Chapter 7

AN INTRODUCTION TO PROBABILITY

As the previous chapters have illustrated, it is often quite easy to come up with physical models that determine the effects that result from various causes — we know how image intensity is determined, for example. The difficulty is that effects could have come from various causes and we would like to know which — for example, is the image dark because the light level is low, or because the surface has low albedo? Ideally, we should like to take our measurements and determine a reasonable description of the world that generated them. Accounting for uncertainty is a crucial component of this process, because of the ambiguity of our measurements. This process of accountancy needs to take into account reasonable preferences about the state of the world — for example, it is less common to see very dark surfaces under very bright lights than it is to see a range of albedoes under a reasonably bright light.

Probability is the proper mechanism for accounting for uncertainty. Axiomatic probability theory is gloriously complicated, and we don't attempt to derive the ideas in detail. Instead, this chapter will first review the basic ideas of probability. We then describe techniques for building probabilistic models and for extracting information from a probabilistic model, all in the context of quite simple examples. In chapter **??**, we show some substantial examples of probabilistic methods; there are other examples scattered about the text by topic.

Discussions of probability are often bogged down with waffle about what probability *means*, a topic that has attracted a spectacular quantity of text. Instead, we will discuss probability as a modelling technique with certain formal, abstract properties — this means we can dodge the question of what the ideas mean and concentrate on the far more interesting question of what they can do for us.

We will develop probability theory in discrete spaces first, because it is possible to demonstrate the underpinning notions without much notation. We then pass to continuous spaces.

7.1 Probability in Discrete Spaces

Generally, a probability model is used to compare various kinds of experimental outcome that can be distinguished. These outcomes are usually called **events**. Now if it is possible to tell whether an event has occurred, it is possible to tell if it has not occurred, too. Furthermore, if it is possible to tell that two events have occurred independently, then it is possible to tell if they have occurred simultaneously.

This motivates a formal structure. We take a discrete space, D, which could be infinite and which represents the world in which experiments occur. Now construct a collection of subsets of D, which we shall call $\mathcal{F}$, with the following properties:

- The empty set is in $\mathcal{F}$ and so is D.
- *Closure under intersection:* if $S_1 \in \mathcal{F}$ and $S_2 \in \mathcal{F}$, then $S_1 \cap S_2 \in \mathcal{F}$.
- *Closure under complements:* if $S_1 \in \mathcal{F}$ then $\overline{S_1} = D - S_1 \in \mathcal{F}$.

The elements of $\mathcal{F}$ correspond to the events. Note that we can we can tell whether any logical combinations of events has occurred, too, because a logical combination of events corresponds to set unions, negations or intersections.

Given a coin that is flipped once,

$$D = \{\texttt{heads}, \texttt{tails}\}$$

There are only two possible sets of events in this case:

$$\{\emptyset, D\}$$

(which implies we flipped the coin, but can't tell what happened!) and

$$\{\emptyset, D, \{\texttt{heads}\}, \{\texttt{tails}\}\}$$

Example 7.1: *The space of events for a single toss of a coin.*

7.1.1 Probability: the P-function

Now we construct a function P, which takes elements of $\mathcal{F}$ to the unit interval. We require that P has some important properties:

- P is defined for every element of $\mathcal{F}$
- $P(\emptyset) = 0$
- $P(D) = 1$
- for $A \in \mathcal{F}$ and $B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Given two coins that are flipped,

$$D = \{\mathtt{hh}, \mathtt{ht}, \mathtt{tt}, \mathtt{th}\}$$

There are rather more possible sets of events in this case. One useful one would be

$$\mathcal{F} = \left\{ \begin{array}{cccc} \emptyset, & D, & & \\ \{\mathtt{hh}\}, & \{\mathtt{ht}\}, & \{\mathtt{tt}\}, & \{\mathtt{th}\}, \\ \{\mathtt{hh}, \mathtt{ht}\}, & \{\mathtt{hh}, \mathtt{th}\}, & \{\mathtt{hh}, \mathtt{tt}\}, & \{\mathtt{ht}, \mathtt{th}\}, \\ \{\mathtt{ht}, \mathtt{tt}\}, & \{\mathtt{th}, \mathtt{tt}\}, & \{\mathtt{hh}, \mathtt{ht}, \mathtt{th}\}, & \{\mathtt{hh}, \mathtt{ht}, \mathtt{tt}\}, \\ \{\mathtt{hh}, \mathtt{th}, \mathtt{tt}\}, & \{\mathtt{ht}, \mathtt{th}, \mathtt{tt}\}\} & & \end{array} \right\}$$

which would correspond to all possible cases. Another (perhaps less useful) structure would be:

$$\mathcal{F} = \{\emptyset, D, \{\mathtt{hh}, \mathtt{ht}\}, \{\mathtt{th}, \mathtt{tt}\}\}$$

which implies that we cannot measure the state of the second coin

Example 7.2: *Two possible spaces of events for a single flip each of two coins.*

which we call *the axiomatic properties of probability.* Note that $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$, because the function takes elements of $\mathcal{F}$ to the unit interval. We call the collection of D, P and $\mathcal{F}$ a **probability model**. We call $P(A)$ the **probability of the event A** — because we are still talking about formal structures, there is absolutely no reason to discuss what this means; it's just a name. Rigorously justifying the properties of P is somewhat tricky — Jaynes' book ([]) is one place to start, as is []. It can be helpful to think of P as a function that measures the size of a subset of D — the whole of D has size one, and the size of the union of two disjoint sets is the sum of their sizes.

In example 1, for the first structure on D, there is only one possible choice of P; for the second, there is a one parameter family of choices, we could choose $P(\mathtt{heads})$ to be an arbitrary number in the unit interval, and the choice of $P(\mathtt{tails})$ follows.

Example 7.3: *The possible P functions for the flip of a single coin.*

In example 2, there is a three-parameter family of choices for P in the case of the first event structure shown in that example — we can choose $P(\mathtt{hh})$, $P(\mathtt{ht})$ and $P(\mathtt{th})$, and all other values will be given by the axioms. For the second event structure in that example, P is the same as that for a single coin (because we can't tell the state of one coin).

Example 7.4: *The P functions for two coins, each flipped once.*

7.1.2 Conditional Probability

If we have some element A of $\mathcal{F}$ where $P(A) \neq 0$ — and this constraint is important — then the collection of sets

$$\mathcal{F}_{\mathcal{A}} = \{u \cap A | u \in \mathcal{F}\}$$

has the same properties as $\mathcal{F}$. Furthermore, the function with domain $\mathcal{F}_A$

$$P_A(C) = \frac{P(C)}{P(A)}$$

(where $C \in \mathcal{F}_A$) also satisfies the axiomatic properties of probability on its domain. We call this function **the conditional probability of C, given A**; it is usually written as $P(C|A)$. If we adopt the metaphor that P measures the size of a set, then the conditional probability measures the size of the set $C \cap A$ relative to A. Notice that

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

an important fact that you should memorize.

Assume that we have a collection of n sets A_i, such that $A_j \cap A_k = \emptyset$ for every $j \neq k$ and $A = \bigcup_i A_i$. The analogy between probability and size motivates the result that

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

a fact well worth remembering.

7.1.3 Choosing P

We have a formal structure — to use it, we need to choose values of P that have useful semantics. There are a variety of ways of doing this, and it is essential to understand that there is no canonical choice. The choice of P is an essential part of the modelling process. A bad choice will lead to an unhelpful or misleading model, and a good choice may lead to a very enlightening model. There are some strategies that help in choosing P.

Symmetry

Many problems have a form of symmetry that means we have no reason to distinguish between certain sets of events. In this case, it is natural to choose P to reflect this fact.

Assume we have a single coin which we will flip, and we can tell the difference between heads and tails. Then

$$\mathcal{F} = \{\emptyset, D, \{\texttt{heads}\}, \{\texttt{tails}\}\}$$

is a reasonable model to adopt. Now this coin is symmetric — there is no reason to distinguish between the heads side and the tails side from a mechanical perspective. Furthermore, the operation of flipping it subjects it to mechanical forces that do not favour one side over the other. In this case, we have no reason to believe that there is any difference between the outcomes, so it is natural to choose

$$P(\texttt{heads}) = P(\texttt{tails})$$

Example 7.5: *Choosing the P function for a single coin flip using symmetry.*

Assume we have a die that we believe to be fair, in the sense that it has been manufactured to have the symmetries of a cube. A symmetry argument allows us to assume that the probability that each face comes up is equal because we have no reason to prefer faces.

Example 7.6: *Choosing the P function for a roll of a die using symmetry.*

Independence

In many probability models, events do not depend on one another. This is reflected in the conditional probability. If there is no interaction between events A and B, then $P(A|B)$ cannot depend on B. This means that $P(A|B) = P(A)$, a property known as **independence**. In turn, if A and B are independent, we have $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$. This property is important, because it reduces the number of parameters that must be chosen in building a probability model.

A more subtle version of this property is **conditional independence**. Formally, A and B are conditionally independent given C if

$$P(A, B, C) = P(A, B|C)P(C) = P(A|C)P(B|C)P(C)$$

Like independence, conditional independence simplifies modelling by (sometimes substantially) reducing the number of parameters that must be chosen in constructing a model.

We adopt the first of the two event structures given for the two coins in example 2 (this is where we can tell the state of both coins). Now we assume that neither coin knows the other's intentions or outcome.
This assumption restricts our choice of probability model quite considerably because it enforces a symmetry. Let us choose

$$P(\{\mathtt{hh}, \mathtt{ht}\}) = p_{1h}$$

and

$$P(\{\mathtt{hh}, \mathtt{th}\}) = p_{2h}$$

Now let us consider conditional probabilities, in particular

$$P(\{\mathtt{hh}, \mathtt{ht}\}|\{\mathtt{hh}, \mathtt{th}\})$$

(which we could interpret as the probability that the first coin comes up heads given the second coin came up heads). If the coins cannot communicate, then this conditional probability should not depend on the conditioning set, which means that

$$P(\{\mathtt{hh}, \mathtt{ht}\}|\{\mathtt{hh}, \mathtt{th}\}) = P(\{\mathtt{hh}, \mathtt{ht}\})$$

In this case, we know that

$$P(\mathtt{hh}) = P(\{\mathtt{hh}, \mathtt{ht}\}|\{\mathtt{hh}, \mathtt{th}\})P(\{\mathtt{hh}, \mathtt{th}\}) = P(\{\mathtt{hh}, \mathtt{ht}\})P(\{\mathtt{hh}, \mathtt{th}\}) = p_{1h}p_{2h}$$

Similar reasoning yields $P(A)$ for all $A \in \mathcal{F}$, so that our assumption that the two coins are independent means that there is now only a two parameter family of probability models to choose from — one parameter describes the first coin, the other describes the second.

Example 7.7: *Choosing the P function for a single flip each of two coins using the idea of independence.*

Frequency:

Data reflecting the relative frequency of events can be easily converted into a form that satisfies the axioms for P, as example 9 indicates.

An interpretation of probability as frequency is consistent, in the sense that if we make repeated, independent trials of a probability model where P has been allocated using frequency data, then the events with the highest probability — which will be long sequences of outcomes — will be those that show the outcomes with about the right frequency. Example 10 illustrates this effect for repeated flips of a single coin.

Saying that the relative frequency of an event is f means that, in a very large number of trials (say, N), we expect that the event occurs in about fN of those

Both I and my neighbour have a lawn; each lawn has its own sprinkler system. There are two reasons that my lawn could be wet in the morning — either it rained in the night, or my sprinkler system came on. There is no reason to believe that the neighbour's sprinkler system comes on at the same times or on the same days as mine does. Neither sprinkler system is smart enough to know whether it has rained. Finally, if it rains, both lawns are guaranteed to get wet; however, if the sprinkler system comes on, there is some probability that the lawn will not get wet (perhaps a jammed nozzle).
A reasonable model has five binary variables (my lawn is wet or not; the neighbour's lawn is wet or not; my sprinkler came on or not; the neighbour's sprinkler came on or not; and it rained or not). D has 32 elements, and the event space is too large to write out conveniently. If there was no independence in the model, specifying P could require 31 parameters.
However, if I know whether it rained in the night, then the state of my lawn is independent of the state of the neighbour's lawn. Our joint probability function is

$$P(\mathtt{W}, \mathtt{W}_n, \mathtt{S}, \mathtt{S}_n, \mathtt{R}) \;=\; P(\mathtt{W}, \mathtt{S}|\mathtt{R})P(\mathtt{W}_n, \mathtt{S}_n|\mathtt{R})P(\mathtt{R})$$

We know that $P(\mathtt{W} = \text{true}, \mathtt{S}|\mathtt{R} = \textit{true}) = P(\mathtt{S})$ (this just says that if it rains, the lawn is going to be wet); a similar observation applies to the neighbour's lawn. The rain and the sprinklers are independent *and* there is a symmetry — both my neighbour's lawn and mine behave in the same. This means that, in total, we need only 5 parameters to specify this model.

Example 7.8: *Simplifying a model using conditional independence: the case of rain, sprinklers and lawns.*

Assume that, in the past, we have flipped the single coin described above many times, and observed that for 51% of these flips it comes up heads, and for 49% it comes up tails. We could choose

$$P(heads) = 0.51 \text{ and } P(tails) = 0.49$$

This choice is a sensible choice, as example 10 indicates.

Example 7.9: *Choosing a P function for a single coin flip using frequency information.*

trials. Now for large n, the expression

$$\left(\begin{array}{c} k \\ n \end{array}\right) p^k(1-p)^{n-k}$$

Now consider a single coin that we flip many times, and where each flip is independent of the other. We set up an event structure that does not reflect the order in which the flips occur. For example, for two flips, we would have:

$$\{\emptyset, D, \texttt{hh}, \texttt{tt}, \{\texttt{ht}, \texttt{th}\}, \{\texttt{hh}, \texttt{tt}\}, \{\texttt{hh}, \texttt{ht}, \texttt{th}\}, \{\texttt{tt}, \texttt{ht}, \texttt{tt}\}\}$$

We assume that $P(\texttt{hh}) = p^2$; a simple computation using the idea of independence yields that $P(\{\texttt{ht}, \texttt{th}\}) = 2p(1-p)$ and $P(\texttt{tt}) = (1-p)^2$. We can generalise this result, to obtain

$$P(k \text{ heads and } n-k \text{ tails in } n \text{ flips}) = \left(\begin{array}{c} k \\ n \end{array}\right) p^k (1-p)^{n-k}$$

Example 7.10: *The probability of various frequencies in repeated coin flips*

(which is what we obtained for the probability of a sequence of trials showing k heads and $n-k$ tails in example 10) has a substantial peak at $p = \frac{k}{n}$. This peak gets very narrow and extremely pronounced as $n \to \infty$. This effect is extremely important, and is consistent with an interpretation of probability as relative frequency:

- firstly, because it means that we assign a high probability to long sequences of coin flips where the event occurs with the "right" frequency
- and secondly, because the probability assigned to these long sequences can also be interpreted as a frequency — essentially, this interpretation means that long sequences where the events occur with the "right" frequency occur far more often than other such sequences (see figure 7.1).

All this means that, if we choose a P function for a coin flip — or some other experiment — on the basis of *sufficiently good* frequency data, then we are very unlikely to see long sequences of coin flips — or repetitions of the experiment — that do not show this frequency.

This interpretation of probability as frequency is widespread, and common. One valuable advantage of the interpretation is that it simplifies estimating probabilities for some sorts of models. For example, given a coin, one could obtain $P(\texttt{heads})$ by flipping the coin many times and measuring the relative frequency with which heads appear.

Subjective probability

It is not always possible to use frequencies to obtain probabilities. There are circumstances in which we would like to account for uncertainty but cannot meaningfully speak about frequencies. For example, it is easy to talk about the probability it will rain tomorrow, but hard to interpret this use of the term as a statement about

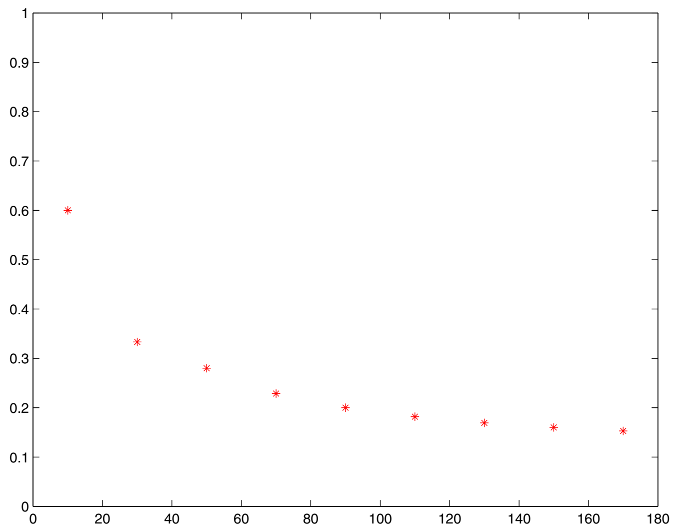

Figure 7.1. We assume that a single flip of a coin has a probability 0.5 of coming up heads. If we interpret probability as frequency, then long sequences of coin flips should almost always have heads appearing about half the time. This plot shows the width of the interval about 0.5 that contains 95% of the probability for various numbers of repeated coin flips. Notice that as the sequence gets longer, the interval gets narrower — one is very likely to observe a frequency of heads in the range $[0.43, 0.57]$ for 170 flips of a coin of this kind.

frequency[1]. An alternative source of P is to regard probability as encoding *degree of belief*. In this approach, which is usually known as **subjective probability**, one chooses P to reflect reasonable beliefs about the situation that applies.

Subjective probability must still satisfy the axioms of probability. It is simply a way of choosing free parameters in a probability model without reference to frequency. The attractive feature of subjective probability is that it emphasizes that a choice of probability model is a *modelling* exercise — there are few circumstances where the choice is canonical. One natural technique to adopt is to choose a function P that yields good behaviour in practice, an approach known as **learning** and discussed in chapter **??**.

[1] One dodge is to assume that there are a very large set of equivalent universes which are the same today. In some of these worlds, it rains tomorrow and in others it doesn't; the frequency with which it rains tomorrow is the probability. This philosophical fiddle isn't very helpful in practice, because we can't actually measure that frequency by looking at these alternative worlds.

A friend with a good reputation for probity and no obvious need for money draws a coin from a pocket, and offers to bet with you on whether it comes up heads or tails — your choice of face. What probability do you ascribe to the event that it comes up heads?
Now an acquaintance draws a coin from a pocket and offers a bet: he'll pay you 15 dollars for your stake of one dollar if the coin comes up heads. What probability do you ascribe to the event that it comes up heads?
Finally you encounter someone in a bar who (it emerges) has a long history of disreputable behaviour and an impressive conviction record. This person produces a coin from a pocket and offers a bet: you pay him 1000 dollars for his stake of one dollar if it lands on its edge and stands there. What probability do you ascribe to the event that it lands on its edge and stands there?
You have to choose your answer for these cases — that's why it's subjective — but you could lose a lot of money learning that the answer in the second case is going to be pretty close to zero and in the third case is pretty close to one.

Example 7.11: *Assigning P functions to two coins from two different sources, using subjective probability.*

7.2 Probability in Continuous Spaces

Much of the discussion above transfers quite easily to a continuous space, as long as we are careful about events. The difficulty is caused by the "size" of continuous spaces — there are an awful lot of numbers between 1.0 and 1.00000001, one for each number between 1.0 and 2.0. For example, if we are observing noise — perhaps by measuring the voltage across the terminals of a warm resistor — the noise will very seldom take the value 1 exactly. It is much more helpful to consider the probability that the value is in the range 1 to $1 + \delta$, for δ a small step.

7.2.1 Event Structures for Continuous Spaces

This observation justifies using events that look like intervals or boxes for continuous spaces. Given a space D, our space of events will be a set $\mathcal{F}$ with the following properties:

- The empty set is in $\mathcal{F}$ and so is D.
- *Closure under finite intersections:* if S_i is a *finite collection* of subsets, and each $S_i \in \mathcal{F}$ then $\cap_i S_i \in \mathcal{F}$.
- *Closure under finite unions:* if S_i is an *finite collection* of subsets, and each $S_i \in \mathcal{F}$ then $\cup_i S_i \in \mathcal{F}$.
- *Closure under complements:* if $S_1 \in \mathcal{F}$ then $\overline{S_1} = D - S_1 \in \mathcal{F}$.

The basic axioms for P apply here too. For D the domain, and A and B events, we have:

- $P(D) = 1$
- $P(\emptyset) = 0$
- for any A, $0 \leq P(A) \leq 1$
- if $A \subset B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The concepts of conditional probability, independence and conditional independence apply in continuous spaces without modification. For example, the conditional probability of an event given another event can be defined by

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

and the conditional probability can be thought of as probability restricted to the set B. Events A and B are independent if and only if

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

and A and B are conditionally independent given C if and only if

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

The main difficulty is expressing the function P in a useful way — it is clearly no longer possible to write down the space of events and give a value of P for each event. We will deal only with R^n, with subsets of this space, or with multiple copies of this space.

7.2.2 Representing a P-function for the Real Line

The set of events for the real line is far too big to write down. All events look like unions of a basic collection of sets. This basic collection consists of:

- individual points (i.e a);
- open intervals (i.e. (a, b));
- half-open intervals (i.e. $(a, b]$ or $[a, b)$);
- and closed intervals (i.e. $[a, b]$).

All of these could extend to infinity. The function P can be represented by a function F with the following properties:

- $F(-\infty) = 0$

- $F(\infty) = 1$
- $F(x)$ is monotonically increasing.

and we interpret $F(x)$ as $P((-\infty, x])$. The function F is referred to as the **cumulative distribution function**.

The value of P for all the basic sets described can be extracted from F, with appropriate attention to limits; for example, $P((a, b]) = F(b) - F(a)$ and $P(a) = \lim_{\epsilon \leftarrow 0^+}(F(a+\epsilon) - F(a))$. Notice that if F is continuous, $P(a) = 0$.

7.2.3 Probability Densities

In R^n, events are unions of elements of a basic collection of sets, too. This basic collection consists of a product of n elements from the basic collection for the real line. A cumulative distribution function can be defined in this case, too. It is given by a function F with the property that $P(\{x_1 \leq u_1, x_2 \leq u_2, \ldots x_n \leq u_n\}) = F(\boldsymbol{u})$. This function is constrained by other properties, too. However, cumulative distribution functions are a somewhat unwieldy way to specify probability.

For the examples we will deal with in continuous spaces, the usual way to specify P is to provide a function p such that

$$P(\text{event}) = \int_{\text{event}} p(u)du$$

This function is referred to as a **probability density function**. Not every probability model admits a density function, but all our cases will. Note that a density function cannot have a negative value, but that its value could be larger than one. In all cases, probability density functions integrate to one, i.e.

$$P(D) = \int_D p(u)du = 1$$

and any non-negative function with this property is a probability density function. The value of the probability density function at a point represents the probability of the event that consists of an infinitesimal neighbourhood at that value, i.e.:

$$p(u_1)du = P(\{u \in [u_1, u_1 + du]\})$$

Conditional probability, independence and conditional independence are ideas that can be translated into properties of probability density functions. In their most useful form, they are properties of random variables.

7.3 Random Variables

Assume that we have a probability model on either a discrete or a continuous domain, $\{D, \mathcal{F}, P\}$. Now let us consider a function of the outcome of an experiment. The values that this function takes on the different elements of D form a new set, which we shall call D'. There is a structure, with the same formal properties as $\mathcal{F}$ on D' defined by the values that this function takes on different elements of $\mathcal{F}$ — call this structure $\mathcal{F}'$.

This function is known as a **random variable**. We can talk about the probability that a random variable takes a particular set of values, because the probability structure carries over. In particular, assume that we have a random variable ξ. If $A' \in \mathcal{F}'$, there is some $A \in \mathcal{F}$ such that $A' = \xi(A)$. This means that

$$P(\{\xi \in A'\}) = P(A)$$

The simplest random variable is given by the identity function — this means that D' is the same as D, and $\mathcal{F}'$ is the same as $\mathcal{F}$. For example, the outcome of a coin flip is a random variable.
Now gamble on the outcome of a coin flip: if it comes up heads, you get a dollar, and if it comes up tails, you pay a dollar. Your income from this gamble is a random variable. In particular, $D' = \{1, -1\}$ and $\mathcal{F}' = \{\emptyset, D', \{1\}, \{-1\}\}$.
Now gamble on the outcome of two coin flips: if both coins come up the same, you get a dollar, and if they come up different, you pay a dollar. Your income from this gamble is a random variable. Again, $D' = \{1, -1\}$ and $\mathcal{F}' = \{\emptyset, D', \{1\}, \{-1\}\}$. In this case, D' is not the same as D and $\mathcal{F}'$ is not the same as $\mathcal{F}$; however, we can still speak about the probability of getting a dollar — which is the same as $P(\{hh, tt\})$.

Example 7.12: *The payoff on a gamble is a random variable.*

Density functions are very useful for specifying the probability model for the value of a random variable. However, they do result in quite curious notations (probability is a topic that seems to encourage creative use of notation). It is common to write the density function for a random variable as p. Thus, the distribution for λ would be written as $p(\lambda)$ — in this case, the name of the *variable* tells you what function is being referred to, rather than the name of the function, which is always p. Some authors resist this convention, but its use is pretty much universal in the vision literature, which is why we adopt it. For similar reasons, we write the probability function for a set of events as P, so that the probability of an event $P(\text{event})$ (despite the fact that different sets of events may have very different probability functions).

7.3.1 Conditional Probability and Independence

Conditional probability is a very useful idea for random variables. Assume we have two random variables, m and n — (for example, the value I read from my rain gauge as m and the value I read on the neighbour's as n). Generally, the probability density function is a function of both variables, $p(m, n)$. Now

$$p(m_1, n_1)dmdn \;=\; P(\{m \in [m_1, m_1 + dm]\} \text{ and } \{n \in [n_1, n_1 + dm]\})$$

$$= P(\{m \in [m_1, m_1 + dm]\} \mid \{n \in [n_1, n_1 + dm]\})P(\{n \in [n_1, n_1 + dm]\})$$

We can define a conditional probability density from this by

$$\begin{aligned} p(m_1, n_1)dmdn &= P(\{m \in [m_1, m_1 + dm]\} \mid \{n \in [n_1, n_1 + dm]\})P(\{n \in [n_1, n_1 + dm]\}) \\ &= (p(m_1|n_1)dm)(p(n_1)dn) \end{aligned}$$

Note that this conditional probability density has the expected property, that

$$p(m|n) = \frac{p(m,n)}{p(n)}$$

Independence and conditional independence carry over to random variables and probability densities without fuss.

We now consider the probability that each of two coins comes up heads, yielding two random variables — the relevant probabilities — which we shall write as p_1 and p_2. Now the density function for these random variables is $p(p_1, p_2)$. There is very little reason to believe that there is any dependency between these coins, so we should be able to write $p(p_1, p_2) = p(p_1)p(p_2)$. Notice that the notation is particularly confusing here; the intended meaning is that $p(p_1, p_2)$ factors, but that the factors are not necessarily equal. *In this case*, a further reasonable modelling step is to assume that $p(p_1)$ is the same function as $p(p_2)$

Example 7.13: *Independence in random variables associated with two coins.*

7.3.2 Expectations

The **expected value** or **expectation** of a random variable (or of some function of the random variable) is obtained by multiplying each value by its probability and summing the results — or, in the case of a continuous random variable, by multiplying by the probability density function and integrating. The operation is known as **taking an expectation**. For a discrete random variable, x, taking the expectation of x yields:

$$\mathrm{E}[x] = \sum_{i \in \text{values}} x_i p(x_i)$$

For a continuous random variable, the process yields

$$\mathrm{E}[x] = \int_D x p(x) dx$$

often referred to as the average, or the mean in polite circles. One model for an expectation is to consider the random variable as a payoff, and regard the expectation as the average reward, per bet, for an infinite number of repeated bets. The expectation of a general function $g(x)$ of a random variable x is written as $\mathrm{E}[g(x)]$.

The **variance** of a random variable x is

$$\mathrm{var}(x) = \mathrm{E}[x^2 - (\mathrm{E}(x))^2]$$

This expectation measures the average deviance from the mean. The variance of a random variable gives quite a strong indication of how common it is to see a value that is significantly different from the mean value. In particular, we have the following useful result:

$$P(\{\mid x - \mathrm{E}[x] \mid \geq \epsilon\}) \leq \frac{\mathrm{var}(x)}{\epsilon^2} \tag{7.3.1}$$

> You and an acquaintance decide to bet on the outcome of a coin flip. You will receive a dollar from your acquaintance if the coin comes up heads, and pay one if it comes up tails. The coin is symmetric.
> This means the expected value of the payoff is
>
> $$1P(\texttt{heads}) - 1P(\texttt{tails}) = 0$$
>
> The variance of the payoff is one, as is the standard deviation.
> Now consider the probability of obtaining 10 dollars in 10 coin flips, with a fair coin. Our random variable x is the income in 10 coin flips. Equation 7.3.1 yields $P(\{\mid x \mid \geq 10\}) \leq \frac{1}{100}$, which is a generous upper bound — the actual probability is of the order of one in a thousand.

Example 7.14: *The expected value of gambling on a coin flip.*

The **standard deviation** is obtained from the variance:

$$\mathrm{sd}(x) = \sqrt{\mathrm{var}(x)} = \sqrt{\mathrm{E}[x^2 - (\mathrm{E}[x])^2]}$$

For a vector of random variables, the **covariance** is

$$\mathrm{cov}(\boldsymbol{x}) = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^t - (\mathrm{E}[x]\mathrm{E}[x]^t)]$$

This matrix (look carefully at the transpose) is symmetric. Diagonal entries are the variance of components of $\boldsymbol{x}$, and must be non-negative. Off-diagonal elements measure the extent to which two variables co-vary. For independent variables, the covariance must be zero. For two random variables that generally have different signs, the covariance can be negative.

Expectations of functions of random variables are extremely useful. The notation for expectations can be a bit confusing, because it is common to omit the density with respect to which the expectation is being taken, which is usually obvious from the context. For example, $\mathrm{E}[x^2]$ is interpreted as

$$\int_D x^2 p(x) dx$$

7.3.3 Joint Distributions and Marginalization

Assume we have a model describing the behaviour of a collection of random variables. We will proceed on the assumption that they are discrete, but (as should be clear by now) the discussion will work for continuous variables if summing is replaced by integration. One way to specify this model is to give the probability distribution for all variables, known in jargon as the **joint probability distribution function** — for concreteness, write this as $P(x_1, x_2, \ldots x_n)$. If the probability distribution is represented by its density function, the density function is usually referred to as the **joint probability density function**. Both terms are often abbreviated as "joint."

As we have already seen, the value of P for some elements of the event space can be determined from the value of P for other elements. This means that if we know

$$P(\{x_1 = a, x_2 = b, \ldots x_n = n\})$$

for each possible value of $a, b, \ldots, n$, then we should know P for a variety of other events. For example, it might be useful to know $P(\{x_1 = a\})$.

It should be obvious that the event structure, while useful, is getting unwieldy as a notation. It is quite common to use a rather sketchy notation to indicate the appropriate event. For example 15, we would write

$$P(\{(\texttt{heads}, I), (\texttt{heads}, II)\}) = P(\texttt{heads})$$

for example. In this notation, the argument of example 15 leads to:

$$P(x_2, \ldots x_n) = \sum_{\text{values of } x_1} P(x_1, x_2, \ldots x_n)$$

This operation is referred to as **marginalisation**.

A similar argument applies to probability density functions, but the operation is now integration. Given a probability density function $p(x_1, x_2, \ldots, x_n)$, we obtain

$$p(x_2, \ldots x_n) = \int_D p(x_1, x_2, \ldots x_n) dx_1$$

7.4 Standard Distributions and Densities

There are a variety of standard distributions that arise regularly in practice. References such as [] give large numbers; we will discuss only the most important cases.

The **uniform distribution** has the same value at each point on the domain. This distribution is often used to express an unwillingness to make a choice or a lack of information. On a continuous space, the uniform distribution has a density function that has the same value at each point. Notice that a uniform density on an infinite continuous domain isn't meaningful, because it could not be scaled to integrate to one.

Let us assume we have a coin which could be from one of two types; the first type of coin is evenly balanced; the other is wildly unbalanced. We flip our coin some number of times, observe the results, and should like to know what type of coin we have. Assume that we flip the coin once. The set of outcomes is

$$D = \{(\texttt{heads}, I), (\texttt{heads}, II), (\texttt{tails}, I), (\texttt{tails}, II)\}$$

An appropriate event space is:

$$\left\{ \begin{array}{cc} \emptyset, & D, \\ \{(\texttt{heads}, I)\}, & \{(\texttt{heads}, II)\}, \\ \{(\texttt{tails}, I)\}, & \{(\texttt{tails}, II)\}, \\ \{(\texttt{heads}, I), (\texttt{heads}, II)\}, & \{(\texttt{tails}, I), (\texttt{tails}, II), \}, \\ \{(\texttt{tails}, I), (\texttt{heads}, I)\}, & \{(\texttt{tails}, II), (\texttt{heads}, II)\}, \\ \{(\texttt{heads}, II), (\texttt{tails}, I), (\texttt{tails}, II)\}, & \{(\texttt{heads}, I), (\texttt{tails}, I), (\texttt{tails}, II)\} \\ \{(\texttt{heads}, I), (\texttt{heads}, II), (\texttt{tails}, II)\} & \{(\texttt{heads}, I), (\texttt{heads}, II), (\texttt{tails}, I)\} \end{array} \right\}$$

In this case, assume that we know $P(\text{face}, \text{type})$, for each face and type. Now, for example, the event that the coin shows heads (whatever the type) is represented by the set

$$\{(\texttt{heads}, I), (\texttt{heads}, II)\}$$

We can compute the probability that the coin shows heads (whatever the type) as follows

$$\begin{aligned} P(\{(\texttt{heads}, I), (\texttt{heads}, II)\}) &= P((\texttt{heads}, I) \cup (\texttt{heads}, II)) \\ &= P((\texttt{heads}, I)) + P((\texttt{heads}, II)) \end{aligned}$$

We can compute the probability that the coin is of type I, etc. with similar ease using the same line of reasoning, which applies quite generally.

Example 7.15: *Marginalising out parameters for two different types of coin.*

The **binomial distribution** applies to situations where one has independent identically distributed samples from a distribution with two values. For example, consider drawing n balls from an urn containing equal numbers of black and white balls. Each time a ball is drawn, its colour is recorded and it is replaced, so that the probability of getting a white ball — which we denote p — is the same for each draw. The binomial distribution gives the probability of getting k white balls

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

The mean of this distribution is np and the variance is $np(1-p)$.

The **Poisson distribution** applies to spatial models that have uniformity properties. Assume that points are placed on the real line randomly in such a way that the expected number of points in an interval is proportional to the length of the interval. The number of points in a unit interval will have a Poisson distribution where

$$P(\{N = x\}) = \frac{\lambda^x e^{-x}}{x!}$$

(where $x = 0, 1, 2 \ldots$ and $\lambda > 0$ is the constant of proportionality). The mean of this distribution is λ and the variance is λ

7.4.1 The Normal Distribution

The probability density function for the **normal distribution** for a single random variable x is

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

The mean of this distribution is μ and the standard deviation is σ. This distribution is widely called a **Gaussian distribution** in the vision community.

The **multivariate normal distribution** for d-dimensional vectors $\boldsymbol{x}$ has probability density function

$$p(\boldsymbol{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\Sigma)} \exp - \left\{ \frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2} \right\}$$

The mean of this distribution is $\boldsymbol{\mu}$ and the covariance is Σ. Again, this distribution is widely called a **Gaussian distribution** in the vision community.

The normal distribution is extremely important in practice, for several reasons:

- The sum of a large number of random variables is normally distributed, pretty much whatever the distribution of the individual random variables. This fact is known as the **central limit theorem**. It is often cited as a reason to model a collection of random effects with a single normal model.
- Many computations that are prohibitively hard for any other case are easy for the normal distribution.
- In practice, the normal distribution appears to give a fair model of some kinds of noise.
- Many probability density functions have a single peak and then die off; a model for such distributions can be obtained by taking a Taylor series of the log of the density at the peak. The resulting model is a normal distribution (which is often quite a good model).

7.5 Probabilistic Inference

Very often, we have a sequence of observations produced by some process whose mechanics we understand, but which has some underlying parameters that we do not know. The problem is to make useful statements about these parameters. For example, we might observe the intensities in an image, which are produced by the interaction of light and surfaces by principles we understand; what we don't know — and would like to know — are such matters as the shape of the surface, the reflectance of the surface, the intensity of the illuminant, etc. Obtaining some representation of the parameters from the data set is known as **inference**. There is no canonical inference scheme; instead, we need to choose some principle that identifies the most desirable set of parameters.

7.5.1 The Maximum Likelihood Principle

A general inference strategy known as **maximum likelihood estimation**, can be described as

> *Choose the world parameters that maximise the probability of the measurement observed*

In the general case, we are choosing

$$\arg\max P(\text{measurements}|\text{parameters})$$

(where the maximum is only over the world parameters because the measurements are known, and arg max means "the argument that maximises"). In many problems, it is quite easy to specify the measurements that will result from a particular setting of model parameters — this means that $P(\text{measurements}|\text{parameters})$, often referred to as the **likelihood**, is easy to obtain. This can make maximum likelihood estimation attractive.

We return to example 15. Now assume that we know some conditional probabilities. In particular, the unbiased coin has $P(\texttt{heads}|I) = P(\texttt{tails}|I) = 0.5$, and the biased coin has $P(\texttt{tails}|II) = 0.2$ and $P(\texttt{heads}|II) = 0.8$.
We observe a series of flips of a single coin, and wish to know what type of coin we are dealing with. One strategy for choosing the type of coin represented by our evidence is to choose either I or II, depending on whether $P(\text{side}|I) > P(\text{side}|II)$. For example, if we observe four heads and one tail in sequence, then $P(\text{hhhht}|II) = (0.8)^4 0.2 = 0.08192$ and $P(\text{hhhht}|I) = 0.03125$, and we choose type II.

Example 7.16: *Maximum likelihood inference on the type of a coin from its behaviour.*

Maximum likelihood is often an attractive strategy, because it can admit quite simple computation. A classical application of maximum likelihood estimation involves estimating the parameters of a normal distribution from a set of samples of that distribution (example 17).

Assume that we have a set of n samples — the i'th of which is x_i — that are known to be independent and to have been drawn from the same normal distribution. The likelihood of our sample is

$$L(x_1, \ldots x_n; \mu, \sigma) = \prod_i p(x_i; \mu, \sigma) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Working with the log of the likelihood will remove the exponential, and not change the position of the maximum. For the log-likelihood, we have

$$Q(x_1, \ldots x_n; \mu, \sigma) = -\sum_i \frac{(x_i - \mu)^2}{2\sigma^2} - n(\frac{1}{2}\log 2 + \frac{1}{2}\log \pi + \log \sigma)$$

and we want the maximum with respect to μ and σ. This must occur when the derivatives are zero, so we have

$$\frac{\partial Q}{\partial \mu} = 2\sum_i \frac{(x_i - \mu)}{2\sigma^2} = 0$$

and a little shuffling of expressions shows that this maximum occurs at

$$\mu = \frac{\sum_i x_i}{n}$$

Similarly

$$\frac{\partial Q}{\partial \sigma} = \frac{\sum_i (x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

and this maximum occurs at

$$\sigma = \frac{\sqrt{\sum_i (x_i - \mu)^2}}{n}$$

Note that this estimate of σ is biased, in that its expected value is

$$\sigma(n/(n-1))$$

and it is more usual to use

$$\frac{\sqrt{\sum_i (x_i - \mu)^2}}{(n-1)}$$

as an estimate.

Example 7.17: *Estimating the parameters of a normal distribution from a series of independent samples from that distribution.*

7.5.2 Priors, Posteriors and Bayes' rule

In example 16, our maximum likelihood estimate incorporates no information about $P(I)$ or $P(II)$ — which can be interpreted as how often coins of type I or type II are handed out, or as our subjective degree of belief that we have a coin of type I or of type II before we flipped the coin. This is unfortunate, to say the least; for example, if coins of type II are rare, we would want to see an awful lot of heads before it would make sense to infer that our coin is of this type. Some quite simple algebra suggests a solution.

Recall that $P(A, B) = P(A|B)P(B)$. This simple observation gives rise to an innocuous looking identity for reversing the order in a conditional probability:

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

This is widely referred to as **Bayes' theorem** or **Bayes' rule**.

Now the interesting property of Bayes' rule is that it tells us which choice of parameters is most probable, given our model and our prior beliefs. Rewriting Bayes' rule gives

$$P(\text{parameters}|\text{measurements}) = \frac{P(\text{measurements}|\text{parameters})P(\text{parameters})}{P(\text{measurements})}$$

The term $P(\text{parameters})$ is referred to as the **prior** (presumably because it describes our knowledge of the world *before* measurements have been taken). The term $P(\text{parameters}|\text{data})$ is usually referred to as the **posterior** (presumably because it describes the probability of various models *after* measurements have been taken). $P(\text{data})$ can be computed by marginalisation (which requires computing a high dimensional integral, often a nasty business) or for some problems can be ignored. As we shall see in following sections, attempting to use Bayes' rule can result in difficult computations — that integral being one — because posterior distributions often take quite unwieldy forms.

7.5.3 Bayesian Inference

The Bayesian philosophy is that

> all information about the world is captured by the posterior.

The first reason to accept this view is that the posterior is a principled combination of prior information about the world and a model of the process by which measurements are generated — i.e. there is no information missing from the posterior, and the information that is there, is combined in a proper manner. The second reason is that the approach appears to produce very good results. The great difficulty is that computing with posteriors can be very difficult — we will discuss mechanisms for computing with posteriors in section **??**.

For example, we could use the study of physics in the last few chapters to get expressions relating pixel values to the position and intensity of light sources, the reflectance and orientation of surfaces, etc. Similarly, we are likely to have some beliefs about the parameters that have nothing to do with the particular values of the measurements that we observe. We know that albedos are never outside the range $[0, 1]$; we expect that illuminants with extremely high exitance are uncommon; and we expect that no particular surface orientation is more common than any other. This means that we can usually cobble up a reasonable choice of $P(\text{parameters})$. This expression is usually referred to as a **prior**, because it represents beliefs about the state of the world before measurements were made.

MAP Inference

An alternative to maximum likelihood inference is to infer a state of the world that maximises the posterior:

> *Choose the world parameters that maximise the conditional probability of the parameters, conditioned on the measurements taking the observed values*

This approach is known as **maximum *a posteriori*** (or MAP) reasoning.

Assume that we have three flips of the coin, and would like to determine whether it has type I or type II. We know that the mint has 3 machines that produce type I coins and 1 machine that produces type II coins, and there is no reason to believe that these machines run at different rates. We therefore assign $P(I) = 0.75$ and $P(II) = 0.25$. Now we observe three heads, in three consecutive flips. The value of the posterior for type I is:

$$\begin{aligned} P(I|\text{hhh}) &= \frac{P(\text{hhh}|I)P(I)}{P(\text{hhh})} \\ &= \frac{P(\text{h}|I)^3 P(I)}{P(\text{hhh}, I) + P(\text{hhh}, II)} \\ &= \frac{P(\text{h}|I)^3 P(I)}{P(\text{hhh}|I)P(I) + P(\text{hhh}|II)P(II)} \\ &= \frac{0.5^3 0.75}{0.5^3 0.75 + 0.8^3 0.25} \\ &= 0.422773 \end{aligned}$$

By a similar argument, the value of the posterior for type II is 0.577227. An MAP inference procedure would conclude the coin is of type II.

Example 7.18: *Determining the type of a coin using MAP inference.*

The denominator in the expression for the posterior can be quite difficult to compute, because it requires a sum over what is potentially a very large number of elements (imagine what would happen if there were many different types of coin). However, knowing this term is not crucial if we wish to isolate the element with the maximum value of the posterior, because it is a constant. Of course, if there are a very large number of events in the discrete space, finding the world parameters that maximise the posterior can be quite tricky.

The Posterior as an Inference

Assume we have a coin which comes from a mint which has a continuous control parameter, λ, which lies in the range $[0, 1]$. This parameter gives the probability that the coin comes up heads, so $P(\texttt{heads}|\lambda) = \lambda$. We know no reason to prefer any one value of λ to any other, so as a prior probability distribution for λ we use the uniform distribution so $p(\lambda) = 1$.
Assume we flip the coin twice, and observe heads twice; what do we know about λ? All our knowledge is captured by the posterior, which is

$$\frac{P(\lambda \in [x, x + dx]|\texttt{hh})}{dx}$$

we shall write this expression as $p(\lambda|\texttt{hh})$. We have

$$\begin{aligned} p(\lambda|\texttt{hh}) &= \frac{p(\texttt{hh}|\lambda)p(\lambda)}{p(\texttt{hh})} \\ &= \frac{p(\texttt{hh}|\lambda)p(\lambda)}{\int_0^1 p(\texttt{hh}|\lambda)p(\lambda)d\lambda} \\ &= \frac{\lambda^2 p(\lambda)}{\int_0^1 p(\texttt{hh}|\lambda)p(\lambda)d\lambda} \\ &= 3\lambda^2 \end{aligned}$$

It is fairly easy to see that if we flip the coin n times, and observe k heads and $n - k$ tails, we have

$$p(\lambda|k \text{ heads and } n - k \text{ tails}) \propto c\lambda^k(1 - \lambda)^{n-k}$$

Example 7.19: *Determining the probability a coin comes up heads from the outcome of a sequence of flips.*

We have argued that choosing parameters that maximise the posterior is a useful inference mechanism. But, as figure 7.2 indicates, the posterior is good for other uses as well. This figure plots the posterior distribution on the probability that a coin comes up heads, given the result of some number of flips. In the figure, the posterior distributions indicate not only the single "best" value for the probability that a coin comes up heads, but also the extent of the uncertainty in that value. For example, inferring a value of this probability after two coin flips leads to a value that is not particularly reliable — the posterior is a rather flat function, and there are many different values of the probability with about the same value of the posterior. Possessing this information allows us to compare this evidence with other sources of evidence about the coin.

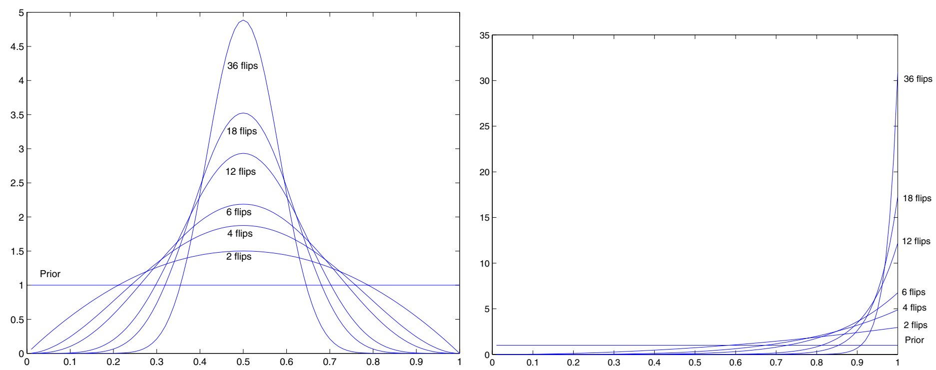

Figure 7.2. On the left, the value of the posterior density for the probability that a coin will come up heads, given an equal number of heads and tails are observed. This posterior is shown for different numbers of observations. With no evidence, the posterior is the prior; but as the quantity of evidence builds up, the posterior becomes strongly peaked — this is because one is very unlikely to observe a long sequence of coin flips where the frequency of heads is very different from the probability of obtaining a head. On the right, a similar plot, but now for the case where every flip comes up heads. As the number of flips builds up, the posterior starts to become strongly peaked near one. This overwhelming of the prior by evidence is a common phenomenon in Bayesian inference.

Bayesian inference is a framework within which it is particularly easy to combine various types of evidence, both discrete and continuous. It is often quite easy to set up the sums.

Bayesian Model Selection

The crucial virtue of Bayesian inference is the accounting for uncertainty shown in examples 20 and 21. We have been able to account for an occasionally untruthful informant and a random measurement; when there was relatively little contradictory evidence from the coin's behaviour, our process placed substantial weight on the informant's testimony, but when the coin disagreed, the informant was discounted. This behaviour is highly attractive, because we are able to combine uncertain sources of information with confidence.

Example 22 shows how to tell whether the informant of examples 20 and 21 is

We use the basic setup of example 19. Assume you have a contact at the coin factory, who will provide a single estimate of λ. Your contact has poor discrimination, and can tell you only whether λ is `low`, `medium` or `high` (i.e in the range $[0, 1/3]$, $(1/3, 2/3)$ or $[2/3, 1]$). You expect that a quarter of the time your contact, not being habitually truthful, will simply guess rather than checking how the coin machine is set. What do you know about λ after a single coin flip, which comes up heads, if your contact says `high`? We need

$$\begin{aligned} p(\lambda|\texttt{high}, \texttt{heads}) &= \frac{p(\texttt{high}, \texttt{heads}|\lambda)p(\lambda)}{p(\texttt{high}, \texttt{heads})} \\ &\propto p(\texttt{high}, \texttt{heads}|\lambda)p(\lambda) \end{aligned}$$

The interesting modelling problem is in $p(\texttt{high}, \texttt{heads}|\lambda)$. This is

$$\begin{aligned} p(\texttt{high}, \texttt{heads}|\lambda) &= p(\texttt{high}, \texttt{heads}|\lambda, \text{truth} = 1)p(\text{truth} = 1) \\ &\quad + p(\texttt{high}, \texttt{heads}|\lambda, \text{truth} = 0)p(\text{truth} = 0) \\ &= p(\texttt{high}, \texttt{heads}|\lambda, \text{truth} = 1)p(\text{truth} = 1) \\ &\quad + p(\texttt{heads}|\lambda, \text{truth} = 0)p(\texttt{high}|\lambda, \text{truth} = 0)p(\text{truth} = 0) \end{aligned}$$

Now from the details of the problem

$$\begin{aligned} p(\text{truth} = 1) &= 0.75 \\ p(\text{truth} = 0) &= 0.25 \\ p(\texttt{heads}|\lambda, \text{truth} = 0) &= \lambda \\ p(\texttt{high}|\lambda, \text{truth} = 0) &= \frac{1}{3} \end{aligned}$$

and the term to worry about is $p(\texttt{high}, \texttt{heads}|\lambda, \text{truth} = 1)$. This term reflects the behaviour of the coin and the informant when the informant is telling the truth; in particular, this term must be zero for $\lambda \in [0, 2/3)$, because in this case λ is not `high`, so we never see a truthful report of `high` with λ in this range. For λ in the `high` range, this term must be λ, because now it is the probability of getting a head with a single flip. Performing the computation, we obtain the posterior graphed in figure 7.3.

Example 7.20: *Determining the type of a coin from a sequence of flips, incorporating information from an occasionally untruthful informant.*

telling the truth or not, given the observations. A useful way to think about this example is to regard it as comparing two *models* (as opposed to the value of a binary parameter within one model). One model has a lying informant, and the other has a truthful informant. The posteriors computed in this example compare how well

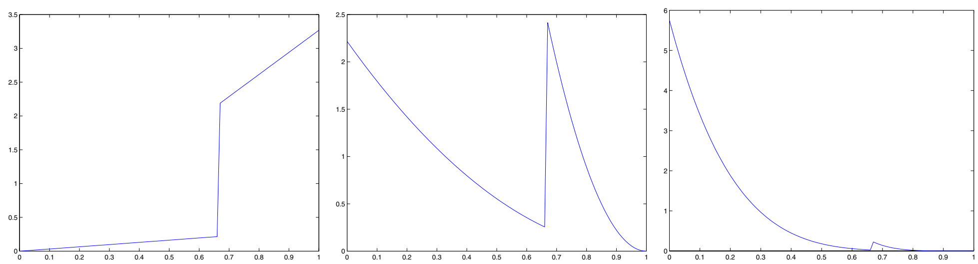

Figure 7.3. On the left, the posterior probability density for the probability a coin comes up heads, given a single flip that shows a head and a somewhat untruthful informant who says `high`, as in example 20. In the center, a posterior probability density for the same problem, but now assuming that we have seen two tails and the informant says `high` (a sketch of the formulation appears in example 21). On the right, a posterior probability density for the case when the coin shows five tails and the informant says `high`. As the number of tails builds up, the weight of the posterior in the `high` region goes down, strongly suggesting the informant is lying.

Now consider what happens in example 20 if the contact says high and we see two tails. We need

$$\begin{aligned} p(\lambda|\texttt{high}, \texttt{tt}) &= \frac{p(\texttt{high}, \texttt{tt}|\lambda)p(\lambda)}{p(\texttt{high}, \texttt{tt})} \\ &\propto p(\texttt{high}, \texttt{tt}|\lambda)p(\lambda) \end{aligned}$$

Now $p(\texttt{high}, \texttt{tt}|\lambda)$ is

$$\begin{aligned} p(\texttt{high}, \texttt{tt}|\lambda) &= p(\texttt{high}, \texttt{tt}|\lambda, \text{truth} = 1)P(\text{truth} = 1) \\ &\quad + p(\texttt{high}, \texttt{tt}|\lambda, \text{truth} = 0)P(\text{truth} = 0) \\ &= p(\texttt{high}, \texttt{tt}|\lambda, \text{truth} = 1)P(\text{truth} = 1) \\ &\quad + p(\texttt{tt}|\lambda, \text{truth} = 0)p(\texttt{high}|\lambda, \text{truth} = 0)P(\text{truth} = 0) \end{aligned}$$

Now $p(\texttt{tt}|\lambda, \text{truth} = 0) = (1 - \lambda)^2$ and the interesting term is $p(\texttt{high}, \texttt{tt}|\lambda, \text{truth} = 1)$. Again, this term reflects the behaviour of the coin and the informant when the informant is telling the truth; in particular, this term must be zero for $\lambda \in [0, 2/3)$, because in this case λ is not high. For λ in the high range, this term must be $(1 - \lambda)^2$, because now it is the probability of getting two tails with two flips. Performing the computation, we obtain the posterior graphed in figure 7.3.

Example 7.21: *Determining the type of a coin from a sequence of flips, incorporating information from an occasionally untruthful informant — II.*

We now need to know whether our informant lied to us. Assume we see a single head and an informant saying `high`, again. The relevant posterior is:

$$\begin{aligned}
P(\text{truth=0}|\texttt{head},\texttt{high}) &= \frac{P(\texttt{head},\texttt{high}|\text{truth=0})P(\text{truth=0})}{P(\texttt{head},\texttt{high})} \\
&= \frac{\int P(\lambda,\texttt{head},\texttt{high}|\text{truth=0})P(\text{truth=0})d\lambda}{P(\texttt{head},\texttt{high})} \\
&= \frac{\int P(\texttt{head},\texttt{high}|\lambda,\text{truth=0})P(\lambda)P(\text{truth=0})d\lambda}{P(\texttt{head},\texttt{high})} \\
&= \frac{1}{1+\frac{\int P(\texttt{head},\texttt{high}|\lambda,\text{truth=1})P(\lambda)d\lambda P(\text{truth=1})}{\int P(\texttt{head},\texttt{high}|\lambda,\text{truth=0})P(\lambda)d\lambda P(\text{truth=0})}}
\end{aligned}$$

Example 7.22: *Is the informant lying?*

different models explain a given data set, given a prior on the *models*. This is a very general problem — usually called **model selection** — with a wide variety of applications in vision:

- **Recognition:** Assume we have a region in an image, and an hypothesis that an object might be present in that region at a particular position and orientation (the hypothesis will have been obtained using methods from section **??**, which aren't immediately relevant). Is there an object there or not? A principled answer involves computing the posterior over two models — that the data was obtained from noise, or from the presence of an object.

- **Are these the same?** Assume we have a set of pictures of surfaces we want to compare. For example, we might want to know if they are the same colour, which would be difficult to answer directly if we didn't know the illuminant. A principled answer involves computing the posterior over two models — that the data was obtained from one surface, or from two (or more).

- **What camera was used?** Assume we have a sequence of pictures of a world. With a certain amount of work, it is usually possible to infer a great deal of information about the shape of the objects from such a sequence (section **??**). The algorithms involved differ quite sharply, depending on the camera model adopted (i.e. perspective, orthographic, etc.). Furthermore, adopting the wrong camera model tends to lead to poor inferences. Determining the right camera model to use is quite clearly a model selection problem.

- **How many segments are there?** We would like to break an image into coherent components, each of which is generated by a probabilistic model. How many components should there be? (section 18.3).

The solution is so absurdly simple *in principle* (in practice, the computations can be quite nasty) that it is easy to expect something more complex, and miss it. We will write out Bayes' rule specialised to this case to avoid this:

$$
\begin{aligned}
P(\text{model}|\text{data}) &= \frac{P(\text{data}|\text{model})}{P(\text{data})} \\
&= \frac{\int P(\text{data}|\text{model}, \text{parameters})P(\text{parameters})d\{\text{parameters}\}}{P(\text{data})} \\
&\propto \int P(\text{data}|\text{model}, \text{parameters})P(\text{parameters})d\{\text{parameters}\}
\end{aligned}
$$

which is exactly the form used in the example. Notice that we are engaging in Bayesian inference here, too, and so can report the MAP solution or report the whole posterior. The latter can be quite helpful when it is difficult to distinguish between models. For example, in the case of the dodgy informant, if $P(\text{truth}{=}0|\text{data}) = 0.5001$, it may be undesirable to conclude the informant is lying — or at least, to take drastic action based on this conclusion. The integral is potentially rather nasty, which means that the method can be quite difficult to use in practice. We will discuss methods for computing the integral in section **??**; useful references include [].

7.5.4 Open Issues

In the rest of the book, we will have regular encounters with practical aspects of the Bayesian philosphy. Firstly, although the posterior encapsulates all information available about the world, we very often need to make discrete decisions — should we shoot it or not? Typically, this decision making process requires some accounting for the cost of false positives and false negatives.

Secondly, how do we build models? There are three basic sources of likelihood functions and priors:

- **Judicious design:** it is possible to come up with models that are too hard to handle computationally. Generally, models on very high-dimensional domains are difficult to deal with, particularly if there is a great deal of interdependence between variables. There is a family of models — commonly known as **graphical models** — for which quite good inference algorithms are known. The underlying principle of this approach is to exploit simplifications due to independence and conditional independence. We describe this approach in chapter **??**, section **??** and in chapter **??**, section **??**, in the context of relevant examples.

- **Physics:** particularly in low-level vision problems, likelihood models follow quite simply from physics. It is hard to give a set of design rules for this strategy. Instead, we illustrate the approach with an example, in section **??**.

- **Learning:** as section **??** suggested, a poor choice of model results in poor performance, and a good choice of model results in good performance. We can use this observation to tune the structure of models if we have a sufficient set of data. We describe aspects of this strategy in chapter **??**.

Finally, the examples above suggest that posteriors can have a nasty functional form. This intuition is correct, and there is a body of technique that can help handle ugly posteriors which we explore in section **??**.

7.6 Discussion

Probabilistic inference techniques lie at the core of any solution to serious vision problems. The great difficulty, in our opinion, is arriving at a model that is both sufficiently accurate and sufficiently compact to allow useful inference. This isn't at all easy. A naive Bayesian view of vision — write out a posterior using the physics of illumination and reflection, guess some reasonable priors, and then study the posterior — very quickly falls apart. In terms of what representation should this posterior be written? and how can we extract information from the posterior? These questions are exciting research topics.

The examples in this chapter are all pretty simple, so as to expose the line of reasoning required. We do some hard examples in chapter **??**. Building and handling complex examples is still very much a research topic; however, probabilistic reasoning of one form or another is now pervasive in vision, which is why it's worth studying.

Exercises

1. The event structure of section 7.1 did not explicitly include unions. Why does the text say that unions are here?
2. In example **??**, if $P(\text{heads}) = p$, what is $P(\text{tails})$?
3. In example 10 show that if $P(\text{hh}) = p^2$ then $P(\{\text{ht}, \text{th}\}) = 2p(1-p)$ and $P(\text{tt}) = (1-p)^2$.
4. In example 10 it says that

   $$P(k \text{ heads and } n-k \text{ tails in } n \text{ flips}) = \begin{pmatrix} k \\ n \end{pmatrix} p^k (1-p)^{n-k}$$

   Show that this is true.
5. A careless study of example 10 often results in quite muddled reasoning, of the following form: I have bet on heads successfully ten times, therefore I should bet on tails next. Explain why this muddled reasoning — which has its own name, **the gambler's fallacy** in some circles, **anti-chance** in others — is muddled.
6. Confirm the count of parameters in example 8.

7. In example 19, what is c?

8. As in example 16, you are given a coin of either type I or type II; you do not know the type. You flip the coin n times, and observe k heads. You will infer the type of the coin using maximum likelihood estimation. for what values of k do you decide the coin is of type I?

9. Compute $P(\text{truth}|\text{high}, \text{coin behaviour})$ for each of the three cases of example 21. You'll have to estimate an integral numerically.

10. In example 22, what is the numerical value of the probability that the informant is lying, given that the informant said `high` and the coin shows a single `tail`? What is the numerical value of the probability that the informant is lying, given that the informant said `high` and the coin shows seven `tail`s in eight flips?

11. The random variable $\boldsymbol{x} = (x_1, x_2, \ldots x_n)^T$ has a normal distribution. Show that the random variable $\hat{\boldsymbol{x}} = (x_2, \ldots, x_n)^T$ has a normal distribution (which is obtained by marginalizing the density). A good way to think about this problem is to consider the mean and covariance of $\hat{\boldsymbol{x}}$, and reason about the behaviour of the integral; a bad way is to storm ahead and try and do the integral.

12. The random variable $\boldsymbol{p}$ has a normal distribution. Furthermore, there are symmetric matrices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ and vectors $\boldsymbol{D}$ and $\boldsymbol{E}$ such that $P(\boldsymbol{d}|\boldsymbol{p})$ has the form

$$-\log P(\boldsymbol{d}|\boldsymbol{p}) = \boldsymbol{p}^T \mathcal{A} \boldsymbol{p} + \boldsymbol{p}^T \mathcal{B} \boldsymbol{d} + \boldsymbol{d}^T \mathcal{C} \boldsymbol{d} + \boldsymbol{p}^T \boldsymbol{D} + \boldsymbol{d}^T \boldsymbol{E} + C$$

(C is the log of the normalisation constant). Show that $P(\boldsymbol{p}|\boldsymbol{d})$ is a normal distribution for any value of $\boldsymbol{d}$. This has the great advantage that inference is relatively easy.

13. x is a random variable with a continuous cumulative distribution function $F(x)$. Show that $u = F(x)$ is a random variable with a uniform density on the range $[0, 1]$. Now use this fact to show that $w = F^{-1}(u)$ is a random variable with cumulative distribution function F.