Appendix A: Mathematical Background

A.1 Introduction

This appendix offers an introduction (or a refresher) to the basic mathematical techniques employed throughout this book. We provide an introductory (but nonexhaustive) overview of four topics: linear algebra, Taylor series approximation, variational calculus, and stochastic dynamics. For each of these techniques, we refer to where it comes into play in the book. Our aim here is to provide a focused introduction—with emphasis on building intuition as opposed to formal and rigorous proofs. The maths required to understand and use Active Inference is not complicated, but its multidisciplinary basis means it is often difficult to find resources that bring together the necessary prerequisites. We hope this appendix goes some way toward remedying this.

A.2 Linear Algebra

A.2.1 The Basics

Linear algebra refers to a notation used to simply and concisely express combinations of multiplications and summations. It relies on matrices and vectors comprising arrays of numbers in structures with multiple rows and columns (or multiple rows and a single column, for a vector). The element of a matrix A in the ith row and jth column is referred to as A_{ij}. The product A of two matrices B and C (or a matrix and vector) is defined as follows:

$$\begin{aligned} &A = BC \\ &\Rightarrow \\ &A_{ij} = \sum_k B_{ik} C_{kj} \end{aligned} \tag{A.1}$$

For this definition to hold, we need the number of columns of B to match the number of rows of C. However, let us instead say that the number of columns of B match the columns in C and we want to express the following sum:

$$A_{ij} = \sum_k B_{ki} C_{kj} \tag{A.2}$$

How would we do this using linear algebraic notation? We need to appeal to another operation that swaps the subscripted indices of B (i.e., reflects the array such that the columns become rows and vice versa). This is the transpose operation, normally expressed using a superscript T:

$$\begin{aligned} B^T_{ik} &\triangleq B_{ki} \\ A = B^T C &\triangleq B \cdot C \\ &\Rightarrow \\ A_{ij} &= \sum_k B_{ki} C_{kj} \end{aligned} \tag{A.3}$$

Equation A.3 shows how we can use the transpose operator to express the summation from equation A.2. The second line highlights an alternative notation using a dot operator. This notation is inspired by the fact that, when B and C have only one column each, equation A.3 reduces to a vector dot product.

Another useful operation is the *trace* operator. This takes the elements along the diagonal of a square matrix and sums them:

$$tr[A] \triangleq \sum_i A_{ii} \tag{A.4}$$

Part of the utility of a trace operator is afforded by the way we can permute elements in the trace of a matrix product:

$$\begin{aligned} tr[ABC] &= \sum_i \sum_j \sum_k A_{ij} B_{jk} C_{ki} \\ &= \sum_k \sum_i \sum_j C_{ki} A_{ij} B_{jk} = tr[CAB] \\ &= \sum_j \sum_k \sum_i B_{jk} C_{ki} A_{ij} = tr[BCA] \end{aligned} \tag{A.5}$$

The main use we will find for this identity in this book is when applied to scalar quantities. A scalar can be viewed as a matrix with only one row and one column. As such, we can apply a trace operator to it, but this will not do anything—we get the same scalar out. This means that, if a matrix product gives rise to a scalar quantity, we can permute the terms as above.

For example, if we have a square matrix B with N columns and rows, and a vector c with N rows, we can use equation A.5 to show the following:

$$\begin{aligned} a &= c \cdot Bc \\ &= tr\left[c^T Bc\right] \\ &= tr\left[Bcc^T\right] \\ &= tr\left[BC\right] \\ C &= c \otimes c \triangleq cc^T \end{aligned} \tag{A.6}$$

This reexpresses a quadratic expression (first line) with the trace of the product of two matrices (penultimate line). The final line defines the outer product (in contrast to the inner dot product). Equation A.6 becomes particularly useful in the context of multivariate normal distributions, as we will come to in section A.2.3.

The final concepts of linear algebra to be aware of are the inverse and determinant of a matrix. An inverse is defined as follows:

$$A^{-1}A = AA^{-1} = I \tag{A.7}$$

Equation A.7 says that the product of a matrix and its inverse is the identity matrix—a square matrix with ones along its main and zeros elsewhere. Multiplying any matrix by the identity matrix returns the original matrix, unchanged. It is the linear algebraic equivalent of scalar multiplication by 1 (which could be interpreted as a 1-dimensional identity matrix). This means that if we multiply something by a matrix, and then by the inverse of that matrix, we end up with the original quantity.

The determinant is a useful quantity but one for which it is harder to develop a clear intuition. The only point at which it appears in this book is as part of the normalizing constant of a multivariate normal distribution. As such, it is worth knowing how it is calculated, but we will not dwell on this concept. The determinant is defined recursively as follows:

$$\left|A\right| \triangleq \sum_i (-1)^{i-1} A_{1i} \left|A_{\setminus(1,i)}\right| \tag{A.8}$$

Here, the notation $A_{\setminus(1,i)}$ means the matrix A with row 1 and column i omitted. For example:

$$
\begin{aligned}
A &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \\
A_{\backslash(1,1)} &= A_{22} \\
A_{\backslash(1,2)} &= A_{21} \\
|A| &= A_{11} \left|A_{22}\right| - A_{12} \left|A_{21}\right| \\
&= A_{11}A_{22} - A_{12}A_{21}
\end{aligned}
\tag{A.9}
$$

This concludes our outline of the basic operations of linear algebra.

A.2.2 Derivatives

Differentiation of matrix and vector quantities follows directly from the application of standard calculus to each element of a matrix. For example, if we have a matrix B whose elements are functions of a scalar x, the derivative of B with respect to x is as follows:

$$
\begin{aligned}
& A(x) = \partial_x B(x) \\
\Rightarrow\ & A(x)_{ij} = \partial_x B(x)_{ij} \\
& \partial_x \triangleq \frac{\partial}{\partial x}
\end{aligned}
\tag{A.10}
$$

However, a few important definitions and identities will be useful in understanding the technical details in this book. The first is how to take derivatives with respect to nonscalar quantities. If we have a vector quantity b that is a function of another vector c, the derivative of b with respect to c is a matrix:

$$
\begin{aligned}
& A = \partial_c b(c) \\
\Rightarrow\ & A_{ij} = \partial_{c_j} b(c)_i
\end{aligned}
\tag{A.11}
$$

We will also make use of the gradient operator, which deals with derivatives with respect to a vector. This is defined as follows:

$$
\begin{aligned}
\nabla_b &= \begin{bmatrix} \partial_{b_1} & \partial_{b_2} & \partial_{b_3} & \cdots \end{bmatrix}^T \\
a &= \nabla_b x(b) \\
&\Rightarrow \\
a_i &= \partial_{b_i} x(b)
\end{aligned}
\tag{A.12}
$$

The definition of the gradient operator as a vector of derivative operators also affords a concise definition of a related quantity—the divergence of a vector function:

$$\nabla_a \cdot b(a) = \sum_i \partial_{a_i} b(a)_i \tag{A.13}$$

There are many useful derivative identities for linear algebraic quantities, but we will not attempt to provide a comprehensive overview; for readers who wish to delve further, we recommend *The Matrix Cookbook* (Petersen and Pedersen 2012). Here, we limit ourselves to two identities that will be particularly useful. The first is the gradient of a quadratic quantity:

$$\begin{aligned} & d(a) = \nabla_a \big(b(a) \cdot Cb(a)\big) \\ & \quad \Rightarrow \\ & d(a)_i = \partial_{a_i} \sum_j \sum_k b(a)_j C_{jk} b(a)_k \\ & \quad = \sum_j \sum_k \Big(\big(\partial_{a_i} b(a)_j\big) C_{jk} b(a)_k + \big(\partial_{a_i} b(a)_k\big) C_{jk} b(a)_j\Big) \\ & \quad \Rightarrow \\ & d(a) = \nabla_a b(a) \cdot \big(C + C^T\big) b(a) \end{aligned} \tag{A.14}$$

Here (and throughout this book), the transposition implied by the dot notation is applied prior to the gradient operator:

$$\nabla_a b(a) \cdot (\cdots) \triangleq \nabla_a b(a)^T (\cdots) \neq \big(\nabla_a b(a)\big)^T (\cdots) \tag{A.15}$$

The identity in equation A.14 is used in the derivation of the belief-update equations for predictive coding in chapter 4. A second useful identity is the derivative of the same quantity with respect to the matrix, C:

$$\begin{aligned} & D(a) = \nabla_C \big(b(a) \cdot Cb(a)\big) \\ & \quad \Rightarrow \\ & D(a)_{ij} = \partial_{C_{ij}} \sum_k \sum_l b(a)_k C_{kl} b(a)_l = b(a)_i b(a)_j \\ & \quad \Rightarrow \\ & D(a) = b(a) \otimes b(a) \end{aligned} \tag{A.16}$$

Here we have used the gradient operator with a matrix subscript to indicate the following:

$$\nabla_C = \begin{bmatrix} \partial_{C_{11}} & \partial_{C_{12}} & \cdots \\ \partial_{C_{21}} & \partial_{C_{22}} & \\ \vdots & & \ddots \end{bmatrix} \tag{A.17}$$

In appendix B, we will see how equation A.16 aids in the estimation of the covariance matrix for a posterior probability.

A.2.3 Probabilities

In the context of probabilistic reasoning, these linear algebraic identities come into play in two important situations. The first is when the random variable we are reasoning about (i.e., the support of a probability distribution) is a vector quantity. The second is when the probability distribution itself is described by sufficient statistics that are vectors, matrices, or higher order tensor quantities.[1] An example of both is the multivariate normal distribution, defined as follows:

$$\begin{aligned} p(x) &= \left(\frac{1}{(2\pi)^k}|\Pi|\right)^{\frac{1}{2}} e^{-\frac{1}{2}(x-\eta)\cdot\Pi(x-\eta)} \\ \dim(x) &= k \end{aligned} \tag{A.18}$$

Here, x is a k-dimensional vector. This means the mode, η, is also a k-dimensional vector. The precision, Π, is the inverse of the covariance—a $k \times k$ dimensional symmetric matrix expressing the dispersion of probability mass around the mode. This appears twice in equation A.18: in the normalizing constant (as a determinant) and in the exponent. Note that the quadratic term in the exponent is a scalar quantity and is therefore susceptible to the identity in equation A.6. This will be important in appendix B.

When dealing with categorical probability distributions, the sufficient statistics of a distribution are simply vectors, matrices, or tensors of probabilities. For example, the probability distribution over the numbers a person could roll on a six-sided die is given by a 6-dimensional vector, with each element of the vector expressing the probability of that number. Things get more interesting in the context of conditional probabilities. For variables o and s, which each take one of several possible values, we can write the conditional probability of o given s as a matrix, A, whose elements are as follows:

$$P(o = i \mid s = j) = A_{ij} \tag{A.19}$$

This says that the probability that o takes its ith possible value if s takes its jth possible value is given by the element of A in the ith row and jth column. Taking this further, we can define conditional probabilities in which there are multiple items in the conditioning set, leading to a tensor structure:

$$P(o = i \mid s_1 = j, s_2 = k, s_3 = l, \ldots) = A_{ijkl\ldots} \tag{A.20}$$

Here we could specify an arbitrary number of variables in the conditioning set, leading to an arbitrary number of indices, and a tensor of arbitrary order. We set out an example of a (T-maze) model in chapter 7 that makes use of a probability tensor of order 3. The principles of this model generalize to any higher order. For a tensor A, we will consistently use the dot notation of equation A.3 to mean summation with respect to the first index:

$$\begin{aligned} A &= B \cdot x \\ &\Rightarrow \\ A_{jklm\ldots} &= \sum\nolimits_i B_{ijklm\ldots} x_i \end{aligned} \tag{A.21}$$

An advantage of this expression of distributions as arrays of numbers is that we can use the definitions in sections A.2.1–A.2.2 to find concise expressions for related quantities. For example, we will often need to compute information-theoretic quantities like entropies for probability distributions. An entropy is a negative expected (average) log probability. If we take the expression in equation A.19, we can find a simple form for its entropy as follows:

$$\begin{aligned} H\big[P(o|s)\big] &\triangleq -\mathbb{E}_{P(o|s)}\big[\ln P(o|s)\big] \\ \mathbf{H}_j &\triangleq H\big[P(o|s=j)\big] \\ &= -\sum\nolimits_i P(o=i|s=j)\ln P(o=i|s=j) \\ &\Rightarrow \\ \mathbf{H} &= -diag(A \cdot \ln A) \end{aligned} \tag{A.22}$$

In equation A.22, *diag* is an operation that takes the diagonal elements of a matrix and stacks them into a vector. This illustrates an example in chapter 4 of defining the expected free energy, in which an appeal to linear algebraic notation offers a concise description of how these quantities may be calculated.

A.3 Taylor Series Approximation

A.3.1 Introduction

Often, it is convenient to simplify the form of a function ($f(x)$) through an approximation (indicated by ^) that is valid in a local region (e.g., the region around a point, a). If we were only interested in the function at a, we

could replace the function with a constant equal to the function evaluated at that point:

$$\hat{f}(x) = f(a) \tag{A.23}$$

However, this is only valid when x is exactly equal to a. In order to make the approximation valid in the region immediately surrounding a, we can add a term to ensure that a small change in x is accompanied by a change in the value of the function consistent with the gradient at a:

$$\begin{aligned} \hat{f}(x) &= f(a) + \varepsilon \partial_x f(x)\big|_{x=a} \\ \varepsilon &\triangleq x - a \end{aligned} \tag{A.24}$$

When x is equal to a, the ε term is zero, consistent with equation A.23. In addition, the first derivative of the original function and of the approximation are equal, when evaluated at a.

Pursuing this approach, we can add an additional term that accounts for the rate of change of the gradient (i.e., the curvature) so that the approximation becomes valid for a greater deviation from a. We do not have to stop here; we could add an arbitrary number of terms to match each successive derivative between the original function and the approximation:

$$\begin{aligned} \hat{f}(x) &= f(a) + \varepsilon \partial_x f(x)\big|_{x=a} + \frac{1}{2}\varepsilon^2 \partial_x^2 f(x)\big|_{x=a} + \cdots \\ &= \sum\nolimits_{n=0} \tfrac{1}{n!} \varepsilon^n \partial_x^n f(x)\big|_{x=a} \end{aligned} \tag{A.25}$$

Equation A.25 shows the Taylor series expansion in one dimension. However, we can generalize this to the multivariate case (where x is a vector) with the following expression:

$$f(x) \approx f(a) + \varepsilon \cdot \nabla_x f(x)\big|_{x=a} + \frac{1}{2}\varepsilon \cdot \nabla_x \left(\nabla_x f(x)\right)^T\Big|_{x=a} \varepsilon + \cdots \tag{A.26}$$

The quantity $\nabla_x(\nabla_x f(x))^T$ is known as a Hessian matrix.

Increasing the number of terms in the series improves the approximation. For our purposes, we need not go beyond the second order (quadratic) expansion. In the following subsections, we highlight the places in this book in which this approximation has been exploited. These include the Laplace approximation, which underwrites the predictive coding schemes described in chapters 4 and 8 and the variational Laplace scheme used for

model-based data analysis described in chapter 9. In addition, the generalized coordinates of motion used to model continuous trajectories (box 4.2) can be interpreted as Taylor series coefficients. We will unpack these applications in sections A.3.2 and A.3.3, respectively.

A.3.2 The Laplace Approximation

An important application of a Taylor series approximation in probabilistic inference is its use in the Laplace approximation. This refers to the use of a Gaussian distribution to approximate a probability distribution (p) in the region surrounding its mode (μ). If we expand the log of a probability distribution using equation A.26, we get the following:

$$\ln p(x) \approx \ln p(\mu) + \varepsilon \cdot \nabla_x \ln p(x)\Big|_{x=\mu} + \frac{1}{2}\varepsilon \cdot \nabla_x \left(\nabla_x \ln p(x)\right)^T\Big|_{x=\mu} \varepsilon \qquad \varepsilon \triangleq x - \mu \tag{A.27}$$

This is simply equation A.26 but with $f(x) = \ln p(x)$ and $a = \mu$. The first term after the approximate equality is constant with respect to x so may be absorbed into a normalizing constant. The second term disappears, as the gradient of the log probability at its mode is zero. Exponentiating both sides leaves us with this:

$$\begin{aligned} p(x) &\approx \frac{1}{Z} e^{-\frac{1}{2}\varepsilon \cdot C^{-1}\varepsilon} \\ &= \mathcal{N}(\mu, C^{-1}) \\ C^{-1} &\triangleq -\nabla_x \left(\nabla_x \ln p(x)\right)^T\Big|_{x=\mu} \end{aligned} \tag{A.28}$$

Equation A.28 says that when we approximate a log probability using a quadratic function, near its mode, the associated probability density is Gaussian. This is the Laplace approximation applied to a probability distribution. However, we can also apply the Laplace approximation to a free energy functional. To provide some intuition for this, we start with a free energy functional (see chapter 4):

$$F[q, y] = \mathbb{E}_{q(x)}[\ln q(x) - \ln p(y, x)] \tag{A.29}$$

Equation A.29 expresses free energy in terms of the expected difference between two log probabilities. The q density is an approximate posterior probability. The p density is a generative model, describing how hidden states

(x) give rise to data (y). As in equation A.27, we can apply a Taylor series expansion to the two log probabilities. Starting with the variational density, we have this:

$$\begin{aligned}
\ln q(x) &\approx \ln q(\mu) + (x-\mu)\cdot \underbrace{\nabla_x \ln q(x)\big|_{x=\mu}}_{0} \\
&\quad + \tfrac{1}{2}(x-\mu)\cdot \nabla_x \left(\nabla_x \ln q(x)\right)^T\Big|_{x=\mu}(x-\mu) \\
\Rightarrow q(x) &\approx \mathcal{N}\left(\mu, \Sigma^{-1}\right) \\
\Sigma^{-1} &= -\nabla_x \left(\nabla_x \ln q(x)\right)^T\Big|_{x=\mu} \\
\mu &= \arg\max_x q(x)
\end{aligned} \tag{A.30}$$

Applying the expectation from equation A.29 to equation A.30, we get this:

$$\begin{aligned}
\mathbb{E}_{q(x)}\left[\ln q(x)\right] &\approx \ln q(\mu) - \tfrac{1}{2}\mathbb{E}_{q(x)}\left[(x-\mu)\cdot\Sigma^{-1}(x-\mu)\right] \\
&= \ln q(\mu) - \tfrac{1}{2} tr\left[\Sigma^{-1}\underbrace{\mathbb{E}_{q(x)}\left[(x-\mu)(x-\mu)^T\right]}_{\Sigma}\right] \\
&= -\tfrac{k}{2}\ln 2\pi - \tfrac{1}{2}\ln\left|\Sigma\right| - \tfrac{k}{2} \\
&= -\tfrac{1}{2}\ln\left(2\pi e\right)^k\left|\Sigma\right|
\end{aligned} \tag{A.31}$$

Here, k is the dimensionality of x. The move from the first to the second line depends on the trace identity in equation A.6. The first two terms in the third line come from the definition of a multivariate normal distribution (equation A.18). Equation A.31 expresses the first term of equation A.27 under the Laplace assumption. The second term of equation A.27 can similarly be expanded around μ:

$$\begin{aligned}
\ln p(y,x) &\approx \ln p(y,\mu) + (\mu - x)\cdot \nabla_x \ln p(y,x)\big|_{x=\mu} \\
&\quad + \frac{1}{2}(\mu-x)\cdot\nabla_x\left(\nabla_x \ln p(y,x)\right)^T\Big|_{x=\mu}(\mu-x) \\
\mathbb{E}_{q(x)}[\ln p(y,x)] &\approx \ln p(y,\mu) + \underbrace{(\mu - \mathbb{E}_{q(x)}[x])}_{0}\nabla_x \ln p(y,x)\big|_{x=\mu} \\
&\quad + \frac{1}{2} tr\left[\mathbb{E}_{q(x)}[(\mu-x)(\mu-x)^T]\nabla_x\left(\nabla_x \ln p(y,x)\right)^T\Big|_{x=\mu}\right] \\
&= \ln p(y,\mu) + \frac{1}{2} tr\left[\Sigma\nabla_x\left(\nabla_x \ln p(y,x)\right)^T\Big|_{x=\mu}\right]
\end{aligned} \tag{A.32}$$

The final equality uses the fact that, if q is a normal distribution, its mean is also its mode. Substituting equations A.31 and A.32 back into equation A.27, we get the Laplace free energy:

$$F[q,y] \approx -\frac{1}{2}\ln(2\pi e)^k|\Sigma| - \ln p(y,\mu) - \frac{1}{2}tr\left[\Sigma\nabla_x\left(\nabla_x \ln p(y,x)\right)^T\Big|_{x=\mu}\right] \quad \text{(A.33)}$$

The trace operator in the last term can be ignored when x is 1-dimensional. The useful thing about this formulation is that if we set the derivative of the free energy with respect to the posterior precision to zero, we find the following:[2]

$$\partial_\Sigma F[q,y] = 0 \Leftrightarrow \Sigma^{-1} = -\nabla_x\left(\nabla_x \ln p(y,x)\right)^T\Big|_{x=\mu} \quad \text{(A.34)}$$

This means that the precision of the posterior is the negative curvature of the log probability of states and data evaluated at the posterior mode. As such, minimizing free energy does not require explicit optimization of the precision—this may be computed analytically from the posterior mean. Furthermore, substitution of A.34 into A.33 reveals that the only term in the free energy that depends on the posterior mean is the log probability over data and states. For details of how this is done to perform inference in continuous state-space models, see chapter 4.

A.3.3 Generalized Coordinates of Motion

In addition to being central to the Laplace approximation, the Taylor series approximation plays another important role in Active Inference. This is in the use of generalized coordinates of motion to represent beliefs about a trajectory through time. In brief, this means drawing inferences not only about the position of a variable (x) but also its velocity (x'), acceleration (x''), and subsequent temporal derivatives. These implicitly represent an approximation to the trajectory that can be made explicit through the following Taylor series:

$$\begin{aligned} x(t) &\approx x(\tau) + \varepsilon x'(t)\big|_{t=\tau} + \frac{1}{2}\varepsilon^2 x''(t)\big|_{t=\tau} + \cdots \\ \varepsilon &= t - \tau \end{aligned} \quad \text{(A.35)}$$

This additionally means we can account for structure in the covariance of random fluctuations, as is necessary in dealing with these fluctuations in biological systems (where fluctuations are themselves generated by dynamical processes). We will discuss this further in section A5. For now, we simply

note that a probability density over the generalized coordinates of motion is equivalent to a distribution over local trajectories constructed by treating the coordinates as coefficients of a Taylor series expansion.

A.4 Variational Calculus

A.4.1 Functional Derivatives

Because Active Inference deals with optimizing beliefs (probability distributions), it is often necessary to talk about the minimization of functionals (functions of functions) with respect to functions. This calls for the concept of a functional (i.e., variational) derivative. The basic problem is finding the function (f) that minimizes a functional (S), normally expressed as an integral[3] of a function that includes f:

$$\begin{aligned} \phi(x) &= \arg\min_{f} S[f(x)] \\ S[f(x)] &\triangleq \int_{x_1}^{x_2} \mathcal{L}(f(x), x)\, dx \end{aligned} \tag{A.36}$$

If we parameterize the function in terms of an arbitrary function (g) that is zero at the extremes of the integral and multiply this by a small number (u), we can take the derivative of S with respect to u:

$$\begin{aligned} f(x,u) &\triangleq \phi(x) + ug(x) \\ \partial_u S[f(x,u)] &= \int_{x_1}^{x_2} \partial_u \mathcal{L}(f(x,u), x)\, dx \\ &= \int_{x_1}^{x_2} \partial_u f(x,u)\, \partial_f \mathcal{L}(f(x,u), x)\, dx \\ &= \int_{x_1}^{x_2} g(x)\, \partial_f \mathcal{L}(f(x,u), x)\, dx \end{aligned} \tag{A.37}$$

When u is zero, f is the function that minimizes the integral. This means equation A.37 should be zero when evaluated at $u = 0$. The condition that must be satisfied for f to minimize S is then as follows:

$$\int_{x_1}^{x_2} g(x)\, \partial_f \mathcal{L}(f(x), x)\, dx \Big|_{f=\phi} = 0 \tag{A.38}$$

For equation A.38 to be true for any arbitrary $g(x)$, the following is implied:[4]

$$\delta_f S \triangleq \partial_f \mathcal{L} = 0 \tag{A.39}$$

Note that, in a physics setting, $\mathcal{L}$ may include the gradient of f in addition to the function itself. The same steps outlined above then give rise to the Euler-Lagrange equation:

$$\begin{aligned} \delta_f S &\triangleq \partial_f \mathcal{L} - \frac{d}{dx}\partial_{f'}\mathcal{L} = 0 \\ f' &\triangleq \partial_x f \end{aligned} \tag{A.40}$$

Depending on whether $\mathcal{L}$ includes the gradient, equations A.39 and A.40 express the notion of a variational (aka functional) derivative.

A.4.2 Variational Bayes

Variational Bayes follows in a relatively straightforward way from the above if we set f to be a factor of an approximate posterior distribution and S to be a free energy functional:

$$\begin{aligned} f(x) &= q_i(x_i) \\ q(x) &= \prod_i q_i(x_i) \\ \mathcal{L}(q_i(x_i), x_i) &= \int q(x)(\ln q(x) - \ln p(y,x))\, dx_{j\neq i} \\ S[q(x)] &= F[q(x), y] \end{aligned} \tag{A.41}$$

The second line here expresses a *mean-field* approximation, in which the approximate posterior is factorized over the variables x. This is often used for reasons of computational tractability. However, this is one of many choices of form for the approximate posterior. Applying equation A.39, we find the form of the approximate posterior that minimizes the free energy (omitting constants):

$$\begin{aligned} \delta_{q_i} F[q,y] &= \ln q_i(x_i) - \int \prod_{j\neq i} q_j(x_j) \ln p(y,x)\, dx_{j\neq i} \\ \delta_{q_i} F[q,y] &= 0 \Leftrightarrow \\ \ln q_i(x_i) &= \mathbb{E}_{q\setminus i}[\ln p(y,x)] \end{aligned} \tag{A.42}$$

The notation $\setminus i$ should be read as "all factors except for the ith factor." Equation A.42 is central to an inference scheme known as variational message passing (Winn and Bishop 2005, Dauwels 2007). This works by optimizing each factor of q independently and relies on p being relatively sparse

(i.e., not every x_i depends on every other x_j). To gain some intuition for this, consider what happens with an (arbitrary) example:

$$\begin{aligned}
&p(y,x) = p(y|x_1)p(x_1|x_2,x_3)p(x_3)p(x_2|x_4)p(x_4)\\
&\qquad\Rightarrow\\
&\ln q(x_1) =\\
&\quad = \mathbb{E}_{q(x_2)q(x_3)q(x_4)}\Big[\ln p(y|x_1) + \ln p(x_1|x_2,x_3) + \underbrace{\ln p(x_3)p(x_2|x_4)p(x_4)}_{\text{constant w.r.t. } x_1}\Big]\\
&\ln q(x_2) =\\
&\quad = \mathbb{E}_{q(x_1)q(x_3)q(x_4)}\Big[\ln p(x_1|x_2,x_3) + \ln p(x_2|x_4) + \underbrace{\ln p(y|x_1)p(x_3)p(x_4)}_{\text{constant w.r.t. } x_2}\Big]\\
&\qquad\vdots
\end{aligned} \tag{A.43}$$

Equation A43 shows what happens when we substitute the density in the first line into equation A.42 for the first two factors of q. Omitting constant terms, we have this:

$$\begin{aligned}
\ln q(x_1) &= \ln p(y|x_1) + \mathbb{E}_{q(x_2)q(x_3)}\big[\ln p(x_1|x_2,x_3)\big]\\
\ln q(x_2) &= \mathbb{E}_{q(x_1)q(x_3)}\big[\ln p(x_1|x_2,x_3)\big] + \mathbb{E}_{q(x_4)}\big[\ln p(x_2|x_4)\big]\\
&\vdots
\end{aligned} \tag{A.44}$$

The terms in the expectation have been simplified by noting the following:

$$\mathbb{E}_{p(b)}\big[f(a)\big] = \int p(b)f(a)\,db = f(a)\underbrace{\int p(b)\,db}_{=1} = f(a) \tag{A.45}$$

This accounts for the simplicity of variational message passing, in which we only need take account of a small subset of beliefs (those about the Markov blanket—see box 4.1) in order to update each belief.

A.5 Stochastic Dynamics

A.5.1 Stochastic Differential Equations

There are a few places in this book where we refer to ideas from the theory of random dynamical systems. In chapter 3, for instance, we highlight the importance of a steady-state distribution to which a random system tends over time and the relationship between these dynamics and the notion of *self-evidencing*. In chapters 4 and 8, we outline how a continuous state-space model may be formulated in terms of stochastic differential equations.

Although this is a fascinating topic (Yuan and Ao 2012), a full dissection of the subtleties of defining stochastic processes is outside the scope of this book. However, it is worth briefly unpacking what we mean by a stochastic differential equation. Put simply, it is a differential equation that is augmented by a random term (ω):

$$\begin{aligned} \dot{x} &= f(x) + \omega \\ \omega &\sim \mathcal{N}(0, \tfrac{1}{2}\Gamma^{-1}) \end{aligned} \tag{A.46}$$

The random term here is chosen to be normally distributed. It has a mean of zero, such that the most likely value for the rate of change of x is simply $f(x)$. The interpretation of equation A.46 is sometimes a little tricky. The best way to dispel any ambiguity is to see it as the limiting case of a discretized scheme:

$$\begin{aligned} \Delta x &= f(x)\Delta\tau + \omega(\Delta\tau)^{\frac{1}{2}} \\ \Delta\tau \to 0 &\Rightarrow \dot{x} = f(x) + \omega \end{aligned} \tag{A.47}$$

Note that if the variance of ω varies with x there are multiple discretizations we could appeal to. The most common choices correspond to Ito and Stratonovich interpretations of a stochastic equation. However, we assume a fixed variance throughout this book—which ensures these interpretations lead to identical results. For the purpose of defining a generative model of the sort found in chapter 8, we just need the probability distribution describing the rate of change of x. From equation A.46, this is simply as follows:

$$p(\dot{x} \mid x) = \mathcal{N}(f(x), \tfrac{1}{2}\Gamma^{-1}) \tag{A.48}$$

This is the form that will be found in the generative models used here. This provides a summary of the distinction between a deterministic and a random dynamical system. If we know the value of x in a deterministic system, then we know its velocity. In a stochastic system, knowing x tells us the distribution of possible velocities we might expect.

A.5.2 Nonequilibrium Steady State

In chapter 3, we see that a system defined such that it descends some energy (or surprise) function maintains its form over time and persists at a (possibly nonequilibrium) steady state. We will briefly unpack what this means here, starting from the idea of a steady state and recovering the surprise-minimizing or "self-evidencing" (Hohwy 2016) dynamics. The starting point

is an alternative expression of the stochastic dynamics in equation A.46 in terms of a deterministic partial differential equation describing how the probability density changes over time. This is known as a Fokker-Planck equation (Risken 1996):

$$\partial_\tau p(x) = \nabla_x \cdot \left(\Gamma \nabla_x p(x) - f(x)p(x)\right) \tag{A.49}$$

The Fokker-Planck equation lets us define a steady state simply by setting the partial derivative of the density with respect to time to be zero:

$$\begin{aligned} \partial_\tau p(x) &= 0 \\ &\Rightarrow \\ \nabla_x \cdot \left(\Gamma \nabla_x p(x) - f(x)p(x)\right) &= 0 \\ &\Rightarrow \\ f(x) &= -(\Gamma - Q(x))\nabla_x \Im(x) \\ \nabla_x \cdot \left(Q(x)\nabla_x p(x)\right) &= 0 \\ \Im(x) &\triangleq -\ln p(x) \end{aligned} \tag{A.50}$$

The third equality here[5] is key, as it says that those systems that maintain steady state must exhibit dynamics that (on average) minimize their surprise ($\Im$). The Q term allows for dynamics along the contours of the surprise, which neither increase nor decrease surprise. This expression underwrites the self-evidencing perspective of Active Inference and is central to the physics of sentient systems. We will not dwell on this here but refer readers to Friston (2019a) for a more comprehensive overview of the consequences of this treatment.

A.5.3 Generalized Coordinates of Motion

As we saw in section A.3.3, we can represent a short trajectory in terms of the coefficients of a Taylor series expansion in time. This raises an interesting question when we translate this into the context of a stochastic setting. When specifying a continuous-time model in terms of generalized coordinates of motion, how do we account for the covariance between the orders of generalized motion? The answer is given in Cox and Miller (1965), which we summarize here. A random process is expressed in generalized coordinates as a vector of the random fluctuations accompanying the flow, the rate of change of that flow, and subsequent temporal derivatives:

$$\begin{aligned}\dot{\tilde{x}} &= \tilde{f}(\tilde{x}) + \tilde{\omega} \\ \tilde{\omega} \triangleq \begin{bmatrix} \omega \\ \omega' \\ \omega'' \\ \omega''' \\ \vdots \end{bmatrix} &= \begin{bmatrix} \omega^{[0]} \\ \omega^{[1]} \\ \omega^{[2]} \\ \omega^{[3]} \\ \vdots \end{bmatrix}\end{aligned} \tag{A.51}$$

The random fluctuations may be characterized as follows:

$$\begin{aligned} p(\tilde{\omega}) &= \mathcal{N}(0, \tilde{\Pi}) \\ \mathbb{E}[\omega^{[0]}(\tau)] &= 0 \\ \mathbb{E}[\omega^{[0]}(\tau) \cdot \omega^{[0]}(\tau)] &= \Sigma \end{aligned} \tag{A.52}$$

Their autocorrelation function is this:

$$\rho(h) \triangleq \Sigma^{-1} \underbrace{\mathbb{E}[\omega^{[0]}(\tau) \cdot \omega^{[0]}(\tau + h)]}_{\text{Covariance}} \tag{A.53}$$

We can multiply both sides of this equation by the variance to show that the covariance between the noise at two time-points may be factorized into an autocorrelation and a variance. We define the *i*th derivative of the random fluctuations as this limiting case:

$$\omega^{[i]}(\tau, \Delta\tau) = \frac{\omega^{[i-1]}(\tau + \Delta\tau) - \omega^{[i-1]}(\tau)}{\Delta\tau} \tag{A.54}$$

Using equations A.52 and A.53, we can express the covariance between a variable and its first temporal derivative:

$$\begin{aligned} \mathbb{E}[\omega^{[1]}(\tau, \Delta\tau) \cdot \omega^{[0]}(\tau + h)] &= \tfrac{1}{\Delta\tau} \mathbb{E}\Big[\big(\omega^{[0]}(\tau + \Delta\tau) - \omega^{[0]}(\tau)\big)\omega^{[0]}(\tau + h)\Big] \\ &= \tfrac{1}{\Delta\tau} \Sigma\big(\rho(h - \Delta\tau) - \rho(h)\big) \end{aligned} \tag{A.55}$$

Taking the limit as the change in time tends to zero:

$$\mathbb{E}[\omega^{[1]}(\tau) \cdot \omega^{[0]}(\tau + h)] = \Sigma \dot{\rho}(h) \tag{A.56}$$

Evaluating at $h = 0$ gives us a covariance of zero, as the instantaneous velocity and position are orthogonal to one another (and the autocorrelation is at a maximum, so its temporal derivative is zero).

We can take this procedure one step further and evaluate the variance of the first derivative:

$$\begin{aligned}
&\mathbb{E}[\omega^{[1]}(\tau,\Delta\tau)\cdot\omega^{[1]}(\tau+h,\Delta\tau)] \\
&= \tfrac{1}{\Delta\tau^2}\sum\mathbb{E}\Big[\big(\omega^{[0]}(\tau+\Delta\tau)-\omega^{[0]}(\tau)\big)\big(\omega^{[0]}(\tau+h+\Delta\tau)-\omega^{[0]}(\tau+h)\big)\Big] \\
&= \sum\tfrac{1}{\Delta\tau}\big(\tfrac{1}{\Delta\tau}\big(\rho(h)-\rho(h-\Delta\tau)\big)-\tfrac{1}{\Delta\tau}\big(\rho(h+\Delta\tau)-\rho(h)\big)\big)
\end{aligned} \tag{A.57}$$

Taking the limit as $\Delta\tau\to 0$, this is as follows:

$$\mathbb{E}[\omega^{[1]}(\tau)\cdot\omega^{[1]}(\tau+h)] = -\sum\ddot{\rho}(h) \tag{A.58}$$

Pursuing this procedure for subsequent derivatives allows us to compute the elements of the generalized precision matrix:

$$\tilde{\Pi} = \Sigma^{-1}\otimes\begin{bmatrix} 1 & 0 & \ddot{\rho}(0) & \\ 0 & -\ddot{\rho}(0) & 0 & \\ \ddot{\rho}(0) & 0 & \dddot{\rho}(0) & \\ & & & \ddots \end{bmatrix}^{-1} \tag{A.59}$$

Choosing the autocorrelation function to be Gaussian, we have the following:

$$\begin{aligned}
&\rho(h) = e^{-\frac{1}{2}\lambda h^2} && \rho(0) = 1 \\
&\dot{\rho}(h) = -\lambda\rho(h) && \ddot{\rho}(0) = 0 \\
&\ddot{\rho}(h) = \lambda(\lambda h^2-1)\rho(h) && \ddot{\rho}(0) = -\lambda \\
&\dddot{\rho}(h) = \lambda^2 h(\lambda h^2-3)\rho(h) && \dddot{\rho}(0) = 0 \\
&\ddddot{\rho}(h) = \lambda^2(\lambda^2 h^4-6\lambda h^2+3)\rho(h) && \ddddot{\rho}(0) = 3\lambda^2
\end{aligned} \tag{A.60}$$

The precision term (λ) can then be thought of as parameterizing the smoothness of the random fluctuations. This may itself be optimized in relation to data through minimization of free energy.