4 The Generative Models of Active Inference

Everything should be made as simple as possible, but not simpler.
—Albert Einstein

4.1 Introduction

This chapter complements the preceding chapters' conceptual treatment of Active Inference with a more formal treatment. Specifically, it sets out the relationship between free energy and Bayesian inference, the form of the generative models typically used in Active Inference, and the dynamics obtained from minimizing free energy for these models. A key focus is on how time is represented in a generative model. We will see the distinction between generative models formulated in continuous time and those that treat time as a sequence of events. Finally, we set out the idea of inferential message passing, which underwrites prominent theories in neurobiology—including predictive coding.

4.2 From Bayesian Inference to Free Energy

In the preceding two chapters, we outlined some of the important connections between Active Inference and other established paradigms in the neurosciences. In chapter 2, we focused on the notion of *the Bayesian brain* (Knill and Pouget 2004, Doya 2007)—one of its closest relatives—which provides a useful way to think about some of the consequences of active inference from a more formal perspective. Specifically, it helps us frame the problems that an agent engaging in Active Inference must solve. Broadly, these are the problem of inferring states of the world (perception) and inferring a course of action (planning). While it is tempting to equate Bayes optimality

with exact Bayesian inference, exact inference is generally computationally intractable or even infeasible. In cognitive psychology and artificial intelligence applications, it is common to consider bounded forms of inference and rationality. We highlighted some examples in chapter 3. Under a Bayesian framework, this translates into using approximate inference. These methods comprise sampling methods and variational methods—on which active inference is based. In this section, we recap the basic elements of Bayesian inference and its variational manifestations (Beal 2003, Wainwright and Jordan 2008). In doing so, we hope to provide some intuition for the role of *free energy* and to emphasize the importance of *generative models* in drawing inferences about the world.

This chapter is more technical than chapters 1–3, appealing to a little linear algebra, differentiation, and the Taylor series expansion. Those readers interested in the details or in need of a refresher may turn to the appendices for the requisite background. Those who do not want to delve into the theoretical underpinnings may skip this chapter. Throughout, we explain the key implications of each equation—so it should be possible to develop an understanding of the important conceptual points herein even without following the formal argument.

A good place to start is Bayes' theorem. Recall from chapter 2 that this theorem expresses an equality between the product of a prior and a likelihood and the product of a posterior and a marginal likelihood. This is reproduced in equation 4.1:

$$\begin{aligned} P(x)P(y|x) &= P(x|y)P(y) \\ P(y) &= \sum_x P(y,x) = \sum_x P(y|x)P(x) \end{aligned} \tag{4.1}$$

The first line of equation 4.1 is Bayes' theorem. The second line shows that the marginal likelihood (or model evidence), $P(y)$, can be computed directly from the prior and likelihood.[1] This makes the point that the prior and likelihood—which together comprise the generative model—are sufficient for us to compute the model evidence and the posterior probability. Despite this, it is not always easy to do so. The summation (or integration, if dealing with continuous variables) in equation 4.1 can be computationally or analytically intractable. One way to resolve this—the starting point of variational inference—is to convert this potentially difficult integration problem into an optimization problem. To understand how this works, we need to appeal to *Jensen's inequality*, which says that "the $\log^2$ of an average

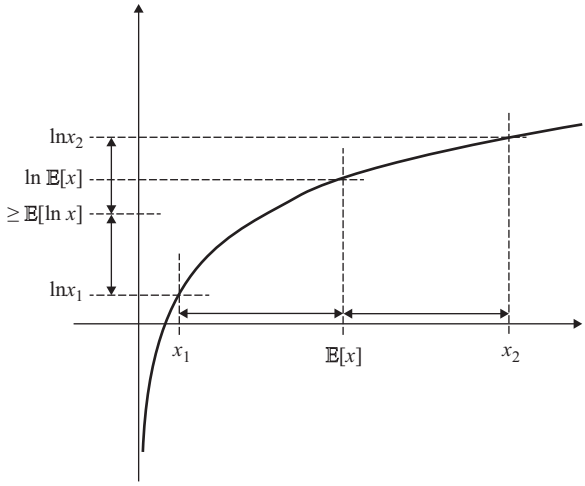

Figure 4.1
Logarithmic function providing intuition for Jensen's inequality. If we had only two data-points (x_1 and x_2), either we could take their average ($\mathbb{E}[x]$) and then find its log, or we could take the log of each data-point and then take the average of these ($\mathbb{E}[\ln x]$). The latter ($\mathbb{E}[\ln x]$) will always be below the former ($\ln \mathbb{E}[x]$), due to the concavity of the logarithmic function, unless the data-points are the same (wherethe log of the average and the average of the log are equal). This inequality holds for any number of data-points.

is always greater than or equal to the average of a log." Figure 4.1 provides a graphical intuition for why this is the case.

To take advantage of this property, we can rewrite equation 4.1 by multiplying the term inside the sum on the second line by an arbitrary function (Q) divided by itself (this is equivalent to multiplying by one, so the equality still holds) and taking the log of each side. Mathematically, this changes nothing. However, we can now interpret the expression as an expectation ($\mathbb{E}$)[3] of a ratio between two probabilities and so exploit Jensen's inequality:

$$\begin{aligned} \ln P(y) &= \ln \sum_x P(y,x) \frac{Q(x)}{Q(x)} \\ &= \ln \mathbb{E}_{Q(x)}\left[\frac{P(y,x)}{Q(x)}\right] \geq \mathbb{E}_{Q(x)}\left[\ln \frac{P(y,x)}{Q(x)}\right] \triangleq -F[Q,y] \end{aligned} \tag{4.2}$$

The second line of this equation uses the fact that we have a log expectation and that, by Jensen's inequality, this must always be greater than or equal to the expectation of the log. This move is sometimes referred to as

importance sampling. The right-hand side of this inequality is known as the (negative) variational free energy:[4] the smaller the free energy, the closer it is to the negative log model evidence. With this in mind, we can rewrite Bayes' theorem (equation 4.1) in logarithmic form, take its average under the posterior distribution, and disclose the relationship between this and the quantities of equation 4.2:

$$\begin{aligned} \ln P(x,y) &= \ln P(y) \quad + \ln P(x \mid y) \Rightarrow \\ \mathbb{E}_{P(x|y)}[\ln P(x,y)] &= \ln P(y) \quad + \mathbb{E}_{P(x|y)}[\ln P(x \mid y)] \\ \mathbb{E}_{Q(x)}[\ln P(x,y)] &= -F[Q,y] + \mathbb{E}_{Q(x)}[\ln Q(x)] \end{aligned} \tag{4.3}$$

The second line follows from the fact that the log probability of y is not a function of x, so taking an expectation under the posterior distribution does not change this quantity. Equation 4.3 provides some intuition for the roles of the free energy and the Q distribution—the two quantities that were difficult to compute without the variational approximation. The former plays the role of the negative log model evidence, while the latter acts as if it were the posterior probability. More formally, we can rearrange the free energy as we did in chapter 2 to quantify the relationship between free energy and model evidence:

$$\begin{aligned} F[Q,y] &= \underbrace{D_{KL}[Q(x) \,||\, P(x \mid y)]}_{\text{Divergence}} - \underbrace{\ln P(y)}_{\text{Log model evidence}} \\ D_{KL}[Q(x) \,||\, P(x \mid y)] &= \mathbb{E}_{Q(x)}\big[\ln Q(x) - \ln P(x \mid y)\big] \end{aligned} \tag{4.4}$$

The first line of equation 4.4 shows the free energy expressed in terms of a KL-Divergence and a negative log evidence. The KL-Divergence is defined in the second line as the expected difference between two log probabilities. This is often used as a measure of how different two probability distributions are from one another.

Sometimes, the use of free energy is motivated directly in terms of this divergence. The argument goes that if our aim is to perform approximate Bayesian inference, we need to find an approximate posterior that best matches the exact posterior. As such, we can select a measure of the divergence between the two—of which the KL-Divergence in equation 4.4 is one example—and minimize this. As we do not know the exact posterior, we cannot use this divergence directly. One solution is to add the log evidence term, which may be combined with the log posterior to form the joint probability (which we do know because this is the generative model). The result is the free energy.

An interesting consequence of this perspective is that there is some ambiguity over which divergence measure to use. If we want to make the approximate and exact posterior as close as possible, we could use the other KL-Divergence, where Q and P are swapped, or choose from a large family of divergences, each of which emphasizes different aspects of the difference between distributions. However, the ideas set out in chapter 3 highlight the importance of self-evidencing for systems engaging in Active Inference. Therefore, we are primarily looking for a tractable evidence maximization scheme and only secondarily looking to minimize the divergence. From this perspective, there is no ambiguity as to which divergence measure to use. This emerges from the use of Jensen's inequality.

4.3 Generative Models

To calculate the free energy, we need three things: data, a family of variational distributions, and a generative model (comprising a prior and a likelihood). In this section, we outline two very general sorts of generative model used for Active Inference and the form the free energy takes in relation to each. The first deals with inferences about categorical variables (e.g., object identity) and is formulated as a sequence of events. The second deals with inferences about continuous variables (e.g., luminance contrast) and is formulated in continuous time using stochastic differential equations. Before specifying the details of these models, we review a graphical formalism that expresses the dependencies implied by a generative model.

Figure 4.2 shows several examples of generative models expressed as factor graphs, chosen to provide some intuition for the sorts of things that may be articulated in this way. These represent the factors (e.g., prior and likelihood) of a generative model as squares and the variables in that model (hidden states or data) in circles. Arrows indicate the direction of causality between these variables. The upper-left graph shows the simplest form these models can take, with a hidden state (x) causing data (y). The prior in this model is shown as factor 1, and the likelihood is factor 2. The other graphs extend this idea by introducing additional variables. In the upper right, z plays the role of a second hidden state, so that y depends on the states of both x and z.

As an example, consider a clinical diagnostic test. In this setting, the simple graph in the upper left can be interpreted as the presence or absence of a

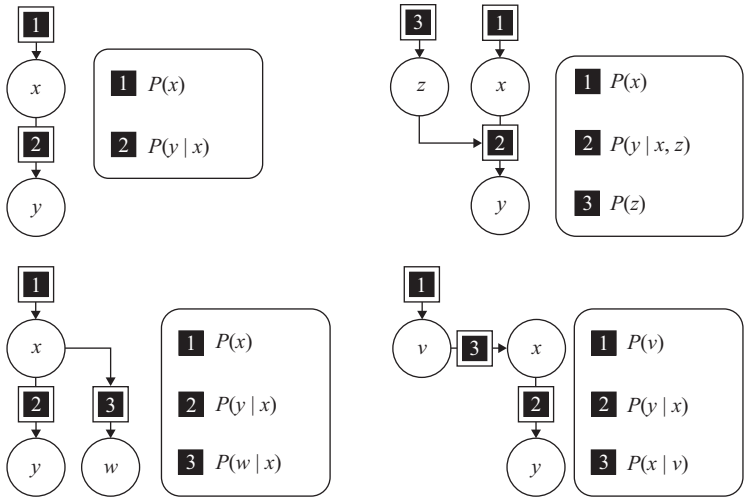

Figure 4.2
Dependencies between variables in a (graphical) probabilistic model. The circles represent random variables (i.e., the things about which we hold beliefs); the squares represent the probability distributions that describe the relationships between these variables. An arrow from one circle to another via a square indicates that the variable in the second circle depends on that in the first circle and that this dependency is captured in the probability distribution represented by the square.

disease (x) and the result of the test (y). The prior is then the prevalence of the disease, while the likelihood specifies the properties of the test. These include its specificity (the probability of a negative result in the absence of the disease) and sensitivity (the probability of a positive result in the presence of the disease). We can then think of the model in terms of the mechanism by which a test result is obtained—going from the top to the bottom of the factor graph. First, we sample a person from a population with known prevalence of a disease. If they have the disease, they will generate a true positive test result with probability given by the test sensitivity, and a false negative otherwise. If they do not have the disease, they will generate a true negative with probability given by the specificity, and a false positive otherwise.

Pursuing the same example, we can interpret the other factor graphs. In the upper-right panel, x and z could be the presence or absence of two different diseases, either of which could give a positive test result. In the lower left, w plays the role of data. Both y and w are generated by x and could represent (for example) two different diagnostic tests that are informative

about the same disease process. Finally, the lower-right graph treats both x and v as hidden states but introduces a hierarchical structure in which v causes x causes y. Here we could think of v as providing a context or a predisposing factor (e.g., genetic polymorphism) for the presence or absence of disease x, which may be tested for by measuring y. In principle, we can add an arbitrary number of variables to this hierarchy.

Generative models of this sort are often used for static perceptual tasks, such as object recognition or cue integration. The generative models used for active inference differ in an important way: they evolve over time as new observations are sampled, and the observations that are added depend (via action) on beliefs about variables in the model. This has two key implications. First, the conditional dependencies include the dependencies of hidden variables at a given time on those at previous times. Second, these models sometimes include hypotheses about "how I am acting" as hidden variables.

Figure 4.3 illustrates the two basic forms of dynamic generative model used in active inference (Friston, Parr, and de Vries 2017) in factor graph form (Loeliger 2004, Loeliger et al. 2007). The upper graph shows a Partially Observable Markov Decision Process (POMDP), which expresses a model in which a sequence of states (s) evolves over time. At each time step, the current state is conditionally dependent on the state at the previous time and on the policy (π) currently being pursued. Policies here may be thought of as indexing alternative trajectories, or sequences of actions, that could be followed. Each time-point is associated with an observation (o) that depends only on the state at that time. This sort of model is very useful in dealing with sequential planning tasks—for example, navigating a maze (Kaplan and Friston 2018)—or decision-making processes that involve selecting between alternatives (e.g., categorization of a scene [Mirza et al. 2016]).

The lower graph in figure 4.3 shows a very similar graphical model but expressed in continuous time. In place of representing a trajectory as a series of states, this model represents the current position, velocity, and acceleration (and successive temporal derivatives) of a state (x). These values (referred to as *generalized coordinates of motion*) can be used to reconstruct a trajectory using a Taylor series expansion (see appendix A for an introduction to Taylor series approximations in this context). The relationship between a state and its temporal derivative here depends on (slowly varying) causes (v) that play a similar role to the policies above. As before,

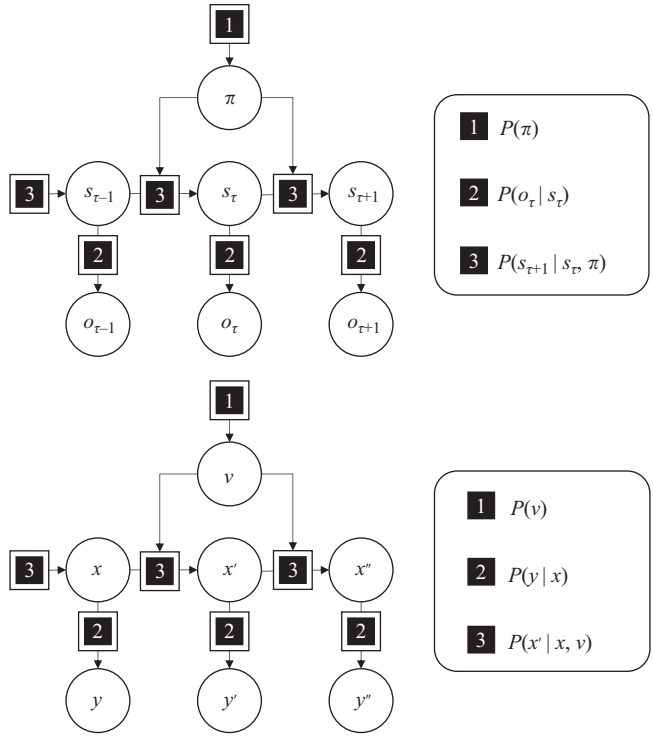

Figure 4.3
Two dynamic generative models (using the same graphical notation as in figure 4.2) that we will appeal to throughout the remainder of this book. *Top:* Partially Observable Markov Decision Process (POMDP), defined in terms of a sequence of states evolving through time (indexed by the subscript). *Bottom:* Continuous-time model, of the sort implied by stochastic differential equations (with the prime notation indicating temporal derivatives).

states generate observations (y). The difference in notation (s, π, o vs. x, v, y) is used to emphasize the difference between categorical variables that evolve in discrete time and continuous variables that evolve in continuous time. Similarly, from here on, we will use lowercase p and q for probability densities over continuous variables and uppercase P and Q for distributions over categorical variables. Sections 4.4 and 4.5 will unpack these models in more detail and will show how minimization of free energy in each case leads to a set of equations that describes the dynamics of inferential processes.

4.4 Active Inference in Discrete Time

In this section, we focus on the discrete-time model outlined above. This is important for understanding a range of cognitive processes that deal with categorical inferences and selection between alternative hypotheses. This formalism additionally facilitates an examination of the classic exploitation-exploration problem and illustrates how active inference resolves this.

4.4.1 Partially Observable Markov Decision Processes

As shown in figure 4.3, a POMDP expresses the evolution over time of a sequence of hidden states that depend on a policy. To specify this process formally, we need to account for the form of each of the square factor nodes in the figure. First, we describe each of these factors. We then combine them to express the joint distribution that constitutes the generative model.

As with the simple example of Bayes' rule given in chapter 2, we can separate the factors into those representing a likelihood and those combining to make a prior. The likelihood is similar to that used before and expresses the probability of an outcome (observable) given a state (hidden). If both the outcomes and states are categorical variables, the likelihood is a categorical distribution, parameterized by a matrix, **A**:

$$\begin{aligned} P(o_\tau | s_\tau) &= Cat(\mathbf{A}) \\ A_{ij} &= P(o_\tau = i | s_\tau = j) \end{aligned} \tag{4.5}$$

The second line here details what is meant by the *Cat* notation (i.e., specification of a categorical distribution). This accounts for the nodes labeled "2" in figure 4.3. The prior over the sequence (expressed using the ~ symbol) of hidden states depends on two things: the prior over the initial state (specified by a vector, **D**) and beliefs about how the state at one time transitions to that at the next (specified as a matrix, **B**):

$$\begin{aligned} P(\tilde{s} | \pi) &= P(s_1) \prod_{\tau=1} P(s_{\tau+1} | s_\tau, \pi) \\ P(s_1) &= Cat(\mathbf{D}) \\ P(s_{\tau+1} | s_\tau, \pi) &= Cat(\mathbf{B}_{\pi\tau}) \end{aligned} \tag{4.6}$$

Together, these account for the "3" nodes in figure 4.3. Note that the transitions are conditionally dependent on the policy chosen. Thus, we can interpret the priors of equation 4.6, combined with the likelihood of equation 4.5, as expressing a model (π) of a behavioral sequence. To allow us to select between these models (i.e., to form a plan), we need a prior belief

about the most probable sequence. For a free energy minimizing creature, a self-consistent prior is that the most probable policies are those that will lead to the lowest expected free energy (G) in the future:

$$\begin{aligned} P(\pi) &= Cat(\boldsymbol{\pi}_0) \\ \boldsymbol{\pi}_0 &= \sigma(-\mathbf{G}) \\ \mathbf{G}_\pi &= G(\pi) = -\mathbb{E}_{\tilde{Q}}[D_{KL}[Q(\tilde{s}|\tilde{o},\pi)\,||\,Q(\tilde{s}|\pi)]] - \mathbb{E}_{\tilde{Q}}[\ln P(\tilde{o}|C)] \\ \tilde{Q}(o_\tau, s_\tau|\pi) &\triangleq P(o_\tau|s_\tau)Q(s_\tau|\pi) \end{aligned} \tag{4.7}$$

This equation, being of fundamental importance to Active Inference, is worth unpacking in more depth. The first two lines express the prior probability for each policy, as parameterized by $\boldsymbol{\pi}_0$, as being related to the negative expected free energy associated with that policy. The softmax function (σ) enforces normalization (i.e., ensures that the probability over policies sums to one). The final two lines of equation 4.7 express the form of the expected free energy.

Note the similarity between this and the functional form of the free energy (equation 4.4)—with a log probability of outcomes and a KL-Divergence. The key difference here is that the expectation is taken with respect to the *posterior predictive* density as defined by the final equality. This distribution expresses a joint probability over future states and observations. Crucially, this means we can compute the expected free energy in the future—something we could not do with the variational free energy, which depends on (present and past) observations. In addition, note the distribution over outcomes depends on parameters (C) and the reversal of the sign of the KL-Divergence, which is a consequence of the expectation under the posterior predictive probability. This last point can cause some confusion, so it is worth spelling out explicitly why this is. In the context of the variational free energy, the KL-Divergence was the expected difference between the log probability of the approximate posterior and the log probability of the exact posterior (equation 4.4). The analogous term in the expected free energy is the expected difference between the approximate posterior and the exact posterior we would get on the basis of the entire trajectory of outcomes, using current posterior beliefs as if they were priors. Unpacking this, we get the following:

$$\begin{aligned} &\mathbb{E}_{\tilde{Q}}\left[\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)\right] \\ &\quad = \mathbb{E}_{Q(\tilde{o}|\pi)}\left[\mathbb{E}_{Q(\tilde{s}|\tilde{o},\pi)}\left[\ln Q(\tilde{s}|\pi) - \ln Q(\tilde{s}|\tilde{o},\pi)\right]\right] \\ &\quad = -\mathbb{E}_{Q(\tilde{o}|\pi)}\left[\mathbb{E}_{Q(\tilde{s}|\tilde{o},\pi)}\left[\ln Q(\tilde{s}|\tilde{o},\pi) - \ln Q(\tilde{s}|\pi)\right]\right] \\ &\quad = -\mathbb{E}_{Q(\tilde{o}|\pi)}\left[D_{KL}[Q(\tilde{s}|\tilde{o},\pi)\,||\,Q(\tilde{s}|\pi)]\right] \end{aligned} \tag{4.8}$$

Here we see that the order in which must take expectations is important. It prompts a reversal in sign relative to the analogous term in the variational free energy. This underwrites an important difference between the two quantities. The expected free energy is minimized by selecting those observations that cause a large change in beliefs, in contrast to the variational free energy that is minimized when observations comply with current beliefs. This is the difference between optimizing beliefs in relation to data that have already been gathered (variational free energy minimization) and selecting those data that will best optimize beliefs (expected free energy minimization).

This reiterates that Active Inference uses two constructs, variational free energy (F) and expected free energy (G), which are mathematically related but play distinct and complementary roles. Variational free energy is the primary quantity that is minimized over time. It is optimized in relation to a generative model, which can include policies (or action sequences). As with all other hidden states, the agent needs to assign a prior probability to policies—because policies are just another random variable in the generative model. Active Inference uses a prior that is (loosely speaking) equivalent to the belief that one will minimize free energy in the future: that is, the expected free energy. In other words, expected free energy furnishes a prior over policies and is therefore a prerequisite in minimizing variational free energy.

In chapter 2 we saw that, as with the variational free energy, the expected free energy can be rearranged in a number of ways to disclose various interpretations. Here, we focus on an interpretation as the difference between the *risk* and the *ambiguity* associated with a policy. This is equivalent to the expression in equation 4.7:

$$\begin{aligned} G(\pi) &= \underbrace{-\mathbb{E}_{\tilde{Q}}[D_{KL}[Q(\tilde{s}|\tilde{o},\pi)\,||\,Q(\tilde{s}|\pi)]]}_{\text{Information gain}} \underbrace{-\mathbb{E}_{\tilde{Q}}[\ln P(\tilde{o}|C)]}_{\text{Pragmatic value}} \\ &= \underbrace{\mathbb{E}_{\tilde{Q}}[H[P(\tilde{o}|\tilde{s})]]}_{\text{Expected ambiguity}} + \underbrace{D_{KL}[Q(\tilde{o}|\pi)\,||\,P(\tilde{o}|C)]}_{\text{Risk}} \end{aligned} \tag{4.9}$$

Recall from chapter 2 that the first of these expresses the trade-off between seeking new information (i.e., exploration) and seeking preferred observations (i.e., exploitation). By minimizing expected free energy, the relative balance between these terms determines whether behavior is predominantly explorative or exploitative. Note that pragmatic value emerges as a prior belief about observations, where the C-parameters of this distribution

may be chosen to reflect the sort of system we are interested in characterizing (in terms of its characteristic or preferred outcome states). Following the second line of equation 4.9, we can rewrite equation 4.7 in linear algebraic form as follows:

$$\begin{aligned}
\boldsymbol{\pi}_0 &= \sigma(-\mathbf{G}) \\
\mathbf{G}_\pi &= \mathbf{H} \cdot \mathbf{s}_{\pi\tau} + \mathbf{o}_{\pi\tau} \cdot \boldsymbol{\varsigma}_{\pi\tau} \\
\boldsymbol{\varsigma}_{\pi\tau} &= \ln \mathbf{o}_{\pi\tau} - \ln \mathbf{C}_\tau \\
\mathbf{H} &= -diag(\mathbf{A} \cdot \ln \mathbf{A}) \\
P(o_\tau \mid C) &= Cat(\mathbf{C}_\tau) \\
Q(o_\tau \mid \pi) &= Cat(\mathbf{o}_{\pi\tau}), \ \mathbf{o}_{\pi\tau} = \mathbf{A}\mathbf{s}_{\pi\tau} \\
Q(s_\tau \mid \pi) &= Cat(\mathbf{s}_{\pi\tau}) \\
Q(s_\tau) &= Cat(\mathbf{s}_\tau), \quad \mathbf{s}_\tau = \sum\nolimits_\pi \boldsymbol{\pi}_\pi \mathbf{s}_{\pi\tau}
\end{aligned} \tag{4.10}$$

The first line of equation 4.10 uses a softmax (normalized exponential) operator to construct a probability distribution (parameterized with sufficient statistics $\boldsymbol{\pi}_0$) that sums to one from the expected free energy vector. Lines two to four express the components of the expected free energy in linear algebraic notation. The fifth line shows that the prior belief about observations is a categorical distribution (whose sufficient statistics are given in the $\mathbf{C}$ vector). The sixth to eighth lines specify the relationship between the linear algebraic quantities and the associated probability distributions. Having completed the specification of the generative model, we can now express the free energy in terms of the variables above:

$$\begin{aligned}
F &= \boldsymbol{\pi} \cdot \mathbf{F} \\
\mathbf{F}_\pi &= \sum\nolimits_\tau \mathbf{F}_{\pi\tau} \\
\mathbf{F}_{\pi\tau} &= \mathbf{s}_{\pi\tau} \cdot (\ln \mathbf{s}_{\pi\tau} - \ln \mathbf{A} \cdot o_\tau - \ln \mathbf{B}_{\pi\tau} \mathbf{s}_{\pi\tau-1})
\end{aligned} \tag{4.11}$$

The decomposition of this into a sum over time is due to the implicit mean-field approximation that assumes we can factorize the approximate posterior into a product of factors:

$$Q(\tilde{s} \mid \pi) = \prod\nolimits_\tau Q(s_\tau \mid \pi) \tag{4.12}$$

In logarithmic form, this becomes a sum, just as in equation 4.11. This factorization is one of many possibilities in variational inference—and represents the simplest option. In practice, this is often nuanced slightly, as detailed in appendix B.

4.4.2 Active Inference in a POMDP

Hitherto, we have defined the four key ingredients for a discrete-time generative model. These are the likelihood (**A**), transition probabilities (**B**), prior beliefs about observations (**C**), and prior belief about the initial state (**D**). Once these probability distributions are specified, a generic message passing scheme can be employed to minimize free energy and solve the POMDP. To make inferences about hidden states under a given policy, we set the rate of change of an auxiliary variable (**v**), which stands in for the log posterior (**s**), to be equal to the negative free energy gradient. A softmax (normalized exponential) function is then used to compute **s** from **v**.

$$\begin{aligned} \mathbf{s}_{\pi\tau} &= \sigma(\mathbf{v}_{\pi\tau}) \\ \dot{\mathbf{v}}_{\pi\tau} &= \boldsymbol{\varepsilon}_{\pi\tau} \triangleq -\nabla_s \mathbf{F}_{\pi\tau} \\ &= \ln \mathbf{A} \cdot o_\tau + \ln \mathbf{B}_{\pi\tau} \mathbf{s}_{\pi\tau-1} + \ln \mathbf{B}_{\pi\tau+1} \cdot \mathbf{s}_{\pi\tau+1} - \ln \mathbf{s}_{\pi\tau} \end{aligned} \tag{4.13}$$

Equation 4.13 can be regarded as an example of variational message passing (see box 4.1). To update beliefs about policies, we find the posterior that minimizes the free energy:

$$\begin{aligned} &\nabla_{\boldsymbol{\pi}} F = 0 \Leftrightarrow \\ &\boldsymbol{\pi} = \sigma(-\mathbf{G} - \mathbf{F}) \end{aligned} \tag{4.14}$$

For the simplest form of POMDP, equations 4.13 and 4.14 can be used to solve an Active Inference problem for any set of probability matrices; these may be thought of as describing perception and planning, respectively. We will unpack this in greater detail in the second part of the book, where we will provide worked examples of Active Inference for perception and planning (and other cognitive functions).

Figure 4.4's graphical representations of equations 4.10, 4.13, and 4.14 hint at possible neuronal implementations of free energy minimization in the brain—if one interprets nodes as neuronal populations, edges as synapses, and messages as synaptic exchanges. In later chapters we will consider the extension of this to factorized state-spaces, deep temporal models, and the optimization of the parameters of the generative model itself (learning).

4.5 Active Inference in Continuous Time

In the previous section, we dealt with the form Active Inference takes under a particular choice of generative model. These POMDPs are a useful way to

Box 4.1
Message passing and inference

Markov blankets

We encountered the concept of a Markov blanket in chapter 3. However, it is worth briefly reviewing the idea here. It relates to a system of multiple interacting variables. A Markov blanket for a given variable comprises a subset of those that interact with it. If we know everything about this subset, knowledge of anything outside this subset does not increase our knowledge of the variable of interest. The relevance here is that we can draw inferences about a variable in a graphical model based on local information about its Markov blanket. The blanket of a variable x are those variables that cause x (*parents*, $\rho(x)$), the variables that are caused by x (*children*, $\kappa(x)$), and the parents of x's children. Using this notation, two of the most common Bayesian message passing schemes used for approximate inference are defined as follows:

Variational message passing

$$\ln Q(x) = \mathbb{E}_{Q(\rho(x))}[\ln P(x \mid \rho(x))] + \mathbb{E}_{\frac{Q(\kappa(x))Q(\rho(\kappa(x)))}{Q(x)}}[\ln P(\kappa(x) \mid \rho(\kappa(x)))]$$

This involves messages from all constituents of the Markov blanket of x, including the parents (via the conditional probability of x given its parents) and the children. The latter depends on the conditional probability of the children of x given all of their parents—which include x. Note the expectation includes the children and parents of the children. As the parents of the children include x, we divide by $Q(x)$ to ensure the expectation includes the blanket only.

Belief propagation

$$\ln Q(x) = \ln \mu_\kappa(x) + \ln \mu_\rho(x)$$
$$\mu_\kappa(x) = \mathbb{E}_{\frac{\mu_\kappa(\kappa(x))\mu_\rho(\kappa(x))}{\mu_x(\kappa(x))}}[P(\kappa(x) \mid \rho(\kappa(x)))]$$
$$\mu_\rho(x) = \mathbb{E}_{\frac{\mu_\rho(\rho(x))\mu_\kappa(\rho(x))}{\mu_x(\rho(x))}}[P(x \mid \rho(x))]$$

This has broadly the same structure as variational message passing but uses a recursive definition of messages such that each message ($\mu_a(b)$ being the message to b from a) depends on other messages (the messages to a). There is a directional aspect to this, such that the message from a to b depends on all messages to a, except for that from b (hence the division in the expectations). NB: The slightly nonstandard use of the expectation operator here allows us to (1) cover both discrete and continuous variables and (2) highlight the formal similarity between variational message passing and belief propagation.

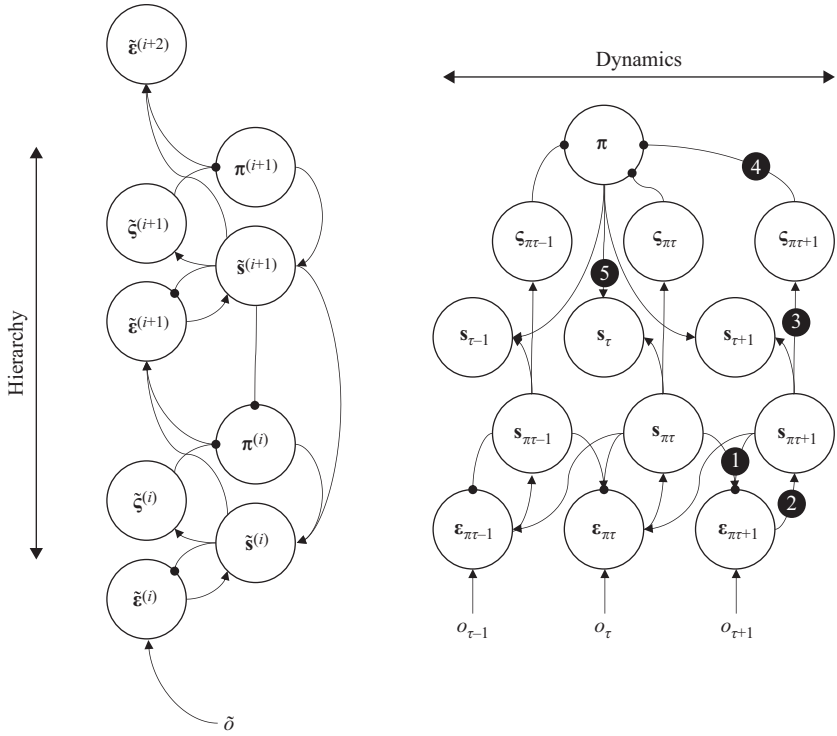

Figure 4.4

Bayesian message passing. *Right:* Dependencies between different variables in the belief-updating scheme outlined in the main text. Intuitively, current beliefs about states (under each policy) at each time are compared with those that would be predicted given beliefs about states at other times (1) and current outcomes to calculate prediction errors. These errors then drive updating in these beliefs (2); given beliefs about states under each policy, we can then calculate the gradients of the expected free energy (3). These are combined with the outcomes predicted under each policy (omitted from the figure) to compute beliefs about policies (4). Using a Bayesian model average, we can then compute posterior beliefs about states averaged over policies (5). This high-level summary of message passing omits some intermediate connections that could be included (e.g., connection (4) could be unpacked to explicitly include computation of the expected free energy). *Left:* This scheme could be expanded hierarchically (collapsing over time steps and policies for simplicity). The key idea is that a higher-level network might predict the states and policies at the lower level and use these to draw inferences about the context in which these occur. We will unpack this idea further in chapter 7.

articulate a range of inference problems, including those that underwrite planning and decision-making. However, when it comes to interacting with a real environment, models described in discrete time with categorical variables fall short. This is because sensory input and motor outputs are continuously evolving variables. To account for this, we now turn to a different sort of generative model. We apply exactly the same idea, a gradient descent on variational free energy, to these models to find the analogous message passing schemes.

4.5.1 A Generative Model for Predictive Coding

To motivate the form of generative model used for continuous states, we start with the following pair of equations:

$$\begin{aligned} \dot{x} &= f(x,v) + \omega_x \\ y &= g(x,v) + \omega_y \end{aligned} \tag{4.15}$$

The first of these expresses the evolution of a hidden state over time, according to a deterministic function ($f(x,v)$) and stochastic fluctuations (ω). The second equation expresses the way in which data are generated from the hidden state. In each case, the fluctuations are assumed normally distributed, giving the following probability densities for the dynamics and likelihood:

$$\begin{aligned} p(\dot{x}|x,v) &= \mathcal{N}(f(x,v), \Pi_x) \\ p(y|x,v) &= \mathcal{N}(g(x,v), \Pi_y) \end{aligned} \tag{4.16}$$

The precision (Π) terms are the inverse covariance of the fluctuations. These two equations form the generative model that underwrite Kalman-Bucy filters in engineering. However, schemes of this sort are limited by the assumption of uncorrelated fluctuations over time (i.e., Wiener assumptions). This is inappropriate for inference in biological systems, where fluctuations are themselves generated by dynamical systems and have a degree of smoothness. We can account for this by considering not only the rate of change of the hidden state and the current value of the data but also their velocities, accelerations, and subsequent temporal derivatives—that is, generalized coordinates of motion (Friston, Stephan et al. 2010; see box 4.2):

Box 4.2
Generalized coordinates of motion

To represent a trajectory in continuous time, generalized coordinates of motion provide a simple parameterization. This is based on a polynomial (Taylor series) expansion around the present time to give a function that lets us extrapolate to the recent past and near future. The plots in figure 4.5 show a trajectory in some space (x) over time (τ) as a solid line. From left to right, they show the trajectory represented in generalized coordinates of motion with one, two, and three coordinates (successive temporal derivatives of x). This is the dashed line. The expansion here is around the initial time point. With each successive generalized coordinate, we get a more accurate approximation of the trajectory into the proximal future. For most applications, around six generalized coordinates are sufficient.

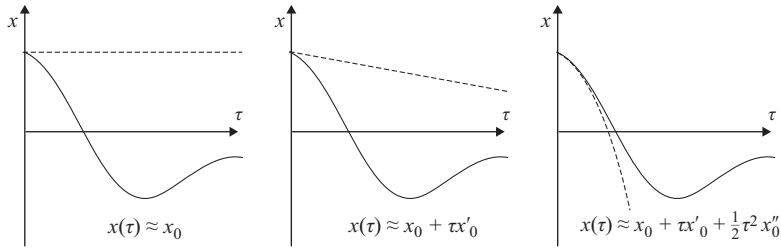

Figure 4.5

$$\begin{aligned}
\dot{x} &= f(x,v) + \omega_x & y &= g(x,v) + \omega_y \\
\dot{x}' &= f'(x',v') + \omega_x' & y' &= g'(x',v') + \omega_y' \\
\dot{x}'' &= f''(x'',v'') + \omega_x'' & y'' &= g''(x'',v'') + \omega_y'' \\
&\vdots & &\vdots \\
\dot{x}^{[i]} &= f^{[i]}(x^{[i]},v^{[i]}) + \omega_x^{[i]} & y^{[i]} &= g^{[i]}(x^{[i]},v^{[i]}) + \omega_y^{[i]} \\
&\vdots & &\vdots
\end{aligned} \tag{4.17}$$

These generalized coordinates can be summarized more succinctly by representing a trajectory (again using the ~ symbol) as a vector with elements corresponding to the successive derivatives above:

$$\left.\begin{aligned} D\tilde{x} &= \tilde{f}(\tilde{x},\tilde{v}) + \tilde{\omega}_x \\ \tilde{y} &= \tilde{g}(\tilde{x},\tilde{v}) + \tilde{\omega}_y \end{aligned}\right\} \Rightarrow \begin{aligned} p(\tilde{x}\,|\,\tilde{v}) &= \mathcal{N}(D\cdot\tilde{f},\tilde{\Pi}_x) \\ p(\tilde{y}\,|\,\tilde{x},\tilde{v}) &= \mathcal{N}(\tilde{g},\tilde{\Pi}_y) \end{aligned} \tag{4.18}$$

In equation 4.18, D is a matrix with ones above the leading diagonal and zeros elsewhere. This effectively shifts all elements of the vector upward and may be thought of as a derivative operator. The generalized precision matrices may be constructed on the basis of the smoothness we assume for the fluctuations, as detailed in appendix B. Equipped with a prior over the hidden cause (v), whose relevance will become clearer below, this lets us write down the free energy for this generative model:

$$\begin{aligned} F[\mu, y] &= -\ln p(\tilde{y}, \tilde{\mu}_x, \tilde{\mu}_v) \\ &= \tfrac{1}{2}\tilde{\varepsilon} \cdot \tilde{\Pi}\tilde{\varepsilon} \\ &= \tfrac{1}{2}\left(\tilde{\varepsilon}_y \cdot \tilde{\Pi}_y \tilde{\varepsilon}_y + \tilde{\varepsilon}_x \cdot \tilde{\Pi}_x \tilde{\varepsilon}_x + \tilde{\varepsilon}_v \cdot \tilde{\Pi}_v \tilde{\varepsilon}_v\right) \\ \tilde{\varepsilon} &= \begin{bmatrix} \tilde{\varepsilon}_y \\ \tilde{\varepsilon}_x \\ \tilde{\varepsilon}_v \end{bmatrix} = \begin{bmatrix} \tilde{y} - \tilde{g}(\tilde{\mu}_x, \tilde{\mu}_v) \\ D\tilde{\mu}_x - \tilde{f}(\tilde{\mu}_x, \tilde{\mu}_v) \\ \tilde{\mu}_v - \tilde{\eta} \end{bmatrix} \\ \tilde{\Pi} &= \begin{bmatrix} \tilde{\Pi}_y & & \\ & \tilde{\Pi}_x & \\ & & \tilde{\Pi}_v \end{bmatrix} \end{aligned} \tag{4.19}$$

In equation 4.19, the μ terms indicate the mode of the approximate posterior density for the x and v terms. The reason the free energy takes such a simple form in the first line is that we have employed a Laplace approximation, as detailed in box 4.3. In brief, this treats all probability densities as Gaussian, which—through a Taylor series expansion—is equivalent to assuming we are operating close to the mode of the distribution. The second line of the equation expresses the log probability in terms of squared precision weighted prediction errors. This omits all terms that are constant with respect to the posterior mode. The third line unpacks this in terms of the log likelihood, log probability of x given v, and log prior of v.

4.5.2 Active Inference as Predictive Coding with Motor Reflexes

Because the variance of the approximate posterior is an analytic function of the mode, under the Laplace approximation, we can optimize the free energy with respect to the mode. A simple way to think about this is that we need only find the *maximum a posteriori* (MAP) estimates[5] for each state. These are the means of the posterior distribution that may be equipped with its precision without need for further inference via the Laplace approximation (see box 4.3).

Box 4.3
The Laplace approximation

Laplace approximations rely on a principle similar to the generalized coordinates of motion described in box 4.2. The idea is that the free energy may be approximated by a quadratic expansion around the posterior mode (μ). In one dimension, this is as follows:

$$\begin{aligned} F[y,q] &= \mathbb{E}_{q(x)}[\ln q(x) - \ln p(y,x)] \\ &\approx \mathbb{E}_{q(x)}\Big[\ln q(\mu) + (x-\mu)\underbrace{\partial_x \ln q(x)\big|_{x=\mu}}_{=0} + \tfrac{1}{2}(x-\mu)^2\,\partial_x^2 \ln q(x)\big|_{x=\mu} \\ &\quad - \ln p(y,\mu) - (x-\mu)\partial_x \ln p(y,x)\big|_{x=\mu} - \tfrac{1}{2}(x-\mu)^2\,\partial_x^2 \ln p(y,x)\big|_{x=\mu} \Big] \end{aligned}$$

The assumption that a quadratic expansion is sufficient is equivalent to saying that we can treat the probabilities as Gaussian (as the log of a Gaussian density is quadratic). Making this explicit, we can simplify the above to the following:

$$\begin{aligned} q(x) &= \mathcal{N}(\mu, \Sigma^{-1}) \\ F[y,\mu] &= -\ln 2\pi\Sigma - \ln p(y,\mu) - \frac{1}{2} tr\Big[\Sigma \partial_x^2 \ln p(y,x)\big|_{x=\mu}\Big] \end{aligned}$$

Under quadratic assumptions, the only term that depends on the mode is the second term. Omitting the other terms leads to the expression in equation 4.19. We can find the precision of the approximate posterior directly, once we know the mode, through the following expansion:

$$\begin{aligned} \ln q(x) &\approx \ln p(x|y) \\ &= \ln p(x,y) - \ln p(y) \\ &\approx \ln p(\mu,y) + (x-\mu)\cdot \underbrace{\partial_x \ln p(x,y)\big|_{x=\mu}}_{=0} \\ &\quad + \tfrac{1}{2}(x-\mu)\cdot \partial_x^2 \ln p(x,y)\big|_{x=\mu}\,(x-\mu) - \ln p(y) \\ &\Rightarrow q(x) \propto e^{-\frac{1}{2}(x-\mu)\cdot\Sigma^{-1}(x-\mu)}, \quad \Sigma^{-1} = -\partial_x^2 \ln p(x,y)\big|_{x=\mu} \end{aligned}$$

This tells us that the posterior precision is simply the second derivative of the joint probability evaluated at the posterior mode.

$$\begin{aligned} \dot{\tilde{\mu}} - D\tilde{\mu} &= -\nabla_{\tilde{\mu}} F \\ &= \nabla_{\tilde{\mu}} \ln p(\tilde{y},\tilde{\mu}) \\ &= -\nabla_{\tilde{\mu}}\, \tilde{\varepsilon}\cdot\tilde{\Pi}\tilde{\varepsilon} \\ \begin{bmatrix} \dot{\tilde{\mu}}_x - D\tilde{\mu}_x \\ \dot{\tilde{\mu}}_v - D\tilde{\mu}_v \end{bmatrix} &= \begin{bmatrix} \nabla_{\tilde{\mu}_x}\tilde{g}\cdot\tilde{\Pi}_y\tilde{\varepsilon}_y - D\cdot\tilde{\Pi}_x\tilde{\varepsilon}_x + \nabla_{\tilde{\mu}_x}\tilde{f}\cdot\tilde{\Pi}_x\tilde{\varepsilon}_x \\ \nabla_{\tilde{\mu}_v}\tilde{g}\cdot\tilde{\Pi}_y\tilde{\varepsilon}_y + \nabla_{\tilde{\mu}_v}\tilde{f}\cdot\tilde{\Pi}_x\tilde{\varepsilon}_x - \tilde{\Pi}_v\tilde{\varepsilon}_v \end{bmatrix} \end{aligned} \tag{4.20}$$

In contrast to the gradient descents we saw for the discrete-time scheme, the left-hand side of equation 4.20 is the difference between the rate of change of μ and the derivative operator applied to this. This is because when the free energy is minimized, it does not make sense for the rate of change of the posterior mode to be zero if the posterior mode associated with rates of change is nonzero. In other words, "the motion of the mode should be the mode of the motion" at the free energy minimum. This ensures $\dot{\mu}^{[i]} = \mu^{[i+1]}$ when free energy is minimized.

We can go one step further than equation 4.20 and treat the hidden cause (v) as if it were data being generated by a higher hierarchical level, with slower dynamics (such that v appears not to change at the lower level). In doing so, we can chain together a hierarchy of equations:

$$\begin{bmatrix} \vdots \\ \dot{\tilde{\mu}}_x^{(i)} - D\tilde{\mu}_x^{(i)} \\ \dot{\tilde{\mu}}_v^{(i)} - D\tilde{\mu}_v^{(i)} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \nabla_{\tilde{\mu}_x^{(i)}} \tilde{g}^{(i)} \cdot \tilde{\Pi}_v^{(i-1)} \tilde{\varepsilon}_v^{(i-1)} - D \cdot \tilde{\Pi}_x^{(i)} \tilde{\varepsilon}_x^{(i)} + \nabla_{\tilde{\mu}_x^{(i)}} \tilde{f}^{(i)} \cdot \tilde{\Pi}_x^{(i)} \tilde{\varepsilon}_x^{(i)} \\ \nabla_{\tilde{\mu}_v^{(i)}} \tilde{g}^{(i)} \cdot \tilde{\Pi}_v^{(i-1)} \tilde{\varepsilon}_v^{(i-1)} + \nabla_{\tilde{\mu}_v^{(i)}} \tilde{f}^{(i)} \cdot \tilde{\Pi}_x^{(i)} \tilde{\varepsilon}_x^{(i)} - \tilde{\Pi}_v^{(i)} \tilde{\varepsilon}_v^{(i)} \\ \vdots \end{bmatrix} \tag{4.21}$$

$$\begin{bmatrix} \tilde{\varepsilon}_x^{(i)} \\ \tilde{\varepsilon}_v^{(i)} \end{bmatrix} = \begin{bmatrix} D\tilde{\mu}_x^{(i)} - f^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}) \\ \tilde{\mu}_v^{(i)} - g^{(i+1)}(\tilde{\mu}_x^{(i+1)}, \tilde{\mu}_v^{(i+1)}) \end{bmatrix}$$

$$\tilde{\varepsilon}_v^{(0)} \triangleq \tilde{\varepsilon}_y$$

Figure 4.6 graphically emphasizes the role of the hidden states (x) in linking together temporal derivatives within one hierarchical level and the role of the hidden causes (v) in linking hierarchical levels together. In this predictive coding scheme (Rao and Ballard 1999, Friston and Kiebel 2009), higher levels send descending predictions to lower levels, which compute errors in these predictions and pass these errors back up the hierarchy to update beliefs.

To complete our overview of predictive coding in the context of Active Inference, we need to incorporate action. Given that our aim is to minimize free energy and that the consequences of action are that we change our sensory data, we have the following:

$$\begin{aligned} \dot{u} &= -\nabla_u F \\ &= -\nabla_u \tilde{y}(u) \cdot \tilde{\Pi}_y \tilde{\varepsilon}_y \end{aligned} \tag{4.22}$$

This equation says that we minimize free energy through action and that the only part of the free energy that depends directly on action is the lowest level of prediction error. In other words, action simply fulfills descending

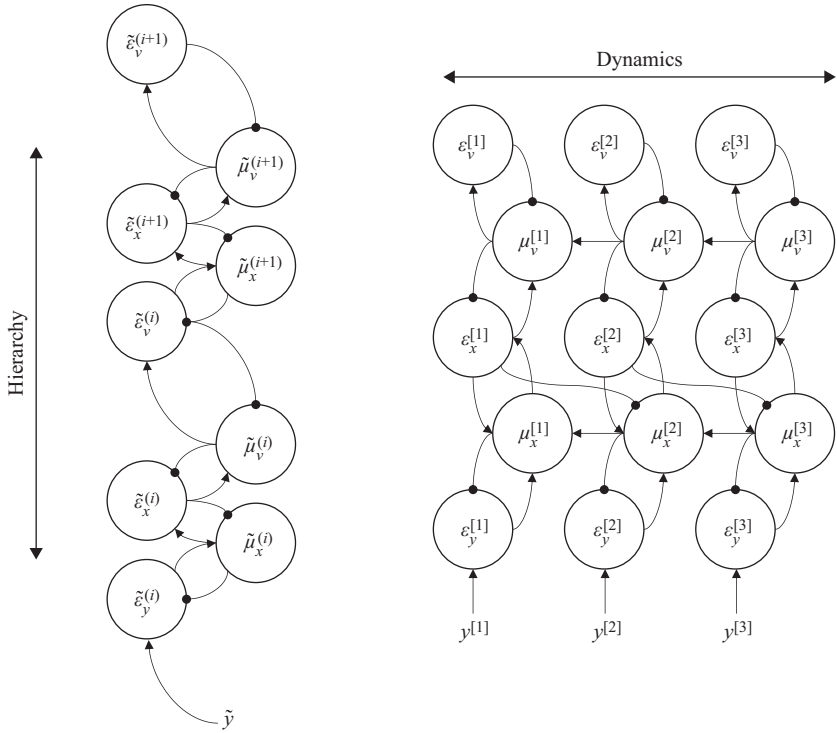

Figure 4.6
Message passing of generalized predictive coding schemes. *Left:* Computation of prediction errors from sensory data, showing how these may be propagated upward through a hierarchy. Higher levels send predictions to the lower levels that may be compared with sensory data to compute these errors. *Right:* A single layer of the hierarchy illustrates how neuronal populations representing different orders of generalized motion interact with one another.

predictions about data through minimizing the error between the predicted and observed sensory consequences of action. One way to think about this is as if we had equipped a predictive coding scheme with classical reflex arcs at the lowest level of the hierarchy (Adams, Shipp, and Friston 2013). In this setting, Active Inference is just predictive coding plus reflex arcs. From a neurobiological perspective, the idea is that sensory afferents enter the brain stem or spinal cord and synapse on motor neurons. Descending predictions of the sensory input are propagated from the cortex to the motor neurons, whose output depends on the difference between their cortical and sensory inputs.

From a computational perspective, a reflex arc is one of the simplest possible forms of controller; these correct deviations in predicted and observed proprioceptive signals. More complex motor behavior requires generating sequences of predictions and fulfilling them in order using reflex arcs. This mechanism sets active inference apart from other schemes for biological motor control, such as optimal control, which are not based on predictive coding and use inverse models and controllers that are more complex than reflex arcs (Friston 2011). Another peculiar characteristic of Active Inference is that it dispenses from notions of value or cost used in optimal control (and reinforcement learning); these are fully absorbed into the (generally more expressive) notion of priors (see chapter 10 for further discussion).

4.6 Summary

This chapter outlined the basic formal ideas that underwrite Active Inference. The key message to take away is that (approximate) Bayesian inference may be framed as minimizing a quantity known as variational free energy. This depends on a generative model that expresses our beliefs about how data are generated. We have looked at two forms of a generative model that may be employed depending on the inference problem at hand: specifically, whether we are interested in categorical or continuous variables. The free energy minimizing solution to either can be unpacked in terms of message passing between populations of neurons, including the generalized predictive coding schemes that follow from continuous models. Finally, we noted that free energy can be minimized not just by changing beliefs—such that they become consistent with data—but also by acting on the world to make data more consistent with beliefs. Over subsequent chapters, we will appeal to the formalisms introduced here and apply them to more concrete settings, providing an opportunity to explore the extensions of the broad concepts set out here.