Chapter 1

Introduction to Probabilistic Modeling

Modeling the world around us and making predictions about the occurrence of events is a multidisciplinary endeavor standing on the solid foundations of probability theory, statistics, and computer science. Although intertwined in the process of modeling, these fields have relatively discernible roles and can be, to a degree, studied individually. Probability theory brings the mathematical infrastructure, firmly grounded in its axioms, for manipulating probabilities and equips us with a broad range of models with well-understood theoretical properties. Statistics contributes frameworks to formulate inference and the process of narrowing down the model space based on the observed data and our experience in order to find, and then analyze, solutions. Computer science provides us with theories, algorithms, and software to manage the data, compute the solutions, and study the relationship between solutions and available resources (time, space, computer architecture, etc.). As such, these three disciplines form the core quantitative framework for all of empirical science and beyond.

Probability theory and statistics have a relatively long history; the formal origins of both can be traced to the 17th century. Probability theory emerged out of efforts to understand games of chance and gambling. The correspondence between Blaise Pascal and Pierre de Fermat in 1654 serves as the oldest record of modern probability theory. Statistics, on the other hand, originated from data collection initiatives and attempts to understand trends in the society (e.g., manufacturing, mortality causes, value of land) and political affairs (e.g., public revenues, taxation, armies). The two disciplines started to merge in the 18th century with the use of data for inferential purposes in astronomy, geography, and social sciences. The increased complexity of models and availability of data in the 19th century emphasized the importance of computing machines. This contributed to establishing the foundations of the field of computer science in the 20th century, which is generally attributed to the introduction of the von Neumann architecture and formalization of the concept of an algorithm. The convergence of the three disciplines has now reached the status of a principled theory of probabilistic inference with widespread applications in science, business, medicine, military, political campaigns, etc. Interestingly, various other disciplines have also contributed to the core of probabilistic modeling. Concepts such as a Boltzmann distribution, a genetic algorithm, or a neural network illustrate the influence of physics, biology, psychology, and engineering.

We will refer to the process of modeling, inference, and decision making based on probabilistic models as *probabilistic reasoning* or reasoning under uncertainty. Some form of reasoning under uncertainty is a necessary component of everyday life. When driving, for example, we often make decisions based on our expectations about which way would be best to take. While these situations do not usually involve an explicit use of probabilities and probabilistic models, an intelligent driverless car such as Google Chauffeur must make use of them. And so must a spam detection software in an email client, a credit card fraud detection system, or an algorithm that infers whether a particular genetic mutation will result in disease. Therefore, we first need to understand the concept of probability and then introduce a formal theory to incorporate evidence (e.g., data collected from instruments) in order to make good decisions in a range of situations. At a basic level, probabilities are used to quantify the chance of the occurrence of events. As Jacob Bernoulli brilliantly put it in his work *The Art of Conjecturing* (1713), "[p]robability, [...] is the degree of certainty, and it differs from the latter as a part differs from the whole". He later adds, "To make a conjecture [prediction] about something is the same as to measure its probability. Therefore, we define the art of conjecturing [science of prediction] or stochastics, as the art of measuring probabilities of things as accurately as possible, to the end that, in judgements and actions, we may always choose or follow that which has been found to be better, more satisfactory, safer, or more carefully considered." The techniques of probabilistic modeling formalize many intuitive concepts. They provide toolkits for rigorous mathematical analysis and inference, often in the presence of evidence, about events influenced by factors that we either do not fully understand or have no control of.

To give a quick insight into the concept of uncertainty and modeling, consider rolling a fair six-sided die. We could accurately predict, or so we think, the outcome of a roll if we carefully incorporated the initial position, force, friction, shape defects, and other physical factors and then executed the experiment. But the physical laws may not be known, they can be difficult to incorporate or such actions may not be allowed by the rules of the experiment. Thus, it is practically useful to simply assume that each outcome is equally likely; in fact, if we rolled the die many times, we would indeed observe that each number is observed roughly equally. Assigning an equal chance (probability) to each outcome of the roll of a die provides an efficient and elegant way of modeling uncertainties inherent to the experiment.

Another, more realistic example in which collecting data provides a basis for simple probabilistic modeling is a situation of driving to work every day and predicting how long it will take us to reach the destination tomorrow. If we recorded the "time to work" for a few months we would observe that trips generally took different times depending on many internal (e.g., preferred speed for the day) and also external factors (e.g., weather, road works, encountering a slow driver). While these events, if known, could be used to predict the exact duration of the commute, it is unrealistic to expect to have full information—rather we have *partial observability.* Therefore, it is useful to provide ways of aggregating external factors via collecting data over a period of time and providing the distribution of the commute time. Such a distribution, in the absence of other information, would then facilitate reasoning about events such as making it on time to an important meeting at 9 am. One way of using the recorded data is to create histograms and calculate percentiles. Another would be to estimate the parameters of some mathematical function that fits the data well. Both approaches are illustrated in Figure 1.1.

As the examples above suggest, the techniques of probabilistic modeling provide a formalism for dealing with repetitive "experiments" influenced by a number of external factors over which we have little control or knowledge. However, we shall see later that probabilities need not be assigned only to events that repeat, but to any event in general. As long as they are assigned according to the formal axioms of probability, we can make inferences because the mathematical formalism does not depend on how the probabilities are assigned.

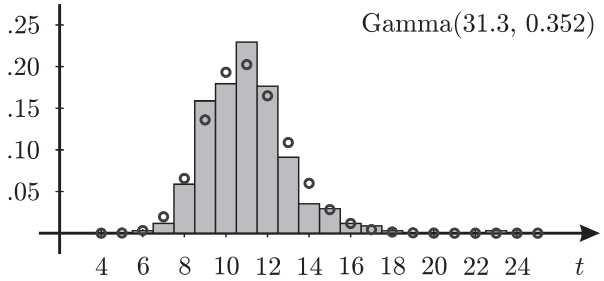


Figure 1.1: A histogram of recordings of the commute time (in minutes) to work. The data set contains 340 measurements collected over one year, for a distance of roughly 3.1 miles. The data was modeled using a gamma family of probability distributions, with the particular location and scale parameters estimated from the raw data. The values of the gamma distribution are shown as dark circles. Although it might seem that fitting the data using a gamma distribution brings little value in this one-dimensional situation, this approach is far superior on high-dimensional data where the number of bins in a multidimensional histogram can be orders of magnitude larger than the data set size.

This gives us an opportunity to incorporate our assumptions and existing knowledge into modeling, including the subjective assessments (beliefs) about occurrence of non-repetitive events. But let us start from the beginning.

1.1 Probability Theory

Probability theory can be seen as a branch of mathematics that deals with set functions. At the heart of probability theory is the concept of an *experiment*. An experiment can be the process of rolling a die, checking the temperature tomorrow or figuring out the location of one's keys. When carried out, each experiment has an *outcome*, which is an element "drawn" from a set of predefined options, potentially infinite in size. The outcome of a roll of a die is a number between one and six; the temperature tomorrow might be a real number; the outcome of the location of one's keys can be a discrete set of places such as a kitchen table, under a couch, in office etc. In many ways, the main goal of probabilistic modeling is to formulate a particular question or a hypothesis pertaining to the physical world as an experiment, collect the data, and then construct a model. Once a model is created, we can compute quantitative measures of sets of outcomes we are interested in and assess the confidence we should have in these measures.

1.1.1 Axioms of probability

We start by introducing the *axioms of probability*. Let the *sample space* (Ω) be a non-empty set of outcomes of the experiment and the *event space* ($\mathcal{A}$) be a non-empty set of subsets of Ω such that

1. $A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A}$
2. $A_1, A_2, \ldots \in \mathcal{A} \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

where A and all A_i's are *events*, and A^c is the complement of A; i.e., $A^c = \Omega - A$. If both conditions hold, $\mathcal{A}$ is called a sigma field, or sigma algebra, and is a set of so-called measurable events.[1] The tuple $(\Omega, \mathcal{A})$ is then called a *measurable space*.

It is important to emphasize that the definition of sigma field requires that $\mathcal{A}$ be closed under both finite and countably infinite number of basic set operations (union, intersection, complementation and set difference). The operations union and complementation are in the definition. For intersection, we can use De Morgan's laws: $\cup A_i = (\cap A_i^c)^c$ and $\cap A_i = (\cup A_i^c)^c$. Any intersection of sets in $\mathcal{A}$ must again be in $\mathcal{A}$ because $\mathcal{A}$ is closed under union and complementation. Therefore, a sigma field is also closed under intersection. Similarly for set difference, we can write $A_1 - A_2 = (A_1 \cap A_2)^c \cap A_1$, which then implies $A_1 - A_2 \in \mathcal{A}$. Because $\mathcal{A}$ is non-empty, we observe that all the above conditions imply that $\Omega \in \mathcal{A}$ and $\varnothing \in \mathcal{A}$, where $\varnothing$ is the empty set.

Let $(\Omega, \mathcal{A})$ be a measurable space. Any function $P : \mathcal{A} \to [0, 1]$ where

1. $P(\Omega) = 1$

2. $A_1, A_2, \ldots \in \mathcal{A},\ A_i \cap A_j = \varnothing\ \forall i, j \quad \Rightarrow \quad P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

is called a *probability measure* or a *probability distribution* and is said to satisfy the axioms of probability.[2] The tuple $(\Omega, \mathcal{A}, P)$ is called the *probability space*.

The beauty of these axioms lies in their compactness and elegance. Many useful expressions can be derived from the axioms of probability. For example, it is obvious that $P(\varnothing) = 0$ or $P(A^c) = 1 - P(A)$. Similarly, closure under infinite unions of disjoint sets (σ-additivity) implies finite closure (additivity), because the remaining sets can be set to the empty set $\varnothing$: $\forall A_1, A_2 \in \mathcal{A}$ with $A_1 \cap A_2 = \varnothing$, set $A_i = \varnothing$ for $i > 2$ to get $P(A_1 \cup A_2) = P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) = P(A_1) + P(A_2)$. Another formula that is particularly important can be derived by considering a *partition* of the sample space; i.e., a set of k non-overlapping sets $\{B_i\}_{i=1}^k$ such that $\Omega = \cup_{i=1}^k B_i$. That is, if A is any set in Ω and if $\{B_i\}_{i=1}^k$ is a partition of Ω it follows that

$$\begin{aligned} P(A) &= P(A \cap \Omega) \\ &= P\left(A \cap \left(\cup_{i=1}^k B_i\right)\right) \\ &= P\left(\cup_{i=1}^k (A \cap B_i)\right) \\ &= \textstyle\sum_{i=1}^k P\left(A \cap B_i\right), \end{aligned} \tag{1.1}$$

where the last line followed from the axioms of probability. We will refer to this expression as the *sum rule*. Another important expression, shown here without derivation, is that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. It holds for any $A, B \in \mathcal{A}$.

It is convenient to separately consider discrete (countable) and continuous (uncountable) sample spaces. A roll of a die draws numbers from a finite space $\Omega = \{1, 2, 3, 4, 5, 6\}$. For finite and other countable sample spaces (e.g., the set of integers $\mathbb{Z}$), $\mathcal{A}$ is usually the power set $\mathcal{P}(\Omega)$. An example of continuous sample space is the set of real numbers $\mathbb{R}$. As we

[1] This terminology is due to historical reasons; so though it sounds complex, one can simply think of a sigma field as the set of events to which we can assign probabilities.

[2] It seems intuitive that the second condition could be replaced with a union of finite sets (the simpler requirement of additivity rather than σ-additivity). However, for sigma fields, closure under *finite* unions may not result in closure under *infinite* unions.

shall see later, for uncountable spaces, $\mathcal{A}$ must be a proper subset of $\mathcal{P}(\Omega)$; i.e., $\mathcal{A} \subset \mathcal{P}(\Omega)$, because there exist sets over which one cannot integrate. Technically, sample spaces can also be mixed; e.g., $\Omega = [0,1] \cup \{2\}$ or $\Omega = [0,1] \times \{0,1\}$. Such spaces are often used in machine learning. The space $\Omega = [0,1] \times \{0,1\}$ is also said to be multidimensional because Ω is a Cartesian product of multiple sets, here $[0,1]$ and $\{0,1\}$. However, the main consideration at this stage will be the distinction between discrete and continuous sample spaces that will give rise to discrete and continuous probability distributions, respectively, and lead to a distinct mathematical treatment.

Owing to many constraints in defining the distribution function P, it is clear that it cannot be chosen arbitrarily. For example, if $\Omega = [0,1]$ and $P([0, \frac{1}{2})) = \frac{1}{2}$, we cannot freely assign $P([\frac{1}{2}, 1]) = \frac{1}{3}$ because probabilities of complement sets must sum to one. It turns out, in practice it is easier to define P indirectly, by selecting a probability mass function or a probability density function. These “helper” functions are defined directly on the sample space where we have fewer restrictions to be concerned with compared to the event space. We address these two ways of defining probability distributions next.

1.1.2 Probability mass functions

Let Ω be a discrete (finite or countably infinite) sample space and $\mathcal{A} = \mathcal{P}(\Omega)$. A function $p : \Omega \to [0,1]$ is called a *probability mass function* (pmf) if

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

The probability of any event $A \in \mathcal{A}$ is defined as

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

It is straightforward to verify that P satisfies the axioms of probability and, thus, is a probability distribution.

Example 1: Consider a roll of a fair six-sided die; i.e., $\Omega = \{1,2,3,4,5,6\}$, and the event space $\mathcal{A} = \mathcal{P}(\Omega)$. What is the probability that the outcome is a number greater than 4?

First, because the die is fair, we know that $p(\omega) = \frac{1}{6}$ for $\forall \omega \in \Omega$. Now, let A be an event in $\mathcal{A}$ that the outcome is greater than 4; i.e., $A = \{5,6\}$. Thus,

$$P(A) = \sum_{\omega \in A} p(\omega) = \frac{1}{3}.$$

It is important to note that P is defined on the elements of $\mathcal{A}$, whereas p is defined on the elements of Ω. That is, $P(\{1\}) = p(1)$, $P(\{2\}) = p(2)$, $P(\{1,2\}) = p(1) + p(2)$, etc. □

In discrete cases, $P(\{\omega\}) = p(\omega)$ for every $\omega \in \Omega$, and the probability of a set is always equal to the sum of probabilities of individual elements. We can define a discrete probability space by providing the tuple $(\Omega, \mathcal{A}, P)$, but it is often much simpler to define P indirectly by assuming that $\mathcal{A} = \mathcal{P}(\Omega)$ and providing a probability mass function p. In this case we say that the probability measure P is induced by a pmf p. In fact, we rarely define $(\Omega, \mathcal{A}, P)$ directly.

A few useful pmfs

Let us now look at some families of functions that are often used to induce discrete probability distributions. This is by no means a comprehensive review of the subject; we shall simply focus on a few basic concepts and will later introduce other distributions as needed. For simplicity, we will often refer to both pmfs and probability distributions they induce as distribution functions.

The *Bernoulli distribution* derives from the concept of a Bernoulli trial, an experiment that has two possible outcomes: success and failure. In a Bernoulli trial, a success occurs with probability α and, thus, failure occurs with probability $1 - \alpha$. A toss of a coin (heads/tails), a basketball game (win/loss), or a roll of a die (even/odd) can all be seen as Bernoulli trials. We model this distribution by setting a sample space to two elements and defining the probability of one of them as α. More specifically, $\Omega = \{\text{success}, \text{failure}\}$ and

$$p(\omega) = \begin{cases} \alpha & \omega = \text{success} \\ 1 - \alpha & \omega = \text{failure} \end{cases}$$

where $\alpha \in (0, 1)$ is a parameter. If we take instead that $\Omega = \{0, 1\}$, we can compactly write the Bernoulli distribution as $p(k) = \alpha^k \cdot (1 - \alpha)^{1-k}$ for $k \in \Omega$. Here we replaced ω with k, which is a more common notation when the sample space is comprised of integers.

To be precise, the Bernoulli distribution as presented above is actually a family of discrete probability distributions, one for each α. We shall refer to each such distribution as Bernoulli(α). However, we do not need to concern ourselves with semantics because the correct interpretation of a family vs. individual distributions should be clear from the context.

The *Binomial distribution* is used to describe a sequence of n independent and identically distributed (i.i.d.) Bernoulli trials. At each value k in the sample space the distribution gives the probability that the success happened exactly k times out of n trials, where of course $0 \leq k \leq n$. More formally, given $\Omega = \{0, 1, \ldots, n\}$, for $\forall k \in \Omega$ the binomial pmf is defined as

$$p(k) = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k},$$

where $\alpha \in (0, 1)$, as before, is the parameter indicating the probability of success in a single trial. Here, the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

enumerates all ways in which one can pick k elements from a list of n elements (e.g., there are 3 different ways in which one can pick $k = 2$ elements from a group of $n = 3$ elements). We will refer to a binomial distribution with parameters n and α as Binomial(n, α). The experiment leading to a binomial distribution can be generalized to a situation with more than two possible outcomes. This experiment results in a multidimensional probability mass function (one dimension per possible outcome) called the multinomial distribution.

The *Poisson distribution* can be derived as a limit of the binomial distribution as $n \to \infty$ with a fixed expected number of successes (λ). Here, $\Omega = \{0, 1, \ldots\}$ and for $\forall k \in \Omega$

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where $\lambda \in (0, \infty)$ is a parameter (the relationship with the binomial distribution can be obtained by taking $\alpha = {}^{\lambda}\!/_{n}$). The Poisson distribution is often used to model counts of events occurring sequentially and independently but with a fixed average (λ) in a particular time interval. Unlike the previous two distributions, Poisson(λ) is defined over an infinite sample space, but still countable.

The *geometric distribution* is also used to model a sequence of independent Bernoulli trials with the probability of success α. At each point $k \in \Omega$, it gives the probability that the first success occurs exactly in the k-th trial. Here, $\Omega = \{1, 2, \ldots\}$ and for $\forall k \in \Omega$

$$p(k) = (1 - \alpha)^{k-1}\alpha,$$

where $\alpha \in (0, 1)$ is a parameter. The geometric distribution, Geometric(α), is defined over an infinite sample space; i.e., $\Omega = \mathbb{N}$.

For the *hypergeometric distribution*, consider a finite population of N elements of two types (e.g., success and failure), K of which are of one type (e.g., success). The experiment consists of drawing n elements, without replacement, from this population such that the elements remaining in the population are equiprobable in terms of being selected in the next draw. The probability of drawing k successes out of n trials can be described as

$$p(k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}},$$

where $0 \leq n \leq N$ and $k \leq n$. The hypergeometric distribution is intimately related to the binomial distribution where the elements are drawn with replacement ($\alpha = {}^{K}\!/_{N}$). There, the probability of drawing a success does not change in subsequent trials. We will refer to the hypergeometric distribution as Hypergeometric(n, N, K).

The *uniform distribution* for discrete sample spaces is defined over a finite set of outcomes each of which is equally likely to occur. Here we can set $\Omega = \{1, \ldots, n\}$; then for $\forall k \in \Omega$

$$p(k) = \frac{1}{n}.$$

The uniform distribution does not contain parameters; it is defined by the size of the sample space. We refer to this distribution as Uniform(n). We will see later that the uniform distribution can also be defined over finite intervals in continuous spaces.

All of the functions above satisfy the definition of a probability mass function, which we can verify by summing over all possible outcomes in the sample space. Four examples are shown in Figure 1.2.

In general, for discrete spaces, we can assign probabilities to outcomes quite freely. For example, one could have a table of 365 values between zero and one for the probability of a birthday falling on each day, as long as the probabilities sum to 1. We will see that for continuous spaces it is more difficult to define consistent probabilities and we will often restrict ourselves to a limited set of known distributions.

1.1.3 Probability density functions

We shall see soon that the treatment of continuous probability spaces is analogous to that of discrete spaces, with probability density functions replacing probability mass functions

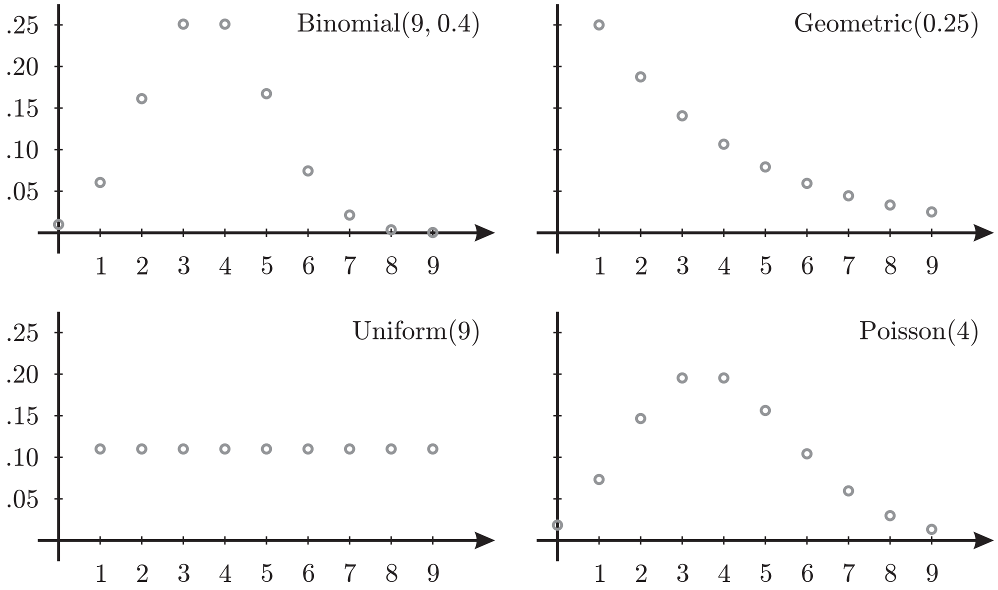


Figure 1.2: Four discrete probability mass functions.

and integrals replacing sums. Mathematically, however, there are fundamental differences between the two situations which we should keep in mind whenever working with continuous spaces. The main obstacle in generalizing the theory to uncountable sample spaces lies in addressing mathematical nuances involving infinitesimal calculus, the countable nature of a sigma field, and limitations of the definition of integrals. For example, there exist sets over which we cannot integrate and thus the set of events $\mathcal{A}$ cannot be the power set of any uncountable set (e.g., $\mathbb{R}$). It is therefore necessary to define an adequate event space which would be applicable to a vast majority of practically important situations.

To illustrate the treatment of continuous spaces, let us for simplicity take that $\Omega = \mathbb{R}$ and define the Borel field. The Borel field on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is a set that contains all points in $\mathbb{R}$, all open, semi-open and closed intervals in $\mathbb{R}$, as well as sets that can be obtained by a countable number of basic set operations on them. By definition, $\mathcal{B}(\mathbb{R})$ is a sigma field; $\mathcal{B}(\mathbb{R})$ is an uncountably infinite set, but still smaller than $\mathcal{P}(\mathbb{R})$. The construction of subsets of $\mathbb{R}$ that are not in $\mathcal{B}(\mathbb{R})$ is difficult and only of theoretical importance (e.g., Vitali sets), but nevertheless, the use of $\mathcal{P}(\mathbb{R})$ as the event space would lead to a flawed theory. Therefore, when discussing probability distributions over continuous sample spaces, we will usually take $\Omega = \mathbb{R}$ to be the sample space and implicitly consider $\mathcal{B}(\mathbb{R})$ to be the event space $\mathcal{A}$.

Let now Ω be a continuous sample space and $\mathcal{A} = \mathcal{B}(\Omega)$. A function $p : \Omega \to [0, \infty)$ is called a *probability density function* (pdf) if

$$\int_\Omega p(\omega) d\omega = 1.$$

The probability of an event $A \in \mathcal{B}(\Omega)$ is defined as

$$P(A) = \int_A p(\omega) d\omega.$$

There are a few mathematical nuances associated with this definition. First, interestingly, the standard Riemann integration does not work for some sets in the Borel field (e.g., how would you integrate over the set of rational or irrational numbers within $[0, 1]$ for any pdf?). For that reason, probability density functions are formally defined using Lebesgue integration which allows us to integrate over all sets in $\mathcal{B}(\Omega)$. Luckily, Riemann integration, when possible, provides identical results as Lebesgue's; thus, it will suffice to use Riemann integration in all situations of our interest.

Second, we mentioned before for pmfs that the probability of a singleton event $\{\omega\}$ is the value of the pmf at the sample point ω; i.e., $P(\{\omega\}) = p(\omega)$. In contrast, the value of a pdf at point ω is not a probability; it can actually be greater than 1. The probability at any single point, but also over any finite or countably infinite set is 0 (i.e., they constitute a set of measure zero). One way to think about the probabilities in continuous spaces is to look at small intervals $A = [x, x + \Delta x]$ as

$$\begin{aligned} P(A) &= \int_x^{x+\Delta x} p(\omega) d\omega \\ &\approx p(x)\Delta x. \end{aligned}$$

Here, a potentially large value of the density function is compensated by the small interval Δx to result in a number between 0 and 1.

A few useful pdfs

Some important probability density functions are reviewed below. As before, the sample space will be defined for each distribution and the Borel field will be implicitly assumed as the event space.

The *uniform distribution* is defined by an equal value of a probability density function over a finite interval in $\mathbb{R}$. Thus, for $\Omega = [a, b]$ the uniform probability density function $\forall \omega \in [a, b]$ is defined as

$$p(\omega) = \frac{1}{b - a}.$$

Note that one can also define Uniform(a, b) by taking $\Omega = \mathbb{R}$ and setting $p(\omega) = 0$ whenever ω is outside of $[a, b]$. This form is convenient because $\Omega = \mathbb{R}$ can then be used consistently for all one-dimensional probability distributions. When we do this, we will refer to the subset of $\mathbb{R}$ where $p(\omega) > 0$ as *support* of the density function.

The *exponential distribution* is defined over a set of non-negative numbers; i.e., $\Omega = [0, \infty)$. Its probability density function is

$$p(\omega) = \lambda e^{-\lambda\omega},$$

where $\lambda > 0$ is a parameter. As before, the sample space can be extended to all real numbers, in which case we would set $p(\omega) = 0$ for $\omega < 0$.

The *Gaussian distribution* or normal distribution is one of the most frequently used probability distributions. It is defined over $\Omega = \mathbb{R}$ as

$$p(\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\omega - \mu)^2}$$

Figure 1.3: Selection of a random number (x) from the unit interval $[0, 1]$.

with two parameters, $\mu \in \mathbb{R}$ and $\sigma > 0$. We will refer to this distribution as Gaussian(μ, σ^2) or $\mathcal{N}(\mu, \sigma^2)$. Both Gaussian and exponential distribution are members of a broader family of distributions called the exponential family. We will see a general definition of this family later.

The *lognormal distribution* is a modification of the normal distribution. Here, for $\Omega = (0, \infty)$ the lognormal density can be expressed as

$$p(\omega) = \frac{1}{\omega\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\ln \omega - \mu)^2},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are parameters. We will refer to this distribution as Lognormal(μ, σ^2) or $\ln \mathcal{N}(\mu, \sigma^2)$.

The *Gumbel distribution* belongs to the class of extreme value distributions. Its probability density function is defined on $\Omega = \mathbb{R}$ as

$$p(\omega) = \frac{1}{\beta} e^{-\frac{\omega - \alpha}{\beta}} e^{-e^{-\frac{\omega - \alpha}{\beta}}},$$

where $\alpha \in \mathbb{R}$ is the location parameter and $\beta > 0$ is the scale parameter. We will refer to this distribution as Gumbel(α, β).

The *Pareto distribution* is useful for modeling events with rare occurrences of extreme values. Its probability density function is defined on $\Omega = [\omega_{\text{min}}, \infty)$ as

$$p(\omega) = \frac{\alpha \omega_{\text{min}}}{\omega^{\alpha+1}},$$

where $\alpha > 0$ is a parameter and $\omega_{\text{min}} > 0$ is the minimum allowed value for ω. We will refer to the Pareto distribution as Pareto$(\alpha, \omega_{\text{min}})$. It leads to a scale-free property when $\alpha \in (0, 2]$.

Example 2: Consider selecting a number (x) between 0 and 1 uniformly randomly (Figure 1.3). What is the probability that the number is greater than $\frac{3}{4}$ or lower than $\frac{1}{4}$?

We know that $\Omega = [0, 1]$. We define an event of interest as $A = [0, \frac{1}{4}) \cup (\frac{3}{4}, 1]$ and calculate its probability as

$$\begin{aligned} P(A) &= \int_0^{1/4} d\omega + \int_{3/4}^{1} d\omega & \qquad \triangleright\ p(\omega) = \frac{1}{b - a} = 1 \\ &= \frac{1}{2}. \end{aligned}$$

Because the probability of any individual event in a continuous case is 0, there is no difference in integration if we consider open or closed intervals. □

1.1.4 Multidimensional distributions

It is often convenient to think of the sample space as a multidimensional space. In the discrete case, one can think of the sample space Ω as a multidimensional array or as a d-dimensional tensor (note: a matrix is a 2D tensor). That is, $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_d$, where Ω_i can be seen as the sample space along dimension i. Then, any function $p : \Omega_1 \times \Omega_2 \times \ldots \times \Omega_d \to [0, 1]$ is called a multidimensional probability mass function if

$$\sum_{\omega_1 \in \Omega_1} \cdots \sum_{\omega_d \in \Omega_d} p\left(\omega_1, \omega_2, \ldots, \omega_d\right) = 1.$$

One example of the multidimensional pmf is the multinomial distribution, which generalizes the binomial distribution to the case when the number of outcomes in any trial is a positive integer $d \geq 2$.

The *multinomial distribution* is used to model a sequence of n independent and identically distributed (i.i.d.) trials with d outcomes. At each point $(k_1, k_2, \ldots, k_d)$ in the sample space, the multinomial pmf gives the probability that the outcome 1 occurred k_1 times, outcome 2 occurred k_2 times, etc. Of course, $0 \leq k_i \leq n$ for $\forall i$ and $\sum_{i=1}^{d} k_i = n$. More formally, given the sample space $\Omega = \{0, 1, \ldots, n\}^d$, the multinomial pmf is defined as

$$p(k_1, k_2, \ldots, k_d) = \begin{cases} \binom{n}{k_1, k_2, \ldots, k_d} \alpha_1^{k_1} \alpha_2^{k_2} \ldots \alpha_d^{k_d} & k_1 + k_2 + \cdots + k_d = n \\ \\ 0 & \text{otherwise} \end{cases}$$

where α_i's are positive coefficients such that $\sum_{i=1}^{d} \alpha_i = 1$. That is, each coefficient α_i gives the probability of outcome i in any trial. The multinomial coefficient

$$\binom{n}{k_1, k_2, \ldots, k_d} = \frac{n!}{k_1! k_2! \cdots k_d!}$$

generalizes the binomial coefficient by enumerating all ways in which one can distribute n balls into d boxes such that the first box contains k_1 balls, the second box k_2 balls, etc. An experiment consisting of n tosses of a fair six-sided die and counting the number of occurrences of each number can be described by a multinomial distribution. Clearly, in this case $\alpha_i = 1/6$, for each $i \in \{1, 2, 3, 4, 5, 6\}$.

In the continuous case, we can think of the sample space as the d-dimensional Euclidean space; i.e., $\Omega = \mathbb{R}^d$ and an event space as $\mathcal{A} = \mathcal{B}(\mathbb{R})^d$. Then, the d-dimensional probability density function can be defined as any function $p : \mathbb{R}^d \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\omega_1, \omega_2, \ldots, \omega_d) d\omega_1 \cdots d\omega_d = 1.$$

The *multivariate Gaussian distribution* is a generalization of the Gaussian or normal distribution to the d-dimensional case, with $\Omega = \mathbb{R}^d$. It is defined as

$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu})\right),$$

with parameters $\boldsymbol{\mu} \in \mathbb{R}^d$ and a positive definite d-by-d matrix $\boldsymbol{\Sigma}$ ($|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$). We will refer to this distribution as Gaussian$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1.1.5 Conditional probabilities

Let $(\Omega, \mathcal{A}, P)$ be a probability space and B an event that already occurred. We are interested in the probability that event A also occurred; i.e., $P(A|B)$. The conditional probability is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{1.2}$$

where $P(B) > 0$. From this expression, which is sometimes referred to as *product rule*, we can now derive two important formulas. The first one is *Bayes' rule*

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

The second formula, referred to as the *chain rule*, applies to a collection of d events $\{A_i\}_{i=1}^{d}$ and can be derived by recursively applying the product rule. Then,

$$P(A_1 \cap A_2 \ldots \cap A_d) = P(A_1)P(A_2|A_1)\ldots P(A_d|A_1 \cap A_2 \ldots \cap A_{d-1}).$$

In some situations we refer to the probability $P(A)$ as *prior probability* because it quantifies the likelihood of occurrence of event A in absence of any other information or evidence. The probability $P(A|B)$ is referred to as *posterior probability* because it quantifies the uncertainty about A in the presence of additional information (event B). The probability $P(B)$ is also an unconditional (prior) probability but in this context can be thought of as the probability of observing evidence B. The product rule from Equation (1.2) has long history; it was first derived by Abraham de Moivre in 1718.

One way to think about conditional probabilities is to consider that the experiment has already been conducted, but that we do not know the outcome yet. For example, a fair die has been rolled and we are interested in an event that the outcome was 4; i.e., $A = \{4\}$. The prior probability of event A is $P(A) = \frac{1}{6}$. But imagine that someone had observed the experiment and told us that the number was even ($B = \{2, 4, 6\}$). The probability after hearing this news becomes $P(A|B) = \frac{1}{3}$. Proper estimation of posterior probabilities from data is central to statistical inference.

1.1.6 Independence of events

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Two events $A, B \in \mathcal{A}$ are defined as *independent* if

$$P(A \cap B) = P(A) \cdot P(B)$$

or, alternatively, if $P(A|B) = P(A)$ or $P(B|A) = P(B)$. More broadly, two or more events are (mutually or jointly) independent, if the probability of intersection of any group of events (of size two, three, etc.) can be expressed as the product of probabilities of individual events. For d events, there are $2^d - d - 1$ independence tests, one for each subset excluding the empty set and singletons.

It is important to distinguish between mutually exclusive events and independent events. Mutually exclusive events are in fact never independent because the knowledge that the outcome of the experiment belongs to event A excludes the possibility that it is in B (Figure 1.4). It is often difficult, and quite non-intuitive, to simply look at events and conclude

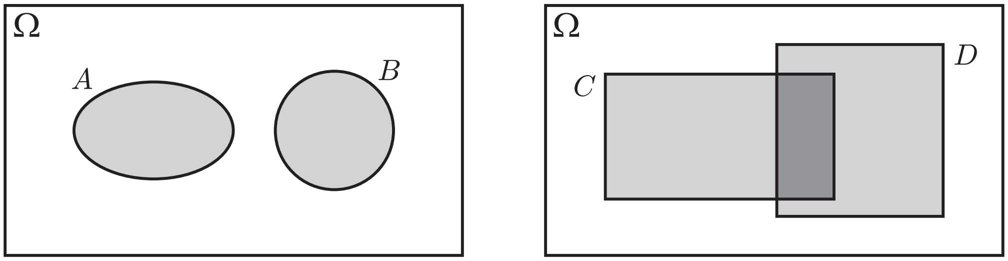


Figure 1.4: Visualization of dependent and independent events. Events A and B on the left are dependent because the occurrence of one excludes the occurrence of the other one. Events C and D on the right are independent. Each event occupies $^1/_4$ of the sample space Ω, while their intersection occupies $^1/_{16}$ of the sample space.

whether they are independent or not. One should (almost) always calculate $P(A \cap B)$ and $P(A) \cdot P(B)$ and numerically verify independence. Sometimes there may exist deep physical reasons why particular events are independent or assumed to be independent. In other occasions it may just be a numerical coincidence.

Let $(\Omega, \mathcal{A}, P)$ be a probability space and A, B, and C some events from $\mathcal{A}$. Events A and B are defined as *conditionally independent* given C if

$$P(A \cap B|C) = P(A|C) \cdot P(B|C)$$

or, alternatively, if $P(A|B \cap C) = P(A|C)$. Independence between events does not imply conditional independence and, likewise, conditional independence between events does not imply their independence. We shall see an example later.

1.1.7 Interpretation of probability

There are two opposing philosophical views of probability, an *objectivist* and a *subjectivist* one. Objectivists see probability as a concept rooted in reality. Their scientific method is based on the existence of an underlying true probability for an experiment or a hypothesis in question; this underlying probability then needs to be estimated from data. An objectivist is restricted by the known facts about reality (assuming these facts are agreed upon) and derives from them to estimate probabilities. On the other end of the spectrum is a purely subjectivist view in which probabilities represent an observer's *degree of belief* or *conviction* about the outcome of the experiment. A subjectivist is unrestricted by the agreed upon facts and can express any views about an experiment because probabilities are inherently related to one's perception. The good news is, this long-standing philosophical debate has almost no bearing on the use of probability theory in practice. No matter how probabilities are assigned, and it mostly happens through a combination of subjective and objective steps, the mechanics of probabilistic manipulations are the same and valid, as long as the assignments adhere to the axioms of probability.

1.2 Random Variables

Until now we operated on relatively simple sample spaces and produced measure functions over sets of outcomes. In many situations, however, we would like to use probabilistic modeling on sets (e.g., a group of people) where elements can be associated with various descriptors. For example, a person may be associated with his/her age, height, citizenship, IQ, or marital status and we may be interested in events related to such descriptors. In other situations, we may be interested in transformations of sample spaces such as those corresponding to digitizing an analog signal from a microphone into a set of integers based on some set of voltage thresholds. The mechanism of a *random variable* facilitates addressing all such situations in a simple, rigorous and unified manner.

A random variable is a variable that, from the observer's point of view, takes values non-deterministically, with generally different preferences for different outcomes. Mathematically, however, it is defined as function that maps one sample space into another, with a few technical caveats we will introduce later. Let us motivate the need for random variables. Consider a probability space $(\Omega, \mathcal{A}, P)$, where Ω is a set of people and let us investigate the probability that a randomly selected person $\omega \in \Omega$ is happy (we may assume we have a diagnostic method to assess any person's status). We start by defining an event A as

$$A = \{\omega \in \Omega : Status(\omega) = \text{happy}\}$$

and simply calculate the probability of this event. This is a perfectly legitimate approach, but it can be much simplified using the random variable mechanism. We first note that, technically, our diagnostic method corresponds to a function $Status : \Omega \to \mathcal{S}$ that maps the sample space Ω to a new binary sample space $\mathcal{S} = \{\text{happy, not happy}\}$. More interestingly, our approach also maps the probability distribution P to a new probability distribution P_{Status} that is defined on some sigma algebra of $\mathcal{S}$; say, $\mathcal{A}_{Status}$ (for the mapping to work as expected, $\mathcal{A}_{Status}$ has to be the power set of $\mathcal{S}$). We can now see that we can calculate $P_{Status}(\{\text{happy}\})$ from the probability of the aforementioned event A; i.e., $P_{Status}(\{\text{happy}\}) = P(A)$. This is a cluttered notation so we may wish to simplify it by using $P(Status = \text{happy})$, where $Status$ is a "random variable".

We will use capital letters $X, Y, \ldots$ to denote random variables (such as $Status$) and lowercase letters $x, y, \ldots$ to indicate elements (such as "happy") of the new spaces $\mathcal{X}, \mathcal{Y} \ldots$ Generally, we will write probabilities as $P(X = x)$, which is a notational relaxation of $P(\{\omega : X(\omega) = x\})$, or $P(X \leq x)$ for $P(\{\omega : X(\omega) \leq x\})$ when the co-domain $\mathcal{X}$ is continuous. We will also refer to the corresponding probability mass or density functions as $p(x)$ or $p_X(x)$ when we need to be more explicit about the random variable. This will indeed happen when x takes a particular value; say, for $x = 1$, we will write $p_X(1)$. Before we proceed to formally define random variables, we shall look at two illustrative examples.

Example 3: Consecutive tosses of a fair coin. Consider a process of three coin tosses and two random variables, X and Y, defined on the sample space. We define X as the number of heads in the first toss and Y as the number of heads over all three tosses. Our goal is to find the probability spaces that are created after the transformations.

First, $\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\}$ and

ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X(\omega)$	1	1	1	1	0	0	0	0
$Y(\omega)$	3	2	2	1	2	1	1	0

Let us only focus on variable Y. Clearly, $Y : \Omega \to \{0, 1, 2, 3\}$ but we also need to find $\mathcal{A}_Y$ and P_Y. To calculate P_Y, a simple approach is to find its pmf $p(y)$. For example, let us calculate $p_Y(2) = P_Y(\{2\})$ as

$$\begin{aligned} P_Y(\{2\}) &= P(Y = 2) \\ &= P(\{\omega : Y(\omega) = 2\}) \\ &= P(\{\text{HHT}, \text{HTH}, \text{THH}\}) \\ &= \frac{3}{8}, \end{aligned}$$

because of the uniform distribution in the original space $(\Omega, \mathcal{A}, P)$. In a similar way, we can calculate that $P(Y = 0) = P(Y = 3) = {}^1\!/_8$, and that $P(Y = 1) = {}^3\!/_8$. In this example, we took that $\mathcal{A} = \mathcal{P}(\Omega)$ and $\mathcal{A}_Y = \mathcal{P}(\mathcal{Y})$. As a final note, we mention that all the randomness is defined in the original probability space $(\Omega, \mathcal{A}, P)$ and that the new probability space $(\mathcal{Y}, \mathcal{A}_Y, P_Y)$ simply inherits it through a deterministic transformation.

□

Example 4: Quantization. Consider $(\Omega, \mathcal{A}, P)$ where $\Omega = [0, 1]$, $\mathcal{A} = \mathcal{B}(\Omega)$, and P is induced by a uniform pdf. Define $X : \Omega \to \{0, 1\}$ as

$$X(\omega) = \begin{cases} 0 & \omega \leq 0.5 \\ 1 & \omega > 0.5 \end{cases}$$

and find the transformed probability space.

Technically, we have changed the sample space to $\mathcal{X} = \{0, 1\}$. For an event space $\mathcal{A}_X = \mathcal{P}(\mathcal{X}) = \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$ we would like to understand the new probability distribution P_X. We have

$$\begin{aligned} p_X(0) &= P_X(\{0\}) \\ &= P(X = 0) \\ &= P(\{\omega : \omega \in [0, 0.5]\}) \\ &= \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} p_X(1) &= P_X(\{1\}) \\ &= P(X = 1) \\ &= P(\{\omega : \omega \in (0.5, 1]\}) \\ &= \frac{1}{2} \end{aligned}$$

From here we can easily see that $P_X(\{0, 1\}) = 1$ and $P_X(\varnothing) = 0$, and so P_X is indeed a probability distribution. Again, P_X is naturally defined using P. Thus, we have transformed the probability space $(\Omega, \mathcal{A}, P)$ into $(\mathcal{X}, \mathcal{A}_X, P_X)$.

□

The mapping from a continuous Ω to other continuous samples spaces is slightly more complicated and will be considered later.

1.2.1 Formal definition of random variable

We now formally define a random variable. Given a probability space $(\Omega, \mathcal{A}, P)$, a random variable X is a function $X : \Omega \to \mathcal{X}$ such that for every $A \in \mathcal{B}(\mathcal{X})$ it holds that $\{\omega : X(\omega) \in A\} \in \mathcal{A}$. It follows that

$$P_X(A) = P(\{\omega : X(\omega) \in A\}).$$

It is important to mention that, by default, we defined the event space of a random variable to be the Borel field of $\mathcal{X}$. This is convenient because a Borel field of a countable set Ω is its power set. Thus, we are working with the largest possible event spaces for both discrete and continuous random variables.

Consider now a *discrete random variable* X defined on $(\Omega, \mathcal{A}, P)$. As we can see from the previous examples, the probability distribution for X can be found as

$$\begin{aligned} p(x) &= P_X(\{x\}) \\ &= P(\{\omega : X(\omega) = x\}) \end{aligned}$$

for $\forall x \in \mathcal{X}$. The probability of an event A can be found as

$$\begin{aligned} P_X(A) &= P(\{\omega : X(\omega) \in A\}) \\ &= \sum_{x \in A} p(x) \end{aligned}$$

for $\forall A \subseteq \mathcal{X}$.

The case of *continuous random variables* is more complicated, but reduces to an approach that is similar to that of discrete random variables. Here we first define a *cumulative distribution function* (cdf) as

$$\begin{aligned} F_X(t) &= P_X(\{x : x \leq t\}) \\ &= P(\{\omega : X(\omega) \leq t\}) \\ &= P(X \leq t), \end{aligned}$$

where $P(X \leq t)$, as before, presents a minor abuse of notation. If the cumulative distribution function is differentiable, the probability density function of a continuous random variable is defined as

$$p(x) = \left.\frac{dF_X(t)}{dt}\right|_{t=x}.$$

Alternatively, if $p(x)$ exists, then

$$F_X(t) = \int_{-\infty}^{t} p(x)\, dx,$$

for each $t \in \mathbb{R}$. Our focus will be exclusively on random variables that have their probability density functions; however, for a more general view, we should always keep in mind “if one exists” when referring to pdfs.

The probability that a random variable will take a value from interval $(a, b]$ can now be calculated as

$$\begin{aligned} P_X((a,b]) &= P(a < X \leq b) \\ &= \int_a^b p\,(x)\,dx \\ &= F_X\,(b) - F_X\,(a)\,, \end{aligned}$$

which follows from the properties of integration.

Suppose now that the random variable X transforms a probability space $(\Omega, \mathcal{A}, P)$ into $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_X)$. To describe the resulting probability space, we commonly use probability mass and density functions inducing P_X. For example, if P_X is induced by a Gaussian distribution with parameters μ and σ^2, we use

$$X : \mathcal{N}(\mu, \sigma^2) \qquad \text{or} \qquad X \sim \mathcal{N}(\mu, \sigma^2).$$

Both notations indicate that the probability density function for the random variable X is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

The Gaussian density implicitly defined that $\mathcal{X} = \mathbb{R}$. This point however is superficial because we can always extend the domain of a density function to $\mathbb{R}$ and set $p(x) = 0$ wherever the original function was not defined.

A group of d random variables $\{X_i\}_{i=1}^d$ defined on the same probability space $(\Omega, \mathcal{A}, P)$ is called a *random vector* or a multivariate (multidimensional) random variable. We have already seen an example of a random vector provided by random variables (X, Y) in Example 3. A generalization of a random vector to infinite sets is referred to as a *random process* or *stochastic process*; i.e., $\{X_i : i \in \mathcal{T}\}$, where $\mathcal{T}$ is an index set usually interpreted as a set of time indices. In the case of discrete time indices (e.g., $\mathcal{T} = \mathbb{N}$) the random process is called a discrete-time random process; otherwise (e.g., $\mathcal{T} = \mathbb{R}$) it is called a continuous-time random process. There are many models in machine learning that deal with temporally connected random variables (e.g., autoregressive models for time series, Markov chains, hidden Markov models, dynamic Bayesian networks). The language of random variables, through stochastic processes, nicely enables formalization of these models. Most of these notes, however, will deal with simpler settings only requiring (i.i.d.) multivariate random variables.

1.2.2 Joint and marginal distributions

Let us first look at two discrete random variables X and Y defined on the same probability space $(\Omega, \mathcal{A}, P)$. We define the *joint probability distribution* $p(x, y)$, or when needed $p_{XY}(x, y)$, of X and Y as

$$\begin{aligned} p(x,y) &= P(X = x, Y = y) \\ &= P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}). \end{aligned}$$

We can extend this to a d-dimensional random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ and define a multidimensional probability mass function as $p(\boldsymbol{x})$ or $p_{\boldsymbol{X}}(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ is a vector of values, such that each x_i is chosen from some $\mathcal{X}_i$.

A *marginal distribution* is defined for a subset of $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ by summing or integrating over the remaining variables. A marginal distribution $p\left(x_i\right)$ or $p_{X_i}\left(x_i\right)$ is defined as

$$p\left(x_i\right) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} p\left(x_1, \ldots, x_d\right),$$

where the j-th variable takes values from $\mathcal{X}_j$. The previous equation directly follows from Equation (1.1) and is also referred to as *sum rule*.

In the continuous case, we define a multidimensional cdf as

$$\begin{aligned} F_{\boldsymbol{X}}(\boldsymbol{t}) &= P_{\boldsymbol{X}}\left(\{\boldsymbol{x} : x_i \leq t_i, i = 1 \ldots d\}\right) \\ &= P\left(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_d \leq t_d\right) \end{aligned}$$

and the probability density function, if it exists, is defined as

$$p\left(\boldsymbol{x}\right) = \left.\frac{\partial^d}{\partial t_1 \cdots \partial t_d} F_{\boldsymbol{X}}\left(t_1, \ldots t_d\right)\right|_{\boldsymbol{t}=\boldsymbol{x}}.$$

The marginal density $p_{X_i}\left(x_i\right)$ is defined as

$$p\left(x_i\right) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{i-1}} \int_{\mathcal{X}_{i+1}} \cdots \int_{\mathcal{X}_d} p\left(\boldsymbol{x}\right) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d$$

If the sample space for each discrete random variable is seen as a countable subset of $\mathbb{R}$, then the probability space for any discrete or continuous d-dimensional random variable $\boldsymbol{X}$ can be defined as $(\mathbb{R}^d, \mathcal{B}(\mathbb{R})^d, P_{\boldsymbol{X}})$.

Example 5: Three tosses of a fair coin (again). Consider two random variables from Example 3 and calculate their probability spaces, joint and marginal distributions. Recall X is the number of heads in the first toss and Y is the number of heads over all three tosses.

A joint probability mass function $p(x, y) = P(X = x, Y = y)$ is shown below

		Y			
		0	1	2	3
X	0	1/8	1/4	1/8	0
	1	0	1/8	1/4	1/8

but let us step back for a moment and show how we can calculate it. Let us consider two sets $A = \{$HHH, HHT, HTH, HTT$\}$ and $B = \{$HHT, HTH, THH$\}$, corresponding to the events that the first toss was heads and that there were exactly two heads over the three tosses, respectively. Now, let us look at the probability of the intersection of A and B

$$\begin{aligned} P\left(A \cap B\right) &= P\left(\{\mathrm{HHT}, \mathrm{HTH}\}\right) \\ &= \frac{1}{4} \end{aligned}$$

We can represent the probability of the logical statement $X = 1 \wedge Y = 2$ as

$$\begin{aligned} p_{XY}\left(1, 2\right) &= P\left(X = 1, Y = 2\right) \\ &= P\left(A \cap B\right) \\ &= P\left(\{\mathrm{HHT}, \mathrm{HTH}\}\right) \\ &= \frac{1}{4}. \end{aligned}$$

The marginal probability distribution can be found in a straightforward way as

$$p\left(x\right)=\sum_{y\in\mathcal{Y}}p\left(x,y\right),$$

where $\mathcal{Y}=\left\{0,1,2,3\right\}$. Thus,

$$\begin{aligned}p_X\left(0\right)&=\sum_{y\in\mathcal{Y}}p_{XY}\left(0,y\right)\\&=\frac{1}{2}.\end{aligned}$$

We note for the end that in the discrete case we have $|\mathcal{X}|\cdot|\mathcal{Y}|-1$ free parameters (because the sum must equal 1) to fully describe the joint distribution $p(x,y)$. Asymptotically, this corresponds to an exponential growth of the number of entries in the table with the number of random variables (d). For example, if $|\mathcal{X}_i|=2$ for $\forall X_i$, there are 2^d-1 free elements in the joint probability distribution. Estimating such distributions from data is intractable and is one form of the *curse of dimensionality*.

□

1.2.3 Conditional distributions

The conditional probability distribution for two random variables X and Y, $p(y|x)$ or $p_{Y|X}(y|x)$, is defined as[3]

$$p(y|x)=\frac{p(x,y)}{p(x)},\tag{1.3}$$

where $p(x)>0$. For discrete spaces, we know that $p(x,y)$ and $p(x)$ are probabilities, which gives the interpretation that $p(y|x)=P(Y=y|X=x)$ as a direct consequence of the product rule from Equation (1.2). For continuous spaces, on the other hand, we shall consider this formula as a definition for mathematical convenience. Equation (1.3) now allows us to calculate the posterior probability of an event A, given some observation x, as

$$P(Y\in A|X=x)=\begin{cases}\sum_{y\in A}p(y|x) & Y:\text{discrete}\\ \\ \int_A p(y|x)dy & Y:\text{continuous}\end{cases}$$

Writing $p(x,y)=p(x|y)p(y)=p(y|x)p(x)$ is called the *product rule*. The extension to more than two variables is straightforward. We can write

$$p(x_d|x_1,\ldots,x_{d-1})=\frac{p(x_1,\ldots,x_d)}{p(x_1,\ldots,x_{d-1})}.$$

By a recursive application of the product rule, we obtain

$$p(x_1,\ldots,x_d)=p(x_1)\prod_{i=2}^{d}p(x_i|x_1,\ldots,x_{i-1})\tag{1.4}$$

[3] It is straightforward to verify that $p(y|x)$ sums (integrates) to 1 over all values $y\in\mathcal{Y}$, and thus satisfies the conditions of a probability mass (density) function.

which is referred to as the *chain rule* or *general product rule*. Using the product rule, we can derive *Bayes' rule*:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{1.5}$$

1.2.4 Independence of random variables

Two random variables are independent if their joint probability distribution can be expressed as

$$p\left(x, y\right) = p(x) \cdot p(y).$$

As before, d random variables are (mutually, jointly) independent if a joint probability distribution of any subset of variables can be expressed as a product of individual (marginal) probability distributions of its components.

Another, different, form of independence can be found even more frequently in probabilistic calculations. It represents independence between variables in the presence of some other random variable (evidence); e.g.,

$$p\left(x, y|z\right) = p(x|z) \cdot p(y|z)$$

and is referred to as *conditional independence*. Interestingly, the two forms of independence are unrelated; i.e., neither one implies the other. We show this in two simple examples from Figure 1.5.

1.2.5 Expectations and moments

Expectations of functions are defined as sums (or integrals) of function values weighted according to the probability mass (or density) function. Given a probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_X)$, we consider a function $f : \mathcal{X} \rightarrow \mathbb{C}$ and define its expectation function as

$$\mathbb{E}\left[f(X)\right] = \begin{cases} \sum_{x \in \mathcal{X}} f(x)p(x) & X : \text{discrete} \\ \\ \int_{\mathcal{X}} f(x)p(x)dx & X : \text{continuous} \end{cases}$$

Note that we use a capital X for the random variable (with $f(X)$ the random variable transformed by f) and lower case x when it is an instance (e.g., $p(x)$ is the probability of a specific outcome). The capital X in $\mathbb{E}\left[f(X)\right]$ also specifies that the summation (integration) is defined over $p(x)$; this will become more important later when we consider multiple random variables. It can happen that $\mathbb{E}\left[f(X)\right] = \pm\infty$; in such cases we say that the expectation does not exist or is not well-defined.[4] For $f(x) = x$, we have a standard expectation $\mathbb{E}\left[X\right] = \sum xp(x)$, or the mean value of X. Using $f(x) = x^k$ results in the k-th moment, $f(x) = \log {}^1\!/_{p(x)}$ gives the well-known entropy function $H(X)$, or differential

[4] There is sometimes disagreement on terminology, and some definitions allow the expected value to be infinite, which, for example, still allows the strong law of large numbers. In that setting, an expectation is not well-defined only if both left and right improper integrals are infinite. For our purposes, this is splitting hairs. An example of an expectation function that does not exist is the variance of the Pareto distribution when $\alpha \in (0, 2]$; thus the *scale-free* terminology for such distributions.

A. X and Y are independent, but not conditionally independent given Z

$P(X=1)$
a

X	$P(Y=1\|X)$
0	b
1	b

X	Y	$P(Z=1\|X,Y)$
0	0	c
0	1	$1-c$
1	0	$1-c$
1	1	c

$$P(Y=y|X=x) = P(Y=y)$$

for example,

$$P(Y=1|X=x) = b$$
$$P(Y=1) = b$$

$$P(Y=y|X=x, Z=z) \neq P(Y=y|Z=z)$$

for example,

$$P(Y=1|X=1, Z=1) = bc/(1-c-b(1-2c))$$
$$P(Y=1|Z=1) = b(1-c-a(1-2c))/d$$

where $d = P(Z=1)$

B. X and Z are conditionally independent given Y, but not independent

$P(X=1)$
a

X	$P(Y=1\|X)$
0	b
1	c

X	Y	$P(Z=1\|X,Y)$
0	0	d
0	1	e
1	0	d
1	1	e

$$P(Z=z|X=x) \neq P(Z=z)$$

for example,

$$P(Z=1|X=1) = d + ce - cd$$
$$P(Z=1) = d + (e-d)(a(c-b)+b)$$

$$P(Z=z|X=x, Y=y) = P(Z=z|Y=y)$$

for example,

$$P(Z=1|X=x, Y=1) = e$$
$$P(Z=1|Y=1) = e$$

Figure 1.5: Independence vs. conditional independence using probability distributions involving three binary random variables. Probability distributions are presented using factorization $p(x,y,z) = p(x)p(y|x)p(z|x,y)$, where all constants $a,b,c,d,e \in [0,1]$. (A) Variables X and Y are independent, but not conditionally independent given Z. When $c = 0$, $Z = X \oplus Y$, where $\oplus$ is an "exclusive or" operator. (B) Variables X and Z are conditionally independent given Y, but are not independent.

$f(x)$	Symbol	Name
x	$\mathbb{E}[X]$	Mean
$(x - \mathbb{E}[X])^2$	$V[X]$	Variance
x^k	$\mathbb{E}[X^k]$	k-th moment; $k \in \mathbb{N}$
$(x - \mathbb{E}[X])^k$	$\mathbb{E}[(X - \mathbb{E}[X])^k]$	k-th central moment; $k \in \mathbb{N}$
e^{tx}	$M_X(t)$	Moment generating function
e^{itx}	$\varphi_X(t)$	Characteristic function
$\log \frac{1}{p(x)}$	$H(X)$	(Differential) entropy
$\log \frac{p(x)}{q(x)}$	$D(p\|\|q)$	Kullback-Leibler divergence
$\left(\frac{\partial}{\partial\theta} \log p(x\|\theta)\right)^2$	$\mathcal{I}(\theta)$	Fisher information

Table 1.1: Some important expectation functions $\mathbb{E}[f(X)]$ for a random variable X described by its distribution $p(x)$. Function $q(x)$ in the definition of the Kullback-Leibler divergence is non-negative and must sum (integrate) to 1; i.e., it is a probability distribution itself. The Fisher information is defined for a family of probability distributions specified by a parameter θ. Note that the moment generating function may not exist for some distributions and all values of t; however, the characteristic function always exists, even when the density function does not.

entropy for continuous random variables, and $f(x) = (x - \mathbb{E}[X])^2$ gives the variance of a random variable X, denoted by $V[X]$. Interestingly, the probability of some event $A \subseteq \mathcal{X}$[5] can also be expressed in the form of expectation; i.e.,

$$P(A) = \mathbb{E}[1(X \in A)],$$

where

$$1(t) = \begin{cases} 1 & t \text{ is true} \\ 0 & t \text{ is false} \end{cases} \tag{1.6}$$

is an indicator function. With this, it is possible to express the cumulative distribution function as $F_X(t) = \mathbb{E}[1(X \in (-\infty, t])]$.

Function $f(x)$ inside the expectation can also be complex-valued. For example, $\varphi_X(t) = \mathbb{E}[e^{itX}]$, where i is the imaginary unit, defines the characteristic function of X. The characteristic function is closely related to the inverse Fourier transform of $p(x)$ and is useful in many forms of statistical inference. Several expectation functions are summarized in Table 1.1.

[5] This notation is a bit loose; we should say $A \in \mathcal{B}(\mathcal{X})$ instead of $A \subseteq \mathcal{X}$. We used it to re-emphasize (through this footnote) that some subsets of continuous sets are not measurable.

$f(x,y)$	Symbol	Name
$(x-\mathbb{E}[X])(y-\mathbb{E}[Y])$	$\mathrm{Cov}[X,Y]$	Covariance
$\frac{(x-\mathbb{E}[X])(y-\mathbb{E}[Y])}{\sqrt{V[X]V[Y]}}$	$\mathrm{Corr}[X,Y]$	Correlation
$\log\frac{p(x,y)}{p(x)p(y)}$	$I(X;Y)$	Mutual information
$\log\frac{1}{p(x,y)}$	$H(X,Y)$	Joint entropy
$\log\frac{1}{p(x\|y)}$	$H(X\|Y)$	Conditional entropy

Table 1.2: Some important expectation functions $\mathbb{E}[f(X,Y)]$ for two random variables, X and Y, described by their joint distribution $p(x,y)$. Mutual information is sometimes referred to as average mutual information.

Given two random variables X and Y and a specific value x assigned to X, we define the conditional expectation as

$$\mathbb{E}[f(Y)|x] = \begin{cases} \sum_{y\in\mathcal{Y}} f(y)p(y|x) & Y: \text{discrete} \\ \\ \int_{\mathcal{Y}} f(y)p(y|x)dy & Y: \text{continuous} \end{cases}$$

where $f:\mathcal{Y}\to\mathbb{C}$ is some function. Again, using $f(y)=y$ results in $\mathbb{E}[Y|x]=\sum yp(y|x)$ or $\mathbb{E}[Y|x]=\int yp(y|x)dy$. We shall see later that under some conditions $\mathbb{E}[Y|x]$ is referred to as the *regression function*. These types of integrals are often seen and evaluated in Bayesian statistics.

For two random variables X and Y we also define

$$\mathbb{E}[f(X,Y)] = \begin{cases} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f(x,y)p(x,y) & X,Y: \text{discrete} \\ \\ \int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)p(x,y)dxdy & X,Y: \text{continuous} \end{cases}$$

Expectations can also be defined over a single variable

$$\mathbb{E}[f(X,y)] = \begin{cases} \sum_{x\in\mathcal{X}} f(x,y)p(x) & X: \text{discrete} \\ \\ \int_{\mathcal{X}} f(x,y)p(x)dx & X: \text{continuous} \end{cases}$$

where $\mathbb{E}[f(X,y)]$ is now a function of y.

We define the covariance function as

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y],\end{aligned}$$

with $\mathrm{Cov}[X,X]=V[X]$ being the variance of the random variable X. Similarly, we define a correlation function as

$$\mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{V[X]\cdot V[Y]}},$$

which is simply a covariance function normalized by the product of standard deviations. Both covariance and correlation functions have wide applicability in statistics, machine learning, signal processing and many other disciplines. Several important expectations for two random variables are listed in Table 1.2.

Example 6: Three tosses of a fair coin (yet again). Consider two random variables from Examples 3 and 5, and calculate the expectation and variance for both X and Y. Then calculate $\mathbb{E}\left[Y|X=0\right]$.

We start by calculating $\mathbb{E}\left[X\right] = 0 \cdot p_X(0) + 1 \cdot p_X(1) = \frac{1}{2}$. Similarly,

$$\begin{aligned}\mathbb{E}\left[Y\right] &= \sum_{y=0}^{3} y \cdot p_Y(y)\\ &= p_Y(1) + 2p_Y(2) + 3p_Y(3)\\ &= \frac{3}{2}\end{aligned}$$

The conditional expectation can be found as

$$\begin{aligned}\mathbb{E}\left[Y|X=0\right] &= \sum_{y=0}^{3} y \cdot p_{Y|X}(y|0)\\ &= p_{Y|X}(1|0) + 2p_{Y|X}(2|0) + 3p_{Y|X}(3|0)\\ &= 1\end{aligned}$$

where $p(y|x) = p(x,y)/p(x)$.

□

In many situations we need to analyze more than two random variables. A simple two-dimensional summary of all pairwise covariance values involving d random variables $X_1, X_2, \ldots, X_d$ is called the covariance matrix. More formally, the covariance matrix is defined as

$$\boldsymbol{\Sigma} = \left[\Sigma_{ij}\right]_{i,j=1}^{d}$$

where

$$\begin{aligned}\Sigma_{ij} &= \mathrm{Cov}[X_i, X_j]\\ &= \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(X_j - \mathbb{E}\left[X_j\right]\right)\right]\end{aligned}$$

with the full matrix written as

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}]\\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top]\\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.\end{aligned}$$

Here, the diagonal elements of a $d \times d$ covariance matrix are individual variance values for each variable X_i and the off-diagonal elements are the covariance values between pairs of variables. The covariance matrix is symmetric and positive semi-definite; i.e., $\boldsymbol{\Sigma} \succeq 0$. We will discuss more about positive semi-definite matrices later in the notes.

Properties of expectations

Here we review, without proofs, some useful properties of expectations. We can generically consider multivariate random variables, $\boldsymbol{X} \in \mathbb{R}^d$ and $\boldsymbol{Y} \in \mathbb{R}^d$, for $d \in \mathbb{N}$, with univariate random variables as a special case. We consider the more general case because it will be useful to start thinking directly in terms of random vectors. For a constant $c \in \mathbb{R}$, it holds that:

1. $\mathbb{E}\left[c\boldsymbol{X}\right] = c\mathbb{E}\left[\boldsymbol{X}\right]$
2. $\mathbb{E}\left[\boldsymbol{X} + \boldsymbol{Y}\right] = \mathbb{E}\left[\boldsymbol{X}\right] + \mathbb{E}\left[\boldsymbol{Y}\right]$
3. $V\left[c\right] = 0$ $\qquad \triangleright$ the variance of a constant is zero
4. $\mathrm{V}[\boldsymbol{X}] \succeq 0$ (i.e., is positive semi-definite), where for $d = 1$, $\mathrm{V}[\boldsymbol{X}] \geq 0$ is a scalar. Note that $\mathrm{V}[\boldsymbol{X}]$ is shorthand for $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}]$.
5. $\mathrm{V}[c\boldsymbol{X}] = c^2\mathrm{V}[\boldsymbol{X}]$.
6. $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})^\top] = \mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]^\top$
7. $\mathrm{V}[\boldsymbol{X} + \boldsymbol{Y}] = \mathrm{V}[\boldsymbol{X}] + \mathrm{V}[\boldsymbol{Y}] + 2\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]$
8. $\mathrm{Cov}[\boldsymbol{X}_1 + \boldsymbol{X}_2 + \ldots + \boldsymbol{X}_m] = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}[\boldsymbol{X}_i, \boldsymbol{X}_j] = \sum_{i=1}^{m}\mathrm{V}[\boldsymbol{X}_i] + 2\sum_{1 \leq i < j \leq m}\mathrm{Cov}[\boldsymbol{X}_i, \boldsymbol{X}_j]$

In addition, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random variables of the same dimension, it holds that:

1. $\mathbb{E}\left[X_iY_j\right] = \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_j\right]$ for all i, j
2. $\mathrm{Cov}[\boldsymbol{X} + \boldsymbol{Y}] = \mathrm{V}[\boldsymbol{X}] + \mathrm{V}[\boldsymbol{Y}]$
3. $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}] = 0$.

1.2.6 Mixtures of distributions

In previous sections we saw that random variables are often described using particular families of probability distributions. This approach can be generalized by considering mixtures of distributions; i.e., linear combinations of other probability distributions. As before, we shall only consider random variables that have their probability mass or density functions.

Given a set of m probability distributions, $\{p_i(x)\}_{i=1}^m$, a finite mixture distribution function, or *mixture model*, $p(x)$ is defined as

$$p(x) = \sum_{i=1}^{m} w_i p_i(x), \tag{1.7}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ is a set of non-negative real numbers such that $\sum_{i=1}^m w_i = 1$. We refer to $\boldsymbol{w}$ as mixing coefficients or, sometimes, as mixing probabilities. A linear combination with such coefficients is called a convex combination. It is straightforward to verify that a function defined in this manner is indeed a probability distribution.

Here we will briefly look into the basic expectation functions of the mixture distribution. Suppose $\{X_i\}_{i=1}^m$ is a set of m random variables described by their respective probability distributions $\{p_{X_i}(x)\}_{i=1}^m$. Suppose also that a random variable X is described by a mixture distribution with coefficients $\boldsymbol{w}$ and probability distributions $\{p_{X_i}(x)\}_{i=1}^m$. Then, assuming continuous random variables defined on $\mathbb{R}$, the expectation function is given as

$$\begin{aligned}\mathbb{E}\left[f(X)\right] &= \int_{-\infty}^{+\infty} f(x)p_X(x)dx \\ &= \int_{-\infty}^{+\infty} f(x)\sum_{i=1}^{m} w_i p_{X_i}(x)dx \\ &= \sum_{i=1}^{m} w_i \int_{-\infty}^{+\infty} f(x)p_{X_i}(x)dx \\ &= \sum_{i=1}^{m} w_i \mathbb{E}[f(X_i)].\end{aligned}$$

We can now apply this formula to obtain the mean, when $f(x) = x$ and the variance, when $f(x) = (x - E[X])^2$, of the random variable X as

$$\mathbb{E}[X] = \sum_{i=1}^{m} w_i \mathbb{E}[X_i],$$

and

$$V[X] = \sum_{i=1}^{m} w_i V[X_i] + \sum_{i=1}^{m} w_i \left(\mathbb{E}[X_i] - \mathbb{E}[X]\right)^2,$$

respectively. A mixture distribution can also be defined for countably and uncountably infinite numbers of components. Such distributions, however, are rare in practice.

Example 7: Signal communications. Consider transmission of a single binary digital signal (bit) over a noisy communication channel shown in Figure 1.6. The magnitude of the signal X emitted by the source is equally likely to be 0 or 1 Volt. The signal is sent over a transmission line (e.g., radio communication, optical fiber, magnetic tape) in which a zero-mean normally distributed noise component Y is added to X. Derive the probability distribution of the signal $Z = X + Y$ that enters the receiver.

We will consider a slightly more general situation where X : Bernoulli(α) and Y : Gaussian(μ, σ^2). To find $p(z)$ we will use characteristic functions of random variables X, Y and Z, written as $\varphi_X(t) = \mathbb{E}[e^{itX}]$, $\varphi_Y(t) = \mathbb{E}[e^{itY}]$ and $\varphi_Z(t) = \mathbb{E}[e^{itZ}]$. Without derivation we write

$$\begin{aligned}\varphi_X(t) &= 1 - \alpha + \alpha e^{it} \\ \varphi_Y(t) &= e^{it\mu - \frac{\sigma^2 t^2}{2}}\end{aligned}$$

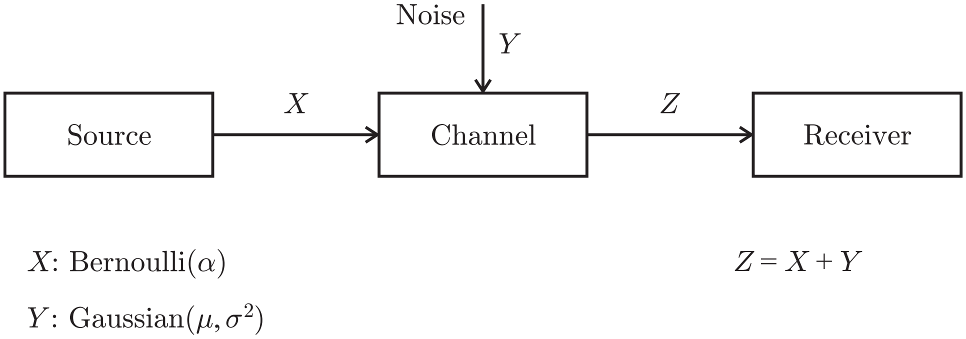


Figure 1.6: A digital signal communication system with additive noise.

and subsequently

$$\begin{aligned}\varphi_Z(t) &= \varphi_{X+Y}(t)\\ &= \varphi_X(t)\cdot\varphi_Y(t)\\ &= \left(1-\alpha+\alpha e^{it}\right)\cdot e^{it\mu-\frac{\sigma^2t^2}{2}}\\ &= \alpha e^{it(\mu+1)-\frac{\sigma^2t^2}{2}} + (1-\alpha)e^{it\mu-\frac{\sigma^2t^2}{2}}.\end{aligned}$$

By performing integration on $\varphi_Z(t)$ we can easily verify that

$$p(z) = \alpha\cdot\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z-\mu-1)^2} + (1-\alpha)\cdot\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z-\mu)^2},$$

which is a mixture of two normal distributions $\mathcal{N}(\mu+1,\sigma^2)$ and $\mathcal{N}(\mu,\sigma^2)$ with coefficients $w_1 = \alpha$ and $w_2 = 1-\alpha$, respectively. Observe that a convex combination of random variables $Z = w_1X + w_2Y$ does not imply $p_Z(x) = w_1p_X(x) + w_2p_Y(x)$.

□

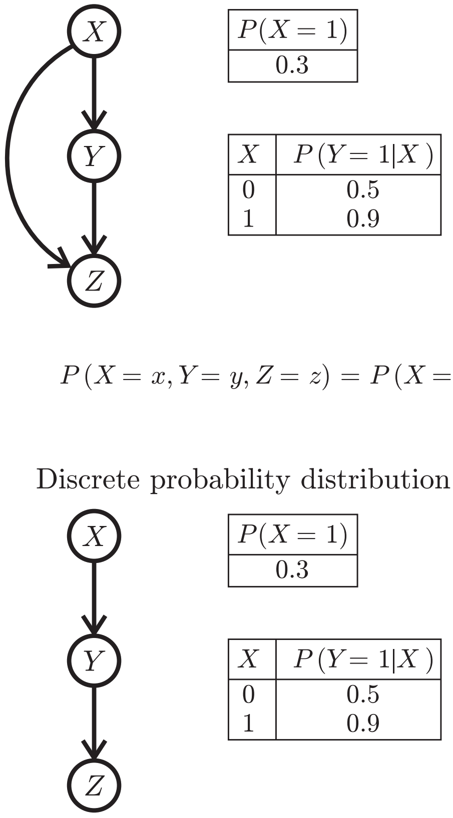
1.2.7 Graphical representation of probability distributions

We saw earlier that a joint probability distribution can be *factorized* using the chain rule from Equation (1.4). Such factorizations can be visualized using a directed graph representation, where nodes represent random variables and edges depict dependence. For example,

$$p(x,y,z) = p(x)p(y|x)p(z|x,y)$$

is shown in Figure 1.7A. Graphical representations of probability distributions using directed acyclic graphs, together with conditional probability distributions, are called *Bayesian networks* or *belief networks*. They facilitate interpretation as well as effective statistical inference.

Visualizing relationships between variables becomes particularly convenient when we want to understand and analyze conditional independence properties of variables. Figure 1.7B shows the same factorization of $p(x,y,z)$ where variable Z is independent of X given Y. To carefully determine conditional independence and dependence properties, however,



A. Discrete probability distribution without conditional independences

$P(X=1)$
0.3

X	$P(Y=1\|X)$
0	0.5
1	0.9

X	Y	$P(Z=1\|X,Y)$
0	0	0.3
0	1	0.1
1	0	0.7
1	1	0.4

$$P(X=x, Y=y, Z=z) = P(X=x)P(Y=y|X=x)P(Z=z|X=x, Y=y)$$

B. Discrete probability distribution; Z is conditionally independent of X given Y

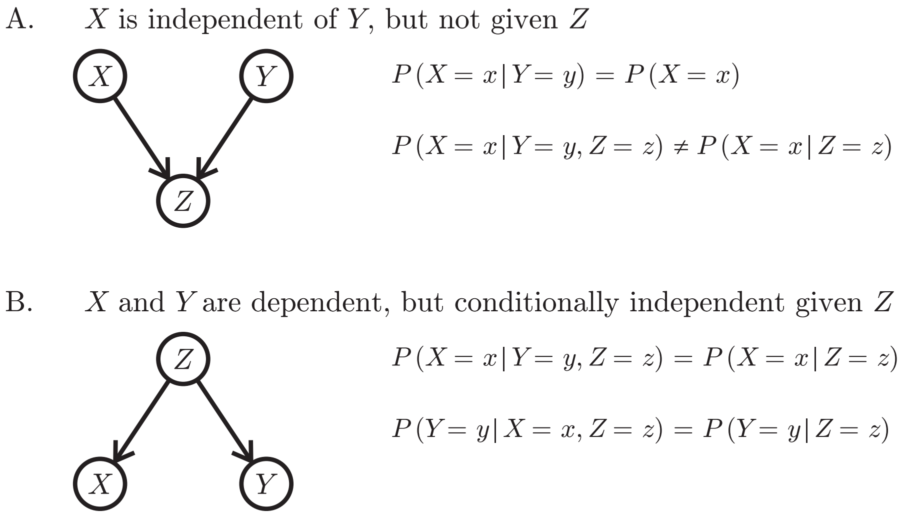
$P(X=1)$
0.3

X	$P(Y=1\|X)$
0	0.5
1	0.9

Y	$P(Z=1\|Y)$
0	0.2
1	0.7

$$P(X=x, Y=y, Z=z) = P(X=x)P(Y=y|X=x)P(Z=z|Y=y)$$

Figure 1.7: Bayesian network: graphical representation of two joint probability distributions for three discrete (binary) random variables (X, Y, Z) using directed acyclic graphs. The probability mass function $p(x, y, z)$ is defined over $\{0, 1\}^3$. (A) Full factorization; (B) Factorization that shows and ensures conditional independence between Z and X, given Y. Each node is associated with a conditional probability distribution. In discrete cases, these conditional distributions are referred to as conditional probability tables.



A. X is independent of Y, but not given Z

$$P(X = x | Y = y) = P(X = x)$$

$$P(X = x | Y = y, Z = z) \neq P(X = x | Z = z)$$

B. X and Y are dependent, but conditionally independent given Z

$$P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

$$P(Y = y | X = x, Z = z) = P(Y = y | Z = z)$$

Figure 1.8: Two examples of Bayesian networks. (A) A model where the lack of an edge between nodes does not indicate independence. Given information about Z, X and Y are actually dependent; i.e., they are conditionally dependent through Z. (B) A model where the lack of an edge between nodes does indicate independence. Given information about Z, X and Y are conditionally independent. We will see this representation later under Naive Bayes models.

one usually uses the *d-separation* rules for belief networks. Though often relationships are intuitive, sometimes dependence properties can get more complicated due to multiple relationships between nodes. For example, in Figure 1.8A, two nodes do not have an edge, but are conditionally dependent through another node. On the other hand, in Figure 1.8B, the absence of an edge does imply conditional independence. We will not further examine d-separation rules at this time; they can easily be found in any standard textbook on graphical models.

Belief networks have a simple, formal definition. Given a set of d random variables $\boldsymbol{X} = (X_1, \ldots, X_d)$, belief networks factorize the joint probability distribution of $\boldsymbol{X}$ as

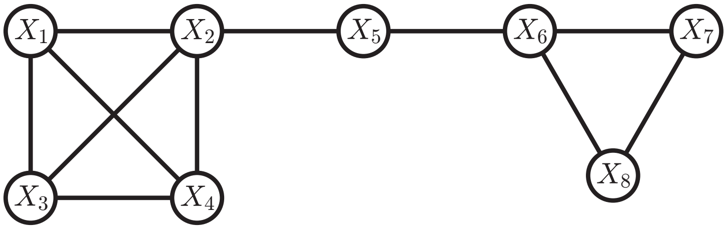
$$p(\boldsymbol{x}) = \prod_{i=1}^{d} p\left(x_i | \boldsymbol{x}_{\text{Parents}(X_i)}\right),$$

where Parents(X) denotes the immediate ancestors of node X in the graph. In Figure 1.7B, node Y is a parent of Z, but node X is not a parent of Z.

It is important to mention that there are multiple (how many?) ways of factorizing a distribution. For example, by reversing the order of variables $p(x, y, z)$ can be also factorized as

$$p(x, y, z) = p(z)p(y|z)p(x|y, z),$$

which has a different graphical representation and its own conditional probability distributions, yet the same joint probability distribution as the earlier factorization. Selecting a proper factorization and estimating the conditional probability distributions from data will be discussed in detail later.



$$\boldsymbol{X}_{C_1} = \{X_1, X_2, X_3, X_4\} \qquad \boldsymbol{X}_{C_3} = \{X_5, X_6\}$$

$$\boldsymbol{X}_{C_2} = \{X_2, X_5\} \qquad \boldsymbol{X}_{C_4} = \{X_6, X_7, X_8\}$$

Figure 1.9: Markov network: graphical representation of a probability distribution using maximum clique decomposition. Shown is a set of eight random variables with their interdependency structure and maximum clique decomposition (a clique is fully connected subgraph of a given graph). A decomposition into maximum cliques covers all vertices and edges in a graph with the minimum number of cliques. Here, the set of variables is decomposed into four maximal cliques $\mathcal{C} = \{C_1, C_2, C_3, C_4\}$.

Undirected graphs can also be used to factorize probability distributions. The main idea here is to decompose graphs into maximal cliques $\mathcal{C}$ (the smallest set of cliques that covers the graph) and express the distribution in the following form

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C),$$

where each $\psi_C(\boldsymbol{x}_C) \geq 0$ is called the clique potential function and

$$Z = \int_{\boldsymbol{x}} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C) d\boldsymbol{x},$$

is called the partition function, used strictly for normalization purposes. In contrast to conditional probability distributions in directed acyclic graphs, the clique potentials usually do not have conditional probability interpretations and, thus, normalization is necessary. One example of a maximum clique decomposition is shown in Figure 1.9.

The potential functions are typically taken to be strictly positive, $\psi_C(\boldsymbol{x}_C) > 0$, and expressed as

$$\psi_C(\boldsymbol{x}_C) = \exp\left(-E(\boldsymbol{x}_C)\right),$$

where $E(\boldsymbol{x}_C)$ is a user-specified energy function on the clique of random variables $\boldsymbol{X}_C$. This leads to the probability distribution of the following form

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \log \psi_C(\boldsymbol{x}_C) \right).$$

As formulated, this probability distribution is called the Boltzmann distribution or the Gibbs distribution.

The energy function $E(\boldsymbol{x})$ must be lower for values of $\boldsymbol{x}$ that are more likely. It also may involve parameters that are then estimated from the available training data. Of course, in a prediction problem, an undirected graph must be created to also involve the target variables, which were here considered to be a subset of $\boldsymbol{X}$.

Consider now any probability distribution over all possible configurations of the random vector $\boldsymbol{X}$ with its underlying graphical representation. If the following property

$$p\left(x_i|\boldsymbol{x}_{-X_i}\right) = p\left(x_i|\boldsymbol{x}_{N(X_i)}\right) \tag{1.8}$$

is satisfied, the probability distribution is referred to as *Markov network* or a *Markov random field*. In the equation above

$$\boldsymbol{X}_{-X_i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$$

and $N(X)$ is a set of random variables neighboring X in the graph; i.e., there exists an edge between X and every node in $N(X)$. The set of random variables in $N(X)$ is also called the Markov blanket of X.

It can be shown that every Gibbs distribution satisfies the property from Equation (1.8) and, conversely, that for every probability distribution for which Equation (1.8) holds can be represented as a Gibbs distribution with some choice of parameters. This equivalence of Gibbs distributions and Markov networks was established by the Hammersley-Clifford theorem.