}{2n} - \frac{m(m-0.5)(m-1)}{3n^2}\right)\end{aligned}$$

- 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 $\geqslant 0.9$ s

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?

Number of Empty Bins

- Let $\mathbb{X}_i$ represent the indicator variable for the i-th bin being empty, i.e., the indicator of the event: $\mathbb{T}_j \neq i$, for all $i \in \{1, \ldots, m\}$
- Note that

$$\mathbb{E}[\mathbb{X}_i] = \mathbb{P}[\mathbb{X}_i = 1] = \left(1 - \frac{1}{n}\right)^m$$

- Let $\mathbb{X} = \sum_{t=0}^{n} \mathbb{X}_i$ represent the number of empty bins

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

- For $m = n$, we expect (roughly) n/e empty bins! For $m = n \log n$, we expect (roughly) 1 empty bin.

Probability of a Bin containing k Balls

- There are $\binom{m}{k}$ ways of choose the balls (indexed by) $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant m$ 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 $\mathbb{X}_{i,=k}$ represent the indicator variable that bin i contains exactly k balls
- Note that, we have

$$\mathbb{P}\left[\mathbb{X}_{i,=k}\right] = \binom{m}{k} \frac{1}{n^k} \left(1 - \frac{1}{n}\right)^{m-k} \tag{1}$$

Max Load

- Let $\mathbb{L}_i$ represent the load of the i-th bin, i.e., the number of balls in the i-th bin
- Note that $\mathbb{L}_i$ is the random variable $\left|\{k \colon \mathbb{T}_k = i\}\right|$
- Let $\mathbb{M}$ be the maximum load of the bins
- That is, $\mathbb{M}$ is the random variable $\max\{\mathbb{L}_1, \ldots, \mathbb{L}_n\}$
- We are interested in understanding how $\mathbb{E}\left[\mathbb{M}\right]$ behaves like

Theorem (Max Load)

For $m = n$, we have

$$\mathbb{E}\left[\mathbb{M}\right] = \Theta\left(\frac{\log n}{\log\log n}\right)$$

- First, we want to show that $\mathbb{M}$ is $\leqslant c\frac{\log n}{\log\log n}$ with ≈ 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 $\mathbb{X}_{i,\geqslant k}$ be the indicator variable for bin i getting $\geqslant k$ balls
- Note that

$$\mathbb{P}\left[\mathbb{X}_{i,\geqslant k} = 1\right] \leqslant \binom{m}{k} \frac{1}{n^k}$$

- Think: Why is this true?
- We will upper bound this probability further

$$\mathbb{P}\left[\mathbb{X}_{i,\geqslant k} = 1\right] \leqslant \binom{m}{k} \frac{1}{n^k} \leqslant \left(\frac{m}{n}\right)^k \frac{1}{k!}$$

- By union bound:

$$\mathbb{P}\left[\exists i \in [n] \colon \mathbb{X}_{i,\geqslant k} = 1\right] \leqslant \left(\frac{m}{n}\right)^k \frac{n}{k!}$$

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

$$\mathbb{P}\left[\exists i \in [n] \colon \mathbb{X}_{i,\geqslant k^*} = 1\right] \leqslant \frac{1}{n}$$

- Negating, we get:

$$\mathbb{P}\left[\forall i \in [n] \colon \mathbb{X}_{i,\geqslant k^*} = 0\right] \geqslant 1 - \frac{1}{n}$$

- Note that the event "$\forall i \in [n] \colon \mathbb{X}_{i,\geqslant k^*} = 0$" implies the event "$\mathbb{M} < k^*$"
- So, we have $\mathbb{P}\left[\mathbb{M} < k^*\right] \geqslant 1 - \frac{1}{n}$
- This implies that $\mathbb{E}\left[\mathbb{M}\right] \leqslant \left(1 - \frac{1}{n}\right)(k^* - 1) + \frac{1}{n} \cdot n \leqslant k^*$

Analysis of Max Load (Lower Bound)

- We are interested in showing that

$$\mathbb{E}\left[\mathbb{M}\right] \geqslant d\frac{\log n}{\log\log n}$$

- 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

$$\begin{aligned}
\mathbb{P}\left[\mathbb{X}_{i,=k}\right] &= \binom{m}{k}\frac{1}{n^k}\left(1-\frac{1}{n}\right)^{m-k} \\
&= \binom{m}{k}\left(\frac{1}{n\left(1-\frac{1}{n}\right)}\right)^k\left(1-\frac{1}{n}\right)^m \\
&\approx \frac{1}{k!}\left(\frac{m}{n-1}\right)^k\left(1-\frac{1}{n}\right)^m \\
&\approx \exp\left(-m/n\right)\cdot\frac{\left(m/n\right)^k}{k!}
\end{aligned}$$

Poisson Distribution

- Let $\mathbb{Y}$ be the distribution over the sample space $\{0, 1, 2, \dots\}$ such that

$$\mathbb{P}[\mathbb{Y} = k] = \exp(-\mu)\frac{\mu^k}{k!}$$

- Prove: This is a probability distribution with mean μ
- This distribution is the Poisson Distribution with mean μ

- **Reality**: The load distribution of n bins when m balls are thrown is represented by the joint random variables $(\mathbb{L}_1, \dots, \mathbb{L}_n)$
- **Approximation**: Consider the distribution $(\mathbb{Y}^{(1)}, \dots, \mathbb{Y}^{(n)})$, where each $\mathbb{Y}^{(i)}$ is an independent Poisson distribution with mean m/n

Theorem (Intuitive: Poisson Approximation)

If f is a "well-behaved" function, then

$$\mathbb{E}\left[f(\mathbb{L}_1, \dots, \mathbb{L}_n)\right] \lesssim \mathbb{E}\left[f(\mathbb{Y}^{(1)}, \dots, \mathbb{Y}^{(n)})\right]$$

Example: We want to show that the max-load is $\geqslant d\frac{\log n}{\log\log n}$ 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 $\mathbb{E}\left[f(\mathbb{Y}^{(1)}, \dots, \mathbb{Y}^{(n)})\right]$ is "small." So, we have $\mathbb{E}\left[f(\mathbb{L}_1, \dots, \mathbb{L}_n)\right]$ is also "small."

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?