Lecture 04: Balls and Bins: Birthday Paradox

Overview

- In today's lecture we will start our study of balls-and-bins problems
- We shall consider a fundamental problem known as the Birthday Paradox

Recall: Inequalities I

- Before we begin, let us recall a few inequalities from previous lectures.
- Using Taylor series, we had concluded the following fact.

Lemma

For any integer $k \geqslant 1$ *and* $x \in [0,1]$, *we have the following bound.*

$$\ln(1-x) \leqslant \left(-x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots - \frac{x^k}{k}\right)$$

- Using Taylor series, we had concluded the following fact.

Lemma

For any integer $k \geqslant 1$ *and* $x \in [0,1/2]$, *we have the following bound.*

$$\left(-x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots - \frac{x^k}{k}\right) - \frac{x^k}{k} \leqslant \ln(1-x)$$

Recall: Inequalities II

- In a previous lecture, we had seen that we can upper and lower-bound summations as integrals.

Lemma

Let c be a positive real number. Then, we have the following upper and lower bounds.

$$\frac{m^{c+1}}{c+1} > \int_1^m x^c \, \mathrm{d}x \geqslant \sum_{i=1}^{m-1} i^c \geqslant \int_0^{m-1} x^c \, \mathrm{d}x = \frac{(m-1)^{c+1}}{c+1}$$

Balls and Bins

- Let us introduce the Balls and Bins Experiment
- Suppose we have n bins and m balls
- We throw m balls into n bins independently and uniformly at random (Note that we have not assumed anything about whether $m < n$ or $m > n$)
- The "load of bin i" refers to the number of balls in bin i
- The "max-load" of the bins refers to the maximum load of the bins

Mathematical Formalization

- Our sample space is $[n]^{\otimes m}$
- Our random variables are $(\mathbb{X}_1, \ldots, \mathbb{X}_m)$, where $\mathbb{X}_i$ represents the bin into which the i-th ball falls. The random variable $\mathbb{X}_i$ is independent and uniformly distributed over $[n]$
- Now, the load of a bin $j \in [n]$ is the number of balls that fall into it. The random variable is represented as follows

$$\mathbb{L}_j = \sum_{i=1}^{m} \mathbf{1}_{\{\mathbb{X}_i = j\}}$$

- The max-load of the bins can be represented as the following random variable.

$$\mathbb{L}_{\max} = \max\{\mathbb{L}_1, \mathbb{L}_2, \ldots, \mathbb{L}_n\}$$

Expected Load I

Let us prove an interesting result about the load of any bin.

Theorem

For any $j \in [n]$, the expected load of the j-th bin is m/n.

Expected Load II

Proof.

$$
\begin{aligned}
\mathbb{E}[\mathbb{L}_j] &= \mathbb{E}\left[\sum_{i=1}^{m} \mathbf{1}_{\{\mathbb{X}_i = j\}}\right], && \text{By definition of the r.v.} \\
&= \sum_{i=1}^{m} \mathbb{E}\left[\mathbf{1}_{\{\mathbb{X}_i = j\}}\right], && \text{By linearity of expectation} \\
&= \sum_{i=1}^{m} \mathbb{P}[\mathbb{X}_i = j], && \text{By properties of indicator variables} \\
&= \sum_{i=1}^{m} \frac{1}{n}, && \text{Because } \mathbb{X}_i \text{ is uniform over } [n] \\
&= \frac{m}{n}
\end{aligned}
$$

$\square$

Expected Load III

- Note that the proof does not rely on the fact that random variables $\mathbb{X}_i$s are independent!

  So, even if the balls are thrown in a "correlated fashion," as long as $\mathbb{P}\left[\mathbb{X}_i = j\right] = 1/n$, for all $i \in [m]$, the proof will hold.

  Consider the following new way of throwing the balls. "Choose a bin uniformly at random and throw all the balls into that bin."

  Note that in this manner of throwing balls, we still have $\mathbb{P}\left[X_i = j\right] = 1/n$, for all $i \in [m]$ and $j \in [n]$. So, the expected number of balls in the j-th bin is still m/n.

English Formulation

- Assume that birthdays are distributed uniformly and independently at random over 365 days of the year
- Suppose we have m people in a room
- What is the probability that there are (at least) two people who share the same birthday?
- Alternatively, what is the probability that all m people have distinct birthdays?

Interestingly, as increase the number m we find that the event of "distinct birthdays" turns from a "likely event" to an "unlikely event" very quickly. Our goal is to study this phenomenon.

Mathematical Formulation

- We shall consider "people" as balls. And "birthdays" as bins.
- We are throwing m balls into n bins
- Note that the event "every ball falls into a distinct bin" is equivalent to the event "$\mathbb{L}_{\max} = 1$"
- So, we are interested in study the following probability

  $$P_{m,n} := \mathbb{P}\left[\mathbb{L}_{\max} = 1\right]$$

  as a function of m and n
- It is clear that for $m = 1$, we have $P_{m,n} = 1$. And, for $m = n + 1$, we have $P_{m,n} = 0$.
- In fact, in the previous lecture, we had calculated this probability exactly

  $$P_{m,n} = \prod_{i=0}^{m-1} \left(1 - \frac{i}{n}\right)$$

Why Bound $P_{m,n}$?

- Note that the exact formula for $P_{m,n}$ is very opaque. We do not understand its properties clearly from that formula.
- Our goal, therefore, is to obtain tight upper and lower bound for this expression using simpler formulas

Upper Bound I

- Let us start with the exact formula

$$P_{m,n} = \prod_{i=0}^{m-1} \left(1 - \frac{i}{n}\right)$$

- We do not like "products of polynomials." Let us turn the expression on the right-hand side into a summation.

$$\ln P_{m,n} = \sum_{i=0}^{m-1} \ln \left(1 - \frac{i}{n}\right)$$

Upper Bound II

- This is still problematic. The right-hand side expressions are "logarithmic." But we can upper bound $\ln(1-x)$ using polynomial in x. For any integer $k \geqslant 1$, we get

$$\begin{aligned}\ln P_{m,n} &= \sum_{i=0}^{m-1} \ln\left(1-\frac{i}{n}\right)\\ &\leqslant \sum_{i=0}^{m-1} -\left(\frac{i}{n}\right) - \left(\frac{i}{n}\right)^2/2 - \cdots - \left(\frac{i}{n}\right)^k/k\end{aligned}$$

- Now we can individually bound the sum $\sum_{i=0}^{m-1} i^c \geqslant \frac{(m-1)^{c+1}}{c+1}$, for each $c \in [k]$. We get

$$\ln P_{m,n} \leqslant -\frac{(m-1)^2}{2n} - \frac{(m-1)^3}{2\cdot 3n^2} - \frac{(m-1)^4}{3\cdot 4n^3} - \cdots - \frac{(m-1)^{k+1}}{k(k+1)n^k}$$

- Please use desmos to see the tightness of this upper-bound.

Upper Bound III

How to use this bound?

- Suppose we want to find out m (as a function of n) such that $P_{m,n} \leqslant 0.1$.
- To find such an m, let us find m such that

$$-\frac{(m-1)^2}{2n} - \frac{(m-1)^3}{2\cdot 3n^2} - \frac{(m-1)^4}{3\cdot 4n^3} - \cdots - \frac{(m-1)^{k+1}}{k(k+1)n^k} = \ln 0.1$$

  For this value of m, we will have $P_{m,n} \leqslant 0.1$.
- Note that if $(m-1) = \beta\sqrt{n}$ then the left hand side of the expression above is

$$-(\beta^2/2) - O(n^{-1/2}) = \ln 0.1$$

  This implies that
  $\beta = \sqrt{-2\ln 0.1 - O(n^{-1/2})} = \sqrt{\ln 100 - O(n^{-1/2})}$
- Conclusion: At $m \geqslant \text{const.}\sqrt{n}$ the probability $P_{m,n}$ falls below 0.1 (i.e., collisions are likely)

- We now prove a lower-bound using similar techniques. Let k be any positive integer.

$$\begin{aligned}
\ln P_{m,n} &= \sum_{i=0}^{m-1} \ln\left(1 - \frac{i}{n}\right) \\
&\geqslant \sum_{i=0}^{m-1} -\frac{i}{n} - \frac{i^2}{2n} - \cdots - \frac{i^k}{kn} - \frac{i^k}{kn} \\
&> -\frac{m^2}{2n} - \frac{m^3}{2 \cdot 3n} - \cdots - \frac{m^{k+1}}{k \cdot (k+1)n} - \frac{m^{k+1}}{k \cdot (k+1)n}
\end{aligned}$$

Lower Bound II

How to use this bound?

- Suppose we want to find out m (as a function of n) such that $P_{m,n} \geqslant 0.9$.
- To find such an m, let us find m such that

$$-\frac{m^2}{2n} - \frac{m^3}{2 \cdot 3n^2} - \cdots - \frac{m^{k+1}}{k \cdot (k+1)n^k} - \frac{m^{k+1}}{k \cdot (k+1)n^k} = \ln 0.9$$

  For this value of m, we will have $P_{m,n} \geqslant 0.9$.
- Note that if $m = \alpha\sqrt{n}$ then, for $k \geqslant 2$, the left hand side of the expression above is

$$-(\alpha^2/2) - O(n^{-1/2}) = \ln 0.9$$

  This implies that $\alpha = \sqrt{\ln(1/0.81) - O(n^{-1/2})}$
- Conclusion: At $m \leqslant \text{const.}\sqrt{n}$ the probability $P_{m,n}$ is above 0.9 (i.e., collisions are unlikely)

Birthday Bound: Conclusion

- So, collisions are unlikely at $m \leqslant \alpha\sqrt{n}$ and are likely at $m \geqslant \beta\sqrt{n}$
- A small increase of $(\beta - \alpha)\sqrt{n}$ in the value of m causes the probability of collisions transition from "low" to "high"
- This surprising phenomenon is referred to as the birthday paradox

Graphs of the Bounds

Check the code for an explanation of the upper and lower bounds on the birthday problem.

- The number n represents the number of bins. You can use the slider the change its values.
- The Y-axis represents probability. The X-axis represents m, the number of balls.
- We are interested in two thresholds. When does $P_{m,n}$ reach 0.9? And, when does $P_{m,n}$ reach 0.1?
- We plot the exact $P_{m,n}$ curve
- The value k represents the parameter k in the approximation used in our lecture today. Increasing k make the upper and lower bounds tighter. You can use the slider to change its value.
- Finally, we have the upper and the lower bounds to the $P_{m,n}$ curve

Alternate Technique to counting Collisions I

- Let $\mathbb{C}_{i,j}$ represent the event that balls i and j fall into the same bin
- Formally, we write this as follows. For $i, j \in [m]$ such that $i < j$ (this restriction avoids double counting) we define

$$\mathbb{C}_{i,j} := \mathbf{1}_{\{\mathbb{X}_i = \mathbb{X}_j\}}$$

- We are interested in the total number of such collisions. That is

$$\mathbb{C} := \sum_{\substack{i,j \in [m] \\ i<j}} \mathbb{C}_{i,j}$$

- Now, we are interested in computing its expected value

Alternate Technique to counting Collisions II

First let us begin with some preliminary observations regarding why $\mathbb{C}$ is a good measure of collisions.

- Note that if there exists a bin with ℓ balls in it, then we have $\mathbb{C} \geqslant \binom{\ell}{2}$
- So, if there exists two balls that collide, then we have $\ell \geqslant 2$ and, hence, $\mathbb{C} \geqslant \binom{2}{2} \geqslant 1$
- Further, we have $\mathbb{C} \geqslant \binom{\mathbb{L}_{\mathsf{max}}}{2}$

Alternate Technique to counting Collisions III

- Now, let us calculate the expected value of $\mathbb{C}$

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{C}\right] &= \mathbb{E}\left[\sum_{\substack{i,j\in[m]\\ i<j}} \mathbf{1}_{\{\mathbb{X}_i=\mathbb{X}_j\}}\right] \\
&= \sum_{\substack{i,j\in[m]\\ i<j}} \mathbb{E}\left[\mathbf{1}_{\{\mathbb{X}_i=\mathbb{X}_j\}}\right] \\
&= \sum_{\substack{i,j\in[m]\\ i<j}} \mathbb{P}\left[\mathbb{X}_i = \mathbb{X}_j\right] \\
&= \sum_{\substack{i,j\in[m]\\ i<j}} \frac{1}{n} = \binom{m}{2}\frac{1}{n}
\end{aligned}
$$

Alternate Technique to counting Collisions IV

- Note that if $m \approx \sqrt{2n}$ then $\mathbb{E}\left[\mathbb{C}\right]$ is (roughly) 1, i.e., we expect two balls to fall in one bin. Earlier we showed that if $m \geqslant \alpha\sqrt{n}$ then the probability of collision is $\geqslant 0.9$, and if $m \leqslant \beta\sqrt{n}$ then the probability of collision is $\leqslant 0.1$. The expected value of collisions becomes 1 in the intermediate zone (please plot this and check)

Note on a subtlety.

- Note that we only rely on the fact that $\mathbb{P}\left[\mathbb{X}_i = \mathbb{X}_j\right] = \frac{1}{n}$, for distinct i and j
- We do not need that all the balls are thrown independently
- It suffices if the random variables $(\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_m)$ are 2-wise independent

Next Lecture

- In the next lecture, we shall study the following quantity

$$\mathbb{E}\left[\mathbb{L}_{\mathsf{max}}\right]$$

- Later in the course, we shall study the concentration of $\mathbb{L}_{\mathsf{max}}$ around the expected value