Part I

Probability

Chapter 1

Probability Spaces

This chapter introduces the basics of discrete probability theory.

1 Probability Spaces and Events

Probability theory is a mathematical study of uncertain situations such as a dice game. In probability theory, we model a situation with an uncertain ***outcome*** as an ***experiment*** and reason carefully about the likelihood of various outcomes in precise mathematical terms.

Example 1.1. Suppose we have two *fair* dice, meaning that each is equally likely to land on any of its six sides. If we toss the dice, what is the chance that their numbers sum to 4? To determine the probability we first notice that there are a total of $6 \times 6 = 36$ distinct outcomes. Of these, only three outcomes sum to 4 (1 and 3, 2 and 2, and 3 and 1). The probability of the event that the number sum up to 4 is therefore

$$\frac{\text{\# of outcomes that sum to } 4}{\text{\# of total possible outcomes}} = \frac{3}{36} = \frac{1}{12}$$

Sample Spaces and Events. A ***sample space*** Ω is an arbitrary and possibly infinite (but countable) set of possible outcomes of a probabilistic experiment. Any experiment will return exactly one outcome from the set. For the dice game, the sample space is the 36 possible outcomes of the dice, and an experiment (roll of the dice) will return one of them. An ***event*** is any subset of Ω, and most often representing some property common to multiple outcomes. For example, an event could correspond to outcomes in which the dice add to 4—this subset would be of size 3. We typically denote events by capital letters from the start of the alphabet, e.g. A, B, C. We often refer to the individual elements of Ω as ***elementary events***. We assign a probability to each event. Our model for probability is defined as follows.

Definition 1.1 (Probability Space)**.** A probability space consists of a ***sample space*** Ω representing the set of possible outcomes, and a ***probability measure***, which is a function $\mathbf{P}$

from all subsets of Ω (the ***events***) to a probability (real number). These must satisfy the following axioms.

- **Nonnegativity:** $\mathbf{P}[A] \in [0, 1]$.
- **Additivity:** for any two disjoint events A and B (i.e., $A \cap B = \emptyset$),

$$\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] \ .$$

- **Normalization:** $\mathbf{P}[\Omega] = 1$.

Note (Infinite Spaces). Probability spaces can have countably infinite outcomes. The additivity rule generalizes to infinite sums, e.g., the probability of the event consisting of the union of infinitely many number of disjoint events is the infinite sum of the probability of each event.

Note. When defining the probability space, we have not specified carefully the exact nature of events, because they may differ based on the experiment and what we are interested in. We do, however, need to take care when setting up the probabilistic model so that we can reason about the experiment correctly. For example, each outcome of the sample space must correspond to one unique actual outcome of the experiment. In other words, they must be mutually exclusive. Similarly, any actual outcome of the experiment must have a corresponding representation in the sample space.

Example 1.2 (Throwing Dice)**.** For our example of throwing two dice, the sample space consists of all of the 36 possible pairs of values of the dice:

$$\Omega = \{(1,1), (1,2), \ldots, (2,1), \ldots, (6,6)\}.$$

Each pair in the sample space corresponds to an outcome of the experiment. The outcomes are mutually exclusive and cover all possible outcomes of the experiment.

For example, having the first dice show up 1 and the second 4 is an outcome and corresponds to the element $(1, 4)$ of the sample space Ω.

The event that the "the first dice is 3" corresponds to the set

$$\begin{aligned} A &= \{(d_1, d_2) \in \Omega \mid d_1 = 3\} \\ &= \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\} \ . \end{aligned}$$

The event that "the dice sum to 4" corresponds to the set

$$\begin{aligned} B &= \{(d_1, d_2) \in \Omega \mid d_1 + d_2 = 4\} \\ &= \{(1,3), (2,2), (3,1)\} \ . \end{aligned}$$

Assuming the dice are unbiased, the probability measure is defined by all elementary events having equal probability, i.e.,

$$\forall x \in \Omega, \quad \mathbf{P}[\{x\}] = \frac{1}{36}.$$

The probability of the event A (that the first dice is 3) is thus

$$\mathbf{P}[A] = \sum_{x \in A} \mathbf{P}[\{x\}] = \frac{6}{36} = \frac{1}{6}.$$

If the dice were biased so the probability of a given value is proportional to that value, then the probability measure would be $\mathbf{P}[\{(x,y)\}] = \frac{x}{21} \times \frac{y}{21}$, and the probability of the event B (that the dice add to 4) would be

$$\mathbf{P}[B] = \sum_{x \in B} \mathbf{P}[\{x\}] = \frac{1 \times 3 + 2 \times 2 + 3 \times 1}{21 \times 21} = \frac{10}{441}.$$

2 Properties of Probability Spaces

Given a probability space, we can prove several properties of probability measures by using the three axioms that they must satisfy.

For example, if for two events A and B. We have

- if $A \subseteq B$, then $\mathbf{P}[A] \leq \mathbf{P}[B]$,
- $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B]$.

2.1 The Union Bound

The union bound, also known as Boole's inequality, is a simple way to obtain an upper bound on the probability of any of a collection of events happening. Specifically for a collection of events $A_0, A_2, \ldots, A_{n-1}$ the bound is:

$$\mathbf{P}\left[\bigcup_{0 \leq i < n} A_i\right] \leq \sum_{i=0}^{n-1} \mathbf{P}[A_i]$$

This bound is true unconditionally. To see why the bound holds we note that the elementary events in the union on the left are all included in the sum on the right (since the union comes from the same set of events). In fact they might be included multiple times in the sum on the right, hence the inequality. In fact the sum on the right could add to more than one, in which case the bound is not useful. The union bound can be useful in generating high-probability bounds for algorithms. For example, when the probability of each of n events is very low, e.g. $1/n^5$ and the sum remains very low, e.g. $1/n^4$.

2.2 Conditional Probability

Conditional probability allows us to reason about dependencies between observations. For example, suppose that your friend rolled a pair of dice and told you that they sum up to

6, what is the probability that one of dice has come up 1? Conditional probability has many practical applications. For example, given that a medical test for a disease comes up positive, we might want to know the probability that the patient has the disease. Or, given that your computer has been working fine for the past 2 years, you might want to know the probability that it will continue working for one more year.

Definition 1.2 (Conditional Probability)**.** For a given probability space, we define the ***conditional probability*** of an event A given B, as the probability of A occurring given that B occurs as

$$\mathbf{P}[A \mid B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}.$$

The conditional probability measures the probability that the event A occurs given that B does. It is defined only when $\mathbf{P}[B] > 0$.

Conditional Probability is a Probability Measure. Conditional probability satisfies the three axioms of probability measures and is itself a probability measure. We can thus treat conditional probabilities just as ordinary probabilities. Intuitively, conditional probability can be thought as a focusing and re-normalization of the probabilities on the assumed event B.

Example 1.3. Consider throwing two fair dice and calculate the probability that the first dice comes us 1 given that the sum of the two dice is 4. Let A be the event that the first dice comes up 1 and B the event that the sum is 4. We can write A and B in terms of outcomes as

$$\begin{aligned} A &= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\} \text{ and} \\ B &= \{(1,3),(2,2),(3,1)\}. \end{aligned}$$

We thus have $A \cap B = \{(1,3)\}$. Since each outcome is equally likely,

$$\mathbf{P}[A \mid B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]} = \frac{|A \cap B|}{|B|} = \frac{1}{3}.$$

2.3 Law of Total Probability

Conditional probabilities can be useful in estimating the probability of an event that may depend on a selection of choices. The total probability theorem can be handy in such circumstances.

Theorem 1.1 (Law of Total Probability)**.** Consider a probabilistic space with sample space Ω and let $A_0, \ldots, A_{n-1}$ be a partition of Ω such that $\mathbf{P}[A_i] > 0$ for all $0 \leq i < n$. For any event B the following holds:

$$\begin{aligned} \mathbf{P}[B] &= \sum_{i=0}^{n-1} \mathbf{P}[B \cap A_i] \\ &= \sum_{i=0}^{n-1} \mathbf{P}[A_i]\,\mathbf{P}[B \mid A_i] \end{aligned}$$

Example 1.4. Your favorite social network partitions your connections into two kinds, near and far. The social network has calculated that the probability that you react to a post by one of your far connections is 0.1 but the same probability is 0.8 for a post by one of your near connections. Suppose that the social network shows you a post by a near and far connection with probability 0.6 and 0.4 respectively.

Let's calculate the probability that you react to a post that you see on the network. Let A_0 and A_1 be the event that the post is near and far respectively. We have $\mathbf{P}[A_0] = 0.6$ and $\mathbf{P}[A_1] = 0.4$. Let B the event that you react, we know that $\mathbf{P}[B \mid A_0] = 0.8$ and $\mathbf{P}[B \mid A_1] = 0.1$.

We want to calculate $\mathbf{P}[B]$, which by total probability theorem we know to be

$$\begin{aligned}\mathbf{P}[B] &= \mathbf{P}[B \cap A_0] + \mathbf{P}[B \cap A_1] \\ &= \mathbf{P}[A_0]\,\mathbf{P}[B \mid A_0] + \mathbf{P}[A_1]\,\mathbf{P}[B \mid A_1]. \\ &= 0.6 \cdot 0.8 + 0.4 \cdot 0.1 \\ &= 0.52.\end{aligned}$$

2.4 Independence

It is sometimes important to reason about the dependency relationship between events. Intuitively we say that two events are independent if the occurrence of one does not affect the probability of the other. More precisely, we define independence as follows.

Definition 1.3 (Independence)**.** Two events A and B are ***independent*** if

$$\mathbf{P}[A \cap B] = \mathbf{P}[A] \cdot \mathbf{P}[B].$$

We say that multiple events $A_0, \ldots, A_{n-1}$ are ***mutually independent*** if and only if, for any non-empty subset $I \subseteq \{0, \ldots, n-1\}$,

$$\mathbf{P}\left[\bigcap_{i \in I} A_i\right] = \prod_{i \in I} \mathbf{P}[A_i].$$

Independence and Conditional Probability. Recall that $\mathbf{P}[A \mid B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}$ when $\mathbf{P}[B] > 0$. Thus if $\mathbf{P}[A \mid B] = \mathbf{P}[A]$ then $\mathbf{P}[A \cap B] = \mathbf{P}[A] \cdot \mathbf{P}[B]$. We can thus define independence in terms of conditional probability but this works only when $\mathbf{P}[B] > 0$.

Example 1.5. For two dice, the events $A = \{(d_1, d_2) \in \Omega \mid d_1 = 1\}$ (the first dice is 1) and $B = \{(d_1, d_2) \in \Omega \mid d_2 = 1\}$ (the second dice is 1) are independent since

$$\begin{array}{rlll} & \mathbf{P}[A] \times \mathbf{P}[B] & = & \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} \\ = & \mathbf{P}[A \cap B] & = & \mathbf{P}[\{(1,1)\}] = \frac{1}{36}. \end{array}$$

However, the event $C \equiv \{X = 4\}$ (the dice add to 4) is not independent of A since

$$\begin{array}{rlll} & \mathbf{P}[A] \times \mathbf{P}[C] & = & \frac{1}{6} \times \frac{3}{36} = \frac{1}{72} \\ \neq & \mathbf{P}[A \cap C] & = & \mathbf{P}[\{(1,3)\}] = \frac{1}{36}. \end{array}$$

A and C are not independent since the fact that the first dice is 1 increases the probability they sum to 4 (from $\frac{1}{12}$ to $\frac{1}{6}$).

Exercise 1.1. For two dice, let A be the event that first roll is 1 and B be the event that the sum of the rolls is 5. Are A and B independent? Prove or disprove.

Consider now the same question but this time define B to be the event that the sum of the rolls is 7.

Chapter 2

Random Variables

This chapter introduces the random variables and their use in probability theory.

Definition 2.1 (Random Variable)**.** A ***random variable*** X is a real-valued function on the outcomes of an experiment, i.e., $X : \Omega \to \mathbb{R}$, i.e., it assigns a real number to each outcome. For a given probability space there can be many random variables, each keeping track of different quantities. We typically denote random variables by capital letters from the end of the alphabet, e.g. X, Y, and Z. We say that a random variable is ***discrete*** if its range is finite or countable infinite. Throughout this book, we only consider discrete random variables.

Example 2.1. For throwing two dice, we can define random variable as the sum of the two dice

$$X(d_1, d_2) = d_1 + d_2 \ ,$$

the product of two dice

$$Y(d_1, d_2) = d_1 \times d_2 \ ,$$

or the value of the first dice the two dice:

$$Z(d_1, d_2) = d_1 \ .$$

Definition 2.2 (Indicator Random Variable)**.** A random variable is called an ***indicator random variable*** if it takes on the value 1 when some condition is true and 0 otherwise.

Example 2.2. For throwing two dice, we can define indicator random variable as getting doubles

$$Y(d_1, d_2) = \begin{cases} 1 & \text{if } d_1 = d_2 \\ 0 & \text{if } d_1 \neq d_2 \ . \end{cases}$$

Using our shorthand, the event $\{X = 4\}$ corresponds to the event "the dice sum to 4".

Notation. For a random variable X and a value $x \in \mathbb{R}$, we use the following shorthand for the event corresponding to X equaling x:

$$\{X = x\} \equiv \{y \in \Omega \mid X(y) = x\} \ ,$$

and when applying the probability measure we use the further shorthand

$$\mathbf{P}\left[X = x\right] \equiv \mathbf{P}\left[\{X = x\}\right] \ .$$

Example 2.3. For throwing two dice, and X being a random variable representing the sum of the two dice, $\{X = 4\}$ corresponds to the event "the dice sum to 4", i.e. the set

$$\{y \in \Omega \mid X(y) = 4\} = \{(1,3),(2,2),(3,1)\} \ .$$

Assuming unbiased coins, we have that

$$\mathbf{P}\left[X = 4\right] = 1/12 \ .$$

Remark. The term random variable might seem counter-intuitive because it is actually a function not a variable. As such, a random variable is not really random, because it is a well defined deterministic function on the sample space. However, when thought in conjunction with the random experiment that selects the events, a random variable takes on its value based on a random process.

1 Probability Mass Function

Definition 2.3 (Probability Mass Function)**.** For a discrete random variable X, we define its ***probability mass function*** or ***PMF***, written $\mathbf{P}_X(\cdot)$, for short as a function mapping each element x in the range of the random variable to the probability of the event $\{X = x\}$, i.e.,

$$\mathbf{P}_X(x) = \mathbf{P}\left[X = x\right].$$

Example 2.4. The probability mass function for the indicator random variable X indicating whether the outcome of a roll of dice is comes up even is

$$\begin{aligned}\mathbf{P}_X(0) &= \mathbf{P}\left[\{X = 0\}\right] = \mathbf{P}\left[\{1,3,5\}\right] = 1/2, \text{ and}\\ \mathbf{P}_X(1) &= \mathbf{P}\left[\{X = 1\}\right] = \mathbf{P}\left[\{2,4,6\}\right] = 1/2.\end{aligned}$$

The probability mass function for the random variable X that maps each outcome in a roll of dice to the smallest Mersenne prime number no less than the outcome is

$$\begin{aligned}\mathbf{P}_X(3) &= \mathbf{P}\left[\{X = 3\}\right] = \mathbf{P}\left[\{1,2,3\}\right] = 1/2, \text{ and}\\ \mathbf{P}_X(7) &= \mathbf{P}\left[\{X = 7\}\right] = \mathbf{P}\left[\{4,5,6\}\right] = 1/2.\end{aligned}$$

Note that much like a probability measure, a probability mass function is a non-negative function. It is also additive in a similar sense: for any distinct x and x', the events $\{X = x\}$ and $\{X = x'\}$ are disjoint. Thus for any set $\bar{x}$ of values of X, we have

$$\mathbf{P}[X \in \bar{x}] = \sum_{x \in \bar{x}} \mathbf{P}_X(x).$$

Furthermore, since X is a function on the sample space, the events corresponding to the different values of X partition the sample space, and we have

$$\sum_x \mathbf{P}_X(x) = 1.$$

These are the important properties of probability mass functions: they are non-negative, normalizing, and are additive in a certain sense.

We can also compute the probability mass function for multiple random variables defined for the same probability space. For example, the ***joint probability mass function*** for two random variables X and Y, written $\mathbf{P}_{X,Y}(x, y)$ denotes the probability of the event $\{X = x\} \cap \{Y = y\}$, i.e.,

$$\mathbf{P}_{X,Y}(x, y) = \mathbf{P}[\{X = x\} \cap \{Y = y\}] = \mathbf{P}[X = x, Y = y].$$

Here $\mathbf{P}[X = x, Y = y]$ is shorthand for $\mathbf{P}[\{X = x\} \cap \{Y = y\}]$.

In our analysis or randomized algorithms, we shall repeatedly encounter a number of well-known random variables and create new ones from existing ones by composition.

2 Bernoulli, Binomial, and Geometric RVs

Bernoulli Random Variable. Suppose that we toss a coin that comes up a head with probability p and a tail with probability $1 - p$. The ***Bernoulli random variable*** takes the value 1 if the coin comes up heads and 0 if it comes up tails. In other words, it is an indicator random variable indicating heads. Its probability mass function is

$$\mathbf{P}_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0. \end{cases}$$

Binomial Random Variable. Consider n Bernoulli trials with probability p. We call the random variable X denoting the number of heads in the n trials as the ***Binomial random variable***. Its probability mass function for any $0 \leq x \leq n$ is

$$\mathbf{P}_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Geometric Random Variable. Consider performing Bernoulli trials with probability p until the coin comes up heads and X denote the number of trials needed to observe the first head. The random variable X is called the ***geometric random variable***. Its probability mass function for any $0 < x$ is

$$\mathbf{P}_X(x) = (1-p)^{x-1}p.$$

3 Functions of Random Variables

It is often useful to "apply" a function to one or more random variables to generate a new random variable. Specifically if we have a function $f : \mathbb{R} \to \mathbb{R}$ and a random variable X we can compose the two giving a new random variable:

$$Y(x) = f(X(x))$$

We often write this shorthand as $Y = f(X)$. Similarly for two random variables X and Y we write $Z = X + Y$ as shorthand for

$$Z(x) = X(x) + Y(x)$$

or equivalently

$$Z = \lambda x.(X(x) + Y(x))$$

The probability mass function for the new variable can be computed by "massing" the probabilities for each value. For example, for a function of a random variable $Y = f(X)$, we can write the probability mass function as

$$\mathbf{P}_Y(y) = \mathbf{P}\left[Y = y\right] = \sum_{x \mid f(x)=y} \mathbf{P}_X(x)\,.$$

Example 2.5. Let X be a Bernoulli random variable with parameter p. We can define a new random variable Y as a transformation of X by a function $f(\cdot)$. For example, $Y = f(X) = 9X + 3$ is random variable that transforms X, e.g., $X = 1$ would be transformed to $Y = 12$. The probability mass function for Y reflects that of X, Its probability mass function is

$$\mathbf{P}_Y(y) = \begin{cases} p & \text{if } y = 12 \\ 1-p & \text{if } y = 3. \end{cases}$$

Example 2.6. Consider the random variable X with the probability mass function

$$\mathbf{P}_X(x) = \begin{cases} 0.25 & \text{if} \quad x = -2 \\ 0.25 & \text{if} \quad x = -1 \\ 0.25 & \text{if} \quad x = 0 \\ 0.25 & \text{if} \quad x = 1 \end{cases}$$

We can calculate the probability mass function for the random variable $Y = X^2$ as follows $\mathbf{P}_Y(y) = \sum_{x \mid x^2=y} \mathbf{P}_X(x)$. This yields

$$\mathbf{P}_Y(y) = \begin{cases} 0.25 & \text{if} \quad y = 0 \\ 0.5 & \text{if} \quad y = 1 \\ 0.25 & \text{if} \quad y = 4. \end{cases}$$

4 Conditioning

In the same way that we can condition an event on another, we can also condition a random variable on an event or on another random variable. Consider a random variable X and an event A in the same probability space, we define the ***conditional probability mass function*** of X conditioned on A as

$$\mathbf{P}_{X \mid A} = \mathbf{P}\left[X = x \mid A\right] = \frac{\mathbf{P}\left[\{X = x\} \cap A\right]}{\mathbf{P}\left[A\right]}.$$

Since for different values of x, $\{X = x\} \cap A$'s are disjoint and since X is a function over the sample space, conditional probability mass functions are normalizing just like ordinary probability mass functions, i.e., $\sum_{x \in X} \mathbf{P}_{X \mid A}(x) = 1$. Thus just as we can treat conditional probabilities as ordinary probabilities, we can treat conditional probability mass functions also as ordinary probability mass functions.

Example 2.7. Roll a pair of dice and let X be the sum of the face values. Let A be the event that the second roll came up 6. We can find the conditional probability mass function

$$\begin{aligned} \mathbf{P}_{X \mid A}(x) &= \frac{\mathbf{P}[\{X=x\} \cap A]}{\mathbf{P}[A]} \\ &= \begin{cases} \frac{1/36}{1/6} = 1/6 & \text{if } x = 7, \ldots, 12. \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Conditional Probability Mass Function. Since random variables closely correspond with events, we can condition a random variable on another. More precisely, let X and Y be two random variables defined on the same probability space. We define the ***conditional probability mass function*** of X with respect to Y as

$$\mathbf{P}_{X \mid Y}(x \mid y) = \mathbf{P}\left[X = x \mid Y = y\right].$$

We can rewrite this as

$$\begin{aligned} \mathbf{P}_{X \mid Y}(x \mid y) &= \mathbf{P}\left[X = x \mid Y = y\right] \\ &= \frac{\mathbf{P}[X=x, Y=y]}{\mathbf{P}[Y=y]} \\ &= \frac{\mathbf{P}_{X,Y}(x,y)}{\mathbf{P}_Y y}. \end{aligned}$$

Conditional PMFs are PMFs. Consider the function $\mathbf{P}_{X \mid Y}(x \mid y)$ for a fixed value of y. This is a non-negative function of x, the event corresponding to different values of x are disjoint, and they partition the sample space, the conditional mass functions are normalizing

$$\sum_x \mathbf{P}_{X \mid Y}(x \mid y) = 1.$$

Conditional probability mass functions thus share the same properties as probability mass functions.

By direct implication of its definition, we can use conditional probability mass functions to calculate joint probability mass functions as follows

$$\mathbf{P}_{X,Y}(x,y) = \mathbf{P}_X(x)\mathbf{P}_{Y \mid X}(y \mid x)$$
$$\mathbf{P}_{X,Y}(x,y) = \mathbf{P}_Y(y)\mathbf{P}_{X \mid Y}(x \mid y).$$

As we can compute total probabilities from conditional ones as we saw earlier in this section, we can calculate marginal probability mass functions from conditional ones:

$$\mathbf{P}_X(x) = \sum_y \mathbf{P}_{X,Y}(x,y) = \sum_y \mathbf{P}_Y(y)\mathbf{P}_{X \mid Y}(x \mid y).$$

5 Independence

As with the notion of independence between events, we can also define independence between random variables and events. We say that a random variable X is ***independent of an event*** A, if

$$\text{for all } x : \mathbf{P}\left[\{X = x\} \cap A\right] = \mathbf{P}\left[X = x\right] \cdot \mathbf{P}\left[A\right] \ .$$

When $\mathbf{P}\left[A\right]$ is positive, this is equivalent to

$$\mathbf{P}_{X \mid A}(x) = \mathbf{P}_X(x).$$

Generalizing this to a pair of random variables, we say a random variable X is ***independent of a random variable*** Y if

$$\text{for all } x, y : \mathbf{P}\left[X = x, Y = y\right] = \mathbf{P}\left[X = x\right] \cdot \mathbf{P}\left[Y = y\right]$$

or equivalently

$$\text{for all } x, y : \mathbf{P}_{X,Y}(x,y) = \mathbf{P}_X(x) \cdot \mathbf{P}_Y(y).$$

In our two dice example, a random variable X representing the value of the first dice and a random variable Y representing the value of the second dice are independent. However X is not independent of a random variable Z representing the sum of the values of the two dice.

Chapter 3

Expectation

This chapter introduces expectation and its use in probability theory.

1 Definitions

The ***expectation*** of a random variable X in a probability space $(\Omega, \mathbf{P})$ is the sum of the random variable over the elementary events weighted by their probability, specifically:

$$\mathbf{E}_{\Omega,\mathbf{P}}[X] = \sum_{y \in \Omega} X(y) \cdot \mathbf{P}\left[\{y\}\right].$$

For convenience, we usually drop the $(\Omega, \mathbf{P}\,[])$ subscript on $\mathbf{E}$ since it is clear from the context.

Example 3.1. Assuming unbiased dice ($\mathbf{P}\left[(d_1, d_2)\right] = 1/36$), the expectation of the random variable X representing the sum of the two dice is:

$$\mathbf{E}\left[X\right] = \sum_{(d_1,d_2)\in\Omega} X(d_1, d_2) \times \frac{1}{36} = \sum_{(d_1,d_2)\in\Omega} \frac{d_1 + d_2}{36} = 7.$$

If we bias the coins so that for each dice the probability that it shows up with a particular value is proportional to the value, we have $\mathbf{P}\left[(d_1, d_2)\right] = (d_1/21) \times (d_2/21)$ and:

$$\mathbf{E}\left[X\right] = \sum_{(d_1,d_2)\in\Omega} \left((d_1 + d_2) \times \frac{d_1}{21} \times \frac{d_2}{21}\right) = 8\,\frac{2}{3}.$$

It is usually more natural to define expectations in terms of the probability mass function of the random variable

$$\mathbf{E}\left[X\right] = \sum_{x} x \cdot \mathbf{P}_X(x).$$

Example 3.2. The expectation of an indicator random variable X is the probability that the associated predicate is true (i.e. that $X = 1$):

$$\begin{aligned}\mathbf{E}[X] &= 0 \cdot \mathbf{P}_X(0) + 1 \cdot \mathbf{P}_X(1). \\ &= \mathbf{P}_X(1).\end{aligned}$$

Example 3.3. Recall that the probability mass function for a Bernoulli random variable is

$$\mathbf{P}_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0. \end{cases}$$

Its expectation is thus

$$E[X] = p \cdot 1 + (1-p) \cdot 0 = p.$$

Example 3.4. Recall that the probability mass function for geometric random variable X with parameter p is

$$\mathbf{P}_X(x) = (1-p)^{x-1}p.$$

The expectation of X is thus

$$\begin{aligned}E[X] &= \sum_{x=1}^{\infty} x \cdot (1-p)^{x-1}p \\ &= p \cdot \sum_{x=1}^{\infty} x \cdot (1-p)^{x-1}\end{aligned}$$

Bounding this sum requires some basic manipulation of sums. Let $q = (1-p)$ and rewrite the sum as $p \cdot \sum_{x=0}^{\infty} xq^{x-1}$. Note now the term xq^{x-1} is the derivative of q^x with respect to q. Since the sum $\sum_{x=0}^{\infty} q^x = 1/(1-q)$, its derivative is $1/(1-q)^2 = 1/p^2$. We thus have conclude that $E[X] = 1/p$.

Example 3.5. Consider performing two Bernoulli trials with probability of success $1/4$. Let X be the random variable denoting the number of heads.

The probability mass function for X is

$$\mathbf{P}_X(x) = \begin{cases} 9/16 & \text{if } x = 0 \\ 3/8 & \text{if } x = 1 \\ 1/16 & \text{if } x = 2. \end{cases}$$

Thus $\mathbf{E}[X] = 0 + 1 \cdot 3/8 + 2 * 1/16 = 7/8$.

2 Composing Expectations

Recall that functions or random variables are themselves random variables (defined on the same probability space), whose probability mass functions can be computed by considering the random variables involved. We can thus also compute the expectation of a random

variable defined in terms of others. For example, we can define a random variable Y as a function of another variable X as $Y = f(X)$. The expectation of such a random variable can be calculated by computing the probability mass function for Y and then applying the formula for expectations. Alternatively, we can compute the expectation of a function of a random variable X directly from the probability mass function of X as

$$E[Y] = E[f(X)] = \sum_x f(x)\mathbf{P}_X(x).$$

Similarly, we can calculate the expectation for a random variable Z defined in terms of other random variables X and Y defined on the same probability space, e.g., $Z = g(X, Y)$, as computing the probability mass function for Z or directly as

$$E[Z] = E[g(X, Y)] = \sum_{x,y} g(x, y)\mathbf{P}_{X,Y}(x, y).$$

These formulas generalize to function of any number of random variables.

3 Linearity of Expectations

An important special case of functions of random variables is the linear functions. For example, let $Y = f(X) = aX + b$, where $a, b \in \mathbb{R}$.

$$\begin{aligned}
\mathbf{E}[Y] = \mathbf{E}[f(X)] &= \mathbf{E}[aX + b] \\
&= \textstyle\sum_x f(x)\mathbf{P}_X(x) \\
&= \textstyle\sum_x (ax + b)\mathbf{P}_X(x) \\
&= a\textstyle\sum_x x\mathbf{P}_X(x) + b\sum_x \mathbf{P}_X(x) \\
&= a\mathbf{E}[X] + b.
\end{aligned}$$

Similar to the example, above we can establish that the linear combination of any number of random variables can be written in terms of the expectations of the random variables. For example, let $Z = aX + bY + c$, where X and Y are two random variables. We have

$$\mathbf{E}[Z] = \mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

The proof of this statement is relatively simple.

$$\begin{aligned}
\mathbf{E}[Z] &= \mathbf{E}[aX + bY + c] \\
&= \textstyle\sum_{x,y} (ax + by + c)\mathbf{P}_{X,Y}(x, y) \\
&= a\textstyle\sum_{x,y} x\mathbf{P}_{X,Y}(x, y) + b\sum_{x,y} y\mathbf{P}_{X,Y}(x, y) + \sum_{x,y} c\mathbf{P}_{X,Y}(x, y) \\
&= a\textstyle\sum_x \sum_y x\mathbf{P}_{X,Y}(x, y) + b\sum_y \sum_x y\mathbf{P}_{X,Y}(x, y) + \sum_{x,y} c\mathbf{P}_{X,Y}(x, y) \\
&= a\textstyle\sum_x x\sum_y \mathbf{P}_{X,Y}(x, y) + b\sum_y y\sum_x \mathbf{P}_{X,Y}(x, y) + \sum_{x,y} c\mathbf{P}_{X,Y}(x, y) \\
&= a\textstyle\sum_x x\mathbf{P}_X(x) + b\sum_y y\mathbf{P}_Y(y) + c \\
&= a\mathbf{E}[X] + b\mathbf{E}[Y] + c.
\end{aligned}$$

An interesting consequence of this proof is that the random variables X and Y do not have to be defined on the same probability space. They can be defined for different experiments and their expectation can still be summed. To see why note that we can define the joint probability mass function $\mathbf{P}_{X,Y}(x, y)$ by taking the Cartesian product of the sample spaces of X and Y and spreading probabilities for each arbitrarily as long as the marginal probabilities, $\mathbf{P}_X(x)$ and $\mathbf{P}_Y(y)$ remain unchanged.

The property illustrated by the example above is known as the ***linearity of expectations***. The linearity of expectations is very powerful often greatly simplifying analysis. The reasoning generalizes to the linear combination of any number of random variables.

Linearity of expectation occupies a special place in probability theory, the idea of replacing random variables with their expectations in other mathematical expressions do not generalize. Probably the most basic example of this is multiplication of random variables. We might ask is $\mathbf{E}[X] \times \mathbf{E}[Y] = \mathbf{E}[X \times Y]$? It turns out it is true when X and Y are independent, but otherwise it is generally not true. To see that it is true for independent random variables we have (we assume x and y range over the values of X and Y respectively):

$$\begin{aligned}
\mathbf{E}[X] \times \mathbf{E}[Y] &= \left(\textstyle\sum_x x\mathbf{P}[\{X = x\}]\right)\left(\textstyle\sum_y y\mathbf{P}[\{Y = y\}]\right) \\
&= \textstyle\sum_x \sum_y (xy\mathbf{P}[\{X = x\}]\,\mathbf{P}[\{Y = y\}]) \\
&= \textstyle\sum_x \sum_y (xy\mathbf{P}[\{X = x\} \cap \{Y = y\}]) \quad \text{due to independence} \\
&= \mathbf{E}[X \times Y]
\end{aligned}$$

For example, the expected value of the product of the values on two (independent) dice is therefore $3.5 \times 3.5 = 12.25$.

Example 3.6. In a previous example , we analyzed the expectation on X, the sum of the two dice, by summing across all 36 elementary events. This was particularly messy for the biased dice. Using linearity of expectations, we need only calculate the expected value of each dice, and then add them. Since the dice are the same, we can in fact just multiply by two. For example for the biased case, assuming X_1 is the value of one dice:

$$\begin{aligned}
\mathbf{E}[X] &= 2\mathbf{E}[X_1] \\
&= 2 \times \textstyle\sum_{d \in \{1,2,3,4,5,6\}} d \times \frac{d}{21} \\
&= 2 \times \frac{1+4+9+16+25+36}{21} \\
&= 8\,\tfrac{2}{3}\,.
\end{aligned}$$

4 Conditional Expectation

Definition 3.1 (Conditional Expectation)**.** We define the conditional expectation of a random variable X for a given value y of Y as

$$\mathbf{E}[X \mid Y = y] = \sum_x x\mathbf{P}_{X|Y}(x \mid y).$$

Theorem 3.1 (Total Expectations Theorem)**.** The expectation of a random variable can be calculated by "averaging" over its conditional expectation given another random variable:

$$\mathbf{E}[X] = \sum_y \mathbf{P}_Y(y)\mathbf{E}[X \mid Y = y].$$

5 Variance and Standard Deviation

Definition 3.2 (Variance)**.** The variance of a random variable is defined as

$$\sigma^2 = \mathbf{E}\left[(X-\mu)^2\right].$$

Definition 3.3 (Standard Deviation)**.** The variance of a random variable is defined as the square-root of its variance.

$$\sigma = \sqrt{\mathbf{E}[(X-\mu)^2]}.$$

6 Markov's Inequality

Consider a non-negative random variable X. We can ask how much can X exceed its expected value. Because the expectation is taken by averaging X over all outcomes, and it cannot take on negative values, X cannot take on a much larger value with significant probability. If it did it would contribute too much to the sum. More generally X cannot be a multiple of β larger than its expectation with probability greater than $1/\beta$. This is because this part on its own would contribute more than $\beta\mathbf{E}[X] \times \frac{1}{\beta} = \mathbf{E}[X]$ to the expectation, which is a contradiction.

Theorem 3.2 (Markov's Inequality)**.** If X is a non-negative random variable, then

$$\mathbf{P}[X \geq \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}.$$

or equivalently (by substituting $\alpha = \beta\mathbf{E}[X]$)

$$\mathbf{P}[X \geq \beta\mathbf{E}[X]] \leq \frac{1}{\beta}.$$

Proof. Let Y be a random variable that lower bounds X as follows

$$Y = \begin{array}{ll} 0 & \text{if } X < \alpha \\ \alpha & \text{otherwise.} \end{array}$$

By definition, $Y \leq X$ and therefore $\mathbf{E}[Y] \leq \mathbf{E}[X]$. Furthermore,

$$\begin{aligned} \mathbf{E}[Y] &= \alpha\mathbf{P}[Y = \alpha] \\ &= \alpha\mathbf{P}[X \geq \alpha] \leq \mathbf{E}[X]. \end{aligned}$$

Thus it follows that $\mathbf{P}[X \geq \alpha] \leq \mathbf{E}[X]/\alpha$. □

Theorem 3.3 (Reverse-Markov Inequality)**.** If X is a random variable that is upper bounded by some constant u then for any $x < u$

$$\mathbf{P}[X \leq x] \leq \frac{\mathbf{E}[u - X]}{u - x}.$$

Proof. Define random variable Y as $Y = u - X$. We have

$$\mathbf{P}[X \leq x] = \mathbf{P}[Y \geq u - x].$$

We know that $Y \geq 0$ and thus we can apply it Markov's inequality to bound this quantity as follows.

$$\begin{aligned} \mathbf{P}[X \leq x] &= \mathbf{P}[Y \geq u - x] \\ &= \frac{\mathbf{E}[Y]}{u-x} \\ &= \frac{u - \mathbf{E}[X]}{u-x} \end{aligned}$$

□

7 Chebyshev's Inequality

Markov's inequality applies only to non-negative random variables. Chebyshev's inequality generalizes Markov's for all random variables. It basically states that a random variable strays away from its mean slowly as controlled by its variance.

Theorem 3.4 (Chebyshev's Inequality)**.** If X is a random variable with expectation μ and variance σ^2, then

$$\mathbf{P}[|X - \mu| \geq \gamma] \leq \frac{\sigma^2}{\gamma^2}.$$

or equivalently

$$\mathbf{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2},$$

Proof. Define the random variable $Y = (X - \mu)^2$ and note that Y is non-negative. Apply now Markov's inequality with $\alpha = \gamma^2$ to obtain

$$\mathbf{P}\left[(X - \mu)^2 \geq \gamma^2\right] \leq \frac{\mathbf{E}\left[(X - \mu)^2\right]}{\gamma^2}.$$

Note now that the event $(X - \mu)^2 \geq \gamma^2$ is the same as $|X - \mu| \geq \gamma$ and by the definition of variance of random variable , we have

$$\mathbf{P}[|X - \mu| \geq \gamma] \leq \frac{\sigma^2}{\gamma^2}.$$

Substituting $\gamma = k\sigma$ yields the second form. □

8 Chernoff Bounds

Chernoff bounds apply when a random variable is the sum of indicatator random variables, e.g., the binomial distribution (the sum of independent Bernoulli trials).

Theorem 3.5 (Chernoff Bound for Binomials)**.** If X is a Binomial random variable with expectation μ, then the following tail bounds hold for any $\delta > 0$

$$\begin{array}{rcll} \mathbf{P}\left[X < (1-\delta)\mu\right] & < & \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} & \text{Lower tail} \\ \mathbf{P}\left[X > (1+\delta)\mu\right] & < & \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} & \text{Upper tail} \end{array}$$

These bounds can be simplified for more constraints values of δ as follows.

$$\begin{array}{rcll} \mathbf{P}\left[X < (1-\delta)\mu\right] & < & e^{\frac{-\mu\delta^2}{2}} & \text{Simplified lower tail, when} 0 < \delta < 1 \\ \mathbf{P}\left[X > (1+\delta)\mu\right] & < & e^{\frac{-\mu\delta^2}{4}} & \text{Simplified upper tail, when } 0 < \delta < 2e-1 \\ \mathbf{P}\left[X > (1+\delta)\mu\right] & < & 2^{-\delta\mu} & \text{Simplified upper tail, when } \delta > 2e-1 \end{array}$$