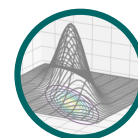


Probability and Distributions



Probability, loosely speaking, concerns the study of uncertainty. Probability can be thought of as the fraction of times an event occurs, or as a degree of belief about an event. We then would like to use this probability to measure the chance of something occurring in an experiment. As mentioned in Chapter 1, we often quantify uncertainty in the data, uncertainty in the machine learning model, and uncertainty in the predictions produced by the model. Quantifying uncertainty requires the idea of a *random variable*, which is a function that maps outcomes of random experiments to a set of properties that we are interested in. Associated with the random variable is a function that measures the probability that a particular outcome (or set of outcomes) will occur; this is called the *probability distribution*.

random variable

probability
distribution

Probability distributions are used as a building block for other concepts, such as probabilistic modeling (Section 8.3), graphical models (Section 8.4) and model selection (Section 8.5). In the next section, we present the three concepts that define a probability space (the sample space, the events and the probability of an event) and how they are related to a fourth concept called the random variable. The presentation is deliberately slightly hand wavy since a rigorous presentation may occlude the intuition behind the concepts. An outline of the concepts presented in this chapter are shown in Figure 6.1.

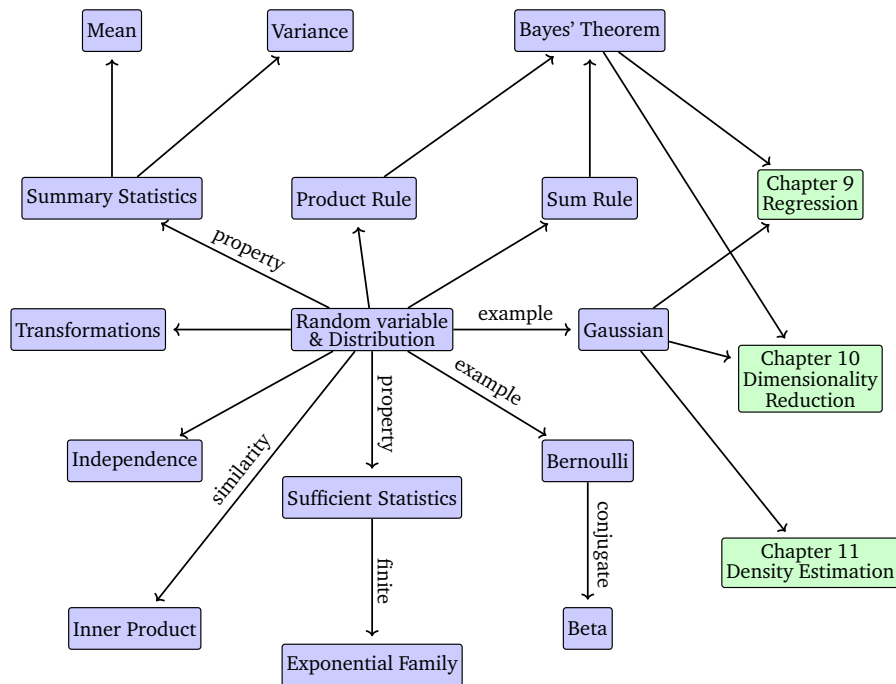
6.1 Construction of a Probability Space

The theory of probability aims at defining a mathematical structure to describe random outcomes of experiments. For example, when tossing a single coin, we cannot determine the outcome, but by doing a large number of coin tosses, we can observe a regularity in the average outcome. Using this mathematical structure of probability, the goal is to perform automated reasoning, and in this sense, probability generalizes logical reasoning (Jaynes, 2003).

6.1.1 Philosophical Issues

When constructing automated reasoning systems, classical Boolean logic does not allow us to express certain forms of plausible reasoning. Consider

Figure 6.1 A mind map of the concepts related to random variables and probability distributions, as described in this chapter.



“For plausible reasoning it is necessary to extend the discrete true and false values of truth to continuous plausibilities.” (Jaynes, 2003)

the following scenario: We observe that A is false. We find B becomes less plausible although no conclusion can be drawn from classical logic. We observe that B is true. It seems A becomes more plausible. We use this form of reasoning daily. We are waiting for a friend, and consider three possibilities; H_1 : she is on time, H_2 : she has been delayed by traffic and H_3 : she has been abducted by aliens. When we observe our friend is late, we must logically rule out H_1 . We also tend to consider H_2 to be more likely, though we are not logically required to do so. Finally, we may consider H_3 to be possible, but we continue to consider it quite unlikely. How do we conclude H_2 is the most plausible answer? Seen in this way, probability theory can be considered a generalization of Boolean logic. In the context of machine learning, it is often applied in this way to formalize the design of automated reasoning systems. Further arguments about how probability theory is the foundation of reasoning systems can be found in (Pearl, 1988).

The philosophical basis of probability and how it should be somehow related to what we think should be true (in the logical sense) was studied by Cox (Jaynes, 2003). Another way to think about it is that if we are precise about our common sense we end up constructing probabilities. E. T. Jaynes (1922–1998) identified three mathematical criteria, which must apply to all plausibilities:

- 1 The degrees of plausibility are represented by real numbers.
- 2 These numbers must be based on the rules of common sense.

- a) Consistency or non-contradiction: when the same result can be reached through different means, the same plausibility value must be found in all cases.
- b) Honesty: All available data must be taken into account.
- c) Reproducibility: If our state of knowledge about two problems are the same, then we must assign the same degree of plausibility to both of them.

The Cox-Jaynes's theorem proves these plausibilities to be sufficient to define the universal mathematical rules that apply to plausibility p , up to transformation by an arbitrary monotonic function. Crucially, these rules are the rules of probability.

Remark. In machine learning and statistics, there are two major interpretations of probability: the Bayesian and frequentist interpretations (Bishop, 2006; Efron and Hastie, 2016). The Bayesian interpretation uses probability to specify the degree of uncertainty that the user has about an event. It is sometimes referred to as “subjective probability” or “degree of belief”. The frequentist interpretation considers the relative frequencies of events of interest to the total number of events that occurred. The probability of an event is defined as the relative frequency of the event in the limit when one has infinite data. \diamond

Some machine learning texts on probabilistic models use lazy notation and jargon, which is confusing. This text is no exception. Multiple distinct concepts are all referred to as “probability distribution”, and the reader has to often disentangle the meaning from the context. One trick to help make sense of probability distributions is to check whether we are trying to model something categorical (a discrete random variable) or something continuous (a continuous random variable). The kinds of questions we tackle in machine learning are closely related to whether we are considering categorical or continuous models.

6.1.2 Probability and Random Variables

There are three distinct ideas, which are often confused when discussing probabilities. First is the idea of a probability space, which allows us to quantify the idea of a probability. However, we mostly do not work directly with this basic probability space. Instead we work with random variables (the second idea), which transfers the probability to a more convenient (often numerical) space. The third idea is the idea of a distribution or law associated with a random variable. We will introduce the first two ideas in this section and expand on the third idea in Section 6.2.

Modern probability is based on a set of axioms proposed by Kolmogorov (Grinstead and Snell, 1997; Jaynes, 2003) that introduce the three concepts of sample space, event space and probability measure. The probability space

	3278	models a real-world process (referred to as an experiment) with random
	3279	outcomes.
	3280	The sample space Ω
sample space	3281	The <i>sample space</i> is the set of all possible outcomes of the experiment,
	3282	usually denoted by Ω . For example, two successive coin tosses have
	3283	a sample space of $\{hh, tt, ht, th\}$, where “h” denotes “heads” and “t”
	3284	denotes “tails”.
	3285	The event space \mathcal{A}
event space	3286	The <i>event space</i> is the space of potential results of the experiment. A
	3287	subset A of the sample space Ω is in the event space \mathcal{A} if at the end
	3288	of the experiment we can observe whether a particular outcome $\omega \in \Omega$
	3289	is in A . The event space \mathcal{A} is obtained by considering the collection of
	3290	subsets of Ω , and for discrete probability distributions (Section 6.2.1)
	3291	\mathcal{A} is often the powerset of Ω .
	3292	The probability P
	3293	With each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the
	3294	probability or degree of belief that the event will occur. $P(A)$ is called
probability	3295	the <i>probability</i> of A .
	3296	The probability of a single event must lie in the interval $[0, 1]$, and the
	3297	total probability over all outcomes in the sample space Ω must be 1, i.e.,
	3298	$P(\Omega) = 1$. Given a probability space (Ω, \mathcal{A}, P) we are interested to use
	3299	it to model some real world phenomenon. In machine learning, we often
	3300	avoid explicitly referring to the probability space, but instead refer to
	3301	probabilities on quantities of interest, which we denote by \mathcal{T} . In this book
target space	3302	we refer to \mathcal{T} as the <i>target space</i> and refer to elements of \mathcal{T} as states. We
	3303	introduce a function $X : \Omega \rightarrow \mathcal{T}$ that takes an element of Ω (an event)
	3304	and returns a particular quantity of interest x , a value in \mathcal{T} . This associa-
random variable	3305	tion or mapping from Ω to \mathcal{T} is called a <i>random variable</i> . For example, in
	3306	the case of tossing two coins and counting the number of heads, a random
	3307	variable X maps to the three possible outcomes: $X(hh) = 2$, $X(ht) = 1$,
	3308	$X(th) = 1$ and $X(tt) = 0$. In this particular case $\mathcal{T} = \{0, 1, 2\}$, and it
	3309	is the probabilities on elements of \mathcal{T} that we are interested in. The name
	3310	“random variable” is a great source of misunderstanding as it is neither
	3311	random nor is it a variable. It is a function. For a finite sample space Ω
	3312	and finite \mathcal{T} , the function corresponding to a random variable is essen-
	3313	tially a lookup table. For any subset $S \subseteq \mathcal{T}$ we associate $P_X(S) \in [0, 1]$
	3314	(the probability) to a particular event occurring corresponding to the ran-
	3315	dom variable X . Example 6.1 provides a concrete illustration of the above
	3316	terminology.
	3317	<i>Remark.</i> The sample space Ω above unfortunately is referred to by dif-
	3318	ferent names in different books. Another common name for Ω is “state
	3319	space” (Jacod and Protter, 2004), but state space is sometimes reserved
	3320	for referring to states in a dynamical system (Hasselblatt and Katok, 2003).

3321 Other names sometimes used to describe Ω are: “sample description space”,
 3322 “possibility space” and “event space”. \diamond

Example 6.1

We assume that the reader is already familiar with computing probabilities of intersections and unions of sets of events. A more gentle introduction to probability with many examples can be found in Chapter 2 of Walpole et al. (2011).

Consider a statistical experiment where we model a funfair game consisting of drawing two coins from a bag (with replacement). There are coins from USA (denoted as \$) and UK (denoted as £) in the bag, and since we draw two coins from the bag, there are four outcomes in total. The state space or sample space Ω of this experiment is then (\$, \$), (\$, £), (£, \$), (£, £). Let us assume that the composition of the bag of coins is such that a draw returns at random a \$ with probability 0.3.

The event we are interested in is the total number of times the repeated draw returns \$. Let us define a random variable X that maps the sample space Ω to \mathcal{T} , that denotes the number of times we draw \$ out of the bag. We can see from the above sample space we can get zero \$, one \$ or two \$s, and therefore $\mathcal{T} = \{0, 1, 2\}$. The random variable X (a function or lookup table) can be represented as a table like below

$$X((\$,\$)) = 2 \quad (6.1)$$

$$X((\$,\pounds)) = 1 \quad (6.2)$$

$$X((\pounds,\$)) = 1 \quad (6.3)$$

$$X((\pounds,\pounds)) = 0. \quad (6.4)$$

Since we return the first coin we draw before drawing the second, this implies that the two draws are independent of each other, which we will discuss in Section 6.4.5. Note that there are two experimental outcomes, which map to the same event, where only one of the draws return \$. Therefore, the probability mass function (Section 6.2.1) of X is given by the calculations below

$$\begin{aligned} P(X = 2) &= P((\$,\$)) \\ &= P(\$) \cdot P(\$) \\ &= 0.3 \cdot 0.3 = 0.09 \end{aligned} \quad (6.5)$$

$$\begin{aligned} P(X = 1) &= P((\$,\pounds) \cup (\pounds,\$)) \\ &= P((\$,\pounds)) + P((\pounds,\$)) \\ &= 0.3 \cdot (1 - 0.3) + (1 - 0.3) \cdot 0.3 = 0.42 \end{aligned} \quad (6.6)$$

$$\begin{aligned} P(X = 0) &= P((\pounds,\pounds)) \\ &= P(\pounds) \cdot P(\pounds) \end{aligned}$$

This toy example is essentially a biased coin flip example.

$$= (1 - 0.3) \cdot (1 - 0.3) = 0.49. \quad (6.7)$$

In the above calculation, we equated two different concepts, the probability of the output of X and the probability of the samples in Ω . For example, in (6.7) we say $P(X = 0) = P((\mathcal{L}, \mathcal{L}))$. Consider the random variable $X : \Omega \rightarrow \mathcal{T}$ and a subset $S \subseteq \mathcal{T}$ (for example a single element of \mathcal{T} , such as the outcome that one head is obtained when tossing two coins). Let $X^{-1}(S)$ be the pre-image of S by X , i.e., the set of elements of Ω that map to S under X ; $\{\omega \in \Omega : X(\omega) \in S\}$. One way to understand the transformation of probability from events in Ω via the random variable X is to associate it with the probability of the pre-image of S (Jacod and Protter, 2004). For $S \subseteq \mathcal{T}$, we have the notation

$$P_X(S) = P(X \in S) = P(X^{-1}(S)) = P(\{\omega \in \Omega : X(\omega) \in S\}). \quad (6.8)$$

3323 The left-hand side of (6.8) is the probability of the set of possible outcomes
 3324 (e.g., number of \$ = 1) that we are interested in. Via the random variable
 3325 X , which maps states to outcomes, we see in the right-hand side of (6.8)
 3326 that this is the probability of the set of states (in Ω) that have the property
 3327 (e.g., \$ \mathcal{L} , \mathcal{L} \$). We say that a random variable X is distributed according
 3328 to a particular probability distribution P_X , which defines the probability
 3329 mapping between the event and the probability of the outcome of the
 3330 random variable. In other words, the function P_X or equivalently $P \circ X^{-1}$
 3331 is the *law* or *distribution* of random variable X .

law
distribution

3332 *Remark.* The target space, that is the range \mathcal{T} of the random variable X ,
 3333 is used to indicate the kind of probability space, i.e., a \mathcal{T} random variable.
 3334 When \mathcal{T} is finite or countably infinite, this is called a discrete random
 3335 variable (Section 6.2.1). For continuous random variables (Section 6.2.2)
 3336 we only consider $\mathcal{T} = \mathbb{R}$ or $\mathcal{T} = \mathbb{R}^D$. \diamond

6.1.3 Statistics

3337
 3338 Probability theory and statistics are often presented together, but they concern
 3339 different aspects of uncertainty. One way of contrasting them is by the
 3340 kinds of problems that are considered. Using probability we can consider
 3341 a model of some process where the underlying uncertainty is captured
 3342 by random variables, and we use the rules of probability to derive what
 3343 happens. Using statistics we observe that something has happened, and
 3344 try to figure out the underlying process that explains the observations. In
 3345 this sense, machine learning is close to statistics in its goals to construct a
 3346 model that adequately represents the process that generated the data. We
 3347 can use the rules of probability to obtain a “best fitting” model for some
 3348 data.

Another aspect of machine learning systems is that we are interested in generalization error (see Chapter 8). This means that we are actually interested in the performance of our system on instances that we will observe in future, which are not identical to the instances that we have seen so far. This analysis of future performance relies on probability and statistics, most of which is beyond what will be presented in this chapter. The interested reader is encouraged to look at the books by Shalev-Shwartz and Ben-David (2014); Boucheron et al. (2013). We will see more about statistics in Chapter 8.

6.2 Discrete and Continuous Probabilities

Let us focus our attention on ways to describe the probability of an event as introduced in Section 6.1. Depending on whether the target space is discrete or continuous the natural way to refer to distributions is different. When the outcome space \mathcal{T} is discrete, we can specify the probability that a random variable X takes a particular value $x \in \mathcal{T}$, denoted as $P(X = x)$. The expression $P(X = x)$ for a discrete random variable X is known as the *probability mass function*. When the outcome space \mathcal{T} is continuous, e.g., the real line \mathbb{R} , it is more natural to specify the probability that a random variable X is in an interval, denoted by $P(a \leq X \leq b)$ for $a < b$. By convention we specify the probability that a random variable X is less than a particular value x , denoted by $P(X \leq x)$. The expression $P(X \leq x)$ for a continuous random variable X is known as the *cumulative distribution function*. We will discuss continuous random variables in Section 6.2.2. We will revisit the nomenclature and contrast discrete and continuous random variables in Section 6.2.3.

Remark. We will use the phrase *univariate* distribution to refer to distributions of a single random variable (whose states are denoted by non-bold x). We will refer to distributions of more than one random variable as *multivariate* distributions, and will usually consider a vector of random variables (whose states are denoted by bold \mathbf{x}). \diamond

probability mass
function

cumulative
distribution function

univariate

multivariate

6.2.1 Discrete Probabilities

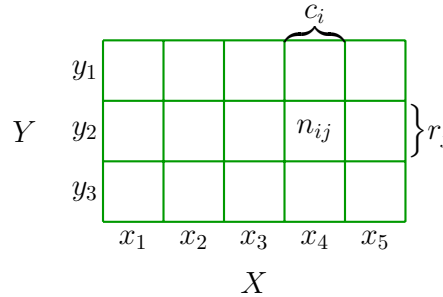
When the target space is discrete, we can imagine the probability distribution of multiple random variables as filling out a (multidimensional) array of numbers. Figure 6.2 shows an example. The target space of the joint probability is the Cartesian product of the target spaces of each of the random variables. We define the *joint probability* as the entry of both values jointly

joint probability

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}, \quad (6.9)$$

Figure 6.2

Visualization of a discrete bivariate probability mass function, with random variables X and Y . This diagram is adapted from Bishop (2006).



where n_{ij} is the number of events with state x_i and y_j and N the total number of events. The joint probability is the probability of the intersection of both events, that is $P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j)$. Figure 6.2 illustrates the *probability mass function* (pmf) of a discrete probability distribution. For two random variables X and Y , the probability that $X = x$ and $Y = y$ is (lazily) written as $p(x, y)$ and is called the joint probability. One can think of a probability as a function that takes state x and y and returns a real number, which is the reason we write $p(x, y)$. The *marginal probability* that X takes the value x irrespective of the value of random variable Y is (lazily) written as $p(x)$. We write $X \sim p(x)$ to denote that the random variable X is distributed according to $p(x)$. If we consider only the instances where $X = x$, then the fraction of instances (the *conditional probability*) for which $Y = y$ is written (lazily) as $p(y | x)$.

Example 6.2

Consider two random variables X and Y , where X has five possible states and Y has three possible states, as shown in Figure 6.2. We denote by n_{ij} the number of events with state $X = x_i$ and $Y = y_j$, and denote by N the total number of events. The value c_i is the sum of the individual frequencies for the i th column, that is $c_i = \sum_{j=1}^3 n_{ij}$. Similarly, the value r_j is the row sum, that is $r_j = \sum_{i=1}^5 n_{ij}$. Using these definitions, we can compactly express the distribution of X and Y .

The probability distribution of each random variable, the marginal probability, which can be seen as the sum over a row or column

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^3 n_{ij}}{N} \quad (6.10)$$

and

$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^5 n_{ij}}{N}, \quad (6.11)$$

where c_i and r_j are the i th column and j th row of the probability table, respectively. By convention for discrete random variables with a finite

number of events, we assume that probabilities sum up to one, that is

$$\sum_{i=1}^5 P(X = x_i) = 1 \quad \text{and} \quad \sum_{j=1}^3 P(Y = y_j) = 1. \quad (6.12)$$

The conditional probability is the fraction of a row or column in a particular cell. For example, the conditional probability of Y given X is

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}, \quad (6.13)$$

and the conditional probability of x given y is

$$P(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}. \quad (6.14)$$

In machine learning, we use discrete probability distributions to model *categorical variables*, i.e., variables that take a finite set of unordered values. They could be categorical features, such as the degree taken at university when used for predicting the salary of a person, or categorical labels, such as letters of the alphabet when doing handwriting recognition. Discrete distributions are also often used to construct probabilistic models that combine a finite number of continuous distributions (Chapter 11).

categorical variables

6.2.2 Continuous Probabilities

We consider real-valued random variables in this section, i.e., we consider target spaces that are intervals of the real line \mathbb{R} . In this book, we pretend that we can perform operations on real random variables as if we have discrete probability spaces with finite states. However, this simplification is not precise for two situations: when we repeat something infinitely often, and when we want to draw a point from an interval. The first situation arises when we discuss generalization errors in machine learning (Chapter 8). The second situation arises when we want to discuss continuous distributions, such as the Gaussian (Section 6.5). For our purposes, the lack of precision allows for a more brief introduction to probability.

Remark. In continuous spaces, there are two additional technicalities, which are counterintuitive. First, the set of all subsets (used to define the event space \mathcal{A} in Section 6.1) is not well behaved enough. \mathcal{A} needs to be restricted to behave well under set complements, set intersections and set unions. Second, the size of a set (which in discrete spaces can be obtained by counting the elements) turns out to be tricky. The size of a set is called its *measure*. For example, the cardinality of discrete sets, the length of an interval in \mathbb{R} and the volume of a region in \mathbb{R}^d are all measures. Sets that behave well under set operations and additionally have a topology are

measure

Borel σ -algebra 3420 called a *Borel σ -algebra*. Betancourt (2018) details a careful construction
 3421 of probability spaces from set theory without being bogged down in tech-
 3422 nicalities. For a more precise construction we refer to Jacod and Protter
 3423 (2004); Billingsley (1995). In this book, we consider real-valued random
 3424 variables with their corresponding Borel σ -algebra. We consider random
 3425 variables with values in \mathbb{R}^D to be a vector of real-valued random vari-
 3426 ables. \diamond

3427 **Definition 6.1** (Probability Density Function). A function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is
 probability density function 3428 called a *probability density function* (pdf) if

- 3429 1 $\forall \mathbf{x} \in \mathbb{R}^D : f(\mathbf{x}) \geq 0$
- 2 Its integral exists and

$$\int_{\mathbb{R}^D} f(\mathbf{x}) d\mathbf{x} = 1. \quad (6.15)$$

3430 For probability mass functions (pmf) of discrete random variables the in-
 3431 tegral in (6.15) is replaced with a sum (6.12).

Observe that the probability density function is any function f that is non-negative and integrates to one. We associate a random variable X with this function f by

$$P(a \leq X \leq b) = \int_a^b f(x) dx, \quad (6.16)$$

3432 where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}$ are outcomes of the continuous random vari-
 3433 able X . States $\mathbf{x} \in \mathbb{R}^D$ are defined analogously by considering a vector
 law 3434 of $x \in \mathbb{R}$. This association (6.16) is called the *law* or *distribution* of the
 3435 random variable X .
 3436 $P(X = x)$ is a set of measure zero.

Remark. In contrast to discrete random variables, the probability of a continuous random variable X taking a particular value $P(X = x)$ is zero. This is like trying to specify an interval in (6.16) where $a = b$. \diamond

3437 **Definition 6.2** (Cumulative Distribution Function). A *cumulative distribu-*
 cumulative 3438 *tion function* (cdf) of a multivariate real-valued random variable X with
 distribution function states $\mathbf{x} \in \mathbb{R}^D$ is given by

$$F_X(\mathbf{x}) = P(X_1 \leq x_1, \dots, X_D \leq x_D), \quad (6.17)$$

3439 where $X = [X_1, \dots, X_D]^\top$, $\mathbf{x} = [x_1, \dots, x_D]^\top$, and the right-hand side
 3440 represents the probability that random variable X_i takes the value smaller
 3441 than or equal to x_i .

There are cdfs which do not have corresponding pdfs.

The cdf can be expressed also as the integral of the probability density function $f(\mathbf{x})$ so that

$$F_X(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \dots, z_D) dz_1 \cdots dz_D. \quad (6.18)$$

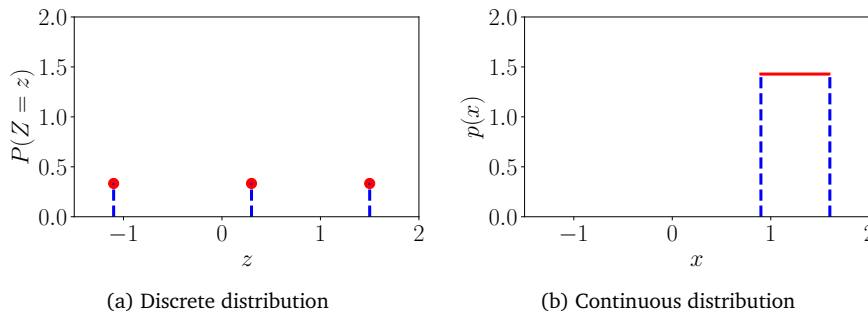


Figure 6.3
Examples of discrete and continuous Uniform distributions. See Example 6.3 for details of the distributions.

Remark. We reiterate that there are in fact two distinct concepts when talking about distributions. First, the idea of a pdf (denoted by $f(x)$) which is a non-negative function that sums to one. Second, the law of a random variable X , that is the association of a random variable X with the pdf $f(x)$. \diamond

For most of this book, we will not use the notation $f(x)$ and $F_X(x)$ as we mostly do not need to distinguish between the pdf and cdf. However, we will need to be careful about pdfs and cdfs in Section 6.7.

6.2.3 Contrasting Discrete and Continuous Distributions

Recall from Section 6.1.2 that probabilities are positive and the total probability sums up to one. For discrete random variables (see (6.12)), this implies that the probability of each state must lie in the interval $[0, 1]$. However, for continuous random variables the normalization (see (6.15)) does not imply that the value of the density is less than or equal to 1 for all values. We illustrate this in Figure 6.3 using the *uniform distribution* for both discrete and continuous random variables.

uniform distribution

Example 6.3

We consider two examples of the uniform distribution, where each state is equally likely to occur. This example illustrates some differences between discrete and continuous probability distributions.

Let Z be a discrete uniform random variable with three states $\{z = -1.1, z = 0.3, z = 1.5\}$. Note that the actual values of these states are not meaningful here, and we deliberately chose numbers to drive home the point that we do not want to use (and should ignore) the ordering of the states. The probability mass function can be represented as a table of probability values.

z	-1.1	0.3	1.5
$P(Z = z)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Table 6.1
Nomenclature for
probability
distributions.

	“point probability”	“interval probability”
discrete	$P(X = x)$ probability mass function	not applicable
continuous	$p(x)$ probability density function	$P(X \leq x)$ cumulative distribution function

Alternatively, we can think of this as a graph (Figure 6.3(a)), where we use the fact that the states can be located on the x -axis, and the y -axis represents the probability of a particular state. The y -axis in Figure 6.3(a) is deliberately extended so that it is the same as in Figure 6.3(b).

Let X be a continuous random variable taking values in the range $0.9 \leq X \leq 1.6$, as represented by Figure 6.3(b). Observe that the height of the density can be greater than 1. However, it needs to hold that

$$\int_{0.9}^{1.6} p(x) dx = 1. \quad (6.19)$$

Remark. There is an additional subtlety with regards to discrete probability distributions. The states z_1, \dots, z_d do not in principle have any structure, i.e., there is usually no way to compare them, for example $z_1 = \text{red}, z_2 = \text{green}, z_3 = \text{blue}$. However, in many machine learning applications discrete states take numerical values, e.g., $z_1 = -1.1, z_2 = 0.3, z_3 = 1.5$, where we could say $z_1 < z_2 < z_3$. Discrete states that assume numerical values are particularly useful because we often consider expected values (Section 6.4.1) of random variables. \diamond

Unfortunately machine learning literature uses notation and nomenclature that hides the distinction between the sample space Ω , the target space \mathcal{T} and the random variable X . For a value x of the set of possible outcomes of the random variable X , i.e., $x \in \mathcal{T}$, $p(x)$ denotes the probability that random variable X has the outcome x . For discrete random variables this is written as $P(X = x)$, which is known as the probability mass function (pmf). The pmf is often referred to as the “distribution”. For continuous variables, $p(x)$ is called the probability density function (often referred to as a density). To muddy things even further the cumulative distribution function $P(X \leq x)$ is often also referred to as the “distribution”. In this chapter, we will use the notation X to refer to both univariate and multivariate random variables, and denote the states by x and \mathbf{x} respectively. We summarize the nomenclature in Table 6.1.

Remark. We will be using the expression “probability distribution” not only for discrete probability mass functions but also for continuous probability density functions, although this is technically incorrect. In line with

We think of the
outcome x as the
argument that
results in the
probability $p(x)$

most machine learning literature, we also rely on context to distinguish the different uses of the phrase probability distribution. \diamond

6.3 Sum Rule, Product Rule and Bayes' Theorem

When we think of probability theory as an extension to logical reasoning. As we discussed in Section 6.1.1, the rules of probability presented here follow naturally from fulfilling the desiderata (Jaynes, 2003, Chapter 2). Probabilistic modeling (Section 8.3) provides a principled foundation for designing machine learning methods. Once we have defined probability distributions (Section 6.2) corresponding to the uncertainties of the data and our problem, it turns out that there are only two fundamental rules, the sum rule and the product rule.

Recall from (6.9) that $p(\mathbf{x}, \mathbf{y})$ is the joint distribution of the two random variables \mathbf{x}, \mathbf{y} . The distributions $p(\mathbf{x})$ and $p(\mathbf{y})$ are the corresponding marginal distributions, and $p(\mathbf{y} | \mathbf{x})$ is the conditional distribution of \mathbf{y} given \mathbf{x} . Given the definitions of the marginal and conditional probability for discrete and continuous random variables in Section 6.2, we can now present the two fundamental rules in probability theory.

The first rule, the *sum rule*, states that

$$p(\mathbf{x}) = \begin{cases} \sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}) & \text{if } \mathbf{y} \text{ is discrete} \\ \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y} & \text{if } \mathbf{y} \text{ is continuous} \end{cases}, \quad (6.20)$$

where \mathcal{Y} is the states of the target space of random variable Y . This means that we sum out (or integrate out) the set of states \mathbf{y} of the random variable Y . The sum rule is also known as the *marginalization property*. The sum rule relates the joint distribution to a marginal distribution. In general, when the joint distribution contains more than two random variables, the sum rule can be applied to any subset of the random variables, resulting in a marginal distribution of potentially more than one random variable. More concretely, if $\mathbf{x} = [x_1, \dots, x_D]^\top$, we obtain the marginal

$$p(x_i) = \int p(x_1, \dots, x_D) d\mathbf{x}_{\setminus i} \quad (6.21)$$

by repeated application of the sum rule where we integrate/sum out all random variables except x_i , which is indicated by $\setminus i$, which reads “all except i ”.

Remark. Many of the computational challenges of probabilistic modeling are due to the application of the sum rule. When there are many variables or discrete variables with many states, the sum rule boils down to performing a high-dimensional sum or integral. Performing high-dimensional sums or integrals is generally computationally hard, in the sense that there is no known polynomial-time algorithm to calculate them exactly. \diamond

These two rules arise naturally (Jaynes, 2003) from the requirements we discussed in Section 6.1.1. sum rule

marginalization property

product rule

The second rule, known as the *product rule*, relates the joint distribution to the conditional distribution via

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x})p(\mathbf{x}). \quad (6.22)$$

3508 The product rule can be interpreted as the fact that every joint distribu-
 3509 tion of two random variables can be factorized (written as a product)
 3510 of two other distributions. The two factors are the marginal distribu-
 3511 tion of the first random variable $p(\mathbf{x})$, and the conditional distribution
 3512 of the second random variable given the first $p(\mathbf{y} | \mathbf{x})$. Since the ordering
 3513 of random variables is arbitrary in $p(\mathbf{x}, \mathbf{y})$ the product rule also implies
 3514 $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x} | \mathbf{y})p(\mathbf{y})$. To be precise, (6.22) is expressed in terms of the
 3515 probability mass functions for discrete random variables. For continuous
 3516 random variables, the product rule is expressed in terms of the probability
 3517 density functions (Section 6.2.3).

Bayes' theorem

In machine learning and Bayesian statistics, we are often interested in making inferences of unobserved (latent) random variables given that we have observed other random variables. Let us assume we have some prior knowledge $p(\mathbf{x})$ about an unobserved random variable \mathbf{x} and some relationship $p(\mathbf{y} | \mathbf{x})$ between \mathbf{x} and a second random variable \mathbf{y} , which we can observe. If we observe \mathbf{y} we can use Bayes' theorem to draw some conclusions about \mathbf{x} given the observed values of \mathbf{y} . *Bayes' theorem* (also: *Bayes' rule* or *Bayes' law*)

$$\underbrace{p(\mathbf{x} | \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y} | \mathbf{x})}^{\text{likelihood}} \overbrace{p(\mathbf{x})}^{\text{prior}}}{\underbrace{p(\mathbf{y})}_{\text{evidence}}} \quad (6.23)$$

is a direct consequence of the product rule in (6.22) since

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x} | \mathbf{y})p(\mathbf{y}) \quad (6.24)$$

and

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x})p(\mathbf{x}) \quad (6.25)$$

so that

$$p(\mathbf{x} | \mathbf{y})p(\mathbf{y}) = p(\mathbf{y} | \mathbf{x})p(\mathbf{x}) \iff p(\mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{y} | \mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \quad (6.26)$$

prior

3518 In (6.23), $p(\mathbf{x})$ is the *prior*, which encapsulates our subjective prior
 3519 knowledge of the unobserved (latent) variable \mathbf{x} before observing any
 3520 data. We can choose any prior that makes sense to us, but it is critical to
 3521 ensure that the prior has a non-zero pdf (or pmf) on all plausible \mathbf{x} , even
 3522 if they are very rare.

likelihood

The likelihood is
 sometimes also
 called the
 “measurement
 model”.

3523 The *likelihood* $p(\mathbf{y} | \mathbf{x})$ describes how \mathbf{x} and \mathbf{y} are related, and in the
 3524 case of discrete probability distributions, it is the probability of the data \mathbf{y}
 3525 if we were to know the latent variable \mathbf{x} . Note that the likelihood is not a

distribution in x , but only in y . We call $p(y|x)$ either the “likelihood of x (given y)” or the “probability of y given x ” but never the likelihood of y (MacKay, 2003).

The *posterior* $p(x|y)$ is the quantity of interest in Bayesian statistics because it expresses exactly what we are interested in, i.e., what we know about x after having observed y . posterior

The quantity

$$p(y) := \int p(y|x)p(x)dx = \mathbb{E}_x[p(y|x)] \quad (6.27)$$

is the *marginal likelihood/evidence*. The right hand side of (6.27) uses the expectation operator which we define in Section 6.4.1. By definition the marginal likelihood integrates the numerator of (6.23) with respect to the latent variable x . Therefore, the marginal likelihood is independent of x and it ensures that the posterior $p(x|y)$ is normalized. The marginal likelihood can also be interpreted as the expected likelihood where we take the expectation with respect to the prior $p(x)$. Beyond normalization of the posterior the marginal likelihood also plays an important role in Bayesian model selection as we will discuss in Section 8.5. Due to the integration in (8.43), the evidence is often hard to compute. marginal likelihood
evidence

Bayes’ theorem (6.23) allows us to invert the relationship between x and y given by the likelihood. Therefore, Bayes’ theorem is sometimes called the *probabilistic inverse*. We will discuss Bayes’ theorem further in Section 8.3. Bayes’ theorem is
also called the
“probabilistic
inverse”
probabilistic inverse

Remark. In Bayesian statistics, the posterior distribution is the quantity of interest as it encapsulates all available information from the prior and the data. Instead of carrying the posterior around, it is possible to focus on some statistic of the posterior, such as the maximum of the posterior, which we will discuss in Section 8.2. However, focusing on some statistic of the posterior leads to loss of information. If we think in a bigger context, then the posterior can be used within a decision making system, and having the full posterior can be extremely useful and lead to decisions that are robust to disturbances. For example, in the context of model-based reinforcement learning, Deisenroth et al. (2015) show that using the full posterior distribution of plausible transition functions leads to very fast (data/sample efficient) learning, whereas focusing on the maximum of the posterior leads to consistent failures. Therefore, having the full posterior can be very useful for a downstream task. In Chapter 9, we will continue this discussion in the context of linear regression. ◇

6.4 Summary Statistics and Independence

We are often interested in summarizing sets of random variables and comparing pairs of random variables. A statistic of a random variable is a deterministic function of that random variable. The summary statistics of a

distribution provides one useful view of how a random variable behaves, and as the name suggests, provides numbers that summarize and characterize the distribution. We describe the mean and the variance, two well-known summary statistics. Then we discuss two ways to compare a pair of random variables: first how to say that two random variables are independent, and second how to compute an inner product between them.

6.4.1 Means and Covariances

Mean and (co)variance are often useful to describe properties of probability distributions (expected values and spread). We will see in Section 6.6 that there is a useful family of distributions (called the exponential family), where the statistics of the random variable capture all possible information.

The concept of the expected value is central to machine learning, and the foundational concepts of probability itself can be derived from the expected value (Whittle, 2000).

expected value

Definition 6.3 (Expected value). The *expected value* of a function $g : \mathbb{R} \rightarrow \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx. \quad (6.28)$$

Correspondingly the expected value of a function g of a discrete random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x) \quad (6.29)$$

where \mathcal{X} is the set of possible outcomes (the target space) of the random variable X .

In this section, we consider discrete random variables to have numerical outcomes. This can be seen by observing that the function g takes real numbers as inputs.

The expected value of a function of a random variable is sometimes referred to as the “law of the unconscious statistician” (Casella and Berger, 2002, Section 2.2).

Remark. We consider multivariate random variables X as a finite vector of univariate random variables $[X_1, \dots, X_n]^\top$. For multivariate random variables, we define the expected value element wise

$$\mathbb{E}_X[g(\mathbf{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D, \quad (6.30)$$

where the subscript \mathbb{E}_{X_d} indicates that we are taking the expected value with respect to the d th element of the vector \mathbf{x} . \diamond

Definition 6.3 defines the meaning of the notation \mathbb{E}_X as the operator

indicating that we should take the integral with respect to the probability density (for continuous distributions) or the sum over all states (for discrete distributions). The definition of the mean (Definition 6.4), is a special case of the expected value, obtained by choosing g to be the identity function.

Definition 6.4 (Mean). The *mean* of a random variable X with states $\mathbf{x} \in \mathbb{R}^D$ is an average and is defined as

$$\mathbb{E}_X[\mathbf{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D, \quad (6.31)$$

where

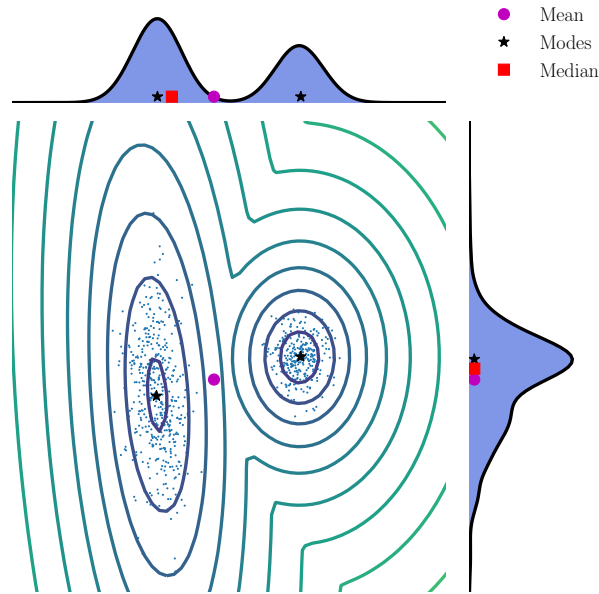
$$\mathbb{E}_{x_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) dx_d & \text{if } X \text{ is a continuous random variable} \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) & \text{if } X \text{ is a discrete random variable} \end{cases} \quad (6.32)$$

for $d = 1, \dots, D$, where the subscript d indicates the corresponding dimension of \mathbf{x} . The integral and sum are over the states \mathcal{X} of the target space of the random variable X .

In one dimension, there are two other intuitive notions of “average”, which are the *median* and the *mode*. The *median* is the “middle” value if we sort the values, i.e., 50% of the values are greater than the median and 50% are smaller than the median. This idea can be generalized to continuous values by considering the value where the cdf (Definition 6.2) is 0.5. For distributions, which are asymmetric or have long tails, the median provides an estimate of a typical value that is closer to human intuition than the mean value. Furthermore, the median is more robust to outliers than the mean. The generalization of the median to higher dimensions is non-trivial as there is no obvious way to “sort” in more than one dimension (Hallin et al., 2010; Kong and Mizera, 2012). The *mode* is the most frequently occurring value. For a discrete random variable, the mode is defined as the value of x having the highest frequency of occurrence. For a continuous random variable, the mode is defined as a peak in the density $p(\mathbf{x})$. A particular density $p(\mathbf{x})$ may have more than one mode, and furthermore there may be a very large number of modes in high dimensional distributions. Therefore finding all the modes of a distribution can be computationally challenging.

Example 6.4

Figure 6.4
Illustration of the mean, mode and median for a two-dimensional dataset, as well as its marginal densities.



Consider the two-dimensional distribution illustrated in Figure 6.4

$$p(\mathbf{x}) = 0.4\mathcal{N}\left(\mathbf{x} \mid \begin{bmatrix} 10 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.6\mathcal{N}\left(\mathbf{x} \mid \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 2.9 \\ -1.1 & 0.7 \end{bmatrix}\right). \quad (6.33)$$

We will define the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ in Section 6.5. Also shown is its corresponding marginal distribution in each dimension. Observe that the distribution is bimodal (has two modes), but one of the marginal distributions is unimodal (has one mode). The horizontal bimodal univariate distribution illustrates that the mean and median can be different from each other. While it is tempting to define the two-dimensional median to be the concatenation of the medians in each dimension, the fact that we cannot define an ordering of two-dimensional points makes it difficult. When we say “cannot define an ordering”, we mean that there is more than one way to define the relation $<$ so that $\begin{bmatrix} 3 \\ 0 \end{bmatrix} < \begin{bmatrix} 2 \\ 3 \end{bmatrix}$.

Remark. The expected value (Definition 6.3) is a linear operator. For example, given a real-valued function $f(\mathbf{x}) = ag(\mathbf{x}) + bh(\mathbf{x})$ where $a, b \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, we obtain

$$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (6.34a)$$

$$= \int [ag(\mathbf{x}) + bh(\mathbf{x})]p(\mathbf{x})d\mathbf{x} \quad (6.34b)$$

$$= a \int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} + b \int h(\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (6.34c)$$

$$= a\mathbb{E}_X[g(\mathbf{x})] + b\mathbb{E}_X[h(\mathbf{x})]. \quad (6.34d)$$

◇

For two random variables, we may wish to characterize their correspondence to each other. The covariance intuitively represents the notion of how dependent random variables are to one another.

Definition 6.5 (Covariance (univariate)). The *covariance* between two univariate random variables $X, Y \in \mathbb{R}$ is given by the expected product of their deviations from their respective means, that is

$$\text{Cov}_{X,Y}[x, y] := \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])]. \quad (6.35)$$

covariance
Terminology:
Covariance of
multivariate random
variables $\text{Cov}[x, y]$
is sometimes
referred to as
cross-covariance,
with covariance
referring to
 $\text{Cov}[x, x]$.

Remark. When the random variable associated with the expectation or covariance is clear by its arguments the subscript is often suppressed (for example $\mathbb{E}_X[x]$ is often written as $\mathbb{E}[x]$). ◇

By using the linearity of expectations, the expression in Definition 6.5 can be rewritten as the expected value of the product minus the product of the expected values, i.e.,

$$\text{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \quad (6.36)$$

The covariance of a variable with itself $\text{Cov}[x, x]$ is called the *variance* and is denoted by $\mathbb{V}_X[x]$. The square-root of the variance is called the *standard deviation* and is often denoted by $\sigma(x)$. The notion of covariance can be generalised to multivariate random variables.

variance
standard deviation

Definition 6.6 (Covariance). If we consider two multivariate random variables X and Y with states $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^E$ respectively, the *covariance* between X and Y is defined as

covariance

$$\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}\mathbf{y}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]^\top = \text{Cov}[\mathbf{y}, \mathbf{x}]^\top \in \mathbb{R}^{D \times E}. \quad (6.37)$$

Definition 6.6 can be applied with the same multivariate random variable in both arguments, which results in a useful concept that intuitively captures the “spread” of a random variable. For a multivariate random variable, the variance describes the relation between individual dimensions of the random variable.

Definition 6.7 (Variance). The *variance* of a random variable X with states $\mathbf{x} \in \mathbb{R}^D$, and a mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ is defined as

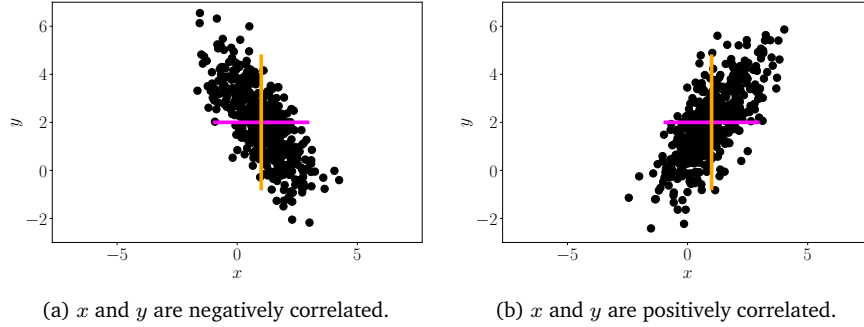
variance

$$\mathbb{V}_X[\mathbf{x}] = \text{Cov}_X[\mathbf{x}, \mathbf{x}] \quad (6.38a)$$

$$= \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \quad (6.38b)$$

Figure 6.5

Two-dimensional datasets with identical means and variances along each axis (colored lines) but with different covariances.



$$= \begin{bmatrix} \text{Cov}[x_1, x_1] & \text{Cov}[x_1, x_2] & \dots & \text{Cov}[x_1, x_D] \\ \text{Cov}[x_2, x_1] & \text{Cov}[x_2, x_2] & \dots & \text{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[x_D, x_1] & \dots & \dots & \text{Cov}[x_D, x_D] \end{bmatrix}. \quad (6.38c)$$

covariance matrix

The matrix (6.38c) is of size $D \times D$ and is called the *covariance matrix* of the multivariate random variable X . The covariance matrix is symmetric and positive definite and tells us something about the spread of the data. The covariance matrix contains the variances of the *marginals*

marginal

$$p(x_i) = \int p(x_1, \dots, x_D) dx_{\setminus i} \quad (6.39)$$

cross-covariance

on its diagonal, where “ $\setminus i$ ” denotes “all variables but i ”. The off-diagonal entries are the *cross-covariance* terms $\text{Cov}[x_i, x_j]$ for $i, j = 1, \dots, D$, $i \neq j$.

When we want to compare the covariances between different pairs of random variables, it turns out that the variance of each random variable affects the value of the covariance. The normalized version of covariance is called the correlation.

correlation

Definition 6.8 (Correlation). The *correlation* between two random variables X, Y is given by

$$\text{corr}[x, y] = \frac{\text{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1]. \quad (6.40)$$

The correlation matrix is the covariance matrix of standardized random variables, $x/\sigma(x)$. In other words, each random variable is divided by its standard deviation (the square root of the variance) in the correlation matrix.

The covariance (and correlation) indicate how two random variables are related, see Figure 6.5. Positive correlation $\text{corr}[x, y]$ means that when x grows then y is also expected to grow. Negative correlation means that as x increases then y decreases.

6.4.2 Empirical Means and Covariances

The definitions in Section 6.4.1 are often also called the *population mean and covariance*, as it refers to the true statistics for the population. In machine learning we need to learn from empirical observations of data. Consider a random variable X . There are two conceptual steps to go from population statistics to the realization of empirical statistics. First we use the fact that we have a finite dataset (of size N) to construct an empirical statistic which is a function of a finite number of identical random variables, X_1, \dots, X_N . Second we observe the data, that is we look at the realization x_1, \dots, x_N of each of the random variables and apply the empirical statistic.

population mean
and covariance

Specifically for the mean (Definition 6.4), given a particular dataset we can obtain an estimate of the mean, which is called the *empirical mean* or *sample mean*. The same holds for the empirical covariance.

empirical mean
sample mean

Definition 6.9 (Empirical Mean and Covariance). The *empirical mean* vector is the arithmetic average of the observations for each variable, and it is defined as

empirical mean

$$\bar{\mathbf{x}} := \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \quad (6.41)$$

where $\mathbf{x}_n \in \mathbb{R}^D$.

Similar to the empirical mean, the *empirical covariance* matrix is a $D \times D$ matrix

empirical covariance

$$\Sigma := \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^\top. \quad (6.42)$$

To compute the statistics for a particular dataset, we would use the realizations (observations) $\mathbf{x}_1, \dots, \mathbf{x}_N$ and use (6.41) and (6.42). Empirical covariance matrices are symmetric, positive semi-definite (see Section 3.2.3).

Throughout the book we use the empirical covariance, which is a biased estimate. The unbiased (sometimes called corrected) covariance has the factor $N - 1$ in the denominator instead of N .

The derivations are exercises at the end of this chapter.

6.4.3 Three Expressions for the Variance

We now focus on a single random variable X and use the empirical formulas above to derive three possible expressions for the variance. The derivation below is the same for the population variance, except that we need to take care of integrals. The standard definition of variance, corresponding to the definition of covariance (Definition 6.5), is the expectation of the squared deviation of a random variable X from its expected value μ , i.e.,

$$\mathbb{V}_X[x] := \mathbb{E}_X[(x - \mu)^2]. \quad (6.43)$$

The expectation in (6.43) and the mean $\mu = \mathbb{E}_X(x)$ are computed using (6.32), depending on whether X is a discrete or continuous random

variable. The variance as expressed in (6.43) is the mean of a new random variable $Z := (X - \mu)^2$.

raw-score formula
for variance

When estimating the variance in (6.43) empirically, we need to resort to a two-pass algorithm: one pass through the data to calculate the mean μ using (6.41), and then a second pass using this estimate $\hat{\mu}$ calculate the variance. It turns out that we can avoid two passes by rearranging the terms. The formula in (6.43) can be converted to the so-called *raw-score formula for variance*:

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2. \quad (6.44)$$

The two terms can
cancel out, resulting
in a loss of
numerical precision
in floating point
arithmetic.

The expression in (6.44) can be remembered as “the mean of the square minus the square of the mean”. It can be calculated empirically in one pass through data since we can accumulate x_i (to calculate the mean) and x_i^2 simultaneously, where x_i is the i^{th} observation. Unfortunately, if implemented in this way, it can be numerically unstable. The raw-score version of the variance can be useful in machine learning, e.g., when deriving the bias-variance decomposition (Bishop, 2006).

A third way to understand the variance is that it is a sum of pairwise differences between all pairs of observations. Consider a sample x_1, \dots, x_N of realizations of random variable X , and we compute the squared difference between pairs of x_i and x_j . By expanding the square we can show that the sum of N^2 pairwise differences is the empirical variance of the observations,

$$\frac{1}{N^2} \sum_{i,j=1}^N (x_i - x_j)^2 = 2 \left[\frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2 \right] \quad (6.45)$$

We see that (6.45) is twice the raw-score expression (6.44). This means that we can express the sum of pairwise distances (of which there are N^2 of them) as a sum of deviations from the mean (of which there are N). Geometrically, this means that there is an equivalence between the pairwise distances and the distances from the center of the set of points. From a computational perspective, this means that by computing the mean (N terms in the summation), and then computing the variance (again N terms in the summation) we can obtain an expression (left-hand side of (6.45)) that has N^2 terms.

6.4.4 Sums and Transformations of Random Variables

We may want to model a phenomenon that cannot be well explained by textbook distributions (we introduce some in Sections 6.5 and 6.6), and hence may perform simple manipulations of random variables (such as adding two random variables).

Consider two random variables X, Y with states $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$. It holds

that

$$\mathbb{E}[\mathbf{x} + \mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{y}] \quad (6.46)$$

$$\mathbb{E}[\mathbf{x} - \mathbf{y}] = \mathbb{E}[\mathbf{x}] - \mathbb{E}[\mathbf{y}] \quad (6.47)$$

$$\mathbb{V}[\mathbf{x} + \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] + \text{Cov}[\mathbf{x}, \mathbf{y}] + \text{Cov}[\mathbf{y}, \mathbf{x}] \quad (6.48)$$

$$\mathbb{V}[\mathbf{x} - \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] - \text{Cov}[\mathbf{x}, \mathbf{y}] - \text{Cov}[\mathbf{y}, \mathbf{x}] \quad (6.49)$$

Mean and (co)variance exhibit some useful properties when it comes to affine transformation of random variables. Consider a random variable X with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and a (deterministic) affine transformation $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ of \mathbf{x} . Then \mathbf{y} is itself a random variable whose mean vector and covariance matrix are given by

$$\mathbb{E}_Y[\mathbf{y}] = \mathbb{E}_X[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mathbb{E}_X[\mathbf{x}] + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \quad (6.50)$$

$$\mathbb{V}_Y[\mathbf{y}] = \mathbb{V}_X[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbb{V}_X[\mathbf{A}\mathbf{x}] = \mathbf{A}\mathbb{V}_X[\mathbf{x}]\mathbf{A}^\top = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top, \quad (6.51)$$

This can be shown directly by using the definition of the mean and covariance.

respectively. Furthermore,

$$\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}(\mathbf{A}\mathbf{x} + \mathbf{b})^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}]^\top \quad (6.52a)$$

$$= \mathbb{E}[\mathbf{x}]\mathbf{b}^\top + \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{A}^\top - \boldsymbol{\mu}\mathbf{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\mathbf{A}^\top \quad (6.52b)$$

$$= \boldsymbol{\mu}\mathbf{b}^\top - \boldsymbol{\mu}\mathbf{b}^\top + (\mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top)\mathbf{A}^\top \quad (6.52c)$$

$$\stackrel{(6.38b)}{=} \boldsymbol{\Sigma}\mathbf{A}^\top, \quad (6.52d)$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ is the covariance of X .

6.4.5 Statistical Independence

Definition 6.10 (Independence). Two random variables X, Y are *statistically independent* if and only if

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}). \quad (6.53)$$

Intuitively, two random variables X and Y are independent if the value of \mathbf{y} (once known) does not add any additional information about \mathbf{x} (and vice versa). If X, Y are (statistically) independent then

- $p(\mathbf{y} | \mathbf{x}) = p(\mathbf{y})$
- $p(\mathbf{x} | \mathbf{y}) = p(\mathbf{x})$
- $\mathbb{V}_{X,Y}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_X[\mathbf{x}] + \mathbb{V}_Y[\mathbf{y}]$
- $\text{Cov}_{X,Y}[\mathbf{x}, \mathbf{y}] = \mathbf{0}$

The last point above may not hold in converse, i.e., two random variables can have covariance zero but are not statistically independent. To understand why, recall that covariance measures only linear dependence. Therefore, random variables that are nonlinearly dependent could have covariance zero.

Example 6.5

Consider a random variable X with zero mean ($\mathbb{E}_X[x] = 0$) and also $\mathbb{E}_X[x^3] = 0$. Let $y = x^2$ (hence, Y is dependent on X) and consider the covariance (6.36) between X and Y . But this gives

$$\text{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x^3] = 0. \quad (6.54)$$

independent and
identically
distributed

3704 In machine learning, we often consider problems that can be modeled
3705 as *independent and identically distributed* random variables, X_1, \dots, X_N .
3706 The word “independent” refers to Definition 6.10, i.e., any pair of random
3707 variables X_i and X_j are independent. The phrase “identically distributed”
3708 means that all the random variables are from the same distribution.

3709 Another concept that is important in machine learning is conditional
3710 independence.

conditionally
independent

Definition 6.11 (Conditional Independence). Two random variables X and Y are *conditionally independent* given Z if and only if

$$p(\mathbf{x}, \mathbf{y} | \mathbf{z}) = p(\mathbf{x} | \mathbf{z})p(\mathbf{y} | \mathbf{z}) \quad \text{for all } \mathbf{z} \in \mathcal{Z}, \quad (6.55)$$

3711 where \mathcal{Z} is the set of states of random variable Z . We write $X \perp\!\!\!\perp Y | Z$ to
3712 denote that X is conditionally independent of Y given Z .

3713 Definition 6.11 requires that the relation in (6.55) must hold true for
3714 every value of \mathbf{z} . The interpretation of (6.55) can be understood as “given
3715 knowledge about \mathbf{z} , the distribution of \mathbf{x} and \mathbf{y} factorizes”. Independence
3716 can be cast as a special case of conditional independence if we write $X \perp\!\!\!\perp$
3717 $Y | \emptyset$.

By using the product rule of probability (6.22) we can expand the left-hand side of (6.55) to obtain

$$p(\mathbf{x}, \mathbf{y} | \mathbf{z}) = p(\mathbf{x} | \mathbf{y}, \mathbf{z})p(\mathbf{y} | \mathbf{z}). \quad (6.56)$$

By comparing the right-hand side of (6.55) with (6.56) we see that $p(\mathbf{y} | \mathbf{z})$ appears in both of them so that

$$p(\mathbf{x} | \mathbf{y}, \mathbf{z}) = p(\mathbf{x} | \mathbf{z}). \quad (6.57)$$

3718 Equation (6.57) provides an alternative definition of conditional indepen-
3719 dence, i.e., $X \perp\!\!\!\perp Y | Z$. This alternative presentation provides the inter-
3720 pretation “given that we know \mathbf{z} , knowledge about \mathbf{y} does not change our
3721 knowledge of \mathbf{x} ”.

3722

6.4.6 Inner Products of Random Variables

Recall the definition of inner products from Section 3.2. We can define an inner product between random variables, which we briefly describe in this

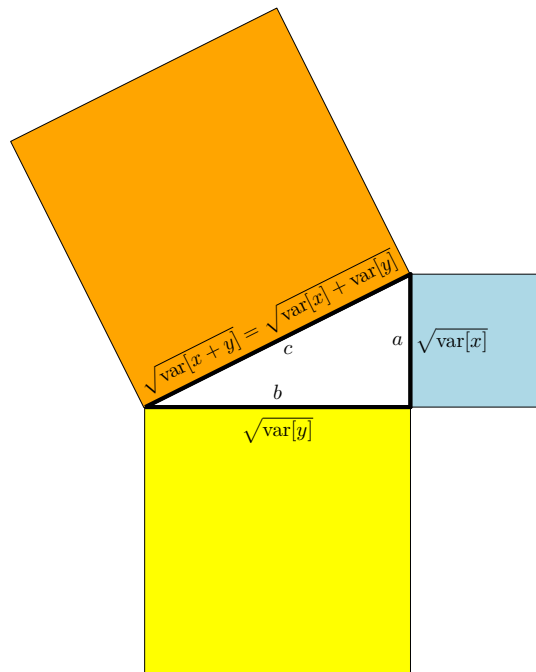


Figure 6.6
Geometry of random variables. If random variables x and y are uncorrelated they are orthogonal vectors in a corresponding vector space, and the Pythagorean theorem applies.

section. If we have two uncorrelated random variables X, Y then

$$\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y] \quad (6.58)$$

Since variances are measured in squared units, this looks very much like the Pythagorean theorem for right triangles $c^2 = a^2 + b^2$.

In the following, we see whether we can find a geometric interpretation of the variance relation of uncorrelated random variables in (6.58). Random variables can be considered vectors in a vector space, and we can define inner products to obtain geometric properties of random variables (Eaton, 2007). If we define

$$\langle X, Y \rangle := \text{Cov}[x, y] \quad (6.59)$$

for zero mean random variables X and Y , we obtain an inner product. we see that the covariance is symmetric, positive definite, and linear in either argument. The length of a random variable is

$$\|X\| = \sqrt{\text{Cov}[x, x]} = \sqrt{\mathbb{V}[x]} = \sigma[x], \quad (6.60)$$

i.e., its standard deviation. The “longer” the random variable, the more uncertain it is; and a random variable with length 0 is deterministic.

If we look at the angle θ between random two random variables X, Y , we get

$$\cos \theta = \frac{\langle X, Y \rangle}{\|X\| \|Y\|} = \frac{\text{Cov}[x, y]}{\sqrt{\mathbb{V}[x] \mathbb{V}[y]}}, \quad (6.61)$$

Inner products between multivariate random variables can be treated in a similar fashion

$$\begin{aligned} \text{Cov}[x, x] &= 0 \iff x = 0 \\ \text{Cov}[\alpha x + z, y] &= \alpha \text{Cov}[x, y] + \text{Cov}[z, y] \text{ for } \alpha \in \mathbb{R}. \end{aligned}$$

which is the correlation (Definition 6.8) between the two random variables. This means that we can think of correlation as the cosine of the angle between two random variables when we consider them geometrically. We know from Definition 3.7 that $X \perp Y \iff \langle X, Y \rangle = 0$. In our case, this means that X and Y are orthogonal if and only if $\text{Cov}[x, y] = 0$, i.e., they are uncorrelated. Figure 6.6 illustrates this relationship.

Remark. While it is tempting to use the Euclidean distance (constructed from the definition of inner products above) to compare probability distributions, it is unfortunately not the best way to obtain distances between distributions. Recall that the probability mass (or density) is positive and needs to add up to 1. These constraints mean that distributions live on something called a statistical manifold. The study of this space of probability distributions is called information geometry. Computing distances between distributions are often done using Kullback-Leibler divergence which is a generalization of distances that account for properties of the statistical manifold. Just like the Euclidean distance is a special case of a metric (Section 3.3) the Kullback-Leibler divergence is a special case of two more general classes of divergences called Bregman divergences and f -divergences. The study of divergences is beyond the scope of this book, and we refer for more details to the recent book by Amari (2016), one of the founders of the field of information geometry. \diamond

6.5 Gaussian Distribution

The Gaussian distribution arises naturally when we consider sums of independent and identically distributed random variables. This is known as the Central Limit Theorem (Grinstead and Snell, 1997). normal distribution

The Gaussian distribution is the most well studied probability distribution for continuous-valued random variables. It is also referred to as the *normal distribution*. Its importance originates from the fact that it has many computationally convenient properties, which we will be discussing in the following. In particular, we will use it to define the likelihood and prior for linear regression (Chapter 9), and consider a mixture of Gaussians for density estimation (Chapter 11).

There are many other areas of machine learning that also benefit from using a Gaussian distribution, for example Gaussian processes, variational inference and reinforcement learning. It is also widely used in other application areas such as signal processing (e.g., Kalman filter), control (e.g., linear quadratic regulator) and statistics (e.g. hypothesis testing).

For a univariate random variable, the Gaussian distribution has a density that is given by

$$p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (6.62)$$

multivariate Gaussian distribution
Also: multivariate normal distribution
mean vector
covariance matrix

The *multivariate Gaussian distribution* is fully characterized by a *mean vector* μ and a *covariance matrix* Σ and defined as

$$p(x | \mu, \Sigma) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right), \quad (6.63)$$

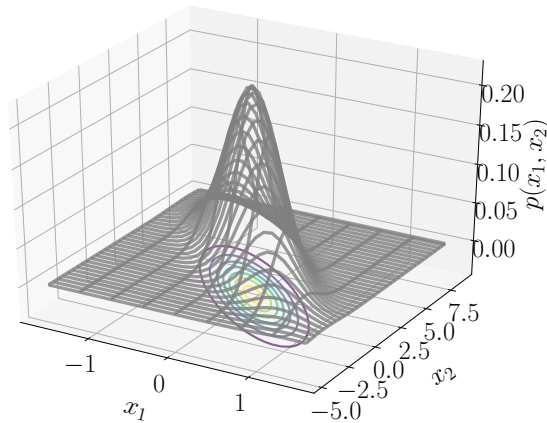


Figure 6.7
Gaussian
distribution of two
random variables
 x, y .

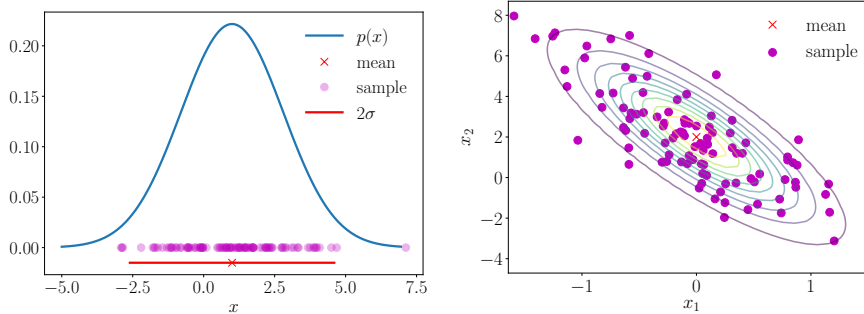


Figure 6.8
Gaussian
distributions
overlaid with 100
samples.

(a) Univariate (1-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance. (b) Multivariate (2-dimensional) Gaussian, viewed from top. The red cross shows the mean and the coloured lines shows the contour lines of the density.

where $\mathbf{x} \in \mathbb{R}^D$. We write $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Figure 6.7 shows a bivariate Gaussian (mesh), with the corresponding contour plot. Figure 6.8 shows a univariate Gaussian and a bivariate Gaussian with corresponding samples. The special case of the Gaussian with zero mean and identity covariance, that is $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, is referred to as the *standard normal distribution*.

Gaussian distributions are widely used in statistical estimation and machine learning because they have closed-form expressions for marginal and conditional distributions. In Chapter 9, we use these closed form expressions extensively for linear regression. A major advantage of modelling with Gaussian distributed random variables is that variable transformations (Section 6.7) are often not needed. Since the Gaussian distribution is fully specified by its mean and covariance we often can obtain the transformed distribution by applying the transformation to the mean and covariance of the random variable.

standard normal
distribution

3776 6.5.1 Marginals and Conditionals of Gaussians are Gaussians

In the following, we present marginalization and conditioning in the general case of multivariate random variables. If this is confusing at first reading, the reader is advised to consider two univariate random variables instead. Let X and Y be two multivariate random variables, which may have different dimensions. We would like to consider the effect of applying the sum rule of probability and the effect of conditioning. We therefore explicitly write the Gaussian distribution in terms of the concatenated states $[\mathbf{x}, \mathbf{y}]^\top$,

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right). \quad (6.64)$$

3777 where $\boldsymbol{\Sigma}_{xx} = \text{Cov}[\mathbf{x}, \mathbf{x}]$ and $\boldsymbol{\Sigma}_{yy} = \text{Cov}[\mathbf{y}, \mathbf{y}]$ are the marginal covariance
3778 matrices of \mathbf{x} and \mathbf{y} , respectively, and $\boldsymbol{\Sigma}_{xy} = \text{Cov}[\mathbf{x}, \mathbf{y}]$ is the cross-
3779 covariance matrix between \mathbf{x} and \mathbf{y} .

The conditional distribution $p(\mathbf{x} | \mathbf{y})$ is also Gaussian (illustrated on the bottom right of Figure 6.9) and given by (derived in Section 2.3 of Bishop (2006))

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \quad (6.65)$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} (\mathbf{y} - \boldsymbol{\mu}_y) \quad (6.66)$$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx}. \quad (6.67)$$

3780 Note that in the computation of the mean in (6.66) the \mathbf{y} -value is an
3781 observation and no longer random.

3782 *Remark.* The conditional Gaussian distribution shows up in many places,
3783 where we are interested in posterior distributions:

- 3784 • The Kalman filter (Kalman, 1960), one of the most central algorithms
3785 for state estimation in signal processing, does nothing but computing
3786 Gaussian conditionals of joint distributions (Deisenroth and Ohlsson,
3787 2011).
- 3788 • Gaussian processes (Rasmussen and Williams, 2006), which are a prac-
3789 tical implementation of a distribution over functions. In a Gaussian pro-
3790 cess, we make assumptions of joint Gaussianity of random variables. By
3791 (Gaussian) conditioning on observed data, we can determine a poste-
3792 rior distribution over functions.
- 3793 • Latent linear Gaussian models (Roweis and Ghahramani, 1999; Mur-
3794 phy, 2012), which include probabilistic principal component analysis
3795 (PPCA) (Tipping and Bishop, 1999). We will look at PPCA in more de-
3796 tail in Section 10.7.

3797 \diamond

The marginal distribution $p(\mathbf{x})$ of a joint Gaussian distribution $p(\mathbf{x}, \mathbf{y})$,

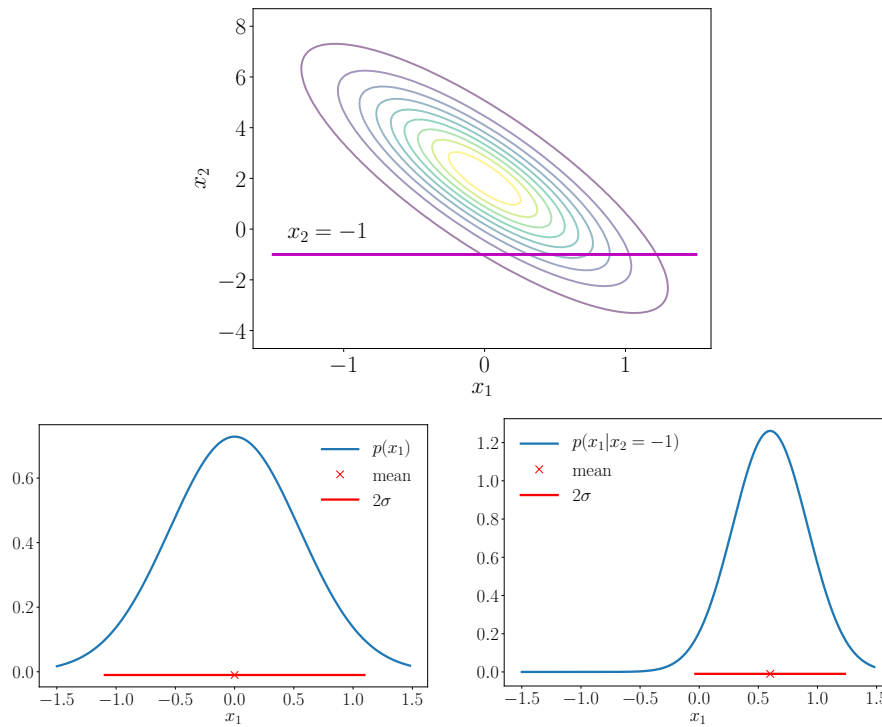


Figure 6.9 Top: Bivariate Gaussian; Bottom left: Marginal of a joint Gaussian distribution is Gaussian; Bottom right: The conditional distribution of a Gaussian is also Gaussian

see (6.64), is itself Gaussian and computed by applying the sum rule (6.20) and given by

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}). \quad (6.68)$$

3798 The corresponding result holds for $p(\mathbf{y})$, which is obtained by marginaliz-
 3799 ing with respect to \mathbf{x} . Intuitively, looking at the joint distribution in (6.64),
 3800 we ignore (i.e., integrate out) everything we are not interested in. This is
 3801 illustrated on the bottom left of Figure 6.9.

Example 6.6

Consider the bivariate Gaussian distribution (illustrated in Figure 6.9)

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right). \quad (6.69)$$

We can compute the parameters of the univariate Gaussian, conditioned on $x_2 = -1$, by applying (6.66) and (6.67) to obtain the mean and variance respectively. Numerically, this is

$$\mu_{x_1 | x_2 = -1} = 0 + (-1)(0.2)(-1 - 2) = 0.6 \quad (6.70)$$

and

$$\sigma_{x_1 | x_2=-1}^2 = 0.3 - (-1)(0.2)(-1) = 0.1. \quad (6.71)$$

Therefore the conditional Gaussian is given by

$$p(x_1 | x_2 = -1) = \mathcal{N}(0.6, 0.1). \quad (6.72)$$

The marginal distribution $p(x_1)$ in contrast can be obtained by applying (6.68), which is essentially using the mean and variance of the random variable x_1 , giving us

$$p(x_1) = \mathcal{N}(0, 0.3) \quad (6.73)$$

6.5.2 Product of Gaussian Densities

For linear regression (Chapter 9), we need to compute a Gaussian likelihood. Furthermore we may wish to assume a Gaussian prior (Section 9.3). The application of Bayes rule to compute the posterior results in a multiplication of the likelihood and the prior, that is the multiplication of two Gaussian densities. The *product* of two Gaussians $\mathcal{N}(\mathbf{x} | \mathbf{a}, \mathbf{A})\mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B})$ is a Gaussian distribution scaled by a $c \in \mathbb{R}$, given by $c\mathcal{N}(\mathbf{x} | \mathbf{c}, \mathbf{C})$ with

$$\mathbf{C} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} \quad (6.74)$$

$$\mathbf{c} = \mathbf{C}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}) \quad (6.75)$$

$$c = (2\pi)^{-\frac{D}{2}} |\mathbf{A} + \mathbf{B}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{a} - \mathbf{b})^\top (\mathbf{A} + \mathbf{B})^{-1} (\mathbf{a} - \mathbf{b})\right). \quad (6.76)$$

The scaling constant c itself can be written in the form of a Gaussian density either in \mathbf{a} or in \mathbf{b} with an “inflated” covariance matrix $\mathbf{A} + \mathbf{B}$, i.e., $c = \mathcal{N}(\mathbf{a} | \mathbf{b}, \mathbf{A} + \mathbf{B}) = \mathcal{N}(\mathbf{b} | \mathbf{a}, \mathbf{A} + \mathbf{B})$.

Remark. For notation convenience, we will sometimes use $\mathcal{N}(\mathbf{x} | \mathbf{m}, \mathbf{S})$ to describe the functional form of a Gaussian density even if \mathbf{x} is not a random variable. We have just done this above when we wrote

$$c = \mathcal{N}(\mathbf{a} | \mathbf{b}, \mathbf{A} + \mathbf{B}) = \mathcal{N}(\mathbf{b} | \mathbf{a}, \mathbf{A} + \mathbf{B}). \quad (6.77)$$

Here, neither \mathbf{a} nor \mathbf{b} are random variables. However, writing c in this way is more compact than (6.76). \diamond

6.5.3 Sums and Linear Transformations

If \mathbf{X}, \mathbf{Y} are independent Gaussian random variables (i.e., the joint is given as $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$) with $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, then $\mathbf{x} + \mathbf{y}$ is also Gaussian distributed and given by

$$p(\mathbf{x} + \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y). \quad (6.78)$$

Knowing that $p(\mathbf{x} + \mathbf{y})$ is Gaussian, the mean and covariance matrix can be determined immediately using the results from (6.46)–(6.49). This property will be important when we consider i.i.d. Gaussian noise acting on random variables as is the case for linear regression (Chapter 9).

Example 6.7

Since expectations are linear operations, we can obtain the weighted sum of independent Gaussian random variables

$$p(a\mathbf{x} + b\mathbf{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y, a^2\boldsymbol{\Sigma}_x + b^2\boldsymbol{\Sigma}_y). \quad (6.79)$$

Remark. A case which will be useful in Chapter 11 is the weighted sum of Gaussian densities. This is different from the weighted sum of Gaussian random variables. \diamond

In Theorem 6.12, the random variable x is from a density which a mixture of two densities $p_1(x)$ and $p_2(x)$, weighted by α . The theorem can be generalized to the multivariate random variable case, since linearity of expectations holds also for multivariate random variables. However the idea of a squared random variable needs to be replaced by $\mathbf{x}\mathbf{x}^\top$.

Theorem 6.12. Consider a weighted sum of two univariate Gaussian densities

$$p(x) = \alpha p_1(x) + (1 - \alpha)p_2(x) \quad (6.80)$$

where the scalar $0 < \alpha < 1$ is the mixture weight, and $p_1(x)$ and $p_2(x)$ are univariate Gaussian densities (Equation (6.62)) with different parameters, that is $(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2)$.

The mean of the mixture x is given by the weighted sum of the means of each random variable,

$$\mathbb{E}[x] = \alpha\mu_1 + (1 - \alpha)\mu_2. \quad (6.81)$$

The variance of the mixture x is the mean of the conditional variance and the variance of the conditional mean,

$$\mathbb{V}[x] = [\alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2] + \left([\alpha\mu_1^2 + (1 - \alpha)\mu_2^2] - [\alpha\mu_1 + (1 - \alpha)\mu_2]^2 \right). \quad (6.82)$$

Proof The mean of the mixture x is given by the weighted sum of the means of each random variable. We apply the definition of the mean (Definition 6.4), and plug in our mixture (Equation (6.80)) above

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)dx \quad (6.83a)$$

$$= \int_{-\infty}^{\infty} \alpha xp_1(x) + (1 - \alpha)xp_2(x)dx \quad (6.83b)$$

$$= \alpha \int_{-\infty}^{\infty} xp_1(x)dx + (1 - \alpha) \int_{-\infty}^{\infty} xp_2(x)dx \quad (6.83c)$$

$$= \alpha\mu_1 + (1 - \alpha)\mu_2. \quad (6.83d)$$

To compute the variance, we can use the raw score version of the variance (Equation (6.44)), which requires an expression of the expectation of the squared random variable. Here we use the definition of an expectation of a function (the square) of a random variable (Definition 6.3).

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x) dx \quad (6.84a)$$

$$= \int_{-\infty}^{\infty} \alpha x^2 p_1(x) + (1 - \alpha) x^2 p_2(x) dx \quad (6.84b)$$

$$= \alpha \int_{-\infty}^{\infty} x^2 p_1(x) dx + (1 - \alpha) \int_{-\infty}^{\infty} x^2 p_2(x) dx \quad (6.84c)$$

$$= \alpha(\mu_1^2 + \sigma_1^2) + (1 - \alpha)(\mu_2^2 + \sigma_2^2). \quad (6.84d)$$

where in the last equality, we again used the raw score version of the variance and rearranged terms such that the expectation of a squared random variable is the sum of the squared mean and the variance.

Therefore the variance is given by subtracting (6.83d) from (6.84d),

$$\mathbb{V}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \quad (6.85a)$$

$$= \alpha(\mu_1^2 + \sigma_1^2) + (1 - \alpha)(\mu_2^2 + \sigma_2^2) - (\alpha\mu_1 + (1 - \alpha)\mu_2)^2 \quad (6.85b)$$

$$= [\alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2] + \left([\alpha\mu_1^2 + (1 - \alpha)\mu_2^2] - [\alpha\mu_1 + (1 - \alpha)\mu_2]^2 \right). \quad (6.85c)$$

For a mixture, the individual components can be considered to be conditional distributions (conditioned on the component identity). The last line is an illustration of the conditional variance formula: “The variance of a mixture is the mean of the conditional variance and the variance of the conditional mean”. \square

Remark. The derivation above holds for any density, but since the Gaussian is fully determined by the mean and variance, the mixture density can be determined in closed form. \diamond

We consider in Example 6.17 a bivariate standard Gaussian random variable X and performed a linear transformation Ax on it. The outcome is a Gaussian random variable with zero mean and covariance AA^\top . Observe that adding a constant vector will change the mean of the distribution, without affecting its variance, that is the random variable $x + \mu$ is Gaussian with mean μ and identity covariance. Therefore, a linear (or affine) transformation of a Gaussian random variable is Gaussian distributed.

Consider a Gaussian distributed random variable $X \sim \mathcal{N}(\mu, \Sigma)$. For a

given matrix \mathbf{A} of appropriate shape, let \mathbf{Y} be a random variable such that $\mathbf{y} = \mathbf{A}\mathbf{x}$ is a transformed version of \mathbf{x} . We can compute the mean of \mathbf{y} by using the fact that the expectation is a linear operator (Equation (6.50)) as follows:

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x}] = \mathbf{A}\mathbb{E}[\mathbf{x}] = \mathbf{A}\boldsymbol{\mu}. \quad (6.86)$$

Similarly the variance of \mathbf{y} can be found by using Equation (6.51):

$$\mathbb{V}[\mathbf{y}] = \mathbb{V}[\mathbf{A}\mathbf{x}] = \mathbf{A}\mathbb{V}[\mathbf{x}]\mathbf{A}^\top = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top. \quad (6.87)$$

This means that the random variable \mathbf{y} is distributed according to

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top). \quad (6.88)$$

Let us now consider the reverse transformation: when we know that a random variable has a mean that is a linear transformation of another random variable. For a given full rank matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ where $M \geq N$, let $\mathbf{y} \in \mathbb{R}^M$ be a Gaussian random variable with mean $\mathbf{A}\mathbf{x}$, i.e.,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x}, \boldsymbol{\Sigma}). \quad (6.89)$$

What is the corresponding probability distribution $p(\mathbf{x})$? If \mathbf{A} is invertible, then we can write $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ and apply the transformation in the previous paragraph. However, in general \mathbf{A} is not invertible, and we use an approach similar to that of the pseudo-inverse (3.57). That is we pre-multiply both sides with \mathbf{A}^\top and then invert $\mathbf{A}^\top\mathbf{A}$ which is symmetric and positive definite, giving us the relation

$$\mathbf{y} = \mathbf{A}\mathbf{x} \iff (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{y} = \mathbf{x}. \quad (6.90)$$

Hence, \mathbf{x} is a linear transformation of \mathbf{y} , and we obtain

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{y}, (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}). \quad (6.91)$$

6.5.4 Sampling from Multivariate Gaussian Distributions

We will not explain the subtleties of random sampling on a computer, and the interested reader is referred to Gentle (2004). In the case of a multivariate Gaussian, this process consists of three stages: first we need a source of pseudo-random numbers that provide a uniform sample in the interval $[0,1]$, second we use a non-linear transformation such as the Box-Müller transform (Devroye, 1986) to obtain a sample from a univariate Gaussian, and third we collate a vector of these samples to obtain a sample from a multivariate standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

For a general multivariate Gaussian, that is where the mean is non-zero and the covariance is not the identity matrix, we use the properties of linear transformations of a Gaussian random variable. Assume we are interested in generating samples $\mathbf{x}_i, i = 1, \dots, n$, from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We would like

3857 to construct the sample from a sampler that provides samples from the
 3858 multivariate standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
 To compute the Cholesky factorization of a matrix, it is required
 3859 that the matrix is symmetric and positive definite
 3860 (Section 3.2.3). Covariance matrices possess this
 3861 property.

3862 To obtain samples from a multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can use
 3863 the properties of a linear transformation of a Gaussian random variable:
 3864 If $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\mu}$, where $\mathbf{A}\mathbf{A}^\top = \boldsymbol{\Sigma}$, is Gaussian distributed
 3865 with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. One convenient choice of \mathbf{A} is to use the Cholesky decomposition (Section 4.3) of the covariance matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$. The Cholesky decomposition has the benefit that \mathbf{A} is triangular, leading to efficient computation.

3866 6.6 Conjugacy and the Exponential Family

3867 Many of the probability distributions “with names” that we find in statistics
 3868 textbooks were discovered to model particular types of phenomena.
 3869 For example we have seen the Gaussian distribution in Section 6.5. The
 3870 distributions are also related to each other in complex ways (Leemis and
 3871 McQueston, 2008). For a beginner in the field, it can be overwhelming to
 3872 figure out which distribution to use. In addition, many of these distributions
 “Computers” were a job description. 3873 were discovered at a time that statistics and computation was done
 3874 by pencil and paper. It is natural to ask what are meaningful concepts
 3875 in the computing age (Efron and Hastie, 2016). In the previous section,
 3876 we saw that many of the operations required for inference can be conveniently
 3877 calculated when the distribution is Gaussian. It is worth recalling
 3878 at this point the desiderata for manipulating probability distributions in
 3879 the machine learning context.

- 3880 1 There is some “closure property” when applying the rules of probability,
 3881 e.g., Bayes’ theorem. By closure we mean that applying a particular
 3882 operation returns an object of the same type.
- 3883 2 As we collect more data, we do not need more parameters to describe
 3884 the distribution.
- 3885 3 Since we are interested in learning from data, we want parameter estimation
 3886 to behave nicely.

exponential family 3887 It turns out that the class of distributions called the *exponential family*
 3888 provides the right balance of generality while retaining favourable computation
 3889 and inference properties. Before we introduce the exponential family,
 3890 let us see three more members of “named” probability distributions, the Bernoulli
 3891 (Example 6.8), Binomial (Example 6.9) and Beta (Example 6.10) distributions.
 3892

Example 6.8

The *Bernoulli distribution* is a distribution for a single binary random

Bernoulli
distribution

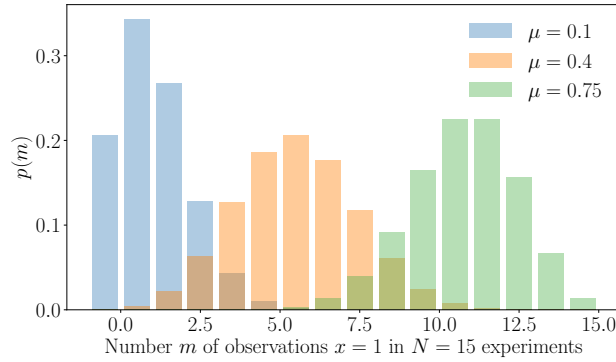


Figure 6.10
Examples of the
Binomial
distribution for
 $\mu \in \{0.1, 0.4, 0.75\}$
and $N = 15$.

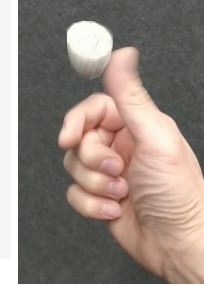
variable X with state $x \in \{0, 1\}$. It is governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $X = 1$. The Bernoulli distribution $\text{Ber}(\mu)$ is defined as

$$p(x | \mu) = \mu^x (1 - \mu)^{1-x}, \quad x \in \{0, 1\}, \quad (6.92)$$

$$\mathbb{E}[x] = \mu, \quad (6.93)$$

$$\mathbb{V}[x] = \mu(1 - \mu), \quad (6.94)$$

where $\mathbb{E}[x]$ and $\mathbb{V}[x]$ are the mean and variance of the binary random variable X .



An example where the Bernoulli distribution can be used is when we are interested in modeling the probability of “head” when flipping a coin.

Remark. The rewriting above of the Bernoulli distribution, where we use Boolean variables as numerical 0 or 1 and express them in the exponents, is a trick that is often used in machine learning textbooks. Another occurrence of this is when expressing the Multinomial distribution. \diamond

Example 6.9

The *Binomial distribution* is a generalization of the Bernoulli distribution to a distribution over integers. In particular, the Binomial can be used to describe the probability of observing m occurrences of $X = 1$ in a set of N samples from a Bernoulli distribution where $p(X = 1) = \mu \in [0, 1]$. The Binomial distribution $\text{Bin}(N, \mu)$ is defined as

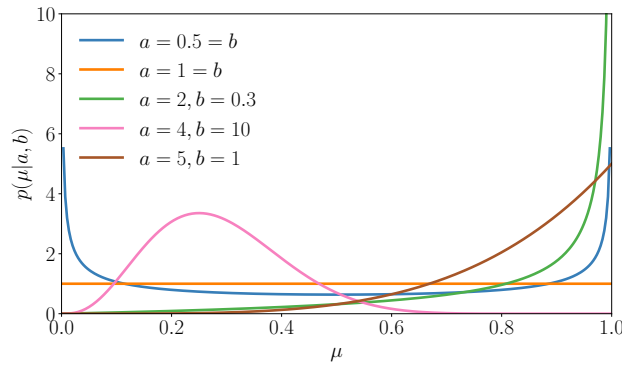
$$p(m | N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}, \quad (6.95)$$

$$\mathbb{E}[m] = N\mu, \quad (6.96)$$

$$\mathbb{V}[m] = N\mu(1 - \mu) \quad (6.97)$$

Binomial
distribution

Figure 6.11
Examples of the
Beta distribution for
different values of α
and β .



where $\mathbb{E}[m]$ and $\mathbb{V}[m]$ are the mean and variance of m , respectively.

3899 An example where the Binomial could be used is if we want to describe
3900 the probability of observing m “heads” in N coin-flip experiments if the
3901 probability for observing head in a single experiment is μ .

Example 6.10

Beta distribution

We may wish to model a continuous random variable on a finite interval. The *Beta distribution* is a distribution over a continuous random variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event (e.g., the parameter governing the Bernoulli distribution). The Beta distribution $\text{Beta}(\alpha, \beta)$ (illustrated in Figure 6.11) itself is governed by two parameters $\alpha > 0$, $\beta > 0$ and is defined as

$$p(\mu | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \quad (6.98)$$

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta}, \quad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (6.99)$$

where $\Gamma(\cdot)$ is the Gamma function defined as

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx, \quad t > 0. \quad (6.100)$$

$$\Gamma(t + 1) = t\Gamma(t). \quad (6.101)$$

Note that the fraction of Gamma functions in (6.98) normalizes the Beta distribution.

3902 Intuitively, α moves probability mass toward 1, whereas β moves prob-
3903 ability mass toward 0. There are some special cases (Murphy, 2012):

- 3904 • For $\alpha = 1 = \beta$ we obtain the uniform distribution $\mathcal{U}[0, 1]$.
- 3905 • For $\alpha, \beta < 1$, we get a bimodal distribution with spikes at 0 and 1.

- For $\alpha, \beta > 1$, the distribution is unimodal.
- For $\alpha, \beta > 1$ and $\alpha = \beta$, the distribution is unimodal, symmetric and centered in the interval $[0, 1]$, i.e., the mode/mean is at $\frac{1}{2}$.

Remark. There is a whole zoo of distributions with names, and they are related in different ways to each other (Leemis and McQueston, 2008). It is worth keeping in mind that each named distribution is created for a particular reason, but may have other applications. Knowing the reason behind the creation of a particular distribution often allows insight into how to best use it. We introduced the above three distributions to be able to illustrate the concepts of conjugacy (Section 6.6.1) and exponential families (Section 6.6.3). \diamond

6.6.1 Conjugacy

According to Bayes' theorem (6.23), the posterior is proportional to the product of the prior and the likelihood. The specification of the prior can be tricky for two reasons: First, the prior should encapsulate our knowledge about the problem before we see any data. This is often difficult to describe. Second, it is often not possible to compute the posterior distribution analytically. However, there are some priors that are computationally convenient: *conjugate priors*.

conjugate priors

Definition 6.13 (Conjugate Prior). A prior is *conjugate* for the likelihood function if the posterior is of the same form/type as the prior.

conjugate

Conjugacy is particularly convenient because we can algebraically calculate our posterior distribution by updating the parameters of the prior distribution.

Remark. When considering the geometry of probability distributions, conjugate priors retain the same distance structure as the likelihood (Agarwal and III, 2010). \diamond

To introduce a concrete example of conjugate priors, we describe below the Binomial distribution (defined on discrete random variables) and the Beta distribution (defined on continuous random variables).

Example 6.11 (Beta-Binomial Conjugacy)

Consider a Binomial random variable $x \sim \text{Bin}(N, \mu)$ where

$$p(x | N, \mu) = \binom{N}{x} \mu^x (1 - \mu)^{N-x} \quad x = 0, 1, \dots, N \quad (6.102)$$

is the probability of finding x times the outcome “head” in N coin flips,

where μ is the probability of a “head”. We place a Beta prior on the parameter μ , that is $\mu \sim \text{Beta}(\alpha, \beta)$ where

$$p(\mu | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \quad (6.103)$$

If we now observe some outcome $x = h$, that is we see h heads in N coin flips, we compute the posterior distribution on μ as

$$p(\mu | x = h, N, \alpha, \beta) \propto p(x | N, \mu) p(\mu | \alpha, \beta) \quad (6.104a)$$

$$= \mu^h (1 - \mu)^{(N-h)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \quad (6.104b)$$

$$= \mu^{h+\alpha-1} (1 - \mu)^{(N-h)+\beta-1} \quad (6.104c)$$

$$\propto \text{Beta}(h + \alpha, N - h + \beta) \quad (6.104d)$$

i.e., the posterior distribution is a Beta distribution as the prior, i.e., the Beta prior is conjugate for the parameter μ in the Binomial likelihood function.

In the following example, we will derive a result that is similar to the Beta-Binomial conjugacy result. Here we will show that the Beta distribution is a conjugate prior for the Bernoulli distribution.

Example 6.12 (Beta-Bernoulli Conjugacy)

Let $x \in \{0, 1\}$ be distributed according to the Bernoulli distribution with parameter $\theta \in [0, 1]$, that is $p(x = 1 | \theta) = \theta$. This can also be expressed as $p(x | \theta) = \theta^x (1 - \theta)^{1-x}$. Let θ be distributed according to a Beta distribution with parameters α, β , that is $p(\theta | \alpha, \beta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}$.

Multiplying the Beta and the Bernoulli distributions, we get

$$p(\theta | x, \alpha, \beta) = p(x | \theta) \times p(\theta | \alpha, \beta) \quad (6.105a)$$

$$\propto \theta^x (1 - \theta)^{1-x} \times \theta^{\alpha-1} (1 - \theta)^{\beta-1} \quad (6.105b)$$

$$= \theta^{\alpha+x-1} (1 - \theta)^{\beta+(1-x)-1} \quad (6.105c)$$

$$\propto p(\theta | \alpha + x, \beta + (1 - x)). \quad (6.105d)$$

The last line above is the Beta distribution with parameters $(\alpha + x, \beta + (1 - x))$.

3936 Table 6.2 lists examples for conjugate priors for the parameters of some
3937 standard likelihoods used in probabilistic modeling. Distributions such as
3938 Multinomial, inverse Gamma, inverse Wishart, and Dirichlet can be found
3939 in any statistical text, and are for example described in Bishop (2006).

3940 The Beta distribution is the conjugate prior for the parameter μ in both
3941 the Binomial and the Bernoulli likelihood. For a Gaussian likelihood func-
3942 tion, we can place a conjugate Gaussian prior on the mean. The reason

Likelihood	Conjugate prior	Posterior
Bernoulli	Beta	Beta
Binomial	Beta	Beta
Gaussian	Gaussian/inverse Gamma	Gaussian/inverse Gamma
Gaussian	Gaussian/inverse Wishart	Gaussian/inverse Wishart
Multinomial	Dirichlet	Dirichlet

Table 6.2 Examples of conjugate priors for common likelihood functions.

why the Gaussian likelihood appears twice in the table is that we need to distinguish the univariate from the multivariate case. In the univariate (scalar) case, the inverse Gamma is the conjugate prior for the variance. In the multivariate case, we use a conjugate inverse Wishart distribution as a prior on the covariance matrix. The Dirichlet distribution is the conjugate prior for the multinomial likelihood function. For further details, we refer to Bishop (2006).

Alternatively, the Gamma prior is conjugate for the precision (inverse variance) in the Gaussian likelihood. Alternatively, the Wishart prior is conjugate for the precision matrix (inverse covariance matrix) in the Gaussian likelihood.

6.6.2 Sufficient Statistics

Recall that a statistic of a random variable is a deterministic function of that random variable. For example if $\mathbf{x} = [x_1, \dots, x_N]^\top$ is a vector of univariate Gaussian random variables, that is $x_n \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean $\hat{\mu} = \frac{1}{N}(x_1 + \dots + x_N)$ is a statistic. Sir Ronald Fisher discovered the notion of *sufficient statistics*: the idea that there are statistics that will contain all available information that can be inferred from data corresponding to the distribution under consideration. In other words sufficient statistics carry all the information needed to make inference about the population, that is they are the statistics that are sufficient to represent the distribution.

sufficient statistics

For a set of distributions parameterized by θ , let X be a random variable with distribution $p(x | \theta_0)$ given an unknown θ_0 . A vector $\phi(x)$ of statistics are called sufficient statistics for θ_0 if they contain all possible information about θ_0 . To be more formal about “contain all possible information”: this means that the probability of x given θ can be factored into a part that does not depend on θ , and a part that depends on θ only via $\phi(x)$. The Fisher-Neyman factorization theorem formalizes this notion, which we state below without proof.

Theorem 6.14 (Fisher-Neyman). *[Theorem 6.5 in Lehmann and Casella (1998)] Let X have probability density function $p(x | \theta)$. Then the statistics $\phi(x)$ are sufficient for θ if and only if $p(x | \theta)$ can be written in the form*

$$p(x | \theta) = h(x)g_\theta(\phi(x)). \quad (6.106)$$

where $h(x)$ is a distribution independent of θ and g_θ captures all the dependence on θ via sufficient statistics $\phi(x)$.

If $p(x | \theta)$ does not depend on θ then $\phi(x)$ is trivially a sufficient statistic for any function ϕ . The more interesting case is that $p(x | \theta)$ is dependent

only on $\phi(x)$ and not x itself. In this case, $\phi(x)$ is a sufficient statistic for x .

In machine learning we consider a finite number of samples from a distribution. One could imagine that for simple distributions (such as the Bernoulli in Example 6.8) we only need a small number of samples to estimate the parameters of the distributions. We could also consider the opposite problem: if we have a set of data (a sample from an unknown distribution), which distribution gives the best fit? A natural question to ask is as we observe more data, do we need more parameters θ to describe the distribution? It turns out that the answer is yes in general, and this is studied in non-parametric statistics (Wasserman, 2007). A converse question is to consider which class of distributions have finite dimensional sufficient statistics, that is the number of parameters needed to describe them do not increase arbitrarily. The answer is exponential family distributions, described in the following section.

6.6.3 Exponential Family

There are three possible levels of abstraction we can have when considering distributions (of discrete or continuous random variables). At level one (the most concrete end of the spectrum), we have a particular named distribution with fixed parameters, for example a univariate Gaussian $\mathcal{N}(0, 1)$ with zero mean and unit variance. In machine learning we often use the second level of abstraction, that is we fix the parametric form (the univariate Gaussian) and infer the parameters from data. For example, we assume a univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$ with unknown mean μ and unknown variance σ^2 , and use a maximum likelihood fit to determine the best parameters (μ, σ^2) . We will see an example of this when considering linear regression in Chapter 9. A third level of abstraction is to consider families of distributions, and in this book, we consider the exponential family. The univariate Gaussian is an example of a member of the exponential family. Many of the widely used statistical models, including all the “named” models in Table 6.2, are members of the exponential family. They can all be unified into one concept (Brown, 1986).

Remark. A brief historical anecdote: like many concepts in mathematics and science, exponential families were independently discovered at the same time by different researchers. In the years 1935–1936, Edwin Pitman in Tasmania, Georges Darrois in Paris, and Bernard Koopman in New York, independently showed that the exponential families are the only families that enjoy finite-dimensional sufficient statistics under repeated independent sampling (Lehmann and Casella, 1998). \diamond

exponential family

An *exponential family* is a family of probability distributions, parameterized by $\theta \in \mathbb{R}^D$, of the form

$$p(x | \theta) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta)) , \quad (6.107)$$

where $\phi(x)$ is the vector of sufficient statistics. In general, any inner product (Section 3.2) can be used in (6.107), and for concreteness we will use the standard dot product here ($\langle \theta, \phi(x) \rangle = \theta^\top \phi(x)$). Note that the form of the exponential family is essentially a particular expression of $g_\theta(\phi(x))$ in the Fisher-Neyman theorem (Theorem 6.14).

The factor $h(x)$ can be absorbed into the dot product term by adding another entry ($\log h(x)$) to the vector of sufficient statistics $\phi(x)$, and constraining the corresponding parameter $\theta_0 = 1$. The term $A(\theta)$ is the normalization constant that ensures that the distribution sums up or integrates to one and is called the *log partition function*. A good intuitive notion of exponential families can be obtained by ignoring these two terms and considering exponential families as distributions of the form

$$p(x | \theta) \propto \exp(\theta^\top \phi(x)). \quad (6.108)$$

For this form of parameterization, the parameters θ are called the *natural parameters*. At first glance it seems that exponential families is a mundane transformation by adding the exponential function to the result of a dot product. However, there are many implications that allow for convenient modelling and efficient computation based on the fact that we can capture information about data in $\phi(x)$.

log partition
function

natural parameters

Example 6.13 (Gaussian as Exponential Family)

Consider the univariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Let $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$. Then by using the definition of the exponential family,

$$p(x | \theta) \propto \exp(\theta_1 x + \theta_2 x^2). \quad (6.109)$$

Setting

$$\theta = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right]^\top \quad (6.110)$$

and substituting into (6.109) we obtain

$$p(x | \theta) \propto \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (6.111)$$

Therefore, the univariate Gaussian distribution is a member of the exponential family with sufficient statistic $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$, and natural parameters given by θ in (6.110).

Example 6.14 (Bernoulli as Exponential Family)

Recall the Bernoulli distribution from Example 6.8

$$p(x | \mu) = \mu^x (1 - \mu)^{1-x}, \quad x \in \{0, 1\}. \quad (6.112)$$

This can be written in exponential family form

$$p(x | \mu) = \exp [\log (\mu^x (1 - \mu)^{1-x})] \quad (6.113)$$

$$= \exp [x \log \mu + (1 - x) \log(1 - \mu)] \quad (6.114)$$

$$= \exp [x \log \mu - x \log(1 - \mu) + \log(1 - \mu)] \quad (6.115)$$

$$= \exp \left[x \log \frac{\mu}{1 - \mu} + \log(1 - \mu) \right]. \quad (6.116)$$

The last line (6.116) can be identified as being in exponential family form (6.107) by observing that

$$h(x) = 1 \quad (6.117)$$

$$\theta = \log \frac{\mu}{1 - \mu} \quad (6.118)$$

$$\phi(x) = x \quad (6.119)$$

$$A(\theta) = -\log(1 - \mu) = \log(1 + \exp(\theta)). \quad (6.120)$$

The relationship between θ and μ is invertible,

$$\mu = \frac{1}{1 + \exp(-\theta)}. \quad (6.121)$$

The relation (6.121) is used to obtain the right equality of (6.120).

sigmoid

Remark. The relationship between the original Bernoulli parameter μ and the natural parameter θ is known as the *sigmoid* or logistic function. Observe that $\mu \in (0, 1)$ but $\theta \in \mathbb{R}$, and therefore the sigmoid function squeezes a real value into the range $(0, 1)$. This property is useful in machine learning, for example it is used in logistic regression (Bishop, 2006, Section 4.3.2), as well as as a nonlinear activation functions in neural networks (Goodfellow et al., 2016, Chapter 6). \diamond

It is often not obvious how to find the parametric form of the conjugate distribution of a particular distribution (for example those in Table 6.2). Exponential families provide a convenient way to find conjugate pairs of distributions. Consider the random variable X is a member of the exponential family (6.107)

$$p(\mathbf{x} | \boldsymbol{\theta}) = h(\mathbf{x}) \exp (\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle - A(\boldsymbol{\theta})) . \quad (6.122)$$

Every exponential family has a conjugate prior (Brown, 1986)

$$p(\boldsymbol{\theta} | \boldsymbol{\gamma}) = h_c(\boldsymbol{\theta}) \exp \left(\left\langle \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\theta} \\ -A(\boldsymbol{\theta}) \end{bmatrix} \right\rangle - A_c(\boldsymbol{\gamma}) \right) , \quad (6.123)$$

where $\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix}$ has dimension $\dim(\boldsymbol{\theta}) + 1$. The sufficient statistics of

the conjugate prior are $\begin{bmatrix} \boldsymbol{\theta} \\ -A(\boldsymbol{\theta}) \end{bmatrix}$. By using the knowledge of the general form of conjugate priors for exponential families, we can derive functional forms of conjugate priors corresponding to particular distributions.

Example 6.15

Recall the exponential family form of the Bernoulli distribution (6.116),

$$p(x | \mu) = \exp \left[x \log \frac{\mu}{1 - \mu} + \log(1 - \mu) \right]. \quad (6.124)$$

The canonical conjugate prior therefore has the same form

$$p(\mu | \gamma, n_0) = \exp \left[n_0 \gamma \log \frac{\mu}{1 - \mu} + n_0 \log(1 - \mu) - A_c(\gamma, n_0) \right], \quad (6.125)$$

which simplifies to

$$p(\mu | \gamma, n_0) = \exp [n_0 \gamma \log \mu + n_0(1 - \gamma) \log(1 - \mu) - A_c(\gamma, n_0)]. \quad (6.126)$$

Putting this in non-exponential family form

$$p(\mu | \gamma, n_0) \propto \mu^{n_0 \gamma} (1 - \mu)^{n_0(1 - \gamma)} \quad (6.127)$$

which is of the same form as the Beta distribution (6.98), with minor manipulations to get the original parametrization (Example 6.12).

Observe that in this example we have derived the form of the Beta distribution by looking at the conjugate prior of the exponential family.

As mentioned in the previous section, the main motivation for exponential families is that they have finite-dimensional sufficient statistics. Additionally, conjugate distributions are easy to write down, and the conjugate distributions also come from an exponential family. From an inference perspective, maximum likelihood estimation behaves nicely because empirical estimates of sufficient statistics are optimal estimates of the population values of sufficient statistics (recall the mean and covariance of a Gaussian). From an optimization perspective, the log-likelihood function is concave allowing for efficient optimization approaches to be applied (Chapter 7).

6.7 Change of Variables/Inverse Transform

It may seem that there are very many known distributions, but in reality the set of distributions for which we have names is quite limited. Therefore, it is often useful to understand how transformed random variables are distributed. For example, assume that X is a random variable distributed according to the univariate normal distribution $\mathcal{N}(0, 1)$, what is

the distribution of X^2 ? Another example, which is quite common in machine learning, is: given that X_1 and X_2 are univariate standard normal, what is the distribution of $\frac{1}{2}(X_1 + X_2)$?

One option to work out the distribution of $\frac{1}{2}(X_1 + X_2)$ is to calculate the mean and variance of X_1 and X_2 and then combine them. As we saw in Section 6.4.4, we can calculate the mean and variance of resulting random variables when we consider affine transformations of random variables. However, we may not be able to obtain the functional form of the distribution under transformations. Furthermore, we may be interested in nonlinear transformations of random variables for which closed-form expressions are not readily available.

Remark (Notation). In this section, we will be explicit about random variables and the values they take. Hence, recall that we use capital letters X, Y to denote random variables and small letters x, y to denote the values in the target space \mathcal{T} that the random variables take. We will explicitly write probability mass functions (pmf) of discrete random variables X as $P(X = x)$. For continuous random variables X (Section 6.2.2), the probability density function (pdf) is written as $f(x)$ and the cumulative distribution function (cdf) is written as $F_X(x)$. \diamond

We will look at two approaches for obtaining distributions of transformations of random variables: a direct approach using the definition of a cumulative distribution function and a change-of-variable approach that uses the chain rule of calculus (Section 5.2.2). The change-of-variable approach is widely used because it provides a “recipe” for attempting to compute the resulting distribution due to a transformation. We will explain the techniques for univariate random variables, and will only briefly provide the results for the general case of multivariate random variables.

Transformations of discrete random variables can be understood directly. Given a discrete random variable X with probability mass function (pmf) $P(X = x)$ (Section 6.2.1), and an invertible function $U(x)$. Consider the transformed random variable $Y := U(X)$, with pmf $P(Y = y)$. Then

$$P(Y = y) = P(U(X) = y) \quad \text{transformation of interest} \quad (6.128a)$$

$$= P(X = U^{-1}(y)) \quad \text{inverse} \quad (6.128b)$$

where we can observe that $x = U^{-1}(y)$. Therefore for discrete random variables, transformations directly change the individual events (with the probabilities appropriately transformed).

6.7.1 Distribution Function Technique

The distribution function technique goes back to first principles, and uses the definition of a cumulative distribution function (cdf) $F_X(x) = P(X \leq x)$ and the fact that its differential is the probability density function (pdf)

Moment generating functions can also be used to study transformations of random variables (Casella and Berger, 2002, Chapter 2).

$f(x)$ (Wasserman, 2004, Chapter 2). For a random variable X and a function U , we find the pdf of the random variable $Y := U(X)$ by

1 Finding the cdf:

$$F_Y(y) = P(Y \leq y) \quad (6.129)$$

2 Differentiating the cdf $F_Y(y)$ to get the pdf $f(y)$.

$$f(y) = \frac{d}{dy} F_Y(y). \quad (6.130)$$

We also need to keep in mind that the domain of the random variable may have changed due to the transformation by U .

Example 6.16

Let X be a continuous random variable with probability density function on $0 \leq x \leq 1$

$$f(x) = 3x^2. \quad (6.131)$$

We are interested in finding the pdf of $Y = X^2$.

The function f is an increasing function of x , and therefore the resulting value of y lies in the interval $[0, 1]$. We obtain

$$F_Y(y) = P(Y \leq y) \quad \text{definition of cdf} \quad (6.132a)$$

$$= P(X^2 \leq y) \quad \text{transformation of interest} \quad (6.132b)$$

$$= P(X \leq y^{\frac{1}{2}}) \quad \text{inverse} \quad (6.132c)$$

$$= F_X(y^{\frac{1}{2}}) \quad \text{definition of cdf} \quad (6.132d)$$

$$= \int_0^{y^{\frac{1}{2}}} 3t^2 dt \quad \text{cdf as a definite integral} \quad (6.132e)$$

$$= [t^3]_{t=0}^{t=y^{\frac{1}{2}}} \quad \text{result of integration} \quad (6.132f)$$

$$= y^{\frac{3}{2}}, \quad 0 \leq y \leq 1. \quad (6.132g)$$

Therefore, the cdf of Y is

$$F_Y(y) = y^{\frac{3}{2}} \quad (6.133)$$

for $0 \leq y \leq 1$. To obtain the pdf, we differentiate the cdf

$$f(y) = \frac{d}{dy} F_Y(y) = \frac{3}{2} y^{\frac{1}{2}} \quad (6.134)$$

for $0 \leq y \leq 1$.

In Example 6.16, we considered a strictly monotonically increasing function $f(x) = 3x^2$. This means that we could compute an inverse function. In general, we require that the function of interest $y = U(x)$ has an inverse $x = U^{-1}(y)$. A useful result can be obtained by considering the

Functions that have inverses are called injective functions (Section 2.7).

4092 cumulative distribution function $F_X(x)$ of a random variable X , and us-
 4093 ing it as the transformation $U(x)$. This leads to the following theorem
 4094 which is called the probability integral transform.

Theorem 6.15. [Theorem 2.1.10 in Casella and Berger (2002)] Let X be a continuous random variable with a strictly monotonic cumulative distribution function $F_X(x)$. Then the random variable Y defined as

$$Y = F_X(x), \quad (6.135)$$

4095 has a uniform distribution.

4096 *Proof* We need to show that the cumulative distribution function of Y
 4097 defines a distribution of a uniform random variable. Recall that by the
 4098 axioms of probability (Section 6.1) probabilities must be non-negative
 4099 and sum/integrate to one. Therefore, the range of possible values of $Y =$
 4100 $F_X(x)$ is the interval $[0, 1]$. For any $F_X(x)$, the inverse $F_X^{-1}(x)$ exists be-
 4101 cause we assumed that $F_X(x)$ is strictly monotonically increasing, which
 4102 we will use in the following.

Given any continuous random variable X , the definition of a cdf gives

$$F_Y(y) = P(Y \leq y) \quad (6.136a)$$

$$= P(F_X(x) \leq y) \quad \text{transformation of interest} \quad (6.136b)$$

$$= P(X \leq F_X^{-1}(y)) \quad \text{inverse exists} \quad (6.136c)$$

$$= F_X(F_X^{-1}(y)) \quad \text{definition of cdf} \quad (6.136d)$$

$$= y, \quad (6.136e)$$

4103 where the last line is due to the fact that $F_X(x)$ composed with its inverse
 4104 results in an identity transformation. The statement $F_Y(y) = y$ along with
 4105 the fact that y lies in the interval $[0, 1]$ means that $F_Y(x)$ is the cdf of the
 4106 uniform random variable on the unit interval. \square

probability integral
transform

4107 Theorem 6.15 is known as the *probability integral transform*, and it is
 4108 used to derive algorithms for sampling from distributions by transforming
 4109 the result of sampling from a uniform random variable (Bishop, 2006).
 4110 The algorithm works by first generating a sample from a uniform distribu-
 4111 tion, then transforming it by the inverse cdf (assuming this is available)
 4112 to obtain a sample from the desired distribution. The probability integral
 4113 transform is also used for hypothesis testing whether a sample comes from
 4114 a particular distribution (Lehmann and Romano, 2005). The idea that the
 4115 output of a cdf gives a uniform distribution also forms the basis of copu-
 4116 las (Nelsen, 2006).

4117 6.7.2 Change of Variables

4118 The distribution function technique in Section 6.7.1 is derived from first
 4119 principles, based on the definitions of cdfs and using properties of in-

verses, differentiation and integration. This argument from first principles relies on two facts:

- 1 We can transform the cdf of Y into an expression that is a cdf of X .
- 2 We can differentiate the cdf to obtain the pdf.

Let us break down the reasoning step by step, with the goal of understanding the more general change of variables approach in Theorem 6.16.

Remark. The name change of variables comes from the idea of changing the variable of integration when faced with a difficult integral. For univariate functions, we use the substitution rule of integration,

$$\int f(g(x))g'(x)dx = \int f(u)du \quad \text{where } u = g(x). \quad (6.137)$$

The derivation of this rule is based on the chain rule of calculus (5.32) and by applying twice the fundamental theorem of calculus. The fundamental theorem of calculus formalizes the fact that integration and differentiation are somehow “inverses” of each other. An intuitive understanding of the rule can be obtained by thinking (loosely) about small changes (differentials) to the equation $u = g(x)$. That is by considering $\Delta u = g'(x)\Delta x$ as a differential of $u = g(x)$. By substituting $u = g(x)$, the argument inside the integral on the right hand side of (6.137) becomes $f(g(x))$. By pretending that the term du can be approximated by $du \approx \Delta u = g'(x)\Delta x$, and that $dx \approx \Delta x$, we obtain (6.137). \diamond

Consider a function of a random variable $Y = U(X)$, where random variable X has states $x \in [a, b]$. By the definition of the cdf, we have

$$F_Y(y) = P(Y \leq y). \quad (6.138)$$

We are interested in a function U of the random variable

$$P(Y \leq y) = P(U(X) \leq y), \quad (6.139)$$

where we assume that the function U is invertible. By applying the inverse U^{-1} to the arguments of $P(U(X) \leq y)$, we obtain

$$P(U(X) \leq y) = P(U^{-1}(U(X)) \leq U^{-1}(y)) = P(X \leq U^{-1}(y)). \quad (6.140)$$

The right most term above is an expression of the cdf of X . Recall the definition of the cdf in terms of the pdf

$$P(X \leq U^{-1}(y)) = \int_a^{U^{-1}(y)} f(x)dx. \quad (6.141)$$

Now we have an expression of the cdf of Y in terms of x :

$$F_Y(y) = \int_a^{U^{-1}(y)} f(x)dx. \quad (6.142)$$

Change of variables in probability relies on the change of variables method in calculus (Tandra, 2014).

To obtain the pdf, we differentiate (6.142) with respect to y .

$$f(y) = \frac{d}{dy} F_y(y) = \frac{d}{dy} \int_a^{U^{-1}(y)} f(x) dx \quad (6.143)$$

Note that the integral on the right hand side is with respect to x , but we need an integral with respect to y because we are differentiating with respect to y . In particular we use (6.137) to get the substitution

$$\int f(U^{-1}(y)) U^{-1'}(y) dy = \int f(x) dx \quad \text{where } x = U^{-1}(y). \quad (6.144)$$

Using (6.144) on the right hand side of (6.143) gives us

$$f(y) = \frac{d}{dy} \int_a^{U^{-1}(y)} f_x(U^{-1}(y)) U^{-1'}(y) dy. \quad (6.145)$$

We then recall that differentiation is a linear operator and we use the subscript x to remind ourselves that $f_x(U^{-1}(y))$ is a function of x and not y . Invoking the fundamental theorem of calculus again gives us

$$f(y) = f_x(U^{-1}(y)) \times \left| \frac{d}{dy} U^{-1}(y) \right|. \quad (6.146)$$

change of variable

This is called the *change of variable* technique. The term $\left| \frac{d}{dy} U^{-1}(y) \right|$ measures how much a unit volume changes when applying U (see also the definition of the Jacobian in Section 5.3). In (6.146) we introduced the absolute value of the differential. For decreasing functions, it turns out that an additional negative sign is needed, and instead of having two types of change-of-variable rules, the absolute value unifies both of them.

Remark. In comparison to the discrete case in (6.128b), we have an additional factor $\left| \frac{d}{dy} U^{-1}(y) \right|$. The continuous case requires more care because $P(Y = y) = 0$ for all y . The probability density function $f(y)$ does not have a description as a probability of an event involving y . \diamond

So far in this section we have been studying univariate change of variables. The case for multivariate random variables is analogous, but complicated by fact that the absolute value cannot be used for multivariate functions. Instead we use the determinant of the Jacobian matrix. Recall from (5.57) that the Jacobian is a matrix of partial derivatives, and that the existence of a non-zero determinant shows that we can invert the Jacobian. Recall the discussion in Section 4.1 that the determinant arises because our differentials (cubes of volume) are transformed into parallelepipeds by the Jacobian. Let us summarize the discussion above in the following theorem, which gives us a recipe for multivariate change of variables.

Theorem 6.16. [Theorem 17.2 in Billingsley (1995)] Let $f(x)$ be the value of the probability density of the multivariate continuous random variable X .

If the vector-valued function $\mathbf{y} = U(\mathbf{x})$ is differentiable and invertible for all values within the domain of \mathbf{x} , then for corresponding values of \mathbf{y} , the probability density of $Y = U(X)$ is given by

$$f(\mathbf{y}) = f_{\mathbf{x}}(U^{-1}(\mathbf{y})) \times \left| \det \left(\frac{\partial}{\partial \mathbf{y}} U^{-1}(\mathbf{y}) \right) \right|. \quad (6.147)$$

4157 The theorem looks intimidating at first glance, but the key point is that
 4158 a change of variable of a multivariate random variable follows the pro-
 4159 cedure of the univariate change of variable. First we need to work out
 4160 the inverse transform, and substitute that into the density of \mathbf{x} . Then we
 4161 calculate the determinant of the Jacobian and multiply the result. The
 4162 following example illustrates the case of a bivariate random variable.

Example 6.17

Consider a bivariate random variable X with states $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and probability density function

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right). \quad (6.148)$$

We use the change-of-variable technique from Theorem 6.16 to derive the effect of a linear transformation (Section 2.7) of the random variable. Consider a matrix $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ defined as

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \quad (6.149)$$

We are interested in finding the probability density function of the transformed bivariate random variable Y with states $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Recall that for change of variables we require the inverse transformation of \mathbf{x} as a function of \mathbf{y} . Since we consider linear transformations, the inverse transformation is given by the matrix inverse (see Section 2.2.2). For 2×2 matrices, we can explicitly write out the formula, given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \quad (6.150)$$

Observe that $ad - bc$ is the determinant (Section 4.1) of \mathbf{A} . The corresponding probability density function is given by

$$f(\mathbf{x}) = f(\mathbf{A}^{-1}\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} \mathbf{y}^\top \mathbf{A}^{-\top} \mathbf{A}^{-1} \mathbf{y}\right). \quad (6.151)$$

The partial derivative of a matrix times a vector with respect to the vector is the matrix itself (Section 5.5) and, therefore,

$$\frac{\partial}{\partial \mathbf{y}} \mathbf{A}^{-1} \mathbf{y} = \mathbf{A}^{-1}. \quad (6.152)$$

Recall from Section 4.1 that the determinant of the inverse is the inverse of the determinant so that the determinant of the Jacobian matrix is given by

$$\det \left(\frac{\partial}{\partial \mathbf{y}} \mathbf{A}^{-1} \mathbf{y} \right) = \frac{1}{ad - bc}. \quad (6.153)$$

We are now able to apply the change-of-variable formula from Theorem 6.16 by multiplying (6.151) with (6.153), which yields

$$f(\mathbf{y}) = f(\mathbf{x}) \left| \det \left(\frac{\partial}{\partial \mathbf{y}} \mathbf{A}^{-1} \mathbf{y} \right) \right| \quad (6.154a)$$

$$= \frac{1}{2\pi} \exp \left(-\frac{1}{2} \mathbf{y}^\top \mathbf{A}^{-\top} \mathbf{A}^{-1} \mathbf{y} \right) |ad - bc|^{-1}. \quad (6.154b)$$

While Example 6.17 is based on a bivariate random variable, which allows to easily compute the matrix inverse, the relation above holds for higher dimensions.

Remark. We saw in Section 6.5 that the density $f(\mathbf{x})$ above is actually the standard Gaussian distribution, and the transformed density $f(\mathbf{y})$ is a bivariate Gaussian with covariance $\Sigma = \mathbf{A} \mathbf{A}^\top$. \diamond

We will use the ideas in this chapter to describe probabilistic modeling in Section 8.3, as well as introduce a graphical language in Section 8.4. We will see direct machine learning applications of these ideas in Chapter 9 and 11.

6.8 Further Reading

This chapter is rather terse at times, Grinstead and Snell (1997) and Walpole et al. (2011) provide more relaxed presentations that are suitable for self study. Readers interested in more philosophical aspects of probability should consider Hacking (2001), whereas a more software engineering approach is presented by Downey (2014). An overview of exponential families can be found in Barndorff-Nielsen (2014). We will see more about how to use probability distributions to model machine learning tasks in Chapter 8. Ironically the recent surge in interest in neural networks has resulted in a broader appreciation of probabilistic models. For example the idea of normalizing flows (Rezende and Mohamed, 2015) relies on change of variables for transforming random variables. An overview of methods for variational inference as applied to neural networks is described in Chapters 16 to 20 of Goodfellow et al. (2016).

We side stepped a large part of the difficulty in continuous random variables by avoiding measure theoretic questions (Billingsley, 1995; Pollard, 2002), and by assuming without construction that we have real numbers,

and ways of defining sets on real numbers as well as their appropriate frequency of occurrence. These details do matter, for example in the specification of conditional probability $p(y|x)$ for continuous random variables x, y (Proschan and Presnell, 1998). The lazy notation hides the fact that we want to specify that $x = x$ (which is a set of measure zero). Furthermore we are interested in the probability density function of y . A more precise notation would have to say $\mathbb{E}_y[f(y) | \sigma(x)]$, where we take the expectation over y of a test function f conditioned on the σ -algebra of x . A more technical audience interested in the details of probability theory have many options (Jacod and Protter, 2004; Jaynes, 2003; MacKay, 2003; Grimmett and Welsh, 2014) including some very technical discussions (Çinlar, 2011; Dudley, 2002; Shirayev, 1984; Lehmann and Casella, 1998; Bickel and Doksum, 2006). An alternative way to approach probability is to start with the concept of expectation, and “work backwards” to derive the necessary properties of a probability space (Whittle, 2000). As machine learning allows us to model more intricate distributions on ever more complex types of data, a developer of probabilistic machine learning models would have to understand these more technical aspects. Machine learning texts with a probabilistic modeling focus include the books by MacKay (2003); Bishop (2006); Murphy (2012); Barber (2012); Rasmussen and Williams (2006).

Exercises

6.1 Consider a mixture of two Gaussian distributions (illustrated in Figure 6.4)

$$0.4\mathcal{N}\left(\begin{bmatrix} 10 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.6\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 2.9 \\ -1.1 & 0.7 \end{bmatrix}\right).$$

- 1 Compute the marginal distributions for each dimension
- 2 Compute the mean, mode and median for each marginal distribution
- 3 Compute the mean and mode for the 2 dimensional distribution

6.2 You have written a computer program that sometimes compiles and sometimes not (code does not change). You decide to model the apparent stochasticity (success vs no success) x of the compiler using a Bernoulli distribution with parameter μ :

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}, \quad x \in \{0, 1\}$$

Choose a conjugate prior for the Bernoulli likelihood and compute the posterior distribution $p(\mu | x_1, \dots, x_N)$.

6.3 There are two bags. The first bag contains 4 mango and 2 apples; the second bag contains 4 mango and 4 apples.

We also have a biased coin, which shows “heads” with probability 0.6 and “tails” with probability 0.4. If the coin shows “heads” we pick a fruit at random from bag 1, otherwise we pick a fruit at random from bag 2.

Your friend flips the coin (you cannot see the result), picks a fruit at random from the corresponding bag, and presents you a mango.

4224 What is the probability that the mango was picked from bag 2?

4225 *Hint: Use Bayes' theorem.*

6.4 Consider the following time-series model:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{w}, & \mathbf{w} &\sim \mathcal{N}(\mathbf{0}, \mathbf{Q}) \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{v}, & \mathbf{v} &\sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \end{aligned}$$

4226 where \mathbf{w}, \mathbf{v} are i.i.d. Gaussian noise variables. Further, assume that $p(\mathbf{x}_0) =$
4227 $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

4228 1 What is the form of $p(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)$? Justify your answer (you do not
4229 have to explicitly compute the joint distribution). (1–2 sentences)

4230 2 Assume that $p(\mathbf{x}_t | \mathbf{y}_1, \dots, \mathbf{y}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$.

4231 a) Compute $p(\mathbf{x}_{t+1} | \mathbf{y}_1, \dots, \mathbf{y}_t)$

4232 b) Compute $p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} | \mathbf{y}_1, \dots, \mathbf{y}_t)$

4233 c) At time $t+1$, we observe the value $\mathbf{y}_{t+1} = \hat{\mathbf{y}}$. Compute $p(\mathbf{x}_{t+1} | \mathbf{y}_1, \dots, \mathbf{y}_{t+1})$.

4234 6.5 Prove the relationship in (6.44), which relates the standard definition of the
4235 variance to the raw score expression for the variance.

4236 6.6 Prove the relationship in (6.45), which relates the pairwise difference be-
4237 tween examples in a dataset with the raw score expression for the variance.

4238 6.7 Express the Bernoulli distribution in the natural parameter form of the ex-
4239ponential family, see (6.107).

4240 6.8 Express the Binomial distribution as an exponential family distribution. Also
4241 express the Beta distribution is an exponential family distribution. Show that
4242 the product of the Beta and the Binomial distribution is also a member of
4243 the exponential family.

4244 6.9 Derive the relationship in Section 6.5.2 in two ways:

4245 1 By completing the square

4246 2 By expressing the Gaussian in its exponential family form

The *product* of two Gaussians $\mathcal{N}(\mathbf{x} | \mathbf{a}, \mathbf{A})\mathcal{N}(\mathbf{x} | \mathbf{b}, \mathbf{B})$ is an unnormalized Gaussian distribution $c\mathcal{N}(\mathbf{x} | \mathbf{c}, \mathbf{C})$ with

$$\mathbf{C} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} \quad (6.155)$$

$$\mathbf{c} = \mathbf{C}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}) \quad (6.156)$$

$$c = (2\pi)^{-\frac{D}{2}} |\mathbf{A} + \mathbf{B}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{a} - \mathbf{b})^\top (\mathbf{A} + \mathbf{B})^{-1}(\mathbf{a} - \mathbf{b})\right). \quad (6.157)$$

4247 Note that the normalizing constant c itself can be considered a (normalized)
4248 Gaussian distribution either in \mathbf{a} or in \mathbf{b} with an “inflated” covariance matrix
4249 $\mathbf{A} + \mathbf{B}$, i.e., $c = \mathcal{N}(\mathbf{a} | \mathbf{b}, \mathbf{A} + \mathbf{B}) = \mathcal{N}(\mathbf{b} | \mathbf{a}, \mathbf{A} + \mathbf{B})$.

6.10 Iterated Expectations.

Consider two random variables x, y with joint distribution $p(x, y)$. Show that:

$$\mathbb{E}_X[x] = \mathbb{E}_Y[\mathbb{E}_X[x | y]]$$

4250 Here, $\mathbb{E}_X[x | y]$ denotes the expected value of x under the conditional distri-
4251bution $p(x | y)$.

6.11 Manipulation of Gaussian Random Variables.

Consider a Gaussian random variable $\mathbf{x} \sim \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, where $\mathbf{x} \in \mathbb{R}^D$. Furthermore, we have

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{w}, \quad (6.158)$$

where $\mathbf{y} \in \mathbb{R}^E$, $\mathbf{A} \in \mathbb{R}^{E \times D}$, $\mathbf{b} \in \mathbb{R}^E$, and $\mathbf{w} \sim \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{Q})$ is independent Gaussian noise. “Independent” implies that \mathbf{x} and \mathbf{w} are independent random variables and that \mathbf{Q} is diagonal.

- 1 Write down the likelihood $p(\mathbf{y} | \mathbf{x})$.
- 2 The distribution $p(\mathbf{y}) = \int p(\mathbf{y} | \mathbf{x})p(\mathbf{x})d\mathbf{x}$ is Gaussian.¹ Compute the mean $\boldsymbol{\mu}_y$ and the covariance $\boldsymbol{\Sigma}_y$. Derive your result in detail.
- 3 The random variable \mathbf{y} is being transformed according to the measurement mapping

$$\mathbf{z} = \mathbf{C}\mathbf{y} + \mathbf{v}, \quad (6.159)$$

where $\mathbf{z} \in \mathbb{R}^F$, $\mathbf{C} \in \mathbb{R}^{F \times E}$, and $\mathbf{v} \sim \mathcal{N}(\mathbf{v} | \mathbf{0}, \mathbf{R})$ is independent Gaussian (measurement) noise.

- Write down $p(\mathbf{z} | \mathbf{y})$.
 - Compute $p(\mathbf{z})$, i.e., the mean $\boldsymbol{\mu}_z$ and the covariance $\boldsymbol{\Sigma}_z$. Derive your result in detail.
- 4 Now, a value $\hat{\mathbf{y}}$ is measured. Compute the posterior distribution $p(\mathbf{x} | \hat{\mathbf{y}})$.²
Hint for solution: Start by explicitly computing the joint Gaussian $p(\mathbf{x}, \mathbf{y})$. This also requires to compute the cross-covariances $\text{Cov}_{\mathbf{x}, \mathbf{y}}[\mathbf{x}, \mathbf{y}]$ and $\text{Cov}_{\mathbf{y}, \mathbf{x}}[\mathbf{y}, \mathbf{x}]$. Then, apply the rules for Gaussian conditioning.

6.12 Probability integral transformation

Given a continuous random variable x , with cdf $F_x(x)$. Show that the random variable $y = F_x(x)$ is uniformly distributed.

¹An affine transformation of the Gaussian random variable \mathbf{x} into $\mathbf{A}\mathbf{x} + \mathbf{b}$ preserves Gaussianity. Furthermore, the sum of this Gaussian random variable and the independent Gaussian random variable \mathbf{w} is Gaussian.

²This posterior is also Gaussian, i.e., we need to determine only its mean and covariance matrix.