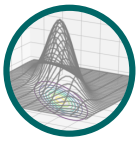
6

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 which 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 state 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 would occlude the main idea. 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, one cannot determine the outcome, but by doing a large number of coin tosses, one 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

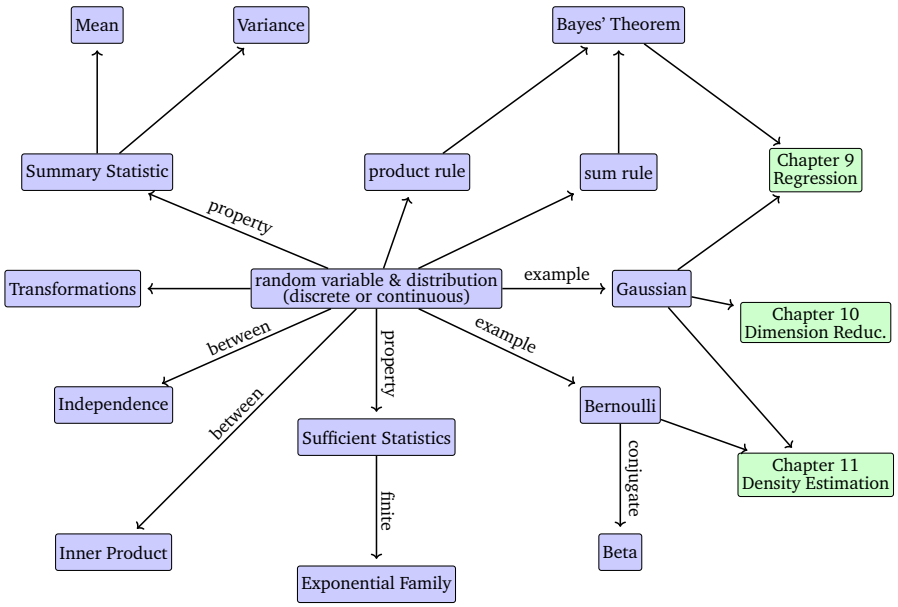


Figure 6.1 A mind map of the concepts related to random variables and probability distributions, as described in this chapter.

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; H1: she is on time, H2: she has been delayed by traffic and H3: she has been abducted by aliens. When we observe our friend is late, we must logically rule out H1. We also tend to consider H2 to be more likely, though we are not logically required to do so. Finally, we may consider H3 to be possible, but we continue to consider it quite unlikely. How do we conclude H2 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).

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

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). The Bayesian interpretation uses probability to specify the degree of uncertainty that the user has about an event, and is sometimes referred to as subjective probability or degree of belief. The frequentist interpretation considers probability to be the relative frequencies of events, in the limit when one has infinite data. ♢

Some machine learning texts on probabilistic models use lazy notation and jargon, which is confusing. 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 continous 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

Modern probability is based on a set of axioms proposed by Kolmogorov (Jacod and Protter, 2004, Chapter 1 and 2) that introduce the three concepts of state space, event space and probability measure. The probability space models a real world process (referred to as an experiment) with random outcomes.

The state space Ω

The *state space* is the set of all possible outcomes of the experiment, usually denoted by Ω. For example, two successive coin tosses have a state space of $\{$hh, tt, ht, th$\}$, where "h" denotes "heads" and "t" denotes "tails". state space

The event space $\mathcal{A}$

The *event space* is the space of potential results of the experiment. A subset A of the state space Ω is in the event space $\mathcal{A}$ if at the end of the experiment we can observe whether a particular state $\omega \in \Omega$ is in A. event space

The event space $\mathcal{A}$ is obtained by considering the collection of subsets of Ω, and for discrete probability distributions (Section 6.2.1) $\mathcal{A}$ is often the powerset of Ω.

The probability P

With each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur. $P(A)$ is called the *probability* of A.

probability

The probability of a single event must lie in the interval $[0, 1]$, and the total probability over all states in the state space Ω must be 1, i.e., $P(\Omega) = 1$. Given a probability space $(\Omega, \mathcal{A}, P)$ we are interested to use it to model some real world phenomenon. In machine learning we often avoid explicitly referring to the probability space, but instead refer to probabilities on quantities of interest which we denote by $\mathcal{T}$. We introduce a function $x : \Omega \to \mathcal{T}$ that takes an element of Ω (an event) and returns a particular quantity of interest, a value in $\mathcal{T}$. For example in the case of tossing two coins and counting the number of heads, a random variable x maps to the three possible outcomes: $x(\mathrm{hh}) = 2$, $x(\mathrm{ht}) = 1$, $x(\mathrm{th}) = 1$ and $x(\mathrm{tt}) = 0$. In this particular case $\mathcal{T} = \{0, 1, 2\}$, and it is the probabilities on elements of $\mathcal{T}$ that we are interested in. This association or mapping from Ω to $\mathcal{T}$ is called a *random variable*. The name "random variable" is a great source of misunderstanding as it is neither random nor is it a variable. It is a function. For a finite state space Ω and finite $\mathcal{T}$, the function corresponding to a random variable is essentially a look up table. For any subset $S \subseteq \mathcal{T}$ we associate $P_x(S) \in [0, 1]$ (the probability) to a particular event occurring corresponding to the random variable x. Example 6.1 provides a concrete example illustrating the above terminology.

random variable

Remark. The state space Ω above unfortunately is referred to by different names in different books. Another common name for Ω is sample space (Grinstead and Snell, 1997; Jaynes, 2003), and state space is sometimes reserved for referring to states in a dynamical system (Hasselblatt and Katok, 2003). Other names sometimes used to describe Ω are: sample description space, possibility space and (very confusingly) event space. ♢

Example 6.1

This toy example is essentially a biased coin flip example.

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 state space Ω to $\mathcal{T}$, that denotes the number of times we draw \$ out of the bag. We can see from the above state space we can get zero \$, one \$ or two \$s, and therefore $\mathcal{T} = \{0, 1, 2\}$. The random variable x (a function or look up table) can be represented as a table like below

$$x((\$, \$)) = 2 \tag{6.1}$$
$$x((\$, £)) = 1 \tag{6.2}$$
$$x((£, \$)) = 1 \tag{6.3}$$
$$x((£, £)) = 0. \tag{6.4}$$

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

$$\begin{aligned} P(x = 2) &= P((\$, \$)) \\ &= P(\$) \times P(\$) \\ &= 0.3 \times 0.3 = 0.09 \end{aligned} \tag{6.5}$$
$$\begin{aligned} P(x = 1) &= P((\$, £) \cup (£, \$)) \\ &= P((\$, £)) + P((£, \$)) \\ &= 0.3 \times (1 - 0.3) + (1 - 0.3) \times 0.3 = 0.42 \end{aligned} \tag{6.6}$$
$$\begin{aligned} P(x = 0) &= P((£, £)) \\ &= P(£) \times P(£) \\ &= (1 - 0.3) \times (1 - 0.3) = 0.49. \end{aligned} \tag{6.7}$$

In the above calculation, notice that we have equated two different concepts, the probability of the output of x and the probability of the states in Ω. For example in (6.7) we say $P(x = 0) = P((£, £))$.

Consider the random variable $x : \Omega \to \mathcal{T}$ 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, that is the set of elements of Ω that map to S under x; $\{\omega \in \Omega : x(\omega) \in S\}$. Another 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). That is for $S \subseteq \mathcal{T}$, we have the

following notation

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

The left hand side of (6.8) is the probability of the set of possible outcomes (e.g. number of heads = 1) that we are interested in. Via the random variable x that maps states to outcomes, we see in the right hand side of (6.8) that this is the probability of the set of states (in Ω) that have the property (e.g. ht, th). We say that a random variable x is distributed according to a particular probability distribution P_x, which defines the probability mapping between the event and the probability of the outcome of the random variable. The two concepts are intertwined, but for ease of presentation we will discuss some properties with respect to random variables and others with respect to their distributions.

Remark. The range of the random variable $\mathcal{T}$ is used to indicate the kind of probability space, that is a $\mathcal{T}$ random variable. When $\mathcal{T}$ is finite or countably infinite, this is called a discrete random variable (Section 6.2.1). For continuous random variables (Section 6.2.2) we only consider $\mathcal{T} = \mathbb{R}$ or $\mathcal{T} = \mathbb{R}^d$. ♢

6.1.3 Statistics

Probability theory and statistics are often presented together, but they concern different aspects of uncertainty. One way of contrasting them is by the kinds of problems that are considered. Using probability we can consider a model of some process where the underlying uncertainty is captured by random variables, and we use the rules of probability to derive what happens. Using statistics we observe that something has happened, and try to figure out the underlying process that explains the observations. In this sense machine learning is close to statistics in its goals, that is to construct a model that adequately represents the process that generated the data. When the machine learning model is a probabilistic model, we can use the rules of probability to calculate the "best fitting" model for some 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.

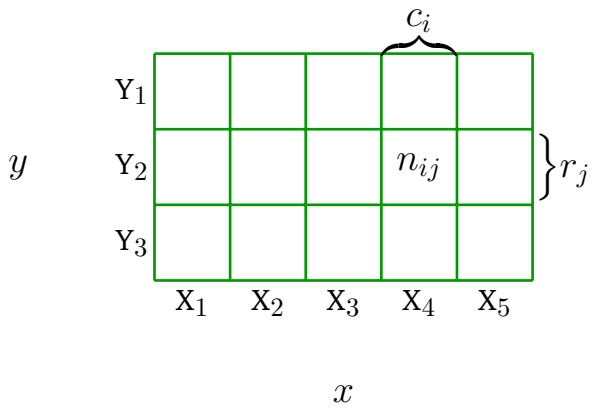


Figure 6.2 Visualization of a discrete bivariate probability mass function, with random variables x and y. This diagram is from Bishop (2006).

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 state 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 $\mathrm{x} \in \mathcal{T}$, denoted as $P(x = \mathrm{x})$. The expression $P(x = \mathrm{x})$ 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. By convention we specify the probability that a random variable x is less than a particular value x, denoted $P(x \leqslant \mathrm{x})$. The expression $P(x \leqslant \mathrm{x})$ 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.

Many probability textbooks tend to use capital letters X for random variables and small letters x for their values.
probability mass function
cumulative distribution function

Remark. We will use the phrase *univariate* distribution to refer to distributions of only one random variable (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 (denoted by bold $\boldsymbol{x}$). ♢

univariate
multivariate

6.2.1 Discrete Probabilities

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

joint probability

$$P(x = \mathrm{x}_i, y = \mathrm{y}_i) = \frac{n_{ij}}{N}\,, \tag{6.9}$$

where n_{ij} is the number of events with $\mathtt{x}_i$ and $\mathtt{y}_j$ and N the total number of events. The joint probability is probability of the intersection of both events, that is $P(x = \mathtt{x}_i, y = \mathtt{y}_i) = P(x = \mathtt{x}_i \cap y = \mathtt{y}_i)$. Figure 6.2 illustrates the *probability mass function* (pmf) of a discrete probability distribution. For two random variables x and y, the probability that $x = \mathtt{x}$ and $y = \mathtt{y}$ is (lazily) written as $p(x, y)$ and is called the joint probability. The *marginal probability* is obtained by summing over a row or column. The *conditional probability* is the fraction of a row or column in a particular cell.

probability mass function

marginal probability

conditional probability

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. The value c_i is the sum of the individual frequencies for the i$^{\text{th}}$ column, that is $c_i = \sum_{j=1}^{3} n_{ij}$. Simularly, 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 by themselves.

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

$$P(x = \mathtt{x}_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N} \tag{6.10}$$

and

$$P(y = \mathtt{y}_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N}, \tag{6.11}$$

where c_i and r_j are the ith column and jth row of the probability table, respectively. By convention for discrete random variables with a finite number of events, we assume that probabilties sum up to one, that is

$$\sum_{i=1}^{5} P(x = \mathtt{x}_i) = 1 \quad \text{and} \quad \sum_{j=1}^{3} P(y = \mathtt{y}_j) = 1\,. \tag{6.12}$$

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

$$P(y = \mathtt{y}_j \,|\, x = \mathtt{x}_i) = \frac{n_{ij}}{c_i}\,, \tag{6.13}$$

and the conditional probability of x given y is

$$P(x = \mathtt{x}_i \,|\, y = \mathtt{y}_j) = \frac{n_{ij}}{r_j}\,. \tag{6.14}$$

The marginal probability that x takes the value $\mathtt{x}$ irrespective of the value of random variable y is (lazily) written as $p(x)$. If we consider only

the instances where $x = \mathtt{x}$, then the fraction of instances (the conditional probability) for which $y = \mathtt{y}$ is written (lazily) as $p(y \mid x)$.

In machine learning, we use discrete probability distributions to model *categorical variables*, i.e., variables that take a finite set of unordered values. These 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 handwritten recognition. Discrete distributions are also often used to construct probabilistic models that combine a finite number of continuous distributions. We will see the Gaussian mixture model in Chapter 11.

categorical variables

6.2.2 Continuous Probabilities

We consider real valued random variables in this section, that is we consider state spaces which are intervals of the real line $\mathbb{R}$. In this book we will 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 error 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 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 furthermore have a topology are called a Borel σ-algebras. Betancourt (2018) details a careful construction of probability spaces from set theory without being bogged down in technicalities. A reader interested in a more precise construction is referred to Jacod and Protter (2004); Billingsley (1995). Further references can be found in the further reading section. In this book, we consider real valued random variables with their corresponding Borel σ-algebra. We consider random variables with values in $\mathbb{R}^d$ to be a vector of real valued random variables. ♢

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

probability density function

1 $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geqslant 0$

2 Its integral exists and

$$\int_{\mathbb{R}^D} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1\,. \tag{6.15}$$

Here, $\boldsymbol{x} \in \mathbb{R}^D$ is a (continuous) random variable. For probability mass functions (pmf) of discrete random variables the integral in (6.15) is replaced with a sum (see (6.12)).

$P(x = \mathrm{x}$ is a set of measure zero.)

In contrast to discrete random variables, the probability of a continuous random variable x taking a particular value $P(x = \mathrm{x})$ is zero. However this does not mean that such events never occur.

cumulative distribution function

Definition 6.2 (Cumulative Distribution Function)**.** A *cumulative distribution function* (cdf) of a multivariate real-valued random variable $\boldsymbol{x} \in \mathbb{R}^D$ is given by

$$F_{\boldsymbol{x}}(\mathrm{x}) = P(x_1 \leqslant \mathrm{x}_1, \ldots, x_D \leqslant \mathrm{x}_D)\,, \tag{6.16}$$

where the right-hand side represents the probability that random variable x_i takes the value smaller than or equal to x_i. This can be expressed also as the integral of the probability density function so that

$$F_{\boldsymbol{x}}(\mathrm{x}) = \int_{-\infty}^{\mathrm{x}} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,. \tag{6.17}$$

6.2.3 Contrasting Discrete and Continuous Distributions

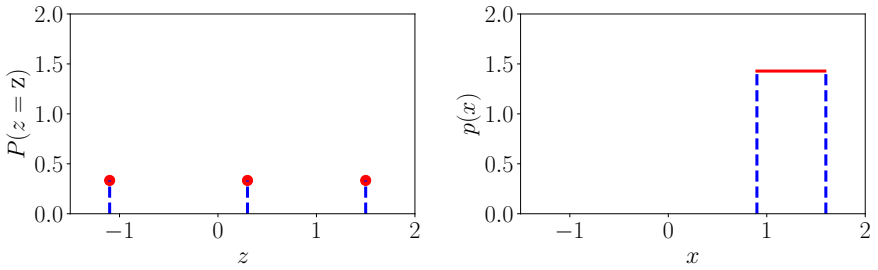
Let us consider both discrete and continuous distributions and contrast them. The aim here is to see that while both discrete and continuous distributions seem to have similar requirements, such as the total probability mass is 1, they are subtly different. Since the total probability mass of a discrete random variable is 1 (see (6.12)) and there are a finite number of states, the probability of each state must be in the interval $[0, 1]$. However, the analogous requirement for continuous random variables (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 the difference 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



(a) Discrete distribution (b) Continuous distribution

ure 6.3 mples of discrete l continuous form tributions. See mple 6.3 for ails of the tributions.

the states. The probability mass function can be represented as a table of probability values.

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

Alternatively, one could 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 is it the same as in Figure 6.3(b).

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

$$\int_{0.9}^{1.6} p(x)\mathrm{d}x = 1\,. \tag{6.18}$$

Remark. There is an additional subtlety with regards to discrete probability distributions. The states $\mathrm{x}_1, \ldots, \mathrm{x}_d$ do not in principle have any structure, that is there is usually no way to compare them, for example $\mathrm{x}_1 = \text{red}, \mathrm{x}_2 = \text{green}, \mathrm{x}_3 = \text{blue}$. However in many machine learning applications the discrete states take numerical values, for example $\mathrm{x}_1 = -1.1, \mathrm{x}_2 = 0.3, \mathrm{x}_3 = 1.5$, where we could say $\mathrm{x}_1 < \mathrm{x}_2 < \mathrm{x}_3$. Discrete states which take numerical values are particularly useful because we often consider expected values (Section 6.4.1) of random variables. ♢

Unfortunately machine learning literature uses notation and nomenclature that hides the distinction between the state space, the set of outcomes and the random variable. For a value x of the set of possible outcomes of the random variable x, that is $\mathrm{x} \in \mathcal{T}$, $p(x)$ denotes the probability that random variable x has the outcome x. For discrete random variables this is written as $P(x = \mathrm{x})$, which is known as the probability mass func-

Table 6.1 Nomenclature for probability distributions.

	"point probability"	"interval probability"
discrete	$P(x = \mathrm{x})$ probability mass function	not applicable
continuous	$p(x)$ probability density function	$P(x \leqslant \mathrm{x})$ cumulative distribution function

tion. This is often referred to as the "distribution". For continuous variables, $p(x)$ is called the probability density function (often referred to as a density), and to muddy things even further the cumulative distribution function $P(x \leqslant \mathrm{x})$ is often also referred to as the "distribution". In this chapter, we will use the notation x or $\boldsymbol{x}$ to refer to univariate and multivariate random variables respectively. We summarise 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. Unfortunately the majority of machine learning literature is also sloppy about the phrase. ♢

6.3 Sum Rule, Product Rule and Bayes' Theorem

When we think of a probabilistic model 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 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, that govern probabilistic inference.

Given the definitions of marginal and conditional probability for discrete and continuous random variables in the previous section, we can now present the two fundamental rules in probability theory. These two rules arise naturally (Jaynes, 2003) from the requirements we discussed in Section 6.1.1. Recall from (6.9) that $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of the two random variables $\boldsymbol{x}, \boldsymbol{y}$, $p(\boldsymbol{x}), p(\boldsymbol{y})$ are the corresponding marginal distributions, and $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$.

sum rule

The first rule, the *sum rule*, is

$$p(\boldsymbol{x}) = \begin{cases} \sum_{\mathrm{Y} \in \boldsymbol{y}} p(\boldsymbol{x}, \mathrm{Y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \int_{\mathrm{Y}} p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases} \tag{6.19}$$

so that we sum out (or integrate out) the set of states Y of the random variable $\boldsymbol{y}$. The sum rule is also known as the *marginalization property*.

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 $\boldsymbol{x} = (x_1, \dots, x_D)$, we obtain the marginal

$$p(x_i) = \int p(x_1, \dots, x_D) \mathrm{d}\boldsymbol{x}_{\setminus i} \tag{6.20}$$

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

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. ♢

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

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \,. \tag{6.21}$$

The product rule can be interpreted as the fact that every joint distribution of two random variables can be factorized (written as a product) of two other distributions. The two factors are the marginal distribution of the first random variable $p(\boldsymbol{x})$, and the conditional distribution of the second random variable given the first $p(\boldsymbol{y} \mid \boldsymbol{x})$. Since the ordering of random variables is arbitrary in $p(\boldsymbol{x}, \boldsymbol{y})$ the product rule also implies $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y})$. To be precise, (6.21) is expressed in terms of the probability mass functions for discrete random variables. For continuous random variables, the product rule is expressed in terms of the probability density functions (Section 6.2.3).

Let us briefly explore how to use probabilistic models to capture uncertainty (Ghahramani, 2015). At the lowest modeling level, measurement noise introduces model uncertainty, for example a camera sensor introduces random error in the value of each pixel it records. We will see in Chapter 9 how to use Gaussian (Section 6.5) noise models for linear regression. At higher modeling levels, we would be interested in modeling the uncertainty of the coefficients in linear regression. This uncertainty captures which values of these parameters will be good at predicting new data. Finally at the highest levels, we may want to capture uncertainties about the model structure. We will discuss model choice in Section 8.5. Once we have the probabilistic models (described in Section 8.3), the basic rules of probability presented in this section are used to infer the unobserved quantities given the observed data.

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 $p(\boldsymbol{x})$ about an unobserved random variable $\boldsymbol{x}$ and some relationship $p(\boldsymbol{y} \mid \boldsymbol{x})$ between $\boldsymbol{x}$ and a second random variable $\boldsymbol{y}$, which we can observe. If we observe $\boldsymbol{y}$ we can use Bayes' theorem to draw some conclusions about $\boldsymbol{x}$ given the observed values of $\boldsymbol{y}$. *Bayes' theorem* (also: *Bayes' rule* or *Bayes' law*)

Bayes' theorem is also called the "probabilistic inverse"

Bayes' theorem

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}} \tag{6.22}$$

is a direct consequence of the product rule in (6.19) since

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y}) \tag{6.23}$$

and

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \tag{6.24}$$

so that

$$p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \iff p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x})}{p(\boldsymbol{y})} . \tag{6.25}$$

prior

In (6.22), $p(\boldsymbol{x})$ is the *prior*, which encapsulates our subjective prior knowledge of the unobserved (latent) variable $\boldsymbol{x}$ before observing any data. We can choose any prior that makes sense to us, but it is critical to ensure that the prior has a non-zero pdf (or pmf) on all plausible $\boldsymbol{x}$, even if they are very rare.

likelihood

The *likelihood* $p(\boldsymbol{y} \mid \boldsymbol{x})$ describes how $\boldsymbol{x}$ and $\boldsymbol{y}$ are related, and it is the probability of the data $\boldsymbol{y}$ if we were to know the latent variable $\boldsymbol{x}$. Note that the likelihood is not a distribution in $\boldsymbol{x}$, but only in $\boldsymbol{y}$. We call $p(\boldsymbol{y} \mid \boldsymbol{x})$ either the "likelihood of $\boldsymbol{x}$ (given $\boldsymbol{y}$)" or the "probability of $\boldsymbol{y}$ given $\boldsymbol{x}$" but never the likelihood of $\boldsymbol{y}$ (MacKay, 2003a).

The likelihood is sometimes also called the "measurement model".

posterior

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

The quantity

$$p(\boldsymbol{y}) := \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{y} \mid \boldsymbol{x})] \tag{6.26}$$

marginal likelihood

evidence

is the *marginal likelihood/evidence*. By definition the marginal likelihood integrates the numerator of (6.22) with respect to the latent variable $\boldsymbol{x}$. Therefore, the marginal likelihood is independent of $\boldsymbol{x}$ and it ensures that the posterior $p(\boldsymbol{x} \mid \boldsymbol{y})$ is normalized. The marginal likelihood can also be interpreted as the expected likelihood where we take the expectation with respect to the prior $p(\boldsymbol{x})$. 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 (8.41), the evidence is often hard to compute.

Bayes' theorem in (6.22) allows us to invert the causal relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ given by the likelihood. Therefore, Bayes' theorem is sometimes 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 9.2.3. However, focusing on the 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 around 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 around in a downstream task can be very useful. 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 main tool we use to compute statistics of a random variable is its expected value with respect to a particular function.

expected value

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

$$\mathbb{E}_x[g(x)] = \int g(x)p(x)\mathrm{d}x\,. \tag{6.27}$$

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

$$\mathbb{E}_x[g(x)] = \sum_{x\in\mathcal{T}} g(x)p(x) \tag{6.28}$$

where $\mathcal{T}$ is the set of possible outcomes of the random variable x.

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

Remark. We consider multivariate random variables $\boldsymbol{x}$ as a finite vector of univariate random variables $[x_1, \ldots, x_n]^\top$. For multivariate random variables, we define the expected value element wise

$$\mathbb{E}_{\boldsymbol{x}}[g(\boldsymbol{x})] = \begin{bmatrix} \mathbb{E}_{x_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{x_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D\,, \tag{6.29}$$

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 $\boldsymbol{x}$. ♢

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).

Definition 6.3 defines the meaning of the notation $\mathbb{E}_x$ and $\mathbb{E}_{\boldsymbol{x}}$ 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.

mean

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

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{x_1}[x_1] \\ \vdots \\ \mathbb{E}_{x_D}[x_D] \end{bmatrix} \in \mathbb{R}^D\,, \tag{6.30}$$

where

$$\mathbb{E}_{x_d}[x_d] := \begin{cases} \displaystyle\int x_d p(x_d)\mathrm{d}x_d & \text{if } x_d \text{ has a continuous domain} \\ \displaystyle\sum_{\mathtt{x}_i} \mathtt{x}_i p(x_d = \mathtt{x}_i) & \text{if } x_d \text{ has a discrete domain} \end{cases} \tag{6.31}$$

for $d = 1, \ldots, D$, where the subscript d indicates the corresponding dimension of $\boldsymbol{x}$, and the integral and sum are over the states $\mathtt{x}_i$.

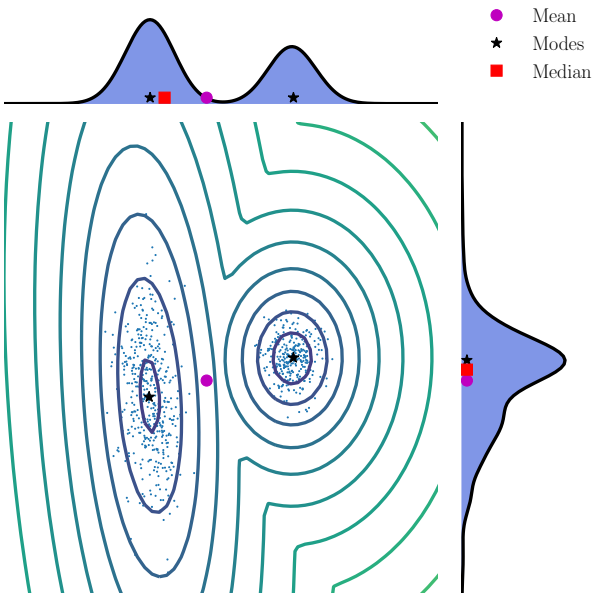


Figure 6.4 Illustration of the mean, mode and median for a two-dimensional dataset, as well as its marginal densities.

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 random variable, the mode is defined as a peak in the density $p(\boldsymbol{x})$. A particular density $p(\boldsymbol{x})$ may have more than one mode, and, therefore, finding the mode may be computationally challenging in high dimensions.

median

mode

mode

Example 6.4
Consider the 2 dimensional distribution illustrated in Figure 6.4

$$\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 1.5\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0 & 2.9\\-1.1 & 0.7\end{bmatrix}\right). \tag{6.32}$$

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 the fact that the mean and median can be quite 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 $<$ such that $\begin{bmatrix}1\\0\end{bmatrix} < \begin{bmatrix}2\\3\end{bmatrix}$.

The mean is recovered if we set the function g in Definition 6.3 to the identity function. This indicates that we can think about functions of random variables, which we will revisit in Section 6.7.

Remark. The expected value is a linear operator. For example, given multivariate real-valued functions $f(\boldsymbol{x}) = ag(\boldsymbol{x}) + bh(\boldsymbol{x})$ where $a, b \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} && \text{(6.33a)}\\ &= \int [ag(\boldsymbol{x}) + bh(\boldsymbol{x})]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} && \text{(6.33b)}\\ &= a\int g(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}x + b\int h(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} && \text{(6.33c)}\\ &= a\mathbb{E}_{\boldsymbol{x}}[g(\boldsymbol{x})] + b\mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x})]\,. && \text{(6.33d)}\end{aligned}$$

♢

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

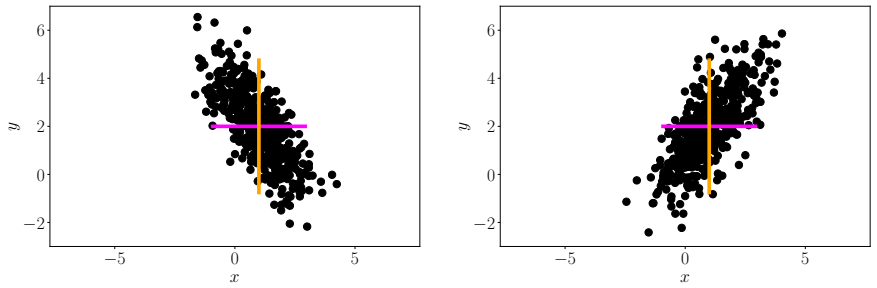
$$\mathrm{Cov}[x, y] = \mathbb{E}\big[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\big]\,. \tag{6.34}$$

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.,

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]\,. \tag{6.35}$$

variance
standard deviation
The covariance of a variable with itself $\mathrm{Cov}[x, x]$ is called the *variance* and is denoted by $\mathbb{V}[x]$. The square root of the variance is called the *standard deviation* and is often denoted by $\sigma(x)$.

When we want to compare the covariances between different pairs of



(a) x and y are negatively correlated.

(b) x and y are positively correlated.

Figure 6.5 Two-dimensional datasets with identical means and variances along each axis (colored lines) but with different covariances.

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.

Definition 6.6 (Correlation)**.** The *correlation* between two random variables x, y is given by

correlation

$$\operatorname{corr}[x, y] = \frac{\operatorname{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}}. \tag{6.36}$$

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 $\operatorname{corr}[x, y]$ means that when x grows then y is also expected to grow. Negative correlation means that as x increases then y decreases.

The notion of covariance can be generalised to multivariate random variables.

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

Definition 6.7 (Covariance)**.** If we consider two random variables $\boldsymbol{x} \in \mathbb{R}^D, \boldsymbol{y} \in \mathbb{R}^E$, the *covariance* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

covariance

$$\operatorname{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}]^\top = \operatorname{Cov}[\boldsymbol{y}, \boldsymbol{x}]^\top \in \mathbb{R}^{D\times E}. \tag{6.37}$$

Here, the subscript makes it explicit with respect to which variable we need to average.

Definition 6.7 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.

Definition 6.8 (Variance)**.** The *variance* of a random variable $\boldsymbol{x} \in \mathbb{R}^D$ with mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ is defined as

variance

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]^\top \tag{6.38a}$$

$$= \begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \dots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \dots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \dots & \dots & \mathrm{Cov}[x_D, x_D] \end{bmatrix} \in \mathbb{R}^{D \times D} . \tag{6.38b}$$

covariance matrix 3461 This matrix is called the *covariance matrix* of the random variable $\boldsymbol{x}$. 3462 The covariance matrix is symmetric and positive definite and tells us some- 3463 thing about the spread of the data.

marginal The covariance matrix contains the variances of the *marginals*

$$p(x_i) = \int p(x_1, \dots, x_D) \mathrm{d}x_{\setminus i} \tag{6.39}$$

on its diagonal, where "$\setminus i$" denotes "all variables but i". The off-diagonal

cross-covariance entries are the *cross-covariance* terms $\mathrm{Cov}[x_i, x_j]$ for $i, j = 1, \dots, D,\ i \neq j$. It generally holds that

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathrm{Cov}_{\boldsymbol{x}}[\boldsymbol{x}, \boldsymbol{x}] . \tag{6.40}$$

3464 We will revisit the idea of covariance again in Section 6.4.5.

3465 6.4.2 Empirical Means and Covariances

population mean 3466 The definitions in Section 6.4.1 are often also called the *population mean*
and covariance 3467 *and covariance*, as it refers to the true statistics for the population. In ma- 3468 chine learning we need to learn from empirical observations of data. Con- 3469 sider a random variable x. There are two conceptual steps to go from 3470 population statistics to the realization of empirical statistics. First we use 3471 the fact that we have a finite dataset (of size N) to construct an empir- 3472 ical statistic which is a function of a finite number of identical random 3473 variables, $x_1, \dots, x_N$. Second we observe the data, that is we look at the 3474 realization of each of the random variables $\mathrm{x}_1, \dots, \mathrm{x}_N$ and apply the em- 3475 pirical statistic.

3476 Specifically for the mean (Definition 6.4), given a particular dataset we
empirical mean 3477 can obtain an estimate of the mean, which is called the *empirical mean* or
sample mean 3478 *sample mean*. The same holds for the empirical covariance.

empirical 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

$$\bar{\boldsymbol{x}} := \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n , \tag{6.41}$$

3479 where $\boldsymbol{x}_n \in \mathbb{R}^D$.

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

matrix

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

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.

In the above definition, we have expressed the statistic in terms of the multivariate random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. To compute the statistics for a particular dataset, we would use the realizations (observations) $\mathtt{x}_1, \ldots, \mathtt{x}_N$ and use (6.41) and (6.42). Empirical covariance matrices are symmetric, positive semi-definite (see Section 3.2.3).

Remark. The notation we use in this book is imprecise in the sense that we do not explicitly make a distinction between the random variable $\boldsymbol{x}_n$ and its deterministic realization $\mathtt{x}_n$. ♢

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.,

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

Depending on whether x is a discrete or continuous random variable, the expectation in (6.43) and the mean $\mu = \mathbb{E}_x(x)$ are computed using (6.31). The variance as expressed in (6.43) is the mean of a new random variable $z := (x-\mu)^2$.

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*:

raw-score formula for variance

$$\mathbb{V}_x[x] = \mathbb{E}_x[x^2] - \mathbb{E}_x[x]^2\,. \tag{6.44}$$

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. 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).

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

A third way to understand the variance is that it is a sum of pairwise differences between all pairs of observations. Consider a sample $x_1, \ldots, x_N$

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,

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

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$. It holds that

$$\mathbb{E}[\boldsymbol{x}+\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{y}] \tag{6.46}$$
$$\mathbb{E}[\boldsymbol{x}-\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{y}] \tag{6.47}$$
$$\mathbb{V}[\boldsymbol{x}+\boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}] \tag{6.48}$$
$$\mathbb{V}[\boldsymbol{x}-\boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}] \tag{6.49}$$

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

$$\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\,, \tag{6.50}$$
$$\mathbb{V}_{\boldsymbol{y}}[\boldsymbol{y}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\,, \tag{6.51}$$

respectively. Furthermore,

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

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}]^\top &(6.52)\\ &= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^\top + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{A}^\top &(6.53)\\ &= \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{A}^\top &(6.54)\end{aligned}$$

$$\overset{(6.38a)}{=} \boldsymbol{\Sigma}\boldsymbol{A}^\top , \tag{6.55}$$

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

6.4.5 Statistical Independence

Definition 6.10 (Independence)**.** Two random variables $\boldsymbol{x}, \boldsymbol{y}$ are *statistically independent* if and only if

statistically independent

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}) . \tag{6.56}$$

Intuitively, two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent if the value of $\boldsymbol{y}$ (once known) does not add any additional information about $\boldsymbol{x}$ (and vice versa).

If $\boldsymbol{x}, \boldsymbol{y}$ are (statistically) independent then

- $p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$
- $p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$
- $\mathbb{V}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] + \mathbb{V}_{\boldsymbol{y}}[\boldsymbol{y}]$
- $\mathrm{Cov}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{0}$

The last point above may not hold in converse, that is 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 non-linearly 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 (6.35) between x and y. But this gives

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x^3] = 0. \tag{6.57}$$

In machine learning we often consider problems that can be modelled as *independent and identically distributed* random variables, $x_1, \ldots, x_N$. The word "independent" refers to Definition 6.10, that is any pair of random variables x_i and x_j are independent. The phrase identically distributed means that all the random variables are from the same distribution.

independent and identically distributed

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

Definition 6.11 (Conditional Independence)**.** Formally, two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are *conditionally independent given* $\boldsymbol{z}$ if and only if

conditionally independent given $\boldsymbol{z}$

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{y} \mid \boldsymbol{z}) \quad \text{for all} \quad z \in \mathcal{A} . \tag{6.58}$$

We write $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{z}$.

Definition 6.11 requires that the relation in (6.58) must hold true for every value of $\boldsymbol{z}$. The interpretation of (6.58) can be understood as "given knowledge about $\boldsymbol{z}$, the distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ factorizes". Independence can be cast as a special case of conditional independence if we write $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \emptyset$.

By using the product rule of probability from (6.21) we can expand the left-hand side of (6.58) to obtain

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) p(\boldsymbol{y} \mid \boldsymbol{z}) \,. \tag{6.59}$$

By comparing the right-hand side of (6.58) with (6.59) we see that $p(\boldsymbol{y} \mid \boldsymbol{z})$ appears in both of them so that

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}) \,. \tag{6.60}$$

Equation (6.60) provides an alternative definition of conditional independence, i.e., $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{z}$. This alternative presentation provides the interpretation "given that we know $\boldsymbol{z}$, knowledge about $\boldsymbol{y}$ does not change our knowledge of $\boldsymbol{x}$".

6.4.6 Inner Products of Random Variables

Recall the definition of inner products from Section 3.2. Another example for defining an inner product between unusual types are random variables or random vectors. If we have two uncorrelated random variables x, y then

$$\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y] \tag{6.61}$$

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

Inner products between multivariate random variables $\boldsymbol{x}, \boldsymbol{y}$ can be treated in a similar fashion

In the following, we see whether we can find a geometric interpretation of the variance relation of uncorrelated random variables in (6.61). 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

$$\langle x, y \rangle := \mathrm{Cov}[x, y] \tag{6.62}$$

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

$\mathrm{Cov}[x, x] > 0$ and $0 \iff x = 0$
$\mathrm{Cov}[\alpha x + z, y] = \alpha \,\mathrm{Cov}[x, y] + \mathrm{Cov}[z, y]$ for $\alpha \in \mathbb{R}$.

$$\|x\| = \sqrt{\mathrm{Cov}[x, x]} = \sqrt{\mathbb{V}[x]} = \sigma[x] \,, \tag{6.63}$$

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

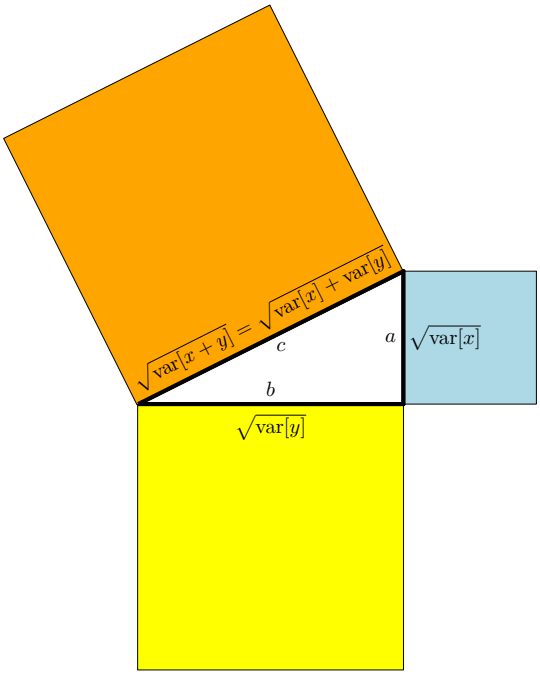


Figure 6.6 Geometry of random variables. If random variables x and y are uncorrelated they are orthogonal vectors in a corresponding vector space, and the Pythagorean theorem applies.

If we look at the angle θ between random two random variables x, y, we get

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\,\|y\|} = \frac{\mathrm{Cov}[x,y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}}\,, \tag{6.64}$$

which is the correlation (Definition 6.6) between the two random variables. This means that we can think of correlation as 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 $\mathrm{Cov}[x,y] = 0$, 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

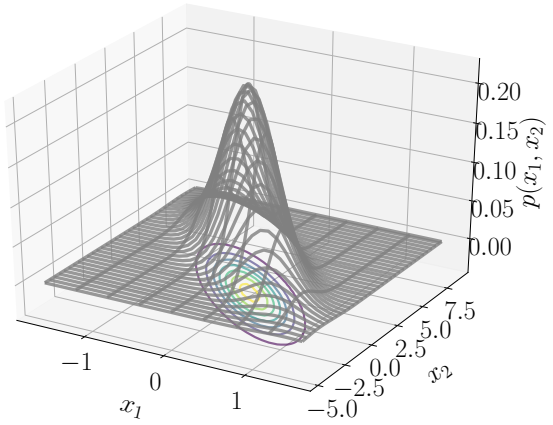


Figure 6.7 Gaussian distribution of two random variables x, y.

f-divergences. The study of divergences is beyond the scope of this book. Interested readers are referred to a recent book (Amari, 2016) written by one of the founders of the field of information geometry. ♢

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 important 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

$$p(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6.65}$$

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

The *multivariate Gaussian distribution* is fully characterized by a *mean vector* $\boldsymbol{\mu}$ and a *covariance matrix* $\boldsymbol{\Sigma}$ and defined as

$$p(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{6.66}$$

where $\boldsymbol{x} \in \mathbb{R}^D$ is a random variable. We write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Figure 6.7 shows a bi-variate Gaussian (mesh), with the

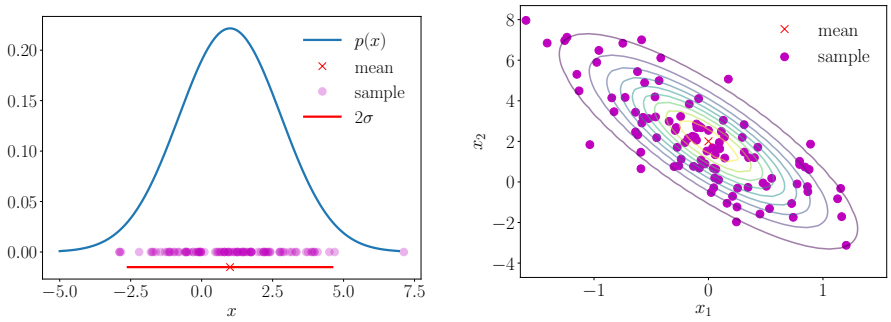


Figure 6.8 Gaussian distributions overlaid with 100 samples. Left: Univariate (1-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance. Right: Multivariate (2-dimensional) Gaussian, viewed from top. The red cross shows the mean and the coloured lines shows the contour lines of the density.

corresponding contour plot. The special case of the Gaussian with zero mean and identity covariance, that is $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, 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.

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 $\boldsymbol{x}$ and $\boldsymbol{y}$ 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 random variable $[\boldsymbol{x}, \boldsymbol{y}]^\top$,

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right). \tag{6.67}$$

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

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

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_{x \mid y}, \boldsymbol{\Sigma}_{x \mid y}) \tag{6.68}$$

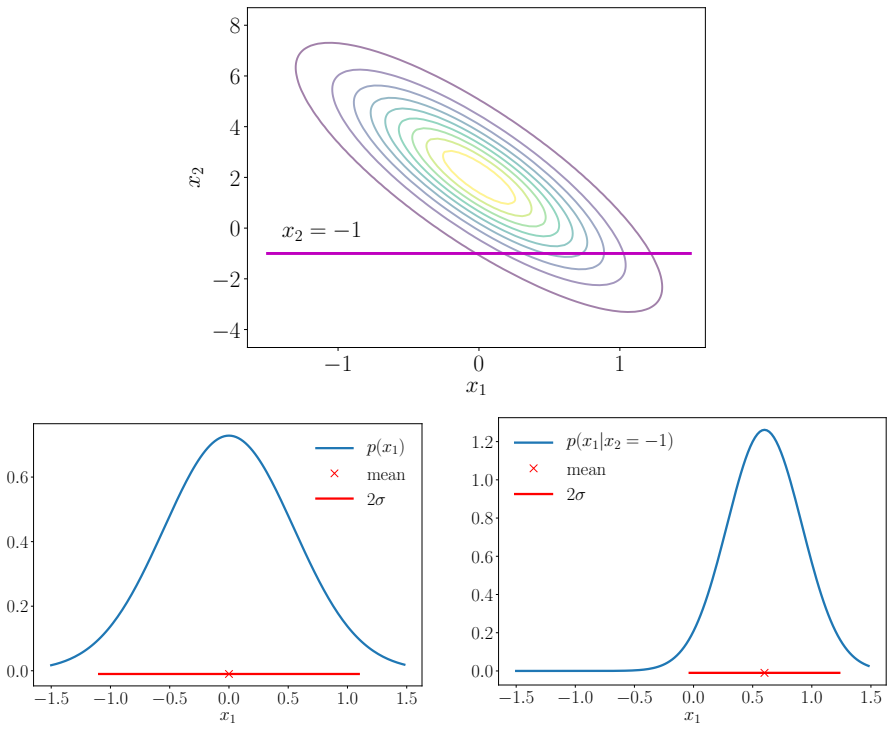


Figure 6.9 Top: Bivariate Gaussian; Bottom left: Marginal of a joint Gaussian distribution is Gaussian; Bottom right: The conditional distribution of a Gaussian is also Gaussian

$$\boldsymbol{\mu}_{x\,|\,y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{6.69}$$

$$\boldsymbol{\Sigma}_{x\,|\,y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\,. \tag{6.70}$$

Note that in the computation of the mean in (6.69) the $\boldsymbol{y}$-value is an observation and no longer random.

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

- The Kalman filter (Kalman, 1960), one of the most central algorithms for state estimation in signal processing, does nothing but computing Gaussian conditionals of joint distributions (Deisenroth and Ohlsson, 2011).
- Gaussian processes (Rasmussen and Williams, 2006), which are a practical implementation of a distribution over functions. In a Gaussian process, we make assumptions of joint Gaussianity of random variables. By (Gaussian) conditioning on observed data, we can determine a posterior distribution over functions.
- Latent linear Gaussian models (Roweis and Ghahramani, 1999; Murphy, 2012), which include probabilistic PCA (Tipping and Bishop, 1999).

♢

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

see (6.67), is itself Gaussian and computed by applying the sum-rule in (6.19) and given by

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x,\, \boldsymbol{\Sigma}_{xx}\big)\,. \tag{6.71}$$

The corresponding result holds for $p(\boldsymbol{y})$, which is obtained by marginalizing with respect to $\boldsymbol{x}$. Intuitively, looking at the joint distribution in (6.67), we ignore (i.e., integrate out) everything we are not interested in. This is illustrated on the bottom left of Figure 6.9.

Example 6.6
Consider the bivariate Gaussian distribution (illustrated in Figure 6.9)

$$p(x, y) = \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0.3 & -1\\-1 & 5\end{bmatrix}\right). \tag{6.72}$$

We can compute the parameters of the univariate Gaussian, conditioned on $y = -1$, by applying (6.69) and (6.70) to obtain the mean and variance respectively. Numerically, this is

$$\mu_{x\,|\,y=-1} = 0 + (-1)(0.2)(-1-2) = 0.6 \tag{6.73}$$

and

$$\sigma^2_{x\,|\,y=-1} = 0.3 - (-1)(0.2)(-1) = 0.1. \tag{6.74}$$

Therefore the conditional Gaussian is given by

$$p(x \,|\, y = -1) = \mathcal{N}\big(0.6,\, 0.1\big). \tag{6.75}$$

The marginal distribution $p(x)$ in contrast can be obtained by applying (6.71), which is essentially using the mean and variance of the random variable x, giving us

$$p(x) = \mathcal{N}\big(0,\, 0.3\big) \tag{6.76}$$

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 $\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\big)\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\big)$ is a Gaussian distribution scaled by a $c \in \mathbb{R}$, given by $c\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{c},\, \boldsymbol{C}\big)$ with

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \tag{6.77}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}) \tag{6.78}$$

$$c = (2\pi)^{-\frac{D}{2}} |\boldsymbol{A} + \boldsymbol{B}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^\top (\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\big)\,. \tag{6.79}$$

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

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

$$c=\mathcal{N}(\boldsymbol{a}\,|\,\boldsymbol{b},\,\boldsymbol{A}+\boldsymbol{B})=\mathcal{N}(\boldsymbol{b}\,|\,\boldsymbol{a},\,\boldsymbol{A}+\boldsymbol{B})\,. \tag{6.80}$$

Here, neither $\boldsymbol{a}$ nor $\boldsymbol{b}$ are random variables. However, writing c in this way is more compact than (6.79). ♢

6.5.3 Sums and Linear Transformations

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

$$p(\boldsymbol{x}+\boldsymbol{y})=\mathcal{N}(\boldsymbol{\mu}_x+\boldsymbol{\mu}_y,\,\boldsymbol{\Sigma}_x+\boldsymbol{\Sigma}_y)\,. \tag{6.81}$$

Knowing that $p(\boldsymbol{x}+\boldsymbol{y})$ 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

$$p(a\boldsymbol{x}+b\boldsymbol{y})=\mathcal{N}(a\boldsymbol{\mu}_x+b\boldsymbol{\mu}_y,\,a^2\boldsymbol{\Sigma}_x+b^2\boldsymbol{\Sigma}_y)\,. \tag{6.82}$$

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. ♢

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 $\boldsymbol{x}\boldsymbol{x}^\top$.

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

$$p(x)=\alpha p_1(x)+(1-\alpha)p_2(x) \tag{6.83}$$

where the scalar $0<\alpha<1$ is the mixture weight, and $p_1(x)$ and $p_2(x)$ are

univariate Gaussian densities (Equation (6.65)) 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,*

$$\mathbb{E}[x] = \alpha\mu_1 + (1-\alpha)\mu_2. \tag{6.84}$$

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

$$\mathbb{V}[x] = \left[\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2\right] + \left(\left[\alpha\mu_1^2 + (1-\alpha)\mu_2^2\right] - \left[\alpha\mu_1 + (1-\alpha)\mu_2\right]^2\right). \tag{6.85}$$

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 (6.83)) above

$$\begin{aligned}
\mathbb{E}[x] &= \int_{-\infty}^{\infty} xp(x)\mathrm{d}x && \text{(6.86a)}\\
&= \int_{-\infty}^{\infty} \alpha x p_1(x) + (1-\alpha)xp_2(x)\mathrm{d}x && \text{(6.86b)}\\
&= \alpha\int_{-\infty}^{\infty} xp_1(x)\mathrm{d}x + (1-\alpha)\int_{-\infty}^{\infty} xp_2(x)\mathrm{d}x && \text{(6.86c)}\\
&= \alpha\mu_1 + (1-\alpha)\mu_2. && \text{(6.86d)}
\end{aligned}$$

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).

$$\begin{aligned}
\mathbb{E}[x^2] &= \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x && \text{(6.87a)}\\
&= \int_{-\infty}^{\infty} \alpha x^2 p_1(x) + (1-\alpha)x^2 p_2(x)\mathrm{d}x && \text{(6.87b)}\\
&= \alpha\int_{-\infty}^{\infty} x^2 p_1(x)\mathrm{d}x + (1-\alpha)\int_{-\infty}^{\infty} x^2 p_2(x)\mathrm{d}x && \text{(6.87c)}\\
&= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2). && \text{(6.87d)}
\end{aligned}$$

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 (6.86d) from (6.87d),

$$\begin{aligned}
\mathbb{V}[x] &= \mathbb{E}[x^2] - (\mathbb{E}[x])^2 && \text{(6.88a)}\\
&= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2) - (\alpha\mu_1 + (1-\alpha)\mu_2)^2 && \text{(6.88b)}\\
&= \left[\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2\right]
\end{aligned}$$

$$+\left(\left[\alpha\mu_1^2 + (1-\alpha)\mu_2^2\right] - \left[\alpha\mu_1 + (1-\alpha)\mu_2\right]^2\right). \tag{6.88c}$$

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". □

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

We consider in Example 6.18 a bivariate standard Gaussian random variable $\boldsymbol{x}$ and performed a linear transformation $\boldsymbol{A}\boldsymbol{x}$ on it. The outcome is a Gaussian random variable with zero mean and covariance $\boldsymbol{A}\boldsymbol{A}^\top$. Observe that adding a constant vector will change the mean of the distribution, without affecting its variance, that is the random variable $\boldsymbol{x} + \boldsymbol{\mu}$ is Gaussian with mean $\boldsymbol{\mu}$ and identity covariance. Therefore, a linear (or affine) transformation of a Gaussian random variable is Gaussian distributed.

Consider a Gaussian distributed random variable $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a given matrix $\boldsymbol{A}$ of appropriate shape, let $\boldsymbol{y}$ be a random variable $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ which is a transformed version of $\boldsymbol{x}$. We can compute the mean of $\boldsymbol{y}$ by using the fact that the expectation is a linear operator (Equation (6.50)) as follows:

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu}. \tag{6.89}$$

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

$$\mathbb{V}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top. \tag{6.90}$$

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

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{\mu},\ \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top). \tag{6.91}$$

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 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ where $m \geqslant n$, let $\boldsymbol{y} \in \mathbb{R}^m$ be a Gaussian random variable with mean $\boldsymbol{A}\boldsymbol{x}$, i.e.,

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x},\ \boldsymbol{\Sigma}). \tag{6.92}$$

What is the corresponding probability distribution $p(\boldsymbol{x})$? If $\boldsymbol{A}$ is invertible, then we can write $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$ and apply the transformation in the previous paragraph. However, in general $\boldsymbol{A}$ is not invertible, and we use an approach similar to that of the pseudo-inverse (3.56). That is we premultiply both sides with $\boldsymbol{A}^\top$ and then invert $\boldsymbol{A}^\top\boldsymbol{A}$ which is symmetric and positive definite, giving us the relation

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \iff (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y} = \boldsymbol{x}\,. \tag{6.93}$$

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

$$p(\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{x} \,|\, (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{y},\, (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\big)\,. \tag{6.94}$$

6.5.4 Sampling from Multivariate Gaussian Distributions

We will not explain the subtleties of random sampling on a computer. 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 $\mathcal{N}\big(\mathbf{0},\, \boldsymbol{I}\big)$.

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 $\boldsymbol{x}_i, i = 1, \ldots, n$, from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We would like to construct the sample from a sampler that provides samples from the multivariate standard normal $\mathcal{N}\big(\mathbf{0},\, \boldsymbol{I}\big)$.

To obtain samples from a multivariate normal $\mathcal{N}\big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)$, we can use the properties of a linear transformation of a Gaussian random variable: If $\boldsymbol{x} \sim \mathcal{N}\big(\mathbf{0},\, \boldsymbol{I}\big)$ then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu}$, where $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$, is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Recall from Section 4.3 that when we can decompose $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$, while there are many possible decompositions, we often choose the Cholesky decomposition. This has the benefit that $\boldsymbol{A}$ is triangular, leading to efficient computation.

To compute the Cholesky factorization of a matrix, it is required that the matrix is symmetric and positive definite (Section 3.2.3). Covariance matrices possess this property.

6.6 Conjugacy and the Exponential Family

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

"Computers" were a job description.

1 There is some "closure property" when applying the rules of probability, e.g., Bayes' theorem. By closure we mean that applying a particular operation returns an object of the same type.

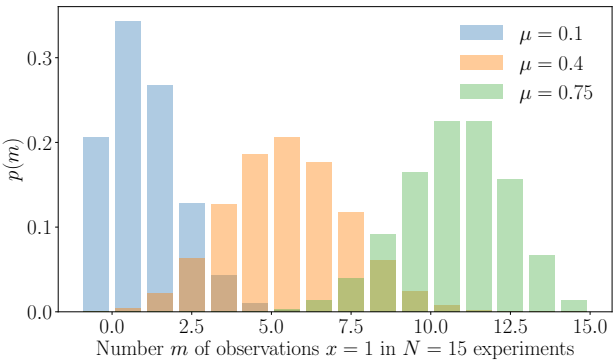


Figure 6.10 Examples of the Binomial distribution for $\mu \in \{0.1, 0.4, 0.75\}$ and $N = 15$.

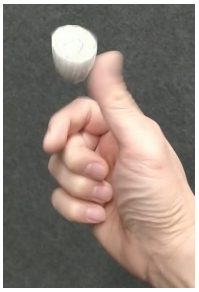
2 As we collect more data, we do not need more parameters to describe the distribution.

3 Since we are interested in learning from data, we want parameter estimation to behave nicely.

exponential family

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

Bernoulli distribution



Example 6.8

The *Bernoulli distribution* is a distribution for a single binary variable $x \in \{0, 1\}$ and is governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $x = 1$. The Bernoulli distribution is defined as

$$p(x \mid \mu) = \mu^x (1 - \mu)^{1-x}, \quad x \in \{0, 1\}, \tag{6.95}$$

$$\mathbb{E}[x] = \mu, \tag{6.96}$$

$$\mathbb{V}[x] = \mu(1 - \mu), \tag{6.97}$$

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 occurence of this is when expressing the Multinomial distribution. ♢

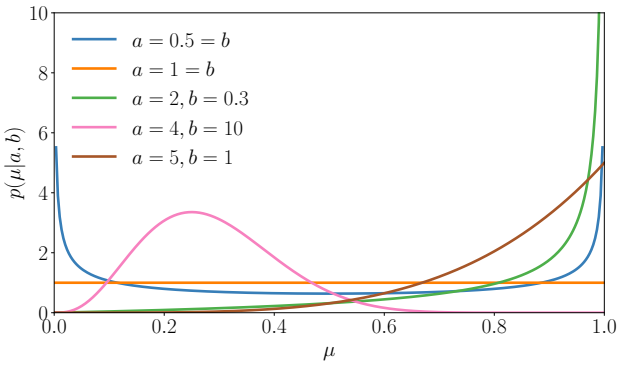


Figure 6.11 Examples of the Beta distribution for different values of α and β.

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 is defined as

Binomial distribution

$$p(m \mid N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}, \tag{6.98}$$

$$\mathbb{E}[m] = N\mu, \tag{6.99}$$

$$\mathbb{V}[m] = N\mu(1 - \mu) \tag{6.100}$$

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

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

Example 6.10

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 (illustrated in Figure 6.11) itself is governed by two parameters $\alpha > 0,\ \beta > 0$ and is defined as

Beta distribution

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1 - \mu)^{\beta-1} \tag{6.101}$$

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta}, \qquad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{6.102}$$

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

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx, \qquad t > 0\,. \tag{6.103}$$

$$\Gamma(t+1) = t\Gamma(t)\,. \tag{6.104}$$

Note that the fraction of Gamma functions in (6.101) normalizes the Beta distribution.

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

- For $\alpha = 1 = \beta$ we obtain the uniform distribution $\mathcal{U}[0, 1]$.
- 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). ♢

6.6.1 Conjugacy

According to Bayes' theorem (6.22), 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 some 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

conjugate **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.

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). ♢

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

$$p(x \mid N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x} \quad x = 0, 1, \ldots, N \tag{6.105}$$

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

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{6.106}$$

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

$$\begin{aligned}
p(\mu \mid x = h, N, \alpha, \beta) &\propto p(x \mid N, \mu) p(\mu \mid \alpha, \beta) && \text{(6.107a)}\\
&= \mu^h (1-\mu)^{(N-h)} \mu^{\alpha-1} (1-\mu)^{\beta-1} && \text{(6.107b)}\\
&= \mu^{h+\alpha-1} (1-\mu)^{(N-h)+\beta-1} && \text{(6.107c)}\\
&\propto \text{Beta}(h+\alpha, N-h+\beta) && \text{(6.107d)}
\end{aligned}$$

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

Multiplying the Beta and the Bernoulli distributions, we get

$$\begin{aligned}
p(\theta \mid x, \alpha, \beta) &= p(x \mid \theta) \times p(\theta \mid \alpha, \beta) && \text{(6.108a)}\\
&\propto \theta^x (1-\theta)^{1-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} && \text{(6.108b)}\\
&= \theta^{\alpha+x-1}(1-\theta)^{\beta+(1-x)-1} && \text{(6.108c)}\\
&\propto p(\theta \mid \alpha + x, \beta + (1-x)). && \text{(6.108d)}
\end{aligned}$$

Table 6.2 Examples of conjugate priors for common likelihood functions.

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

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

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

The Beta distribution is the conjugate prior for the parameter μ in both the Binomial and the Bernoulli likelihood. For a Gaussian likelihood function, we can place a conjugate Gaussian prior on the mean. The reason 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 $\boldsymbol{x} = [x_1, \ldots, x_N]^\top$ is a vector of univariate Gaussian random variables, that is $x_n \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean $\hat{\mu} = \frac{1}{N}(x_1 + \cdots + x_N)$ 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 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)**.** *Let x have probability density function $p(x \mid \theta)$. Then the statistics $\phi(x)$ are sufficient for θ if and only if $p(x \mid \theta)$ can be written in the form*

$$p(x \mid \theta) = h(x)g_\theta(\phi(x)). \tag{6.109}$$

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

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

ing 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 Darmois 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). ♢

exponential family

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

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right) , \tag{6.110}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is the vector of sufficient statistics. In general, any inner product (Section 3.2) can be used in (6.110), and for concreteness we will use the standard dot product here ($\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})$). 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 $h(\boldsymbol{x})$ can be absorbed into the dot product term by adding another entry ($\log h(\boldsymbol{x})$) to the vector of sufficient statistics $\boldsymbol{\phi}(\boldsymbol{x})$, and constraining the corresponding parameter $\theta_0 = 1$. The term $A(\boldsymbol{\theta})$ 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

log partition function

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})\right) . \tag{6.111}$$

natural parameters

For this form of parameterization, the parameters $\boldsymbol{\theta}$ 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 $\boldsymbol{\phi}(\boldsymbol{x})$.

Example 6.13 (Gaussian as Exponential Family)

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

$$p(x \mid \boldsymbol{\theta}) \propto \exp(\theta_1 x + \theta_2 x^2) . \tag{6.112}$$

Setting

$$\theta = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]^\top \tag{6.113}$$

and substituting into (6.112) we obtain

$$p(x \mid \boldsymbol{\theta}) \propto \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{6.114}$$

Therefore, the univariate Gaussian distribution is a member of the exponential family with sufficient statistic $\boldsymbol{\phi}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$.

Example 6.14 (Bernoulli as Exponential Family)
Recall the Bernoulli distribution from Example 6.8

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}, \quad x \in \{0, 1\}. \tag{6.115}$$

This can be written in exponential family form

$$\begin{aligned} p(x \mid \mu) &= \exp\left[\log\left(\mu^x (1-\mu)^{1-x}\right)\right] && (6.116)\\ &= \exp\left[x \log \mu + (1-x)\log(1-\mu)\right] && (6.117)\\ &= \exp\left[x \log \mu - x \log(1-\mu) + \log(1-\mu)\right] && (6.118)\\ &= \exp\left[x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right]. && (6.119) \end{aligned}$$

The last line (6.119) can be identified as being in exponential family form (6.110) by observing that

$$\begin{aligned} h(x) &= 1 && (6.120)\\ \theta &= \log \frac{\mu}{1-\mu} && (6.121)\\ \phi(x) &= x && (6.122)\\ A(\theta) &= -\log(1-\mu) = \log(1+\exp(\theta)). && (6.123) \end{aligned}$$

The relationship between θ and μ is invertible,

$$\mu = \frac{1}{1+\exp(-\theta)}. \tag{6.124}$$

This relation is used to obtain the right equality of (6.123).

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, 1995, Section 4.3.2), as well as as a nonlinear activation functions in neural networks (Goodfellow et al., 2016, Chapter 6). ♢ sigmoid

It is often not obvious how to find the parametric form of the conju-

gate distribution of a particular distribution. Exponential families provide a convenient way to find conjugate pairs of distributions. Consider the random variable $\boldsymbol{x}$ distributed as an exponential family (6.110)

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right) . \tag{6.125}$$

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

$$p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) = h_c(\boldsymbol{\theta}) \exp\left(\left\langle \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \gamma_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\theta} \\ -A(\boldsymbol{\theta}) \end{bmatrix} \right\rangle - A_c(\boldsymbol{\gamma})\right) , \tag{6.126}$$

where $\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \gamma_2 \end{bmatrix}$ 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 (6.119),

$$p(x \mid \mu) = \exp\left[x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right] . \tag{6.127}$$

The canonical conjugate prior therefore has the same form

$$p(\mu \mid \gamma, n_0) = \exp\left[n_0 \gamma \log \frac{\mu}{1-\mu} + n_0 \log(1-\mu) - A_c(\gamma, n_0)\right] , \tag{6.128}$$

which simplifies to

$$p(\mu \mid \gamma, n_0) = \exp\left[n_0 \gamma \log \mu + n_0(1-\gamma)\log(1-\mu) - A_c(\gamma, n_0)\right] . \tag{6.129}$$

Putting this in non-exponential family form

$$p(\mu \mid \gamma, n_0) \propto \mu^{n_0 \gamma}(1-\mu)^{n_0(1-\gamma)} \tag{6.130}$$

which is of the same form as the Beta distribution (6.101), 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, we will use small letters x, y to denote random variables and small capital letters X, Y to denote the values that the random variables take. We will explicitly write probability mass functions (pmf) of discrete random variables x as $P(x = \mathrm{X})$. For continuous random variables x, the probability density function (pdf) is written as $f(x)$ and the the cumulative distribution function (cdf) is written as $F_x(x \leqslant \mathrm{X})$. ♢

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.

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

As mentioned in Section 6.1, random variables and probability distributions are closely associated with each other. It is worth carefully teasing apart the two ideas, and in doing so we will motivate why we need to transform random variables.

Example 6.16
Consider a medical test that returns the number of cancerous cells that can be found in the biopsy. The state space is the set of non-negative integers. The random variable x is the *square* of the number of cancerous cells. Given that we know the probability distribution corresponding to the number of cancerous cells in a biopsy, how do we obtain the distribution of random variable x?

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

$$P(y = \mathrm{y}) = P(U(x) = \mathrm{y}) \qquad \text{transformation of interest} \tag{6.131a}$$
$$= P(x = U^{-1}(\mathrm{y})) \qquad \text{inverse} \tag{6.131b}$$

where we can observe that $\mathrm{x} = U^{-1}(\mathrm{y})$. 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) $F_x(\mathrm{x}) = P(x \leqslant \mathrm{x})$ and the fact that its differential is the probability density function (pdf) $f(x)$ (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:

$$F_y(\mathrm{y}) = P(y \leqslant \mathrm{y}) \tag{6.132}$$

2 Differentiating the cdf $F_y(\mathrm{y})$ to get the pdf $f(y)$.

$$f(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_y(\mathrm{y})\,. \tag{6.133}$$

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

Example 6.17
Let x be a continuous random variable with probability density function on $0 \leqslant x \leqslant 1$

$$f(x) = 3x^2\,. \tag{6.134}$$

We are interested in finding the pdf of $y = x^2$.

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

$$\begin{aligned}
F_y(\mathrm{Y}) &= P(y \leqslant \mathrm{Y}) && \text{definition of cdf} && \text{(6.135a)}\\
&= P(x^2 \leqslant \mathrm{Y}) && \text{transformation of interest} && \text{(6.135b)}\\
&= P(x \leqslant \mathrm{Y}^{\frac{1}{2}}) && \text{inverse} && \text{(6.135c)}\\
&= F_x(\mathrm{Y}^{\frac{1}{2}}) && \text{definition of cdf} && \text{(6.135d)}\\
&= \int_0^{\mathrm{Y}^{\frac{1}{2}}} 3t^2 \mathrm{d}t && \text{cdf as a definite integral} && \text{(6.135e)}\\
&= \left[t^3\right]_{t=0}^{t=\mathrm{Y}^{\frac{1}{2}}} && \text{result of integration} && \text{(6.135f)}\\
&= \mathrm{Y}^{\frac{3}{2}}\,, \quad 0 \leqslant \mathrm{Y} \leqslant 1\,. && && \text{(6.135g)}
\end{aligned}$$

Therefore, the cdf of y is

$$F_y(\mathrm{Y}) = \mathrm{Y}^{\frac{3}{2}} \tag{6.136}$$

for $0 \leqslant \mathrm{Y} \leqslant 1$. To obtain the pdf, we differentiate the cdf

$$f(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_y(\mathrm{Y}) = \frac{3}{2} y^{\frac{1}{2}} \tag{6.137}$$

for $0 \leqslant y \leqslant 1$.

In Example 6.17, 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 cumulative distribution function $F_x(\mathrm{X})$ of a random variable x, and using it as the transformation $U(x)$. This leads to the following theorem which is called the probability integral transform. The result is the basis of generating samples from distributions whose cdfs are known by first generating a sample from a uniform distribution and then transforming it by the inverse cdf.

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

Theorem 6.15. *This is Theorem 2.1.10 in Casella and Berger (2002). Let x be a continuous random variable with a strictly monotonic cumulative distribution function $F_x(\cdot)$. Then the random variable y defined as*

$$y = F_x(\mathrm{X})\,, \tag{6.138}$$

has a uniform distribution.

Proof We need to show that the cumulative distribution function of y defines a distribution of a uniform random variable. Recall that by the axioms of probability (Section 6.1) probabilities must be non-negative and

sum/integrate to one. Therefore, the range of possible values of $y = F_x(\mathrm{X})$ is the interval $[0, 1]$. For any $F_x(\cdot)$, the inverse $F_x^{-1}(\cdot)$ exists because we assumed that $F_x(\cdot)$ is strictly monotonically increasing, which we will use in the following.

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

$$\begin{aligned} F_y(\mathrm{Y}) &= P(y \leqslant \mathrm{Y}) && && \text{(6.139a)} \\ &= P(F_x(\mathrm{X}) \leqslant \mathrm{Y}) && \text{transformation of interest} && \text{(6.139b)} \\ &= P(x \leqslant F_x^{-1}(\mathrm{Y})) && \text{inverse exists} && \text{(6.139c)} \\ &= F_x(F_x^{-1}(\mathrm{Y})) && \text{definition of cdf} && \text{(6.139d)} \\ &= \mathrm{Y}\,, && && \text{(6.139e)} \end{aligned}$$

where the last line is due to the fact that $F_x(\cdot)$ composed with its inverse results in an identity transformation. The statement $F_y(\mathrm{Y}) = \mathrm{Y}$ along with the fact that y lies in the interval $[0, 1]$ means that $F_y(\cdot)$ is the cdf of the uniform random variable on the unit interval. □

probability integral transform

Theorem 6.15 is known as the *probability integral transform*, and it is used to derive algorithms for sampling from distributions by transforming the result of sampling from a uniform random variable (Bishop, 2006). It is also used for hypothesis testing whether a sample comes from a particular distribution (Lehmann and Romano, 2005). The idea that the output of a cdf gives a uniform distribution also forms the basis of copulas (Nelsen, 2006).

6.7.2 Change of Variables

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

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

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,

$$\int f(g(x))g'(x)\mathrm{d}x = \int f(u)\mathrm{d}u \quad \text{where} \quad u = g(x). \tag{6.140}$$

The derivation of this rule is based on the chain rule of calculus (5.31) 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 subsituting $u = g(x)$, the argument inside the integral on the right hand side of (6.140) becomes $f(g(x))$. By pretending that the term $\mathrm{d}u$ can be approximated by $\mathrm{d}u \approx \Delta u = g'(x)\Delta x$, and that $\mathrm{d}x \approx \Delta x$, we obtain (6.140). ♢

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

$$F_y(\mathrm{Y}) = P(y \leqslant \mathrm{Y})\,. \tag{6.141}$$

We are interested in a function U of the random variable

$$P(y \leqslant \mathrm{Y}) = P(U(x) \leqslant \mathrm{Y})\,, \tag{6.142}$$

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

$$P(U(x) \leqslant \mathrm{Y}) = P(U^{-1}(U(x)) \leqslant U^{-1}(\mathrm{Y})) = P(x \leqslant U^{-1}(\mathrm{Y}))\,, \tag{6.143}$$

which is an expression of the cdf of x. Recall the definition of the cdf in terms of the pdf

$$P(x \leqslant U^{-1}(\mathrm{Y})) = \int_a^{U^{-1}(\mathrm{Y})} f(x)\mathrm{d}x\,. \tag{6.144}$$

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

$$F_y(\mathrm{Y}) = \int_a^{U^{-1}(\mathrm{Y})} f(x)\mathrm{d}x\,. \tag{6.145}$$

To obtain the pdf, we differentiate (6.145) with respect to y.

$$f(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_y(\mathrm{Y}) = \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{U^{-1}(\mathrm{Y})} f(x)\mathrm{d}x \tag{6.146}$$

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 (6.140) to get the substitution

$$\int f(U^{-1}(y))U^{-1\prime}(y)\mathrm{d}y = \int f(x)\mathrm{d}x \quad \text{where} \quad x = U^{-1}(y). \tag{6.147}$$

Using (6.147) on the right hand side of (6.146) gives us

$$f(y) = \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{U^{-1}(\mathrm{Y})} f_x(U^{-1}(y))U^{-1\prime}(y)\mathrm{d}y. \tag{6.148}$$

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

$$f(y) = f_x(U^{-1}(y)) \times \left| \frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y) \right| . \tag{6.149}$$

change of variable

This is called the *change of variable* technique. The term $\left| \frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y) \right|$ measures how much a unit volume changes when applying U (see also the Remark on page 151. In (6.149) 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 (6.131b), we have an additional factor $\left| \frac{\mathrm{d}}{\mathrm{d}y} U^{-1}(y) \right|$. The continuous case requires more care because $P(y = \mathrm{Y}) = 0$ for all Y. The probability density function $f(y)$ does not have a description as a probability of an event involving y. ♢

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 (5.56) 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 $f(\boldsymbol{x})$ be the value of the probability density of the multivariate continuous random variable $\boldsymbol{x}$. If the vector-valued function $\boldsymbol{y} = U(\boldsymbol{x})$ is differentiable and invertible for all values within the domain of $\boldsymbol{x}$, then for corresponding values of $\boldsymbol{y}$, the probability density of $\boldsymbol{y} = U(\boldsymbol{x})$ is given by*

$$f(\boldsymbol{y}) = f_{\boldsymbol{x}}(U^{-1}(\boldsymbol{y})) \times \left| \det \left(\frac{\partial}{\partial \boldsymbol{y}} U^{-1}(\boldsymbol{y}) \right) \right| . \tag{6.150}$$

The theorem looks intimidating at first glance, but the key point is that a change of variable of a multivariate random variable follows the procedure of the univariate change of variable. First we need to work out the inverse transform, and substitute that into the density of $\boldsymbol{x}$. Then we calculate the determinant of the Jacobian and multiply the result. The following example illustrates the case of a bivariate random variable.

Example 6.18

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

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right). \tag{6.151}$$

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 $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ defined as

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{6.152}$$

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

Recall that for change of variables we require the inverse transformation of $\boldsymbol{x}$ as a function of $\boldsymbol{y}$. 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

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{6.153}$$

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

$$f(\boldsymbol{x}) = f(\boldsymbol{A}^{-1}\boldsymbol{y}) = \frac{1}{2\pi} \exp\left(-\tfrac{1}{2}\boldsymbol{y}^\top \boldsymbol{A}^{-\top}\boldsymbol{A}^{-1}\boldsymbol{y}\right). \tag{6.154}$$

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

$$\frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{A}^{-1}\boldsymbol{y} = \boldsymbol{A}^{-1}. \tag{6.155}$$

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

$$\det\left(\frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{A}^{-1}\boldsymbol{y}\right) = \frac{1}{ad - bc}. \tag{6.156}$$

We are now able to apply the change-of-variable formula from Theorem 6.16 by multiplying (6.154) with (6.156), which yields

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) \left|\det\left(\frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{A}^{-1}\boldsymbol{y}\right)\right| \tag{6.157a}$$

$$= \frac{1}{2\pi} \exp\left(-\tfrac{1}{2}\boldsymbol{y}^\top \boldsymbol{A}^{-\top}\boldsymbol{A}^{-1}\boldsymbol{y}\right) |ad - bc|^{-1}. \tag{6.157b}$$

While Example 6.18 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 $f(\boldsymbol{x})$ above is actually the standard Gaussian distribution, and the transformed density $f(\boldsymbol{y})$ is a bivariate Gaussian with covariance $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$. ♢

6.8 Further Reading

This chapter is rather terse at times, Grinstead and Snell (1997) and Walpole et al. (2011) provides 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 probabilisitic 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 $p(y \mid x)$ for continuous random variables x, y (Proschan and Presnell, 1998). The lazy notation hides the fact that we want to specify that $x = \mathrm{x}$ (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 $\mathbb{E}_y[f(y) \mid \sigma(x)]$, 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, 2003b; Grimmett and Welsh, 2014) including some very technical discussions (Çinlar, 2011; Dudley, 2002; Shiryayev, 1984; Lehmann and Casella, 1998; Bickel and Doksum, 2006). 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 books with a probabilistic modelling focus includes MacKay (2003b); 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)

$$\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 1.5\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0 & 2.9\\-1.1 & 0.7\end{bmatrix}\right).$$

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 $p(\mu|x_1, \ldots, x_N)$.

6.3 Consider the following time-series model:

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

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

1 What is the form of $p(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$? Justify your answer (you do not have to explicitly compute the joint distribution). (1–2 sentences)
2 Assume that $p(\boldsymbol{x}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$.
   a) Compute $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$
   b) Compute $p(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$
   c) At time $t+1$, we observe the value $\boldsymbol{y}_{t+1} = \hat{\boldsymbol{y}}$. Compute $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t+1})$.

6.4 Prove the relationship in Equation 6.44, which relates the standard definition of the variance to the raw score expression for the variance.

6.5 Prove the relationship in Equation 6.45, which relates the pairwise difference between examples in a dataset with the raw score expression for the variance.

6.6 Express the Bernoulli distribution in the natural parameter form of the exponential family (Equation (6.110)).

6.7 Express the Binomial distribution as an exponential family distribution. Also express the Beta distribution is an exponential family distribution. Show that the product of the Beta and the Binomial distribution is also a member of the exponential family.

6.8 Derive the relationship in Section 6.5.2 in two ways:

1 By completing the square
2 By expressing the Gaussian in its exponential family form

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

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \tag{6.158}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}) \tag{6.159}$$

$$c = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A}+\boldsymbol{B}|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(\boldsymbol{a}-\boldsymbol{b})^\top(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{a}-\boldsymbol{b})\right). \tag{6.160}$$

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

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

$$\mathbb{E}_x[x] = \mathbb{E}_y\big[\mathbb{E}_x[x|y]\big]$$

Here, $\mathbb{E}_x[x|y]$ denotes the expected value of x under the conditional distribution $p(x|y)$.

6.10 **Manipulation of Gaussian Random Variables.**
Consider a Gaussian random variable $\boldsymbol{x} \sim \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x,\, \boldsymbol{\Sigma}_x\big)$, where $\boldsymbol{x} \in \mathbb{R}^D$. Furthermore, we have

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w}\,, \tag{6.161}$$

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

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

$$\boldsymbol{z} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{v}\,, \tag{6.162}$$

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

- Write down $p(\boldsymbol{z}|\boldsymbol{y})$.
- Compute $p(\boldsymbol{z})$, i.e., the mean $\boldsymbol{\mu}_z$ and the covariance $\boldsymbol{\Sigma}_z$. Derive your result in detail.

4 Now, a value $\hat{\boldsymbol{y}}$ is measured. Compute the posterior distribution $p(\boldsymbol{x}|\hat{\boldsymbol{y}})$.[2]
Hint for solution: Start by explicitly computing the joint Gaussian $p(\boldsymbol{x}, \boldsymbol{y})$. This also requires to compute the cross-covariances $\mathrm{Cov}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}, \boldsymbol{y}]$ and $\mathrm{Cov}_{\boldsymbol{y},\boldsymbol{x}}[\boldsymbol{y}, \boldsymbol{x}]$. Then, apply the rules for Gaussian conditioning.

6.11 **Probability integral transformation**
Given a continous random variable x, with cdf $F_x(x)$. Show that the random variable $y = F_x(x)$ is uniformly distributed.

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

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