33

Variational Methods

Variational methods are an important technique for the approximation of complicated probability distributions, having applications in statistical physics, data modelling and neural networks.

▶ 33.1 Variational free energy minimization

One method for approximating a complex distribution in a physical system is *mean field theory*. Mean field theory is a special case of a general *variational free energy* approach of Feynman and Bogoliubov which we will now study. The key piece of mathematics needed to understand this method is Gibbs' inequality, which we repeat here.

Gibbs' inequality first appeared in equation (1.24); see also exercise 2.26 (p.37).

The relative entropy between two probability distributions $Q(x)$ and $P(x)$ that are defined over the same alphabet $\mathcal{A}_X$ is

$$D_{\mathrm{KL}}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}. \tag{33.1}$$

The relative entropy satisfies $D_{\mathrm{KL}}(Q||P) \geq 0$ (Gibbs' inequality) with equality only if $Q = P$. In general $D_{\mathrm{KL}}(Q||P) \neq D_{\mathrm{KL}}(P||Q)$.

In this chapter we will replace the log by ln, and measure the divergence in nats.

Probability distributions in statistical physics

In statistical physics one often encounters probability distributions of the form

$$P(\mathbf{x} \,|\, \beta, \mathbf{J}) = \frac{1}{Z(\beta, \mathbf{J})} \exp[-\beta E(\mathbf{x}; \mathbf{J})] \,, \tag{33.2}$$

where for example the state vector is $\mathbf{x} \in \{-1, +1\}^N$, and $E(\mathbf{x}; \mathbf{J})$ is some energy function such as

$$E(\mathbf{x}; \mathbf{J}) = -\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n. \tag{33.3}$$

The partition function (normalizing constant) is

$$Z(\beta, \mathbf{J}) \equiv \sum_{\mathbf{x}} \exp[-\beta E(\mathbf{x}; \mathbf{J})] \,. \tag{33.4}$$

The probability distribution of equation (33.2) is complex. Not unbearably complex – we can, after all, evaluate $E(\mathbf{x}; \mathbf{J})$ for any particular $\mathbf{x}$ in a time

polynomial in the number of spins. But evaluating the normalizing constant $Z(\beta, \mathbf{J})$ is difficult, as we saw in Chapter 29, and describing the properties of the probability distribution is also hard. Knowing the value of $E(\mathbf{x}; \mathbf{J})$ at a few arbitrary points $\mathbf{x}$, for example, gives no useful information about what the average properties of the system are.

An evaluation of $Z(\beta, \mathbf{J})$ would be particularly desirable because from Z we can derive all the thermodynamic properties of the system.

Variational free energy minimization is a method for *approximating* the complex distribution $P(\mathbf{x})$ by a simpler ensemble $Q(\mathbf{x}; \boldsymbol{\theta})$ that is parameterized by adjustable parameters $\boldsymbol{\theta}$. We adjust these parameters so as to get Q to best approximate P, in some sense. A by-product of this approximation is a lower bound on $Z(\beta, \mathbf{J})$.

The variational free energy

The objective function chosen to measure the quality of the approximation is the *variational free energy*

$$\beta\tilde{F}(\boldsymbol{\theta}) = \sum_{\mathbf{x}} Q(\mathbf{x}; \boldsymbol{\theta}) \ln \frac{Q(\mathbf{x}; \boldsymbol{\theta})}{\exp[-\beta E(\mathbf{x}; \mathbf{J})]}. \tag{33.5}$$

This expression can be manipulated into a couple of interesting forms: first,

$$\beta\tilde{F}(\boldsymbol{\theta}) = \beta \sum_{\mathbf{x}} Q(\mathbf{x}; \boldsymbol{\theta}) E(\mathbf{x}; \mathbf{J}) - \sum_{\mathbf{x}} Q(\mathbf{x}; \boldsymbol{\theta}) \ln \frac{1}{Q(\mathbf{x}; \boldsymbol{\theta})} \tag{33.6}$$

$$\equiv \beta \left\langle E(\mathbf{x}; \mathbf{J}) \right\rangle_Q - S_Q, \tag{33.7}$$

where $\left\langle E(\mathbf{x}; \mathbf{J}) \right\rangle_Q$ is the average of the energy function under the distribution $Q(\mathbf{x}; \boldsymbol{\theta})$, and S_Q is the entropy of the distribution $Q(\mathbf{x}; \boldsymbol{\theta})$ (we set k_{B} to one in the definition of S so that it is identical to the definition of the entropy H in Part I).

Second, we can use the definition of $P(\mathbf{x} \,|\, \beta, \mathbf{J})$ to write:

$$\beta\tilde{F}(\boldsymbol{\theta}) = \sum_{\mathbf{x}} Q(\mathbf{x}; \boldsymbol{\theta}) \ln \frac{Q(\mathbf{x}; \boldsymbol{\theta})}{P(\mathbf{x} \,|\, \beta, \mathbf{J})} - \ln Z(\beta, \mathbf{J}) \tag{33.8}$$

$$= D_{\mathrm{KL}}(Q||P) + \beta F, \tag{33.9}$$

where F is the true free energy, defined by

$$\beta F \equiv -\ln Z(\beta, \mathbf{J}), \tag{33.10}$$

and $D_{\mathrm{KL}}(Q||P)$ is the relative entropy between the approximating distribution $Q(\mathbf{x}; \boldsymbol{\theta})$ and the true distribution $P(\mathbf{x} \,|\, \beta, \mathbf{J})$. Thus by Gibbs' inequality, the variational free energy $\tilde{F}(\boldsymbol{\theta})$ is bounded below by F and attains this value only for $Q(\mathbf{x}; \boldsymbol{\theta}) = P(\mathbf{x} \,|\, \beta, \mathbf{J})$.

Our strategy is thus to vary $\boldsymbol{\theta}$ in such a way that $\beta\tilde{F}(\boldsymbol{\theta})$ is minimized. The approximating distribution then gives a simplified approximation to the true distribution that may be useful, and the value of $\beta\tilde{F}(\boldsymbol{\theta})$ will be an upper bound for βF. Equivalently, $\tilde{Z} \equiv e^{-\beta\tilde{F}(\boldsymbol{\theta})}$ is a lower bound for Z.

Can the objective function $\beta\tilde{F}$ be evaluated?

We have already agreed that the evaluation of various interesting sums over $\mathbf{x}$ is intractable. For example, the partition function

$$Z = \sum_{\mathbf{x}} \exp(-\beta E(\mathbf{x}; \mathbf{J})), \tag{33.11}$$

the energy

$$\langle E \rangle_P = \frac{1}{Z} \sum_{\mathbf{x}} E(\mathbf{x}; \mathbf{J}) \exp(-\beta E(\mathbf{x}; \mathbf{J})) \,, \tag{33.12}$$

and the entropy

$$S \equiv \sum_{\mathbf{x}} P(\mathbf{x} \,|\, \beta, \mathbf{J}) \ln \frac{1}{P(\mathbf{x} \,|\, \beta, \mathbf{J})} \tag{33.13}$$

are all presumed to be impossible to evaluate. So why should we suppose that this objective function $\beta \tilde{F}(\boldsymbol{\theta})$, which is also defined in terms of a sum over all $\mathbf{x}$ (33.5), should be a convenient quantity to deal with? Well, for a range of interesting energy functions, and for sufficiently simple approximating distributions, the variational free energy *can* be efficiently evaluated.

▶ 33.2 Variational free energy minimization for spin systems

An example of a tractable variational free energy is given by the spin system whose energy function was given in equation (33.3), which we can approximate with a *separable* approximating distribution,

$$Q(\mathbf{x}; \mathbf{a}) = \frac{1}{Z_Q} \exp\left(\sum_n a_n x_n \right). \tag{33.14}$$

The variational parameters $\boldsymbol{\theta}$ of the variational free energy (33.5) are the components of the vector $\mathbf{a}$. To evaluate the variational free energy we need the entropy of this distribution,

$$S_Q = \sum_{\mathbf{x}} Q(\mathbf{x}; \mathbf{a}) \ln \frac{1}{Q(\mathbf{x}; \mathbf{a})}, \tag{33.15}$$

and the mean of the energy,

$$\langle E(\mathbf{x}; \mathbf{J}) \rangle_Q = \sum_{\mathbf{x}} Q(\mathbf{x}; \mathbf{a}) E(\mathbf{x}; \mathbf{J}). \tag{33.16}$$

The entropy of the separable approximating distribution is simply the sum of the entropies of the individual spins (exercise 4.2, p.68),

$$S_Q = \sum_n H_2^{(e)}(q_n), \tag{33.17}$$

where q_n is the probability that spin n is $+1$,

$$q_n = \frac{e^{a_n}}{e^{a_n} + e^{-a_n}} = \frac{1}{1 + \exp(-2a_n)}, \tag{33.18}$$

and

$$H_2^{(e)}(q) = q \ln \frac{1}{q} + (1 - q) \ln \frac{1}{(1 - q)}. \tag{33.19}$$

The mean energy under Q is easy to obtain because $\sum_{m,n} J_{mn} x_m x_n$ is a sum of terms each involving the product of two *independent* random variables. (There are no self-couplings, so $J_{mn} = 0$ when $m = n$.) If we define the mean value of x_n to be $\bar{x}_n$, which is given by

$$\bar{x}_n = \frac{e^{a_n} - e^{-a_n}}{e^{a_n} + e^{-a_n}} = \tanh(a_n) = 2q_n - 1, \tag{33.20}$$

we obtain

$$\langle E(\mathbf{x};\mathbf{J})\rangle_Q = \sum_{\mathbf{x}} Q(\mathbf{x};\mathbf{a}) \left[-\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n \right] \quad (33.21)$$

$$= -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n. \quad (33.22)$$

So the variational free energy is given by

$$\beta \tilde{F}(\mathbf{a}) = \beta \langle E(\mathbf{x};\mathbf{J})\rangle_Q - S_Q = \beta \left(-\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n \right) - \sum_n H_2^{(e)}(q_n). \quad (33.23)$$

We now consider minimizing this function with respect to the variational parameters $\mathbf{a}$. If $q = 1/(1+e^{-2a})$, the derivative of the entropy is

$$\frac{\partial}{\partial q} H_2^e(q) = \ln \frac{1-q}{q} = -2a. \quad (33.24)$$

So we obtain

$$\frac{\partial}{\partial a_m} \beta \tilde{F}(\mathbf{a}) = \beta \left[-\sum_n J_{mn} \bar{x}_n - h_m \right] \left(2 \frac{\partial q_m}{\partial a_m} \right) - \ln \left(\frac{1-q_m}{q_m} \right) \left(\frac{\partial q_m}{\partial a_m} \right)$$

$$= 2 \left(\frac{\partial q_m}{\partial a_m} \right) \left[-\beta \left(\sum_n J_{mn} \bar{x}_n + h_m \right) + a_m \right]. \quad (33.25)$$

This derivative is equal to zero when

$$a_m = \beta \left(\sum_n J_{mn} \bar{x}_n + h_m \right). \quad (33.26)$$

So $\tilde{F}(\mathbf{a})$ is extremized at any point that satisfies equation (33.26) and

$$\bar{x}_n = \tanh(a_n). \quad (33.27)$$

The variational free energy $\tilde{F}(\mathbf{a})$ may be a multimodal function, in which case each stationary point (maximum, minimum or saddle) will satisfy equations (33.26) and (33.27). One way of using these equations, in the case of a system with an arbitrary coupling matrix $\mathbf{J}$, is to update each parameter a_m and the corresponding value of $\bar{x}_m$ using equation (33.26), one at a time. This *asynchronous updating of the parameters* is guaranteed to decrease $\beta \tilde{F}(\mathbf{a})$.

Equations (33.26) and (33.27) may be recognized as the *mean field* equations for a spin system. The variational parameter a_n may be thought of as the strength of a fictitious field applied to an isolated spin n. Equation (33.27) describes the mean response of spin n, and equation (33.26) describes how the field a_m is set in response to the mean state of all the other spins.

The variational free energy derivation is a helpful viewpoint for mean field theory for two reasons.

1. This approach associates an objective function $\beta \tilde{F}$ with the mean field equations; such an objective function is useful because it can help identify alternative dynamical systems that minimize the same function.

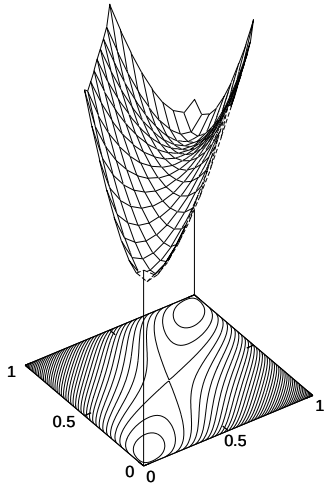


Figure 33.1. The variational free energy of the two-spin system whose energy is $E(\mathbf{x}) = -x_1 x_2$, as a function of the two variational parameters q_1 and q_2. The inverse-temperature is $\beta = 1.44$. The function plotted is

$$\beta \tilde{F} = -\beta \bar{x}_1 \bar{x}_2 - H_2^{(e)}(q_1) - H_2^{(e)}(q_2),$$

where $\bar{x}_n = 2q_n - 1$. Notice that for fixed q_2 the function is convex $\smile$ with respect to q_1, and for fixed q_1 it is convex $\smile$ with respect to q_2.

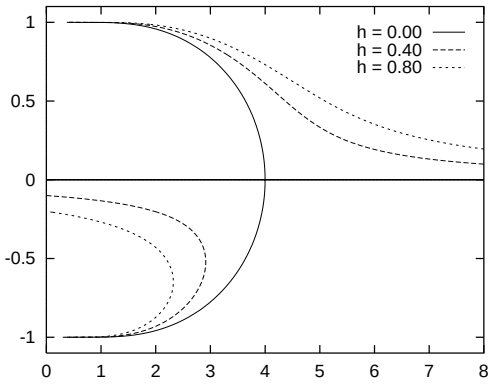


Figure 33.2. Solutions of the variational free energy extremization problem for the Ising model, for three different applied fields h. Horizontal axis: temperature $T = 1/\beta$. Vertical axis: magnetization $\bar{x}$. The critical temperature found by mean field theory is $T_c^{\rm mft} = 4$.

2. The theory is readily generalized to other approximating distributions. We can imagine introducing a more complex approximation $Q(\mathbf{x}; \boldsymbol{\theta})$ that might for example capture correlations among the spins instead of modelling the spins as independent. One could then evaluate the variational free energy and optimize the parameters $\boldsymbol{\theta}$ of this more complex approximation. The more degrees of freedom the approximating distribution has, the tighter the bound on the free energy becomes. However, if the complexity of an approximation is increased, the evaluation of either the mean energy or the entropy typically becomes more challenging.

▶ 33.3 Example: mean field theory for the ferromagnetic Ising model

In the simple Ising model studied in Chapter 31, every coupling J_{mn} is equal to J if m and n are neighbours and zero otherwise. There is an applied field $h_n = h$ that is the same for all spins. A very simple approximating distribution is one with just a single variational parameter a, which defines a separable distribution

$$Q(\mathbf{x}; a) = \frac{1}{Z_Q} \exp\left(\sum_n a x_n \right) \tag{33.28}$$

in which all spins are independent and have the same probability

$$q_n = \frac{1}{1 + \exp(-2a)} \tag{33.29}$$

of being up. The mean magnetization is

$$\bar{x} = \tanh(a) \tag{33.30}$$

and the equation (33.26) which defines the minimum of the variational free energy becomes

$$a = \beta \left(CJ\bar{x} + h \right), \tag{33.31}$$

where C is the number of couplings that a spin is involved in – $C = 4$ in the case of a rectangular two-dimensional Ising model. We can solve equations (33.30) and (33.31) for $\bar{x}$ numerically – in fact, it is easiest to vary $\bar{x}$ and solve for β – and obtain graphs of the free energy minima and maxima as a function of temperature as shown in figure 33.2. The solid line shows $\bar{x}$ versus $T = 1/\beta$ for the case $C = 4, J = 1$.

When $h = 0$, there is a pitchfork bifurcation at a critical temperature $T_c^{\rm mft}$. [A pitchfork bifurcation is a transition like the one shown by the solid lines in

figure 33.2, from a system with one minimum as a function of a (on the right) to a system (on the left) with two minima and one maximum; the maximum is the middle one of the three lines. The solid lines look like a pitchfork.] Above this temperature, there is only one minimum in the variational free energy, at $a = 0$ and $\bar{x} = 0$; this minimum corresponds to an approximating distribution that is uniform over all states. Below the critical temperature, there are two minima corresponding to approximating distributions that are symmetry-broken, with all spins more likely to be up, or all spins more likely to be down. The state $\bar{x} = 0$ persists as a stationary point of the variational free energy, but now it is a local *maximum* of the variational free energy.

When $h > 0$, there is a global variational free energy minimum at any temperature for a positive value of $\bar{x}$, shown by the upper dotted curves in figure 33.2. As long as $h < JC$, there is also a second local minimum in the free energy, if the temperature is sufficiently small. This second minimum corresponds to a self-preserving state of magnetization in the opposite direction to the applied field. The temperature at which the second minimum appears is smaller than $T_c^{\rm mft}$, and when it appears, it is accompanied by a saddle point located between the two minima. A name given to this type of bifurcation is a saddle-node bifurcation.

The variational free energy per spin is given by

$$\beta\tilde{F} = \beta\left(-\frac{C}{2}J\bar{x}^2 - h\bar{x}\right) - H_2^{(e)}\left(\frac{\bar{x}+1}{2}\right). \tag{33.32}$$

Exercise 33.1.[2] Sketch the variational free energy as a function of its one parameter $\bar{x}$ for a variety of values of the temperature T and the applied field h.

Figure 33.2 reproduces the key properties of the real Ising system – that, for $h = 0$, there is a critical temperature below which the system has long-range order, and that it can adopt one of two macroscopic states. However, by probing a little more we can reveal some inadequacies of the variational approximation. To start with, the critical temperature $T_c^{\rm mft}$ is 4, which is nearly a factor of 2 greater than the true critical temperature $T_c = 2.27$. Also, the variational model has equivalent properties in any number of dimensions, including $d = 1$, where the true system does not have a phase transition. So the bifurcation at $T_c^{\rm mft}$ should not be described as a phase transition.

For the case $h = 0$ we can follow the trajectory of the global minimum as a function of β and find the entropy, heat capacity and fluctuations of the approximating distribution and compare them with those of a real 8×8 fragment using the matrix method of Chapter 31. As shown in figure 33.3, one of the biggest differences is in the fluctuations in energy. The real system has large fluctuations near the critical temperature, whereas the approximating distribution has no correlations among its spins and thus has an energy-variance which scales simply linearly with the number of spins.

▶ 33.4 Variational methods in inference and data modelling

In statistical data modelling we are interested in the posterior probability distribution of a parameter vector $\mathbf{w}$ given data D and model assumptions $\mathcal{H}$, $P(\mathbf{w} \mid D, \mathcal{H})$.

$$P(\mathbf{w} \mid D, \mathcal{H}) = \frac{P(D \mid \mathbf{w}, \mathcal{H})P(\mathbf{w} \mid \mathcal{H})}{P(D \mid \mathcal{H})}. \tag{33.33}$$

In traditional approaches to model fitting, a single parameter vector $\mathbf{w}$ is optimized to find the mode of this distribution. What is really of interest is

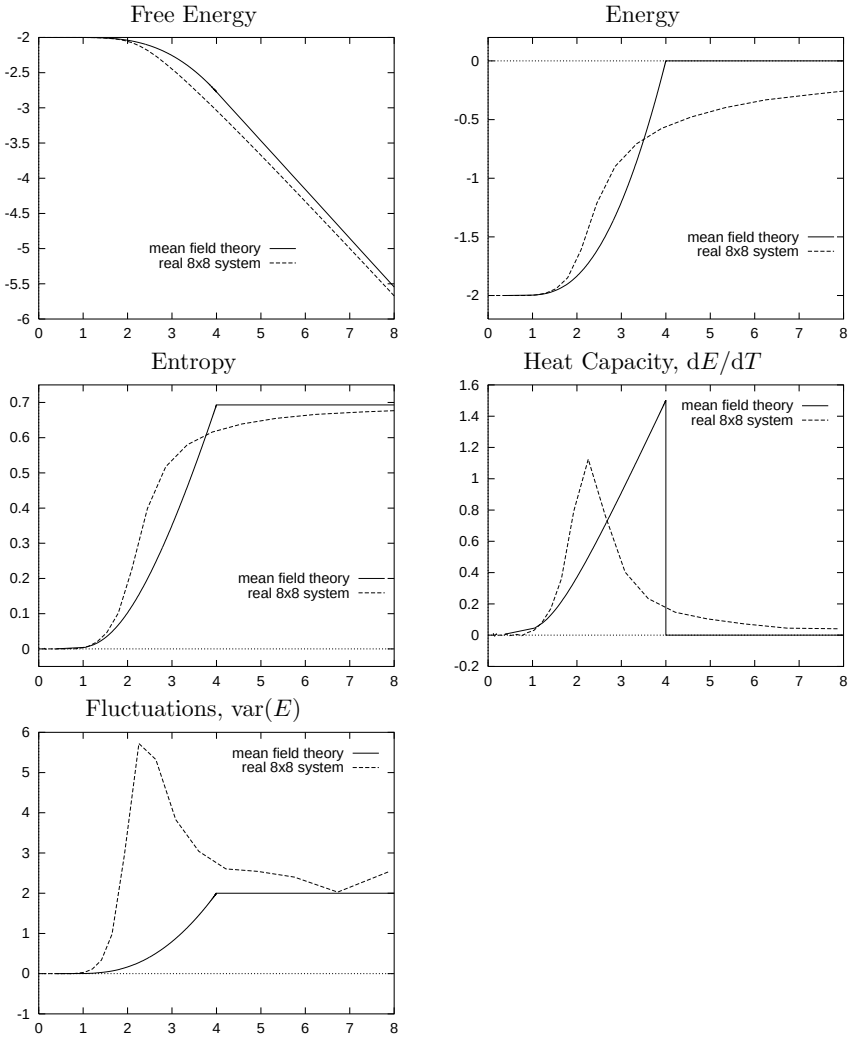


Figure 33.3. Comparison of approximating distribution's properties with those of a real 8×8 fragment. Notice that the variational free energy of the approximating distribution is indeed an upper bound on the free energy of the real system. All quantities are shown 'per spin'.

the whole distribution. We may also be interested in its normalizing constant $P(D\,|\,\mathcal{H})$ if we wish to do model comparison. The probability distribution $P(\mathbf{w}\,|\,D,\mathcal{H})$ is often a complex distribution. In a variational approach to inference, we introduce an approximating probability distribution over the parameters, $Q(\mathbf{w};\boldsymbol{\theta})$, and optimize this distribution (by varying its own parameters $\boldsymbol{\theta}$) so that it approximates the posterior distribution of the parameters $P(\mathbf{w}\,|\,D,\mathcal{H})$ well.

One objective function we may choose to measure the quality of the approximation is the variational free energy

$$\tilde{F}(\boldsymbol{\theta}) = \int \mathrm{d}^k\mathbf{w}\; Q(\mathbf{w};\boldsymbol{\theta}) \ln \frac{Q(\mathbf{w};\boldsymbol{\theta})}{P(D\,|\,\mathbf{w},\mathcal{H})P(\mathbf{w}\,|\,\mathcal{H})}. \tag{33.34}$$

The denominator $P(D\,|\,\mathbf{w},\mathcal{H})P(\mathbf{w}\,|\,\mathcal{H})$ is, within a multiplicative constant, the posterior probability $P(\mathbf{w}\,|\,D,\mathcal{H}) = P(D\,|\,\mathbf{w},\mathcal{H})P(\mathbf{w}\,|\,\mathcal{H})/P(D\,|\,\mathcal{H})$. So the variational free energy $\tilde{F}(\boldsymbol{\theta})$ can be viewed as the sum of $-\ln P(D\,|\,\mathcal{H})$ and the relative entropy between $Q(\mathbf{w};\boldsymbol{\theta})$ and $P(\mathbf{w}\,|\,D,\mathcal{H})$. $\tilde{F}(\boldsymbol{\theta})$ is bounded below by $-\ln P(D\,|\,\mathcal{H})$ and only attains this value for $Q(\mathbf{w};\boldsymbol{\theta}) = P(\mathbf{w}\,|\,D,\mathcal{H})$. For certain models and certain approximating distributions, this free energy, and its derivatives with respect to the approximating distribution's parameters, can be evaluated.

The approximation of posterior probability distributions using variational free energy minimization provides a useful approach to approximating Bayesian inference in a number of fields ranging from neural networks to the decoding of error-correcting codes (Hinton and van Camp, 1993; Hinton and Zemel, 1994; Dayan *et al.*, 1995; Neal and Hinton, 1998; MacKay, 1995a). The method is sometimes called *ensemble learning* to contrast it with traditional learning processes in which a single parameter vector is optimized. Another name for it is *variational Bayes*. Let us examine how ensemble learning works in the simple case of a Gaussian distribution.

▶ 33.5 The case of an unknown Gaussian: approximating the posterior distribution of μ and σ

We will fit an approximating ensemble $Q(\mu,\sigma)$ to the posterior distribution that we studied in Chapter 24,

$$P(\mu,\sigma\,|\,\{x_n\}_{n=1}^N) = \frac{P(\{x_n\}_{n=1}^N\,|\,\mu,\sigma)P(\mu,\sigma)}{P(\{x_n\}_{n=1}^N)} \tag{33.35}$$

$$= \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{N(\mu-\bar{x})^2+S}{2\sigma^2}\right) \frac{1}{\sigma_\mu}\frac{1}{\sigma}}{P(\{x_n\}_{n=1}^N)}. \tag{33.36}$$

We make the single assumption that the approximating ensemble is separable in the form $Q(\mu,\sigma) = Q_\mu(\mu)Q_\sigma(\sigma)$. No restrictions on the functional form of $Q_\mu(\mu)$ and $Q_\sigma(\sigma)$ are made.

We write down a variational free energy,

$$\tilde{F}(Q) = \int \mathrm{d}\mu\,\mathrm{d}\sigma\; Q_\mu(\mu)Q_\sigma(\sigma) \ln \frac{Q_\mu(\mu)Q_\sigma(\sigma)}{P(D\,|\,\mu,\sigma)P(\mu,\sigma)}. \tag{33.37}$$

We can find the optimal separable distribution Q by considering separately the optimization of $\tilde{F}$ over $Q_\mu(\mu)$ for fixed $Q_\sigma(\sigma)$, and then the optimization of $Q_\sigma(\sigma)$ for fixed $Q_\mu(\mu)$.

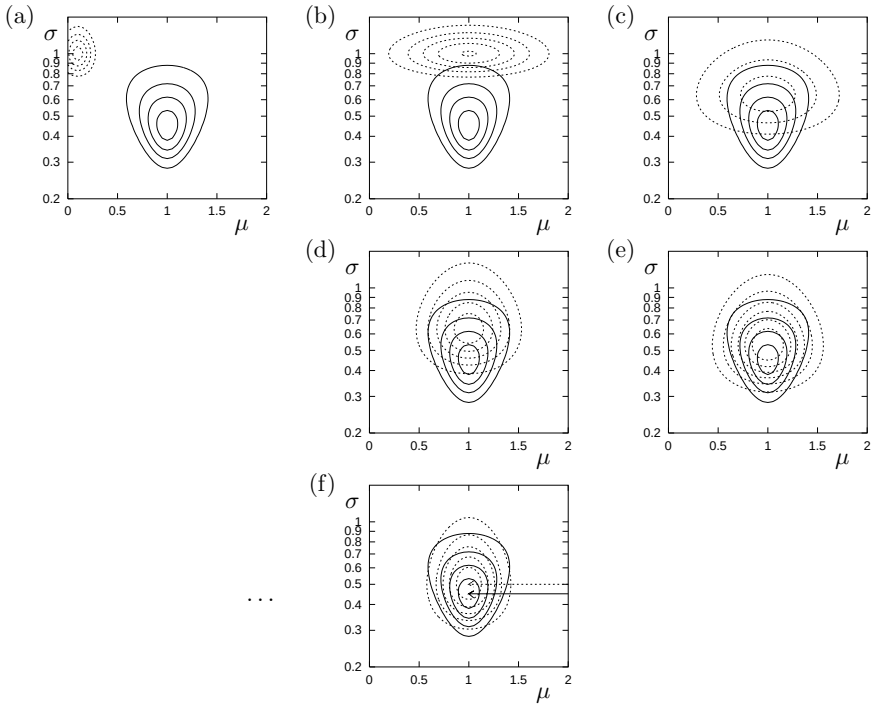


Figure 33.4. Optimization of an approximating distribution. The posterior distribution $P(\mu, \sigma \mid \{x_n\})$, which is the same as that in figure 24.1, is shown by solid contours. (a) Initial condition. The approximating distribution $Q(\mu, \sigma)$ (dotted contours) is an arbitrary separable distribution. (b) Q_μ has been updated, using equation (33.41). (c) Q_σ has been updated, using equation (33.44). (d) Q_μ updated again. (e) Q_σ updated again. (f) Converged approximation (after 15 iterations). The arrows point to the peaks of the two distributions, which are at $\sigma_N = 0.45$ (for P) and $\sigma_{N-1} = 0.5$ (for Q).

Optimization of $Q_\mu(\mu)$

As a functional of $Q_\mu(\mu)$, $\tilde{F}$ is:

$$\tilde{F} = -\int \mathrm{d}\mu\, Q_\mu(\mu) \left[\int \mathrm{d}\sigma\, Q_\sigma(\sigma) \ln P(D \mid \mu, \sigma) + \ln[P(\mu)/Q_\mu(\mu)] \right] + \kappa \quad (33.38)$$

$$= \int \mathrm{d}\mu\, Q_\mu(\mu) \left[\int \mathrm{d}\sigma\, Q_\sigma(\sigma) N\beta \frac{1}{2}(\mu - \bar{x})^2 + \ln Q_\mu(\mu) \right] + \kappa', \quad (33.39)$$

where $\beta \equiv 1/\sigma^2$ and κ denote constants that do not depend on $Q_\mu(\mu)$. The dependence on Q_σ thus collapses down to a simple dependence on the mean

$$\bar{\beta} \equiv \int \mathrm{d}\sigma\, Q_\sigma(\sigma) 1/\sigma^2. \quad (33.40)$$

Now we can recognize the function $-N\bar{\beta}\frac{1}{2}(\mu - \bar{x})^2$ as the logarithm of a Gaussian identical to the posterior distribution for a particular value of $\beta = \bar{\beta}$. Since a relative entropy $\int Q \ln(Q/P)$ is minimized by setting $Q = P$, we can immediately write down the distribution $Q_\mu^{\mathrm{opt}}(\mu)$ that minimizes $\tilde{F}$ for fixed Q_σ:

$$Q_\mu^{\mathrm{opt}}(\mu) = P(\mu \mid D, \bar{\beta}, \mathcal{H}) = \mathrm{Normal}(\mu; \bar{x}, \sigma^2_{\mu|D}). \quad (33.41)$$

where $\sigma^2_{\mu|D} = 1/(N\bar{\beta})$.

Optimization of $Q_\sigma(\sigma)$

We represent $Q_\sigma(\sigma)$ using the density over β, $Q_\sigma(\beta) \equiv Q_\sigma(\sigma)\,|\mathrm{d}\sigma/\mathrm{d}\beta|$. As a functional of $Q_\sigma(\beta)$, $\tilde{F}$ is (neglecting additive constants):

The prior $P(\sigma) \propto 1/\sigma$ transforms to $P(\beta) \propto 1/\beta$.

$$\tilde{F} = -\int \mathrm{d}\beta\, Q_\sigma(\beta) \left[\int \mathrm{d}\mu\, Q_\mu(\mu) \ln P(D \mid \mu, \sigma) + \ln[P(\beta)/Q_\sigma(\beta)] \right] \quad (33.42)$$

$$= \int \mathrm{d}\beta\, Q_\sigma(\beta) \left[(N\sigma^2_{\mu|D} + S)\beta/2 - \left(\tfrac{N}{2} - 1\right) \ln \beta + \ln Q_\sigma(\beta) \right], \quad (33.43)$$

where the integral over μ is performed assuming $Q_\mu(\mu) = Q_\mu^{\rm opt}(\mu)$. Here, the β-dependent expression in square brackets can be recognized as the logarithm of a gamma distribution over β – see equation (23.15) – giving as the distribution that minimizes $\tilde{F}$ for fixed Q_μ:

$$Q_\sigma^{\rm opt}(\beta) = \Gamma(\beta; b', c'), \tag{33.44}$$

with

$$\frac{1}{b'} = \frac{1}{2}(N\sigma_{\mu|D}^2 + S) \quad \text{and} \quad c' = \frac{N}{2}. \tag{33.45}$$

In figure 33.4, these two update rules (33.41, 33.44) are applied alternately, starting from an arbitrary initial condition. The algorithm converges to the optimal approximating ensemble in a few iterations.

Direct solution for the joint optimum $Q_\mu(\mu)Q_\sigma(\sigma)$

In this problem, we do not need to resort to iterative computation to find the optimal approximating ensemble. Equations (33.41) and (33.44) define the optimum implicitly. We must simultaneously have $\sigma_{\mu|D}^2 = 1/(N\bar{\beta})$, and $\bar{\beta} = b'c'$. The solution is:

$$1/\bar{\beta} = S/(N-1). \tag{33.46}$$

This is similar to the true posterior distribution of σ, which is a gamma distribution with $c' = \frac{N-1}{2}$ and $1/b' = S/2$ (see equation 24.13). This true posterior also has a mean value of β satisfying $1/\bar{\beta} = S/(N-1)$; the only difference is that the approximating distribution's parameter c' is too large by $1/2$.

> The approximations given by variational free energy minimization always tend to be more compact than the true distribution.

In conclusion, ensemble learning gives an approximation to the posterior that agrees nicely with the conventional estimators. The approximate posterior distribution over β is a gamma distribution with mean $\bar{\beta}$ corresponding to a variance of $\sigma^2 = S/(N-1) = \sigma_{N-1}^2$. And the approximate posterior distribution over μ is a Gaussian with mean $\bar{x}$ and standard deviation $\sigma_{N-1}/\sqrt{N}$.

The variational free energy minimization approach has the nice property that it is parameterization-independent; it avoids the problem of basis-dependence from which MAP methods and Laplace's method suffer.

A convenient software package for automatic implementation of variational inference in graphical models is `VIBES` (Bishop *et al.*, 2002). It plays the same role for variational inference as `BUGS` plays for Monte Carlo inference.

▶ 33.6 Interlude

One of my students asked:

> How do you ever come up with a useful approximating distribution, given that the true distribution is so complex you can't compute it directly?

Let's answer this question in the context of Bayesian data modelling. Let the 'true' distribution of interest be the posterior probability distribution over a set of parameters $\mathbf{x}$, $P(\mathbf{x} \mid D)$. A standard data modelling practice is to find a single, 'best-fit' setting of the parameters, $\mathbf{x}^*$, for example, by finding the maximum of the likelihood function $P(D \mid \mathbf{x})$, or of the posterior distribution.

One interpretation of this standard practice is that the full description of our knowledge about $\mathbf{x}$, $P(\mathbf{x} \mid D)$, is being approximated by a delta-function, a probability distribution concentrated on $\mathbf{x}^*$. From this perspective, *any* approximating distribution $Q(\mathbf{x}; \boldsymbol{\theta})$, no matter how crummy it is, *has* to be an improvement on the spike produced by the standard method! So even if we use only a simple Gaussian approximation, we are doing well.

We now study an application of the variational approach to a realistic example – data clustering.

▶ 33.7 K-means clustering and the expectation–maximization algorithm as a variational method

In Chapter 20, we introduced the soft K-means clustering algorithm, version 1. In Chapter 22, we introduced versions 2 and 3 of this algorithm, and motivated the algorithm as a maximum likelihood algorithm.

K-means clustering is an example of an 'expectation–maximization' (EM) algorithm, with the two steps, which we called 'assignment' and 'update', being known as the 'E-step' and the 'M-step' respectively.

We now give a more general view of K-means clustering, due to Neal and Hinton (1998), in which the algorithm is shown to optimize a variational objective function. Neal and Hinton's derivation applies to any EM algorithm.

The probability of everything

Let the parameters of the mixture model – the means, standard deviations, and weights – be denoted by $\boldsymbol{\theta}$. For each data point, there is a missing variable (also known as a latent variable), the class label k_n for that point. The probability of everything, given our assumed model $\mathcal{H}$, is

$$P(\{\mathbf{x}^{(n)}, k_n\}_{n=1}^N, \boldsymbol{\theta} \mid \mathcal{H}) \;=\; P(\boldsymbol{\theta} \mid \mathcal{H}) \prod_{n=1}^{N} \Big[P(\mathbf{x}^{(n)} \mid k_n, \boldsymbol{\theta}) P(k_n \mid \boldsymbol{\theta}) \Big]. \quad (33.47)$$

The posterior probability of everything, given the data, is proportional to the probability of everything:

$$P(\{k_n\}_{n=1}^N, \boldsymbol{\theta} \mid \{\mathbf{x}^{(n)}\}_{n=1}^N, \mathcal{H}) \;=\; \frac{P(\{\mathbf{x}^{(n)}, k_n\}_{n=1}^N, \boldsymbol{\theta} \mid \mathcal{H})}{P(\{\mathbf{x}^{(n)}\}_{n=1}^N \mid \mathcal{H})}. \quad (33.48)$$

We now approximate this posterior distribution by a separable distribution

$$Q_k(\{k_n\}_{n=1}^N)\, Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \quad (33.49)$$

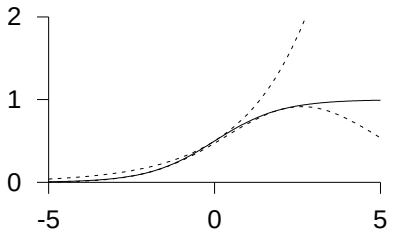
and define a variational free energy in the usual way:

$$\tilde{F}(Q_k, Q_{\boldsymbol{\theta}}) = \sum_{\{k_n\}} \int \mathrm{d}^D \boldsymbol{\theta}\; Q_k(\{k_n\}_{n=1}^N)\, Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{Q_k(\{k_n\}_{n=1}^N)\, Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\{\mathbf{x}^{(n)}, k_n\}_{n=1}^N, \boldsymbol{\theta} \mid \mathcal{H})}. \quad (33.50)$$

$\tilde{F}$ is bounded below by minus the evidence, $\ln P(\{\mathbf{x}^{(n)}\}_{n=1}^N \mid \mathcal{H})$. We can now make an iterative algorithm with an 'assignment' step and an 'update' step. In the assignment step, $Q_k(\{k_n\}_{n=1}^N)$ is adjusted to reduce $\tilde{F}$, for fixed $Q_{\boldsymbol{\theta}}$; in the update step, $Q_{\boldsymbol{\theta}}$ is adjusted to reduce $\tilde{F}$, for fixed Q_k.

If we wish to obtain exactly the soft K-means algorithm, we impose a further constraint on our approximating distribution: $Q_{\boldsymbol{\theta}}$ is constrained to be a delta function centred on a point estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$:

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \quad (33.51)$$



Upper bound

$$\frac{1}{1+e^{-a}} \leq \exp(\mu a - H_2^e(\mu)) \qquad \mu \in [0,1]$$

Lower bound

$$\frac{1}{1+e^{-a}} \geq g(\nu)\exp\left[(a-\nu)/2 - \lambda(\nu)(a^2-\nu^2)\right]$$

where $\lambda(\nu) = [g(\nu) - 1/2]/2\nu$.

Figure 33.5. Illustration of the Jaakkola–Jordan variational method. Upper and lower bounds on the logistic function (solid line)

$$g(a) \equiv \frac{1}{1+e^{-a}}.$$

These upper and lower bounds are exponential or Gaussian functions of a, and so easier to integrate over. The graph shows the sigmoid function and upper and lower bounds with $\mu = 0.505$ and $\nu = -2.015$.

Unfortunately, this distribution contributes to the variational free energy an infinitely large integral $\int \mathrm{d}^D\boldsymbol{\theta}\, Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, so we'd better leave that term out of $\tilde{F}$, treating it as an additive constant. [Using a delta function $Q_{\boldsymbol{\theta}}$ is not a good idea if our aim is to minimize $\tilde{F}$!] Moving on, our aim is to derive the soft K-means algorithm.

▷ Exercise 33.2.[2] Show that, given $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, the optimal Q_k, in the sense of minimizing $\tilde{F}$, is a separable distribution in which the probability that $k_n = k$ is given by the responsibility $r_k^{(n)}$.

▷ Exercise 33.3.[3] Show that, given a separable Q_k as described above, the optimal $\boldsymbol{\theta}^*$, in the sense of minimizing $\tilde{F}$, is obtained by the update step of the soft K-means algorithm. (Assume a uniform prior on $\boldsymbol{\theta}$.)

Exercise 33.4.[4] We can instantly improve on the infinitely large value of $\tilde{F}$ achieved by soft K-means clustering by allowing $Q_{\boldsymbol{\theta}}$ to be a more general distribution than a delta-function. Derive an update step in which $Q_{\boldsymbol{\theta}}$ is allowed to be a separable distribution, a product of $Q_\mu(\{\mu\})$, $Q_\sigma(\{\sigma\})$, and $Q_\pi(\pi)$. Discuss whether this generalized algorithm still suffers from soft K-means's 'kaboom' problem, where the algorithm glues an ever-shrinking Gaussian to one data point.

Sadly, while it sounds like a promising generalization of the algorithm to allow $Q_{\boldsymbol{\theta}}$ to be a non-delta-function, and the 'kaboom' problem goes away, other artefacts can arise in this approximate inference method, involving local minima of $\tilde{F}$. For further reading, see (MacKay, 1997a; MacKay, 2001).

▶ 33.8 Variational methods other than free energy minimization

There are other strategies for approximating a complicated distribution $P(\mathbf{x})$, in addition to those based on minimizing the relative entropy between an approximating distribution, Q, and P. One approach pioneered by Jaakkola and Jordan is to create adjustable upper and lower bounds Q^U and Q^L to P, as illustrated in figure 33.5. These bounds (which are unnormalized densities) are parameterized by variational parameters which are adjusted in order to obtain the tightest possible fit. The lower bound can be adjusted to *maximize*

$$\sum_{\mathbf{x}} Q^L(\mathbf{x}), \tag{33.52}$$

and the upper bound can be adjusted to *minimize*

$$\sum_{\mathbf{x}} Q^U(\mathbf{x}). \tag{33.53}$$

Using the normalized versions of the optimized bounds we then compute approximations to the predictive distributions. Further reading on such methods can be found in the references (Jaakkola and Jordan, 2000a; Jaakkola and Jordan, 2000b; Jaakkola and Jordan, 1996; Gibbs and MacKay, 2000).

Further reading

The Bethe and Kikuchi free energies

In Chapter 26 we discussed the sum–product algorithm for functions of the factor-graph form (26.1). If the factor graph is tree-like, the sum–product algorithm converges and correctly computes the marginal function of any variable x_n and can also yield the joint marginal function of subsets of variables that appear in a common factor, such as $\mathbf{x}_m$.

The sum–product algorithm may also be applied to factor graphs that are not tree-like. If the algorithm converges to a fixed point, it has been shown that that fixed point is a stationary point (usually a minimum) of a function of the messages called the Kikuchi free energy. In the special case where all factors in factor graph are functions of one or two variables, the Kikuchi free energy is called the Bethe free energy.

For articles on this idea, and new approximate inference algorithms motivated by it, see Yedidia (2000); Yedidia *et al.* (2000); Welling and Teh (2001); Yuille (2001); Yedidia *et al.* (2001b); Yedidia *et al.* (2001a).

▶ 33.9 Further exercises

Exercise 33.5.[2, p.435] This exercise explores the assertion, made above, that the approximations given by variational free energy minimization always tend to be more compact than the true distribution. Consider a two dimensional Gaussian distribution $P(\mathbf{x})$ with axes aligned with the directions $\mathbf{e}^{(1)} = (1, 1)$ and $\mathbf{e}^{(2)} = (1, -1)$. Let the variances in these two directions be σ_1^2 and σ_2^2. What is the optimal variance if this distribution is approximated by a *spherical* Gaussian with variance σ_Q^2, optimized by variational free energy minimization? If we instead optimized the objective function

$$G = \int \mathrm{d}\mathbf{x}\, P(\mathbf{x}) \ln \frac{P(\mathbf{x})}{Q(\mathbf{x}; \sigma^2)}, \tag{33.54}$$

what would be the optimal value of σ^2? Sketch a contour of the true distribution $P(\mathbf{x})$ and the two approximating distributions in the case $\sigma_1/\sigma_2 = 10$.

[Note that in general it is not possible to evaluate the objective function G, because integrals under the true distribution $P(\mathbf{x})$ are usually intractable.]

Exercise 33.6.[2, p.436] What do you think of the idea of using a variational method to optimize an approximating distribution Q which we then use as a proposal density for importance sampling?

Exercise 33.7.[2] Define the *relative entropy* or *Kullback–Leibler divergence* between two probability distributions P and Q, and state Gibbs' inequality.

Consider the problem of approximating a joint distribution $P(x, y)$ by a separable distribution $Q(x, y) = Q_X(x)Q_Y(y)$. Show that if the objec-

tive function for this approximation is

$$G(Q_X, Q_Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{Q_X(x) Q_Y(y)}$$

that the minimal value of G is achieved when Q_X and Q_Y are equal to the marginal distributions over x and y.

Now consider the alternative objective function

$$F(Q_X, Q_Y) = \sum_{x,y} Q_X(x) Q_Y(y) \log_2 \frac{Q_X(x) Q_Y(y)}{P(x,y)};$$

the probability distribution $P(x,y)$ shown in the margin is to be approximated by a separable distribution $Q(x,y) = Q_X(x)Q_Y(y)$. State the value of $F(Q_X, Q_Y)$ if Q_X and Q_Y are set to the marginal distributions over x and y.

$P(x,y)$		x			
		1	2	3	4
y	1	1/8	1/8	0	0
	2	1/8	1/8	0	0
	3	0	0	1/4	0
	4	0	0	0	1/4

Show that $F(Q_X, Q_Y)$ has three distinct minima, identify those minima, and evaluate F at each of them.

▶ 33.10 Solutions

Solution to exercise 33.5 (p.434). We need to know the relative entropy between two one-dimensional Gaussian distributions:

$$\begin{aligned}
&\int \mathrm{d}x \, \mathrm{Normal}(x; 0, \sigma_Q) \ln \frac{\mathrm{Normal}(x; 0, \sigma_Q)}{\mathrm{Normal}(x; 0, \sigma_P)} \\
&= \int \mathrm{d}x \, \mathrm{Normal}(x; 0, \sigma_Q) \left[\ln \frac{\sigma_P}{\sigma_Q} - \frac{1}{2} x^2 \left(\frac{1}{\sigma_Q^2} - \frac{1}{\sigma_P^2} \right) \right] \qquad (33.55) \\
&= \frac{1}{2} \left(\ln \frac{\sigma_P^2}{\sigma_Q^2} - 1 + \frac{\sigma_Q^2}{\sigma_P^2} \right). \qquad (33.56)
\end{aligned}$$

So, if we approximate P, whose variances are σ_1^2 and σ_2^2, by Q, whose variances are both σ_Q^2, we find

$$F(\sigma_Q^2) = \frac{1}{2} \left(\ln \frac{\sigma_1^2}{\sigma_Q^2} - 1 + \frac{\sigma_Q^2}{\sigma_1^2} + \ln \frac{\sigma_2^2}{\sigma_Q^2} - 1 + \frac{\sigma_Q^2}{\sigma_2^2} \right); \qquad (33.57)$$

differentiating,

$$\frac{\mathrm{d}}{\mathrm{d} \ln(\sigma_Q^2)} F = \frac{1}{2} \left[-2 + \left(\frac{\sigma_Q^2}{\sigma_1^2} + \frac{\sigma_Q^2}{\sigma_2^2} \right) \right], \qquad (33.58)$$

which is zero when

$$\frac{1}{\sigma_Q^2} = \frac{1}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right). \qquad (33.59)$$

Thus we set the approximating distribution's inverse variance to the mean inverse variance of the target distribution P.

In the case $\sigma_1 = 10$ and $\sigma_2 = 1$, we obtain $\sigma_Q \simeq \sqrt{2}$, which is just a factor of $\sqrt{2}$ larger than σ_2, pretty much *independent* of the value of the larger standard deviation σ_1. *Variational free energy minimization typically leads to approximating distributions whose length scales match the shortest length scale of the target distribution.* The approximating distribution might be viewed as *too compact*.

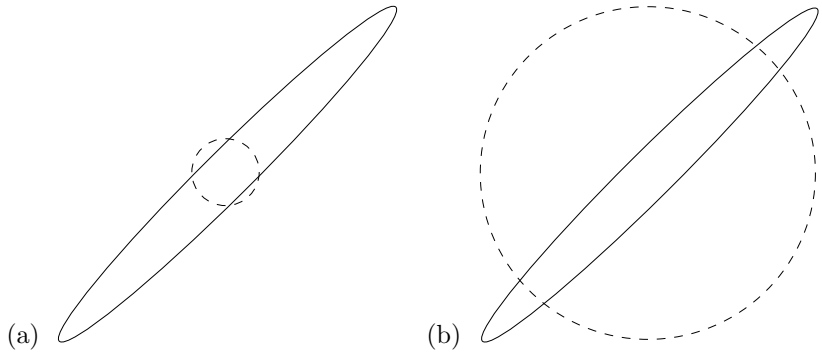


Figure 33.6. Two separable Gaussian approximations (dotted lines) to a bivariate Gaussian distribution (solid line). (a) The approximation that minimizes the variational free energy. (b) The approximation that minimizes the objective function G. In each figure, the lines show the contours at which $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = 1$, where $\mathbf{A}$ is the inverse covariance matrix of the Gaussian.

In contrast, if we use the objective function G then we find:

$$G(\sigma_Q^2) = \frac{1}{2}\left(\ln \sigma_Q^2 + \frac{\sigma_1^2}{\sigma_Q^2} + \ln \sigma_Q^2 + \frac{\sigma_2^2}{\sigma_Q^2}\right) + \text{constant}, \tag{33.60}$$

where the constant depends on σ_1 and σ_2 only. Differentiating,

$$\frac{\mathrm{d}}{\mathrm{d}\ln \sigma_Q^2} G = \frac{1}{2}\left[2 - \left(\frac{\sigma_1^2}{\sigma_Q^2} + \frac{\sigma_2^2}{\sigma_Q^2}\right)\right], \tag{33.61}$$

which is zero when

$$\sigma_Q^2 = \frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right). \tag{33.62}$$

Thus we set the approximating distribution's variance to the mean variance of the target distribution P.

In the case $\sigma_1 = 10$ and $\sigma_2 = 1$, we obtain $\sigma_Q \simeq 10/\sqrt{2}$, which is just a factor of $\sqrt{2}$ smaller than σ_1, independent of the value of σ_2.

The two approximations are shown to scale in figure 33.6.

Solution to exercise 33.6 (p.434). The best possible variational approximation is of course the target distribution P. Assuming that this is not possible, a good variational approximation is *more compact* than the true distribution. In contrast, a good sampler is *more heavy tailed* than the true distribution. An over-compact distribution would be a lousy sampler with a large variance.