6

Probability and Distributions

Probability, loosely speaking, is 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 the introduction (Chapter 1), we would often like to quantify uncertainty: 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 real numbers. Associated with the random variable is a number corresponding to each possible mapping of outcomes to real numbers. This set of numbers specifies the probability of occurrence, and is called the *probability distribution*.

random variable

probability distribution

Probability distributions are used as a building block for other concepts, such as model selection (Section 8.4) and graphical models (Section 8.5). In this section, we present the three concepts that define a probability space: the state space, the events and the probability of an event. The presentation is deliberately slightly hand wavy since a rigorous presentation would occlude the main idea.

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 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

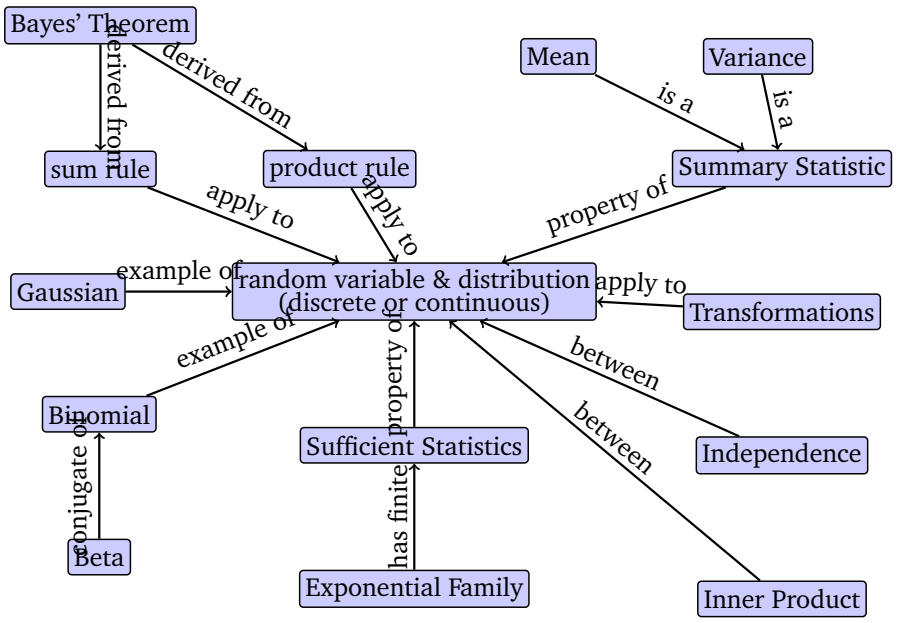

Figure 6.1 A mind map of the concepts related to random variables and probability distributions, as described in this chapter.

observe that B is true. It seems A becomes more plausible. We use this form of reasoning daily: Our friend is late. We have three hypotheses H1, H2, H3. Was she H1 abducted by aliens, H2 abducted by kidnappers or H3 delayed by traffic. How do we conclude H3 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 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.
    1. Consistency or non-contradiction: when the same result can be reached through different means, the same plausibility value must be found in all cases.
    2. Honesty: All available data must be taken into account.
    3. 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 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 The frequentist interpretation considers probability to be the relative frequencies of events, in the limit when one has infinite data. ♢

It is worth noting that some machine learning literature 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 state space Ω

state space

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".

The events $\mathcal{A}$

events

The *events* can be observed after the experiment is done, i.e., they are realizations of an experiment. The event space is often denoted by $\mathcal{A}$ and is also often the set of all subsets of Ω. In the two coins example, one possible element of $\mathcal{A}$ is the event when both tosses are the same, that is $\{$hh, tt$\}$.

The probability $P(A)$

With each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the probability or 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 must sum to 1, i.e.,

$\sum_{A\in\mathcal{A}} P(A) = 1$. We associate this number (the probability) to a particular event occurring, and intuitively understand this as the chance that this event occurs. This association or mapping is called a *random variable*. This brings us back to the concepts at the beginning of this chapter, where we can see that a random variable is a map from Ω to $\mathbb{R}$. The name "random variable" is a great source of misunderstanding as it is neither random nor is it a variable. It is a function.

random variable

We omit the definition of a random variable as this will become too technical for the purpose of this book.

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. ♢

We say that a random variable is distributed according to a particular probability distribution, which defines the probability mapping between the event and the probability of the event. 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. An outline of the concepts presented in this chapter are shown in Figure 6.1.

6.1.3 Statistics

Probability theory and statistics are often presented together, and in some sense they are intertwined. 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. 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.

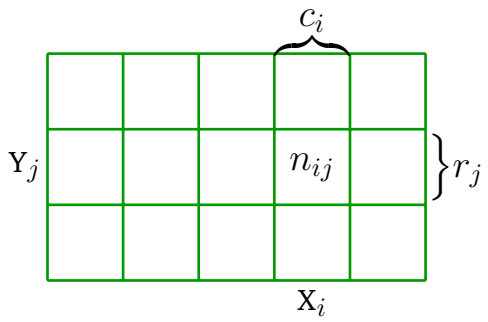

Figure 6.2 Visualization of a discrete bivariate probability mass function, with random variables x and y. This diagram is from Bishop (2006).

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 state space Ω is discrete, we can specify the probability that a random variable x takes a particular value $\mathrm{x} \in \Omega$, 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*. We will discuss discrete random variables in the following subsection. When the state space Ω is continuous, for example 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

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. We define the *joint probability* as the entry of both values jointly.

joint probability

$$P(x = \mathrm{x}_i, y = \mathrm{y}_i) = \frac{n_{ij}}{N} \,. \tag{6.1}$$

To be precise, the above table defines the *probability mass function* (pmf) of a discrete probability distribution. For two random variables x and y, the probability that $x = \mathrm{x}$ and $y = \mathrm{y}$ 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 6.1
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 probabilities 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

$$P(x = \mathrm{X}_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N} \tag{6.2}$$

and

$$P(y = \mathrm{Y}_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N}, \tag{6.3}$$

where c_i and r_j are the ith column and jth row of the probability table, respectively. Recall that by the axioms of probability (Section 6.1) we require that the probabilities sum up to one, that is

$$\sum_{i=1}^{3} P(x = \mathrm{X}_i) = 1 \quad \text{and} \quad \sum_{j=1}^{5} P(y = \mathrm{Y}_j) = 1\,. \tag{6.4}$$

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

$$p(y = \mathrm{Y}_j \,|\, x = \mathrm{X}_i) = \frac{n_{ij}}{c_i}\,, \tag{6.5}$$

and the conditional probability of x given y is

$$p(x = \mathrm{X}_i \,|\, y = \mathrm{Y}_j) = \frac{n_{ij}}{r_j}\,, \tag{6.6}$$

The marginal probability that x takes the value X irrespective of the value of random variable y is (lazily) written as $p(x)$. If we consider only the instances where $x = \mathrm{X}$, then the fraction of instances (the conditional probability) for which $y = \mathrm{Y}$ is written (lazily) as $p(y \,|\, x)$.

Example 6.2
Consider a statistical experiment where we perform a medical test for cancer two times. There are two possible outcomes for each test, and hence there are four outcomes in total. The state space or sample space Ω of this experiment is then (cancer, cancer), (cancer, healthy), (healthy, cancer), (healthy, healthy). The event we are interested in is the total

This toy example is essentially a coin flip example.

number of times the repeated medical test returns a cancerous answer, where we can see from the above state space can occur in no test, either one of the tests or both tests. Therefore the event space $\mathcal{A}$ is 0, 1, 2. Let variable x denote the number of times the medical test returns "cancer". Then x is a random variable (a function) that counts the number of times "cancer" appears. It can be represented as a table as below

$$x((\text{cancer}, \text{cancer})) = 2 \tag{6.7}$$
$$x((\text{cancer}, \text{healthy})) = 1 \tag{6.8}$$
$$x((\text{healthy}, \text{cancer})) = 1 \tag{6.9}$$
$$x((\text{healthy}, \text{healthy})) = 0. \tag{6.10}$$

Let us assume that this useless test returns at random a value of "cancer" with probability 0.3, ignoring any real world information. This assumption also implies that the two tests are independent of each other, which we will discuss in Section 6.4.3. Note that since there are two states which map to the same event, where only one of the tests say "cancer". Therefore the probability mass function of x is given by the table below

$$P(x = 2) = 0.09 \tag{6.11}$$
$$P(x = 1) = 0.42 \tag{6.12}$$
$$P(x = 0) = 0.49. \tag{6.13}$$

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 gender of a person 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 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

When we consider real valued random variables, that is when we consider state spaces which are intervals of the real line $\mathbb{R}$ we have corresponding definitions to the discrete case (Section 6.2.1). We will sweep measure theoretic considerations under the carpet in this book, and pretend as if we can perform operations 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

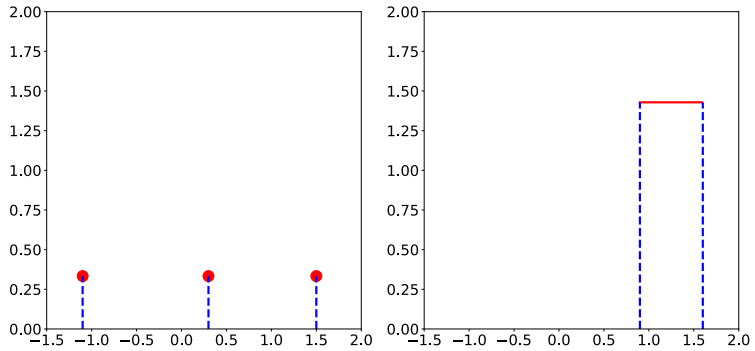

Figure 6.3 Examples of Uniform distributions. (left) discrete, (right) continuous. See example for details of the distributions.

when we want to discuss continuous distributions such as the Gaussian (Section 6.6). For our purposes, the lack of precision allows a more brief introduction to probability. A reader interested a measure based approach is referred to Billingsley (1995).

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

$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) d\boldsymbol{x} = 1\,. \tag{6.14}$$

Here, $\boldsymbol{x} \in \mathbb{R}^D$ is a (continuous) random variable. For discrete random variables, the integral in (6.14) is replaced with a sum.

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

cumulative distribution function

$$F_{\boldsymbol{x}}(\mathrm{x}) = P(x_1 \leqslant \mathrm{x}_1, \ldots, x_D \leqslant \mathrm{x}_D) \tag{6.15}$$

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

$$F_{\boldsymbol{x}}(\mathrm{x}) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{6.16}$$

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 (Equation (6.4)), and there are a finite number of states, the probability of each state must lie in the interval $[0, 1]$. However the analogous requirement for continuous random variables (Equation (6.14)) does not imply that the value of the density is less than 1 for all values. We illustrate this 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 used numbers to drive home the point that we do not want to use (and should ignore) the ordering of the states. The probability mass function can be represented as a table of probability values.

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

Alternatively one could think of this as a graph (left of Figure 6.3), 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 the left of Figure 6.3 is deliberately extended such that is it the same as the right figure.

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

$$\int_{0.9}^{1.6} p(x)\mathrm{d}x = 1\,. \tag{6.17}$$

Very often the literature uses lazy notation and nomenclature that can be confusing to a beginner. For a value x of a state space Ω, $p(x)$ denotes the probability that random variable x takes value x, i.e., $P(x = \mathrm{x})$, which is known as the probability mass function. 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 make things even more confusing the cumulative distribution function $P(x \leqslant \mathrm{x})$ is often also referred to as the "distribution". In this chapter we often will use the

	"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

Table 6.1 Nomenclature for probability distributions.

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 distributions but also for continuous probability density functions, although this is technically incorrect. However, this is consistent with the majority of machine learning literature. ♢

6.3 Sum Rule, Product Rule and Bayes' Theorem

When we think of a probabilisitic 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 modelling 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.

Before we define the sum rule and product rule, let us briefly explore how to use probabilistic models to capture uncertainty (Ghahramani, 2015). At the lowest modelling level, measurement noise introduces model uncertainty. for example the measurement error in a camera sensor. We will see in Chapter 9 how to use Gaussian (Section 6.6) noise models for linear regression. At higher modelling levels, we would be interested to model 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 discuss model selection issues in Chapter 8. Once we have the probabilistic models, the basic rules of probability presented in this section are used to infer the unobserved quantities given the observed data. The same rules of probability are used for inference (transforming prior probabilities to posterior probabilities) and learning (estimating the likelihood of the model for a given dataset).

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 that $p(x, y)$ is the joint distribution of the two

random variables x, y, $p(x), p(y)$ are the corresponding marginal distributions, and $p(y \mid x)$ is the conditional distribution of y given x.

sum rule

The first rule, the *sum rule* is expressed for discrete random variables as

$$p(x) = \sum_y p(x, y) \qquad \text{sum rule/marginalization property} . \quad (6.18)$$

marginalization property

The sum above is over the set of states of the random variable Y. The sum rule is also known as the *marginalization property*. For continuous probability distributions, the sum is replaced by an integral

$$p(x) = \int_y p(x, y) dy . \quad (6.19)$$

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.

Remark. Many of the computational challenges of probabilistic modelling 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 are generally computationally hard, in the sense that there is no known polynomial time algorithm to calculate them exactly. ♢

product rule

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

$$p(x, y) = p(y \mid x)p(x) \qquad \text{product rule.} \quad (6.20)$$

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 $p(x)$, and the conditional distribution of the second random variable given the first $p(y \mid x)$. Observe that since the ordering of random variables is arbitrary in $p(x, y)$ the product rule also implies $p(x, y) = p(x \mid y)p(y)$. To be precise, Equation (6.20) 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 (recall the discussion in Section 6.2.3).

In machine learning and Bayesian statistics, we are often interested in making inferences of random variables given that we have observed other random variables. Let us assume, we have some prior knowledge $p(x)$ about a random variable x and some relationship $p(y \mid x)$ between x and a second random variable y. If we now observe y, we can use Bayes' theorem to draw some conclusions about x given the observed values of y. *Bayes'*

Bayes' theorem is also called the "probabilistic inverse"

Bayes' theorem

theorem or *Bayes' law*

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(y)} \tag{6.21}$$

is a direct consequence of the sum and product rules in (6.18)–(6.20).

Example 6.4 (Applying the Sum and Product Rule)
We prove Bayes' theorem by using the sum and product rule. First we observe that we can apply the product rule in two ways,

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y})\,. \tag{6.22}$$

Simple algebra then gives us (6.21). Very often in machine learning, the evidence term $p(\boldsymbol{x})$ is hard to estimate, and we rewrite it by using the sum and product rule.

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

We now have an alternative formulation

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

In Equation (6.21), $p(y)$ is the *prior*, which encapsulates our prior knowledge of y, $p(x \mid y)$ is the *likelihood* that describes how x and y are related. The quantity $p(x)$ is the marginal likelihood or *evidence* and is a normalizing constant (independent of y). The *posterior* $p(\boldsymbol{x} \mid \boldsymbol{y})$ expresses exactly what we are interested in, i.e., what we know about x if we observe y. We will see an application of this in Maximum-A-Posteriori estimation (Section 9.2.3).

prior
likelihood
The likelihood is sometimes also called the "measurement model".
evidence
posterior

6.4 Summary Statistics and Independence

We are often interested in summarizing and contrasting random variables. A statistic of a random variable is a deterministic function of that random variable. The summary statistics of a distribution provide one useful view how a random variable behaves, and as the name suggests, provides numbers that summarize the distribution. The following describes 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 6.7 that there is a useful family of distributions (called the exponential family) where the statistics of the random variable capture all the possible information. The definitions in this section are stated for a general multivariate continuous random variable, because it is more intuitive to think about means and covariances in terms of real numbers. Analogous definitions exist for discrete random variables where the integral is replaced by a sum.

In one dimension, the mean value is the average value. It is the value obtained by summing up all values and dividing by the number of items. In more than one dimension, the sum becomes vector addition and the idea still holds. To account for the fact that we are dealing with a continuous random variable $\boldsymbol{x} \in \mathbb{R}^D$ with a particular density $p(\boldsymbol{x})$, the sum becomes an integral, and the addition is weighted by the density.

mean

Definition 6.3 (Mean)**.** The *mean* of a random variable $\boldsymbol{x} \in \mathbb{R}^D$ is defined as

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} = \begin{bmatrix} \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_D] \end{bmatrix} \in \mathbb{R}^D \,, \tag{6.25}$$

where the subscript indicates the corresponding dimension of $\boldsymbol{x}$.

median
mode

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, that is intuitively it is a typical value. For distributions which are asymmetric or has long tails, the median provides an estimate of a typical value that is closer to human intuition than the mean value. The mode is the most frequently occurring value, which is the highest 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.

The generalization of the median to higher dimensions is non-trivial, as there is no obvious way to "sort" in more than one dimension.

The definition of the mean (Definition 6.3), is actually a special case of an incredibly useful concept: the expected value.

expected value

Definition 6.4 (Expected value)**.** The *expected value* of a function g of a random variable $\boldsymbol{x} \sim p(\boldsymbol{x})$ is given by

$$\mathbb{E}_{\boldsymbol{x}}[g(\boldsymbol{x})] = \int g(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \,. \tag{6.26}$$

The mean is recovered if we set the function g in Definition 6.4 to the identity function. This indicates that we can think about functions of random variables, which we will revisit in Section 6.5.

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

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

$$\begin{aligned}\mathbb{E}_x[f(x)] &= \int f(x)p(x)dx && (6.27)\\ &= \int [ag(x) + bh(x)]p(x)dx && (6.28)\\ &= a\int g(x)p(x)dx + b\int h(x)p(x)dx && (6.29)\\ &= a\mathbb{E}_x[g(x)] + b\mathbb{E}_x[h(x)] && (6.30)\end{aligned}$$

This linear relationship holds in higher dimensions as well. ♢

For two random variables, we may wish to figure out their correspondence to each other.

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

$$\mathrm{Cov}[x, y] = \mathbb{E}\big[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\big]\,. \tag{6.31}$$

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

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]\,. \tag{6.32}$$

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 denoted $\sigma(x)$. variance standard deviation

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

Definition 6.6 (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

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}]^\top = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]^\top \in \mathbb{R}^{D\times E}\,. \tag{6.33}$$

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

Covariance intuitively represents the notion of how dependent random variables are to one another. We will revisit the idea of covariance again in Section 6.4.3

Definition 6.6 can be applied with the same multivariate random variable in both arguments, which results in a useful concept that intuitively captures the "spread" of a random variable.

Definition 6.7 (Variance)**.** The *variance* of a random variable $\boldsymbol{x} \in \mathbb{R}^D$ variance

with mean vector $\boldsymbol{\mu}$ is defined as

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]^\top \tag{6.34}$$

$$= \begin{bmatrix} \mathrm{Cov}[x_1,x_1] & \mathrm{Cov}[x_1,x_2] & \dots & \mathrm{Cov}[x_1,x_D] \\ \mathrm{Cov}[x_2,x_1] & \mathrm{Cov}[x_2,x_2] & \dots & \mathrm{Cov}[x_2,x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D,x_1] & \dots & \dots & \mathrm{Cov}[x_D,x_D] \end{bmatrix} \in \mathbb{R}^{D\times D}\,. \tag{6.35}$$

covariance matrix

This matrix is called the *covariance matrix* of the random variable $\boldsymbol{x}$. The covariance matrix is symmetric and positive definite and tells us something about the spread of the data.

The covariance matrix contains the variances of the marginals $p(x_i) = \int p(x_1,\dots,x_D)\mathrm{d}x_{\setminus i}$ on its diagonal, where "$\setminus i$" denotes "all variables but i". The off-diagonal terms contain the *cross-covariance* terms $\mathrm{Cov}[x_i,x_j]$ for $i,j=1,\dots,D,\ i\neq j$.

cross-covariance

It generally holds that

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathrm{Cov}_{\boldsymbol{x}}[\boldsymbol{x},\boldsymbol{x}]\,. \tag{6.36}$$

population mean and covariance

The definitions above are often also called the *population mean and covariance*. For a particular set of data we can obtain an estimate of the mean, which is called the *empirical mean* or *sample mean*. The same holds for the empirical covariance.

empirical mean
sample mean

Definition 6.8 (Empirical Mean and Covariance)**.** The *empirical mean* vector is the arithmetic average of the observations for each variable, and is written

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\,. \tag{6.37}$$

empirical covariance

The *empirical covariance* is a $K\times K$ matrix

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\bar{\boldsymbol{x}})(\boldsymbol{x}_n-\bar{\boldsymbol{x}})^\top. \tag{6.38}$$

Empirical covariance matrices are positive semi-definite (see Section 3.2.3).

We use the sample covariance in this book. The unbiased (sometimes called corrected) covariance has the factor $N-1$ in the denominator.

6.4.2 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 one needs 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. That is

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{6.39}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the mean. Observe that the variance as expressed above is the mean of a new random variable $z = (x-\mu)^2$.

When estimating this empirically, we need to resort to a two pass algorithm: one pass through the data to calculate the mean μ using (6.37), 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 (6.39) can be converted to the so called raw score formula for variance

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 . \tag{6.40}$$

This expression in (6.40) can be remembered as "the mean of the square minus the square of the mean". It can be calculated 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 is numerically unstable. The raw score version of the variance can be useful in machine learning, for example when deriving the bias-variance decomposition (Bishop, 2006).

The two terms can cancel out, resulting is 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. By expanding the square we can show that the sum of pairwise differences is two times the raw score expression,

$$\frac{1}{N^2}\sum_{i,j=1}^{N}(x_i-x_j)^2 = 2\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right] \tag{6.41}$$

Observe that (6.41) is twice of (6.40). 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.

6.4.3 Statistical Independence

Definition 6.9 (Independence)**.** Two random variables $\boldsymbol{x}, \boldsymbol{y}$ are *statistically independent* if and only if

statistically independent

$$p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})\,. \tag{6.42}$$

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

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

Note that the last point above may not hold in converse, that is two random variables can have covariance zero but are not statistically independent.

correlation

Remark. Let us briefly mention the relationship between *correlation* and *covariance*. 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. ♢

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

conditionally independent given $\boldsymbol{z}$

Definition 6.10 (Conditional Independence)**.** Formally, $\boldsymbol{x}$ and $\boldsymbol{y}$ are *conditionally independent given* $\boldsymbol{z}$ if and only if

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{y} \mid \boldsymbol{z}) \,. \tag{6.43}$$

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

Note that the definition of conditional independence above requires that the relation in Equation (6.43) must hold true for every value of $\boldsymbol{z}$. The interpretation of Equation (6.43) above 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 (Equation (6.20)), we can expand the left hand side of Equation 6.43 to obtain

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) p(\boldsymbol{y} \mid \boldsymbol{z}). \tag{6.44}$$

By comparing the right hand side of Equation (6.43) with Equation (6.44), we see that $p(\boldsymbol{y} \mid \boldsymbol{z})$ appears in both, and therefore

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}). \tag{6.45}$$

Equation (6.45) above provides an alternative definition of conditional independence, that is $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{z}$. This alternative presentation provides the interpretation: "given that we know $\boldsymbol{z}$, knowledge about $\boldsymbol{y}$ does not change our knowledge of $\boldsymbol{x}$".

6.4.4 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 Section 6.6 and 6.7), 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.

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}]^\top && (6.52)\\ &= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^\top + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{A}^\top && (6.53)\\ &= \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{A}^\top && (6.54)\\ &\overset{(6.34)}{=} \boldsymbol{\Sigma}\boldsymbol{A}^\top\,, && (6.55)\end{aligned}$$

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

6.4.5 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

$$\mathbb{V}[x+y] = \mathbb{V}[x] + \mathbb{V}[y] \tag{6.56}$$

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

In the following, we see whether we can find a geometric interpretation of the variance relation of uncorrelated random variables in (6.56).

Random variables can be considered vectors in a vector space, and we

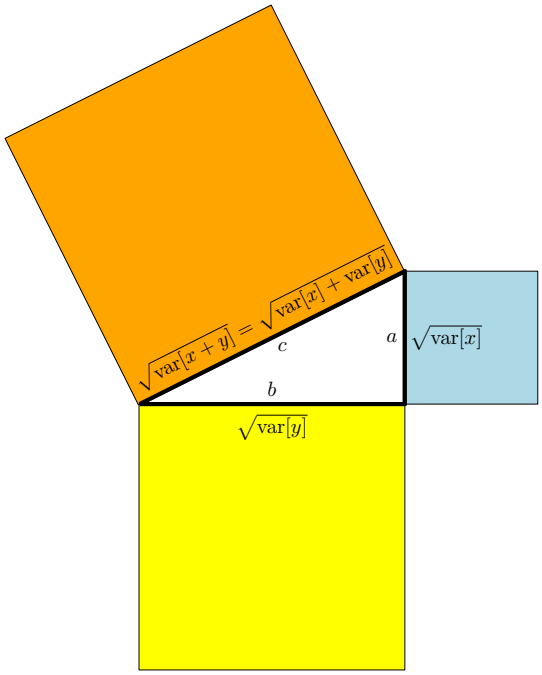

Figure 6.4 Geometry of random variables. If random variables x and y are uncorrelated they are orthogonal vectors in a corresponding vector space, and the Pythagorean theorem applies.

can define inner products to obtain geometric properties of random variables. If we define

$$\langle x, y \rangle := \mathrm{Cov}[x, y] \tag{6.57}$$

we see that the covariance is symmetric, positive definite[1], and linear in either argument[2] The length of a random variable is

$$\|x\| = \sqrt{\mathrm{Cov}[x, x]} = \sqrt{\mathbb{V}[x]} = \sigma[x]\,, \tag{6.58}$$

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

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

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\|\,\|y\|} = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}}\,. \tag{6.59}$$

We know from Definition 3.6 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 6.4 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

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

distributions. Due to the fact that the probability mass (or density) needs to add up to 1, distributions live in a subspace which is called a manifold. The study of this space of probability distributions is called information geometry. Computing distances between distributions are done using Bregman divergences or f-divergences, which 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 Change of Variables/Inverse transform

It may seem that there are very many known distributions to a beginner, but in reality the set of distributions for which we have names are quite limited. Therefore it is often useful to understand how transformations of 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 $\frac{1}{2}(x_1 + x_2)$?

Remark. 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 covariance 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 other transformations (for example nonlinear) of random variables. ♢

In this section, we need to be explicit about random variables and the values they take, and 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 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.

One can also use the moment generating function to study transformations of random variables (Casella and Berger, 2002, Chapter 2).

As mentioned in the introductory comments in this chapter, 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 6.5
It is worth contrasting this example with the example in Section 6.2.1. Consider a medical test the returns the number of cancerous cells that can be found in the biopsy. The state space the 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?

Remark. An analogy to object oriented programming may provide an alternative view for computer scientists. The distinction between random variables and probability distributions can be thought of as the distinction between objects and classes. A probability distribution defines the behaviour (the probability) corresponding to a particular statistical experiment, which quantifies the uncertainty associated with the experiment. A random variable is a particular instantiation of this statistical experiment, which follows the probabilities defined by the distribution. ♢

Transformations of discrete random variables can be understood directly. Given a discrete random variable x with probability mass function $p_x(\mathrm{x})$ (Section 6.2.1), and an invertible function $g(x)$ with inverse $h(\cdot)$. Let y be the random variable transformed by $g(x)$, that is $y = g(x)$. Then

$$p_y(\mathrm{y}) = p_x(h(\mathrm{y})). \tag{6.60}$$

This can be seen by the following short derivation,

$$\begin{aligned} p_y(\mathrm{y}) &= P(y = \mathrm{y}) && \text{definition of pmf} && (6.61)\\ &= P(g(x) = \mathrm{y}) && \text{transformation of interest} && (6.62)\\ &= P(x = h(\mathrm{y})) && \text{inverse} && (6.63)\\ &= p_x(h(\mathrm{y})) && \text{definition of pmf.} && (6.64) \end{aligned}$$

Therefore for discrete random variables, transformations directly change the individual probability of events. The following discussion focuses on continuous random variables and we will need both probability density functions $p(x)$ and cumulative distribution functions $P(x \leqslant \mathrm{x})$.

6.5.1 Distribution Function Technique

The distribution function technique goes back to first principles, and uses the definition of a cumulative distribution function (cdf) and the fact that its differential is the probability density function (pdf) (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:

$$F_y(\mathrm{Y}) = P(y \leqslant \mathrm{Y}) \tag{6.65}$$

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

$$f(y) = \frac{d}{dy} F_y(\mathrm{Y}) \tag{6.66}$$

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

Example 6.6
Let x be a continuous random variable with the following probability density function on $0 < x < 1$

$$f(x) = 3x^2. \tag{6.67}$$

What is the pdf of $y = x^2$?

Note that the function f is an increasing function of x and also the resulting value of y is in the interval $(0, 1)$.

$$\begin{aligned}
F_y(\mathrm{Y}) &= P(y \leqslant \mathrm{Y}) && \text{definition of cdf} && (6.68)\\
&= P(x^2 \leqslant \mathrm{Y}) && \text{transformation of interest} && (6.69)\\
&= P(x \leqslant \mathrm{Y}^{\frac{1}{2}}) && \text{inverse} && (6.70)\\
&= P_x(\mathrm{Y}^{\frac{1}{2}}) && \text{definition of cdf} && (6.71)\\
&= \int_0^{\mathrm{Y}^{\frac{1}{2}}} 3t^2 \mathrm{d}t && \text{cdf as a definite integral} && (6.72)\\
&= \left[t^3\right]_{t=0}^{t=\mathrm{Y}^{\frac{1}{2}}} && \text{result of integration} && (6.73)\\
&= \mathrm{Y}^{\frac{3}{2}}, \quad 0 < \mathrm{Y} < 1 && && (6.74)
\end{aligned}$$

Therefore the cdf of y is

$$F_y(\mathrm{Y}) = \mathrm{Y}^{\frac{3}{2}} \tag{6.75}$$

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

$$f_y(\mathrm{Y}) = \frac{d}{d\mathrm{Y}} F_y(\mathrm{Y}) = \frac{3}{2}\mathrm{Y}^{\frac{1}{2}} \tag{6.76}$$

for $0 < \mathrm{Y} < 1$.

In the previous example, we considered a monotonically increasing function x^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 $x = U^{-1}(y)$. One useful result that can be obtained by applying the technique above when the transformation of interest is the cumulative distri-

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

bution function of the random variable itself (Casella and Berger, 2002, Theorem 2.1.10).

Theorem 6.11. *Let x be a continous random variable with cumulative distribution function $F_x(\cdot)$. Then the random variable y defined as*

$$y = F_x(\mathrm{X}), \tag{6.77}$$

has a uniform distribution.

Proof We need to show that the cumulative distribution function (cdf) of y defines a distribution of a uniform random variable. Recall that by the axioms of probability (Section 6.1) that probabilities must be non-negative and sum to one. Therefore the range of possible values of $y = F_x(\mathrm{X})$ is in the interval $[0, 1]$. Note that for any $F_x(\cdot)$, the inverse $F_x^{-1}(\cdot)$ exists because cdfs are monotone increasing, which we will use in the following proof. Given any continuous random variable x, the definition of a cdf gives

$$\begin{aligned}
F_y(\mathrm{Y}) &= P(y \leqslant \mathrm{Y}) && && (6.78)\\
&= P(F_x(\mathrm{X}) \leqslant \mathrm{Y}) && \text{transformation of interest} && (6.79)\\
&= P(x \leqslant F_x^{-1}(\mathrm{Y})) && \text{inverse exists} && (6.80)\\
&= F_x(F_x^{-1}(\mathrm{Y})) && \text{definition of cdf} && (6.81)\\
&= \mathrm{Y}, && && (6.82)
\end{aligned}$$

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

This result (Theorem 6.11) is known as the *probability integral transform*, and is used to derive algorithms for sampling from distributions by transforming the result of sampling from a uniform random variable. 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).

6.5.2 Change of Variables

The argument from first principles in the previous section 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 deriving a more general approach called change of variables.

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

$$F_y(\mathrm{Y}) = P(y \leqslant \mathrm{Y})\,. \tag{6.83}$$

We are interested in a function U of the random variable

$$P(y \leqslant \mathrm{Y}) = P(U(x) \leqslant \mathrm{Y})\,, \tag{6.84}$$

and we assume that the function U is invertible. By multiplying both sides with the inverse

$$P(U(x) \leqslant y) = P(U^{-1}(U(x)) \leqslant U^{-1}(\mathrm{Y})) = P(x \leqslant U^{-1}(\mathrm{Y})) \tag{6.85}$$

we obtain an expression of the cdf of x. Recall the definition of the cdf in terms of the pdf

$$P(x \leqslant U^{-1}(\mathrm{Y})) = \int_a^{U^{-1}(\mathrm{Y})} f(x)\mathrm{d}x\,. \tag{6.86}$$

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

$$F_y(\mathrm{Y}) = \int_a^{U^{-1}(\mathrm{Y})} f(x)\mathrm{d}x \tag{6.87}$$

To obtain the pdf, we differentiate the expression above with respect to y. Since the expression is in terms of x, we apply the chain rule of calculus from (5.56) and obtain

$$f_y(\mathrm{Y}) = \frac{d}{d\mathrm{Y}} F_y(\mathrm{Y}) = \frac{d}{d\mathrm{Y}} \int_a^{U^{-1}(\mathrm{Y})} f(x)\mathrm{d}x \tag{6.88}$$

$$= f_x(U^{-1}(\mathrm{Y})) \times \left| \det\left(\frac{d}{d\mathrm{Y}} U^{-1}(\mathrm{Y}) \right) \right|. \tag{6.89}$$

The fact that integration and differentiation are somehow "inverses" of each other is due to a deep result called the Fundamental Theorem of Calculus.

This is called the *change of variable* technique. The term $\left|\frac{d}{d\mathrm{Y}} U^{-1}(\mathrm{Y})\right|$ measures how much a unit volume changes when applying U. Recall from Section 4.1 that the existence of the determinant shows that we can invert the Jacobian. Recall further that the determinant arises because our differentials (cubes of volume) are transformed into parallelepipeds by the determinant. In the last expression above, we have 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.

change of variable

Remark. Observe that in comparison to the discrete case in Equation (6.60), we have an additional factor $\left|\frac{d}{dy} U^{-1}(y)\right|$. The continuous case requires more care because $P(y = \mathrm{Y}) = 0$ for all Y. The probability density function $f_y(\mathrm{Y})$ 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 Equation (5.68) that the Jacobian is a matrix of partial derivatives. Let us summarize the discussion above in the following theorem which describes the recipe for multivariate change of variables.

Theorem 6.12. *Let $f_{\boldsymbol{x}}(\mathrm{X})$ be the value of the probability density of the multivariate continuous random variable $\boldsymbol{x}$ at X. If the vector valued function $\boldsymbol{y} = U(\boldsymbol{x})$ is differentiable and invertible for all values within the range of $\boldsymbol{x}$, then for corresponding values of $\boldsymbol{y}$, the probability density of $y = U(x)$ is given by*

$$f_{\boldsymbol{y}}(\mathrm{Y}) = f_{\boldsymbol{x}}(U^{-1}(\mathrm{Y})) \times \left| \det \left(\frac{\partial}{\partial \mathrm{Y}} U^{-1}(\mathrm{Y}) \right) \right| . \tag{6.90}$$

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

Example 6.7

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

$$f_{\boldsymbol{x}} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) = \frac{1}{2\pi} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) . \tag{6.91}$$

We use the change of variable technique (Theorem 6.12) to derive the effect of an linear transformation (Section 2.7) of the random variables. Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ defined as

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} . \tag{6.92}$$

What is the probability density function of the resulting 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 are considering linear transformations, the inverse transformation is given matrix inverse from Section 2.2.2. For 2×2 matrices, we can explicitly write out the formula,

given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{A}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{6.93}$$

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

$$f_{\boldsymbol{x}}(\mathrm{X}) = f_{\boldsymbol{x}}(\boldsymbol{A}^{-1}\mathrm{Y}) \tag{6.94}$$

$$= \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathrm{Y}^\top \boldsymbol{A}^{-\top}\boldsymbol{A}^{-1}\mathrm{Y}\right). \tag{6.95}$$

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

$$\frac{\partial}{\partial \mathrm{Y}} \boldsymbol{A}^{-1}\mathrm{Y} = \boldsymbol{A}^{-1}. \tag{6.96}$$

Recall from Section 4.1 that the determinant of the inverse is the inverse of the determinant, and therefore the determinant of the Jacobian matrix is given by

$$\left|\tfrac{\partial}{\partial \mathrm{Y}} \boldsymbol{A}^{-1}\mathrm{Y}\right| = ad - bc. \tag{6.97}$$

We are now able to apply the change of variable formula from Theorem 6.12, by multiplying Equation (6.95) with Equation (6.97),

$$f_y(\mathrm{Y}) = f_x(\mathrm{X}) \times \left|\left|\tfrac{\partial}{\partial \mathrm{Y}} \boldsymbol{A}^{-1}\mathrm{Y}\right|\right| \tag{6.98}$$

$$= \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathrm{Y}^\top \boldsymbol{A}^{-\top}\boldsymbol{A}^{-1}\mathrm{Y}\right)(ad - bc). \tag{6.99}$$

While the example above is based on a bivariate random variable so that we can compute the matrix inverse in closed form, the relation above holds true for higher dimensions.

Remark. We will see in Section 6.6 that the density $f_{\boldsymbol{x}}(\mathrm{X})$ above is actually the standard Gaussian distribution, and the transformed density $f_{\boldsymbol{y}}(\mathrm{Y})$ is a bivariate Gaussian with covariance $\boldsymbol{\Sigma} = \boldsymbol{A}^\top \boldsymbol{A}$. The linear transformation $\boldsymbol{A}$ turns out to correspond to the Cholesky factorization (Section 4.3) of $\boldsymbol{\Sigma}$. ♢

6.6 Gaussian 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

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

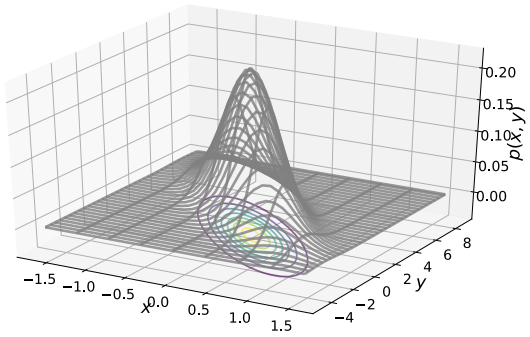

Figure 6.5 Gaussian distribution of two random variables x, y.

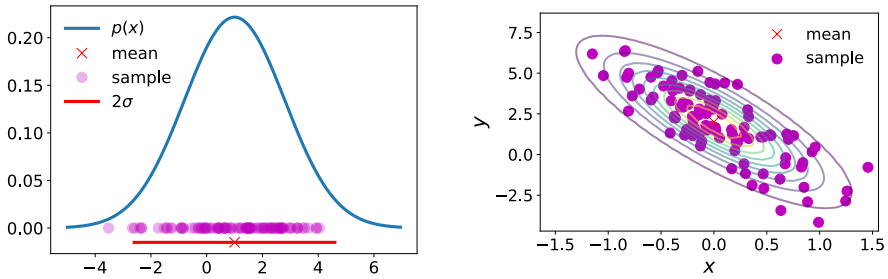

Figure 6.6 Gaussian distributions overlaid with 100 samples. Left: Univariate (1-dimensional) Gaussian; The red cross shows and mean and the red line the extent of the variance. Right: Multivariate (2-dimensional) Gaussian, viewed from top. The red cross shows the mean and the coloured lines contour lines of the density.

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

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6.100}$$

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

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{6.101}$$

where $\boldsymbol{x} \in \mathbb{R}^D$ is a random variable. We write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Figure 6.5 shows a bi-variate Gaussian (mesh), with the corresponding contour plot. The special case of the Gaussian with zero mean and identity variance, that is $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, is referred to as the *standard normal distribution*.

standard normal distribution

Gaussian distributions are widely used in statistical estimation and ma-

chine 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 6.5) 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.

6.6.1 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 $[\boldsymbol{x}, \boldsymbol{y}]^\top$,

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right). \tag{6.102}$$

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 6.7) and given by

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{\mu}_{x \mid y}, \boldsymbol{\Sigma}_{x \mid y}\big) \tag{6.103}$$

$$\boldsymbol{\mu}_{x \mid y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{6.104}$$

$$\boldsymbol{\Sigma}_{x \mid y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\,. \tag{6.105}$$

Note that in the computation of the mean in (6.104) 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

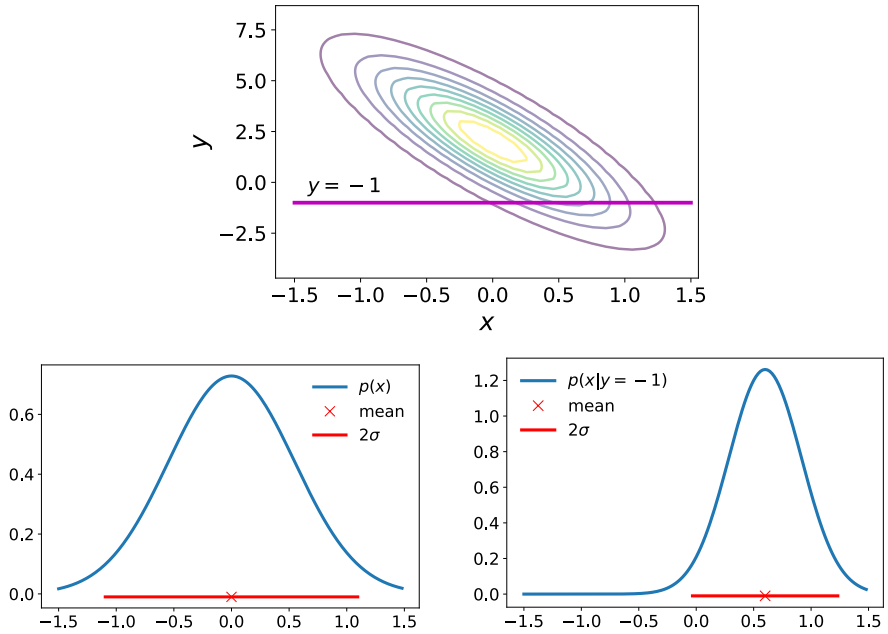

Figure 6.7 Top: Bivariate Gaussian; Bottom left: Marginal of a joint Gaussian distribution is Gaussian; Bottom right: The conditional distribution of a Gaussian is also Gaussian

(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 (6.102), is itself Gaussian and computed by applying the sum-rule in (6.18) and given by

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x, \, \boldsymbol{\Sigma}_{xx}\right). \tag{6.106}$$

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 (6.102), we ignore (i.e., integrate out) everything we are not interested in. This is illustrated on the bottom left of Figure 6.7.

Example 6.8

Consider the bivariate Gaussian distribution (illustrated in Figure 6.7)

$$p(x, y) = \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0.3 & -1\\-1 & 5\end{bmatrix}\right). \tag{6.107}$$

We can compute the parameters of the univariate Gaussian, conditioned

on $y = -1$, by applying (6.104) and (6.105) to obtain the mean and variance respectively. Numerically, this is

$$\mu_{x\,|\,y=-1} = 0 + (-1)(0.2)(-1-2) = 0.6 \tag{6.108}$$

and

$$\sigma^2_{x\,|\,y=-1} = 0.3 - (-1)(0.2)(-1) = 0.1. \tag{6.109}$$

Therefore the conditional Gaussian is given by

$$p(x\,|\,y=-1) = \mathcal{N}\big(0.6,\, 0.1\big). \tag{6.110}$$

The marginal distribution $p(x)$ in contrast can be obtained by applying (6.106), which is essentially using the mean and variance of the random variable x, giving us

$$p(x) = \mathcal{N}\big(0,\, 0.3\big) \tag{6.111}$$

6.6.2 Product of Gaussians

In machine learning, we often assume that examples are perturbed by Gaussian noise, leading to a Gaussian likelihood for linear regression. 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 Gaussians. 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 an unnormalized Gaussian distribution $c\mathcal{N}\big(\boldsymbol{x}\,|\,\boldsymbol{c},\,\boldsymbol{C}\big)$ with

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \tag{6.112}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}) \tag{6.113}$$

$$c = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A}+\boldsymbol{B}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{a}-\boldsymbol{b})^\top(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{a}-\boldsymbol{b})\big)\,. \tag{6.114}$$

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

Remark. For notation convenience, we will sometimes use $\mathcal{N}\big(\boldsymbol{x}\,|\,\boldsymbol{m},\,\boldsymbol{S}\big)$ 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

$$c = \mathcal{N}\big(\boldsymbol{a}\,|\,\boldsymbol{b},\,\boldsymbol{A}+\boldsymbol{B}\big) = \mathcal{N}\big(\boldsymbol{b}\,|\,\boldsymbol{a},\,\boldsymbol{A}+\boldsymbol{B}\big)\,. \tag{6.115}$$

Here, neither $\boldsymbol{a}$ nor $\boldsymbol{b}$ are random variables. However, writing c in this way is more compact than (6.114). ♢

6.6.3 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 $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, then $\boldsymbol{x} + \boldsymbol{y}$ is also Gaussian distributed and given by

$$p(\boldsymbol{x} + \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y) . \tag{6.116}$$

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 6.9
Since expectations are linear operations, we can obtain the weighted sum of independent Gaussian random variables

$$p(a\boldsymbol{x} + b\boldsymbol{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y, a\boldsymbol{\Sigma}_x + b\boldsymbol{\Sigma}_y) . \tag{6.117}$$

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 6.13, the random variable z is from the mixture density of the two random variables x and y. 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 requires more care.

Theorem 6.13. *Consider a weighted sum of two univariate Gaussian densities*

$$p(z) = \alpha p(x) + (1 - \alpha)p(y) \tag{6.118}$$

where the scalar $0 < \alpha < 1$ is the mixture weight, and $p(x)$ and $p(y)$ are univariate Gaussian densities (Equation (6.100)) with different parameters, that is $(\mu_x, \sigma_x^2) \neq (\mu_y, \sigma_y^2)$.

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

$$\mathbb{E}[z] = \alpha\mu_x + (1 - \alpha)\mu_y. \tag{6.119}$$

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

$$\mathbb{V}[z] = \left[\alpha\sigma_x^2 + (1 - \alpha)\sigma_y^2\right] + \left(\left[\alpha\mu_x^2 + (1 - \alpha)\mu_y^2\right] \left[\alpha\mu_x + (1 - \alpha)\mu_y\right]^2\right) . \tag{6.120}$$

Proof The mean of the mixture z is given by the weighted sum of the

means of each random variable. We apply the definition of the mean (Definition 6.3), and plug in our mixture (Equation (6.118)) above

$$\mathbb{E}[z] = \int_{-\infty}^{\infty} zp(z)\mathrm{d}z \tag{6.121}$$

$$= \int_{-\infty}^{\infty} \alpha zp(x) + (1-\alpha)zp(y)\mathrm{d}z \tag{6.122}$$

$$= \alpha\int_{-\infty}^{\infty} zp(x)\mathrm{d}z + (1-\alpha)\int_{-\infty}^{\infty} zp(y)\mathrm{d}z \tag{6.123}$$

$$= \alpha\mu_x + (1-\alpha)\mu_y. \tag{6.124}$$

To compute the variance, we can use the raw score version of the variance (Equation (6.40)), 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 6.4).

$$\mathbb{E}[z^2] = \int_{-\infty}^{\infty} z^2p(z)\mathrm{d}z \tag{6.125}$$

$$= \int_{-\infty}^{\infty} \alpha z^2p(x) + (1-\alpha)z^2p(y)\mathrm{d}z \tag{6.126}$$

$$= \alpha\int_{-\infty}^{\infty} z^2p(z)\mathrm{d}z + (1-\alpha)\int_{-\infty}^{\infty} z^2p(y)\mathrm{d}z \tag{6.127}$$

$$= \alpha(\mu_x^2+\sigma_x^2) + (1-\alpha)(\mu_y^2+\sigma_y^2). \tag{6.128}$$

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 the two terms above

$$\mathbb{V}[z] = \mathbb{E}[z^2] - (\mathbb{E}[z])^2 \tag{6.129}$$

$$= \alpha(\mu_x^2+\sigma_x^2) + (1-\alpha)(\mu_y^2+\sigma_y^2) - (\alpha\mu_x + (1-\alpha)\mu_y)^2 \tag{6.130}$$

$$= \left[\alpha\sigma_x^2 + (1-\alpha)\sigma_y^2\right] + \left(\left[\alpha\mu_x^2 + (1-\alpha)\mu_y^2\right]\left[\alpha\mu_x + (1-\alpha)\mu_y\right]^2\right). \tag{6.131}$$

Observe 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. ♢

Recall the example in Section 6.5, where we considered a bivariate standard Gaussian random variable X and performed a linear transformation $\boldsymbol{A}X$ on it. The outcome was a Gaussian random variable with zero mean

and covariance $\boldsymbol{A}^\top\boldsymbol{A}$. 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 $\boldsymbol{x} \sim \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)$. 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:

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu}. \tag{6.132}$$

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

$$\mathbb{V}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top. \tag{6.133}$$

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

$$p(\boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{A}\boldsymbol{\mu},\, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big). \tag{6.134}$$

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 matrix $\boldsymbol{A}$ of appropriate shape, let $\boldsymbol{y}$ be a Gaussian random variable with mean $\boldsymbol{A}\boldsymbol{x}$, i.e.,

$$p(\boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{y} \,|\, \boldsymbol{A}\boldsymbol{x},\, \boldsymbol{\Sigma}\big). \tag{6.135}$$

What is the corresponding probability distribution $p(\boldsymbol{x})$? If $\boldsymbol{A}$ is invertible, then we can write $\boldsymbol{x} = \boldsymbol{A}^{-1}y$ and apply the transformation in the previous paragraph. However in general $\boldsymbol{A}$ is not invertible, and we use an approach similar to the that of the pseudo-inverse (Equation 3.54). That is we pre-multiply both sides with $\boldsymbol{A}^\top$ and then invert $\boldsymbol{A}^\top\boldsymbol{A}$ which is symmetric and positive definite, giving us the relation

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \iff (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y} = \boldsymbol{x}\,. \tag{6.136}$$

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

$$p(\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{x} \,|\, (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y},\, (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\big)\,. \tag{6.137}$$

6.6.4 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 $\mathcal{N}\big(\boldsymbol{0},\, \boldsymbol{I}\big)$.

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 $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

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.

To obtain samples from a multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can use the properties of a linear transformation of a Gaussian random variable: If $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 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 $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$ is the Cholesky factorization of $\boldsymbol{\Sigma}$.

6.7 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. 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.
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.

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.

exponential family

Example 6.10

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

Bernoulli distribution

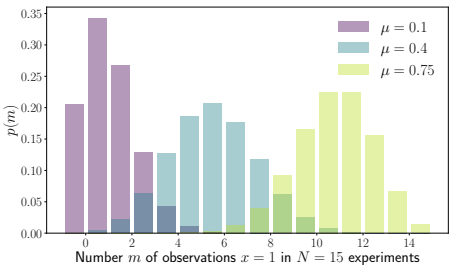

Figure 6.8 Examples of the Binomial distribution for $\mu \in \{0.1, 0.4, 0.75\}$ and $N = 15$.

as

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}, \quad x \in \{0, 1\}, \tag{6.138}$$

$$\mathbb{E}[x] = \mu, \tag{6.139}$$

$$\mathbb{V}[x] = \mu(1-\mu), \tag{6.140}$$

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.

Example 6.11

Binomial distribution

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

$$p(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}, \tag{6.141}$$

$$\mathbb{E}[m] = N\mu, \tag{6.142}$$

$$\mathbb{V}[m] = N\mu(1-\mu) \tag{6.143}$$

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 6.12

The Beta distribution is a distribution over a continuous variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event

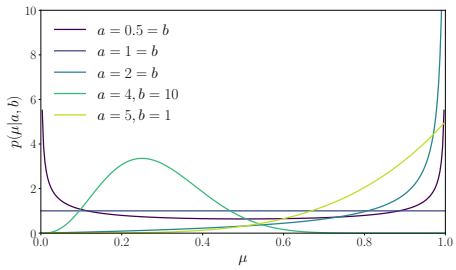

Figure 6.9 Examples of the Beta distribution for different values of α and β.

(e.g., the parameter governing the Bernoulli distribution). The Beta distribution (illustrated in Figure 6.9) itself is governed by two parameters $\alpha > 0,\ \beta > 0$ and is defined as

$$p(\mu \,|\, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{6.144}$$

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta}\,, \qquad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{6.145}$$

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

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx, \qquad t > 0\,. \tag{6.146}$$

$$\Gamma(t+1) = t\Gamma(t)\,. \tag{6.147}$$

Note that the fraction of Gamma functions in (6.144) 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 6.7.1) and exponential families (Section 6.153). ♢

6.7.1 Conjugacy

According to Bayes' theorem (6.21), 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 6.14 (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 6.13 (Beta-Binomial Conjugacy)
Consider a Binomial random variable $x \sim \text{Bin}(m \,|\, N, \mu)$ where

$$p(x \,|\, \mu, N) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \;\propto\; \mu^a (1-\mu)^b \tag{6.148}$$

for some constants a, b. We place a Beta prior on the parameter μ:

$$\text{Beta}(\mu \,|\, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \;\propto\; \mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{6.149}$$

If we now observe some outcomes $\boldsymbol{x} = (x_1, \ldots, x_N)$ of a repeated coin-flip experiment with h heads and t tails, we compute the posterior distribution on μ as

$$p(\mu \,|\, \boldsymbol{x} = h) \;\propto\; p(\boldsymbol{x} \,|\, \mu) p(\mu \,|\, \alpha, \beta) = \mu^h (1-\mu)^t \mu^{\alpha-1}(1-\mu)^{\beta-1} \tag{6.150}$$

$$= \mu^{h+\alpha-1}(1-\mu)^{t+\beta-1} \;\propto\; \text{Beta}(h+\alpha, t+\beta) \tag{6.151}$$

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.

Table 6.2 lists examples for conjugate priors for the parameters of some of standard likelihoods used in probabilistic modeling. Distributions such

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

Table 6.2 Examples of conjugate priors for common likelihood functions.

as Multinomial, inverse Gamma, inverse Wishart, and Dirichlet can be found in any statistical text, and is 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 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.

6.7.2 Sufficient Statistics

Recall that a statistic of a random variable is a deterministic function of that random variable. For example if $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$ 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 6.15 (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*

$$p(x \mid \theta) = h(x)g_\theta(\phi(x)). \tag{6.152}$$

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

Note that 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.

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.

6.7.3 Exponential Family

At this point it is worth being a bit careful by discussing three possible levels of abstraction we can have when considering distributions (of discrete or continuous random variables). At 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 fix the parametric form (the univariate Gaussian) and infer the parameters from data. For example, we assume a univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$ with unknown mean μ and unknown variance σ^2, and use a maximum likelihood fit to determine the best parameters (μ, σ^2). We will see an example of this when considering linear regression in Chapter 9. A third level of abstraction is to consider families of distributions, and in this book, we consider the exponential family. The univariate Gaussian is an example of a member of the exponential family. Many of the widely used statistical models, including all the "named" models in Table 6.2, are members of the exponential family. They can all be unified into one concept (Brown, 1986).

Remark. A brief historical anecdote: like many concepts in mathematics and science, exponential families were independently discovered at the same time by different researchers. In the years 1935–1936, Edwin Pitman in Tasmania, Georges 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

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right) , \tag{6.153}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is the vector of sufficient statistics. In general, any inner prod-

uct (Section 3.2) can be used in (6.153), and for concreteness we will use the standard dot product here. Note that the form of the exponential family is essentially a particular expression of $g_\theta(\phi(x))$ in the Fisher-Neyman theorem (Theorem 6.15).

The factor $h(\boldsymbol{x})$ can be absorbed into the dot product term by adding another entry to the vector of sufficient statistics $\log h(\boldsymbol{x})$, and constraining the corresponding parameter $\theta = 1$. 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

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})\right). \tag{6.154}$$

For this form of parameterization, the parameters $\boldsymbol{\theta}$ are called the *natural paramters*. 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 to the fact that we can capture information about data in $\boldsymbol{\phi}(\boldsymbol{x})$.

natural paramters

Example 6.14 (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,

$$p(x \mid \boldsymbol{\theta}) \propto \exp(\theta_1 x + \theta_2 x^2). \tag{6.155}$$

Setting

$$\theta = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]^\top \tag{6.156}$$

and substituting into (6.155) we obtain

$$p(x \mid \boldsymbol{\theta}) \propto \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{6.157}$$

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

Exponential families also provide a convenient way to find conjugate pairs of distributions. In the following example, we will derive a result that is similar to the Beta-Binomial conjugacy result of Section 6.7.1. Here we will show that the Beta distribution is a conjugate prior for the Bernoulli distribution.

Example 6.15 (Beta-Bernoulli Conjugacy)
Let $x \in \{0, 1\}$ be distributed according to the Bernoulli distribution with parameter $\theta \in [0, 1]$, that is $P(x = 1 \,|\, \theta) = \theta$. This can also be expressed as $P(x \,|\, \theta) = \theta^x(1 - \theta)^{1-x}$. Let θ be distributed according to a Beta distribution with parameters α, β, that is $p(\theta \,|\, \alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$.

Multiplying the Beta and the Bernoulli distributions, we get

$$p(\theta \,|\, x, \alpha, \beta) = P(x \,|\, \theta) \times p(\theta \,|\, \alpha, \beta) \tag{6.158}$$
$$\propto \theta^x(1 - \theta)^{1-x} \times \theta^{\alpha-1}(1 - \theta)^{\beta-1} \tag{6.159}$$
$$= \theta^{\alpha+x-1}(1 - \theta)^{\beta+(1-x)-1} \tag{6.160}$$
$$\propto p(\theta \,|\, \alpha + x, \beta + (1 - x)). \tag{6.161}$$

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

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. ♢

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).

6.8 Further Reading

Probabilistic models in machine learning Bishop (2006); Murphy (2012) provide a way for users to capture uncertainty about data and predictive models in a principled fashion. Ghahramani (2015) presents a short review of probabilistic models in machine learning. This chapter is rather terse at times, and Grinstead and Snell (1997) provides a more relaxed presentation that is 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).

Given a probabilistic model, we may be lucky enough to be able to compute parameters of interest analytically. However in general analytic solutions are rare and computational methods such as sampling (Brooks et al.,

2011) and variational inference (Blei et al., 2017) are used. 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).

A more technical audience interested in the details of probability theory have many options (Jacod and Protter, 2004; Jaynes, 2003; Mackay, 2003) including some very technical discussions (Dudley, 2002; Shiryayev, 1984; Lehmann and Casella, 1998; Bickel and Doksum, 2006). We side stepped a large part of the difficulty by glossing over 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. As machine learning allows us to model move intricate distributions on ever move 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 (2003); Bishop (2006); Murphy (2012); Barber (2012); Rasmussen and Williams (2006).

Exercises

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

6.2 Consider the following time-series model:

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

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

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 $p(\boldsymbol{x}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t) = \mathcal{N}\big(\boldsymbol{\mu}_t,\, \boldsymbol{\Sigma}_t\big)$.
    1. Compute $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$
    2. Compute $p(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$
    3. At time $t{+}1$, we observe the value $\boldsymbol{y}_{t+1} = \hat{\boldsymbol{y}}$. Compute $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t+1})$.

6.3 Prove the relationship in Equation 6.40, which relates the standard definition of the variance to the raw score expression for the variance.

6.4 Prove the relationship in Equation 6.41, which relates the pairwise difference between examples in a dataset with the raw score expression for the variance.

6.5 Express the Bernoulli distribution in the natural parameter form of the exponential family (Equation (6.153)).

6.6 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.

6.7 **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 $p(x|y)$.

6.8 **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

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w}\,, \tag{6.162}$$

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.[3] 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

$$\boldsymbol{z} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{v}\,, \tag{6.163}$$

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 $p(\boldsymbol{x}|\hat{\boldsymbol{y}})$.[4]
   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.

[3] 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.

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