Lecture 10: Birthday Paradox

Problem Statement

- Let S be a set of size n
- Suppose $(X_1, X_2, \ldots, X_n)$ are identical and independent distributions, such that X_i is the uniform distribution over the set S
- We say that a Collision has happened if there exists $i \neq j$ such that $X_i = X_j$
- We want to understand the probability

$$\mathbb{P}\left[Collision\right]$$

as a function of k and n

Example: Birthday Problem

- Assume that the birthdays of people are uniformly distributed over 365 days
- Given a sample of k randomly chosen people, what is the probability that two people share the same birthday?

Example: Hash Function Collision

- Let $f : D \rightarrow R$ be a function from the domain D and range R
- Assume that if $x \in D$ is picked uniformly at random from D, then $f(x)$ is uniformly at random in R
- How many samples $\{x_1, \ldots, x_k\}$ should you obtain before discovering a collision of f?

Summary of the Result

- We shall explore the asymptotic behavior of $\mathbb{P}[\mathit{Collision}]$ as $n \to \infty$
- We shall show that if $k \leqslant c_1\sqrt{n}$ then $\mathbb{P}[\mathit{Collision}] \leqslant 0.1$, for a suitable constant c_1
- We shall also show that if $k \geqslant c_2\sqrt{n}$ then $\mathbb{P}[\mathit{Collision}] \geqslant 0.9$, for a suitable constant c_2
- Intuitively, sampling only (roughly) $\sqrt{n}$ samples, the $\mathbb{P}[\mathit{Collision}]$ suddenly <u>transitions</u> from 0.1 to 0.9!

Inequalities

- We shall use the following inequalities without proof

$$\exp\left(-x-\frac{3}{4}x^2\right) \leqslant 1-x \leqslant \exp(-x) \leqslant 1-x+x^2/2$$

- The red inequality holds for $x \in [0, c]$, where c is a suitable constant in the range $(0, 1)$
- The remaining inequalities hold for all $x \in [0, 1]$
- These inequalities can be proven using The Remainder Theorem for Taylor Expansion of Functions
- It is recommended to plot these functions and verify the inequalities

Calculating the probability Expression I

- It will be easy to calculate $\mathbb{P}[NoCollision]$
- Note that $\mathbb{P}[NoCollision] = \mathbb{P}\left[\forall i \neq j \colon X_i \neq X_j\right]$
- This is identical to the probability that all the following events hold simultaneously
  - $X_2 \neq X_1$ (call this event E_2)
  - $X_3 \neq X_1$ and $X_3 \neq X_2$ (call this event E_3)
  - $X_4 \neq X_1$, $X_4 \neq X_2$, and $X_4 \neq X_3$ (call this event E_4)
  - and so on ...

Calculating the probability Expression II

- So, we are interested in computing

$$\mathbb{P}\left[E_2, E_3, E_4, \ldots, E_k\right]$$

- By Chain Rule, this expression is identical to

$$\mathbb{P}\left[E_2\right] \cdot \mathbb{P}\left[E_3 | E_2\right] \cdot \mathbb{P}\left[E_4 | E_2, E_3\right] \cdots \mathbb{P}\left[E_k | E_2, E_3, \ldots, E_{k-1}\right]$$

- Note that $\mathbb{P}\left[E_2\right] = (n-1)/n$ (because X_2 can take any value other than the value taken by X_1)
- Note that $\mathbb{P}\left[E_3 | E_2\right] = (n-2)/n$ (because the event E_2 implies that X_1 and X_2 have distinct values, and X_3 needs to avoid the two values taken by X_1 and X_2)
- Similarly, we have $\mathbb{P}\left[E_4 | E_2, E_3\right] = (n-3)/n$(because the event E_2 and E_3 imply that X_1, X_2, and X_3 have distinct values, and X_4 needs to avoid the three values taken by X_1, X_2, and X_3)

- Extending this logic, for all $i \in \{2, \ldots, k\}$, we can conclude that

$$\mathbb{P}\left[E_i | E_2, E_3, \ldots, E_{i-1}\right] = \frac{n-(i-1)}{n} = 1 - \frac{i-1}{n}$$

- Now, we can calculate

Final Result

$$\begin{aligned}\mathbb{P}\left[NoCollision\right] &= \mathbb{P}\left[E_2, E_3, \ldots, E_k\right] \\ &= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) \\ &= \prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right)\end{aligned}$$

Upper-bounding the probability of No Collision

$$\begin{aligned}
\prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right) &\leqslant \prod_{i=1}^{k-1}\exp\left(-\frac{i}{n}\right), && \text{Using } 1-x \leqslant \exp(-x) \\
&= \exp\left(-\sum_{i=1}^{k-1}\frac{i}{n}\right) \\
&= \exp\left(-\frac{(k-1)k}{2n}\right) \\
&\leqslant 1-\frac{(k-1)k}{2n}+\frac{(k-1)^2k^2}{8n^2}, && \text{Using } \exp(-x) \leqslant 1-x+x^2/2
\end{aligned}$$

Lower-bounding the probability of No Collision

$$\prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right) \geqslant \prod_{i=1}^{k-1}\exp\left(-\frac{i}{n}-\frac{3i^2}{4n^2}\right),$$

Using $1-x \geqslant \exp(-x-3x^2/4)$

We can use this inequality

because we shall only use $k = o(n)$

$$= \exp\left(-\sum_{i=1}^{k-1}\frac{i}{n}+\frac{3i^2}{4n^2}\right)$$

$$= \exp\left(-\frac{(k-1)k}{2n}-\frac{(k-1)(k-1/2)k}{4n^2}\right)$$

$$\geqslant 1-\frac{(k-1)k}{2n}-\frac{(k-1)(k-1/2)k}{4n^2},$$

Using $\exp(-x) \geqslant 1-x$

Summary of the Bounds

$$1-\frac{(k-1)k}{2n}+\frac{(k-1)^2k^2}{8n^2} \geqslant \mathbb{P}\left[NoCollision\right] \geqslant 1-\frac{(k-1)k}{2n}-\frac{(k-1)(k-1/2)k}{4n^2}$$

Or, equivalently

$$\frac{(k-1)k}{2n}-\frac{(k-1)^2k^2}{8n^2} \leqslant \mathbb{P}\left[Collision\right] \leqslant \frac{(k-1)k}{2n}+\frac{(k-1)(k-1/2)k}{4n^2}$$

- So, we can choose $k = c_1\sqrt{n}$ such that $\mathbb{P}\left[Collision\right] \leqslant 0.1$ and we can choose $k = c_2\sqrt{n}$ such that $\mathbb{P}\left[Collision\right] \geqslant 0.9$
- Plot and verify these bounds

- Recommended Exercise: Use the fact that $1 - x \leqslant \exp\left(-x - x^2/2\right)$ to obtain a better upper bound