CHAPTER 10

Variational Methods

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

10.1 Variational free energy minimization

One well known method for approximating a complex distribution in a physical system is ‘mean field theory’. Mean field theory is in fact 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 (equation (1.24), exercise 20), which we repeat here.

The relative entropy or Kullback-Leibler divergence between two probability distributions $Q(x)$ and $P(x)$ that are defined over the same alphabet $\mathcal{X}$ is

$$D_{\text{KL}}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}. \quad (10.1)$$

The relative entropy satisfies $D_{\text{KL}}(Q||P) \geq 0$ (Gibbs’ inequality) with equality only if $Q=P$. Note that in general $D_{\text{KL}}(Q||P) \neq D_{\text{KL}}(P||Q)$.

10.1.1 Probability distributions in statistical physics

In statistical physics one often encounters probability distributions of the form

$$P(x|\beta,J) = \frac{1}{Z(\beta,J)} \exp[-\beta E(x;J)], \quad (10.2)$$

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

$$E(x;J) = -\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n. \quad (10.3)$$

The partition function is

$$Z(\beta,J) = \sum_x \exp[-\beta E(x;J)] \quad (10.4)$$

The probability distribution of equation (10.2) is complex. Not unbearably complex; after all, we can, for any $x$, evaluate $E(x;J)$ in a time polynomial in the number of spins. But evaluating
the normalizing constant $Z(\beta, \mathbf{J})$ is difficult, 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 the partition function 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 $Q(\mathbf{x}; \theta)$ that is parameterized by adjustable parameters $\theta$. We adjust these parameters so as to get $Q$ to best approximate $P$, in some sense. A by-product of this approximation is a bound on $Z(\beta, \mathbf{J})$.

### 10.2 The variational free energy

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

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

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

$$
\begin{align*}
\beta \tilde{F}(\theta) &= \beta \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) E(\mathbf{x}; \mathbf{J}) - \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) \log \frac{1}{Q(\mathbf{x}; \theta)} \\
&= \beta \langle E(\mathbf{x}; \mathbf{J}) \rangle_Q - S_Q,
\end{align*}
$$

where $\langle E(\mathbf{x}; \mathbf{J}) \rangle_Q$ is the average of the energy function under the distribution $Q(\mathbf{x}; \theta)$, and $S_Q$ is the entropy of the distribution $Q(\mathbf{x}; \theta)$ (we omit the factor of $k_B$ in the definition of $S$ so that it is identical to the definition of the entropy $H$ in part 1).

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

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

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

where $F$ is the true free energy, defined by

$$
\beta F \equiv -\log Z(\beta, \mathbf{J}),
$$

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

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

### 10.2.1 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

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

1011011110010 © David J.C. MacKay: April 10, 1997

121
the energy
\[ \langle E \rangle_p = \frac{1}{Z} \sum_x E(x; J) \exp(-\beta E(x; J)) \] (10.12)
and the entropy
\[ S = \sum_x P(x|\beta, J) \log \frac{1}{P(x|\beta, J)} \] (10.13)
are all presumed to be impossible to evaluate. So why should we suppose that this objective function \( \beta \tilde{F}(\theta) \), which is also defined in terms of a sum over all \( x \) (equation (10.5)), 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.

### 10.3 Variational free energy minimization for Ising models

An example of a tractable variational free energy is given by the spin system whose energy function was given in equation (10.3), which we can approximate with a separable approximating distribution
\[ Q(x; a) = \frac{1}{Z_Q} \exp \left( \sum_n a_n x_n \right). \] (10.14)
The variational parameters \( \theta \) are here termed \( a \). To evaluate the variational free energy we need the entropy of this distribution,
\[ S_Q = \sum_x Q(x; \theta) \log \frac{1}{Q(x; \theta)} \] (10.15)
and the mean of the energy
\[ \langle E(x; J) \rangle_Q = \sum_x Q(x; \theta) E(x; J). \] (10.16)
The entropy of the separable approximating distribution is simply the sum of the entropies of the individual spins (exercise 15),
\[ S_Q = \sum_n H_2^{(c)}(q_n) \] (10.17)
where \( q_n \) is the probability that spin \( n \) is +1,
\[ q_n = \frac{e^{a_n}}{e^{a_n} + e^{-a_n}} = \frac{1}{1 + \exp(-2a_n)} \] (10.18)
and
\[ H_2^{(c)}(q) = q \log \frac{1}{q} + (1 - q) \log \frac{1}{1 - q}, \] (10.19)
all logs being natural logarithms. The mean energy 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 \( \bar{x}_n \), with
\[ \bar{x}_n = \frac{e^{a_n} - e^{-a_n}}{e^{a_n} + e^{-a_n}} = \tanh(a_n) = 2q_n - 1, \] (10.20)
we obtain
\[
\langle E(\mathbf{x}; \mathbf{J}) \rangle_Q = \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) \left[ -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n \right] \tag{10.21}
\]
\[
= -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n. \tag{10.22}
\]

So the variational free energy is given by
\[
\beta \bar{F}(\mathbf{a}) = \beta \langle E(\mathbf{x}; \mathbf{J}) \rangle_Q - S_Q = \beta \left( -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_n \right) - \sum_n H_2(\epsilon)(q_n). \tag{10.23}
\]

We now consider minimizing this function with respect to the variational parameters \( \mathbf{a} \). Noting that when \( q = 1/(1 + e^{-2a}) \), the derivative of the entropy is
\[
\frac{\partial}{\partial q} H_2(q) = \log \frac{1 - q}{q} = -2a,
\]
we obtain
\[
\frac{\partial}{\partial a_m} \beta \bar{F}(\mathbf{a}) = \beta \left[ -\sum_n J_{mn} \bar{x}_n - h_m \right] \left( 2 \frac{\partial q_m}{\partial a_m} \right) - \log \left( \frac{1 - q_m}{q_m} \right) \left( \frac{\partial q_m}{\partial a_m} \right) \tag{10.25}
\]
\[
= 2 \left( \frac{\partial q_m}{\partial a_m} \right) \left[ -\beta \left( \sum_n J_{mn} \bar{x}_n + h_m \right) + a_m \right]. \tag{10.26}
\]

This derivative is set to zero when
\[
a_m = \beta \left( \sum_n J_{mn} \bar{x}_n + h_m \right). \tag{10.27}
\]

This equation and the definition
\[
\bar{x}_n = \tanh(a_n) \tag{10.28}
\]
define the solution to our variational free energy minimization. The variational free energy \( \bar{F}(\mathbf{a}) \) may be a multimodal function in which case each stationary point (maximum, minimum or saddle) will satisfy equations (10.27) and (10.28). 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 (10.27) (one at a time). This asynchronous updating of the parameters is guaranteed to decrease \( \beta \bar{F}(\mathbf{a}) \).

The equations (10.27) and (10.28) 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 an imaginary field applied to spin \( n \). Equation (10.28) describes the mean response of spin \( n \), and equation (10.27) 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 theory associates an objective function \( \beta \bar{F} \) 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 \( Q(\mathbf{x}; \theta) \) that might, for example, capture correlations among the spins instead of modelling them as independent. One could then evaluate the variational free energy and optimize the parameters \( \theta \) of this more complex approximation. The more degrees of freedom the approximating distribution has, the tighter the bound on the free energy becomes. Typically as the complexity of an approximation is increased the evaluation of either the mean energy or the entropy becomes more challenging.

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

\[
P(\mathbf{w}|D, \mathcal{H}) = \frac{P(D|\mathbf{w}, \mathcal{H})P(\mathbf{w}|\mathcal{H})}{P(D|\mathcal{H})}. \tag{10.29}
\]

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 the whole distribution; and its normalizing constant \( P(D|\mathcal{H}) \) may also be of interest if we wish to do model comparison. The probability distribution \( P(\mathbf{w}|D, \mathcal{H}) \) is often a complex distribution, so we may be interested in working in terms of an approximating ensemble \( Q(\mathbf{w}; \theta) \), that is, a probability distribution over the parameters, and optimize the ensemble (by varying its own parameters \( \theta \)) 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

\[
\tilde{F}(\theta) = \int d^k \mathbf{w} Q(\mathbf{w}; \theta) \log \frac{Q(\mathbf{w}; \theta)}{P(D|\mathbf{w}, \mathcal{H})P(\mathbf{w}|\mathcal{H})}. \tag{10.30}
\]

The numerator \( P(D|\mathbf{w}, \mathcal{H})P(\mathbf{w}|\mathcal{H}) \) is, within a multiplicative constant, equal to 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 \( \tilde{F}(\theta) \) can be viewed as the sum of \( - \log P(D|\mathcal{H}) \) and the relative entropy between \( Q(\mathbf{w}; \theta) \) and \( P(\mathbf{w}|D, \mathcal{H}) \). \( \tilde{F}(\theta) \) is bounded below by \( - \log P(D|\mathcal{H}) \) and only attains this value for \( Q(\mathbf{w}; \theta) = P(\mathbf{w}|D, \mathcal{H}) \).

For certain models and certain approximating distributions, this free energy, and its derivatives with respect to the ensemble'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 1993; MacKay 1995). We have given this idea the name of ‘ensemble learning’ in contrast with traditional learning processes in which a single parameter vector is optimized. Let us examine how ensemble learning works in the simple case of a Gaussian distribution.

10.5 Approximating the posterior distribution of \( \mu \) and \( \sigma \)

We will fit an approximating ensemble \( Q(\mu, \sigma) \) to the posterior distribution that we studied in chapter 8,

\[
P(\mu, \sigma|\{x_n\}_{n=1}^N) = \frac{P(\{x_n\}_{n=1}^N|\mu, \sigma)P(\mu, \sigma)}{P(\{x_n\}_{n=1}^N)} \tag{10.31}
\]
\[
\frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left( -\frac{(\mu - \mu_0)^2}{2\sigma^2} \right) \frac{1}{\sigma^{N/2}} \cdot 
\]

Let us 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,

\[
\tilde{F}(Q) = \int d\mu \, d\sigma \, Q_\mu(\mu)Q_\sigma(\sigma) \log \frac{Q_\mu(\mu)Q_\sigma(\sigma)}{P(\mu, \sigma)P(\mu, \sigma)}. \tag{10.33}
\]

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 \(Q_\mu(\mu)\).

### 10.5.1 Optimization of \(Q_\mu(\mu)\)

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

\[
\tilde{F} = -\int d\mu \, Q_\mu(\mu) \left[ \int d\sigma \, Q_\sigma(\sigma) \log P(D|\mu, \sigma) + \log\left[ P(\mu)/Q(\mu) \right] \right] + \text{const.} \tag{10.34}
\]

\[
= \int d\mu \, Q_\mu(\mu) \left[ \int d\sigma \, Q_\sigma(\sigma) \frac{1}{2}N\beta(\mu - \bar{x})^2 + \log Q(\mu) \right] + \text{const}'. \tag{10.35}
\]

The dependence on \(Q_\sigma\) thus collapses down to a simple dependence on the mean

\[
\bar{\beta} \equiv \int d\sigma \, Q_\sigma(\sigma)/\sigma^2. \tag{10.36}
\]

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

\[
Q_\mu^{\text{opt}}(\mu) = P(\mu|D, \bar{\beta}, \mathcal{H}) = \text{Normal}(\mu; \bar{x}, \sigma_{\mu|D}^2). \tag{10.37}
\]

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

### 10.5.2 Optimization of \(Q_\sigma(\sigma)\)

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

\[
\tilde{F}(Q) = -\int d\sigma \, Q_\sigma(\sigma) \left[ \int d\mu \, Q_\mu(\mu) \log P(D|\mu, \sigma) + \log\left[ P(\sigma)/Q_\sigma(\sigma) \right] \right] + \text{const.} \tag{10.38}
\]

\[
= \int d\sigma \, Q_\sigma(\sigma) \left[ (N\sigma_{\mu|D}^2 + S)\beta/2 - \left( \frac{N}{2} - 1 \right) \log \beta + \log Q_\sigma(\sigma) \right] + \text{const}'. \tag{10.39}
\]

where the integral over \(\mu\) is performed assuming \(Q_\mu(\mu) = Q_\mu^{\text{opt}}(\mu)\). Here, the \(\beta\)-dependent expression in the brackets can be recognized as the log of a gamma distribution over \(\beta\) (see equation (8.27)), giving as the distribution that minimizes \(\tilde{F}\) for fixed \(Q_\mu\): \(Q_\sigma^{\text{opt}}(\beta) = \Gamma(\beta; b', c')\), with \(1/b' = \frac{1}{2}(N\sigma_{\mu|D}^2 + S)\) and \(c' = N/2\).

1011011111010  © David J.C. MacKay: April 10, 1997  125
10.5.3 Joint optimum $Q_\mu(\mu)Q_\sigma(\sigma)$

We now have an implicit equation for the optimal approximating ensemble, with $\sigma^2_{D|\mathcal{O}} = 1/(N\bar{\beta})$, and $\bar{\beta} = b'e'$. The solution is:

$$1/\bar{\beta} = S/(N - 1).$$  (10.40)

This is similar to the true posterior distribution of $\sigma$, which is a gamma distribution with $c' = \frac{N-1}{2}$ and $1/b' = 1/b + S/2$. This true posterior also has a mean value of $\beta$ 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 $\beta$ is a gamma distribution with mean $\bar{\beta}$ corresponding to a variance of $\sigma^2 = S/(N - 1) = \sigma^2_{D|\mathcal{O}}$. And the approximate posterior distribution over $\mu$ is a Gaussian with mean $\bar{x}$ and standard deviation $\sigma_{\mu|\mathcal{O}}/\sqrt{N}$.

10.6 Exercises

**Exercise 109:** 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 $P(x)$ with axes aligned with the directions $e^{(1)} = (1, 1)$ and $e^{(2)} = (1, -1)$. Let the variances in these two directions be $\sigma_1^2$ and $\sigma_2^2$. What is the optimal variance if this distribution is approximated by a spherical Gaussian with variance $\sigma_D^2$, optimized by variational free energy minimization? If we instead optimized the objective function

$$G = \int dx\, P(x) \log \frac{P(x)}{Q(x; \sigma^2)},$$  (10.41)

what would be the optimal value of $\sigma^2$? Sketch a contour of the true distribution $P(x)$ 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 $P(x)$ are intractable.]

**Exercise 110:** Compare and contrast what makes an approximation $Q$ a good variational approximation for a distribution $P$ (by the variational free energy criterion), and what would make it a good sampler for importance sampling.

References


