CHAPTER 2

ADAPTIVE LEARNING

At the heart of any learning problem is a *belief model*, which is some type of statistical model that characterizes what we know about the behavior of our system. However, unlike the usual statistical model which is just a point estimate, our belief models will always include an explicit representation of the uncertainty in our model. Thus, if we have an estimate that driving a particular route is likely to take 25 minutes, our belief model will include a distribution about what the true mean might be. These beliefs will evolve as new information arrives, and our goal in this book is to acquire information that has the biggest impact on the quality of our decisions.

There are two perspectives that we can take when forming a belief, known as the *frequentist view* and the *Bayesian view*. In the frequentist view, we begin with no knowledge at all about our parameters, and our beliefs are formed entirely by the results of experiments that we run. Since experiments are inherently noisy, we can repeat a series of experiments and obtain different estimates of the parameters. If we repeat the experiments often enough, we can form a frequency distribution of any parameter that we estimate using the experiments.

In the Bayesian view, we start with initial beliefs about parameters, known as the *prior distribution*, which is formed before we make any observations. After an experiment, we combine the prior distribution with an experiment to form a posterior distribution. This then becomes the next prior distribution. This distribution is our *distribution of belief* about the true values of a set of parameters, formed from a combination of our initial belief and subsequent observations. By contrast, in the frequentist view, we form a probability distribution of estimates of parameters that reflect the variation in the observations.

Optimal Learning. By Warren B. Powell and Ilya O. Ryzhov

Copyright © 2018 John Wiley & Sons, Inc.

We begin with an overview of major classes of belief models, including lookup table, parametric and nonparametric models.

2.1 BELIEF MODELS

At the heart of any learning problem is some sort of statistical model we refer to as the *belief model*, which captures not only our estimate of how a function or system will respond to a controllable input x, but also the uncertainty in our estimate of the function.

We can represent our function or system as $f(x)$. Often, this is a function of a controllable input x (drug dosages, price of a product, concentration of a chemical), and a random variable W which covers any uncontrollable (we sometimes refer to these as exogenous) inputs. In principle, we might write

$$f(x) = \mathbb{E}F(x, W),$$

where the expectation $\mathbb{E}$ averages over all possible outcomes W.

Most of the time in this book we are going to treat x as being a discrete choice. We can write the set of choices as $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$. These might be discrete choices (color, choice of material, choice of drug), or discretized values of a continuous parameter (price, length, dosage). If x is multidimensional, we might view $\mathcal{X}$ as a sample of possible values of x. However, we will occasionally address situations where x is continuous, either scalar, or a continuous vector.

Belief models can be described as coming in one of three flavors:

Lookup tables Assume that we have a discrete set of choices $x \in \mathcal{X} = \{x_1, \ldots, x_M\}$ and let

$$\mu_x = f(x) = \mathbb{E}_W F(x, W).$$

A lookup table representation refers to an estimate $\bar{\mu}^n_x \approx \mu_x$ for each $x \in \mathcal{X}$. So, if $M = 100$, then we need to estimate 100 different parameters.

Parametric models There are settings where the set $\mathcal{X}$ can become quite large, as occurs when x is multidimensional, or continuous. In such a setting, we might want to write our belief model as

$$\begin{aligned} \bar{f}(x|\theta) &\approx \mathbb{E}F(x, W) \\ &= \sum_{f \in \mathcal{F}} \theta_f \phi_f(x), \end{aligned} \tag{2.1}$$

where $\phi_f(x),\ f \in \mathcal{F}$ is a set of features. For example, x might be a choice of a movie, while $\phi_f(x)$ captures features such as genre. Now, instead of having a belief μ_x for every movie x, we just have to estimate a vector of parameters $(\theta_f),\ f \in \mathcal{F}$ for a presumably small set of features. Equation (2.1) illustrates a linear model, which means it is linear in θ (the features $\phi_f(x)$ may be highly nonlinear in x). We may also need to use a nonlinear model such as

$$\bar{f}(x|\theta) = \begin{cases} 1 & \text{If } x < \theta^{min} \\ 0 & \text{If } \theta^{min} < x < \theta^{max} \\ \text{-1} & \text{If } x > \theta^{max} \end{cases}$$

Another example might be a logistic function

$$\bar{f}(x|\theta) = \frac{e^{\theta_0+\theta_1 x}}{1+e^{\theta_0+\theta_1 x}}.$$

Nonparametric models Nonparametric models allow us to create estimates without having to assume a functional form such as our linear model above, but the price is high. Imagine we have a set of observations $(x^n, \hat{f}^n),\ n = 1, \ldots, N$. We can create an approximation $\bar{f}(x)$ by using a local average around x, consisting of an average of points $\hat{f}^n$ with weights that are inversely proportional to the distance $\|x - x^n\|$. Nonparametric models are quite flexible, but can be hard to use, and as a result will not play a major role in our treatment.

The remainder of this chapter will focus on methods for updating lookup table belief models. For this setting, we are going to use W_x^n as the observation of our function $f(x)$, which means that if we choose to evaluate the function at $x = x^n$, then we are going to observe

$$\begin{aligned} W_{x^n}^{n+1} &= f(x^n) + \varepsilon^{n+1} \\ &= \mu_{x^n} + \varepsilon^{n+1}. \end{aligned}$$

Below, we describe how to use frequentist and Bayesian statistics to produce estimates of μ_x for each x.

2.2 THE FREQUENTIST VIEW

The frequentist view is arguably the approach that is most familiar to people with an introductory course in statistics. Assume we are trying to estimate the mean μ of a random variable W which might be the performance of a device or policy. Let W^n be the nth sample observation. Also let $\bar{\mu}^n$ be our estimate of μ, and $\hat{\sigma}^{2,n}$ be our estimate of the variance of W. We know from elementary statistics that we can write $\bar{\mu}^n$ and $\hat{\sigma}^{2,n}$ using

$$\bar{\mu}^n = \frac{1}{n}\sum_{m=1}^{n} W^m \tag{2.2}$$

$$\hat{\sigma}^{2,n} = \frac{1}{n-1}\sum_{m=1}^{n} (\hat{f}^m - \bar{\mu}^n)^2. \tag{2.3}$$

The estimate $\bar{\mu}^n$ is a random variable (in the frequentist view) because it is computed from other random variables, namely $W^1, W^2, \ldots, W^n$. Imagine if we had 100 people each choose a sample of n observations of W. We would obtain 100 different estimates of $\bar{\mu}^n$, reflecting the variation in our observations of W. The best estimate of the variance of the estimator $\bar{\mu}^n$ is easily found to be

$$\bar{\sigma}^{2,n} = \frac{1}{n}\hat{\sigma}^{2,n}.$$

Note that as $n \to \infty$, $\bar{\sigma}^{2,n} \to 0$, but $\hat{\sigma}^{2,n} \to \sigma^2$ where σ^2 is the true variance of W. If σ^2 is known, there would be no need to compute $\hat{\sigma}^{2,n}$ and $\bar{\sigma}^{2,n}$ would be given as above with $\hat{\sigma}^{2,n} = \sigma^2$.

We can write these expressions recursively using

$$\bar{\mu}^n = \left(1 - \frac{1}{n}\right)\bar{\mu}^{n-1} + \frac{1}{n}W^n, \tag{2.4}$$

$$\hat{\sigma}^{2,n} = \begin{cases} \frac{1}{n}(W^n - \bar{\mu}^{n-1})^2 & n = 2, \\ \frac{n-2}{n-1}\hat{\sigma}^{2,n-1} + \frac{1}{n}(W^n - \bar{\mu}^{n-1})^2 & n > 2. \end{cases} \tag{2.5}$$

We will often speak of our belief state which captures what we know about the parameters we are trying to estimate. Given our observations, we would write our belief state as

$$B^{freq,n} = (\bar{\mu}^n, \hat{\sigma}^{2,n}, n).$$

Equations (2.4) and (2.5) describe how our belief state evolves over time.

The belief state is supposed to communicate a probability distribution as opposed to statistics such as mean and variance. When we are forming an average, we can apply the law of large numbers and assume that our estimate $\bar{\mu}^n$ is approximately normally distributed. This is true exactly if W is normally distributed, but it is generally a very good approximation even when W is described by other distributions.

2.3 THE BAYESIAN VIEW

The Bayesian perspective casts a different interpretation on the statistics we compute which is particularly useful in certain problem settings in optimal learning. While the frequentist perspective starts with no knowledge about a function or response (which is the case in many settings), the Bayesian perspective leverages prior knowledge that comes from past experience or an understanding of the dynamics of a problem.

For example, let x be the price of a product, and let μ_x be the sales when we offer the product at a price x. With frequentist learning, we start knowing nothing about μ_x, but as we collect data to construct estimates $\bar{\mu}^n_x$ of mu_x, we may find that $\bar{\mu}^n_x$ does not necessarily decline with x. In a Bayesian model, we would at least start with a prior $\bar{\mu}^0_x$ where this is the case.

The biggest difference between Bayesian and frequentist learning is the Bayesian prior, which is the initial distribution of belief about the truth. The prior is how we are able to communicate domain knowledge, if this is available. A prior is critical in the context of virtually any problem where experiments are really expensive (such as laboratory or field experiments), or when there is significant risk, as in health decisions.

The second difference between the two perspectives is the view of the truth. In the Bayesian view, the truth μ_x is handled as a random variable. For example, we might model μ_x as being normally distributed with mean $\bar{\mu}^0_x$ and variance $\bar{\sigma}^{2,0}_x$. As we collect information, this distribution changes, where we can guarantee that the variance will steadily shrink. In the frequentist perspective, the truth is an unknown number, and our estimate of this number is a random variable that reflects the variation in our observations. We understand, for example, that if 100 people all collected data to produce an estimate, each of these 100 estimates would be different, and could be described by some distribution (often normal).

The Bayesian perspective is well suited to information collection in settings where information is expensive (or where there is a significant risk involved if we make the wrong decision). It is in these settings where there is the incentive to bring in prior knowledge, but these are also the settings where we are likely to have domain knowledge. By contrast, there are settings within search algorithms, or on the internet, where information is relatively much easier to obtain, and while this does not mean it is free, it is often easier to run a set of experiments to form initial estimates than to depend on a prior from an exogenous source.

We note a subtle change in notation from the frequentist perspective, where $\bar{\mu}^n$ was our statistic giving our estimate of μ. In the Bayesian view, we let $\bar{\mu}^n$ be our estimate of the mean of the random variable μ after we have made n observations. It is important to remember that μ is a random variable whose distribution reflects our prior belief about μ. The parameter $\bar{\mu}^0$ is not a random variable. This is our initial estimate of the mean of our prior distribution. By contrast, $\bar{\mu}^n$, for $n \geq 1$, is a random variable for the same reason that $\bar{\mu}^n$ is random in the frequentist view: $\bar{\mu}^n$ is computed from a series of random observations $W^1, W^2, \ldots, W^n$, and therefore the distribution of $\bar{\mu}^n$ reflects the distribution of all of our experiments. However, in the Bayesian perspective we are primarily interested in the mean and variance of μ.

Below we first use some simple expressions from probability to illustrate the effect of collecting information. We then give the Bayesian version of (2.4) and (2.5) for the case of independent beliefs, where observations of one choice do not influence our beliefs about other choices. We follow this discussion by giving the updating equations for correlated beliefs, where an observation of μ_x for alternative x tells us something about $\mu_{x'}$. We round out our presentation by touching on other important types of distributions.

2.3.1 The updating equations for independent beliefs

We begin by assuming (as we do through most of our presentation) that our random variable W is normally distributed. Let σ_W^2 be the variance of W, which captures the noise in our ability to observe the true value. To simplify the algebra, we define the *precision* of W as

$$\beta^W = \frac{1}{\sigma_W^2}.$$

Precision has an intuitive meaning: smaller variance means that the observations will be closer to the unknown mean, that is, they will be more precise. Now let $\bar{\mu}^n$ be our estimate of the true mean μ after n observations, and let β^n be the precision of this estimate. That is, having already seen the values $W^1, W^2, ..., W^n$, we believe that the mean of μ is $\bar{\mu}^n$, and the variance of μ is $\frac{1}{\beta^n}$. We say that we are "at time n" when this happens; note that all quantities that become known at time n are indexed by the superscript n, so the observation W^{n+1} is not known until time $n + 1$. Higher precision means that we allow for less variation in the unknown quantity, that is, we are more sure that μ is equal to $\bar{\mu}^n$. After observing W^{n+1}, the updated mean and precision of our estimate of μ is given by

$$\bar{\mu}^{n+1} = \frac{\beta^n \bar{\mu}^n + \beta^W W^{n+1}}{\beta^n + \beta^W}, \tag{2.6}$$

$$\beta^{n+1} = \beta^n + \beta^W. \tag{2.7}$$

Equation (2.6) can be written more compactly as

$$\bar{\mu}^{n+1} = (\beta^{n+1})^{-1} \left(\beta^n \bar{\mu}^n + \beta^W W^{n+1}\right). \tag{2.8}$$

There is another way of expressing the updating which provides insight into the structure of the flow of information. First define

$$\tilde{\sigma}^{2,n} = \mathit{Var}^n[\bar{\mu}^{n+1}] \tag{2.9}$$
$$= \mathit{Var}^n[\bar{\mu}^{n+1} - \bar{\mu}^n] \tag{2.10}$$

where $\mathit{Var}^n[\cdot] = Var[\cdot \,|\, W^1, ..., W^n]$ denotes the variance of the argument given the information we have through n observations. For example,

$$\mathit{Var}^n[\bar{\mu}^n] = 0$$

since, given the information after n observations, $\bar{\mu}^n$ is a number that we can compute deterministically from the prior history of observations.

The parameter $\tilde{\sigma}^{2,n}$ can be described as the variance of $\bar{\mu}^{n+1}$ given the information we have collected through iteration n, which means the only random variable is W^{n+1}. Equivalently, $\tilde{\sigma}^{2,n}$ can be thought of as the *change* in the variance of $\bar{\mu}^n$ as a result of the observation of W^{n+1}. Equation (4.6) is an equivalent statement since, given the information collected up through iteration n, $\bar{\mu}^n$ is deterministic and is therefore a constant. We use equation (4.6) to offer the interpretation that $\tilde{\sigma}^{2,n}$ is the *change* in the variance of our estimate of the mean of μ.

It is possible to write $\tilde{\sigma}^{2,n}$ in different ways. For example, we can show that

$$\tilde{\sigma}^{2,n} = \bar{\sigma}^{2,n} - \bar{\sigma}^{2,n+1} \tag{2.11}$$
$$= \frac{(\bar{\sigma}^{2,n})}{1 + \sigma^2_W/\bar{\sigma}^{2,n}} \tag{2.12}$$
$$= (\beta^n)^{-1} - (\beta^n + \beta^W)^{-1}. \tag{2.13}$$

The proof of (2.11) is given in Section 2.8.1. Equations (2.12) and (2.13) come directly from (2.11) and (2.7), using either variances or precisions.

Just as we let $\mathit{Var}^n[\cdot]$ be the variance given what we know after n experiments, let $\mathbb{E}^n$ be the expectation given what we know after n experiments. That is, if $W^1, \ldots, W^n$ are the first n experiments, we can write

$$\mathbb{E}^n \bar{\mu}^{n+1} \equiv \mathbb{E}(\bar{\mu}^{n+1} | W^1, \ldots, W^n) = \bar{\mu}^n.$$

We note in passing that $\mathbb{E}\bar{\mu}^{n+1}$ refers to the expectation before we have made any experiments, which means $W^1, \ldots, W^n$ are all random, as is W^{n+1}. By contrast, when we compute $\mathbb{E}^n \bar{\mu}^{n+1}$, $W^1, \ldots, W^n$ are assumed fixed, and only W^{n+1} is random. By the same token, $\mathbb{E}^{n+1}\bar{\mu}^{n+1} = \bar{\mu}^{n+1}$, where $\bar{\mu}^{n+1}$ is some number which is fixed because we assume that we already know $W^1, \ldots, W^{n+1}$. It is important to realize that when we write an expectation, we have to be explicit about what is random (that is, what variable(s) we are averaging over), and what we are conditioning on.

Using this property, we can write $\bar{\mu}^{n+1}$ in a different way that brings out the role of $\tilde{\sigma}^n$. Assume that we have made n observations and let

$$Z = \frac{\bar{\mu}^{n+1} - \bar{\mu}^n}{\tilde{\sigma}^n}.$$

We note that Z is a random variable only because we have not yet observed W^{n+1}. Normally we would index $Z = Z^{n+1}$ since it is a random variable that depends on W^{n+1}, but we are going to leave this indexing implicit. It is easy to see that $\mathbb{E}^n Z = 0$ and

$Var^n Z = 1$. Also, since W^{n+1} is normally distributed, then Z is normally distributed, which means $Z \sim N(0,1)$. This means that we can write

$$\bar{\mu}^{n+1} = \bar{\mu}^n + \tilde{\sigma}^n Z. \tag{2.14}$$

Equation (2.14) makes it clear how $\bar{\mu}^n$ evolves over the observations. It also reinforces the idea that $\tilde{\sigma}^n$ is the change in the variance due to a single observation.

Equations (2.6) and (2.7) are the Bayesian counterparts of (2.4) and (2.5), although we have simplified the problem a bit by assuming that the variance of W is known. The belief state in the Bayesian view (with normally distributed beliefs) is given by

$$B^{Bayes,n} = (\bar{\mu}^n, \beta^n).$$

As we show below in Section 2.8.2, if our prior belief about μ is normally distributed with mean $\bar{\mu}^n$ and precision β^n, and if W is normally distributed, then our posterior belief after $n+1$ observations is also normally distributed with mean $\bar{\mu}^{n+1}$ and precision β^{n+1}. We often use the term *Gaussian prior*, when we want to say that our prior is normally distributed. We also allow ourselves a slight abuse of notation: we use $\mathcal{N}\left(\mu, \sigma^2\right)$ to mean a normal distribution with mean μ and variance σ^2, but we also use the notation $\mathcal{N}\left(\mu, \beta\right)$ when we are using the precision instead of the variance.

Needless to say, it is especially convenient if the prior distribution and the posterior distribution are of the same basic type. When this is the case, we say that the prior and posterior are *conjugate*, or that they are a *conjugate family*. This happens in a few special cases when the prior distribution and the distribution of W are chosen in a specific way.

The property that the posterior distribution is in the same family as the prior distribution is called *conjugacy*. The normal distribution is unusual in that the conjugate family is the same as the sampling family (the distribution of the experiment W is also normal). For this reason, this class of models is sometimes referred to as the "normal-normal" model (this phraseology becomes clearer below when we discuss other combinations).

In some cases, we may impose conjugacy as an approximation. For example, it might be the case that the prior distribution on μ is normal, but the distribution of the observation W is not normal (for example, it might be nonnegative). In this case, the posterior may not even have a convenient analytical form. But we might feel comfortable approximating the posterior as a normal distribution, in which case we would simply use (2.54)-(2.55) to update the mean and variance and then assume that the posterior distribution is normal.

2.3.2 Updating for correlated normal priors

A particularly important problem class in optimal learning involves problems where there are multiple choices, where our beliefs about the choices are correlated. Some examples of correlated beliefs are as follows:

- We are interested in finding the price of a product that maximizes total revenue. We believe that the function $R(p)$ that relates revenue to price is continuous. Assume that we set a price p^n and observe revenue R^{n+1} that is higher than we had expected. If we raise our estimate of the function $R(p)$ at the price p^n, our beliefs about the revenue at nearby prices should be higher.

- We choose five people for the starting lineup of our basketball team and observe total scoring for one period. We are trying to decide if this group of five people is better than another lineup that includes three from the same group with two different people.

If the scoring of these five people is higher than we had expected, we would probably raise our belief about the other group, since there are three people in common.

- A physician is trying to treat diabetes using a treatment of three drugs, where she observes the drop in blood sugar from a course of a particular treatment. If one treatment produces a better-than-expected response, this would also increase our belief of the response from other treatments that have one or two drugs in common.

- We are trying to find the highest concentration of a virus in the population. If the concentration of one group of people is higher than expected, our belief about other groups that are close (either geographically, or due to other relationships) would also be higher.

Correlated beliefs are a particularly powerful device in optimal learning, allowing us to generalize the results of a single observation to other alternatives that we have not directly measured.

Let $\bar{\mu}_x^n$ be our belief about alternative x after n experiments. Now let

$$Cov^n(\mu_x, \mu_y) \quad = \quad \text{the covariance in our belief about } \mu_x \text{ and } \mu_y.$$

We let Σ^n be the covariance matrix, with element $\Sigma^n_{xy} = Cov^n(\mu_x, \mu_y)$. Just as we defined the precision β_x^n to be the reciprocal of the variance, we are going to define the precision matrix B^n to be

$$B^n = (\Sigma^n)^{-1}.$$

Let e_x be a column vector of zeroes with a 1 for element x, and as before we let W^{n+1} be the (scalar) observation when we decide to measure alternative x. We could label W^{n+1} as W_x^{n+1} to make the dependence on the alternative more explicit. For this discussion, we are going to use the notation that we choose to measure x^n and the resulting observation is W^{n+1}. If we choose to measure x^n, we can also interpret the observation as a column vector given by $W^{n+1}e_{x^n}$. Keeping in mind that $\bar{\mu}^n$ is a column vector of our beliefs about the expectation of μ, the Bayesian equation for updating this vector in the presence of correlated beliefs is given by

$$\bar{\mu}^{n+1} \quad = \quad (B^{n+1})^{-1}\left(B^n\bar{\mu}^n + \beta^W W^{n+1}e_{x^n}\right), \tag{2.15}$$

where B^{n+1} is given by

$$B^{n+1} \quad = \quad (B^n + \beta^W e_{x^n}(e_{x^n})^T). \tag{2.16}$$

Note that $e_x(e_x)^T$ is a matrix of zeroes with a one in row x, column x, whereas β^W is a scalar giving the precision of our experimental outcome W.

It is possible to perform these updates without having to deal with the inverse of the covariance matrix. This is done using a result known as the Sherman-Morrison formula. If A is an invertible matrix (such as Σ^n) and u is a column vector (such as e_x), the Sherman-Morrison formula is

$$[A + uu^T]^{-1} = A^{-1} - \frac{A^{-1}uu^TA^{-1}}{1 + u^TA^{-1}u}. \tag{2.17}$$

Let $\lambda^W = \sigma_W^2 = 1/\beta^W$ be the variance of our experimental outcome W^{n+1}. We are going to simplify our notation by assuming that our experimental variance is the same across all alternatives x, but if this is not the case, we can replace λ^W with λ^W_x throughout. Using the Sherman-Morrison formula, and letting $x = x^n$, we can rewrite the updating equations as

$$\begin{aligned} \bar{\mu}^{n+1}(x) &= \bar{\mu}^n + \frac{W^{n+1} - \bar{\mu}^n_x}{\lambda^W + \Sigma^n_{xx}} \Sigma^n e_x, & (2.18) \\ \Sigma^{n+1}(x) &= \Sigma^n - \frac{\Sigma^n e_x (e_x)^T \Sigma^n}{\lambda^W + \Sigma^n_{xx}}. & (2.19) \end{aligned}$$

where we express the dependence of $\bar{\mu}^{n+1}(x)$ and $\Sigma^{n+1}(x)$ on the alternative x which we have chosen to measure.

To illustrate, assume that we have three alternatives with mean vector

$$\bar{\mu}^n = \begin{bmatrix} 20 \\ 16 \\ 22 \end{bmatrix}.$$

Assume that $\lambda^W = 9$ and that our covariance matrix Σ^n is given by

$$\Sigma^n = \begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix}.$$

Assume that we choose to measure $x = 3$ and observe $W^{n+1} = W_3^{n+1} = 19$. Applying equation (2.18), we update the means of our beliefs using

$$\begin{aligned} \bar{\mu}^{n+1}(3) &= \begin{bmatrix} 20 \\ 16 \\ 22 \end{bmatrix} + \frac{19 - 22}{9 + 15} \begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 20 \\ 16 \\ 22 \end{bmatrix} + \frac{-3}{24} \begin{bmatrix} 3 \\ 4 \\ 15 \end{bmatrix} \\ &= \begin{bmatrix} 19.625 \\ 15.500 \\ 20.125 \end{bmatrix}. \end{aligned}$$

The update of the covariance matrix is computed using

$$
\begin{aligned}
\Sigma^{n+1}(3) &= \begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix} - \frac{\begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}[0\ 0\ 1]\begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix}}{9+15} \\
&= \begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix} - \frac{1}{24}\begin{bmatrix} 3 \\ 4 \\ 15 \end{bmatrix}[3\ 4\ 15] \\
&= \begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix} - \frac{1}{24}\begin{bmatrix} 9 & 12 & 45 \\ 12 & 16 & 60 \\ 45 & 60 & 225 \end{bmatrix} \\
&= \begin{bmatrix} 12 & 6 & 3 \\ 6 & 7 & 4 \\ 3 & 4 & 15 \end{bmatrix} - \begin{bmatrix} 0.375 & 0.500 & 1.875 \\ 0.500 & 0.667 & 2.500 \\ 1.875 & 2.500 & 9.375 \end{bmatrix} \\
&= \begin{bmatrix} 11.625 & 5.500 & 1.125 \\ 5.500 & 6.333 & 1.500 \\ 1.125 & 1.500 & 5.625 \end{bmatrix}.
\end{aligned}
$$

These calculations are fairly easy, which means we can execute them even if we have thousands of alternatives. But we will run up against the limits of computer memory if the number of alternatives is in the 10^5 range or more, which arises when we consider problems where an alternative x is itself a multidimensional vector.

2.3.3 Bayesian updating with an uninformative prior

What if we truly have no prior information about a parameter before we start collecting information? We can use what is known in the Bayesian statistics literature as an uninformative prior, which is equivalent to a normal density with infinite variance (or zero precision). We note that it is not necessary for the prior to be a true density (which integrates to 1.0). For example, our prior on a random variable x can be $f(x) = .01$ for all $-\infty < x < \infty$, which of course integrates to infinity. This is known as an *improper prior*.

When we look at the Bayesian updating equations in (2.6) and (2.7), we see that if we use $\beta^0 = 0$, then it simply means that we put no weight on the initial estimates. It is easy to see that if $\beta^0 = 0$, then $\bar{\mu}^1 = W^1$ (the first observation) and $\beta^1 = \beta^W$ (the precision of our experiments).

The problem with uninformative priors is that we have no guidance at all regarding our first experiment x^0. Fortunately, in most applications of information collection we start with some prior knowledge, but this is not always the case.

Another strategy we can use if we have no prior information is known as *empirical Bayes*. Put simply, empirical Bayes requires that we collect a small initial sample, and then use the results of this sample to form a “prior” (obviously, this “prior” is formed *after* we calculate our small sample, but before we do any guided learning). Empirical Bayes sounds like a different name for frequentist (essentially our prior belief is created using a frequentist procedure), but the interpretation of what is random is different than with a frequentist model. The main distinguishing feature of the Bayesian approach is that it puts a number on the likelihood that the unknown value takes on a certain value or falls within a particular interval.

2.4 BAYESIAN UPDATING FOR SAMPLED NONLINEAR MODELS

There will be many instances where we need to work with models that are nonlinear in the parameter vector θ. Some examples are

---

■ EXAMPLE 2.1

Pricing a product on the internet - Imagine that we set a price x^n for day n and then observe total revenue Y_t. We might model the demand for the product using

$$D(x|\theta) = \theta_1 e^{-\theta_2(x-\theta_3)}.$$

The revenue would then be given by $f(x|\theta) = xD(x|\theta)$. We might set a price x^n for day x and then observe revenue Y^n. After N days, we have a dataset $(Y^n, x^n), n = 1, \ldots, N$.

■ EXAMPLE 2.2

Assume a physical has to choose from a set of treatments $\mathcal{X} = \{x_1, \ldots, x_M\}$ for reducing blood sugar. We are interested in estimating the probability that a treatment x will be judged a success.

We might begin by constructing a utility function that captures the attractiveness of the product as a function of price p, which we might write as

$$U(x|\theta) = \theta_0 + \sum_{m=1}^{M} \theta_m \mathbb{1}_{x=x_m}.$$

where

$$\mathbb{1}_{x=x_m} = \begin{cases} 1 & \text{If } x = x_m \\ 0 & \text{Otherwise.} \end{cases}$$

Now assume that we model the probability of the treatment x being effective is given by

$$P(x) = \frac{e^{U(x|\theta)}}{1 + e^{U(x|\theta)}}.$$

Normally the estimation of nonlinear models can become quite complex, but we are going to use a technique known as a *sampled belief model*. We assume that we are given a nonlinear function $f(x|\theta)$ where x is the input (in our examples this was either a price or medical treatment), and then we observe a response Y. We then assume that θ is one of the set $\Theta = \{\theta_1, \ldots, \theta_K\}$, initially with probability $P[\theta = \theta_k] = p^0_k$. We refer to $p^0 = (p^0_k), k = 1, \ldots, K$ as the prior.

The next step is to design the updating equations for the probability vector p^n after running an experiment with $x = x^n$ and observing a response y^{n+1}. We start with Bayes theorem

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

Now interpret the event A as the event that $\theta = \theta_k$, and B as the new information y^{n+1}. We are going to let H^n be the history of our experiments where

$$H^n = (S^0, x^0, \hat{y}^1, x^1, \hat{y}^2, \ldots, x^{n-1}, \hat{y}^n).$$

We can write our belief probabilities as

$$p_k^n = \mathbb{P}[\theta = \theta_k | H^n].$$

We note that we can write Bayes' theorem where all of the probabilities are conditioned on a third event C, as in

$$\mathbb{P}(A|B,C) = \frac{\mathbb{P}(B|A,C)\mathbb{P}(A|C)}{\mathbb{P}(B|C)}.$$

In our setting, the new event C is our history H^n. Adapting this to our setting gives us

$$\mathbb{P}[\theta = \theta_k | y^{n+1} = y, H^n] = \frac{\mathbb{P}[\hat{y}^{n+1} = y|\theta_k, H^n]\mathbb{P}[\theta = \theta_k|H^n]}{\mathbb{P}[\hat{y}^{n+1} = y|H^n]}. \tag{2.20}$$

Let $f^y(\hat{y} = y|\theta)$ be the distribution of the random observation $\hat{y}$ given θ, or

$$f^y(\hat{y} = y|\theta) = \mathbb{P}[\hat{y} = y|\theta].$$

The conditioning on the history H^n affects the probability that θ takes on a particular value. We can rewrite equation (2.20) as

$$p_k^{n+1} \quad = \quad \frac{f^y(\hat{y}^{n+1} = y|\theta_k)p_k^n}{f^y(\hat{y}^{n+1} = y)} \tag{2.21}$$

where

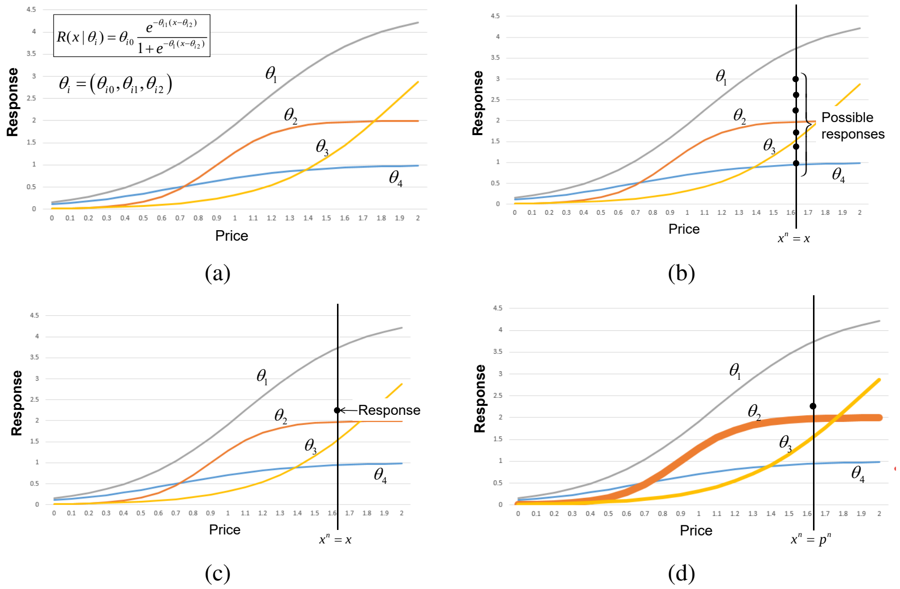
$$f^y(\hat{y}^{n+1} = y) = \sum_{k=1}^{K} f^y(\hat{y}^{n+1} = y|\theta_k)p_k^n. \tag{2.22}$$

We are exploiting the fact that the distribution of $\hat{y}^{n+1}$ depends only on θ, so we can drop the dependence on H^n when we are also conditioning on θ, since the history only affects the probabilities p_k^n.

We again note that we assume that $f^y(y^{n+1}|\theta_k)$ is a known distribution that can be easily computed from the structure of the problem. For example, it might be a normal density, a Poisson distribution, or a logistic regression.

Equation 2.21 is quite easy to compute (given $f^y(y^{n+1}|\theta_k)$) when we use a sampled belief model. Without the sampled belief model, the expectation in equation (2.22) can become quite complex (imagine what happens when θ is a vector). By contrast, this equation is easy to compute for the sampled belief model, even if θ is a high-dimensional vector.

The process is illustrated in figure 2.1 for a problem where we are trying to show the market response to the price using a logistics curve. The figure shows (a) the initial, uniform prior of a sampled belief model with four possible values of θ, (b) the possible outcomes from an experiment x, (c) the actual result of an experiment, and (d) the posterior distribution which are indicated by the thickness of the lines.



(a)

(b)

(c)

(d)

Figure 2.1 Illustration of Bayesian updating for a sampled belief model. Starting from top left: a) initial uniform prior, b) set of potential outcomes from an experiment x, c) the response $\hat{y}^{n+1}$, d) the posterior (width proportional to probability).

2.5 THE EXPECTED VALUE OF INFORMATION

The previous section described how we update the mean and variance (or precision) for a particular observation of W. It is useful to see what happens when we look at the *expected* change in the mean and variance when we average over all possible realizations.

Let μ be the unknown true value of some quantity. Then, μ is a random variable with (assumed) mean $\bar{\mu}^0$ and variance $\bar{\sigma}^{2,0}$. Let W be a random observation of whatever we are measuring, from which we might update our estimate of the mean and variance. Note that

$$\mathbb{E}W = \mathbb{E}_\mu \mathbb{E}_{W|\mu}(W \mid \mu) = \mathbb{E}\mu = \bar{\mu}^0.$$

We observe that

$$\begin{aligned} \mathbb{E}\mu &= \bar{\mu}^0, \\ Var(\mu) &= \bar{\sigma}^{2,0}. \end{aligned}$$

Assume that we observe $W = w$. For example, we might assume that the travel time on a link has mean $\bar{\mu}^0 = 22$ minutes, but we observe an actual travel time of $W = w = 27$ minutes. Then, $\mathbb{E}(\mu|W = w)$ would be the updated mean, and $Var(\mu|W = w)$ would be the updated variance.

Now let's consider what happens *on average* to our estimates of the mean and variance when we consider all possible outcomes of our observation of W. Let $\mathbb{E}_\mu\mu$ be the expected value of μ (over our density function for μ), which is equal to $\bar{\mu}^0$. Now let $\mathbb{E}_\mu(\mu|W)$ be the expected value of μ given a particular observation of W. The expectation $\mathbb{E}_\mu(\mu|W)$ is a random variable (since it depends on W), but we are interested in its expectation over all possible values of W, which we can write as $\mathbb{E}_W\mathbb{E}_\mu(\mu|W)$. We can compute this by taking expectations of equation (2.8) given what we know at iteration n (which means that only W^{n+1} is random). We start by observing that

$$\mathbb{E}_\mu(\mu|W) = \bar{\mu}^1 = (\beta^1)^{-1}\left(\beta^0\bar{\mu}^n + \beta^W W\right),$$

where $\beta^1 = \beta^0 + \beta^W$. We then take expectations of both sides over the random observation W to obtain

$$\begin{aligned}\mathbb{E}_W\bar{\mu}^1 &= \mathbb{E}_W\left((\beta^1)^{-1}\left(\beta^0\bar{\mu}^0 + \beta^W W\right)\right)\\ &= (\beta^1)^{-1}\left(\beta^0\bar{\mu}^0 + \beta^W\mathbb{E}W\right)\\ &= (\beta^1)^{-1}\left(\beta^0\bar{\mu}^0 + \beta^W\bar{\mu}^0\right)\\ &= (\beta^1)^{-1}\left(\beta^0 + \beta^W\right)\bar{\mu}^0\\ &= \bar{\mu}^0.\end{aligned}$$

This result seems to be saying that collecting an observation of W does not change our belief of the true mean μ. This is not the case. As we saw in the previous section, a particular observation of W will, in fact, change our belief about the mean of μ. But if we look at all possible realizations of W, before the observation occurs, *on average* our estimate of the mean does not change.

This simple equation provides an insight into priors and learning. Imagine if

$$\mathbb{E}\mu = \mathbb{E}_W\mathbb{E}_\mu(\mu|W) = \bar{\mu}^0 + a.$$

That is, observing W will, on average, shift our belief about the true mean from $\bar{\mu}^0$ to $\bar{\mu}^0 + a$, where a is a constant which would have to be known before we run our experiment. If this were true (for a other than zero), then this would mean that our initial estimate $\bar{\mu}^0$ is not a true prior. That is, we could shift our prior by a so that it becomes an unbiased estimate of the mean.

Now consider what happens to the variance after an experiment. We use some basic relationships from probability to obtain

$$\begin{aligned}Var(\bar{\mu}) &= \mathbb{E}_\mu(\bar{\mu}^2) - (\mathbb{E}_\mu(\mu))^2 && (2.23)\\ &= \mathbb{E}_\mu(\bar{\mu}^2) - (\mathbb{E}_\mu(\mu))^2 + (\mathbb{E}_\mu(\mu))^2 - (\mathbb{E}_\mu(\mu))^2 && (2.24)\\ &= \mathbb{E}_W[(\mathbb{E}_\mu(\bar{\mu}^2|W))] - \mathbb{E}_W[(\mathbb{E}_\mu(\mu|W))^2] + \mathbb{E}_W[(\mathbb{E}_\mu(\mu|W))^2] - (\mathbb{E}_W[\mathbb{E}_\mu(\mu|W)])^2 && (2.25)\\ &= \mathbb{E}_W[(\mathbb{E}_\mu(\bar{\mu}^2|W)) - (\mathbb{E}_\mu(\mu|W))^2] + \mathbb{E}_W[(\mathbb{E}_\mu(\mu|W))^2] - (\mathbb{E}_W[\mathbb{E}_\mu(\mu|W)])^2 && (2.26)\\ &= \mathbb{E}_W[Var(\mu|W)] + Var[\mathbb{E}_\mu(\mu|W)]. && (2.27)\end{aligned}$$

Equation (2.23) is the definition of the variance. In (2.24), we add and subtract $(\mathbb{E}_\mu(\mu|W))^2$. In (2.25), we then condition throughout on W and then take the expectation over W. In (2.26), we pull the $\mathbb{E}_W$ out front of the first two terms, setting up the final (and classic) result given by equation (2.28). Our interest is primarily in equation (2.26). Above, we pointed out that $\mathbb{E}_W \mathbb{E}_\mu(\mu|W) = \mathbb{E}_\mu \mu$. This is not the case with the variance, where equation (2.28) tells us that

$$\mathbb{E}_W[\mathit{Var}(\mu|W)] = \mathit{Var}(\mu) - \mathit{Var}[\mathbb{E}_\mu(\mu|W)]. \tag{2.28}$$

This means that the variance after an experiment will, on average, always be smaller than the original variance. Of course, it might be the case that $\mathit{Var}[\mathbb{E}_\mu(\mu|W)] = 0$. This would happen if W were an irrelevant piece of information. For example, assume that μ is our estimate of the travel time on a path in a network, and W is an observation of the change in the S&P stock index yesterday. The S&P stock index does not tell us anything about the travel time on the path, which means that $\mathbb{E}_\mu(\mu|W)$ is a constant (in our example, it would be $\bar{\mu}^0$). Clearly, $\mathit{Var}(\bar{\mu}^0) = 0$, since $\bar{\mu}^0$ is not a random variable (it is just a number).

We collect observations one at a time. So, the above discussion continues to apply after we observe $W^1, W^2, ..., W^n$. Since the posterior mean $\bar{\mu}^n$ is a known, fixed quantity at time n, we can simply view it as a new prior. Our problem essentially restarts after each observation, just with different parameters for our distribution of belief. The advantage of recursive updating is that it allows us to turn a problem with a long-time horizon into a sequence of small problems – a concept that will also inform our solution techniques in later chapters.

2.6 UPDATING FOR NON-GAUSSIAN PRIORS

So far, we have considered a setting where the random observations W^n are assumed to come from a normal distribution, whose mean is the true value that we wish to estimate. We have also used a normal distribution to describe our prior belief about the unknown value. These are two very different normal distributions, and it is important to distinguish between them. We use the term "sampling distribution" to refer to the distribution of the observations W^n. This distribution depends on certain unknown parameters, like the value μ in the preceding section. Although we do not know this distribution, we have a way of obtaining samples from it.

By contrast, the term "prior distribution" describes the distribution of our own beliefs about the unknown parameters of the sampling distribution. Unlike the sampling distribution, which is determined by nature and unknown to us, the prior distribution is something that we construct to encode our own uncertainty about the truth. The prior distribution changes as we accumulate observations, reflecting the changes in our beliefs and the reduction in our uncertainty that result from collecting new information.

In the preceding section, both the sampling distribution and the prior distribution are assumed to be normal. This model is particularly intuitive and versatile because the unknown parameter in the sampling distribution is precisely the mean of the sampled observations. The mean $\bar{\mu}^0$ of the prior distribution can easily be interpreted as our "best guess" as to the unknown value, with σ^0 representing our uncertainty. Thus, whenever we have a rough idea of what the unknown mean might be, a normal distribution provides a very intuitive and understandable way to encode that idea in a mathematical model.

Unfortunately, the normal-normal model (normal prior, normal samples) is not always suitable. For one thing, a normal sampling distribution assumes that our random observations can take on any real value. In reality, this might not be the case: for instance, we might be observing the waiting times of customers in a service center, where we are trying to estimate the service rate. Waiting times are always positive, so an exponential distribution would seem to be a more appropriate choice than a normal distribution. Even more troubling is a situation where our observations are obviously discrete. For instance, we might be observing the success or failure of a medical test, where the outcome is 0 or 1. A normal distribution is certainly a poor choice for a sampling model in this setting.

The normal distribution is not always the best choice for a prior, either. For example, if we are observing exponentially distributed service times, the service rate is necessarily a strictly positive number. If we were to put a normal prior on the service rate, then even with a positive prior mean, we would essentially be allowing the possibility that our exponential sampling distribution has a negative rate, which is impossible. Our uncertainty about the service rate should be encoded using a different kind of distribution that accounts for the fact that the rate must be positive. In the case of 0/1 observations (success or failure of a test), the sample mean is the probability of success. We know that this number must be between 0 and 1, so a normal distribution is again not the best choice to represent our beliefs.

In this section, we discuss several other possible learning models where the sampling and prior distributions are not normal. However, all the distributions that we consider will retain the conjugacy property, which means that the posterior distribution is of the same type as the prior distribution.

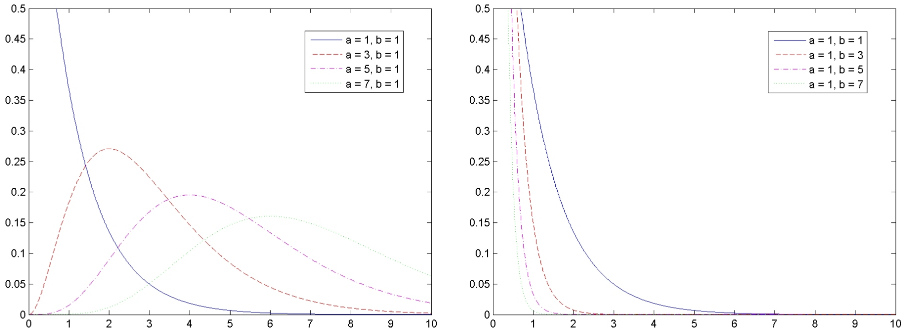
2.6.1 The gamma-exponential model

The gamma-exponential model is one possible choice for a situation where the observations are continuous and positive. Suppose that we are trying to estimate the service time distribution at a car repair shop by combining our prior belief with observations of actual service times. We feel comfortable assuming that the sampling distribution governing the service times is exponential with parameter λ. The service rate is the unknown value that we wish to estimate.

Since λ is itself unknown, we view it as a random variable. Clearly it is not appropriate to assume that it follows a normal distribution, since this would mean that we believe that λ might be negative. Assume instead that λ comes from a gamma distribution with parameters a and b. This distribution is given by

$$f(x|a,b) = b(bx)^{a-1}\frac{e^{-bx}}{\Gamma(a)}$$

where a is typically an integer (as it will be for our applications) and $\Gamma(a) = (a-1)!$. Figure 2.2 illustrates several examples of the gamma density. If $a = 1$, the gamma is an exponential distribution. For $a > 1$, it takes on a skewed shape which approaches a normal distribution for larger values of a.



2.2(a) Varying a

2.2(b) Varying b

Figure 2.2 Illustration of a family of gamma distributions a) varying a and b) varying b.

The mean of this distribution is given by

$$\mathbb{E}(\lambda) = \frac{a^0}{b^0}. \tag{2.29}$$

The quantities a^0 and b^0 should be chosen by us in such a way so that (2.29) represents our initial beliefs about the service rate.

We can let a^n/b^n be our estimate of λ after n observations, as before. After observing W^{n+1}, we update our beliefs using the equations

$$a^{n+1} = a^n + 1, \tag{2.30}$$

$$b^{n+1} = b^n + W^{n+1}. \tag{2.31}$$

Our belief about λ is that it follows a gamma distribution with parameters a^{n+1} and b^{n+1}. Despite the complexity of the gamma density, the updating equations governing the way in which we learn about λ are actually quite simple.

Equations (2.30) and (2.31) give a simple explanation of the gamma prior that makes the parameters a^n and b^n seem a bit less mysterious. Essentially, b^n is the sum of the first n service times (plus some prior constant b^0), whereas a^n is roughly equal to n (again, plus a prior constant). Thus, after n observations, our estimate of the service rate is given by

$$\mathbb{E}\left(\lambda \,|\, W^1, ..., W^n\right) = \frac{a^n}{b^n}.$$

This estimate is roughly the number of customers that were served per single unit of time. This is precisely the meaning of a service rate.

While the gamma-exponential model (gamma prior, exponential sampling distribution) is useful for modeling problems with continuous, positive observations, it is incapable of handling correlated beliefs. There is no easy multivariate analog for the gamma distribution, the way there is with the normal distribution, and thus no analog of the correlated normal updating equations (2.18)-(2.19). In a setting where there are multiple unknown values with heavy correlations, it is important to consider the tradeoff between using a multivariate normal model to capture the correlations, and using a different type of model to more accurately represent the individual distributions of the alternatives.

2.6.2 The gamma-Poisson model

The gamma-Poisson model is similar to the gamma-exponential model, but the observations are now assumed to be discrete. For example, we may now be interested in the arrival rate of customers to the car repair shop, rather than the service time. Suppose that the total number of customers N that visit the shop in a single day follows a Poisson distribution with rate λ customers per day. Our observations are now the actual numbers of customers that arrive on different days. If the arrival rate is λ, the distribution of N follows the Poisson distribution given by

$$\mathbb{P}[N = x] = \frac{\lambda^x e^{-\lambda}}{x!},$$

where $x = 0, 1, \ldots$. The problem is that we do not know λ, and we wish to estimate it from observations N^n where N^n is the observed number of arrivals on the n^{th} day.

Once again, we assume that λ comes from a gamma distribution with parameters a^0 and b^0. The prior distribution changes after each observation according to the equations

$$\begin{aligned} a^{n+1} &= a^n + N^{n+1}, && (2.32)\\ b^{n+1} &= b^n + 1. && (2.33) \end{aligned}$$

After n observations, our estimate of the Poisson rate,

$$\mathbb{E}\left(\lambda \,|\, W^1, ..., W^n\right) = \frac{a^n}{b^n},$$

is roughly equal to the average number of customers that arrived per day. This is in line with the meaning of the Poisson rate.

The gamma-Poisson case highlights the distinction between the sampling distribution and the prior. While the individual Poisson observations are discrete, the Poisson rate itself can be any positive real number, and thus can be modeled using the gamma distribution.

2.6.3 The Pareto-uniform model

Suppose that W is uniform on the interval $[0, B]$, where B is unknown. Our problem is thus to estimate the maximum of a uniform distribution. This problem is the continuous version of a production estimation problem, in which we can observe a sequence of serial numbers, and the goal is to guess the highest serial number produced. We can also use this model to estimate the maximum possible demand for a product or other extreme values.

We assume that B comes from a Pareto distribution with parameters $b > 0$ and $\alpha > 1$. The density of this distribution is given by

$$f(x|\alpha, b) = \begin{cases} \frac{\alpha b^\alpha}{x^{\alpha+1}} & \text{if } x > b \\ 0 & \text{otherwise.} \end{cases}$$

Thus, our prior estimate of B using priors $\alpha = \alpha^0$ and $b = b^0$ is given by

$$\mathbb{E}\left(B\right) = \frac{\alpha^0 b^0}{\alpha^0 - 1}.$$

The parameter b^0 estimates the $\frac{\alpha^0 - 1}{\alpha^0}$-quantile of the uniform distribution, and α^0 gives us the multiplier used to obtain the estimate of the maximum.

Although this model looks somewhat peculiar, it also has the conjugacy property. Our beliefs continue to have a Pareto distribution as we make observations, and the parameters of the distribution evolve according to the equations

$$\begin{aligned} b^{n+1} &= \max\left(b^n, W^{n+1}\right), && (2.34) \\ \alpha^{n+1} &= \alpha^n + 1. && (2.35) \end{aligned}$$

Thus, b^n is roughly the maximum of the first n uniform observations. Our beliefs tell us that the true maximum of the distribution must be larger than b^n. However, if we have made many observations, it is likely that b^n is fairly close to the maximum. The degree of this "closeness" is represented by α^n.

2.6.4 Models for learning probabilities*

In many problems, our objective is to learn the probability that a certain event will occur, rather than the economic value of the event. For example, in a medical setting, our observations might simply be whether or not a certain medical treatment is successful. Such an observation can be modeled as a Bernoulli random variable, which is equal to 1 (success) with probability ρ, and 0 (failure) with probability $1 - \rho$. The success probability ρ is the unknown true value in this case.

We assume that ρ comes from a beta distribution with parameters α and β. Recall that the beta density is given by

$$f\left(x|\alpha,\beta\right) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}\left(1-x\right)^{\beta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

As before, $\Gamma(y) = y!$ when y is integer. In this setting, α and β are always integer. Figure 2.3 illustrates the beta distribution for different values of α and β.

Our prior estimate of ρ using $\alpha = \alpha^0$ and $\beta = \beta^0$ is given by

$$\mathbb{E}\left(\rho\right) = \frac{\alpha^0}{\alpha^0 + \beta^0}. \tag{2.36}$$

Thus, α^0 and β^0 are weights that, when normalized, give us the probabilities of success and failure, respectively. If α^0 is large relative to β^0, this means that we believe success to be more likely than failure.

A common trait of all the learning models we have discussed thus far is that, while the prior or sampling distributions can have fairly complicated densities, the resulting updating equations are simple and often have an intuitive meaning. This is also the

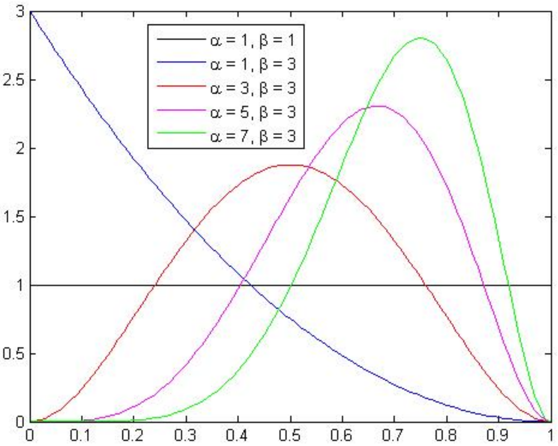


Figure 2.3 Illustration of a family of beta distributions.

case for the beta-Bernoulli model. The conjugacy property holds, and the parameters evolve according to the equations

$$\alpha^{n+1} = \alpha^n + W^{n+1}, \tag{2.37}$$

$$\beta^{n+1} = \beta^n + \left(1 - W^{n+1}\right), \tag{2.38}$$

where the observations W^n are 1 or 0, indicating a success or a failure. We see that the parameters α^n and β^n roughly keep track of the number of successes and failures in n observations. For instance, the parameter α^n is the number of successes in n observations, plus a prior constant α^0. This prior constant serves as a scaling factor of sorts. To see the importance of this scaling, consider the following three scenarios:

Scenario 1:	$\alpha^0 = \beta^0 = 0.1$
Scenario 2:	$\alpha^0 = \beta^0 = 1$
Scenario 3:	$\alpha^0 = \beta^0 = 10$

In each scenario, our estimate of ρ is $1/2$, because the prior constants α^0 and β^0 are equal. However, suppose now that $W^1 = 1$ in all three cases, and consider how our beliefs change:

Scenario 1:	$\alpha^1 = 1.1, \beta^1 = 0.1,$	$\mathbb{E}\left(\rho \,\vert\, W^1\right) = 0.916$
Scenario 2:	$\alpha^1 = 2, \beta^1 = 1,$	$\mathbb{E}\left(\rho \,\vert\, W^1\right) = 0.666$
Scenario 3:	$\alpha^0 = 11, \beta^0 = 10,$	$\mathbb{E}\left(\rho \,\vert\, W^1\right) = 0.524$

In the first scenario, observing a single success leads us to greatly increase our estimate of the success probability. In the other two scenarios, we also increase our estimate of ρ, but by a much smaller amount. Thus, the prior values of α^0 and β^0 can be viewed as a measure of our confidence in our estimate of ρ. High values of α^0 and β^0 show that we are very confident in our prior estimate, and hence this estimate is not likely to change by very much. Low values of α^0 and β^0 show that we have little prior knowledge about ρ, and our prior estimate can easily be changed by only a few observations.

The beta-Bernoulli model can be easily generalized to a multivariate setting. Suppose that, instead of a simple 0/1 value, each observation can be classified as belonging to one of K different categories. For example, instead of merely measuring the success or failure of a medical treatment, we also consider cases where the treatment is generally effective, but causes certain side effects. This result should be viewed differently from either total failure or unqualified success. Consequently, we now have more than two possible outcomes.

We model our observations as individual trials from a multinomial distribution with K categories. The probability that an observation belongs to category $k = 1, ..., K$ is $P(W^n = k) = \rho_k$, with each $\rho_k \in [0, 1]$ and $\sum_{k=1}^K \rho_k = 1$. The unknown true values are now the probabilities ρ_k. Let us use $\rho = (\rho_1, ..., \rho_K)$ to denote the vector containing all these probabilities.

Our prior is the multivariate generalization of the beta distribution, called the *Dirichlet distribution*. This distribution has a vector of parameters $\alpha = (\alpha_1, ..., \alpha_K)$, with one parameter for each category, satisfying $\alpha_k \geq 0$ for all k. The Dirichlet density is given by

$$f(x) = \begin{cases} \frac{\Gamma(\alpha_1 + ... + \alpha_K)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K x_k^{\alpha_k - 1} & \text{if } x_k \geq 0 \text{ for all } k \text{ and } \sum_{k=1}^K x_k = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.39}$$

Essentially, the Dirichlet density is a probability distribution for the probability that an observation belongs in a particular category. The density can only be non-zero on the set of points x (the probabilities) in a k-dimensional space such that every component of x is positive, and the sum of the components is 1 (since the sum of the probabilities of being in each category has to sum to 1). For example, in two dimensions, this set is the part of the line $x_1 + x_2 = 1$ that lies in the non-negative quadrant.

By computing the marginal densities of (2.39), it can be shown that our estimate of the probability of observing an outcome in category k given a prior $\alpha = \alpha^0$ is

$$\mathbb{E}(\rho_k) = \frac{\alpha_k^0}{\alpha_1^0 + ... + \alpha_K^0},$$

a straightforward generalization of (2.36). Just as with the beta-Bernoulli model, the prior values α_k^0 represent the weight that we want to assign to the kth category. Large values of α_k^0 relative to $\alpha_{k'}^0$ indicate that we are more likely to observe category k than category k'. Our earlier discussion of scaling applies here as well.

To update our beliefs, we apply the equation

$$\alpha_k^{n+1} = \begin{cases} \alpha_k^n + 1 & \text{if } W^{n+1} \text{ belongs to category } k \\ \alpha^n & \text{otherwise.} \end{cases} \tag{2.40}$$

We can write this more concisely if we model W as taking values of the form e_k, where e_k is a K-vector of zeroes, with only the kth component equal to 1. Then, the probability mass function of W is given by $P(W = e_k) = \rho_k$, and (2.40) can be rewritten as

$$\alpha^{n+1} = \alpha^n + W^{n+1}. \tag{2.41}$$

It is important to remember that (2.41) is a vector equation, where only one component of W^{n+1} is equal to 1, and the other components are equal to zero. Simply put, if observation $n+1$ belongs to category k, we increment α_k^n by 1 and leave the other components of α^n unchanged.

2.6.5 Learning an unknown variance*

Our last learning model takes us back to the basic setting of one-dimensional Gaussian priors from Section 2.3.1. As before, we assume that the observation $W \sim \mathcal{N}\left(\mu, \beta^W\right)$, where $\beta^W = 1/\sigma_W^2$ is the precision. However, we will now suppose that both the true mean μ and the precision β^W are unknown. We will have to learn both of these quantities at the same time.

It is easy to imagine applications where β^W is unknown. In fact, the precision of an observation is often more difficult to estimate than the mean. For example, in finance, the return of a stock can be directly observed from market data, but the volatility has to be indirectly inferred from the returns. We often assume that β^W is known because it makes our model cleaner, but even then, in practice the value that we plug in for this quantity will be some sort of statistical estimate.

Because the mean and precision are both estimated using the same data, our beliefs about these two quantities are correlated. We create a joint prior distribution on $\left(\mu, \beta^W\right)$ in the following way. First, the marginal distribution of β^W is $Gamma\left(a, b\right)$, where $a, b > 0$ are prior parameters of our choosing. Next, given that $\beta^W = r$, the conditional distribution of μ is $\mathcal{N}\left(\theta, \tau r\right)$, where $-\infty < \theta < \infty$ and $\tau > 0$ are also prior parameters. Note that τr denotes the conditional precision of μ, not the conditional variance. We can write the joint density of $\left(\mu, \beta^W\right)$ as

$$f\left(x, r \,|\, \theta, \tau, a, b\right) = \frac{1}{\sqrt{2\pi\tau^{-1}r^{-1}}} \frac{b\left(br\right)^{a-1} e^{-br}}{\Gamma\left(a\right)} e^{-\frac{(x-\theta)^2}{2\tau^{-1}r^{-1}}}.$$

This is widely known as a "normal-gamma" distribution. It is closely related to Student's t-distribution (often simply called the t-distribution), used in statistics to estimate the mean of a sample under unknown variance. In fact, if $\left(\mu, \beta^W\right)$ is normal-gamma with parameters θ, τ, a and b, the marginal distribution of μ can be connected back to the t distribution. The random variable $\sqrt{\frac{\tau a}{b}}\left(\mu - \theta\right)$ follows the standard t distribution with $2a$ degrees of freedom, analogous to expressing a Gaussian random variable in terms of the standard Gaussian distribution.

The estimates of the unknown quantities that we obtain from the normal-gamma distribution are given by

$$\mathbb{E}\mu = \theta, \qquad \mathbb{E}\beta^W = \frac{a}{b}.$$

The parameter τ only affects the amount of uncertainty in our beliefs, with

$$Var\left(\mu\right) = \frac{b}{\tau\left(a-1\right)}, \qquad Var\left(\beta^W\right) = \frac{a}{b^2}.$$

Recall that τ affects the precision, so lower τ leads to more uncertainty and higher prior variance.

Like the other distributions considered in this chapter, the normal-gamma prior is conjugate when combined with normal observations. Suppose that $(\bar{\mu}^n, \tau^n, a^n, b^n)$ represent our beliefs after n observations, and we make an observation $W^{n+1} \sim \mathcal{N}\left(\mu, \beta^W\right)$. Then, the posterior distribution of $\left(\mu, \beta^W\right)$ is normal-gamma with parameters

$$\bar{\mu}^{n+1} = \frac{\tau^n \bar{\mu}^n + W^{n+1}}{\tau^n + 1}, \tag{2.42}$$

$$\tau^{n+1} = \tau^n + 1, \tag{2.43}$$

$$a^{n+1} = a^n + \frac{1}{2}, \tag{2.44}$$

$$b^{n+1} = b^n + \frac{\tau^n \left(W^{n+1} - \bar{\mu}^n\right)^2}{2\left(\tau^n + 1\right)}. \tag{2.45}$$

The equations are more complicated than their analogs in Section 2.3.1. However, (2.42) is actually a straightforward generalization of (2.6), replacing the precisions β^n and β^W from the known-variance model by scale factors, τ^n for the prior precision and 1 for the observation. In this way, we can see that the parameter τ^n is roughly equal to n, plus a prior constant.

Later on, we will use the normal-gamma model in a setting where, instead of collecting one observation at a time, we can obtain a batch of k observations simultaneously. In this case, we can easily update our beliefs by calculating (2.42)-(2.45) k times for the individual observations $W^{n+1}, ..., W^{n+k}$. It is instructive, however, to look at the equivalent "batch version" of the updating equations. Let $\bar{W}^{n,k} = \frac{1}{k}\sum_{i=1}^{k} W^{n+i}$ be the average of our k observations. Then, the posterior distribution of $\left(\mu, \beta^W\right)$ is normal-gamma with parameters

$$\bar{\mu}^{n+k} = \frac{\tau^n \bar{\mu}^n + k\bar{W}^{n,k}}{\tau^n + k}, \tag{2.46}$$

$$\tau^{n+k} = \tau^n + k, \tag{2.47}$$

$$a^{n+k} = a^n + \frac{k}{2}, \tag{2.48}$$

$$b^{n+k} = b^n + \frac{1}{2}\sum_{i=1}^{k}\left(W^{n+i} - \bar{W}^{n,k}\right)^2 + \frac{\tau^n k \left(\bar{W}^{n,k} - \bar{\mu}^n\right)^2}{2\left(\tau^n + k\right)}. \tag{2.49}$$

The differences between (2.42)-(2.45) and (2.46)-(2.49) are mostly straightforward. For example, in (2.46), the scale factor of k observations is simply k times the scale factor of a single observation. Notice, however, that (2.49) now involves a sum of squared errors $\left(W^{n+i} - \bar{W}^{n,k}\right)^2$ for $i = 1, ..., k$.

This sum of squares will be automatically computed if we apply (2.45) k times in a row, much like the recursive expression in (2.5) computes the sum of squares in (2.3). In this way, we see that the frequentist and Bayesian models are estimating the variance in roughly the same way. The Bayesian model simply adds a correction to the frequentist estimate in the form of the last term in (2.45), which uses the prior information $\bar{\mu}^n$ and τ^n. Larger values of τ^n correspond to a more precise prior and will place more weight on this term.

The Bayesian model allows us to learn the unknown variance from a single observation, whereas the frequentist model requires at least two. Essentially, the prior stands in for the missing "first" observation. The difference is mostly cosmetic: in practice, the prior is frequently constructed using old or preliminary observations. The true difference between the frequentist and Bayesian philosophies is not in the updating equations, but in the philosophical interpretation of μ as an "unknown, but fixed number" (frequentist) or a "random variable" (Bayesian).

2.6.6 The normal as an approximation

The list of different priors and sampling distributions can make this entire problem of deriving posteriors seem quite complicated. Perhaps it is not surprising that we keep coming back to the normal distribution, drawing on the power of the central limit theorem.

Consider the setting of counting successes and failures. Let $\bar{p}^n$ be our probability of a success, and let W^{n+1} be our binary outcome capturing whether the experiment was a success ($W^{n+1} = 1$) or a failure ($W^{n+1} = 0$). The estimated probability $\bar{p}^n$ is a random variable with a beta distribution, but for n large enough (and this does not have to be very large), it is also accurately approximated by a normal distribution. If we then observe a binary W^{n+1}, the posterior distribution will not be normal, but it will be approximately normal.

This is a reason why the normal distribution is so widely used. With the notable exception of our sampled beliefs for nonlinear models, the normal distribution will be our most common reference distribution in this book.

2.7 MONTE CARLO SIMULATION

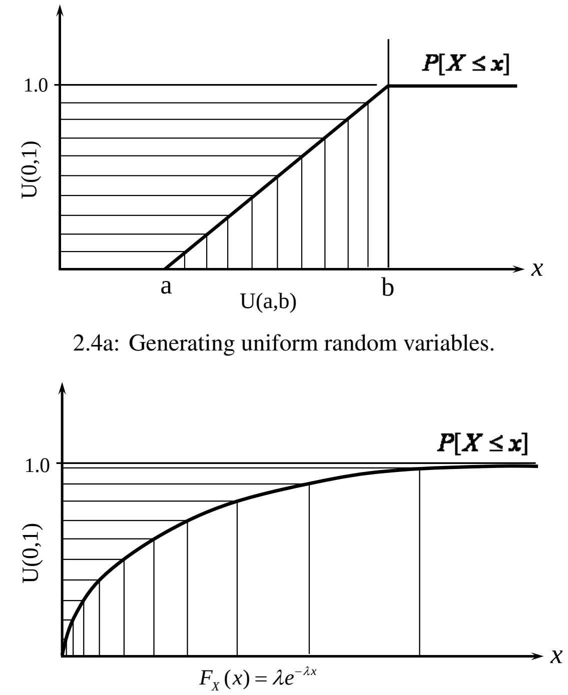
There are many settings where we need to take a sample realization of a random variable, a process known widely as Monte Carlo simulation. There are a number of good books on this subject. This section provides only a brief introduction to some elementary methods for generating samples of a random variable.

All computer languages include a utility for generating a random number that is uniformly distributed between 0 and 1. For example, in Microsoft Excel, this function is called "RAND()", and we can generate a random number from this function by entering "=RAND()" in a cell.

We can use this simple function to generate random variables with virtually any distribution. Let U be a random variable that is uniformly distributed between 0 and 1. If we want a random variable that is uniformly distributed between a and b, we first compute U and then calculate

$$X = a + (b - a)U.$$

It is fairly easy to generate random variables with general distributions if we can compute the inverse cumulative distribution function. Let $F(x) = P[X \leq x]$ be the cdf of a random variable X. If we have a way of generating a sample realization of X with the cumulative distribution $F(x)$, then if we let $Y = F(X)$ (where X is a



2.4a: Generating uniform random variables.

2.4a: Generating exponentially-distributed random variables.

Figure 2.4 Generating uniformly and exponentially distributed random variables using the inverse cumulative distribution method.

realization of our random variable), then it is a simple probability exercise to show that Y is uniformly distributed between 0 and 1. Now assume that we can find the inverse of the cumulative distribution function, which we represent as $F^{-1}(u)$ for a value $0 \le u \le 1$. If U is a random variable that is uniformly distributed between 0 and 1, then if we compute X using

$$X = F_X^{-1}(U),$$

we can show that X has the cumulative distribution $F(x)$.

For example, consider the case of an exponential density function $\lambda e^{-\lambda x}$ with cumulative distribution function $1 - e^{-\lambda x}$. Setting $U = 1 - e^{-\lambda x}$ and solving for x gives

$$X = -\frac{1}{\lambda}\ln(1 - U).$$

Since $1 - U$ is also uniformly distributed between 0 and 1, we can use

$$X = -\frac{1}{\lambda}\ln(U).$$

Figure 2.4 demonstrates how we can use a random variable that is uniformly distributed between 0 and 1 to create a random variable that is uniformly distributed between a and b (in figure 2.4a), and a random variable that has an exponential distribution (in figure 2.4b).

We can use this method to compute random variables with a normal distribution. Excel and MATLAB, for example, have a function called $\texttt{NORMINV}(p, \mu, \sigma)$ where p is a probability, μ is the mean of a normally distributed random variable with standard deviation σ. Writing

$$x = \texttt{NORMINV}(p, \mu, \sigma)$$

in MATLAB (in Excel, you would enter "$= \texttt{NORMINV}(p, \mu, \sigma)$" in a cell) generates the value of a normally distributed random variable X such that $P(X \leq x) = p$. To generate a random variable that is normally distributed with mean μ and standard deviation σ, simply generate a uniform random variable U and them compute

$$x = \texttt{NORMINV}(U, \mu, \sigma).$$

Now imagine that we want to generate a Monte Carlo sample for a vector of correlated random variables. Let μ be a M-dimensional vector of means, and let Σ be a $M \times M$ covariance matrix. We would like to find a sample realization $u \sim N(\mu, \Sigma)$. This can be done very simply. Let C be the "square root" of Σ which is computed using Cholesky decomposition. In MATLAB, this done using

$$C = \texttt{chol}(\Sigma).$$

The result is an upper triangular matrix C, which is sometimes called the square root of Σ because

$$\Sigma = CC^T.$$

Let Z be an M-dimensional vector of independent, standard normal variables generated as we just described above. Given Z, we can compute a sample realization of μ using

$$u = \mu + CZ.$$

The vector u satisfies $\mathbb{E}u = 0$. To find the variance, we use

$$Cov(u) = Var(u + CZ) = CCov(Z)C^T.$$

Since the elements of the vector Z are independent with variance 1, $Cov(Z) = I$ (the identity matrix), which means $Cov(u) = CC^T = \Sigma$.

To illustrate, assume our vector of means is given by

$$\mu = \begin{bmatrix} 10 \\ 3 \\ 7 \end{bmatrix}.$$

our covariance matrix is given by

$$\Sigma = \begin{bmatrix} 9 & 3.31 & 0.1648 \\ 3.31 & 9 & 3.3109 \\ 0.1648 & 3.3109 & 9 \end{bmatrix}.$$

The Cholesky decomposition computed by MATLAB using $C = \texttt{chol}(\Sigma)$ is

$$C = \begin{bmatrix} 3 & 1.1033 & 0.0549 \\ 0 & 3 & 1.1651 \\ 0 & 0 & 3 \end{bmatrix}.$$

Imagine that we generate a vector Z of independent standard normal deviates

$$Z = \begin{bmatrix} 1.1 \\ -0.57 \\ 0.98 \end{bmatrix}.$$

Using this set of sample realizations of Z, a sample realization u would be

$$u = \begin{bmatrix} 10.7249 \\ 2.4318 \\ 7.9400 \end{bmatrix}.$$

Using computers to generate random numbers has proven to be an exceptionally powerful tool in the analysis of stochastic systems. Not surprisingly, then, the field has matured into a rich and deep area of study. This presentation is little more than a hint at the many tools available to help with the process of generating random numbers.

2.8 WHY DOES IT WORK?*

2.8.1 Derivation of $\tilde{\sigma}$

An important quantity in optimal learning is the variance $\tilde{\sigma}^n_x$ of $\bar{\mu}^{n+1}_x$ given what we know after n experiments.

Proposition 2.8.1 *The variance of $\bar{\mu}^{n+1}$, defined as*

$$\begin{aligned} \tilde{\sigma}^n &= Var^n[\bar{\mu}^{n+1}] \\ &= Var^n[\bar{\mu}^{n+1} - \bar{\mu}^n], \end{aligned}$$

is given by $(\tilde{\sigma}^n_x)^2 = (\sigma^n_x)^2 - (\sigma^{n+1}_x)^2$.

Proof: Keep in mind that after n experiments, $\bar{\mu}^n_x$ is deterministic. We are dealing with two random variables: the truth μ_x, and the estimate $\bar{\mu}^{n+1}_x$ after we have made n experiments (the $n + 1$st experiment, W^{n+1}, is unknown). We begin with the relation

$$(\bar{\mu}^{n+1}_x - \mu_x) = (\bar{\mu}^{n+1}_x - \bar{\mu}^n_x) + (\bar{\mu}^n_x - \mu_x). \tag{2.50}$$

Recall that $(\sigma^{n+1}_x)^2 = \mathbb{E}^{n+1}\left[(\bar{\mu}^{n+1}_x - \mu_x)^2\right]$. Squaring both sides of (2.50) and taking the conditional expectation $\mathbb{E}^{n+1}(\cdot) = \mathbb{E}(\cdot|W^1, \ldots, W^{n+1})$ gives

$$\begin{aligned} (\sigma^{n+1}_x)^2 &= \mathbb{E}^{n+1}\left[(\bar{\mu}^n_x - \mu_x)^2\right] + 2\mathbb{E}^{n+1}\left[(\bar{\mu}^n_x - \mu_x)(\bar{\mu}^{n+1}_x - \bar{\mu}^n_x)\right] \\ &\quad + \mathbb{E}^{n+1}\left[(\bar{\mu}^{n+1}_x - \bar{\mu}^n_x)^2\right] \\ &= \mathbb{E}^{n+1}\left[(\bar{\mu}^n_x - \mu_x)^2\right] + 2(\bar{\mu}^n_x - \bar{\mu}^{n+1}_x)(\bar{\mu}^{n+1}_x - \bar{\mu}^n_x) + (\bar{\mu}^{n+1}_x - \bar{\mu}^n_x)^2 \\ &= \mathbb{E}^{n+1}\left[(\bar{\mu}^n_x - \mu_x)^2\right] - (\bar{\mu}^{n+1}_x - \bar{\mu}^n_x)^2. \end{aligned}$$

Keep in mind that $\mathbb{E}^{n+1}\bar{\mu}_x^{n+1} = \bar{\mu}_x^{n+1}$ and $\mathbb{E}^{n+1}\mu_x = \bar{\mu}_x^{n+1}$. We then observe that while $\bar{\mu}_x^{n+1}$ is random given the first n observations, σ_x^{n+1} is deterministic, because σ_x^{n+1} does not depend on W^{n+1} - it only depends on our decision of what to measure x^n. Using this property, we can take the expectation given $W^1, \ldots, W^n$ to obtain

$$\begin{aligned}\mathbb{E}^n(\sigma_x^{n+1})^2 = (\sigma_x^{n+1})^2 &= \mathbb{E}^n\left[\mathbb{E}^{n+1}\left[(\bar{\mu}_x^n - \mu_x)^2\right]\right] - \mathbb{E}^n\left[(\bar{\mu}_x^{n+1} - \bar{\mu}_x^n)^2\right]\\ &= \mathbb{E}^n\left[(\bar{\mu}_x^n - \mu_x)^2\right] - \mathbb{E}^n\left[(\bar{\mu}_x^{n+1} - \bar{\mu}_x^n)^2\right]\\ &= (\sigma_x^n)^2 - (\tilde{\sigma}_x^n)^2.\end{aligned}$$

2.8.2 Derivation of Bayesian updating equations for independent beliefs

Bayesian analysis begins with a simple formula that everyone learns in their first probability course. Given the events A and B, the basic properties of conditional probability imply

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

which implies

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

This expression is famously known as Bayes' theorem. In a learning setting, the event A refers to an experiment (or some type of new information), while B refers to the event that a parameter (say, the mean of a distribution) takes on a particular value. $P(B)$ refers to our initial (or prior) distribution of belief about the unknown parameter before we run an experiment, and $P(B|A)$ is the distribution of belief about the parameter after the experiment has been run. For this reason, $P(B|A)$ is known as the *posterior distribution*.

We can apply the same idea for continuous variables. We replace B with the event that $\mu = u$ (to be more precise, we replace B with the event that $u \leq \mu \leq u + du$), and A with the event that we observed $W = w$. Let $g(u)$ be our prior distribution of belief about the mean μ, and let $g(u|w)$ be the posterior distribution of belief about μ given that we observed $W = w$. We then let $f(w|u)$ be the distribution of the random variable W if $\mu = u$. We can now write our posterior $g(u|w)$, which is the density of μ given that we observe $W = w$, as

$$g(u|w) = \frac{f(w|u)g(u)}{f(w)},$$

where $f(w)$ is the unconditional density of the random variable W which we compute using

$$f(w) = \int_u f(w|u)g(u).$$

Equation (2.51) gives us the density of μ given that we have observed $W = w$.

We illustrate these calculations by assuming that our prior $g(u)$ follows the normal distribution with mean $\bar{\mu}^0$ and variance $\bar{\sigma}^{2,0}$, given by

$$g(u) = \frac{1}{\sqrt{2\pi}\sigma^0} \exp\left(-\frac{1}{2}\frac{(u-\bar{\mu}^0)^2}{\bar{\sigma}^{2,0}}\right).$$

We further assume that the observation W is also normally distributed with mean μ and variance σ^2_ϵ, which is sometimes referred to as the experimental or observation error. The conditional distribution $f(w|u)$ is

$$f(w|u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(w-u)^2}{\sigma^2_\epsilon}\right).$$

We can compute $f(w)$ from $f(w|u)$ and $g(u)$, but it is only really necessary to find the density $g(u|w)$ up to a normalization constant ($f(w)$ is part of this normalization constant). For this reason, we can write

$$g(u|w) \propto f(w|u)g(u). \tag{2.51}$$

Using this reasoning, we can drop coefficients such as $\frac{1}{\sqrt{2\pi}\sigma^0}$, and write

$$\begin{aligned} g(u|w) &\propto \left[\exp\left(-\frac{1}{2}\frac{(w-u)^2}{\sigma^2}\right)\right] \cdot \left[\exp\left(-\frac{1}{2}\frac{(u-\bar{\mu}^0)^2}{\bar{\sigma}^{2,0}}\right)\right], \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{(w-u)^2}{\sigma^2} + \frac{(u-\bar{\mu}^0)^2}{\bar{\sigma}^{2,0}}\right)\right]. \end{aligned} \tag{2.52}$$

After some algebra, we find that

$$g(u|w) \propto \exp -\frac{1}{2}\beta^1(u-\bar{\mu}^1)^2 \tag{2.53}$$

where

$$\bar{\mu}^1 = \frac{(\beta^W w + \beta^0\bar{\mu}^0)}{\alpha + \beta^0}, \tag{2.54}$$

$$\beta^1 = \beta^W + \beta^0. \tag{2.55}$$

The next step is to find the normalization constant (call it K) which we do by solving

$$K\int_u g(u|w)du = 1.$$

We could find the normalization constant by solving the integral and picking K so that $g(u|w)$ integrates to 1, but there is an easier way. What we are going to do is look around for a known probability density function with the same structure as (2.53), and then simply use its normalization constant. It is fairly easy to see that (2.53) corresponds to a normal distribution, which means that the normalization constant $K = \sqrt{\frac{\beta^1}{2\pi}}$. This means that our posterior density is given by

$$g(u|w) = \sqrt{\frac{\beta^1}{2\pi}} \exp -\frac{1}{2}\beta^1(u-\bar{\mu}^1)^2. \tag{2.56}$$

From equation (2.56), we see that the posterior density $g(u|w)$ is also normally distributed with mean $\bar{\mu}^1$ given by (2.54), and precision β^1 given by (2.55) (it is only now that we see our choice of notation $\bar{\mu}^1$ and β^1 in equations (2.54) and (2.55) was not an accident). This means that as long as we are willing to stay with our assumption of normality, then we need only to carry the mean and variance (or precision). The implication is that we can write our belief state as $B^n = (\bar{\mu}^n, \beta^n)$ (or $B^n = (\bar{\mu}^n, \bar{\sigma}^{2,n})$), and that (2.54)-(2.55) is our belief transition function.

Our derivation above was conducted in the context of the normal distribution, but we followed certain steps that can be applied to other distributions. These include:

1) We have to be given the prior $g(u)$ and the conditional density $f(w|u)$ of the experimental outcome.

2) We use equation (2.51) to find the posterior up to the constant of proportionality, as we did in (2.52) for the normal distribution.

3) We then manipulate the result in the hope of finding a posterior distribution, recognizing that we can discard terms that do not depend on u (these are absorbed into the normalization constant). If we are lucky, we will find that we have a conjugate family, and that we end up with the same class of distribution we started with for the prior. Otherwise, we are looking for a familiar distribution so that we do not have to compute the normalization constant ourselves.

4) We identify the transition equations that relate the parameters of the posterior to the parameters of the prior and the distribution of the experimental outcome, as we did with equations (2.54)-(2.55).

2.9 BIBLIOGRAPHIC NOTES

Sections 2.2-2.3 - There are numerous books on both frequentist and Bayesian statistics. An excellent reference for frequentist statistics is Hastie et al. (2009), which is available as of this writing as a free download in PDF form. An excellent reference for Bayesian statistics is Gelman et al. (2004). Another classical reference, with a focus on using Bayesian statistics to solve decision problems, is DeGroot (1970).

Section 2.6 - See Gelman et al. (2004) for a thorough treatment of Bayesian models.

Section 2.7 - There are a number of outstanding references on Monte Carlo simulation, including Banks et al. (1996), Roberts & Casella (2004), Glasserman (2004) and Rubinstein & Kroese (2008), to name a few.

PROBLEMS

2.1 In a spreadsheet, implement both the batch and recursive formulas for the frequentist estimates mean and variance of a set of random numbers (just use RAND() to produce random numbers between 0 and 1). Use a sequence of 20 random numbers. Note that Excel has functions to produce the batch estimates (AVERAGE and VAR) of the mean and variance. Compare your results to Bayesian estimates of the mean and variance assuming that your prior is a mean of .5 and a variance of .2 (the prior estimate of the mean is correct, while the variance is incorrect).

2.2 Download the spreadsheet from the book website:

```
http://optimallearning.princeton.edu/exercises/FiveAlternative.xls
```

On day 0, the spreadsheet shows your initial estimate of the travel time on each path. In the column painted yellow, enter the path that you wish to follow the *next* day. You will see an observed time, and in the column to the right, we record the time you experience the next day when you follow your chosen path. To the right there is a work area where you can code your own calculations. Assume all random variables are normally distributed.

a) In the work area, use the Bayesian updating formulas to compute updated estimates of the mean and variance. Assume that the standard deviation of the observed travel time for any path is 5 minutes.

b) Using your estimates from part a, simulate a policy where you always choose the path that you thought was fastest. Record your total travel time, and the path that you thought was best.

c) You should find that you are getting stuck on one path, and that you do not "discover" the best path (you can quickly find that this is path 1). Suggest a policy that could be applied to any dataset (there can not be any hint that you are using your knowledge of the best path). Report your total travel time and the final path you choose.

2.3 You are trying to determine the distribution of how much people weigh in a population. Your prior distribution of belief about the mean μ is that it is normally distributed with mean 180 and standard deviation 40. You then observe the weights of n students drawn from this population. The average weight in the sample is $\bar{y}$ = 150 pounds. Assume that the weights are normally distributed with unknown mean μ and a known standard deviation of 20 pounds.

a) Give your posterior distribution for μ given the sample you have observed. Note that your answer will be a function of n.

b) A new observation is made and has a weight of $\tilde{y}$ pounds. Give the posterior distribution for $\tilde{y}$ (again, your answer will be a function of n).

c) Give a 95 percent posterior confidence intervals for μ and $\tilde{y}$ if $n = 10$.

d) Repeat (c) with $n = 100$.

2.4 Show that, for a fixed $k > 1$, equation (2.49) is equivalent to applying (2.45) k times in a row. Use the equivalence of (2.3) and (2.5). Now implement both the recursive and batch formulas in a spreadsheet and verify that they produce the same numbers.

2.5 In this problem, you will derive the Bayesian updating equations (2.30) and (2.31) for the gamma-exponential model. Suppose that W is a continuous random variable that follows an exponential distribution with parameter λ. The parameter λ is also random, reflecting our distribution of belief. We say that λ has a gamma prior by writing $\lambda \sim Gamma(a, b)$, meaning that our distribution of belief is gamma. Each time we observe W, we use this observation to update our belief about the true distribution of W, reflecting our uncertainty about λ.

a) Write down the prior density $f(u)$. What does u refer to?

b) Write down the conditional density $g(w|u)$ of the observation.

c) Write down the unconditional density $g(w)$ of the observation. (Hint: Start with $g(w|u)$ and take an expectation over the prior. Remember that the gamma function has the property that $\Gamma(a) = (a-1)\Gamma(a-1)$.)

d) Apply Bayes' rule to get $f(u|w)$, the posterior density after the observation. What distribution does this density correspond to? How does this verify equations (2.30) and (2.31)?

2.6 In this problem, you will see how the updating equations you derived in exercise 2.5 works in practice. Suppose that the customer service time at a certain store is exponentially distributed, and we are using the gamma-exponential model to learn the service rate as we observe the service times of individual customers.

a) Let U be a random variable that is uniformly distributed between 0 and 1. Let $R = -\frac{1}{\lambda}\log U$. Show that R follows an exponential distribution with parameter λ. This gives us a way to create samples from any exponential distribution by transforming samples from a uniform distribution (see section 2.4).

b) In a spreadsheet, use the above method to simulate 10 observations from an exponential distribution with parameter $\lambda = 3$. Now suppose that we do not know that λ has this value, and model our beliefs using a gamma prior with parameters $\alpha^0 = 1$ and $\beta^0 = 0.2$. What is our best initial estimate of λ? In your spreadsheet, record the values of α^n and β^n, as well as the resulting estimate of λ, for each n. Copy the results into a table and hand in a paper copy.