Chapter 1

Basic Probability Theory

In this chapter we introduce the mathematical framework of probability theory, which makes it possible to reason about uncertainty in a principled way using set theory. Appendix A contains a review of basic set-theory concepts.

1.1 Probability spaces

Our goal is to build a mathematical framework to represent and analyze uncertain phenomena, such as the result of rolling a die, tomorrow's weather, the result of an NBA game, etc. To this end we model the phenomenon of interest as an **experiment** with several (possibly infinite) mutually exclusive **outcomes**.

Except in simple cases, when the number of outcomes is small, it is customary to reason about sets of outcomes, called *events*. To quantify how likely it is for the outcome of the experiment to belong to a specific event, we assign a **probability** to the event. More formally, we define a **measure** (recall that a measure is a function that maps sets to real numbers) that assigns probabilities to each event of interest.

More formally, the experiment is characterized by constructing a **probability space**.

Definition 1.1.1 (Probability space)**.** *A probability space is a triple* $(\Omega, \mathcal{F}, \mathrm{P})$ *consisting of:*

- *A* ***sample space*** Ω*, which contains all possible outcomes of the experiment.*
- *A set of events* $\mathcal{F}$*, which must be a* ***σ-algebra*** *(see Definition 1.1.2 below).*
- *A* ***probability measure*** P *that assigns probabilities to the events in* $\mathcal{F}$ *(see Definition 1.1.4 below).*

Sample spaces may be **discrete** or **continuous**. Examples of discrete sample spaces include the possible outcomes of a coin toss, the score of a basketball game, the number of people that show up at a party, etc. Continuous sample spaces are usually intervals of $\mathbb{R}$ or $\mathbb{R}^n$ used to model time, position, temperature, etc.

The term σ-algebra is used in measure theory to denote a collection of sets that satisfy certain conditions listed below. Don't be too intimidated by it. It is just a sophisticated way of stating that if we assign a probability to certain events (for example *it will rain tomorrow* or *it will*

snow tomorrow) we also need to assign a probability to their complements (i.e. *it will not rain tomorrow* or *it will not snow tomorrow*) and to their union (*it will rain or snow tomorrow*).

Definition 1.1.2 (σ-algebra)**.** *A σ-algebra $\mathcal{F}$ is a collection of sets in Ω such that:*

1. *If a set $S \in \mathcal{F}$ then $S^c \in \mathcal{F}$.*
2. *If the sets $S_1, S_2 \in \mathcal{F}$, then $S_1 \cup S_2 \in \mathcal{F}$. This also holds for infinite sequences; if $S_1, S_2, \ldots \in \mathcal{F}$ then $\cup_{i=1}^{\infty} S_i \in \mathcal{F}$.*
3. *$\Omega \in \mathcal{F}$.*

If our sample space is discrete, a possible choice for the σ-algebra is the **power set** of the sample space, which consists of all possible sets of elements in the sample space. If we are tossing a coin and the sample space is

$$\Omega := \{\text{heads}, \text{tails}\}, \tag{1.1}$$

then the power set is a valid σ-algebra

$$\mathcal{F} := \{\text{heads or tails}, \text{heads}, \text{tails}, \emptyset\}, \tag{1.2}$$

where $\emptyset$ denotes the empty set. However, in many cases σ-algebras do not contain every possible set of outcomes.

Example 1.1.3 (Cholesterol)**.** A doctor is interested in modeling the cholesterol levels of her patients probabilistically. Every time a patient visits her, she tests their cholesterol level. Here the *experiment* is the cholesterol test, the outcome is the measured cholesterol level, and the sample space Ω is the positive real line. The doctor is mainly interested in whether the patients to have low, borderline-high, or high cholesterol. The event L (low cholesterol) contains all outcomes below 200 mg/dL, the event B (borderline-high cholesterol) contains all outcomes between 200 and 240 mg/dL, and the event H (high cholesterol) contains all outcomes above 240 mg/dL. The σ-algebra $\mathcal{F}$ of possible events therefore equals

$$\mathcal{F} := \{L \cup B \cup H, L \cup B, L \cup H, B \cup H, L, B, H, \emptyset\}. \tag{1.3}$$

The events are a partition of the sample space, which simplifies deriving the corresponding σ-algebra. △

The role of the probability measure P is to quantify how likely we are to encounter each of the events in the σ-algebra. Intuitively, the probability of an event A can be interpreted as the fraction of times that the outcome of the experiment is in A, as the number of repetitions tends to infinity. It follows that probabilities should always be nonnegative. Also, if two events A and B are disjoint (their intersection is empty), then

$$P(A \cup B) = \frac{\text{outcomes in } A \text{ or } B}{\text{total}} \tag{1.4}$$

$$= \frac{\text{outcomes in } A + \text{outcomes in } B}{\text{total}} \tag{1.5}$$

$$= \frac{\text{outcomes in } A}{\text{total}} + \frac{\text{outcomes in } B}{\text{total}} \tag{1.6}$$

$$= P(A) + P(B). \tag{1.7}$$

Probabilities of unions of disjoint events should equal the sum of the individual probabilities. Additionally, the probability of the whole sample space Ω should equal one, as it contains all outcomes

$$P(\Omega) = \frac{\text{outcomes in } \Omega}{\text{total}} \tag{1.8}$$

$$= \frac{\text{total}}{\text{total}} \tag{1.9}$$

$$= 1. \tag{1.10}$$

These conditions are necessary for a measure to be a valid probability measure.

Definition 1.1.4 (Probability measure)**.** *A probability measure is a function defined over the sets in a σ-algebra $\mathcal{F}$ such that:*

1. $\mathrm{P}(S) \geq 0$ *for any event* $S \in \mathcal{F}$.

2. *If the sets* $S_1, S_2, \ldots, S_n \in \mathcal{F}$ *are disjoint (i.e.* $S_i \cap S_j = \emptyset$ *for* $i \neq j$*) then*

$$\mathrm{P}\left(\cup_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1.11}$$

Similarly, for a countably infinite sequence of disjoint sets $S_1, S_2, \ldots \in \mathcal{F}$

$$\mathrm{P}\left(\lim_{n\to\infty} \cup_{i=1}^{n} S_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1.12}$$

3. $\mathrm{P}(\Omega) = 1$.

The two first axioms capture the intuitive idea that the probability of an event is a measure such as mass (or length or volume): just like the mass of any object is nonnegative and the total mass of several distinct objects is the sum of their masses, the probability of any event is nonnegative and the probability of the union of several disjoint objects is the sum of their probabilities. However, in contrast to mass, the amount of probability in an experiment cannot be unbounded. If it is highly likely that it will rain tomorrow, then it cannot be also very likely that it will *not* rain. If the probability of an event S is large, then the probability of its complement S^c must be small. This is captured by the third axiom, which normalizes the probability measure (and implies that $\mathrm{P}(S^c) = 1 - \mathrm{P}(S)$).

It is important to stress that the probability measure does *not* assign probabilities to individual outcomes, but rather to events in the σ-algebra. The reason for this is that when the number of possible outcomes is uncountably infinite, then one cannot assign nonzero probability to all the outcomes and still satisfy the condition $\mathrm{P}(\Omega) = 1$. This is not an exotic situation, it occurs for instance in the cholesterol example where any positive real number is a possible outcome. In the case of discrete or countable sample spaces, the σ-algebra may equal the power set of the sample space, which means that we do assign probabilities to events that only contain a single outcome (e.g. the coin-toss example).

Example 1.1.5 (Cholesterol (continued))**.** A valid probability measure for Example 1.1.3 is

$$\mathrm{P}(L) = 0.6, \quad \mathrm{P}(B) = 0.28, \quad \mathrm{P}(H) = 0.12. \tag{1.13}$$

Using the properties, we can determine for instance that $\mathrm{P}(B \cup H) = 0.6 + 0.28 = 0.88$. △

Definition 1.1.4 has the following consequences:

$$\mathrm{P}(\emptyset) = 0, \tag{1.14}$$
$$A \subseteq B \quad \text{implies} \quad \mathrm{P}(A) \leq \mathrm{P}(B), \tag{1.15}$$
$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{1.16}$$

We omit the proofs (try proving them on your own).

1.2 Conditional probability

Conditional probability is a crucial concept in probabilistic modeling. It allows us to update probabilistic models when additional information is revealed. Consider a probabilistic space $(\Omega, \mathcal{F}, \mathrm{P})$ where we find out that the outcome of the experiment belongs to a certain event $S \in \mathcal{F}$. This obviously affects how likely it is for any other event $S' \in \mathcal{F}$ to have occurred: we can rule out any outcome not belonging to S. The updated probability of each event is known as the **conditional probability** of S' **given** S. Intuitively, the conditional probability can be interpreted as the fraction of outcomes in S that are also in S',

$$\mathrm{P}\left(S'|S\right) = \frac{\text{outcomes in } S' \text{ and } S}{\text{outcomes in } S} \tag{1.17}$$
$$= \frac{\text{outcomes in } S' \text{ and } S}{\text{total}} \frac{\text{total}}{\text{outcomes in } S} \tag{1.18}$$
$$= \frac{\mathrm{P}(S' \cap S)}{\mathrm{P}(S)}, \tag{1.19}$$

where we assume that $\mathrm{P}(S) \neq 0$ (later on we will have to deal with the case when S has zero probability, which often occurs in continuous probability spaces). The definition is rather intuitive: S is now the new sample space, so if the outcome is in S' then it must belong to $S' \cap S$. However, just using the probability of the intersection would underestimate how likely it is for S' to occur because the sample space has been reduced to S. Therefore we normalize by the probability of S. As a sanity check, we have $\mathrm{P}(S|S) = 1$ and if S and S' are disjoint then $\mathrm{P}(S'|S) = 0$.

The conditional probability $\mathrm{P}(\cdot|S)$ is a valid probability measure in the probability space $(S, \mathcal{F}_S, \mathrm{P}(\cdot|S))$, where $\mathcal{F}_S$ is a σ-algebra that contains the intersection of S and the sets in $\mathcal{F}$. To simplify notation, when we condition on an intersection of sets we write the conditional probability as

$$\mathrm{P}(S|A, B, C) := \mathrm{P}(S|A \cap B \cap C), \tag{1.20}$$

for any events S, A, B, C.

Example 1.2.1 (Flights and rain)**.** JFK airport hires you to estimate how the punctuality of flight arrivals is affected by the weather. You begin by defining a probability space for which the sample space is

$$\Omega = \{\text{late and rain}, \text{late and no rain}, \text{on time and rain}, \text{on time and no rain}\} \tag{1.21}$$

and the σ-algebra is the power set of Ω. From data of past flights you determine that a reasonable estimate for the probability measure of the probability space is

$$\mathrm{P}(\text{late, no rain}) = \frac{2}{20}, \qquad \mathrm{P}(\text{on time, no rain}) = \frac{14}{20}, \tag{1.22}$$

$$\mathrm{P}(\text{late, rain}) = \frac{3}{20}, \qquad \mathrm{P}(\text{on time, rain}) = \frac{1}{20}. \tag{1.23}$$

The airport is interested in the probability of a flight being late if it rains, so you define a new probability space conditioning on the event *rain*. The sample space is the set of all outcomes such that *rain* occurred, the σ-algebra is the power set of $\{\text{on time}, \text{late}\}$ and the probability measure is $\mathrm{P}(\cdot|\text{rain})$. In particular,

$$\mathrm{P}(\text{late}|\text{rain}) = \frac{\mathrm{P}(\text{late, rain})}{\mathrm{P}(\text{rain})} = \frac{3/20}{3/20 + 1/20} = \frac{3}{4} \tag{1.24}$$

and similarly $\mathrm{P}(\text{late}|\text{no rain}) = 1/8$.

△

Conditional probabilities can be used to compute the intersection of several events in a structured way. By definition, we can express the probability of the intersection of two events $A, B \in \mathcal{F}$ as follows,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) \tag{1.25}$$

$$= \mathrm{P}(B)\,\mathrm{P}(A|B)\,. \tag{1.26}$$

In this formula $\mathrm{P}(A)$ is known as the **prior** probability of A, as it captures the information we have about A before anything else is revealed. Analogously, $\mathrm{P}(A|B)$ is known as the **posterior** probability. These are fundamental quantities in Bayesian models, discussed in Chapter 10. Generalizing (1.25) to a sequence of events gives the *chain rule*, which allows to express the probability of the intersection of multiple events in terms of conditional probabilities. We omit the proof, which is a straightforward application of induction.

Theorem 1.2.2 (Chain rule)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $S_1, S_2, \ldots$ *a collection of events in* $\mathcal{F}$*,*

$$\mathrm{P}(\cap_i S_i) = \mathrm{P}(S_1)\,\mathrm{P}(S_2|S_1)\,\mathrm{P}(S_3|S_1 \cap S_2)\cdots \tag{1.27}$$

$$= \prod_i \mathrm{P}\left(S_i|\cap_{j=1}^{i-1} S_j\right). \tag{1.28}$$

Sometimes, estimating the probability of a certain event directly may be more challenging than estimating its probability conditioned on simpler events. A collection of disjoint sets $A_1, A_2, \ldots$ such that $\Omega = \cup_i A_i$ is called a **partition** of Ω. The law of total probability allows us to pool conditional probabilities together, weighting them by the probability of the individual events in the partition, to compute the probability of the event of interest.

Theorem 1.2.3 (Law of total probability)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and let the collection of disjoint sets* A_1, A_2, $\ldots \in \mathcal{F}$ *be any partition of* Ω*. For any set* $S \in \mathcal{F}$

$$\mathrm{P}(S) = \sum_i \mathrm{P}(S \cap A_i) \tag{1.29}$$

$$= \sum_i \mathrm{P}(A_i)\,\mathrm{P}(S|A_i)\,. \tag{1.30}$$

Proof. This is an immediate consequence of the chain rule and Axiom 2 in Definition 1.1.4, since $S = \cup_i S \cap A_i$ and the sets $S \cap A_i$ are disjoint. □

Example 1.2.4 (Aunt visit)**.** Your aunt is arriving at JFK tomorrow and you would like to know how likely it is for her flight to be on time. From Example 1.2.1, you recall that

$$\mathrm{P}(\text{late}|\text{rain}) = 0.75, \quad \mathrm{P}(\text{late}|\text{no rain}) = 0.125. \tag{1.31}$$

After checking out a weather website, you determine that $\mathrm{P}(\text{rain}) = 0.2$.

Now, how can we integrate all of this information? The events *rain* and *no rain* are disjoint and cover the whole sample space, so they form a partition. We can consequently apply the law of total probability to determine

$$\mathrm{P}(\text{late}) = \mathrm{P}(\text{late}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain}) \tag{1.32}$$

$$= 0.75 \cdot 0.2 + 0.125 \cdot 0.8 = 0.25. \tag{1.33}$$

So the probability that your aunt's plane is late is 1/4.

△

It is crucial to realize that in general $\mathrm{P}(A|B) \neq \mathrm{P}(B|A)$: most players in the NBA probably own a basketball ($\mathrm{P}(\text{owns ball}|\text{NBA})$ is large) but most people that own basketballs are not in the NBA ($\mathrm{P}(\text{NBA}|\text{owns ball})$ is small). The reason is that the prior probabilities are very different: $\mathrm{P}(\text{NBA})$ is much smaller than $\mathrm{P}(\text{owns ball})$. However, it is possible to *invert* conditional probabilities, i.e. find $\mathrm{P}(A|B)$ from $\mathrm{P}(B|A)$, as long as we take into account the priors. This straightforward consequence of the definition of conditional probability is known as Bayes' rule.

Theorem 1.2.5 (Bayes' rule)**.** *For any events* A *and* B *in a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B|A)}{\mathrm{P}(B)}, \tag{1.34}$$

as long as $\mathrm{P}(B) > 0$.

Example 1.2.6 (Aunt visit (continued))**.** You explain the probabilistic model described in Example 1.2.4 to your cousin Marvin who lives in California. A day later, you tell him that your aunt arrived late but you don't mention whether it rained or not. After he hangs up, Marvin wants to figure out the probability that it rained. Recall that the probability of rain was 0.2, but since your aunt arrived late he should update the estimate. Applying Bayes' rule and the

law of total probability:

$$\begin{aligned} \mathrm{P}\left(\text{rain}|\text{late}\right) &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}\right)} && (1.35)\\ &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)+\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} && (1.36)\\ &= \frac{0.75\cdot 0.2}{0.75\cdot 0.2+0.125\cdot 0.8} = 0.6. && (1.37) \end{aligned}$$

As expected, the probability that it rained increases under the assumption that your aunt is late.

△

1.3 Independence

As discussed in the previous section, conditional probabilities quantify the extent to which the knowledge of the occurrence of a certain event affects the probability of another event. In some cases, it makes no difference: the events are **independent**. More formally, events A and B are independent if and only if

$$\mathrm{P}\left(A|B\right) = \mathrm{P}\left(A\right). \tag{1.38}$$

This definition is not valid if $\mathrm{P}\left(B\right) = 0$. The following definition covers this case and is otherwise equivalent.

Definition 1.3.1 (Independence)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are independent if and only if*

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A\right)\mathrm{P}\left(B\right). \tag{1.39}$$

Example 1.3.2 (Congress)**.** We consider a data set compiling the votes of members of the U.S. House of Representatives on two issues in 1984 [1]. The issues are cost sharing for a water project (issue 1) and adoption of the budget resolution (issue 2). We model the behavior of the congressmen probabilistically, defining a sample space where each outcome is a sequence of votes. For instance, a possible outcome is *issue 1 = yes, issue 2 = no*. We choose the σ-algebra to be the power set of the sample space. To estimate the probability measure associated to different events, we just compute the fraction of their occurrence in the data.

$$\begin{aligned} \mathrm{P}\left(\text{issue 1} = \text{yes}\right) &\approx \frac{\text{members voting yes on issue 1}}{\text{total votes on issue 1}} && (1.40)\\ &= 0.597, && (1.41)\\ \mathrm{P}\left(\text{issue 2} = \text{yes}\right) &\approx \frac{\text{members voting yes on issue 2}}{\text{total votes on issue 2}} && (1.42)\\ &= 0.417, && (1.43)\\ \mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes}\right) &\approx \frac{\text{members voting yes on issues 1 and 2}}{\text{total members voting on issues 1 and 2}} && (1.44)\\ &= 0.069. && (1.45) \end{aligned}$$

[1] The data is available here.

Based on these data, we can evaluate whether voting behavior on the two issues was dependent. In other words, if we know how a member voted on issue 1, does this provide information about how they voted on issue 2? The answer is yes, since

$$P\left(\text{issue 1} = \text{yes}\right) \mathrm{P}\left(\text{issue 2} = \text{yes}\right) = 0.249 \tag{1.46}$$

is very different from $\mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes}\right)$. If a member voted yes on issue 1, they were less likely to vote yes on issue 2. △

Similarly, we can define **conditional independence** between two events given a third event. A and B are conditionally independent given C if and only if

$$\mathrm{P}\left(A|B,C\right) = \mathrm{P}\left(A|C\right), \tag{1.47}$$

where $\mathrm{P}\left(A|B,C\right) := \mathrm{P}\left(A|B \cap C\right)$. Intuitively, this means that the probability of A is not affected by whether B occurs or not, *as long as* C *occurs*.

Definition 1.3.3 (Conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. Two events* $A, B \in \mathcal{F}$ *are conditionally independent given a third event* $C \in \mathcal{F}$ *if and only if*

$$\mathrm{P}\left(A \cap B|C\right) = \mathrm{P}\left(A|C\right) \mathrm{P}\left(B|C\right). \tag{1.48}$$

Example 1.3.4 (Congress (continued))**.** The main factor that determines how members of congress vote is political affiliation. We therefore incorporate it into the probabilistic model in Example 1.3.2. Each outcome now consists of the votes for issues 1 and 2, and also the affiliation of the member, e.g. *issue 1 = yes*, *issue 2 = no*, *affiliation = republican*, or *issue 1 = no*, *issue 2 = no*, *affiliation = democrat*. The σ-algebra is the power set of the sample space. We again estimate the values of the probability measure associated to different events using the data:

$$\mathrm{P}\left(\text{issue 1} = \text{yes} \,|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issue 1}}{\text{total republican votes on issue 1}} \tag{1.49}$$

$$= 0.134, \tag{1.50}$$

$$\mathrm{P}\left(\text{issue 2} = \text{yes} \,|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issue 2}}{\text{total republican votes on issue 2}} \tag{1.51}$$

$$= 0.988, \tag{1.52}$$

$$\mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes} \,|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issues 1 and 2}}{\text{republicans voting on both issues}}$$

$$= 0.134. \tag{1.53}$$

Based on these data, we can evaluate whether voting behavior on the two issues was dependent *conditioned on the member being a republican*. In other words, if we know how a member voted on issue 1 and that they are a republican, does this provide information about how they voted on issue 2? The answer is no, since

$$\mathrm{P}\left(\text{issue 1} = \text{yes} \,|\, \text{republican}\right) \mathrm{P}\left(\text{issue 2} = \text{yes} \,|\, \text{republican}\right) = 0.133 \tag{1.54}$$

is very close to $\mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes} \,|\, \text{republican}\right)$. The votes are approximately independent given the knowledge that the member is a republican. △

As suggested by Examples 1.3.2 and 1.3.4, independence does not imply conditional independence or vice versa. This is further illustrated by the following examples. From now on, to simplify notation, we write the probability of the intersection of several events in the following form

$$\mathrm{P}\left(A, B, C\right) := \mathrm{P}\left(A \cap B \cap C\right). \tag{1.55}$$

Example 1.3.5 (Conditional independence does not imply independence)**.** Your cousin Marvin from Exercise 1.2.6 always complains about taxis in New York. From his many visits to JFK he has calculated that

$$\mathrm{P}\left(\text{taxi}|\text{rain}\right) = 0.1, \quad \mathrm{P}\left(\text{taxi}|\text{no rain}\right) = 0.6, \tag{1.56}$$

where *taxi* denotes the event of finding a free taxi after picking up your luggage. Given the events *rain* and *no rain*, it is reasonable to model the events *plane arrived late* and *taxi* as conditionally independent,

$$\mathrm{P}\left(\text{taxi}, \text{late}|\text{rain}\right) = \mathrm{P}\left(\text{taxi}|\text{rain}\right)\mathrm{P}\left(\text{late}|\text{rain}\right), \tag{1.57}$$

$$\mathrm{P}\left(\text{taxi}, \text{late}|\text{no rain}\right) = \mathrm{P}\left(\text{taxi}|\text{no rain}\right)\mathrm{P}\left(\text{late}|\text{no rain}\right). \tag{1.58}$$

The logic behind this is that the availability of taxis after picking up your luggage depends on whether it's raining or not, but not on whether the plane is late or not (we assume that availability is constant throughout the day). Does this assumption imply that the events are independent?

If they were independent, then knowing that your aunt was late would give no information to Marvin about taxi availability. However,

$$\mathrm{P}\left(\text{taxi}\right) = \mathrm{P}\left(\text{taxi}, \text{rain}\right) + \mathrm{P}\left(\text{taxi}, \text{no rain}\right) \quad \text{(by the law of total probability)} \tag{1.59}$$

$$= \mathrm{P}\left(\text{taxi}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right) + \mathrm{P}\left(\text{taxi}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right) \tag{1.60}$$

$$= 0.1 \cdot 0.2 + 0.6 \cdot 0.8 = 0.5, \tag{1.61}$$

$$\mathrm{P}\left(\text{taxi}|\text{late}\right) = \frac{\mathrm{P}\left(\text{taxi}, \text{late}, \text{rain}\right) + \mathrm{P}\left(\text{taxi}, \text{late}, \text{no rain}\right)}{\mathrm{P}\left(\text{late}\right)} \quad \text{(by the law of total probability)}$$

$$= \frac{\mathrm{P}\left(\text{taxi}|\text{rain}\right)\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right) + \mathrm{P}\left(\text{taxi}|\text{no rain}\right)\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)}{\mathrm{P}\left(\text{late}\right)}$$

$$= \frac{0.1 \cdot 0.75 \cdot 0.2 + 0.6 \cdot 0.125 \cdot 0.8}{0.25} = 0.3. \tag{1.62}$$

$\mathrm{P}\left(\text{taxi}\right) \neq \mathrm{P}\left(\text{taxi}|\text{late}\right)$ so the events are *not* independent. This makes sense, since if the airplane is late, it is more probable that it is raining, which makes taxis more difficult to find.

△

Example 1.3.6 (Independence does not imply conditional independence)**.** After looking at your probabilistic model from Example 1.2.1 your contact at JFK points out that delays are often caused by mechanical problems in the airplanes. You look at the data and determine that

$$\mathrm{P}\left(\text{problem}\right) = \mathrm{P}\left(\text{problem}|\text{rain}\right) = \mathrm{P}\left(\text{problem}|\text{no rain}\right) = 0.1, \tag{1.63}$$

so the events *mechanical problem* and *rain in NYC* are independent, which makes intuitive sense. After some more analysis of the data, you estimate

$$\mathrm{P}\left(\text{late}|\text{problem}\right) = 0.7, \quad \mathrm{P}\left(\text{late}|\text{no problem}\right) = 0.2, \quad \mathrm{P}\left(\text{late}|\text{no rain}, \text{problem}\right) = 0.5.$$

The next time you are waiting for Marvin at JFK, you start wondering about the probability of his plane having had some mechanical problem. Without any further information, this probability is 0.1. It is a sunny day in New York, but this is of no help because according to the data (and common sense) the events *problem* and *rain* are independent.

Suddenly they announce that Marvin's plane is late. Now, what is the probability that his plane had a mechanical problem? At first thought you might apply Bayes' rule to compute $\mathrm{P}\left(\text{problem}|\text{late}\right) = 0.28$ as in Example 1.2.6. However, you are not using the fact that it is sunny. This means that the rain was not responsible for the delay, so intuitively a mechanical problem should be more likely. Indeed,

$$\begin{aligned}
\mathrm{P}\left(\text{problem}|\text{late}, \text{no rain}\right) &= \frac{\mathrm{P}\left(\text{late}, \text{no rain}, \text{problem}\right)}{\mathrm{P}\left(\text{late}, \text{no rain}\right)} && (1.64)\\
&= \frac{\mathrm{P}\left(\text{late}|\text{no rain}, \text{problem}\right)\mathrm{P}\left(\text{no rain}\right)\mathrm{P}\left(\text{problem}\right)}{\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} && \text{(by the Chain Rule)}\\
&= \frac{0.5 \cdot 0.1}{0.125} = 0.4. && (1.65)
\end{aligned}$$

Since $\mathrm{P}\left(\text{problem}|\text{late}, \text{no rain}\right) \neq \mathrm{P}\left(\text{problem}|\text{late}\right)$ the events *mechanical problem* and *rain in NYC* are *not* conditionally independent given the event *plane is late*.

△