Fall 2018

Probability Theory Review

Aleksandar Nikolov

1 Basic Notions: Sample Space, Events

A *probability space* $(\Omega, \mathbb{P})$ consists of a finite or countable set[1] Ω called the *sample space*, and the *probability function* $\mathbb{P} : \Omega \to \mathbb{R}$ such that for all $\omega \in \Omega$, $\mathbb{P}(\omega) \geq 0$ and $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$. We call an element $\omega \in \Omega$ a *sample point*, or *outcome*, or *simple event*. You should think of a sample space as modeling some random "experiment": Ω contains *all possible* outcomes of the experiment, and $\mathbb{P}(\omega)$ gives the probability that we are going to get outcome ω. Note that we never speak of probabilities except in relation to a sample space.

At this point we give a few examples:

1. Consider a random experiment in which we toss a single fair coin. The two possible outcomes are that the coin comes up heads (H) or tails (T), and each of these outcomes is equally likely. Then the probability space is $(\Omega, \mathbb{P})$, where $\Omega = \{H, T\}$ and $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$.

2. Consider a random experiment in which we toss a single coin, but the coin lands heads with probability $\frac{2}{3}$. Then, once again the sample space is $\Omega = \{H, T\}$ but the probability function is different: $\mathbb{P}(H) = \frac{2}{3}$, $\mathbb{P}(T) = \frac{1}{3}$.

3. Consider a random experiment in which we toss a fair coin three times, and each toss is independent of the others. The coin can come up heads all three times, or come up heads twice and then tails, etc. Then $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ and each sample point has probability $\frac{1}{8}$.

4. Consider a random experiment in which we roll two dice: one black and one white. The sample space is $\Omega = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\}$, i.e. all pairs of numbers (x, y) where x is the value we got from the black die, and y is the value we got from the white die (for example). If the dice are fair, then the probability of any sample point is $\frac{1}{36}$. Here the outcomes $(1, 3)$ and $(3, 1)$ are not the same, because in one case the black die landed on a 1, and in the other it landed on a 3.

5. Consider a random experiment in which we roll two white dice, and assume that we cannot distinguish which one is which. Then the sample space Ω is equal to the set of all multisets $\{x, y\}$ of two elements from $\{1, \ldots, 6\}$. Here "multiset" means that we can have $x = y$, but $\{x, y\}$ and $\{y, x\}$ are considered the same set. For example, in this case, the outcomes $\{1, 3\}$ and $\{3, 1\}$ are the same. The size of Ω is $\frac{6 \times 5}{2} + 6 = 21$, and, if the dice are fair, the probability of any outcome of the type $\{x, x\}$ is $\frac{1}{36}$, and the probability of any other outcome is $\frac{1}{18}$.

[1] A set Ω is countable if you can list all its elements in some order, i.e. $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$. Not every set is countable, for example the real numbers are not countable. Do not worry about this, we almost always deal with finite sample spaces.

When we have that $\mathbb{P}(\omega) = \mathbb{P}(\omega')$ for every two $\omega, \omega' \in \Omega$, we say that the probability is *uniform* over Ω.

An *event* A is a set of sample points, i.e. $A \subseteq \Omega$. We define the probability of an event A to be the sum of the probabilities of its elements, i.e.

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega).$$

This definition extends the probability function to all subsets of Ω. You should think of an event A as something that happens if the outcome ω of the experiment lies in the set of outcomes A. For instance, in the third example above the event that the first coin lands heads is given by the set of outcomes $A = \{HHH, HHT, HTH, HTT\}$. The probability of A is

$$\mathbb{P}(A) = \mathbb{P}(HHH) + \mathbb{P}(HHT) + \mathbb{P}(HTH) + \mathbb{P}(HTT) + \mathbb{P}(THH) = 4 \times \frac{1}{8} = \frac{1}{2}.$$

We define the *complement* of an event A to be $\bar{A} = \Omega \setminus A$. Often instead of $\bar{A}$, we write "not A". We have $\mathbb{P}(\text{not } A) = \mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.

It is easy to see (by drawing the Venn diagram) that for any two events A and B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Because probabilities are always non-negative, this implies the *union bound*:

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

From this inequality, a straightforward induction argument shows that, for any events $A_1, \ldots, A_k$, we have

$$\mathbb{P}(A_1 \cup \ldots \cup A_k) \leq \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_k).$$

We say that two events A and B are *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$. In the case of two disjoint events the union bound becomes an equality: if A and B are disjoint then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Note that sometimes we write $\mathbb{P}(A \text{ and } B)$ instead of $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A \text{ or } B)$ instead of $\mathbb{P}(A \cup B)$.

2 Conditional Probability and Independence

We often need to compute the probability of some event given that we already know that some other event holds. To capture this, we define the probability of an event A *conditional* on an event B by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

The conditional probability $\mathbb{P}(A \mid B)$ is defined only when $\mathbb{P}(B) > 0$.

It follows from the definition that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B).$$

Applying this formula inductively, we get the chain rule

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_k) = \mathbb{P}(A_1 \mid A_2 \cap A_3 \cap \cdots \cap A_k) \cdot \mathbb{P}(A_2 \mid A_3 \cap \cdots \cap A_k) \cdots \mathbb{P}(A_k).$$

Another easy to derive but important formula is Bayes' rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}.$$

Yet another useful formula is that

$$\mathbb{P}(A) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B) + \mathbb{P}(A \mid \bar{B}) \cdot \mathbb{P}(\bar{B})$$

when $0 < \mathbb{P}(B) < 1$.

We say that two events A and B are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

If $\mathbb{P}(B) > 0$, this implies $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

More generally, a set of events $A_1, A_2, \ldots, A_k$ are *mutually independent* if for all $I \subseteq \{1, \ldots, k\}$ we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Note that, for example, if $k = 3$, it is possible to have three events A_1, A_2, A_3 such that A_i and A_j are independent for all $i, j \in \{1, 2, 3\}$, but A_1, A_2, A_3 are *not* mutually independent. For example, take the sample space $\Omega = \{0, 1\}^3$, where

$$\mathbb{P}(000) = \frac{1}{4} \qquad \mathbb{P}(011) = \frac{1}{4},$$
$$\mathbb{P}(101) = \frac{1}{4} \qquad \mathbb{P}(110) = \frac{1}{4},$$

and all other elements of Ω have probability 0. This corresponds to picking the first two bits independently and uniformly from $\{0, 1\}$ and setting the third bit to equal 1 if exactly one of the first two bits equals 1 (i.e. the third bit is the exclusive-or of the first two). Define the event A_i to hold if and only if the i-th bit equals 1. Then A_1 and A_2 are independent by construction, and you can also verify that A_1 and A_3 are independent, and so are A_2 and A_3. However, whether A_1 and A_2 hold completely determines A_3, and, for example

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0,$$

whereas mutual independence would require that the probability is $\frac{1}{8}$. The takeaway message from this example is that it is important to verify *all* the equalities in the definition of mutual independence.

3 Random Variables and Expectation

A *random variable* defined on the probability space $(\Omega, \mathbb{P})$ is a function with domain Ω. Usually we will deal with real-valued random variables $X : \Omega \to \mathbb{R}$. Say X takes values in the set $\mathcal{X}$. Then the *probability distribution* of X is the function $p : \mathcal{X} \to [0, 1]$ defined by

$$p(x) = \mathbb{P}(X = x).$$

Above we used the usual convention that $\mathbb{P}(X = x)$ means $\mathbb{P}(\{\omega : X(\omega) = x\})$. Similarly we write $\mathbb{P}(X \leq x)$ for $\mathbb{P}(\{\omega : X(\omega) \leq x\})$, etc. When $p(x)$ is equal to $\frac{1}{|\mathcal{X}|}$ for all elements of $\mathcal{X}$, we say that X is *uniformly distributed*.

For example, let us go back to the experiment in which we toss two fair coins, one after the other. We can define a random variable X equal to the number of heads, i.e. $X(HH) = 2$, $X(HT) = 1$, $X(TH) = 1$, $X(TT) = 0$. The random variable X takes the values $0, 1, 2$, and has probability distribution

$$p(0) = \frac{1}{4} \quad p(1) = \frac{1}{2} \quad p(2) = \frac{1}{4}.$$

We say that random variables $X_1, \ldots, X_k$ defined on the same probability space are *mutually independent* if for any combination of values $x_1, \ldots, x_k$ we have

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \prod_{i=1}^{k} \mathbb{P}(X_i = x_i).$$

The *expected value* (or just expectation) of a random variable X is defined as

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) = \sum_{x \in \mathcal{X}} xp(x),$$

where $\mathcal{X}$ is the set of all values taken by X and p is the distribution of X. The second sum follows by rearranging the terms of the first sum. This second formula makes it clear that the expectation of a random variable is entirely determined by its probability distribution. In the example above, the expected value of the number of heads when we toss two fair coins is

$$0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) = 1.$$

You should think of expected value as the average of the values taken by the variable X, where the average is weighted by the probability distribution. When X is uniformly distributed, the expected value is equal to the (standard unweighted) average of the values taken by X. The expected value is <u>not</u> in general the the most likely value taken by X, and actually may not at all be a value taken by X: for example, it could be a rational number even when X takes only integer values.

A very important property of the expected value is *linearity of expectation*. If we have random variables $X_1, \ldots, X_k$ defined on the same probability space, then

$$\mathbb{E}[X_1 + \ldots + X_k] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_k].$$

This requires no special assumption whatsoever on the random variables: they do <u>not</u> need to be independent. Linearity of expectation is very useful in calculations. In the example above, we can define the random variable X_1 to equal 1 if the first coin is heads and 0 otherwise, and also define X_2 to equal 1 if the second coin is heads and 0 otherwise. Then the number of heads is $X = X_1 + X_2$, and we have

$$\mathbb{E}[X] = \mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = \frac{1}{2} + \frac{1}{2} = 1,$$

where we used

$$\mathbb{E}[X_1] = 0 \cdot \mathbb{P}(X_1 = 0) + 1 \cdot \mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 1) = \frac{1}{2},$$

and similarly for X_2.

When we do have *independent* random variables $X_1, \ldots, X_k$, we have that

$$\mathbb{E}\left[\prod_{i=1}^{k} X_i\right] = \prod_{i=1}^{k} \mathbb{E}[X_i].$$

It is convenient to define expectation also with respect to conditional probabilities. We write

$$\mathbb{E}[X \mid A] = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x \mid A)$$

for the *conditional expectation* of X with respect to the event A. Here, as before, $\mathcal{X}$ is the set of values taken by X.

We can also define the conditional expectation of a random variable X with respect to another random variable Y. However, in this case the expected value is itself a function. Let Y take values in the set $\mathcal{Y}$. Then $\mathbb{E}[X \mid Y]$ is the function defined on $\mathcal{Y}$ by

$$\mathbb{E}[X \mid Y](y) = \mathbb{E}[X \mid Y = y].$$

We can think of $\mathcal{Y}$ as a sample space with probability function $\mathbb{P}(y) = p(y)$, where p is the distribution of Y. Then $\mathbb{E}[X \mid Y]$ is a random variable defined on this probability space. Since it is a random variable, it also has an expected value, and a calculation shows that

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

This is the *law of total expectation.*

4 Some Common Probability Distributions

We could give a large list here, but instead we will list only two: binomial and geometric. Remember that we already defiend a **uniform** random variable. For your classwork you will not need to memorize names and parameters of probability distributions. Instead you will need to understand the basic properties and definitions in the previous sections and to be able to use them in calculations.

Binomial Distribution Suppose we toss a coin n times, where each toss is independent and the probability that a single toss lands on heads is p. Let X be the number of times the coin lands on heads. The probability distribution of X is called the *binomial distribution.* We have

$$p(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$
$$\mathbb{E}[X] = np.$$

The probability p is called the *success probability.*

Geometric Distribution Suppose we repeatedly toss a coin, where each toss is independent, and the probability that a single toss lands on heads is p. Let X be the random variable equal to the number of tosses up until and including the first heads. The probability distribution of X is called the *geometric distribution*. We have

$$p(k) = \mathbb{P}(X = k) = (1-p)^{k-1}p;$$
$$\mathbb{E}[X] = \frac{1}{p}.$$

The probability p is called the *success probability*.

We see one way to derive the formula for the expectation in the exercises below.

5 Exercises

Exercise 1. *In this experiment you choose* 3 *random bits, i.e. a string* $x \in \{0,1\}^3$*, from the following probability distribution:*

- *with probability* $1/3$ *all* 3 *bits are* 1*;*
- *with probability* $1/3$ *all* 3 *bits are* 0*;*
- *with probability* $1/3$ *each bit is chosen to be* 0 *or* 1 *with equal probability, independently from the others.*

Answer the following questions.

a. *Specify fully the sample space and the probability of each outcome.*

b. *What is the probability that all* 3 *bits are* 1*?*

Exercise 2. *In an experiment, we flip a fair coin ten consecutive times. Compute the probabilities of each of the following events:*

a. *the number of heads and the number of tails are equal;*

b. *there are more heads than tails;*

c. *the* i*-th flip and the* $(11-i)$*-th flip are the same for all* $i \in \{1, 2, \ldots, 5\}$*;*

d. *We flip at least four consecutive heads.*

Exercise 3. *Recall that a permutation of* $\{1, \ldots, n\}$ *is just the integers from* 1 *to* n *arranged in a sequence so each one of them appears exactly once.*

a. *How many permutations of* $\{1, \ldots, n\}$ *are there?*

b. *Suppose we have an experiment in which we draw a random permutation of* $\{1, \ldots, n\}$ *so that each permutation has equal probability. Write down the sample space and probability function when* $n = 3$*.*

c. *A permutation π has a fixed point at i if $\pi_i = i$. In the experiment from the previous subproblem, define the random variable X to equal the number of fixed points of a random permutation. What are the probability distribution and the expected value of X when $n = 3$.*

d. *What is the probability that $\pi_i = i$ for a random permutation π of $\{1, \ldots, n\}$?*

e. *Use linearity of expectation to compute the expected value of the random variable X from subproblem c. for any n.*

Exercise 4. *Suppose Alice and Bob repeatedly play a game in which either one of them wins or they draw. Assume that in each game, independently of the other games, Alice and Bob draw with probability p, Alice wins with probability $(1-p)/2$, and Bob wins with probability $(1-p)/2$. Assume, finally, that Alice and Bob repeatedly play the game until one of them wins.*

1. *What is the probability that Alice wins in the final game?*
2. *What is the probability that Bob wins in the final game?*
3. *What is the expected value of the number of games they play?*

Exercise 5. *We throw n balls into m bins, every ball is equal likely to fall into any bin, and the bins in which the different balls land are mutually independent.*

1. *What is the probability that the i-th bin is empty after all the balls are thrown?*
2. *What is the expected value of the number of empty bins?*

Exercise 6. *We want to test n people for a certain genetic condition, but each test is expensive, so we want to minimize the number of tests. Consider the following two strategies:*

i) *We test each person separately, and conduct a total of n tests.*

ii) *We divide the people into disjoint groups of k (if k does not divide n then there is one remaining group with fewer people). For any group, we pool the blood samples of all k people in the group and analyze them together. This only takes one test. If the test is negative, we are done for this group. If the test is positive for the genetic condition, then we test each person in the group, resulting in k additional tests, and a total of $k + 1$ tests.*

Assume that each person has the genetic condition with probability p, independently of all the others. Answer the following questions about strategy ii):

a *What is the probability that the test for a pooled sample of k people is positive for the genetic condition?*

b *What is the expected number X of tests performed under the second strategy? Assume that k divides N.*

c *Show that when $p \leq \frac{1}{2}$, there is a choice of k such that $\mathbb{E}[X] \leq 3n\sqrt{p} + 2np$. You can use the inequalities*

$$e^{-2p} \leq 1 - p \leq e^{-p},$$

valid for for $0 \leq p \leq \frac{1}{2}$. Give a formula to choose k given p.

Notice that this can be a lot fewer tests than the trivial strategy i) if the genetic condition is rare.

Exercise 7. *Let X be a random variable that takes values in $\mathbb{N} = \{1, 2, 3, \ldots\}$.*

a. *Prove that*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i).$$

(You can assume that the expectation exists, and that the infinite sum on the right hand side converges.)

b. *What is $\mathbb{P}(X \geq i)$ for a geometrically distributed random variable X with success probability p?*

c. *Use the previous two subquestions to show that $\mathbb{E}[X] = \frac{1}{p}$ for a geometric random variable X with success probability p.*

Exercise 8. *Show that if two events X and Y are independent, then X and $\bar{Y}$ are also independent.*