32

Exact Monte Carlo Sampling

32.1 The problem with Monte Carlo methods

For high-dimensional problems, the most widely used random sampling methods are Markov chain Monte Carlo methods like the Metropolis method, Gibbs sampling, and slice sampling.

The problem with all these methods is this: yes, a given algorithm can be guaranteed to produce samples from the target density $P(\mathbf{x})$ asymptotically, 'once the chain has converged to the equilibrium distribution'. But if one runs the chain for too short a time T, then the samples will come from some other distribution $P^{(T)}(\mathbf{x})$. For how long must the Markov chain be run before it has 'converged'? As was mentioned in Chapter 29, this question is usually very hard to answer. However, the pioneering work of Propp and Wilson (1996) allows one, for certain chains, to answer this very question; furthermore Propp and Wilson show how to obtain 'exact' samples from the target density.

32.2 Exact sampling concepts

Propp and Wilson's *exact sampling method* (also known as 'perfect simulation' or 'coupling from the past') depends on three ideas.

Coalescence of coupled Markov chains

First, if several Markov chains starting from different initial conditions share a single random-number generator, then their trajectories in state space may *coalesce*; and, having, coalesced, will not separate again. If *all* initial conditions lead to trajectories that coalesce into a single trajectory, then we can be sure that the Markov chain has 'forgotten' its initial condition. Figure 32.1a-i shows twenty-one Markov chains identical to the one described in section 29.4, which samples from $\{0, 1, \ldots, 20\}$ using the Metropolis algorithm (figure 29.12, p.368); each of the chains has a different initial condition but they are all driven by a single random number generator; the chains coalesce after about 80 steps. Figure 32.1(a-ii) shows the same Markov chains with a different random number seed; in this case, coalescence does not occur until 400 steps have elapsed (not shown). Figure 32.1b shows similar Markov chains, each of which has identical proposal density to those in section 29.4 and figure 32.1a; but in figure 32.1b, the proposed move at each step, 'left' or 'right', is obtained in the same way by all the chains at any timestep, independent of the current state. This coupling of the chains changes the statistics of coalescence. Because two neighbouring paths only merge when a rejection occurs, and rejections only occur at the walls (for this particular Markov chain), coa-

lescence will occur only when the chains are all in the leftmost state or all in the rightmost state.

Coupling from the past

How can we use the coalescence property to find an exact sample from the equilibrium distribution of the chain? The state of the system at the moment when complete coalescence occurs is not a valid sample from the equilibrium distribution; for example in figure 32.1b, final coalescence always occurs when the state is against one of the two walls, because trajectories only merge at the walls. So sampling forward in time until coalescence occurs is not a valid method.

The second key idea of exact sampling is that we can obtain exact samples by sampling *from a time T_0 in the past, up to the present.* If coalescence has occured, the present sample is an unbiased sample from the equilibrium distribution; if not, we restart the simulation from a time T_0 further into the past, *reusing the same random numbers.* The simulation is repeated at a sequence of ever more distant times T_0, with a doubling of T_0 from one run to the next being a convenient choice. When coalescence occurs at a time before 'the present', we can record $x(0)$ as an *exact sample* from the equilibrium distribution of the Markov chain.

Figure 32.2 shows two exact samples produced in this way. In the leftmost panel of figure 32.2a, we start twenty-one chains in all possible initial conditions at $T_0 = -50$ and run them forward in time. Coalescence does not occur. We restart the simulation from all possible initial conditions at $T_0 = -100$, and reset the random number generator in such a way that the random numbers generated at each time t (in particular, from $t = -50$ to $t = 0$) will be identical to what they were in the first run. Notice that the trajectories produced from $t = -50$ to $t = 0$ by these runs that started from $T_0 = -100$ are identical to a *subset* of the trajectories in the first simulation with $T_0 = -50$. Coalescence still does not occur, so we double T_0 again to $T_0 = -200$. This time, all the trajectories coalesce and we obtain an exact sample, shown by the arrow. If we pick an earlier time such as $T_0 = -500$, all the trajectories must still end in the same point at $t = 0$, since all trajectories must pass through some state at $t = -200$, and all those states lead to the same final point. So if we ran the Markov chain for an infinite time in the past, from any initial condition, it would end in the same state. Figure 32.2b shows an exact sample produced in the same way with the Markov chains of figure 32.1b.

This method, called *coupling from the past*, is important because it allows us to obtain exact samples from the equilibrium distribution; but, as described here, it is of little practical use, since we are obliged to simulate chains starting in *all* initial states. In the examples shown, there are only twenty-one states, but in any realistic sampling problem there will be an utterly enormous number of states – think of the 2^{1000} states of a system of 1000 binary spins, for example. The whole point of introducing Monte Carlo methods was to try to avoid having to visit all the states of such a system!

Monotonicity

Having established that we can obtain valid samples by simulating forward from times in the past, starting in *all* possible states at those times, the third trick of Propp and Wilson, which makes the exact sampling method useful in practice, is the idea that, for some Markov chains, it may be possible to detect coalescence of all trajectories *without simulating all those trajectories.*
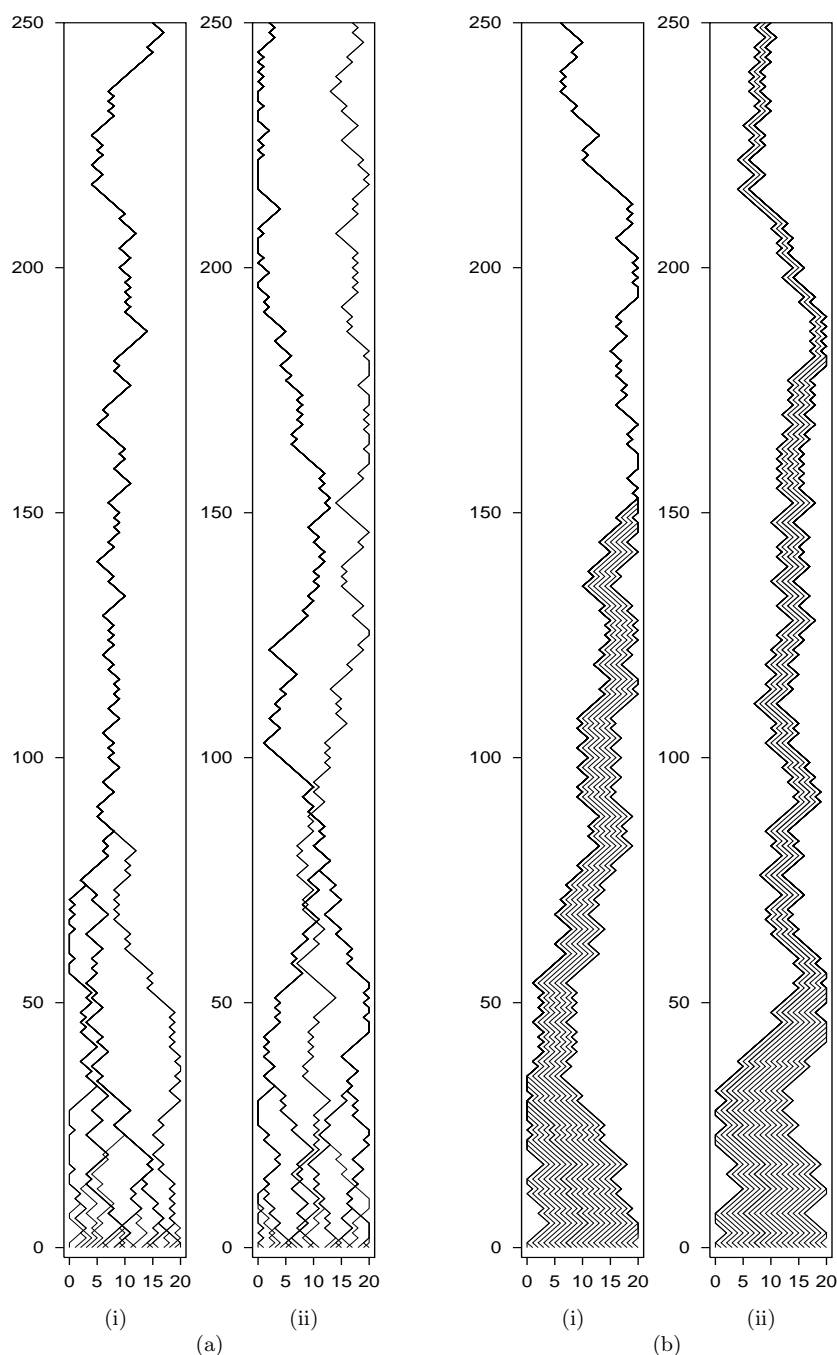
Figure 32.1. Coalescence, the first idea behind the exact sampling method. In the leftmost panel, coalescence occurred within 100 steps. Different coalescence properties are obtained depending on the way each state uses the random numbers it is supplied with. (a) Two runs of a Metropolis simulator in which the random bits that determine the proposed step depend on the current state; a different random number seed was used in each case. (b) In this simulator the random proposal ('left' or 'right') is the same for all states. In each panel, one of the paths, the one starting at location $x = 8$, has been highlighted.
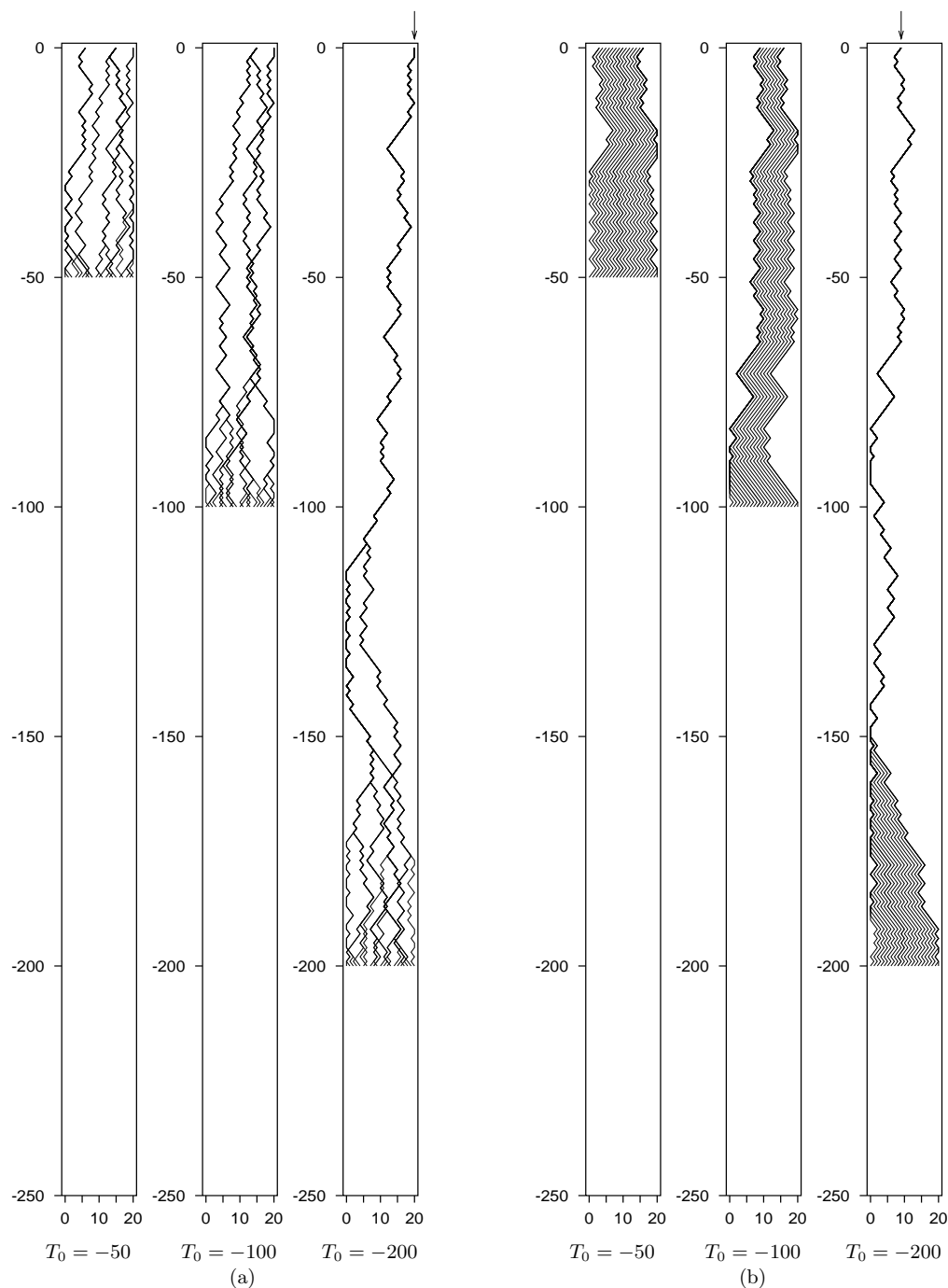
$T_0 = -50$	$T_0 = -100$	$T_0 = -200$	$T_0 = -50$	$T_0 = -100$	$T_0 = -200$
	(a)			(b)	

Figure 32.2. 'Coupling from the past', the second idea behind the exact sampling method.
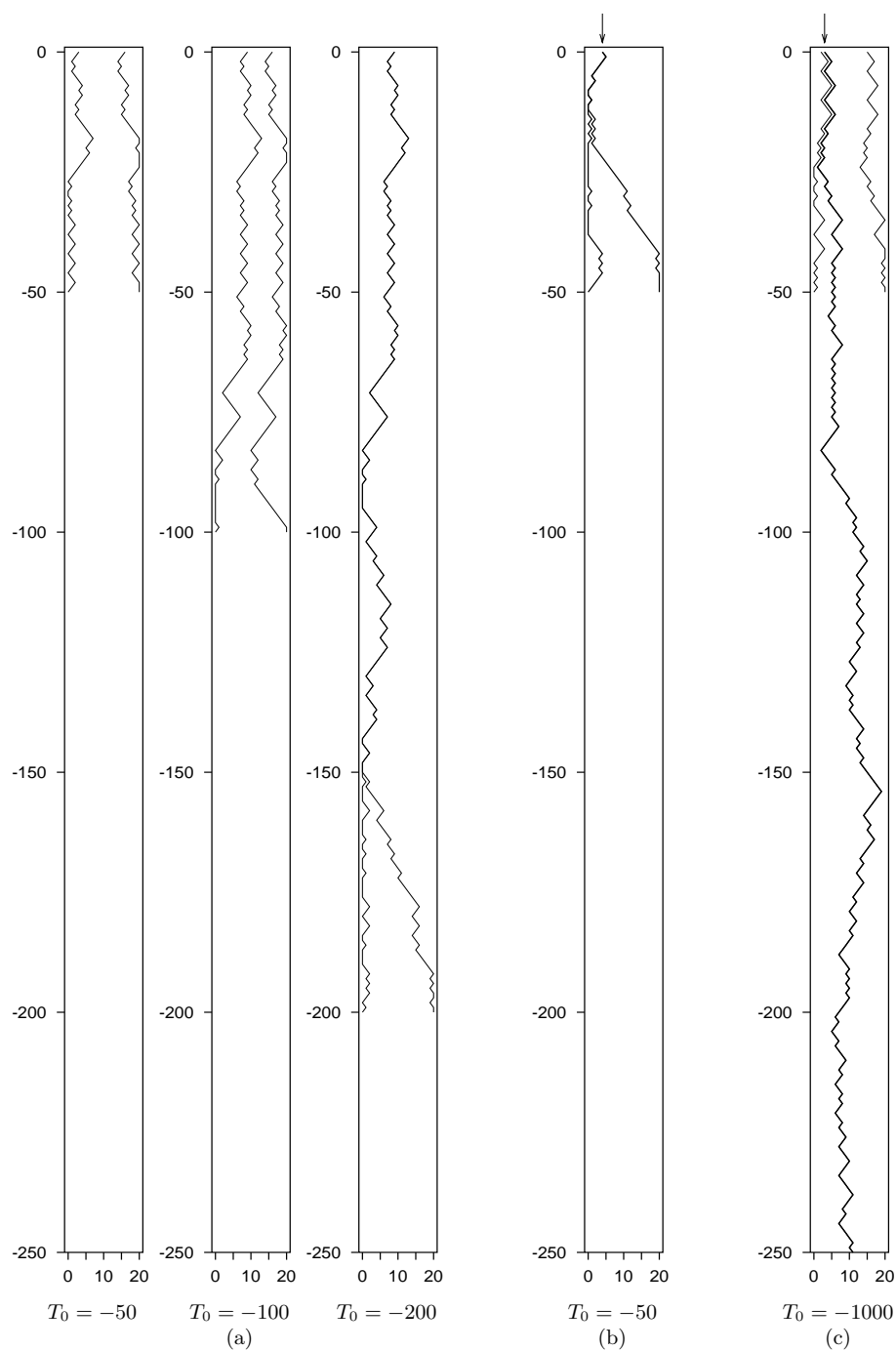
Figure 32.3. (a) Ordering of states, the third idea behind the exact sampling method. The trajectories shown here are the left-most and right-most trajectories of figure 32.2b. In order to establish what the state at time zero is, we only need to run simulations from $T_0 = -50$, $T_0 = -100$, and $T_0 = -200$, after which point coalescence occurs.

   (b,c) Two more exact samples from the target density, generated by this method, and different random number seeds. The initial times required were $T_0 = -50$ and $T_0 = -1000$, respectively.

This property holds, for example, in the chain of figure 32.1b, which has the property that *two trajectories never cross*. So if we simply track the two trajectories starting from the leftmost and rightmost states, we will know that coalescence of *all* trajectories has occurred when *those two* trajectories coalesce. Figure 32.3b illustrates this idea by showing only the left-most and right-most trajectories of figure 32.2b. Figure 32.3(c,d) shows two more exact samples from the same equilibrium distribution generated by running the 'coupling from the past' method starting from the two end-states alone. In (c), two runs coalesced starting from $T_0 = -50$; in (d), it was necessary to try times up to $T_0 = -1000$ to achieve coalescence.

32.3 Exact sampling from interesting distributions

In the toy problem we studied, the states could be put in a one-dimensional order such that no two trajectories crossed. The states of many interesting state spaces can also be put into a *partial order* and coupled Markov chains can be found that respect this partial order. [An example of a partial order on the four possible states of two spins is this: $(+,+) > (+,-) > (-,-)$; and $(+,+) > (-,+) > (-,-)$; and the states $(+,-)$ and $(-,+)$ are not ordered.] For such systems, we can show that coalescence has occurred merely by verifying that coalescence has occurred for all the histories whose initial states were 'maximal' and 'minimal' states of the state space.

As an example, consider the Gibbs sampling method applied to a ferromagnetic Ising spin system, with the partial ordering of states being defined thus: state $\mathbf{x}$ is 'greater than or equal to' state $\mathbf{y}$ if $x_i \geq y_i$ for all spins i. The maximal and minimal states are the the all-up and all-down states. The Markov chains are coupled together as shown in algorithm 32.4. Propp and Wilson (1996) show that exact samples can be generated for this system, although the time to find exact samples is large if the Ising model is below its critical temperature, since the Gibbs sampling method itself is slowly-mixing under these conditions. Propp and Wilson have improved on this method for the Ising model by using a Markov chain called the single-bond heat bath algorithm to sample from a related model called the random cluster model; they show that exact samples from the random cluster model can be obtained rapidly and can be converted into exact samples from the Ising model. Their ground-breaking paper includes an exact sample from a 16-million-spin Ising model at its critical temperature. A sample for a smaller Ising model is shown in figure 32.5.

A generalization of the exact sampling method for 'non-attractive' distributions

The method of Propp and Wilson for the Ising model, sketched above, can only be applied to probability distributions that are, as they call them, 'attractive'. Rather than define this term, let's say what it means, for practical purposes: the method can be applied to spin systems in which all the couplings are positive (e.g., the ferromagnet), and to a few special spin systems with negative couplings (e.g., as we already observed in Chapter 31, the rectangular ferromagnet and antiferromagnet are equivalent); but it cannot be applied to general spin systems in which some couplings are negative, because in such systems the trajectories followed by the all-up and all-down states are not guaranteed to be upper and lower bounds for the set of all trajectories. Fortunately, however, we do not need to be so strict. It is possible to re-

Compute $a_i := \sum_j J_{ij} x_j$
Draw u from Uniform$(0, 1)$
If $u < 1/(1 + e^{-a_i})$
  $x_i := +1$
Else
  $x_i := -1$

Algorithm 32.4. Gibbs sampling coupling method. The Markov chains are coupled together by having all chains update the same spin i at each time step and having all chains sharing a common sequence of random numbers u.
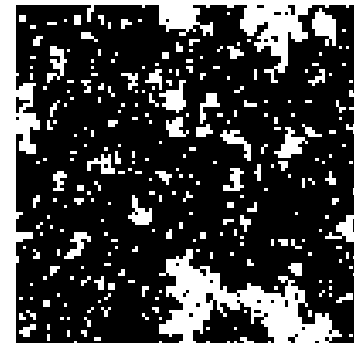

Figure 32.5. An exact sample from the Ising model at its critical temperature, produced by D.B. Wilson. Such samples can be produced within seconds on an ordinary computer by exact sampling.

express the Propp and Wilson algorithm in a way that generalizes to the case
of spin systems with negative couplings. The idea the *summary state* version
of the exact sampling method is still that we keep track of bounds on the set
of all trajectories, and detect when these bounds are equal, so as to find exact
samples. But the bounds will not themselves be actual trajectories, and they
will not necessarily be *tight* bounds.

Instead of simulating two trajectories, each of which moves in a state space
$\{-1, +1\}^N$, we simulate one *trajectory envelope* in an augmented state space
$\{-1, +1, ?\}^N$, where the symbol ? denotes 'either -1 or $+1$'. We call the state
of this augmented system the 'summary state'. An example summary state of
a six-spin system is ++-?+?. This summary state is shorthand for the set of
states

$$++-+++,\ ++-++-,\ ++--++,\ ++--+-\ .$$

The update rule at each step of the Markov chain takes a single spin, enu-
merates all possible states of the neighbouring spins that are compatible with
the current summary state, and, for each of these local scenarios, computes
the new value (+ or -) of the spin using Gibbs sampling (coupled to a random
number u as in algorithm 32.4). If all these new values agree, then the new
value of the updated spin in the summary state is set to the unanimous value
(+ or -). Otherwise, the new value of the spin in the summary state is '?'. The
initial condition, at time T_0, is given by setting all the spins in the summary
state to '?', which corresponds to considering all possible start configurations.

In the case of a spin system with positive couplings, this summary state
simulation will be identical to the simulation of the uppermost state and low-
ermost states, in the style of Propp and Wilson, with coalescence occuring
when all the '?' symbols have disappeared. The summary state method can
be applied to general spin systems with any couplings. The only shortcoming
of this method is that the envelope may describe an unnecessarily large set of
states, so there is no guarantee that the summary state algorithm will con-
verge; the time for coalescence to be *detected* may be considerably larger than
the actual time taken for the underlying Markov chain to coalesce.

The summary state scheme has been applied to exact sampling in belief
networks by Harvey and Neal (2000), and to the triangular antiferromagnetic
Ising model by Childs *et al.* (2001).

Further reading

For further reading, impressive pictures of exact samples from other distribu-
tions, and generalizations of the exact sampling method, browse the perfectly-
random sampling website.[1]

For beautiful exact-sampling demonstrations running live in your web-
browser, see Jim Propp's website.[2]

Other uses for coupling

The idea of coupling together Markov chains by having them share a random
number generator has other applications beyond exact sampling. Pinto and
Neal (2001) have shown that the accuracy of estimates obtained from a Markov
chain Monte Carlo simulation (the second problem discussed in section 29.1,

---

[1] http://www.dbwilson.com/exact/
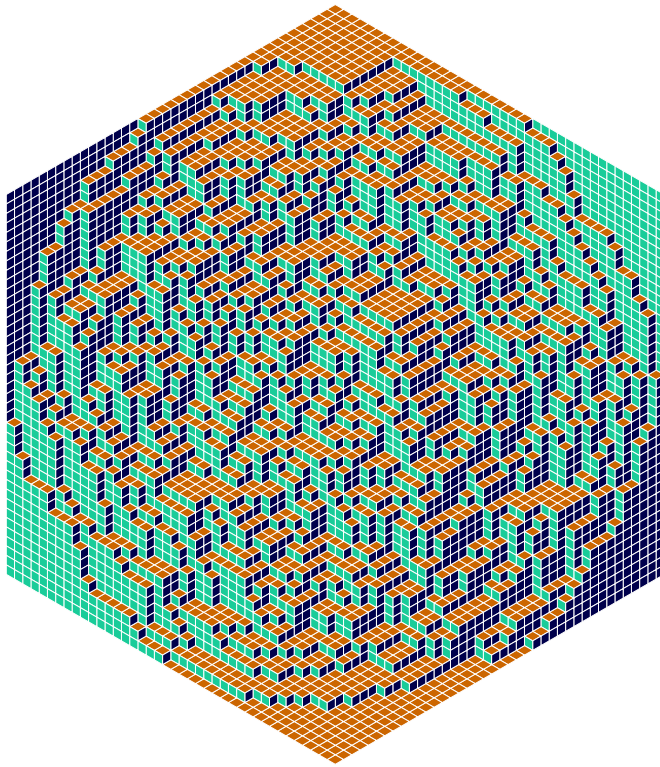[2] http://www.math.wisc.edu/∼propp/tiling/www/applets/

Figure 32.6. A perfectly random
tiling of a hexagon by lozenges,
provided by J.G. Propp and
D.B. Wilson.

p.357), using the estimator

$$\hat{\Phi}_P \equiv \frac{1}{T} \sum_t \phi(\mathbf{x}^{(t)}), \qquad (32.1)$$

can be improved by coupling the chain of interest, which converges to P, to a
second chain, which generates samples from a second, simpler distribution, Q.
The coupling must be set up in such a way that the states of the two chains
are strongly correlated. The idea is that we first estimate the expectations of
a function of interest, ϕ, under P and under Q in the normal way (32.1) and
compare the estimate under Q, $\hat{\Phi}_Q$, with the true value of the expectation
under Q, Φ_Q which we assume can be evaluated exactly. If $\hat{\Phi}_Q$ is an overes-
timate then it is likely that $\hat{\Phi}_P$ will be an overestimate too. The difference
$(\hat{\Phi}_Q - \Phi_Q)$ can thus be used to correct $\hat{\Phi}_P$.

32.4 Exercises

▷ Exercise 32.1.[2, p.421] Is there any relationship between the probability dis-
tribution of the time taken for all trajectories to coalesce, and the equi-
libration time of a Markov chain? Prove that there is a relationship, or
find a single chain that can be realized in two different ways that have
different coalescence times.

▷ Exercise 32.2.[2] Imagine that Fred ignores the requirement that the random
bits used at some time t, in every run from increasingly distant times
T_0, must be identical, and makes a coupled-Markov-chain simulator that
uses fresh random numbers every time T_0 is changed. Describe what
happens if Fred applies his method to the Markov chain that is intended
to sample from the uniform distribution over the states 0, 1, and 2, using
the Metropolis method, driven by a random bit source as in figure 32.1b.

Exercise 32.3.[5] Investigate the application of perfect sampling to linear regression in Holmes and Mallick (1998) and try to generalize it.

Exercise 32.4.[3] The concept of coalescence has many applications. Some surnames are more frequent than others, and some die out altogether. Make a model of this process; how long will it take until everyone has the same surname?

Similarly, variability in any particular portion of the human genome (which forms the basis of forensic DNA fingerprinting) is inherited like a surname. A DNA fingerprint is like a string of surnames. Should the fact that these surnames are subject to coalescences, so that some surnames are by chance more prevalent than others, affect the way in which DNA fingerprint evidence is used in court?

▷ Exercise 32.5.[2] How can you use a coin to create a random ranking of 3 people? Construct a solution that uses exact sampling. For example, you could apply exact sampling to a Markov chain in which the coin is repeatedly used alternately to decide whether to switch first and second, then whether to switch second and third.

Exercise 32.6.[5] Finding the partition function Z of a probability distribution is a difficult problem. Many Markov chain Monte Carlo methods produce valid samples from a distribution without ever finding out what Z is.

Is there any probability distribution and Markov chain such that either the time taken to produce a perfect sample or the number of random bits used to create a perfect sample are related to the value of Z? Are there some situations in which the time to coalescence conveys information about Z?

32.5 Solutions

Solution to exercise 32.1 (p.420). It is perhaps surprising that there is no direct relationship between the equilibration time and the time to coalescence. A simple example that proves this is the case of the uniform distribution over the integers $\mathcal{A} = \{0, 1, 2, \ldots, 20\}$. A Markov chain that converges to this distribution in exactly one iteration is the chain for which the probability of state s_{t+1} given s_t is the uniform distribution, for all s_t. Such a chain can be coupled to a random number generator in two ways: (a) we could draw a random integer $u \in \mathcal{A}$, and set s_{t+1} equal to u regardless of s_t; or (b) we could draw a random integer $u \in \mathcal{A}$, and set s_{t+1} equal to $(s_t + u) \bmod 21$. Method (b) would produce a cohort of trajectories locked together, similar to the trajectories in figure 32.1, except that no coalescence ever occurs. Thus, while the equilibration times of methods (a) and (b) are both one, the coalescence times are respectively one and infinity.

It seems plausible on the other hand that coalescence time provides some sort of upper bound on equilibration time.

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,

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

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

where $\langle E(\mathbf{x}; \mathbf{J}) \rangle_Q$ 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 only attains this value 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 $\beta \tilde{F}$ 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

$$
\begin{aligned}
\langle E(\mathbf{x}; \mathbf{J}) \rangle_Q &= \sum_{\mathbf{x}} Q(\mathbf{x}; \mathbf{a}) \left[-\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n \right] \quad (33.21) \\
&= -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n. \quad (33.22)
\end{aligned}
$$

So the variational free energy is given by

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

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

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

So we obtain

$$
\begin{aligned}
\frac{\partial}{\partial a_m} \beta \tilde{F}(\mathbf{a}) &= \beta \left[-\sum_n J_{mn} \bar{x}_n - h_m \right] \left(2 \frac{\partial q_m}{\partial a_m} \right) - \ln \left(\frac{1-q_m}{q_m} \right) \left(\frac{\partial q_m}{\partial a_m} \right) \\
&= 2 \left(\frac{\partial q_m}{\partial a_m} \right) \left[-\beta \left(\sum_n J_{mn} \bar{x}_n + h_m \right) + a_m \right]. \tag{33.25}
\end{aligned}
$$

This derivative is equal to zero when

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

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

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

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.

2. The theory is readily generalized to other approximating distributions. We can imagine introducing a more complex approximation $Q(\mathbf{x}; \boldsymbol{\theta})$ that

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. Horizontal axis: temperature $T = 1/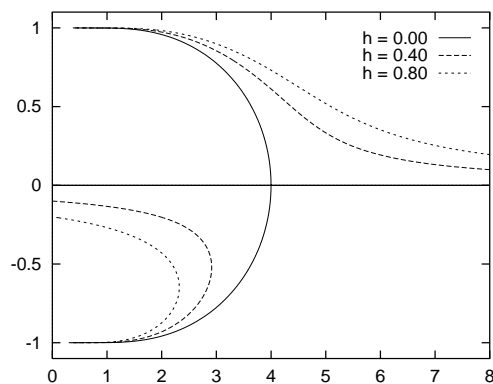\beta$. Vertical axis: magnetization $\bar{x}$. The critical temperature found by mean field theory is $T_c^{\mathrm{mft}} = 4$.

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

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

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^{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^{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

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

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^{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^{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 $\mathcal{H}$, $P(\mathbf{w}\,|\,D,\mathcal{H})$.

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

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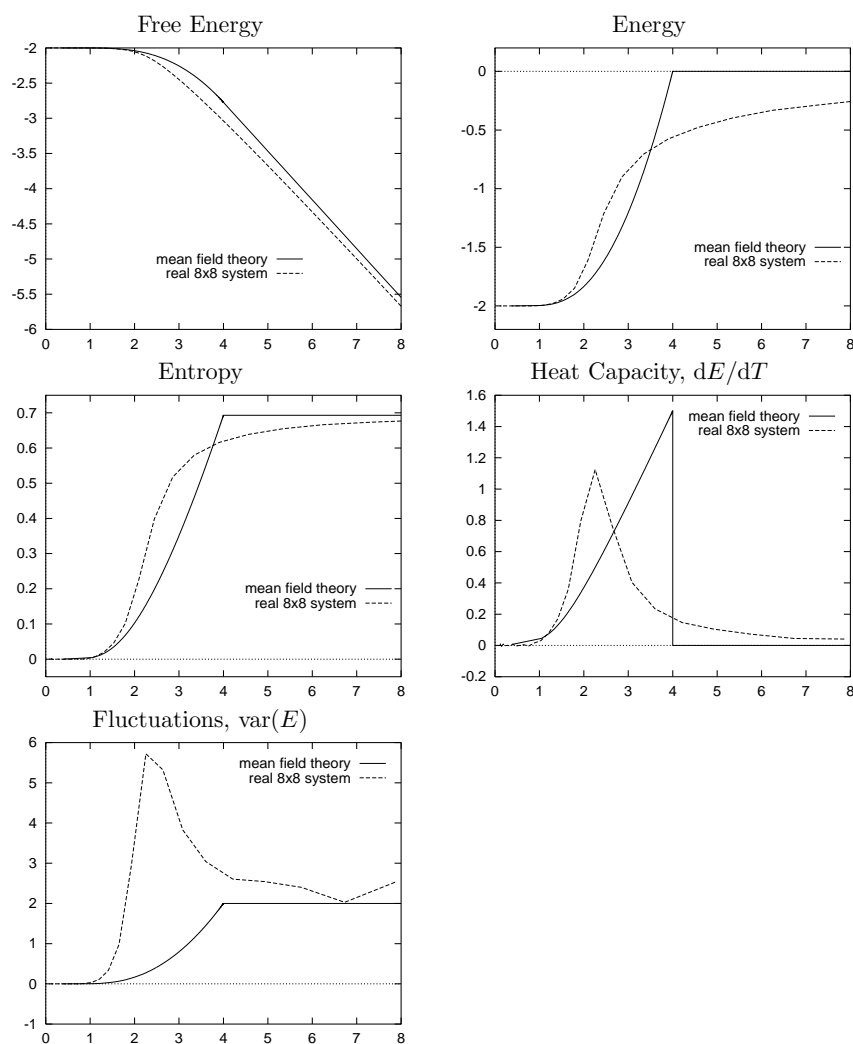
Figure 33.3. Comparison of approximating distribution's properties with those of a real 8×8 fragment. Notice that the variational free energy of the approximating distribution is indeed an upper bound on the free energy of the real system. All quantities are shown 'per spin'.

the whole distribution. We may also be interested in its normalizing constant $P(D \mid \mathcal{H})$ if we wish to do model comparison. The probability distribution $P(\mathbf{w} \mid D, \mathcal{H})$ 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 $P(\mathbf{w} \mid D, \mathcal{H})$ well.

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

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

The denominator $P(D \mid \mathbf{w}, \mathcal{H}) P(\mathbf{w} \mid \mathcal{H})$ is, within a multiplicative constant, equal to the posterior probability $P(\mathbf{w} \mid D, \mathcal{H}) = P(D \mid \mathbf{w}, \mathcal{H}) P(\mathbf{w} \mid \mathcal{H}) / P(D \mid \mathcal{H})$. So the variational free energy $\tilde{F}(\boldsymbol{\theta})$ can be viewed as the sum of $-\ln P(D \mid \mathcal{H})$ and the relative entropy between $Q(\mathbf{w}; \boldsymbol{\theta})$ and $P(\mathbf{w} \mid D, \mathcal{H})$. $\tilde{F}(\boldsymbol{\theta})$ is bounded below by $-\ln P(D \mid \mathcal{H})$ and only attains this value for $Q(\mathbf{w}; \boldsymbol{\theta}) = P(\mathbf{w} \mid D, \mathcal{H})$. 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,

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

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

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,

$$\tilde{F}(Q) = \int d\mu \, d\sigma \, Q_\mu(\mu) Q_\sigma(\sigma) \ln \frac{Q_\mu(\mu) Q_\sigma(\sigma)}{P(D \mid \mu, \sigma) P(\mu, \sigma)}. \qquad (33.37)$$

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 $Q_\mu(\mu)$.
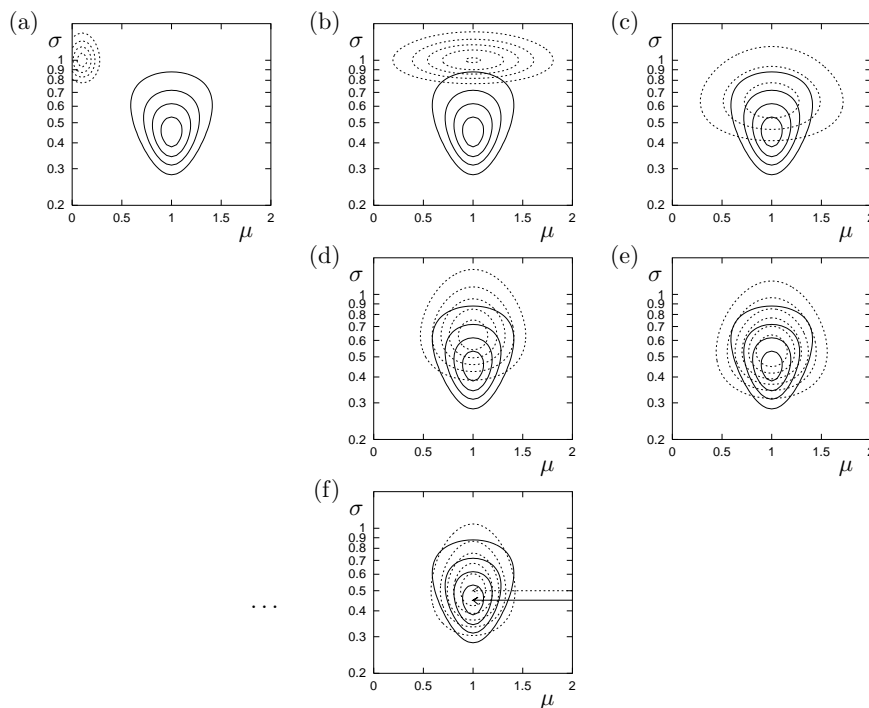
(a)

(b)

(c)

Figure 33.4. Optimization of an approximating distribution. The posterior distribution $P(\mu, \sigma \,|\, \{x_n\})$, 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).

(d)

(e)

(f)

...

Optimization of $Q_\mu(\mu)$

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

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

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

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

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

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 $Q_\mu^{\text{opt}}(\mu)$ that minimizes $\tilde{F}$ for fixed Q_σ:

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

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

Optimization of $Q_\sigma(\sigma)$

As a functional of $Q_\sigma(\sigma)$, $\tilde{F}$ is (neglecting additive constants):

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

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

where the integral over μ is performed assuming $Q_\mu(\mu) = Q_\mu^{\text{opt}}(\mu)$. 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:

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

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 and Winn, 2000; Bishop *et al.*, 2002; Bishop and Winn, 2003). 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 $\mathbf{x}$, $P(\mathbf{x}\,|\,D)$. 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 $P(D \,|\, \mathbf{x})$, or of the posterior distribution. One interpretation of this standard practice is that the full description of our knowledge about $\mathbf{x}$, $P(\mathbf{x} \,|\, D)$, 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

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

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

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

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:

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

$\tilde{F}$ is bounded below by minus the evidence, $\ln P(\{\mathbf{x}^{(n)}\}_{n=1}^N \,|\, \mathcal{H})$. 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

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

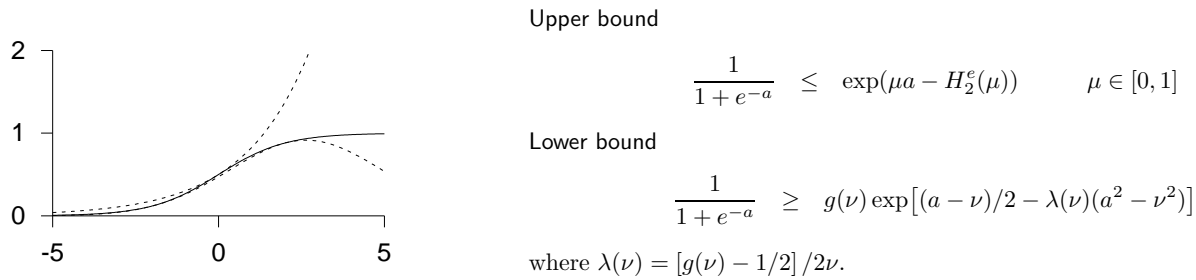where $\lambda(\nu) = [g(\nu) - 1/2]/2\nu$.

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 $\int d^D\boldsymbol{\theta}\, Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, 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.[4] 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.* (2000c); Welling and Teh (2001); Yuille (2001); Yedidia *et al.* (2000b); Yedidia *et al.* (2000a).

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

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

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.

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

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

33.10 Solutions

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

$$\int dx\, \text{Normal}(x; 0, \sigma_Q) \ln \frac{\text{Normal}(x; 0, \sigma_Q)}{\text{Normal}(x; 0, \sigma_P)}$$

$$= \int dx\, \text{Normal}(x; 0, \sigma_Q) \left[\ln \frac{\sigma_P}{\sigma_Q} - \frac{1}{2}x^2 \left(\frac{1}{\sigma_Q^2} - \frac{1}{\sigma_P^2} \right) \right] \quad (33.55)$$

$$= \frac{1}{2} \left(\ln \frac{\sigma_P^2}{\sigma_Q^2} - 1 + \frac{\sigma_Q^2}{\sigma_P^2} \right). \quad (33.56)$$

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

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

differentiating,

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

which is zero when

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

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.*