2

Probability and machine learning principles

Probability will be a recurring theme throughout the thesis. Indeed, its influence is surprisingly pervasive in the material making up the remaining chapters. The quantum mechanical underpinnings of the NMR effect which are outlined in chapter 3 describe probabilistic behaviours; the diffusion of water in the brain measured by dMRI is fundamentally a stochastic process; probabilistic sampling techniques are important to some of the tractography methods described in chapter 5; and the machine learning methods that we apply to the problem of tract selection in chapter 8 are probabilistic by their nature.

In this chapter we lay out the theory of probability and describe those machine learning and inference methods upon which later chapters are dependent. General references for this material include MacKay (2003) and Bishop (2006).

2.1 Fundamentals of probability theory

Consider a nondeterministic experiment, such as rolling a fair die. The result of this experiment on any given trial will be one of exactly six possibilities, representing the number of spots on the uppermost face of the die. Moreover, each of these possibilities is equally likely; so over a very large number of trials, all six will occur an approximately equal number of times. This kind of experiment is represented mathematically by a random variable, which we call X. The set of possible outcomes, or *sample space*, relating to X is $\{1,2,3,4,5,6\}$. The probability of each of these outcomes on a single trial is, of course, 1/6.

In general, we denote the sample space for a discrete random variable, X, as $\mathcal{A}_X = \{a_i\}$, where each member of the set has a corresponding probability, p_i. We write

$$\Pr(x = a_i) = p_i \,,$$

where "Pr" represents "the probability that", and x represents a particular outcome. The result $x = a_i$ is an example of an *event*, a concept which can generally encapsulate the occurrence of any subset of the sample space: if E is some subset of $\mathcal{A}_X$, we have

$$\Pr(E) = \Pr(x \in E) = \sum_{a_i \in E} \Pr(x = a_i) \,. \tag{2.1}$$

In the example of the die, if $E = \{1,2\}$, then $\Pr(E)$—the probability that the outcome of a trial is *either* 1 or 2—is the sum of $\Pr(x = 1)$ and $\Pr(x = 2)$, i.e. 1/3.

The basic axioms of probability state that the probability of any event is greater than or equal to zero, with the latter representing an impossible event; and that the probability of the whole sample space is unity—i.e. every outcome must be drawn from the space. That is,

$$\forall E \subseteq \mathcal{A}_X \,.\, \Pr(E) \geq 0 \qquad\qquad \Pr(\mathcal{A}_X) = 1 \,. \tag{2.2}$$

Naturally, Eq. (2.2) additionally implies that $\Pr(E) \leq 1$. In general, given any pair of events, E_1 and E_2, the probability of their union is given by

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2) \,. \tag{2.3}$$

This third axiom follows straight from Eq. (2.1). In the special case in which $\Pr(E_1 \cap E_2) = 0$, the two events cannot occur simultaneously and are therefore mutually exclusive.

We now consider another experiment, represented by the random variable Y, which consists of flipping a coin. The sample space for this variable can be represented as $\mathcal{A}_Y = \{0, 1\}$, where 0 represents a tail and 1 a head. If we perform both experiments together, what is the probability that the die roll produces a 6 *and* the coin toss gives a head? We represent this **joint probability** as $\Pr(x = 6, y = 1)$. Since the roll of the die and the coin toss can be assumed to have no influence on each other, the two events are *independent* and the joint probability is simply the product of the individual probabilities. For the case of two events that are not independent, we need to introduce the concept of a **conditional probability**, which is defined by

$$\Pr(x = a_i \,|\, y = b_j) \equiv \frac{\Pr(x = a_i, y = b_j)}{\Pr(y = b_j)} \quad \text{if } \Pr(y = b_j) \neq 0 \,,$$

and should be interpreted as "the probability that $x = a_i$ given that $y = b_j$". Hence, if we omit the particular value of each outcome to indicate the general case, it follows by trivial rearrangement that

$$\Pr(x, y) = \Pr(x \,|\, y) \Pr(y) \,, \tag{2.4}$$

which is called the *product rule* for probabilities. Consequently, the following are equivalent statements of independence between X and Y:

$$\Pr(x, y) = \Pr(x) \Pr(y) \qquad\qquad \Pr(x \,|\, y) = \Pr(x) \,.$$

Finally, given a group of joint probabilities, $\Pr(x, y)$, we can calculate the so-called **marginal probability**, $\Pr(x)$, by summing over all possible values of y; an operation known as marginalisation:

$$\Pr(x) \equiv \sum_{y \in \mathcal{A}_Y} \Pr(x, y) \,.$$

It follows from Eq. (2.4) that

$$\Pr(x) = \sum_{y \in \mathcal{A}_Y} \Pr(x \,|\, y) \Pr(y) \,, \tag{2.5}$$

a relationship which is called the *sum rule* for probabilities. These basic rules for combining probabilities together are extremely important in machine learning.

2.2 Probability distributions

Every random variable has associated with it a probability distribution, which can be used to assign to some interval $[\alpha, \beta]$ over the set of real numbers, $\mathbb{R}$, a probability that the corresponding outcome will fall within that interval on any given trial. For the example of our fair die, the distribution is easily defined as

$$P(x) = \begin{cases} \Pr(x) & \text{if } x \in \mathcal{A}_X \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

In this case, $P(x)$ is called the **probability mass function** (p.m.f.) for X. It then follows from Eqs (2.1) and (2.6) that

$$\Pr(\alpha \leq x \leq \beta) = \sum_{a_i \leq \beta} P(a_i) - \sum_{a_i < \alpha} P(a_i) = \sum_{a_i \leq \beta} \Pr(a_i) - \sum_{a_i < \alpha} \Pr(a_i) \,,$$

where $\Pr(a_i)$ is shorthand for $\Pr(x = a_i)$, and $a_i \in \mathcal{A}_X$ in all cases.

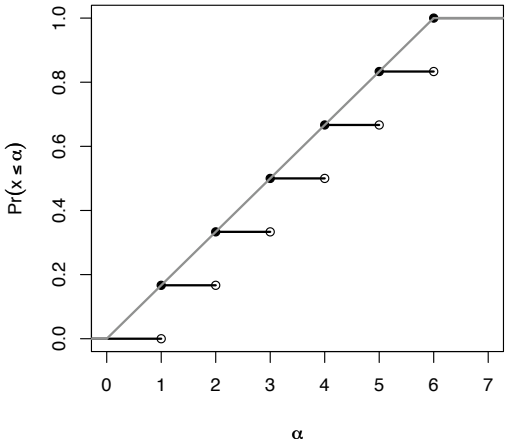

Figure 2.1: Cumulative distribution functions for discrete (black) and continuous (grey) uniform distributions. Open and closed circles indicate that each interval with a particular cumulative probability is open at one end and closed at the other; i.e. $\Pr(x \leq \alpha) = 1/6$ for $\alpha \in [1,2)$, for the discrete case.

It may appear that the distribution function buys us nothing over the individual probabilities for each outcome—after all, its only addition is to make explicit the fact that the probability of an outcome outside of the sample space is zero, a fact which follows uncontroversially from Eq. (2.2). However, the significance of probability distributions is far more obvious when we deal with continuous random variables.

Consider a continuous analogue of the die-rolling scenario, in which the outcome can be *any* real number in the interval $[0,6]$. The distribution for this continuous random variable is now defined by a **probability density function** (p.d.f.); specifically

$$P(x) = \begin{cases} \frac{1}{6} & \text{if } 0 \leq x \leq 6 \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

Notice that this distribution, while similar to the p.m.f. for the discrete case, is nonzero at an infinite number of points. As a result, the value of this p.d.f. at any given point does *not* represent a probability—if it did, the sum of probabilities across the sample space would be infinite, which defies the axioms of Eq. (2.2). Instead, the p.d.f. represents probability *density*, which is related to probability through integration:

$$\Pr(\alpha \leq x \leq \beta) = \int_{\alpha}^{\beta} P(x)\,\mathrm{d}x\,.$$

Consequently, the probability of an outcome having any particular value—i.e. $\Pr(x = \alpha)$—is *zero* for all values of $\alpha \in \mathbb{R}$, both within and outside the sample space, for any continuous random variable.

Every interval over the real numbers is a subset of $\mathbb{R}$, so the normalisation axiom in Eq. (2.2) implies that

$$\int_{-\infty}^{\infty} P(x)\,\mathrm{d}x = \int_{\mathcal{A}_X} P(x)\,\mathrm{d}x = 1\,, \tag{2.8}$$

since that part of the integral that is outside the sample space will be equal to zero.

The difference between the discrete and continuous versions of the distribution are most easily illustrated by comparing their *cumulative distribution functions* (c.d.f.s), which map each real number, α, to the probability that x is less than or equal to α. These functions are shown graphically in Fig. 2.1. It can be seen that $\Pr(x \leq \alpha)$ increases in jumps for the discrete case and smoothly for the continuous case; but for all integer values of α, the value of the c.d.f. is the same in both cases. Note that the c.d.f. is zero for all values below the lower bound of the sample space, and unity for all values above its upper bound, in each case.

Probability distributions are not only used in relation to static events—it is also common to consider a sequence of random variables, $(X(t))$, which are parameterised by t, often representing time in some sense. This parameter may be discrete or continuous. Such a collection of related variables, used to represent the state of some time-dependent system, is called a *stochastic process*. The evolution of such a process over time is then described by conditional distributions, such as $P(X(t)|X(t-1), X(t-2), \ldots)$ for the discrete-time case.

A final foundational concept with regard to probability distributions is that of the **expectation** of a random variable, which is essentially a weighted mean value over the sample space. For a discrete random variable, X, it is defined as

$$\langle X \rangle = \sum_{x \in \mathcal{A}_X} x P(x) \,, \tag{2.9}$$

and equivalently, using an integral, for the continuous case. Note that the expectation is a property of the random variable—or equivalently, its distribution—rather than of any outcome. We can also find the expectation of a function of X with respect to its probability distribution:

$$\langle f(X) \rangle = \sum_{x \in \mathcal{A}_X} f(x) P(x) \,. \tag{2.10}$$

If we know the distribution of a particular random variable, we can deduce the distribution of other random variables related to it. Let us assume that $X \sim U(0,6)$, which is shorthand to say that X is *uniformly* distributed over the sample space $[0,6]$, as described by Eq. (2.7). We now wish to know the distribution of the random variable

$$Y = 1 - \sqrt{X} \,.$$

We cannot find the distribution of Y by simply mapping the sample space accordingly, because this nonlinear function of X cannot be expected to have a uniform distribution itself; and even if it were linear, we would still need to ensure that the new distribution remains properly normalised. Instead, the rules of integration by substitution (Riley *et al.*, 2002) tell us that, using the Leibniz notation,

$$\mathrm{d}x = \frac{\partial x}{\partial y} \mathrm{d}y = 2(y-1)\,\mathrm{d}y \,;$$

so from Eq. (2.8),

$$\int_0^6 \frac{1}{6} \mathrm{d}x = \int_1^{1-\sqrt{6}} \frac{2(y-1)}{6} \mathrm{d}y = \int_{1-\sqrt{6}}^1 \frac{1-y}{3} \mathrm{d}y = 1 \,.$$

The distribution for Y is thus $P(y) = (1-y)/3$, and the sample space is $\mathcal{A}_Y = [1-\sqrt{6}, 1]$. It should be noted that substitutions for functions of more than one original variable are more complex, requiring the calculation of a Jacobian matrix of partial derivatives.

This process of finding the distribution of one random variable from that of another is very important when artificially sampling from a distribution. We sometimes wish to generate data with a certain distribution without truly sampling the value of an appropriate random variable many times; and while computing environments typically provide a method to generate uniformly distributed pseudorandom numbers, an appropriate transformation is needed to turn these into samples from the distribution of interest.

2.3 Inference and learning

So far we have talked about probabilities in terms of the chance of a particular event happening, on average, as a result of running a trial of a particular experiment. This interpretation of probability is the classical *frequentist* interpretation. However, there is an alternative, and broader, interpretation of probability which includes the sense of a *degree of belief*. Consider, for example, the relationship between the fact that the sky is cloudy and the fact that it is raining. Intuitively, if we are told that the sky is cloudy then it seems much more likely

that it is raining than if we are told that the sky is clear, or if we know nothing at all about state of the sky. However, the proposition "it is raining" cannot be strictly represented by a random variable since the experiment required to find an outcome (for example, going outside to look) is deterministic. Either it is raining or it isn't—there can be no two ways about it. It is also unrepeatable, since it is fixed to a particular time and we cannot sample the state of the weather *right now* many times. However, if we allow the broader interpretation of probability, we can admit a conditioned probability Pr(raining|cloudy), which represents how strongly we believe our proposition, given the truth of another proposition which says "the sky is cloudy". Moreover, we can use a distribution over the state space, in this case {raining,not raining}, to encapsulate the uncertainty we have about the proposition.

If this talk of using some propositions to inform others sounds like logical deduction, it is no coincidence. Some authors who subscribe to this broader, *Bayesian*, interpretation of probability—notably Jaynes (2003)—have been keen to frame it as a form of logical framework for the uncertain propositions that are common in science.

Note that before we are told about the state of the sky, it cannot influence our belief of whether it is raining or not. As a result, the **prior probability** that it is raining, Pr(raining), may be assumed to take the value 0.5, indicating total uncertainty. The distribution is then uniform over the two outcomes, which is an *uninformative* prior distribution because it tells us nothing except the size of the state space, which we already know. On the other hand, it may be that assumptions and information unrelated to the sky conditions could be incorporated into the prior distribution. Say, for example, that weather records tell us that it typically rains 20 per cent of the time—in that case we might instead use the prior Pr(raining) = 0.2. This is a trivial case of *inference*, whereby we use sample data—the weather records—to infer the nature of the distribution that is used to predict future weather. Note that we need to make an assumption, that previous weather will be representative of the future, in order to do even this simple an inference. In general, the making of assumptions is a prerequisite for inference.

Let's say that we have encoded our prior knowledge in a distribution of some kind. Now, introducing the knowledge that it is cloudy will alter the plausibility of the proposition that it is raining, but how? Given the fact that joint probabilities are symmetric, i.e. $P(x,y) = P(y,x)$, the relationship between the prior probability and the conditioned **posterior probability** can be established straight from Eq. (2.4). It is

$$\Pr(\text{raining}|\text{cloudy}) = \frac{\Pr(\text{cloudy}|\text{raining})\Pr(\text{raining})}{\Pr(\text{cloudy})} .$$

This relationship is the extremely important result known as **Bayes' rule**, after the 18th century mathematician and clergyman, the Rev. Thomas Bayes. It is significant because it describes a mathematical way to use relevant information to update the level of belief in a proposition—that is, to *learn*.

It turns out that the rules for manipulating probabilities that we looked at earlier can be applied to probability densities as well as probabilities, although showing that this is the case requires a more formal exploration of probability in terms of measure theory, which is beyond our scope here (see Kingman & Taylor, 1966). The same applies to Bayes' rule, so we can write in general,

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} . \tag{2.11}$$

The denominator of this equation, known in this context as the **evidence**, is commonly expanded using Eq. (2.5), in which the sum is replaced by an integral for the continuous case:

$$P(y) = \int_{\mathcal{A}_X} P(y|x)P(x) . \tag{2.12}$$

At this point, having introduced the Bayesian interpretation of probability, we will drop the notational distinction between distribution variables (including random variables) and outcome variables which has been used so far. This is common practice in the literature, and it helps to reduce the quantity of notation needed for dealing with more complex problems.

2.4 Maximum likelihood

We now have the tools in place to consider a more practically interesting example. Let us say that we have a random variable, x. We suspect that x is approximately normally distributed; that is, $x \sim N(\mu, \sigma^2)$, where μ (the mean) and σ^2 (the variance) are parameters of the distribution. We do not know what these parameters are, but if we want to make predictions about x we will need to know them. The definition of the normal, or Gaussian, distribution tells us that

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \tag{2.13}$$

In order to make any progress towards establishing μ and σ, we need some information. Let us assume that we have a data set, $D = \{d_i\}$ for $i \in \{1..N\}$, of example values of x. Since we are working on the assumption that x has the distribution given above, these data are assumed to be *samples* from the distribution. We assume that each sample has no dependence on any other, and that the values of μ and σ did not vary across the sample, a combination called the assumption of *independent and identically distributed* (i.i.d.) data. Hence, the product rule gives us a joint distribution for the whole sample data set:

$$P(D|\mu,\sigma) = \prod_{i=1}^{N} P(d_i|\mu,\sigma) . \tag{2.14}$$

The distribution given in Eq. (2.14) may not appear to get us any closer to an actual estimate for the parameters. But note that, from Eqs (2.11) and (2.12),

$$P(\mu,\sigma|D) = \frac{P(D|\mu,\sigma)P(\mu,\sigma)}{\iint P(D|\mu,\sigma)P(\mu,\sigma)\,\mathrm{d}\mu\,\mathrm{d}\sigma} . \tag{2.15}$$

Note that the distribution $P(D|\mu,\sigma)$, which is known as the **likelihood** of the parameters, is meaningful in a frequentist sense, since the elements of the data set are sample outcomes of the random variable x. However, the prior and posterior distributions over the parameters possess only Bayesian significance, since their values are fixed but unknown.

It makes intuitive sense to use as an estimate of the parameters those values which sit at the mode—that is, the point of maximal probability density—of the posterior distribution $P(\mu,\sigma|D)$. This approach amounts to finding the *most likely* values of the parameters in light of the sample data available. If we have no prior information about the parameters, so that $P(\mu,\sigma)$ is uninformative, then maximising the posterior is equivalent to maximising the likelihood, since the evidence is a normalisation factor that is not dependent on the values chosen for μ and σ. Hence, we can find a *maximum likelihood estimator* for the parameters by maximising the value of Eq. (2.14) with respect to them.

In practice, it is often mathematically easier to maximise the (natural) logarithm of the likelihood. This is valid because $\ln n$ will always increase when n increases—we say that the logarithm is a *monotonically increasing* function. Elementary calculus tells us that at the maximum of a function its derivative is zero, so from Eqs (2.13) and (2.14), our estimator of μ is given when

$$\frac{\partial}{\partial\mu}\left(-\frac{1}{2}\sum_{i=1}^{N}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right) = 0 .$$

Solving this equation gives us the value of the estimator for μ as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} d_i .$$

The "hat" notation is commonly used to indicate an estimate. Note that this maximum likelihood (ML) estimate is exactly equal to the mean of the sample. The maximum likelihood

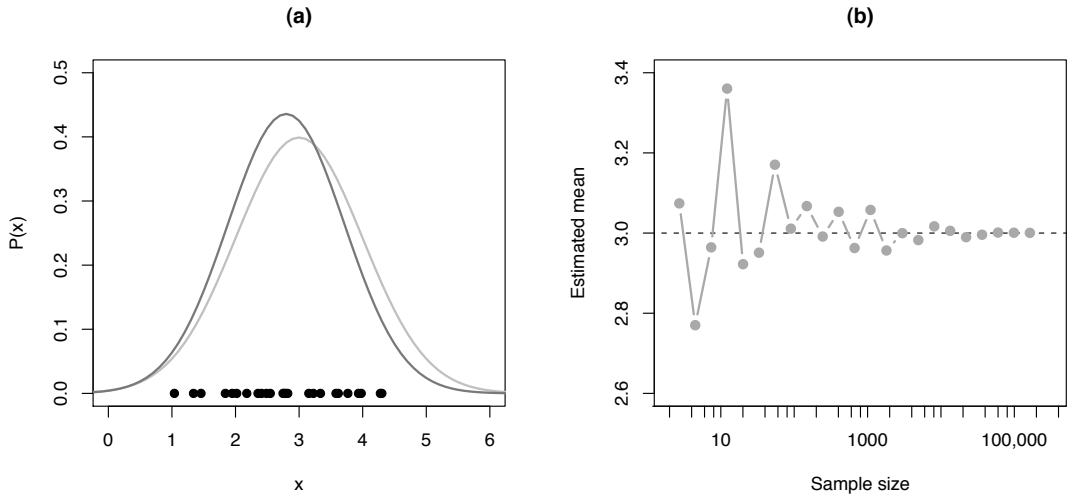

Figure 2.2: Maximum likelihood estimation for a Gaussian distribution. **(a)** A set of sample data (black points), the generating distribution (light grey line) and estimated distribution (dark grey line). **(b)** The estimated mean approaches the generative mean as the size of the sample vector increases.

variance also turns out, in this case, to be the given by the (biased) variance of the sample, viz.

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{\mu})^2 \,.$$

The two parameters can be estimated separately because $\hat{\mu}$ has no dependence on $\hat{\sigma}$. It is possible to demonstrate, by taking second derivatives, that these estimates really represent a maximum in the likelihood function.

Let's take a step back at this point and consider what we have done. We were given a set of sample values of x. We hypothesised, and thereafter assumed, that the samples were drawn from a Gaussian distribution with unknown mean and variance. In the language of machine learning, this Gaussian distribution is our **model** for the data, and μ and σ are parameters associated with that model. We have no direct way of establishing the values of these parameters, but we used the observed data and Bayes' rule, which can be summarised in words as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \,,$$

to learn the most likely estimate for the parameters given the observed data. Since our model describes a distribution which could be used to generate data like D, it is called a *generative* model.

The process is illustrated by Fig. 2.2(a). A sample of 25 points are shown in black—these were sampled from a Gaussian distribution with mean 3 and variance 1, whose p.d.f. is shown by the lighter curve. The learnt model distribution is the darker curve. It can be seen that the peak of the distribution—the mode, which is equal to the mean for a Gaussian distribution—is slightly offset from that of the generating distribution, and the "broadness" of the curve—which indicates the variance—is slightly less. Nevertheless, the estimated distribution may be considered a satisfactory approximation, and thus useful for predicting the general behaviour of the variable x. Not surprisingly, increasing the size of the sample vector will produce maximum likelihood estimators that are closer, on average, to the generative parameters—as demonstrated by Fig. 2.2(b). This effect is called the law of large numbers.

It should be remembered that the maximum likelihood method implicitly assumes that the priors in Eq. (2.15) are uninformative. If, on the other hand, meaningful prior information *is*

available, and we wish to take a more firmly Bayesian approach, we can calculate the maximum of the posterior distribution with the prior distribution incorporated into it. This more general approach to choosing an estimate for the parameters is called the maximum *a posteriori* (MAP) method, and it allows us to influence the parameter estimate based on what we know in advance.

2.5 Expectation–Maximisation

Unfortunately, it is quite easy to find cases in which simple maximum likelihood estimation is insufficient to find an estimate for a set of parameters. Consider the two-dimensional, or bivariate, version of the Gaussian distribution described by Eq. (2.13). It is

$$P(x,y|\boldsymbol{\mu},\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma^2}\right). \tag{2.16}$$

This is effectively a special case of Eq. (2.14), because we are treating the x and y dimensions as independent. This will only be the case if the covariance between x and y is zero; but we make that assumption here to avoid overcomplication. Note also that the mean, $\boldsymbol{\mu} = (\mu_x, \mu_y)$, is now a vector quantity since it has a component in each dimension. Consider now

$$P(x,y|\theta) = aP_1(x,y|\theta) + (1-a)P_2(x,y|\theta)\,, \tag{2.17}$$

where each of P_1 and P_2 have the distribution given in Eq. (2.16), and $\theta = \{\boldsymbol{\mu}_1, \sigma_1, \boldsymbol{\mu}_2, \sigma_2\}$ is a collection of all the parameters of this model. Eq. (2.17) is called a Gaussian *mixture model*, because it is made up of a combination of two independent Gaussian distributions over the same parameter space. The parameter a, which must be in the interval $[0,1]$ to ensure that the overall distribution is properly normalised, is called the mixture coefficient. We include it in the set $\phi = \{\boldsymbol{\mu}_1, \sigma_1, \boldsymbol{\mu}_2, \sigma_2, a\}$, a superset of θ.

In a generative sense, any sample data point must be drawn from exactly one of the component distributions, P_1 and P_2. We say there is a **latent variable**, which we denote z_i, associated with each data point, $\mathbf{d}_i$. We can characterise this variable by defining

$$z_i = \begin{cases} 1 & \text{if } \mathbf{d}_i \text{ was drawn from } P_1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

By analogy with the maximum likelihood estimation process for a single Gaussian distribution, we might expect to be able to infer the mean and variance of P_1 according to

$$\hat{\boldsymbol{\mu}}_1 = \frac{\sum_{i=1}^N z_i \mathbf{d}_i}{\sum_{i=1}^N z_i} \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N z_i \|\mathbf{d}_i - \hat{\boldsymbol{\mu}}_1\|^2}{\sum_{i=1}^N z_i}\,, \tag{2.19}$$

where $\|\cdot\|$ is the Euclidean norm; and similarly for P_2. (Note that $\sum_i z_i$ is equal to the number of data points that were drawn from P_1.) However, without any knowledge of the set $Z = \{z_i\}$, Eq. (2.19) cannot be evaluated, and so no estimate for ϕ can be calculated. Conversely, if ϕ were known then Z could be inferred, but we have neither.

The **Expectation–Maximisation** (EM) method provides a way to estimate both ϕ and Z simultaneously, thus sidestepping the problem of their mutual dependency (Dempster *et al.*, 1977). The method is initialised by choosing a first estimate, $\hat{\phi}$, for the parameters. After that, an expectation step, or "E-step", and a maximisation step, or "M-step", are applied iteratively until some termination criterion is met. Each E-step calculates a posterior distribution for Z based on the current parameter estimate, while the M-step updates the parameters.

We once again assume that the elements of our data set, $D = \{\mathbf{d}_i\}$, are i.i.d., and hence the values of z_i are also independent. As a result, the posterior over Z can be expanded to

$$P(Z|D,\hat{\phi}) = \prod_{i=1}^N P(z_i|\mathbf{d}_i,\hat{\phi})\,, \tag{2.20}$$

and so we can consider the posterior for each z_i individually. Bayes' rule gives us

$$P(z_i|\mathbf{d}_i,\hat{\phi}) = \frac{P(\mathbf{d}_i|z_i,\hat{\phi})P(z_i|\hat{\phi})}{\sum_{z_i} P(\mathbf{d}_i|z_i,\hat{\phi})P(z_i|\hat{\phi})}\,, \tag{2.21}$$

where $\sum_{z_i}$ is shorthand for the sum over the sample space of z_i. Note that the distributions over z_i are discrete, so the prior $P(z_i = 1)$ is meaningful, and will in general be nonzero. Its exact value will be given by the current estimate for the mixture coefficient, $\hat{a}$, which is updated by the M-step below; and $P(z_i = 0)$ follows directly by normalisation.

Observe that the particular case $P(\mathbf{d}_i|z_i = 1,\hat{\phi})$ is equivalent to $P_1(\mathbf{d}_i|\hat{\theta})$, a fact that follows straight from the definition of z_i in Eq. (2.18). As a result, we can expand Eq. (2.21) by exhaustive enumeration of the two outcomes, as follows.

$$P(z_i = 1|\mathbf{d}_i,\hat{\phi}) = \frac{\hat{a}P_1(\mathbf{d}_i|\hat{\theta})}{\hat{a}P_1(\mathbf{d}_i|\hat{\theta}) + (1-\hat{a})P_2(\mathbf{d}_i|\hat{\theta})} \tag{2.22}$$

$$P(z_i = 0|\mathbf{d}_i,\hat{\phi}) = \frac{(1-\hat{a})P_2(\mathbf{d}_i|\hat{\theta})}{\hat{a}P_1(\mathbf{d}_i|\hat{\theta}) + (1-\hat{a})P_2(\mathbf{d}_i|\hat{\theta})} \tag{2.23}$$

The job of the M-step is to refine our current estimate for $\hat{\phi}$. In order to do this, we need concrete values for each z_i. Since the E-step has already calculated posterior distributions for z_i in Eqs (2.22) and (2.23), we simply take as our z_i values the expectations of these distributions:

$$\langle z_i\rangle = \sum_{z_i} z_i P(z_i) = P(z_i = 1)\,.$$

Note that due to the nature of the definition of z_i, this expectation is equal to the value of $P(z_i = 1)$ calculated in Eq. (2.22). Hence, using these values for z_i, we can update our estimates for the means and variances of P_1 and P_2 with ML, according to Eq. (2.19).

All that remains for the M-step is to update $\hat{a}$, the remaining element of $\hat{\phi}$. Our estimate for this parameter is the expected mean value of the set of latent variables, given by

$$\hat{a} = \left\langle \frac{1}{N}\sum_{i=1}^{N} z_i \right\rangle = \frac{1}{N}\sum_{i=1}^{N}\langle z_i\rangle = \frac{1}{N}\sum_{i=1}^{N} P(z_i = 1)\,.$$

Fig. 2.3 shows a graphical representation of the process, in which each small filled circle represents a data point. The posterior distribution over each latent variable, as calculated by the E-step, is indicated by a colour, with pure red indicating that $P(z_i = 1|\mathbf{d}_i,\hat{\phi}) = 1$, and pure blue indicating the opposite definite outcome. Hence, the shade of each data point represents how likely it is to be drawn from each of the component distributions. It can be seen that after a single iteration of the algorithm, the estimated component distributions, which are updated by the M-step, have a large variance and significant overlap; and as a result the assignment of data to each component is uncertain, so all points appear in shades of purple. By contrast, after 11 further iterations, the algorithm has converged to a stable solution and most points appear red or blue, since they are much more likely to be from one component distribution than the other. There is just one point that remains ambiguous.

A useful way to gauge the progress of the algorithm is to plot the overall data log-likelihood (DLL), given by

$$\ln P(D|\hat{\phi}) = \sum_{i=1}^{N} \ln P(\mathbf{d}_i|\hat{\phi}) = \sum_{i=1}^{N} \ln\left(\sum_{z_i} P(\mathbf{d}_i|z_i,\hat{\phi})P(z_i|\hat{\phi})\right),$$

which can be calculated after each iteration of the algorithm. The DLL gives us an idea of how well the current model explains the data. Since EM is a maximum likelihood technique—differing practically from the simpler ML estimation of §2.4 in that it can cope with models that

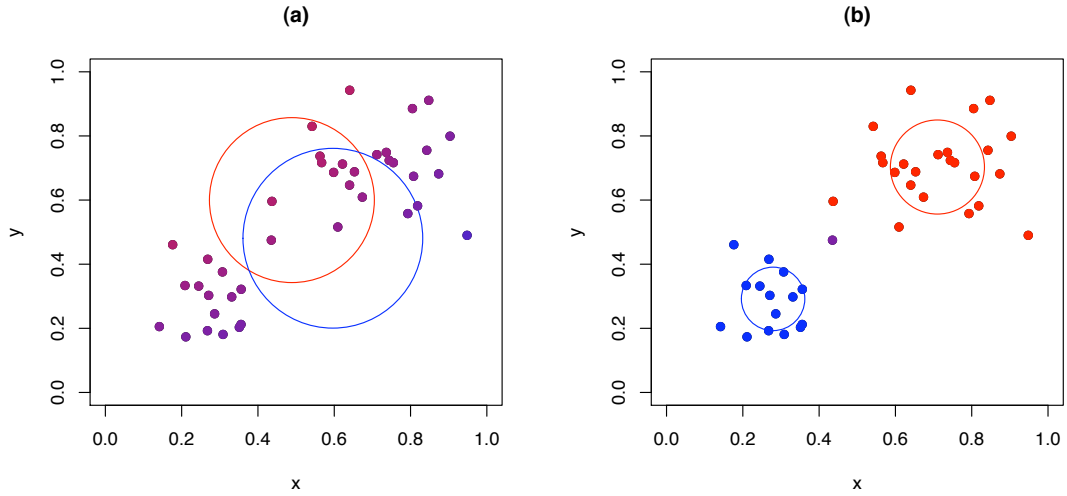

Figure 2.3: Results of applying Expectation–Maximisation to a Gaussian mixture model, after one iteration **(a)** and at convergence **(b)**. Each large circle represents a component distribution, centred at the mean and with radius equal to one standard deviation. Data points with z_i closer to 1 are more red, and those closer to 0 are more blue. The generating distribution has parameters $\mu_1 = (0.3, 0.3)$, $\mu_2 = (0.7, 0.7)$, $\sigma_1 = \sigma_2 = 0.1$, and $a = 0.5$.

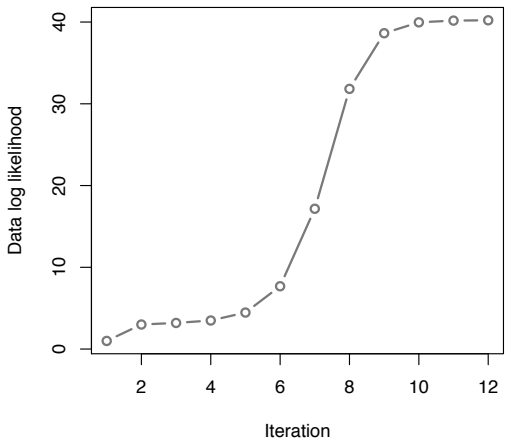

Figure 2.4: Typical plot of data log-likelihood as the Expectation–Maximisation algorithm progresses.

include latent variables—we might expect that the DLL would be at its peak when the algorithm terminates.

An example plot of DLL is shown in Fig. 2.4. The first M-step produces a very large increase in DLL (not shown), after which there is a general increase, ending with a final asymptotic convergence on a maximum likelihood value. Note is that there is *never* a drop in DLL from one iteration to the next. This is guaranteed by the theory of the EM method, which is beyond our scope here (see Bishop, 2006).

2.6 Sampling methods

Up to this point we have dealt with very simple, analytically tractable model distributions; and moreover we have been happy to work with a single estimate for the parameters of the model. However, a maximum likelihood estimator for the parameters does not always exist; and in practice it is often useful to be able to fully characterise a distribution over the model parameter space—that is, the joint sample space of all parameters.

Consider a general case in which we have a scalar valued quantity, x, modelled by a distribution with parameter set θ. The now-familiar Bayes' rule defines the posterior distribution for the parameter set according to

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}\,, \tag{2.24}$$

where

$$P(x) = \int_{\mathcal{A}_\theta} P(x|\theta)P(\theta)\,\mathrm{d}\theta\,. \tag{2.25}$$

If we can evaluate the normalisation constant, Eq. (2.25), analytically then it will be possible to characterise Eq. (2.24) exactly. The full posterior distribution over θ would then be able to provide information on not only the most likely value of θ—i.e. the mode of the distribution—but also on the extent to which such an estimate is likely to be valid or useful. For example, the distribution might have multiple modes, in which case taking a single estimate for the parameters may be inappropriate.

The problem is that for a complicated likelihood function, the integral in Eq. (2.25) may be impossible to evaluate analytically, putting exact marginalisation out of reach. Similar problems occur when trying to find the expectation of a function with respect to a complex distribution. In such cases, it may instead be practical to approximately infer the *target density* over θ by drawing samples from it. Given a set of these samples, $\{\theta^{(i)}\}$ for $i \in \{1..N\}$, the approximation is then a probability mass function of the form

$$\hat{P}(\theta) = \frac{1}{N}\sum_{i=1}^{N} P_\delta(\theta|\theta^{(i)})\,, \tag{2.26}$$

where $P_\delta(\theta)$ is a p.m.f. analogue of the Dirac delta function:

$$P_\delta(\theta|\theta^{(i)}) = \begin{cases} 1 & \text{if } \theta = \theta^{(i)} \\ 0 & \text{otherwise.} \end{cases}$$

This is the principle of so-called Monte Carlo (MC) methods, which include the sampling techniques described below (for a review see Andrieu *et al.*, 2003). Of course, the approach presupposes that it is possible to evaluate the distribution of interest, but this is the case often enough for the assumption to be tenable for a wide range of practical problems. In fact, it is sufficient to evaluate the target density to within a multiplicative constant, since the approximating p.m.f., Eq. (2.26), is self-normalising. This is extremely useful, because it obviates the need to evaluate the evidence term in Eq. (2.24) when sampling from the posterior distribution.

Moreover, by the law of large numbers the expectation of some function, f, with respect to $\hat{P}(\theta)$ will converge towards the expectation of the same function with respect to the true, continuous distribution for θ as N increases:

$$\langle f(\theta)\rangle_{\hat{P}(\theta)} = \frac{1}{N}\sum_{i=1}^{N} f(\theta)P_\delta(\theta\,|\,\theta^{(i)}) = \frac{1}{N}\sum_{i=1}^{N} f(\theta^{(i)}) \xrightarrow{N\to\infty} \langle f(\theta)\rangle_{P(\theta)} = \int_{\mathcal{A}_\theta} f(\theta)P(\theta)\,\mathrm{d}\theta\,.$$

The issue now becomes one of choosing samples: how can we efficiently generate pseudorandom numbers which accurately represent the unknown target distribution? We are generally primarily interested in regions of the parameter space in which $P(\theta)$ is relatively large, but how can we identify such places without evaluating the distribution everywhere?

The naïve method of sampling at every point on a grid throughout the space will quickly become unfeasible, especially if the space has high dimensionality—that is, if there are a large number of parameters. The next most simple approach is to choose points randomly and uniformly from the parameter space, and sample the distribution at those points. However, since areas of high probability density are usually concentrated in a small region of the space, the number of samples required to ensure that this *typical set* is reached at least a few times will still often be prohibitively large.

2.6.1 Rejection sampling

A more sophisticated general approach to the sampling problem is to avoid sampling directly from the unknown target density, $P(x)$, and instead sample from a known, simpler *proposal density*. In particular, if we can evaluate $\tilde{P}(x) = zP(x)$, where z is an unknown constant, and we can find a proposal density, $Q(x)$, and a finite positive real number, k, such that $\tilde{P}(x) \leq kQ(x)$ for all real x, then we can apply a method known as *rejection sampling*.

Fig. 2.5(a) shows a situation in which this approach is appropriate. In this case the target density is a Gaussian mixture with component means at $x = 3$ and $x = 5$; and the proposal density is a simple Gaussian distribution, centred at $x = 3.5$, with $k = 2$. In a one-dimensional case such as this, it is easy to see by inspection that the proposal density is always greater than the target density.

The process for generating N samples from the target density is given by Algorithm 2.1. In common with most MC methods, the rejection sampling algorithm involves the use of (uniformly distributed) random numbers. At each step, a candidate sample, x^*, is generated from the proposal distribution and a random number, u, is drawn from a uniform distribution over [0,1]. Then, if

$$u < \frac{\tilde{P}(x^*)}{kQ(x^*)}\,,$$

the sample is "accepted" as a sample from $\tilde{P}(x)$; otherwise it is rejected and another candidate sample is drawn. The significance of this acceptance criterion is shown by Fig. 2.5(a): it amounts to a test of whether the quantity $ukQ(x^*)$, which is uniformly distributed between

Require: $k \in (0, \infty)$
$i \leftarrow 0$
repeat
  Sample $x^* \sim Q(x)$ and $u \sim U(0,1)$
  if $ukQ(x^*) < \tilde{P}(x^*)$ **then**
    $i \leftarrow i + 1$
    $x^{(i)} \leftarrow x^*$ [Accept x^*]
  else
    Reject x^*
  end if
until $i = N$

Algorithm 2.1: Rejection sampling for N samples.

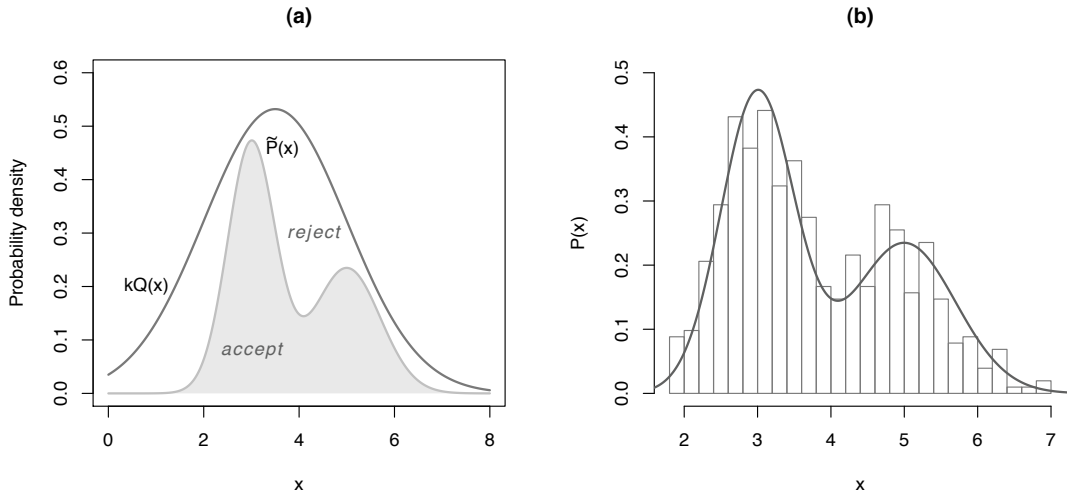

Figure 2.5: Rejection sampling for a univariate Gaussian mixture. **(a)** The target and proposal densities. Samples from the proposal density will be accepted if $ukQ(x^*) < \tilde{P}(x^*)$—this corresponds to the shaded area under the target curve. **(b)** Histogram of the accepted samples, overlaid with the exact target density. In this case 51% of samples from the proposal density were accepted.

zero and the value of the proposal density at $x = x^*$, falls below the target density. Thus more samples will be accepted in regions where the two densities are very similar, and far fewer in areas where $\tilde{P}(x) \ll kQ(x)$. As a result, the technique is most efficient when the proposal density closely approximates the target density. In particular, the two should have as large an overlap in their typical sets as possible. This is certainly the case in our example: both densities are defined for all real numbers, but the vast majority of the probability mass is in the interval [0,8]. A uniform proposal density is the worst case, in which case rejection sampling is equivalent to uniform sampling.

After choosing 1000 samples from the proposal distribution, of which 51% were accepted, Fig. 2.5(b) shows a histogram of the accepted samples for a single run of our example case. It can be seen that the normalised histogram agrees quite well with the true target distribution, which is overlaid.

The probability that any given candidate sample is accepted is given by the expectation of the density ratio with respect to the proposal distribution:

$$\Pr(\text{accepted}) = \int_{\mathcal{A}_x} \frac{\tilde{P}(x)}{kQ(x)} Q(x)\,\mathrm{d}x = \frac{1}{k}\int_{\mathcal{A}_x} \tilde{P}(x)\,\mathrm{d}x = \frac{z}{k}\,.$$

Hence in the example, where $z = 1$ and $k = 2$, we expect around half of samples to be accepted. However, this relationship highlights a crucial shortcoming of rejection sampling—as k increases, fewer and fewer samples will be accepted, so the run time required to obtain a reasonable sample size from the target density will also increase. For target distributions over high-dimensional sample spaces, it may be hard to find an appropriate value for k at all; but even if one can be found it will tend to be large, making the method impractical. In such cases, it will be necessary to be more clever about the choice of sampling locations.

2.6.2 Markov chain Monte Carlo

A *Markov chain* is a particular type of discrete-time stochastic process in which the state of the system at time t is dependent only on its state at the previous time step, $t-1$. That is,

$$P(x(t)\,|\,x(t-1), x(t-2), \ldots, x(0)) = P(x(t)\,|\,x(t-1))\,; \tag{2.27}$$

Initialise $x^{(0)}$
for $i \in \{1..N\}$ **do**
  Sample $x^* \sim Q(x|x^{(i-1)})$ and $u \sim U(0,1)$
  if $u < A(x^*, x^{(i-1)})$ **then**
    $x^{(i)} \leftarrow x^*$
  else
    $x^{(i)} \leftarrow x^{(i-1)}$
  end if
end for

Algorithm 2.2: The Metropolis and Metropolis–Hastings algorithms. The difference between the two methods is in the choice of acceptance function, A.

the so-called Markov property. The distribution on the right hand side of Eq. (2.27) is called a *transition kernel*.

A subclass of MC techniques called **Markov chain Monte Carlo** (MCMC) methods are designed such that the set of samples drawn forms a Markov chain with the target density as an invariant distribution. Details on how this is achieved can be found in more complete treatments of MCMC methods, such as Neal (1993).

The *Metropolis algorithm* (Metropolis *et al.*, 1953) is an early MCMC method which assumes that the proposal density from which candidate samples, x^*, are sampled is symmetric in the sense that

$$Q(x^* \,|\, x^{(i)}) = Q(x^{(i)} \,|\, x^*) \,.$$

Under these circumstances a candidate sample drawn from this distribution is accepted with probability

$$A(x^*, x^{(i)}) = \min\left\{1, \frac{\tilde{P}(x^*)}{\tilde{P}(x^{(i)})}\right\} \,, \tag{2.28}$$

where $\tilde{P}(x)$ is proportional to the target density, $P(x)$, as before. If the candidate sample is accepted then it becomes the new sample, $x^{(i)}$; if not, then the new sample is *the same* as the previous one: $x^{(i)} = x^{(i-1)}$. Thus the effect of rejecting a sample differs from the rejection sampling approach in that a new sample is *always* created on each step of the algorithm.

It can be seen directly from Eq. (2.28) that if the value of the target density at x^* is greater than that at $x^{(i)}$, then the sample will always be accepted. On the other hand, if the proposed new sample location represents a substantial drop in probability density, then it is very unlikely to be accepted, and the chain is most likely to remain in its previous state. The result of this policy is that the chain will spend most time in regions of the sample space where the target density is high-valued, as we require.

The Metropolis algorithm was later generalised by W. Keith Hastings to include the case in which the proposal distribution is not symmetric (Hastings, 1970). In this case the acceptance probability is given by

$$A(x^*, x^{(i)}) = \min\left\{1, \frac{\tilde{P}(x^*)Q(x^{(i)}|x^*)}{\tilde{P}(x^{(i)})Q(x^*|x^{(i)})}\right\} \,. \tag{2.29}$$

Algorithm 2.2 describes the Metropolis and Metropolis–Hastings algorithms, given appropriate forms for A. It is important to note that unlike the rejection sampling method,

Initialise $\mathbf{x}^{(0)}$
for $i \in \{1..N\}$ **do**
  Sample $x_1^{(i)} \sim P(x_1 \,|\, x_2^{(i-1)}, x_3^{(i-1)}, \ldots, x_n^{(i-1)})$
  Sample $x_2^{(i)} \sim P(x_2 \,|\, x_1^{(i-1)}, x_3^{(i-1)}, \ldots, x_n^{(i-1)})$
  etc.
end for

Algorithm 2.3: Gibbs sampling over a vector quantity, $\mathbf{x}$.

Metropolis–Hastings generates correlated, rather than independent, samples. However, if a subset consisting of, say, every 50th sample is taken, then these may be considered to be close enough to independent for most practical purposes. The proportion of samples which may be kept whilst retaining approximate independence will depend on the exact form of the proposal density, as will the performance of the method in approximating its target. In particular, if the variance of the proposal density is very large, few candidate samples will be accepted, resulting in highly correlated samples; and if it is very small then some significant regions of the parameter space may be left unexplored.

The extension of these methods to the multivariate case where each sample is a vector, $\mathbf{x}^{(i)}$, just requires that the proposal distribution be defined in the appropriate number of dimensions. There is no change needed to the algorithms themselves. However, under a popular special case of the Metropolis–Hastings algorithm called *Gibbs sampling*, each element of such a vector is sampled from a different proposal distribution (Geman & Geman, 1984). This method requires that the conditional distributions of each element in the sample vector given all other elements be known, because these are used as the proposal distributions (see Algorithm 2.3). It can be shown that under these circumstances, the acceptance probability for samples is unity, and so this method is highly efficient.

2.7 Summary

In this chapter we have reviewed the basic principles of probability, and explained how the strict, frequentist interpretation of probability can be broadened to encompass any proposition with which uncertainty is associated. We have also looked at the basic mechanisms of inference and learning from data, which typically involve the use of Bayes' rule. The rationale for maximum likelihood and maximum *a posteriori* parameter estimates has been explained, and methods for calculating such estimates, including the Expectation–Maximisation approach, have been outlined. Finally, we explored ways in which a probability distribution, whose exact form cannot be calculated analytically, can be approximated efficiently from data. The probabilistic perspective will appear commonly throughout the remainder of this thesis, and we will outline techniques which rely on some of the tools and ideas marshalled above.