

# Review: Basics of Probability

An Algorithmist's Toolkit (CSCI 2952T)

D Ellis Herskowitz  
Brown University

The goal of this is to quickly remind you of the basics of probability, much of which we will later use in class. I'll assume that you've seen most of this before and won't provide proofs. If you're curious about how to prove something, though, feel free to ask!

For each of the problems you are asked to solve, you are free to assume any of the preceding facts in the document (other than the fact itself).

## 1 Discrete Probability Spaces

The basic object of study of (discrete) probability is discrete probability spaces, as defined below.<sup>1</sup>

**Definition 1.1 (Discrete Probability Space)** *A discrete probability space consists of*

1. *A (countable) set  $S$  called the sample space.*
2. *A probability function  $P : S \rightarrow [0, 1]$  satisfying  $\sum_{o \in S} P(o) = 1$ .*

A single element  $o \in S$  is referred to as an “outcome” of the sample space.

As an example, if we consider rolling a six-sided die, then a reasonable sample space would be  $S = \{1, 2, 3, 4, 5, 6\}$  and if the die is fair then  $P(o) = 1/6$  for every  $o \in S$ .

### 1.1 Events

Events allow us to formalize the idea of *something* happening in our probability space.

**Definition 1.2 (Event)** *An event is a subset  $E \subseteq S$ .*

We can extend the definition of probability from an outcome to an event as follows.

**Definition 1.3 (Event Probability)** *The probability of event  $E \subseteq S$  is*

$$P(E) := \sum_{o \in E} P(o).$$

Given two events  $E_1$  and  $E_2$ , the union of these events naturally corresponds to an outcome of  $E_1$  or  $E_2$  occurring and the intersection corresponds to both  $E_1$  and  $E_2$  occurring. For this reason, we will often write “ $E_1$  or  $E_2$ ” or “ $E_1$  and  $E_2$ ” to stand in for the events  $E_1 \cup E_2$  and  $E_1 \cap E_2$  respectively.

The union bound gives a quick and easy upper bound on the “or” of two events.

---

<sup>1</sup>Recall that a set is countable if it is finite or has cardinality equal to that of the natural numbers.

**Theorem 1.4 (Two Event Union Bound)** *Given any events  $E_1, E_2$  we have*

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2).$$

**Problem 1**

Prove the two event union bound.

More generally, it upper bounds the probability of a collection of events by the sum of their probabilities.

**Theorem 1.5 (Fully General Union Bound)** *Given any events  $E_1, E_2, \dots, E_k$  we have*

$$P(E_1 \cup E_2 \cup \dots \cup E_k) \leq \sum_{i=1}^k P(E_i).$$

**Problem 2**

Prove the fully general union bound using the two event union bound.

## 1.2 Conditioning and Independence

Conditioning on an event allows us to update our probabilities under the assumption that the event happened. In particular, we can define a new probability space assuming event  $B$ . The probabilities of this space are given as below.

**Definition 1.6 (Conditional Probability of an Outcome)** *Given discrete probability space  $(S, P)$  and event  $B \subseteq S$ , we define the conditional probability of outcome  $o$  as*

$$P_B(o) := \begin{cases} \frac{P(o)}{P(B)} & \text{if } o \in B \\ 0 & \text{if } o \notin B. \end{cases}$$

We will refer to  $(S, P_B)$  as the result of conditioning on  $B$ —as below, this result is itself just a new discrete probability space.

**Problem 3**

Given any discrete probability space  $(S, P)$  and any event  $B \subseteq S$ , we have that  $(S, P_B)$  is a discrete probability space.

We can extend the notion of conditional probabilities from outcomes to events as below.

**Definition 1.7 (Conditional Probability of an Event)** *Given discrete probability space  $(S, P)$  and event  $B \subseteq S$ , we define the conditional probability of event  $A$  as*

$$P(A | B) := \frac{P(A \cap B)}{P(B)}.$$

It is not too hard to see that the probability of  $A$  conditioned on  $B$  is just the probability of  $A$  in the new probability space  $(S, P_B)$ .

### Problem 4

Prove that for any events  $A$  and  $B$  we have  $P(A | B) = P_B(A)$ .

Informally, two events are independent if knowing one happened does not change the probability of the other.

**Definition 1.8 (Independent Events)** *Events  $A$  and  $B$  are independent iff*

$$P(A | B) = P(A)$$

or  $P(B) = 0$ .

Another common characterization of independent events is as below.

### Problem 5

$A$  and  $B$  are independent iff  $P(A \cap B) = P(A) \cdot P(B)$ .

## 2 Random Variables

Random variables allow us to associate numbers with outcomes.

**Definition 2.1 (Random Variable)** *A random variable is a function  $X : S \rightarrow \mathbb{R}$ .*

A very common event is one built from a random variable and a predicate as below.<sup>2</sup>

**Definition 2.2 (Events from Predicates)** *Given a predicate  $p$ , we let  $p(X)$  be the event*

$$\{o \in S : p(X(o))\}.$$

As an example, “ $X = i$ ” is used to stand for the event  $\{o \in S : X(o) = i\}$  and “ $X \leq i$ ” is used to stand for the event  $\{o \in S : X(o) \leq i\}$ .

### 2.1 Expectation

The expectation of a random variable is its (weighted) average.

**Definition 2.3 (Expectation)** *The expectation of random variable  $X$  is defined as*

$$\mathbb{E}[X] := \sum_{o \in S} X(o) \cdot P(o).$$

A useful alternate definition of the expectation is as below.

### Problem 6

Given any random variable  $X$  in a discrete probability space, we have  $\mathbb{E}[X] = \sum_i i \cdot P(X = i)$ .

Linearity of expectation is one of the most useful algebraic properties of expectation.

<sup>2</sup>Recall that a predicate is a function which takes in values and outputs a proposition—namely, a statement that is either true or false.

**Theorem 2.4 (Linearity of Expectation)** For any random variables  $X$  and  $Y$  and reals  $a, b \in \mathbb{R}$  we have

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

In general, expectation does not distribute over multiplication as it does over addition—however, it does if the random variables are independent.

**Theorem 2.5** For any independent random variables  $X$  and  $Y$  we have

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

Just as conditioning on an event gives us new probabilities for outcomes and events, conditioning on an event gives us new expectations as below. In fact, these expectations are just expectations in the probability space resulting from conditioning.

**Definition 2.6 (Conditional Expectation)** The expectation of random variable  $X$  conditioned on event  $B$  is

$$\mathbb{E}[X | B] := \sum_{o \in B} X(o) \cdot P_B(o).$$

It is not too hard to see that, just as there is an alternate characterization of expectation, there is an alternate characterization of conditional expectation

**Theorem 2.7** Given any random variable  $X$  and event  $B$  we have

$$\mathbb{E}[X | B] := \sum_i i \cdot P(X = i | B).$$

## 2.2 Variance and Covariance

The variance tells us how far, on average, a random variable is from its expectation.

**Definition 2.8 (Variance)** The variance of random variable  $X$  is

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

A useful alternate characterization of variance is as below.

**Theorem 2.9** For any random variable  $X$  we have

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

### Problem 7

Show Theorem 2.9.

The covariance tells us how far, on average, two random variables are from each other.

**Definition 2.10 (Covariance)** The covariance of two random variables  $X$  and  $Y$  is

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Variance is not linear like expectation. However, one can pull out reals up to a squaring.

**Theorem 2.11** For any random variables  $X$  and real  $c \in \mathbb{R}$  we have

$$\text{Var}(cX) = c^2 \text{Var}(X).$$

Likewise, it does distribute across addition up to a covariance term.

**Theorem 2.12** For any random variables  $X$  and  $Y$  we have

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y).$$

Furthermore, if the random variables are independent then covariance is 0 and so it does distribute across addition in this case.

**Theorem 2.13** For any independent random variables  $X$  and  $Y$  we have

$$\text{Cov}(X, Y) = 0.$$

#### Problem 8

Show Theorem 2.13.

**Theorem 2.14** For any independent random variables  $X$  and  $Y$  we have

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

## 2.3 PMFs, CDFs and Distributions

The probability mass function (PMF) of a discrete random variable gives the probability that it takes on a particular value.

**Definition 2.15 (Probability Mass Function (PMF))** Given a discrete random variable  $X$ , the PMF of  $X$  is the function  $P_X$  which on input  $i \in \mathbb{R}$  gives

$$P_X(i) = P(X = i).$$

The cumulative distribution function (CDF) gives the probability a discrete random variable takes on at most some value.

**Definition 2.16 (Cumulative Distribution Function (CDF))** Given random variable  $X$ , the cumulative distribution function (CDF) of  $X$  is the function  $C_X$  which on  $i \in \mathbb{R}$  gives

$$C_X(i) := \sum_{a \leq i} P_X(a).$$

The word “distribution” is often used to either describe the PMF or the CDF of a random variable.

It is worth noting that the support of a random variable and its PMF induce a new discrete probability space, as below.

**Definition 2.17 (Random Variable Support)** Given discrete probability space  $(S, P)$  and a random variable  $X$ , the support of  $X$  is all values the random variable takes on, namely

$$\text{supp}(X) := \{X(o) : o \in S\}.$$

The following gives the new discrete probability space.

**Theorem 2.18** Given any discrete probability space  $(S, P)$  and random variable  $X$ , it holds that  $(\text{supp}(X), P_X)$  is a discrete probability space.

### 3 Continuous RVs

For the most part, we will be working with discrete probability spaces. However, there are a few places where we will use continuous probability spaces.

In rough terms, a continuous probability space is one with a sample space that is uncountable—that is, cardinality on the order of  $|\mathbb{R}|$ . These continuous spaces replace the probability function  $P$  with a “probability measure”. Furthermore, unlike in the discrete case where we talk about the probability of outcomes and events, here we generally only talk about the probability of an event as specified by the probability measure. Furthermore, the valid events of a continuous spaces are not arbitrary subsets of  $2^\Omega$ , but rather a special collection of subsets of  $2^\Omega$  which form what is called a  $\sigma$ -algebra.

Random variables likewise get extended. The distribution of random variable  $X$ , namely  $P_X$ , has as its analogue the probability density function  $\phi_X$ . Likewise, the CDF and expectation generalize but become integrals rather than sums. Like in the discrete case where the PMF sums to 1, the PDF must always integrate to 1.

We won’t need the details of this formalism, but the following table compares the two cases.

Discrete Probability Space	Continuous Probability Space
Sample space $\Omega$ is countable	Sample space $\Omega$ is uncountable
Probability function $P$	Probability Measure $P$
An event is any subset of $2^\Omega$	An event is an element of a $\sigma$ -algebra $\subseteq 2^\Omega$
PMF $P_X$	PDF $\phi_X$
CDF $C_X(i) = \sum_{a \leq i} P_X(a)$	CDF $C_X(i) = \int_{-\infty}^i \phi_X(a) da$
$\mathbb{E}[X] = \sum_i i \cdot P(X = i)$	$\mathbb{E}[X] = \int_{-\infty}^{\infty} i \cdot \phi(i) di$
$\sum_i P_X(i) = 1$	$\int_{-\infty}^{\infty} \phi(i) di = 1$

#### 3.1 Gaussians

The main continuous random variable we will make use of is the Gaussian.

**Definition 3.1 (Gaussian PDF)** A continuous random variable  $X$  is said to be Gaussian if its PDF on  $i \in \mathbb{R}$  is

$$\phi_X(i) = \frac{1}{\sqrt{2\pi}} e^{-i^2/2}.$$

It is a surprisingly interesting argument to show that the PDF of a Gaussian indeed integrates to 1.

**Theorem 3.2 (Gaussian PDF is Actually a PDF)** *Letting  $\phi$  be the Gaussian PDF as in Definition 3.1, we have*

$$\int_{-\infty}^{\infty} \phi(x) dx = 1.$$

Since  $\phi_X$  is symmetric around the  $y$ -axis, it is easy to check that if  $X$  is Gaussian then  $\mathbb{E}[X] = 0$ . Likewise, with a bit more work one can show that  $\text{Var}(X) = 1$  for a Gaussian random variable. As such, if  $X$  is a random variable with a Gaussian PDF, then we write  $X \sim N(0, 1)$ . Such an  $X$  is referred to as a standard Gaussian.

A non-standard Gaussian is as follows.

**Definition 3.3 (Non-Standard Gaussian)**  *$Y$  is a non-standard Gaussian if it is of the form  $\mu + \sigma X$  where  $X \sim N(0, 1)$  is a standard Gaussian.*

It is easy to check that such a Gaussian has expectation and variance  $\mu$  and  $\sigma^2$  respectively, hence for such a Gaussian we write  $Y \sim N(\mu, \sigma^2)$ .

We will make use of 2 key properties of Gaussians. The first of these states that the direction that a vector whose entries are iid Gaussians is uniformly random. Equivalently, any two vectors with the same norm are equally likely when coordinates are drawn as iid Gaussians (you should convince yourself this is equivalent).<sup>3</sup>

**Theorem 3.4 (Rotational Symmetry of Gaussians)** *Suppose that  $u = (u_1, u_2, \dots, u_k)$  is a random vector in  $\mathbb{R}^k$  where each  $u_i \sim N(0, 1)$  independently. Then for any  $x, y \in \mathbb{R}^k$  with  $\|x\| = \|y\|$  we have*

$$P(u = x) = P(u = y).$$

Second, if we sum together two scaled Gaussians, the result is a Gaussian.

**Theorem 3.5 (Sum of Gaussians is Gaussian)** *If  $X \sim N(0, 1)$  and  $Y \sim N(0, 1)$  and  $X$  and  $Y$  are independent then, for any reals  $a, b \in \mathbb{R}$ , we have  $aX + bY \sim N(0, a^2 + b^2)$*

It's a fun (but tricky) exercise to show Theorem 3.5 using Theorem 3.4.

---

<sup>3</sup>This is actually somewhat incorrect as stated. In particular, for a continuous random vector the probability of taking any exact value is zero, so expressions like  $P(u = x)$  should be interpreted informally.