Randomized Algorithms and Probabilistic Techniques

September 16, 2021

1 08/24

This course will involve only a few techniques with a variety of applications. Many domains, such as Distributed Computing, Networks, etc., all depend on randomized algorithms.

We will focus on a few techniques that lead to understanding.

1. Union Bound
2. Linearity of Expectation
3. Markov's Inequality
4. Chernoff Bounds

These four things prove very useful in the design and understanding of algorithms.

1.1 Basic Definitions

Definition 1.1

A *Sample Space* is a set S whose elements consist of *simple events* (also called *elementary* events). When S is finite or countably infinite, we say it is a *Discrete* sample space.

Definition 1.2

An *event* in a sample space S is a subset of S.

Definition 1.3

A *Probability Distribution* on S is a function $\mathbb{P} : 2^S \to [0, 1]$ that satisfies

1. $\mathbb{P}(S) = 1$
2. If E_1, E_2, ... are pairwise disjoint events (i.e., $E_i \cap E_j = \emptyset$ for all pairs i, j, also called mutually exclusive), ed by some finite or countably infinite set I, then

$$\mathbb{P}\left(\bigcup_{i \in I} E_i\right) = \sum_{i \in I} \mathbb{P}(E_i)$$

1.2 Conditional Probability

The expression $\mathbb{P}(E_2 \mid E_1)$ is read "the probability of E_2 given E_1." For example, if we randomly select a person from Texas, we might write

$$E_1 = \text{person chosen is in Houston}$$
$$E_2 = \text{person chosen is a UH student}$$

in which case $\mathbb{P}(E_2 \mid E_1)$ is simply the probability that a randomly selected person from Texas is a UH student *given that they are in Houston.*

Definition 1.4 ▶ Conditional Probability

The conditional probability $\mathbb{P}(E_2 \mid E_1)$ is defined

$$\mathbb{P}(E_2 \mid E_1) = \frac{\mathbb{P}(E_2 \cap E_1)}{\mathbb{P}(E_1)}$$

Intuitively, this can be thought of as taking the probability that E_2 and E_1 occur and "normalizing it" by dividing by the probability that E_1 occurs.

1.3 Independence

Definition 1.5 ▶ Independent Events

Two events, E_1 and E_2, are *independent* if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\,\mathbb{P}(E_2)$$

or, equivalently, if

$$\mathbb{P}(E_1 \mid E_2) = \mathbb{P}(E_1)$$

For example, suppose C_1 and C_2 are the outcomes of two fair coin tosses. These are independent, since

$$\mathbb{P}(C_1 = H \cap C_2 = T) = \mathbb{P}(C_1 = H)\,\mathbb{P}(C_2 = T) = \frac{1}{4}$$

Independence is not always related to physical independence. For example, say we are given a fair die and let

$$\begin{aligned} E_1 &= \text{roll is even} \\ E_2 &= \text{roll is less than or equal to } 4 \end{aligned}$$

In this case, we can enumerate the sample space and explicitly determine $\mathbb{P}(E_1)$, $\mathbb{P}(E_2)$, $\mathbb{P}(E_1 \cap E_2)$, and $\mathbb{P}(E_1)\,\mathbb{P}(E_2)$, to see if the events are independent:

$$\begin{aligned} E_1 &= \{2, 4, 6\} \\ E_2 &= \{1, 2, 3, 4\} \\ E_1 \cap E_2 &= \{2, 4\} \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}(E_1) &= \frac{3}{6} = \frac{1}{2} \\ \mathbb{P}(E_2) &= \frac{4}{6} = \frac{2}{3} \\ \mathbb{P}(E_1 \cap E_2) &= \frac{2}{6} = \frac{1}{3} \\ \mathbb{P}(E_1)\,\mathbb{P}(E_2) &= \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3} \end{aligned}$$

Thus, we see the events are independent. On the other hand, if E_2 is the event that the roll is *strictly less* than 4, we have

$$\begin{aligned} E_1 &= \{2, 4, 6\} \\ E_2 &= \{1, 2, 3\} \\ E_1 \cap E_2 &= \{2\} \end{aligned}$$

Then

$$\begin{aligned}\mathbb{P}(E_1) &= \frac{3}{6} = \frac{1}{2}\\ \mathbb{P}(E_2) &= \frac{3}{6} = \frac{1}{2}\\ \mathbb{P}(E_1 \cap E_2) &= \frac{1}{6}\\ \mathbb{P}(E_1)\,\mathbb{P}(E_2) &= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}\end{aligned}$$

and we see that the events are *not independent*.

1.4 The Inclusion-Exclusion Principle

A basic result in set theory is that

$$\begin{aligned}|A \cup B| &= |A| + |B| - |A \cap B|\\ |A \cup B \cup C| &= |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|\end{aligned}$$

which can be generalized to an arbitrary finite union by

$$\bigcup_{i=1}^{n} A_i = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \le i_1 < \cdots < i_k \le n} |A_{i_1} \cap \cdots \cap A_{i_k}| \right)$$

or equivalently

$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{\emptyset \ne J \subseteq \{1,\ldots,n\}} (-1)^{|J|+1} \left| \bigcap_{j \in J} A_j \right|$$

This yields the probability formulas

Theorem 1.1 ▶ Inclusion-Exclusion Principle

For any events E_1, E_2, …, E_n,

$$\begin{aligned}\mathbb{P}(E_1 \cup E_2) &= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)\\ \mathbb{P}\left(\bigcup_{i=1}^{n} E_i \right) &= \sum_{\emptyset \ne J \subseteq \{1,\ldots,n\}} (-1)^{|J|+1}\, \mathbb{P}\left(\bigcap_{j \in J} E_j \right)\end{aligned}$$

1.5 Union Bound

Theorem 1.2 ▶ Union Bound

Let E_1, E_2, …, be any countable set of events. Then

$$\mathbb{P}\Big(\bigcup E_i \Big) \le \sum \mathbb{P}(E_i)$$

While this bound is often not very precise, it is useful in many cases where the events E_i are "bad" and we can bound the likelihood of a single E_i. This allows us to bound the likelihood of *any* E_i.

1.6 Conditioning on Multiple Events (Chain Rule)

By repeatedly applying the definition of conditional probability, we have

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) &= \prod_{i=1}^{n} \mathbb{P}(E_i \mid E_1 \cap E_2 \cap \cdots \cap E_{i-1})\\ &= \mathbb{P}(E_1)\,\mathbb{P}(E_2 \mid E_1) \cdots \mathbb{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1})\end{aligned}$$

1.7 Birthday Paradox

Problem 1.1 ▶ Birthday Paradox

Suppose n people are in a room and they have birthdays chosen uniformly at random from the 365 calendar days. What is the probability that two people share a birthday?

The probability that *no two* people share a birthday can be calculated by considering the following events:

$$\begin{aligned} E_1 &= \text{person 1 has a birthday} \\ E_2 &= \text{person 2 has a different birthday than person 1} \\ &\vdots \\ E_i &= \text{person } i \text{ has a different birthday than people 1 through } i-1 \end{aligned}$$

Specifically, the probability that no two people share a birthday is

$$\begin{aligned} \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) &= \mathbb{P}(E_1)\,\mathbb{P}(E_2 \mid E_1) \cdots \mathbb{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \\ &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{365-n+1}{365}\right) \end{aligned}$$

For what value of n is the above probability less than $1/2$?

2 08/26

2.1 Birthday Paradox

The probability that no two people share a birthday is

$$\begin{aligned} \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) &= \mathbb{P}(E_1)\,\mathbb{P}(E_2 \mid E_1) \cdots \mathbb{P}(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \\ &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{365-n+1}{365}\right) \end{aligned}$$

Applying the inequality $1 - x \leq e^{-x}$ when $|x| < 1$

$$\begin{aligned} &\leq 1 \cdot e^{-\frac{1}{365}} \cdot e^{-\frac{2}{365}} \cdots e^{-\frac{n-1}{365}} \\ &= e^{-\frac{n(n-1)}{2 \cdot 365}} \end{aligned}$$

Setting

$$e^{-\frac{n(n-1)}{2 \cdot 365}} \leq \frac{1}{2}$$

yields $n \geq 23$.

2.2 Randomized Algorithms

Problem 2.1 ▶ Toy Problem

Let A be an array of n numbers with the property that at least one number occurs at least $n/2$ times and the rest are distinct. Our goal is to find the element with duplicate entries.

Any deterministic algorithm will require at least $n/2 + 1$ operations. However, we can construct a randomized algorithm that only requires $\mathcal{O}(\log n)$ operations.

We do this by *random sampling*. Assume that choosing a single element in $\{a_1, a_2, \ldots, a_n\}$ requires $\mathcal{O}(1)$ time and that we sample *with replacement*. We will frequently use sampling with replacement because it allows the samples to be independent, and therefore simplifies the analysis. Now, consider the following algorithm:

Algorithm 1 Randomized Algorithm for Toy Problem

```
Function RAND-DUPLICATE-TOY(arr):
    loop:
        sample_1 ← SAMPLE(arr)
        sample_2 ← SAMPLE(arr)
        if sample_1 ≠ sample_2:
            return sample_1
```

This algorithm simply chooses two elements (not necessarily distinct) at random from the array and repeats until the two elements are identical, after which it outputs the sampled element as the duplicate.

Let us analyze the probability that this algorithm outputs the duplicate value after a single iteration. Suppose the two values chosen are a_i and a_j. Then

$$\begin{aligned}\mathbb{P}(a_i = a_j \text{ and } i \neq j) &= \frac{n/2}{n} \cdot \frac{n/2 - 1}{n} \\ &= \frac{1}{2}\left(\frac{1}{2} - \frac{1}{n}\right) \\ &= \frac{1}{4} - \frac{1}{2n} \\ &\geq \frac{3}{2} \text{ for } n \geq 5\end{aligned}$$

What is the probability, then, that our algorithm fails after k iterations of Algorithm 1? These events are independent, hence the probability is bounded above by

$$\begin{aligned}\mathbb{P}(\text{failure}) &\leq \left(1 - \frac{3}{20}\right)^k \\ &\leq e^{-\frac{3}{2}k}\end{aligned}$$

Setting $k = \frac{20}{3} \ln n$,

$$\begin{aligned}&= e^{-\ln n} \\ &= \frac{1}{n}\end{aligned}$$

2.3 Concepts in Randomized Algorithms

Definition 2.1

An algorithm is *Monte Carlo* if it has a non-zero probability of outputting an incorrect solution. If an algorithm will output the correct solution with probability 1, it is *Las Vegas*.

Definition 2.2 ▶ High Probability

Given an input of size n, we say an event occurs *with high probability* (whp) if it occurs with probability $1 - 1/n^c$ for some $c > 0$.

Problem 2.2 ▶ Modified Toy Problem

Let A be an array of n numbers with the property that *either* all elements are distinct *or* $n/2$ elements are repeated.

Let us define

$$\begin{aligned}E_1 &= \text{event that exactly } n/2 \text{ elements are distinct} \\ E_2 &= \text{event that all } n \text{ elements are distinct}\end{aligned}$$