EECS 70 Discrete Mathematics and Probability Theory
Spring 2021 Shyam Parekh and Satish Rao

# Note 14

## Conditional Probability, Independence, and Combinations of Events

One of the key properties of coin flips is independence: if you flip a fair coin ten times and get ten $T$'s, this does not make it more likely that the next coin flip will be $H$'s. It still has exactly 50% chance of being $H$.

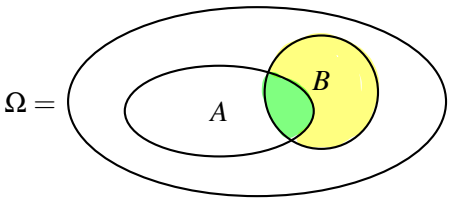
By contrast, suppose while dealing cards, the first ten cards are all red (hearts or diamonds). What is the chance that the next card is red? We started with exactly 26 red cards and 26 black cards. But after dealing the first ten cards, we know that the deck has 16 red cards and 26 black cards. So the chance that the next card is red is $\frac{16}{42}$. So unlike the case of coin flips, now the chance of drawing a red card is no longer independent of the previous card that was dealt. This is the phenomenon we will explore in this note on conditional probability.

## 1 Conditional Probability

Let's consider an example with a smaller sample space. Suppose we toss $m = 4$ labeled balls into $n = 3$ labeled bins; this is a uniform sample space with $3^4 = 81$ sample points. From the previous note, we already know that the probability the first bin is empty is $(1-\frac{1}{3})^4 = (\frac{2}{3})^4 = \frac{16}{81}$. What is the probability of this event *given* that the second bin is empty? Let $A$ denote the event that the first bin is empty, and $B$ the event that the second bin is empty. In the language of conditional probability, we wish to compute the probability $\mathbb{P}[A|B]$, which we read as "the conditional probability of $A$ given $B$."

↳ Probability of A, given B

How should we compute $\mathbb{P}[A|B]$? Since event $B$ is guaranteed to happen, we need to look not at the whole sample space $\Omega$, but at the smaller sample space consisting only of the sample points in $B$. In terms of the picture below, we are no longer looking at the large oval, but only the oval labeled $B$:




What should be the probability of each sample point $\omega \in B$ given that the event $B$ occurs? If they all simply inherited their probabilities from $\Omega$, then the sum of these probabilities would be $\sum_{\omega \in B} \mathbb{P}[\omega] = \mathbb{P}[B]$, which in general is less than 1. So, to get the correct normalization, we need to *scale* the probability of each sample point by $\frac{1}{\mathbb{P}[B]}$. That is, for each sample point $\omega \in B$, the new probability becomes

$$\mathbb{P}[\omega|B] = \frac{\mathbb{P}[\omega]}{\mathbb{P}[B]}.$$

Note: $\omega \in B$

Now it is clear how to compute $\mathbb{P}[A|B]$: namely, we just sum up these scaled probabilities over all sample

points that lie in both $A$ and $B$:

$$\mathbb{P}[A|B] = \sum_{\omega \in A \cap B} \mathbb{P}[\omega|B] = \sum_{\omega \in A \cap B} \frac{\mathbb{P}[\omega]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

**Definition 14.1** (Conditional Probability)**.** *For events $A, B \subseteq \Omega$ in the same probability space such that $\mathbb{P}[B] > 0$, the conditional probability of $A$ given $B$ is*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Returning to our example of balls and bins, to compute $\mathbb{P}[A|B]$ we need to figure out $\mathbb{P}[A \cap B]$. But $A \cap B$ is the event that both the first two bins are empty, i.e., all four balls fall in the third bin. So $\mathbb{P}[A \cap B] = \frac{1}{81}$ (why?). Therefore,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{1/81}{16/81} = \frac{1}{16}.$$

Not surprisingly, $\mathbb{P}[A|B] = \frac{1}{16} = 0.0625$ is quite a bit less than $\mathbb{P}[A] = \frac{16}{81} \approx 0.1975$; knowing that bin 2 is empty makes it significantly less likely that bin 1 will be empty.

### Example: Card Dealing

Let's apply the ideas discussed above to compute the probability that, when dealing 2 cards and the first card is known to be an ace, the second card is also an ace. Let $B$ be the event that the first card is an ace, and let $A$ be the event that the second card is an ace.

To compute $\mathbb{P}[A|B]$, we need to figure out $\mathbb{P}[A \cap B]$. This is the probability that both cards are aces. Note that there are $52 \cdot 51$ sample points in the sample space, since each sample point is a sequence of two cards. A sample point is in $A \cap B$ if both cards are aces. This can happen in $4 \cdot 3 = 12$ ways.

Since each sample point is equally likely, $\mathbb{P}[A \cap B] = \frac{12}{52 \cdot 51}$, while $\mathbb{P}[B]$, the probability of drawing an ace in the first trial, is $\frac{4}{52}$. Therefore,

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{3}{51},$$

which is less than $\mathbb{P}[A] = \frac{4}{52} = \frac{1}{13}$ (see Exercise 1). Hence, if the first card is an ace, it makes it less likely that the second card is also an ace.

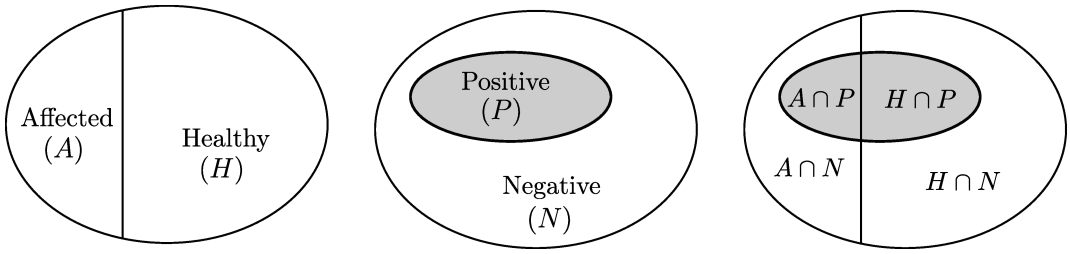
## 2 Bayesian Inference

Now that we have introduced the notion of conditional probability, we can see how it is used in real world settings. Conditional probability is at the heart of a subject called *Bayesian inference*, used extensively in fields such as machine learning, communications and signal processing. Bayesian inference is a way to *update knowledge* after making an observation. For example, we may have an estimate of the probability of a given event $A$. After event $B$ occurs, we can update this estimate to $\mathbb{P}[A|B]$. In this interpretation, $\mathbb{P}[A]$ can be thought of as a *prior* probability: our assessment of the likelihood of an event of interest, $A$, *before* making an observation. It reflects our prior knowledge. $\mathbb{P}[A|B]$ can be interpreted as the *posterior* probability of $A$ after the observation. It reflects our updated knowledge.

Here is an example of where we can apply such a technique. A pharmaceutical company is marketing a new test for a certain medical disorder. According to clinical trials, the test has the following properties:

1. When applied to an affected person, the test comes up positive in 90% of cases, and negative in 10% (these are called "false negatives").
2. When applied to a healthy person, the test comes up negative in 80% of cases, and positive in 20% (these are called "false positives").

Suppose that the incidence of the disorder in the US population is 5%; this is our prior knowledge. When a random person is tested and the test comes up positive, how can we update the probability that the person has the disorder? (Note that this is presumably *not* the same as the simple probability that a random person in the population has the disorder, which is just $\frac{1}{20}$.) The implicit probability space here is the entire US population with equal probabilities.




The sample space here consists of all people in the US — denote their number by $N$ (so $N \approx 325$ million). Let $A$ be the event that a person chosen at random is affected, and $B$ be the event that a person chosen at random tests positive. Now we can rewrite the information above as:

- $\mathbb{P}[A] = 0.05$, (5% of the U.S. population is affected)
- $\mathbb{P}[B|A] = 0.9$, (90% of the affected people test positive)
- $\mathbb{P}[B|\overline{A}] = 0.2$, (20% of healthy people test positive)

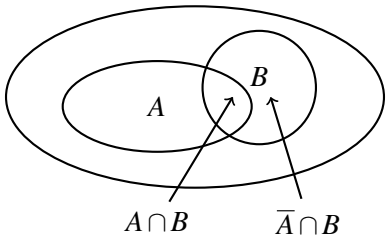
We want to calculate $\mathbb{P}[A|B]$. We can proceed as follows:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}. \quad \text{Bayes Rule} \tag{1}$$

We obtained the second equality above by applying the definition of conditional probability:

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]}.$$

To evaluate (1), we need to compute $\mathbb{P}[B]$. This is the probability that a random person tests positive. To compute this, we can sum two values: the probability $\mathbb{P}[\overline{A} \cap B]$ that a healthy person tests positive and the probability $\mathbb{P}[A \cap B]$ that an affected person tests positive. We can sum because the events $\overline{A} \cap B$ and $A \cap B$ do not intersect:



By again applying the definition of conditional probability, we have:

$$\begin{aligned}\mathbb{P}[B] &= \mathbb{P}[A \cap B] + \mathbb{P}[\overline{A} \cap B] = \mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|\overline{A}]\mathbb{P}[\overline{A}] \\ &= \mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|\overline{A}](1 - \mathbb{P}[A]).\end{aligned} \tag{2}$$

Total Probability Rule

Combining (1) and (2), we have expressed $\mathbb{P}[A|B]$ in terms of $\mathbb{P}[A], \mathbb{P}[B|A]$ and $\mathbb{P}[B|\overline{A}]$:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|\overline{A}](1 - \mathbb{P}[A])}. \tag{3}$$

By plugging in the values written above, we obtain $\mathbb{P}[A|B] = \frac{9}{47} \approx 0.19$.

Equation (3) is useful for many inference problems. We are given $\mathbb{P}[A]$, which is the (unconditional) probability that the event of interest, $A$, happens. We are given $\mathbb{P}[B|A]$ and $\mathbb{P}[B|\overline{A}]$, which quantify how noisy the observation is. (If $\mathbb{P}[B|A] = 1$ and $\mathbb{P}[B|\overline{A}] = 0$, for example, the observation is completely noiseless.) Now we want to calculate $\mathbb{P}[A|B]$, the probability that the event of interest happens given we made the observation. Equation (3) allows us to do just that.

Of course, (1), (2) and (3) are derived from the basic axioms of probability and the definition of conditional probability, and are therefore true with or without the above Bayesian inference interpretation. However, this interpretation is very useful when we apply probability theory to study inference problems.

# 3 Bayes' Rule and Total Probability Rule

Equations (1) and (2) are very useful in their own right. The former is called **Bayes' Rule** and the latter is called the **Total Probability Rule**. Bayes' Rule is useful when one wants to calculate $\mathbb{P}[A|B]$ but one is given $\mathbb{P}[B|A]$ instead, i.e., it allows us to "flip" things around.

The Total Probability Rule is an application of the strategy of "dividing into cases." There are two possibilities: either an event $A$ happens or $A$ does not happen. If $A$ happens, the probability that $B$ happens is $\mathbb{P}[B|A]$. If $A$ does not happen, the probability that $B$ happens is $\mathbb{P}[B|\overline{A}]$. If we know or can easily calculate these two probabilities and also $\mathbb{P}[A]$, then the Total Probability Rule yields the probability of event $B$.

## 3.1 Examples

**Tennis Match**

You are about to play a tennis match against a randomly chosen opponent and you wish to calculate your probability of winning. You know your opponent will be one of two people, $X$ or $Y$. If person $X$ is chosen, you will win with probability 0.7. If person $Y$ is chosen, you will win with probability 0.3. Your opponent is chosen by flipping a biased coin such that the probability of choosing $X$ is 0.6.

Let's first determine which events we are interested in. Let $A$ be the event that you win. Let $B_X$ be the event that person $X$ is chosen, and let $B_Y$ be the event that person $Y$ is chosen. We wish to calculate $\mathbb{P}[A]$. Here is what we know so far:

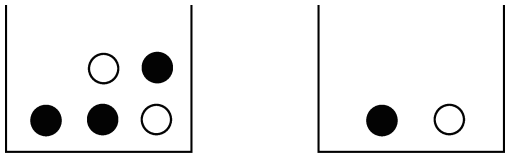
- $\mathbb{P}[A|B_X] = 0.7$, (if person $X$ is chosen, you win with probability 0.7)
- $\mathbb{P}[A|B_Y] = 0.3$, (if person $Y$ is chosen, you win with probability 0.3)

- $\mathbb{P}[B_X] = 0.6$, (person $X$ is chosen with probability 0.6)
- $\mathbb{P}[B_Y] = 0.4$, (person $Y$ is chosen with probability 0.4)

By using the Total Probability Rule, we have $\mathbb{P}[A] = \mathbb{P}[A|B_X]\mathbb{P}[B_X] + \mathbb{P}[A|B_Y]\mathbb{P}[B_Y]$, and plugging in the known values gives $\mathbb{P}[A] = (0.7 \times 0.6) + (0.3 \times 0.4) = 0.54$.

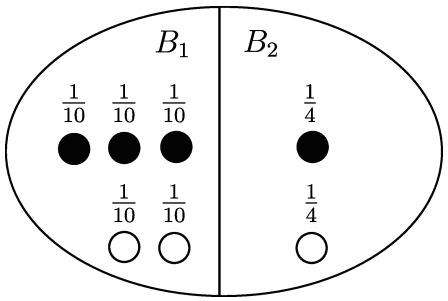
**Balls and Bins**

Imagine we have the following two bins each containing some number of black and white balls:




Suppose one of the two bins is chosen with equal probability and a ball is drawn from the chosen bin uniformly at random. What is the probability that we picked Bin 1 given that a white ball was drawn, i.e., $\mathbb{P}[\text{Bin 1} \mid \bigcirc]$?

Is the answer $\frac{2}{3}$, since we know that there are a total of three white balls, two of which are in Bin 1? This reasoning is *incorrect*. Instead, what we should do is appropriately scale each sample point as the following picture shows:




This diagram shows that the sample space $\Omega$ consists of the outcomes in event $B_1$ (corresponding to Bin 1) and event $B_2$ (corresponding to Bin 2), i.e., $\Omega = B_1 \cup B_2$. We can use the definition of conditional probability to see that

$$\mathbb{P}[\text{Bin1} \mid \bigcirc] = \frac{\frac{1}{10} + \frac{1}{10}}{\frac{1}{10} + \frac{1}{10} + \frac{1}{4}} = \frac{\frac{2}{10}}{\frac{9}{20}} = \frac{4}{9}.$$

Let us try to achieve this probability using Bayes' Rule. To apply Bayes' Rule, we need to compute $\mathbb{P}[\bigcirc|\text{Bin 1}]$, $\mathbb{P}[\text{Bin 1}]$, and $\mathbb{P}[\bigcirc]$. Here, $\mathbb{P}[\bigcirc|\text{Bin 1}]$ is the chance that we pick a white ball given that we picked Bin 1, which is $\frac{2}{5}$. $\mathbb{P}[\text{Bin 1}] = \frac{1}{2}$, as given in the description of the problem. Finally, $\mathbb{P}[\bigcirc]$ can be computed using the Total Probability Rule:

$$\mathbb{P}[\bigcirc] = \mathbb{P}[\bigcirc|\text{Bin 1}] \times \mathbb{P}[\text{Bin 1}] + \mathbb{P}[\bigcirc|\text{Bin 2}] \times \mathbb{P}[\text{Bin 2}] = \frac{2}{5} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{9}{20}.$$

Observe that we can apply the Total Probability Rule here because $\mathbb{P}[\text{Bin 1}]$ is the complement of $\mathbb{P}[\text{Bin 2}]$. Finally, upon plugging the above values into Bayes' Rule, we obtain the probability that we picked Bin 1

given that we picked a white ball:

$$\mathbb{P}[\text{Bin 1} \mid \bigcirc] = \frac{\frac{2}{5} \times \frac{1}{2}}{\frac{9}{20}} = \frac{\frac{2}{10}}{\frac{9}{20}} = \frac{4}{9}.$$

All we have done above is to combine Bayes' Rule and the Total Probability Rule; this is also how we obtained (3). Equivalently, we could have plugged in the appropriate values to (3).

## 3.2 Maximum Likehood or Maximum A Postieri (MAP)

In inference, a *maximum likelihood estimate (MLE)* predicts a parameter which maximizes the probability of the data, the *maximum a postieri (MAP) estimate* predicts the parameter which maximizes the conditional probability of the parameter given the data. They are equivalent when every value of the parameter has equal prior probability.

For the balls in bin example above, the data that was observed was the white ball and the parameter we estimated is which bin it came from. In this situation, we have $\mathbb{P}[\bigcirc \mid \text{Bin 1}] = 2/5$ and that $\mathbb{P}[\bigcirc \mid \text{Bin 2}] = 1/2$. Thus, the MLE is Bin 2. In this case, the prior probabilities of each bin is uniform so the MAP estimate is also Bin 2.

If one chose the bins in a non-uniform fashion, in particular, if $\mathbb{P}[\text{Bin 1}] > 5/9$, the MAP estimate of the parameter would be Bin 1. The greater prior probability on Bin 1 compensates for the fact that Bin 1 is less likely to produce a white ball. The threshold of $5/9$ comes from comparing $\mathbb{P}[\text{Bin 1} \cap \bigcirc]$ to $\mathbb{P}[\text{Bin 2} \cap \bigcirc]$. That is, for Bin 1 to be the MAP estimate, we have that

$$\mathbb{P}[\text{Bin 1} \cap \bigcirc] = \mathbb{P}[\text{Bin 1} \mid \bigcirc] \times \mathbb{P}[\text{Bin 1}] = \frac{2}{5} \times \mathbb{P}[\text{Bin 1}] > \frac{1}{2} \times (1 - \mathbb{P}[\text{Bin 1}]) = \mathbb{P}[\text{Bin 2} \mid \bigcirc] \times \mathbb{P}[\text{Bin 2}] = Pr[\text{Bin 2} \cap \bigcirc]$$

Simplifying the inequality in the middle above yields that when $\mathbb{P}[\text{Bin 1}] > \frac{5}{9}$ the MAP estimate is Bin 1. Below that the MAP estimate and the MLE will be Bin 2.

In general, if one has prior probabilities one should use the MAP estimator. This is what we do in this class. In practice, one does use the maximum likelihood estimate as that is easier to conclude from data at times and the assumption of uniform priors is made implicitly.

## 3.3 Generalization

We now consider Bayes' Rule and the Total Probability Rule in a more general context. First, we define a *partition* of an event as follows.

**Definition 14.2** (Partition of an event)**.** *We say that an event $A$ is partitioned into $n$ events $A_1, \ldots, A_n$ if*

1. $A = A_1 \cup A_2 \cup \cdots \cup A_n$, *and*
2. $A_i \cap A_j = \varnothing$ *for all $i \neq j$ (i.e., $A_1, \ldots, A_n$ are mutually exclusive).*

*In other words, each outcome in $A$ belongs to exactly one of the events $A_1, \ldots, A_n$.*

Now, let $A_1, \ldots, A_n$ be a partition of the sample space $\Omega$. Then, the **Total Probability Rule** for any event $B$ is

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B \cap A_i] = \sum_{i=1}^{n} \mathbb{P}[B|A_i]\,\mathbb{P}[A_i], \tag{4}$$

$\mathbb{P}[B] = \mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|\bar{A}]\mathbb{P}[A]$

while **Bayes' Rule**, assuming $\mathbb{P}[B] \neq 0$, is given by

$$\mathbb{P}[A_i|B] = \frac{\mathbb{P}[B|A_i]\,\mathbb{P}[A_i]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A_i]\,\mathbb{P}[A_i]}{\sum_{j=1}^{n}\mathbb{P}[B|A_j]\,\mathbb{P}[A_j]}, \tag{5}$$

where the second equality follows from the Total Probability Rule.

# 4 Combinations of Events

In most applications of probability in Computer Science, we are interested in things like $\mathbb{P}[\bigcup_{i=1}^{n} A_i]$ and $\mathbb{P}[\bigcap_{i=1}^{n} A_i]$, where the $A_i$ are simple events (i.e., we know or can easily compute $\mathbb{P}[A_i]$). The intersection $\bigcap_i A_i$ corresponds to the logical AND of the events $A_i$, while the union $\bigcup_i A_i$ corresponds to their logical OR. As an example, if $A_i$ denotes the event that a failure of type $i$ happens in a certain system, then $\bigcup_i A_i$ is the event that the system fails.

In general, computing the probabilities of such combinations can be very difficult. In this section, we discuss some situations where it can be done. Let's start with independent events, for which intersections are quite simple to compute.

## 4.1 Independent Events

**Definition 14.3** (Independence)**.** *Two events $A, B$ in the same probability space are said to be independent if $\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B]$.*

The intuition behind this definition is the following. Suppose that $\mathbb{P}[B] > 0$. Then we have

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A] \times \mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A].$$

Thus independence has the natural meaning that "the probability of $A$ is not affected by whether or not $B$ occurs." (By a symmetrical argument, we also have $\mathbb{P}[B|A] = \mathbb{P}[B]$ provided $\mathbb{P}[A] > 0$.) For events $A, B$ such that $\mathbb{P}[B] > 0$, the condition $\mathbb{P}[A|B] = \mathbb{P}[A]$ is actually *equivalent* to the definition of independence.

Several of our previously mentioned random experiments consist of independent events. For example, if we flip a coin twice, the event of obtaining heads in the first trial is independent of the event of obtaining heads in the second trial. The same applies for two rolls of a die; the outcomes of each trial are independent.

The above definition generalizes to any finite set of events:

**Definition 14.4** (Mutual independence)**.** *Events $A_1, \ldots, A_n$ are said to be mutually independent if for **every** subset $I \subseteq \{1, \ldots, n\}$ with size $|I| \geq 2$,*

$$\mathbb{P}[\cap_{i \in I} A_i] = \prod_{i \in I} \mathbb{P}[A_i]. \tag{6}$$

An equivalent definition of mutual independence is as follows.

**Definition 14.5** (Mutual independence)**.** *Events $A_1, \ldots, A_n$ are said to be mutually independent if for all $B_i \in \{A_i, \overline{A_i}\}, i = 1, \ldots, n$,*

$$\mathbb{P}[B_1 \cap \cdots \cap B_n] = \prod_{i=1}^{n} \mathbb{P}[B_i]. \tag{7}$$

**Remarks.**

1. In Definition 14.4, (6) needs to hold for *every* subset $I$ of $\{1,\ldots,n\}$ with size $|I| \geq 2$. The cases of $|I| = 0$ and $|I| = 1$ are omitted, as they impose no constraints.

2. Note that (6) imposes $2^n - n - 1$ constraints on the probability distribution, while (7) defines $2^n$ constraints. It turns out that exactly $n+1$ constraints implied by (7) are actually redundant.

For mutually independent events $A_1,\ldots,A_n$, it is not hard to check from the definition of conditional probability that, for any $1 \leq i \leq n$ and any subset $I \subseteq \{1,\ldots,n\} \setminus \{i\}$, we have

$$\mathbb{P}[A_i | \bigcap_{j\in I} A_j] = \mathbb{P}[A_i].$$

Note that the independence of every pair of events (so-called *pairwise independence*) does *not* necessarily imply mutual independence. As illustrated in the following example, it is possible to construct three events $A, B, C$ such that each *pair* is independent but the triple $A, B, C$ is *not* mutually independent.

**Example: Pairwise Independent but Not Mutually Independent**

Suppose you toss a fair coin twice and let $A$ be the event that the first flip is $H$ and $B$ be the event that the second flip is $H$. Now let $C$ be the event that both flips are the same (i.e., both $H$'s or both $T$'s). Of course $A$ and $B$ are independent. What is more interesting is that so are $A$ and $C$: given that the first toss came up $H$, the chance of the second flip being the same as the first is still $1/2$. Another way of saying this is that $\mathbb{P}[A \cap C] = \mathbb{P}[A]\mathbb{P}[C] = 1/4$ since $A \cap C$ is the event that the first flip is $H$ and the second is also $H$. By the same reasoning $B$ and $C$ are also independent. On the other hand, $A$, $B$ and $C$ are not mutually independent. For example, if we are given that $A$ and $B$ occurred, then the probability that $C$ occurs is 1. So, even though $A$, $B$ and $C$ are not mutually independent, every pair of them are independent. In other words, $A$, $B$ and $C$ are pairwise independent but not mutually independent.

## 4.2 Correlated Events

If a pair of events are not independent, then we can refer to them as correlated. In particular, events $A$ and $B$ are *positively correlated* (respectively *negatively correlated*) if $Pr[A|B] > Pr[A]$ (respectively $Pr[A|B] < Pr[B]$.)

Of course, this definition only makes sense if when $Pr[A|B] > Pr[A]$ we also have $Pr[B|A] > Pr[B]$. This can be shown as follows:

$$\begin{aligned} Pr[B \mid A] &= \frac{Pr[A \mid B]Pr[B]}{Pr[A]} && (8)\\ &= \frac{Pr[A \mid B]}{Pr[A]} Pr[B] && (9)\\ &> Pr[B] \end{aligned}$$

The last inequality follows from $Pr[A|B] > Pr[A]$.

## 4.3 Intersections of Events

Computing the probability of an intersection of mutually independent events is easy; it follows from the definition. We simply multiply the probabilities of each event. How do we compute the probability of an

intersection for events that are not mutually independent? From the definition of conditional probability, we immediately have the following Product Rule (sometimes also called the chain rule) for computing the probability of an intersection of events.

**Theorem 14.1** (Product Rule). *For any events $A, B$, we have*

Chain Rule

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B|A]. \qquad \mathbb{P}[A \cap B] = \mathbb{P}[B|A]\,\mathbb{P}[A]$$

*More generally, for any events $A_1, \ldots, A_n$,*

$$\mathbb{P}[\textstyle\bigcap_{i=1}^{n} A_i] = \mathbb{P}[A_1] \times \mathbb{P}[A_2|A_1] \times \mathbb{P}[A_3|A_1 \cap A_2] \times \cdots \times \mathbb{P}[A_n|\textstyle\bigcap_{i=1}^{n-1} A_i].$$

*Proof.* The first assertion follows directly from the definition of $\mathbb{P}[B|A]$ (and is in fact a special case of the second assertion with $n = 2$).

To prove the second assertion, we will use induction on $n$ (the number of events). The base case is $n = 1$, and corresponds to the statement that $\mathbb{P}[A] = \mathbb{P}[A]$, which is trivially true. For an arbitrary $n > 1$, assume (the inductive hypothesis) that

$$\mathbb{P}[\textstyle\bigcap_{i=1}^{n-1} A_i] = \mathbb{P}[A_1] \times \mathbb{P}[A_2|A_1] \times \cdots \times \mathbb{P}[A_{n-1}|\textstyle\bigcap_{i=1}^{n-2} A_i].$$

Now we can apply the definition of conditional probability to the two events $A_n$ and $\bigcap_{i=1}^{n-1} A_i$ to deduce that

$$\begin{aligned}\mathbb{P}[\textstyle\bigcap_{i=1}^{n} A_i] = \mathbb{P}[A_n \cap (\textstyle\bigcap_{i=1}^{n-1} A_i)] &= \mathbb{P}[A_n|\textstyle\bigcap_{i=1}^{n-1} A_i] \times \mathbb{P}[\textstyle\bigcap_{i=1}^{n-1} A_i] \\ &= \mathbb{P}[A_n|\textstyle\bigcap_{i=1}^{n-1} A_i] \times \mathbb{P}[A_1] \times \mathbb{P}[A_2|A_1] \times \cdots \times \mathbb{P}[A_{n-1}|\textstyle\bigcap_{i=1}^{n-2} A_i],\end{aligned}$$

where in the last line we have used the inductive hypothesis. This completes the proof by induction. □

The Product Rule is particularly useful when we can view our sample space as a sequence of choices. The next few examples illustrate this point.

### Example: Coin Tosses

Toss a fair coin three times. Let $A$ be the event that all three tosses are heads. Then $A = A_1 \cap A_2 \cap A_3$, where $A_i$ is the event that the $i$th toss comes up heads. We have

$$\begin{aligned}\mathbb{P}[A] &= \mathbb{P}[A_1] \times \mathbb{P}[A_2|A_1] \times \mathbb{P}[A_3|A_1 \cap A_2] \\ &= \mathbb{P}[A_1] \times \mathbb{P}[A_2] \times \mathbb{P}[A_3] \\ &= \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \tfrac{1}{8}.\end{aligned}$$

The second line here follows from the fact that the tosses are mutually independent. Of course, we already know that $\mathbb{P}[A] = \frac{1}{8}$ from our definition of the probability space in the previous note. Another way of looking at this calculation is that it justifies our definition of the probability space, and shows that it was consistent with assuming that the coin flips are mutually independent.

If the coin is biased with heads probability $p$, we get, again using independence,

$$\mathbb{P}[A] = \mathbb{P}[A_1] \times \mathbb{P}[A_2] \times \mathbb{P}[A_3] = p^3.$$

More generally, the probability of any sequence of $n$ tosses containing $k$ heads and $n-k$ tails is $p^k(1-p)^{n-k}$. This is in fact the reason we defined the probability space this way: we defined the sample point probabilities so that the coin tosses would behave independently.

### Example: Monty Hall Revisited

Recall the Monty Hall problem from the previous note: there are three doors and the probability that the prize is behind any given door is $\frac{1}{3}$. There are goats behind the other two doors. The contestant picks a door randomly, and the host opens one of the other two doors, revealing a goat. How do we calculate intersections in this setting? For example, what is the probability that the contestant chooses door 1, the prize is behind door 2, and the host chooses door 3?

Let $C_i$ be the event that the contestant chooses door $i$, let $P_i$ be the event that the prize is behind door $i$, and let $H_i$ be the event that the host chooses door $i$. Then, by the Product Rule,

$$\mathbb{P}[C_1 \cap P_2 \cap H_3] = \mathbb{P}[C_1] \times \mathbb{P}[P_2|C_1] \times \mathbb{P}[H_3|C_1 \cap P_2].$$

The probability of $C_1$ is $\frac{1}{3}$, since the contestant is choosing the door at random. The probability of $P_2$ given $C_1$ is still $\frac{1}{3}$ since they are independent. The probability of the host choosing door 3 given events $C_1$ and $P_2$ is 1; the host cannot choose door 1 since the contestant has already chosen it, and the host cannot choose door 2 since the host must reveal a goat (and not the prize). Therefore,

$$\mathbb{P}[C_1 \cap P_2 \cap H_3] = \frac{1}{3} \times \frac{1}{3} \times 1 = \frac{1}{9}.$$

Observe that we needed conditional probability in this setting; had we simply multiplied the probabilities of each event, we would have obtained $\frac{1}{27}$ since the probability of $H_3$ is also $\frac{1}{3}$ (can you figure out why?). What if we changed the situation, and instead asked for the probability that the contestant chooses door 1, the prize is behind door 1, and the host chooses door 3? We can use the same technique as above, but our final answer will be different. This is left as an exercise (see Exercise 4).

Now, noting that $\mathbb{P}[H_3|C_1] = \frac{1}{2}$ (see Exercise 5) and $\mathbb{P}[C_1 \cap H_3] = \mathbb{P}[C_1]\mathbb{P}[H_3|C_1] = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$, we obtain

$$\mathbb{P}[P_2 \mid C_1 \cap H_3] = \frac{\mathbb{P}[C_1 \cap P_2 \cap H_3]}{\mathbb{P}[C_1 \cap H_3]} = \frac{\frac{1}{9}}{\frac{1}{6}} = \frac{2}{3},$$

which formally justifies the intuitive answer described in the previous note.

### Example: Poker Hands

Let's use the Product Rule to compute the probability of a flush in a different way. This is equal to $4 \times \mathbb{P}[A]$, where $A$ is the probability of a Hearts flush. Intuitively, this should be clear since there are 4 suits; we'll see why this is formally true in the next section. We can write $A = \bigcap_{i=1}^{5} A_i$, where $A_i$ is the event that the $i$th card we pick is a Heart. So we have

$$\mathbb{P}[A] = \mathbb{P}[A_1] \times \mathbb{P}[A_2|A_1] \times \cdots \times \mathbb{P}[A_5|\textstyle\bigcap_{i=1}^{4} A_i].$$

Clearly $\mathbb{P}[A_1] = \frac{13}{52} = \frac{1}{4}$. What about $\mathbb{P}[A_2|A_1]$? Well, since we are conditioning on $A_1$ (the first card is a Heart), there are only 51 remaining possibilities for the second card, 12 of which are Hearts. So $\mathbb{P}[A_2|A_1] = \frac{12}{51}$. Similarly, $\mathbb{P}[A_3|A_1 \cap A_2] = \frac{11}{50}$, and so on. So we get

$$4 \times \mathbb{P}[A] = 4 \times \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48},$$

which is exactly the same fraction we computed in the previous note.

So now we have two methods of computing probabilities in many of our sample spaces. It is useful to keep these different methods around, both as a check on your answers and because in some cases one of the methods is easier to use than the other.

## 4.4 Unions of Events

You are in Las Vegas, and you spy a new game with the following rules. You pick a number between 1 and 6. Then three dice are thrown. You win if and only if your number comes up on at least one of the dice.

The casino claims that your odds of winning are 50%, using the following argument. Let $A$ be the event that you win. We can write $A = A_1 \cup A_2 \cup A_3$, where $A_i$ is the event that your number comes up on die $i$. Clearly $\mathbb{P}[A_i] = \frac{1}{6}$ for each $i$. Therefore, they claim

$$\mathbb{P}[A] = \mathbb{P}[A_1 \cup A_2 \cup A_3] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + \mathbb{P}[A_3] = 3 \times \frac{1}{6} = \frac{1}{2}.$$

Is this calculation correct? Well, suppose instead that the casino rolled six dice, and again you win if and only if your number comes up at least once. Then the analogous calculation would say that you win with probability $6 \times \frac{1}{6} = 1$, i.e., certainly! The situation becomes even more ridiculous when the number of dice gets bigger than 6.

The problem is that the events $A_i$ are *not disjoint*: i.e., there are some sample points that lie in more than one of the $A_i$. (We could get really lucky and our number could come up on two of the dice, or all three.) So, if we add up the $\mathbb{P}[A_i]$, we are counting some sample points more than once.

Fortunately, in Note 11 we learned about the Principle of Inclusion-Exclusion which allows us to deal with this kind of situation. In the proof of Theorem 11.3, it was shown that every $\omega \notin A_1 \cup \cdots \cup A_n$ is not counted by the Inclusion-Exclusion formula, while every $\omega \in A_1 \cup \cdots \cup A_n$ is counted exactly once by the formula. Hence, rather than summing (with appropriate signs) the cardinality of $\cap_{i \in S} A_i$ in the Inclusion-Exclusion formula, if we instead sum their probability $\mathbb{P}[\cap_{i \in S} A_i]$, we obtain the following useful formula for computing $\mathbb{P}[A_1 \cup \cdots \cup A_n]$:

**Theorem 14.2** (Inclusion-Exclusion). *Let $A_1, \ldots, A_n$ be events in some probability space, where $n \geq 2$. Then, we have*

$$\mathbb{P}[A_1 \cup \cdots \cup A_n] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq \{1,\ldots,n\} : |S| = k} \mathbb{P}[\cap_{i \in S} A_i]. \tag{10}$$

The right hand side of (10) can be written as

$$\mathbb{P}[\textstyle\bigcup_{i=1}^{n} A_i] = \displaystyle\sum_{i=1}^{n} \mathbb{P}[A_i] - \sum_{i<j} \mathbb{P}[A_i \cap A_j] + \sum_{i<j<k} \mathbb{P}[A_i \cap A_j \cap A_k] - \cdots + (-1)^{n-1} \mathbb{P}[A_1 \cap A_2 \cap \cdots \cap A_n],$$

where $\sum_{i<j}$ denotes summing over all $i, j \in \{1, \ldots, n\}$ such that $i < j$, and so on. That is, to compute $\mathbb{P}[\bigcup_i A_i]$, we start by summing the event probabilities $\mathbb{P}[A_i]$, then we *subtract* the probabilities of all pairwise intersections, then we *add* back in the probabilities of all three-way intersections, and so on. You might like to verify it for the special case $n = 3$ by drawing a Venn diagram. You might also like to prove the formula for general $n$ by induction (in similar fashion to the proof of the Product Rule above).

Using (10), the probability we get lucky in the new game in Las Vegas is given by

$$\mathbb{P}[A_1 \cup A_2 \cup A_3] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + \mathbb{P}[A_3] - \mathbb{P}[A_1 \cap A_2] - \mathbb{P}[A_1 \cap A_3] - \mathbb{P}[A_2 \cap A_3] + \mathbb{P}[A_1 \cap A_2 \cap A_3].$$

Now the nice thing here is that the events $A_i$ are mutually independent (the outcome of any die does not depend on that of the others), so $\mathbb{P}[A_i \cap A_j] = \mathbb{P}[A_i]\mathbb{P}[A_j] = (\frac{1}{6})^2 = \frac{1}{36}$, and similarly $\mathbb{P}[A_1 \cap A_2 \cap A_3] = (\frac{1}{6})^3 = \frac{1}{216}$. So, we get

$$\mathbb{P}[A_1 \cup A_2 \cup A_3] = \left(3 \times \tfrac{1}{6}\right) - \left(3 \times \tfrac{1}{36}\right) + \tfrac{1}{216} = \tfrac{91}{216} \approx 0.42.$$

So your odds are quite a bit worse than what the casino is claiming!

When $n$ is large (i.e., we are interested in the union of many events), the Inclusion-Exclusion formula is essentially useless because it involves computing the probability of the intersection of every non-empty subset of the events: and there are $2^n - 1$ of these! Sometimes we can just look at the first few terms of it and forget the rest: note that successive terms actually give us an overestimate and then an underestimate of the answer, and these estimates both get better as we go along.

However, in many situations we can get a long way by just looking at the first term:

1. (**Mutually exclusive events**) If the events $A_1, \ldots, A_n$ are mutually exclusive (i.e., $A_i \cap A_j = \varnothing$ for all $i \neq j$), then
$$\mathbb{P}[\bigcup_{i=1}^n A_i] = \sum_{i=1}^n \mathbb{P}[A_i].$$
[Note that we have already used this fact several times in our examples, e.g., in claiming that the probability of a flush is four times the probability of a Hearts flush — clearly flushes in different suits are disjoint events.]

2. (**Union bound**) Let $A_1, \ldots, A_n$ be events in some probability space. Then, for all $n \in \mathbb{Z}^+$,
$$\mathbb{P}[\bigcup_{i=1}^n A_i] \leq \sum_{i=1}^n \mathbb{P}[A_i]. \tag{11}$$
This merely says that adding up the $\mathbb{P}[A_i]$ can only *over*estimate the probability of the union. Crude as it may seem, we will later see how to use the union bound effectively in Computer Science examples.

# 5 Exercises

1. Suppose five cards are dealt from a standard deck of playing cards. Let $A_i$ be the event that the $i$th card is an ace. Show that $\mathbb{P}[A_i] = \frac{1}{13}$ for all $i = 1, \ldots, 5$.
2. Show (4) and (5).
3. Show that Definitions 14.4 and 14.5 of mutual independence are equivalent.
4. In the Monty Hall problem, find $\mathbb{P}[C_1 \cap P_1 \cap H_3]$.
5. In the Monty Hall problem, show that $\mathbb{P}[H_3 \mid C_1] = \frac{1}{2}$.
6. Prove the union bound (11).