Ensemble Learning in Bayesian Neural Networks

David Barber
Institute for Adaptive and Neural Computation,
Division of Informatics, University of Edinburgh,
5 Forrest Hill, Edinburgh, EH1 2QL, U.K.

Christopher M. Bishop
Microsoft Research,
7 J J Thomson Avenue, Cambridge, CB23 0FB, U.K.
http://research.microsoft.com/∼cmbishop

Published in *Neural Networks and Machine Learning* (1998), C. M. Bishop (Ed.),
Springer, 215–237.

Abstract

Bayesian treatments of learning in neural networks are typically based either on a local Gaussian approximation to a mode of the posterior weight distribution, or on Markov chain Monte Carlo simulations. A third approach, called *ensemble learning*, was introduced by Hinton and van Camp (1993). It aims to approximate the posterior distribution by minimizing the Kullback-Leibler divergence between the true posterior and a parametric approximating distribution. The original derivation of a deterministic algorithm relied on the use of a Gaussian approximating distribution with a *diagonal* covariance matrix and hence was unable to capture the posterior correlations between parameters. In this chapter we show how the ensemble learning approach can be extended to full-covariance Gaussian distributions while remaining computationally tractable. We also extend the framework to deal with hyperparameters, leading to a simple re-estimation procedure. One of the benefits of our approach is that it yields a strict lower bound on the marginal likelihood, in contrast to other approximate procedures.

1 Introduction

Bayesian techniques have been successfully applied to neural networks in the context of both regression and classification problems (MacKay 1992; Neal 1996). In contrast to the maximum likelihood approach, which finds a single estimate for the regression parameters, the Bayesian approach yields a posterior distribution of network parameters, $P(\mathbf{w}|D)$, conditional on the training data D, and predictions are expressed in terms of expectations with respect to this posterior distribution (Bishop 1995). However, the corresponding integrals over weight space are analytically intractable.

One well-established procedure for approximating these integrals, known as Laplace's method, is to model the posterior distribution by a Gaussian, centred at a mode of $p(\mathbf{w}|D)$, in which the covariance of the Gaussian is determined by the local curvature of the posterior distribution (MacKay 1992). With the further assumption that the variance of the distribution is small,

and hence that the network function can be linearized in the neighbourhood of the mode, the required integrations can be performed analytically.

More recent approaches have used Markov chain Monte Carlo simulations to generate samples from the posterior (Neal 1996). However, such techniques can be computationally expensive, and they also suffer from the difficulty of assessing convergence.

A third approach, called ensemble learning, was introduced by Hinton and van Camp (1993) and again involves finding a simple, analytically tractable, approximation to the true posterior distribution. Unlike Laplace's method, however, the approximating distribution is fitted globally, rather than locally, by minimizing a Kullback-Leibler divergence. Hinton and van Camp (1993) showed that, in the case of a Gaussian approximating distribution with a *diagonal* covariance, a deterministic learning algorithm could be derived. This approach removes the constraint that the mode of the approximating Gaussian must coincide with a mode of the posterior. However, the restriction to a diagonal covariance prevents the model from capturing the (often very strong) posterior correlations between the parameters. MacKay (1995) suggested a modification to the algorithm by including a linear preprocessing of the inputs to achieve a somewhat richer class of approximating distributions, although this was not implemented. In this chapter we show that the ensemble learning approach can be extended to allow a Gaussian approximating distribution with a *general* covariance matrix, while still leading to a tractable algorithm (Barber and Bishop 1998). Our focus is on the essential principles of the approach, with the mathematical details relegated to the Appendix.

1.1 Bayesian Neural Networks

Consider a two-layer feed-forward network having H hidden units and a single output whose value is given by

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{H} v_i \sigma(\mathbf{u}_i^{\mathrm{T}} \mathbf{x}) \tag{1}$$

where $\mathbf{w} \equiv \{\mathbf{u}_i, v_i\}$ is a k-dimensional vector representing all of the adaptive parameters in the model, $\mathbf{x}$ is the input vector, $\{\mathbf{u}_i\}, i = 1, \ldots, H$ are the input-to-hidden weights, and $\{v_i\}, i = 1, \ldots, H$ are the hidden-to-output weights. The extension to multiple outputs is straightforward. For reasons of analytical tractability, we choose the sigmoidal hidden-unit activation function $\sigma(a)$ to be the 'erf' (cumulative Gaussian) function

$$\sigma(a) = \sqrt{\frac{2}{\pi}} \int_0^a \exp\left(-s^2/2\right)\, ds \tag{2}$$

which (with an appropriate linear re-scaling) is quantitatively very similar to the standard logistic sigmoid. Hidden unit biases are accounted for by appending the input vector with an additional input whose value is always unity.

The data set itself consists of N pairs of input vectors and corresponding target output values $D = \{\mathbf{x}^\mu, t^\mu\}, \mu = 1, \ldots, N$. We make the standard assumption of Gaussian noise on the target values, with variance β^{-1}. The likelihood function for $\mathbf{w}$ and β is then

$$P\left(D|\mathbf{w}, \beta\right) = \frac{\exp(-\beta E_D)}{Z_D} \tag{3}$$

where $Z_D = (2\pi/\beta)^{N/2}$ is a normalizing factor, and E_D is the 'training error' defined to be

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{\mu}\left(f(\mathbf{x}^{\mu},\mathbf{w}) - t^{\mu}\right)^2. \tag{4}$$

The prior distribution over weights is chosen to be a Gaussian of the form

$$P(\mathbf{w}|\mathbf{A}) = \frac{\exp\left(-E_W(\mathbf{w})\right)}{Z_P} \tag{5}$$

where $E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}$, $\mathbf{A}$ is a matrix of hyperparameters, and $Z_P = (2\pi)^{k/2}\left|A\right|^{-1/2}$ is the normalizing factor. From Bayes' theorem, the posterior distribution over weights can then be written

$$P(\mathbf{w}|D,\beta,\mathbf{A}) = \frac{1}{Z_F}\exp\left(-\beta E_D(\mathbf{w}) - E_W(\mathbf{w})\right) \tag{6}$$

where Z_F is a normalizing constant defined by

$$Z_F = \int \exp\left(-\beta E_D(\mathbf{w}) - E_W(\mathbf{w})\right)\, d\mathbf{w}. \tag{7}$$

Note that this integration is analytically intractable, and so Z_F cannot be evaluated explicitly.

Predictions for a new input (for given β and $\mathbf{A}$) are given by integration over the posterior distribution of weights. For instance the predictive mean is given by

$$\langle f(\mathbf{x})\rangle = \int f(\mathbf{x},\mathbf{w})P(\mathbf{w}|D,\beta,\mathbf{A})\, d\mathbf{w}. \tag{8}$$

This represents an integration over a high-dimensional space, weighted by a posterior distribution $P(\mathbf{w}|D,\beta,\mathbf{A})$. Since the posterior is exponentially small except in narrow regions whose locations are unknown a-priori, the accurate evaluation of such integrals is very difficult.

So far we have treated the hyperparameters β and $\mathbf{A}$ as if they are constants. A Bayesian treatment of β and $\mathbf{A}$ is given in Section 2.2.

1.2 Laplace's Method

As the number N of data points is increased, the posterior distribution approaches a Gaussian (Walker 1969) whose variance goes to zero in the limit $N \to \infty$. This motivates Laplace's method which seeks to approximate the posterior distribution with a Gaussian. In order to calculate this Gaussian approximation, we write the posterior distribution in the form

$$P(\mathbf{w}|D,\beta,\mathbf{A}) = \exp\left(-\phi\left(\mathbf{w}\right)\right) \tag{9}$$

and expand ϕ around a mode of the distribution, $\mathbf{w}_* = \arg\min\phi\left(\mathbf{w}\right)$, so that

$$\phi\left(\mathbf{w}\right) \approx \phi\left(\mathbf{w}_*\right) + \frac{1}{2}\left(\mathbf{w}-\mathbf{w}_*\right)^{\mathrm{T}}\mathbf{H}\left(\mathbf{w}-\mathbf{w}_*\right). \tag{10}$$

Here we have defined

$$\mathbf{H} = \nabla\nabla\phi\left(\mathbf{w}\right)\big|_{\mathbf{w}_*} \tag{11}$$

which is the local Hessian matrix. This local expansion defines a Gaussian approximation to the distribution $P(\mathbf{w}|D, \beta, \mathbf{A})$ of the form

$$P(\mathbf{w}|D, \beta, \mathbf{A}) \simeq \frac{|\mathbf{H}|^{1/2}}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{w}_*)^{\mathrm{T}} \mathbf{H} (\mathbf{w} - \mathbf{w}_*)\right\}. \tag{12}$$

The expected value of $f(\mathbf{x}, \mathbf{w})$, as required in (8), can be evaluated by making a further local linearization of the function $f(\cdot, \mathbf{w})$ around the point $\mathbf{w}_*$. In a practical implementation, a standard non-linear optimization algorithm such as conjugate gradients is used to find a mode $\mathbf{w}_*$ of the log posterior distribution. The local Hessian $\mathbf{H}$ can then be evaluated efficiently using an extension of the back-propagation procedure (Bishop 1992; Pearlmutter 1994) or by using one of several approximation schemes (Bishop 1995).

So far in this discussion of Laplace's method we have assumed that the hyper-parameters β and $\mathbf{A}$ are fixed. In a fully Bayesian treatment we would define prior distributions of the hyper-parameters and then integrate them out. Since exact integration is analytically intractable, MacKay (1992) uses an approximation called type-II maximum likelihood (Berger 1985) which involves estimating specific values for the hyper-parameters by maximizing the marginal likelihood $P(D|\beta, \mathbf{A})$ with respect to β and $\mathbf{A}$. The marginal likelihood is given by

$$P(D|\beta, \mathbf{A}) = \int P(D|\mathbf{w}, \beta) P(\mathbf{w}|\mathbf{A})\, d\mathbf{w}. \tag{13}$$

Again, this integration is analytically intractable. However, using the local Laplace approximation the marginal likelihood in (13) can be approximated analytically. The conditions for stationarity with respect to $\mathbf{A}$ and β then lead to simple re-estimation formulae for the hyper-parameters expressed in terms of the eigenvalue/eigenvector decomposition of the Hessian matrix. This treatment of hyper-parameters is called the *evidence* framework by MacKay (1992) and involves alternating the optimization of $\mathbf{w}$ for fixed hyper-parameters with re-estimation of the hyper-parameters by re-evaluating the Hessian matrix for the new value of $\mathbf{w}$.

The various approximations involved in this approach all improve as the number of data points $N \to \infty$. However, for a finite data set it can be difficult to asses the accuracy of the method. One obvious limitation is that it only takes account of the behaviour of the posterior distribution at the mode. A review of this framework is given by MacKay (1995).

1.3 Markov Chain Monte Carlo Methods

In recent years, Monte Carlo methods have been applied extensively in Bayesian statistics. The central idea is to replace integrals weighted by the posterior distribution, such as that in (8), by finite sums, so that

$$\int P(\mathbf{w}|D, \beta, \mathbf{A}) g(\mathbf{w})\, d\mathbf{w} \approx \frac{1}{m} \sum_{i=1}^{m} g(\mathbf{w}_i) \tag{14}$$

where the vectors $\mathbf{w}_i$ are samples from the posterior distribution $P(\mathbf{w}|D, \beta, \mathbf{A})$, and $g(\mathbf{w})$ is some function. In principle, this procedure is exact in the limit $m \to \infty$. The great difficulty, however, is in finding a representative set of samples $\{\mathbf{w}_i\}$. One of the most successful approaches in the context of neural networks is that of *hybrid Monte Carlo* (Duane, Kennedy, Pendleton, and Roweth 1987; Neal 1996).

2 Ensemble Learning

We now introduce the technique of ensemble learning for Bayesian neural networks. This is a special case of the general framework of variational methods for approximate inference and learning in probabilistic models, which are reviewed by Jordan, Gharamani, Jaakkola, and Saul (1998) and Bishop (1998b).

Consider the logarithm of the marginal likelihood, given by (13). We introduce a distribution $Q(\mathbf{w})$ which is intended to provide an approximation to the true posterior distribution. Then it is easily verified that

$$\begin{aligned} \ln P(D|\beta, \mathbf{A}) &= \ln \int P(D|\mathbf{w}, \beta) P(\mathbf{w}|\mathbf{A})\, d\mathbf{w} \\ &= \ln \int \frac{P(D|\mathbf{w}, \beta) P(\mathbf{w}|\mathbf{A})}{Q(\mathbf{w})} Q(\mathbf{w})\, d\mathbf{w} \\ &\geq \int \ln \left\{ \frac{P(D|\mathbf{w}, \beta) P(\mathbf{w}|\mathbf{A})}{Q(\mathbf{w})} \right\} Q(\mathbf{w})\, d\mathbf{w} \\ &= \mathcal{F}[Q]\,. \end{aligned} \tag{15}$$

where we have made use of Jensen's inequality together with the convexity of the function $\ln(\cdot)$. We have therefore obtained a rigorous lower bound $\mathcal{F}[Q]$ on the logarithm of the marginal likelihood. The difference between $\ln P(D|\beta, \mathbf{A})$ and $\mathcal{F}[Q]$ is easily seen to be the Kullback-Leibler divergence between the distribution $Q(\mathbf{w})$ and the true posterior

$$\mathrm{KL}(Q\|P) = \int Q(\mathbf{w}) \ln \left\{ \frac{Q(\mathbf{w})}{P(\mathbf{w}|D, \beta, \mathbf{A})} \right\} d\mathbf{w}. \tag{16}$$

It is a well known result that $\mathrm{KL}(Q\|P) \geq 0$, with equality if and only if $P(\mathbf{w}|D, \beta, \mathbf{A}) = Q(\mathbf{w})\ \forall \mathbf{w}$. The relationship between the various quantities is illustrated in Figure 1.

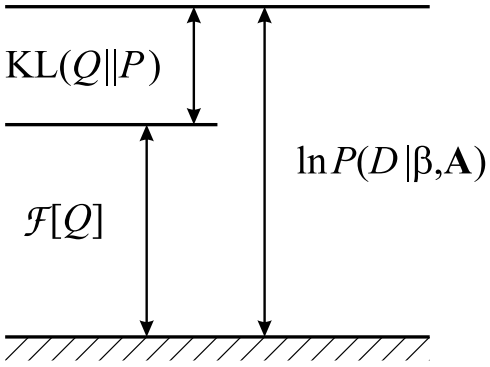


Figure 1: The quantity $\mathcal{F}[Q]$ provides a rigorous lower bound on the log marginal likelihood $\ln P(D|\beta, \mathbf{A})$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between the approximating distribution $Q(\mathbf{w})$ and the true posterior $P(\mathbf{w}|D, \beta, \mathbf{A})$.

We have already noted that the marginal likelihood $P(D|\beta, \mathbf{A})$ involves an intractable integration over $\mathbf{w}$. Our goal in the ensemble learning approach is to choose a form for the $Q(\mathbf{w})$

distribution such that the lower bound $\mathcal{F}[Q]$ can be evaluated efficiently. In particular, if we can find a parametric family of distributions, then we can adapt the parameters to find the tightest lower bound within this family. Maximizing the lower bound $\mathcal{F}[Q]$ with respect to the parameters of Q is equivalent to minimizing the Kullback-Leibler divergence (16). The richer the family of Q distributions considered, the better the resulting bound will be. In the extreme case of a completely general class of distributions we obtain the tightest possible bound in which Q is given by the true posterior distribution, the Kullback-Leibler divergence vanishes, and the lower bound equals the true log marginal likelihood. Of course in this case there is no benefit to the variational approach since the lower bound will necessarily be intractable.

The key to a successful application of variational methods therefore lies in the choice of the Q distribution, which should be as close to the true posterior distribution as possible while leading to an analytically tractable integration. Note that we are at liberty to use as rich a family of approximating distributions as we please (there is no 'over-fitting') and we are limited only by computational resources and the requirement of analytical tractability in the evaluation of the lower bound $\mathcal{F}[Q]$.

There is an interesting relation between the variational framework and the EM (expectation-maximization) algorithm, as pointed out by Neal and Hinton (1998). If we use the type II maximum likelihood procedure to set the values of the hyper-parameters, as discussed in Section 1.2, then we can regard the weight vector $\mathbf{w}$ to be a missing (hidden) variable. The standard EM algorithm for solving the maximum likelihood problem (Dempster, Laird, and Rubin 1977) alternates between an E-step in which the posterior distribution of the hidden variables is used to evaluate the expectation of the complete-data log likelihood, and an M-step in which the expected complete-data log likelihood is maximized with respect to the model parameters. This corresponds to alternate maximization of $\mathcal{F}$ with respect to a free-form Q distribution (E-step) for fixed hyper-parameters, and with respect to the hyper-parameters (M-step) for fixed Q. It is well known that the EM algorithm can be generalized by increasing rather than maximizing the expected complete-data log likelihood in the M-step to give the generalized EM (GEM) algorithm (Dempster, Laird, and Rubin 1977) while still retaining a guarantee that the true log likelihood will be increased at each step (unless already at a maximum). Neal and Hinton (1998) pointed out that the EM algorithm can also be generalized to allow for partial E-steps by maximizing $\mathcal{F}$ with respect to a constrained family of distributions Q. Although this does not guarantee to increase the true log likelihood, it is guaranteed to increase a lower bound on the true log likelihood, and the existence of the function $\mathcal{F}$ ensures that the overall algorithm will be stable.

2.1 Gaussian Variational Distributions

We turn now to the problem of making a suitable choice for the family of Q distributions in the context of variational learning for neural networks. Hinton and van Camp (1993) were the first to consider this problem. By restricting attention to a Gaussian Q distribution having a *diagonal* covariance matrix they were able to obtain a tractable algorithm which involved the use of pre-computed two-dimensional look-up tables.

MacKay (1995) further noted that a somewhat more general class of Gaussian approximating distributions can be considered by allowing for linear transformations of the input variables. However, even with this generalization, this approach is incapable of capturing the typically strong correlations amongst arbitrary subsets of the network parameters.

In this chapter we show that such restrictions are unnecessary, and that by careful analytical treatment it is possible to use an arbitrary Gaussian distribution for Q while still obtaining a tractable algorithm. Here we give an overview of the key ideas, and defer a more detailed discussion of the analysis to the Appendix.

We consider a Q distribution given by a Gaussian with mean $\overline{\mathbf{w}}$ and covariance $\mathbf{C}$. From the definition of $\mathcal{F}$ in (15), and making use of (3) and (5), we obtain

$$\mathcal{F}[Q] = -\int Q(\mathbf{w}) \ln Q(\mathbf{w})\, d\mathbf{w} - \int Q(\mathbf{w}) \left\{E_W + E_D\right\}\, d\mathbf{w} - \ln Z_P - \ln Z_D. \tag{17}$$

The first term in (17) is the entropy of a Gaussian distribution, and is easily evaluated to give

$$-\int Q(\mathbf{w}) \ln Q(\mathbf{w})\, d\mathbf{w} = \frac{1}{2} \ln |\mathbf{C}| + \frac{k}{2} \left(1 + \ln 2\pi\right). \tag{18}$$

The prior term $E_W(\mathbf{w})$ is quadratic in $\mathbf{w}$, and integrates to give

$$\int Q(\mathbf{w}) E_W(\mathbf{w})\, d\mathbf{w} = \mathrm{Tr}\left(\mathbf{CA}\right) + \frac{1}{2} \overline{\mathbf{w}}^{\mathrm{T}} \mathbf{A} \overline{\mathbf{w}}. \tag{19}$$

This leaves the data dependent term in (17) which we write as

$$\int Q(\mathbf{w}) E_D(\mathbf{w})\, d\mathbf{w} = \frac{1}{2} \sum_{\mu=1}^{N} l(\mathbf{x}^\mu, t^\mu) \tag{20}$$

where

$$l(\mathbf{x}, t) = \int Q(\mathbf{w}) f(\mathbf{x}, \mathbf{w})^2\, d\mathbf{w} - 2t \int Q(\mathbf{w}) f(\mathbf{x}, \mathbf{w})\, d\mathbf{w} + t^2. \tag{21}$$

For clarity, we concentrate only on the first term in (21), as the calculation of the term linear in $f(\mathbf{x}, \mathbf{w})$ is similar, though simpler. Writing the Gaussian integral over Q as an average, $\langle\ \rangle$, the first term of (21) becomes

$$\left\langle f(\mathbf{x}, \mathbf{w})^2 \right\rangle = \sum_{i,j=1}^{H} \left\langle v_i v_j \sigma(\mathbf{u}_i{}^{\mathrm{T}} \mathbf{x}) \sigma(\mathbf{u}_j{}^{\mathrm{T}} \mathbf{x}) \right\rangle. \tag{22}$$

To simplify the notation, we denote the set of input-to-hidden weights $\{\mathbf{u}_i\}_{i=1}^H$ by $\mathbf{u}$ and the set of hidden-to-output weights, $\{v_i\}_{i=1}^H$ by $\mathbf{v}$. Similarly, we partition the covariance matrix $\mathbf{C}$ into blocks, $\mathbf{C}_{uu}$, $\mathbf{C}_{uv}$, $\mathbf{C}_{vv}$, and $\mathbf{C}_{vu} = \mathbf{C}_{uv}^{\mathrm{T}}$. For convenience, we denote the scalar product $\mathbf{x}^{\mathrm{T}} \mathbf{u}_i$ by $\mathbf{u}^{\mathrm{T}} \mathbf{x}^i$ where we define $\mathbf{x}^i$ to be a vector of the same dimensions as the concatenated vector $\mathbf{u}$ with zero components everywhere except for those that correspond to hidden unit i, which contain the vector $\mathbf{x}$.

As the components of $\mathbf{v}$ do not enter the non-linear sigmoid functions, we can directly integrate over $\mathbf{v}$, so that each term in the summation (22) gives

$$\left\langle \left(\theta_{ij} + (\mathbf{u} - \overline{\mathbf{u}})^{\mathrm{T}} \boldsymbol{\Psi}_{ij} (\mathbf{u} - \overline{\mathbf{u}}) + \boldsymbol{\Omega}_{ij}^{\mathrm{T}} (\mathbf{u} - \overline{\mathbf{u}})\right) \sigma\left(\mathbf{u}^{\mathrm{T}} \mathbf{x}^i\right) \sigma\left(\mathbf{u}^{\mathrm{T}} \mathbf{x}^j\right) \right\rangle \tag{23}$$

where

$$\begin{aligned} \theta_{ij} &= \left(\mathbf{C}_{vv} - \mathbf{C}_{vu} \mathbf{C}_{uu}{}^{-1} \mathbf{C}_{uv}\right)_{ij} + \overline{v}_i \overline{v}_j \\ \boldsymbol{\Psi}_{ij} &= \mathbf{C}_{uu}{}^{-1} \mathbf{C}_{u,v=i} \mathbf{C}_{v=j,u} \mathbf{C}_{uu}{}^{-1} \\ \boldsymbol{\Omega}_{ij} &= \mathbf{C}_{uu}{}^{-1} \mathbf{C}_{u,v=j} \overline{v}_i + \mathbf{C}_{uu}{}^{-1} \mathbf{C}_{u,v=i} \overline{v}_j \end{aligned}$$

and the expectation in (23) is over a Gaussian distribution in $\mathbf{u}$ with mean $\overline{\mathbf{u}}$ and covariance $\mathbf{C}_{uu}$. Although the remaining integrations in (23) over $\mathbf{u}$ are not analytically tractable, we can make use of techniques discussed in the Appendix to reduce them to one-dimensional integrals. For example

$$\left\langle \sigma\left(\mathbf{z}^{\mathrm{T}}\mathbf{a}+a_0\right)\sigma\left(\mathbf{z}^{\mathrm{T}}\mathbf{b}+b_0\right)\right\rangle_{\mathbf{z}} = \left\langle \sigma\left(z\|\mathbf{a}\|+a_0\right)\sigma\left(\frac{z\mathbf{a}^{\mathrm{T}}\mathbf{b}+b_0\|\mathbf{a}\|}{\sqrt{\|\mathbf{a}\|^2\left(1+\|\mathbf{b}\|^2\right)-\left(\mathbf{a}^{\mathrm{T}}\mathbf{b}\right)^2}}\right)\right\rangle_z \tag{24}$$

where $\mathbf{a}$, $\mathbf{b}$ are vectors and a_0, b_0 are scalar offsets. The average on the left of (24) is over an isotropic multi-dimensional Gaussian, $P(\mathbf{z}) \propto \exp(-\mathbf{z}^{\mathrm{T}}\mathbf{z}/2)$, while the average on the right is over the one-dimensional Gaussian $P(z) \propto \exp(-z^2/2)$. The resulting integrals can then be evaluated using standard numerical techniques.

Similar analytical techniques can be used to evaluate the derivatives of the KL divergence with respect to both the mean and covariance matrix (Appendix A.3). Together with the KL divergence, these derivatives are then used in a scaled conjugate gradient optimizer to find the parameters $\overline{\mathbf{w}}$ and $\mathbf{C}$ that represent the best Gaussian fit. It is worthwhile to note that fixed point equations are also derivable for this method. Consider for example optimizing (17) with respect to the parameters $\overline{\mathbf{w}}$. These parameters appear only in the average of the terms E_W and E_D. Using (19) and differentiating with respect to $\overline{\mathbf{w}}$, we then obtain the following condition for an extremum,

$$\mathbf{A}\overline{\mathbf{w}} = -\nabla_{\overline{\mathbf{w}}}\int Q(\mathbf{w})E_D d\mathbf{w} \tag{25}$$

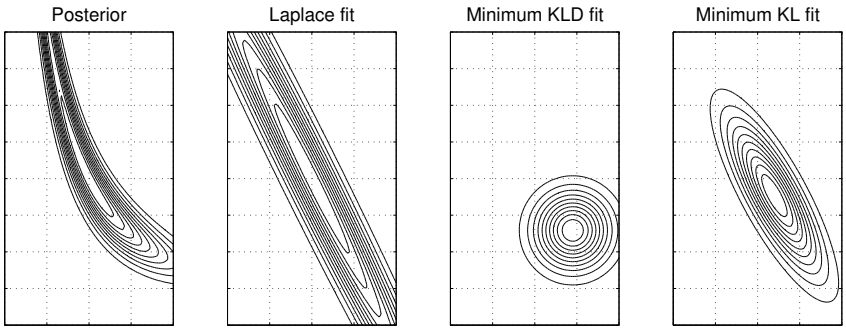
This suggests the iterative solution $\overline{\mathbf{w}}^{new} = -\mathbf{A}^{-1}\nabla_{\overline{\mathbf{w}}^{old}}\int Q(\mathbf{w})E_D d\mathbf{w}$. Similar procedures can be constructed for other variables and can be expected to improve convergence in the KL optimization.

The number of parameters in the covariance matrix scales quadratically with the number of weight parameters. We therefore have also implemented a version with a constrained covariance matrix

$$\mathbf{C} = \mathrm{diag}(d_1^2, \ldots, d_n^2) + \sum_{i=1}^{s}\mathbf{s}_i\mathbf{s}_i^{\mathrm{T}} \tag{26}$$

which is the form of covariance used in factor analysis (Bishop 1998a). This reduces the number of independent parameters in the covariance matrix from $k(k+1)/2$ to $k(s+1) - s(s+1)/2$, which is now linear in k. Thus, the number of parameters can be controlled by changing s and, unlike a diagonal covariance matrix, this model can still capture the strongest of the posterior correlations. For $s = k-1$ we obtain a completely general positive-definite covariance matrix. The value of s should be as large as possible, subject only to computational cost limitations. There is no 'over-fitting' as s is increased since more flexible distributions $Q(\mathbf{w})$ simply give better approximations to the true posterior. The diagonal Gaussian model of Hinton and van Camp (1993) is recovered when $s = 0$.

We illustrate the optimization of the KL divergence in Figure 2 using a synthetic example involving two parameters, thereby allowing the posterior distribution, and various approximations to the posterior, to be plotted directly. The training data set, consisting of 6 points, was



(a) (b) (c) (d)

Figure 2: Comparison of various approximations to the posterior distribution for a synthetic regression problem involving two adaptive parameters (details given in the text). (a) The true posterior distribution. (b) The local Gaussian approximation obtained by Laplace's method, giving a Kullback-Leibler (KL) divergence value of 41. (c) The minimum KL fit obtained with a diagonal covariance Gaussian (KLD), giving a residual KL value of 4.6. (d) The minimum KL fit obtained using a full covariance Gaussian distribution, giving a residual KL value of 3.9.

generated by sampling the single input variable x in the range $(-2, 2)$, computing the corresponding values of the function $y = -2.5\sigma(-0.13x)$, and adding Gaussian noise with standard deviation 0.1.

2.2 Hyperparameter Adaptation

So far, we have treated the hyperparameters β and $\mathbf{A}$ as fixed. We now extend the ensemble learning formalism to include hyperparameters within the Bayesian framework. For simplicity, we consider a standard isotropic prior covariance matrix of the form $\mathbf{A} = \alpha\mathbf{I}$, and introduce hyperpriors given by Gamma distributions

$$\ln p(\alpha) = \ln\left\{\alpha^{a-1}\exp\left(-\frac{\alpha}{b}\right)\right\} + \text{const.} \tag{27}$$

$$\ln p(\beta) = \ln\left\{\beta^{c-1}\exp\left(-\frac{\beta}{d}\right)\right\} + \text{const.} \tag{28}$$

where a, b, c, d are constants. The joint posterior distribution of the weights and hyperparameters is given by

$$p(\mathbf{w}, \alpha, \beta|D) \propto p(D|\mathbf{w}, \beta)\, p(\mathbf{w}|\alpha)\, p(\alpha)\, p(\beta) \tag{29}$$

in which

$$\ln p(D|\mathbf{w}, \beta) = -\beta E_D + \frac{N}{2}\ln\beta + \text{const.} \tag{30}$$

$$\ln p(\mathbf{w}|\alpha) = -\alpha\|\mathbf{w}\|^2 + \frac{k}{2}\ln\alpha + \text{const.} \tag{31}$$

We follow MacKay (1995,1996) by modelling the joint posterior $p(\mathbf{w}, \alpha, \beta|D)$ by a factorized approximating distribution of the form

$$Q(\mathbf{w})R(\alpha)S(\beta) \tag{32}$$

where $Q(\mathbf{w})$ is a Gaussian distribution as before, and the functional forms of R and S are left unspecified. We then maximize the lower bound

$$\mathcal{F}[Q,R,S] = -\iiint Q(\mathbf{w})R(\alpha)S(\beta)\ln\left\{\frac{P(D,\mathbf{w},\alpha,\beta)}{Q(\mathbf{w})R(\alpha)S(\beta)}\right\} d\mathbf{w}\,d\alpha\,d\beta. \tag{33}$$

Consider first the dependence of (33) on $Q(\mathbf{w})$

$$\begin{aligned}\mathcal{F}[Q] &= -\iiint QRS\left\{\beta E_D(\mathbf{w}) + \frac{\alpha}{2}\|\mathbf{w}\|^2 + \ln Q(\mathbf{w})\right\} d\mathbf{w}\,d\alpha\,d\beta + \text{const.} \\ &= -\int Q(\mathbf{w})\left\{\overline{\beta}E_D(\mathbf{w}) + \frac{\overline{\alpha}}{2}\|\mathbf{w}\|^2 + \ln Q(\mathbf{w})\right\} d\mathbf{w} + \text{const.}\end{aligned} \tag{34}$$

where $\overline{\alpha} = \int R(\alpha)\alpha\,d\alpha$ and $\overline{\beta} = \int S(\beta)\beta\,d\beta$. We see that (34) takes the same form as (17), except that the fixed hyperparameters are now replaced with their average values. Thus the optimization of $Q(\mathbf{w})$ will proceed as discussed previously. To calculate the average values of the hyperparameters, we next consider the dependence of the functional $\mathcal{F}$ on $R(\alpha)$

$$\begin{aligned}\mathcal{F}[R] &= \iiint QRS\left\{-\frac{\alpha}{2}\|\mathbf{w}\|^2 + \frac{k}{2}\ln\alpha + (a-1)\ln\alpha \right. \\ &\qquad \left. -\frac{\alpha}{b} - \ln R\right\} d\mathbf{w}\,d\alpha\,d\beta + \text{const.} \\ &= \int R(\alpha)\left\{\frac{\alpha}{s} + (r-1)\ln\alpha - \ln R(\alpha)\right\} d\alpha + \text{const.}\end{aligned} \tag{35}$$

where $r = k/2 + a$ and $2/s = \|\overline{\mathbf{w}}\|^2 + \text{Tr}\,(\mathbf{C}) + 2/b$. We recognize (35) as the negative Kullback-Leibler divergence between $R(\alpha)$ and a Gamma distribution. Thus the optimum $R(\alpha)$ is also Gamma distributed, with

$$R(\alpha) \propto \alpha^{r-1}\exp\left(-\frac{\alpha}{s}\right). \tag{36}$$

We therefore obtain

$$\overline{\alpha} = rs. \tag{37}$$

A similar procedure for $S(\beta)$ gives

$$\overline{\beta} = uv \tag{38}$$

where $u = N/2 + c$ and $1/v = \langle E_D\rangle + 1/d$, in which $\langle E_D\rangle$ has already been calculated during the optimization of $Q(\mathbf{w})$.

This defines an iterative procedure in which we start by initializing the hyperparameters (using the mean of the hyperprior distributions) and then alternately optimize the KL divergence over $Q(\mathbf{w})$ using the techniques discussed in Section 2.1 and re-estimate $\overline{\alpha}$ and $\overline{\beta}$ using (37) and (38).

2.3 Illustrative Application

As a demonstration of our method on a standard benchmark problem, we applied the ensemble learning procedure to the Boston Housing data set. This is a problem with 13 inputs and one output, for which the data can be obtained from the DELVE archive[1](we took 128 training

[1]See `http://www.cs.utoronto.ca/~delve/`

Method	Test Error
Ensemble ($s = 1$)	0.22
Ensemble (diagonal)	0.28
Laplace	0.33

Table 1: Comparison of ensemble learning using the covariance matrix (26) with $s = 1$, ensemble learning using a diagonal covariance, and Laplace's method. The test error is defined to be the mean squared error over the test set.

points and 250 test points). We trained a network of four hidden units, with covariance matrix given by (26) with $s = 1$, and choose broad hyperpriors on α and β by setting $a = 0.25$ and $b = 400$ in (27), and $c = 0.05$ and $d = 2000$ in (28). Predictions were made by evaluating the integral in (8) with $P(\mathbf{w}|D, \beta, \mathbf{A})$ replaced by $Q(\mathbf{w})$, and the hyperparameters fixed at their average values. The required integration over $\mathbf{w}$ can be done analytically (see the Appendix) as a consequence of the form of the sigmoid function given in (2).

We compared the performance of the KL method against the Laplace framework of MacKay (1992) which also treats hyperparameters through a re-estimation procedure. In addition we also evaluated the performance of the ensemble learning method using a diagonal covariance matrix. Results are summarized in Table 1.

3 Discussion

In this chapter we have reviewed the framework of ensemble learning and discussed its application to feed-forward neural networks. We have shown that earlier approaches based on constrained Gaussian variational distributions can be extended to general Gaussian distributions while remaining computationally tractable.

The ensemble learning approach has the virtue of maintaining a rigorous lower bound on the marginal likelihood, and in this sense can be regarded as a more controlled approximation than that provided by the Laplace expansion. One drawback, however, is that it is generally computationally more costly than the Laplace method.

An additional limitation of the ensemble approach (in common with the Laplace method) is that, with a Gaussian approximating distribution, it cannot effectively model a multi-modal posterior distribution. This problem can be tackled by considering a Q distribution which is a probabilistic *mixture* of Gaussian distributions of the form

$$Q(\mathbf{w}) = \sum_{i=1}^{L} \pi_i Q_i(\mathbf{w}) \tag{39}$$

where each component $Q_i(\mathbf{w})$ is a Gaussian with its own mean $\overline{\mathbf{w}}_i$ and covariance $\mathbf{C}_i$, and the mixing coefficients π_i are also adaptive parameters. If we consider using the mixture distribution (39) to evaluate the lower bound (17) we see that the data and prior terms can be obtained trivially by forming linear combinations of the earlier results. The entropy term, however, cannot be evaluated analytically. To tackle this we can follow the approach of Jaakkola and

Jordan (1998) and make further use of variational methods to obtain a lower bound on $\mathcal{F}[Q]$, which will therefore also be a lower bound on the log marginal likelihood. This approach has already been applied successfully to the use of mixture distributions in variational treatments of sigmoidal belief networks (Bishop, Lawrence, Jaakkola, and Jordan 1998) and Boltzmann machines (Lawrence, Bishop, and Jordan 1998). The extension to ensemble learning for neural networks using mixtures of Gaussians is straightforward in principle, although it will increase the complexity and computational cost of the algorithm compared to the use of a single Gaussian Q distribution. Other advances are also possible using tractable approximating distributions of a similar analytic form to the posterior, but with limited interactions between some variables. See Barber and Wiegerinck (1998) for an example of such an approach.

In this chapter we have focussed on the application of ensemble learning methods to feed-forward neural networks. It should be clear, however, that the general approach has much wider applicability. A central issue in any new application is the choice of a suitable form for the variational Q distribution, which should sufficiently flexible to give good performance and yet lead to an analytically tractable lower bound.

An interesting example, discussed by Barber and Schottky (1998), uses a particular form of radial basis function (RBF) network. They consider a model in which the network output is given by

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{H} u_i \exp\left\{-\lambda_i \left(\mathbf{x} - \mathbf{c}_i\right)^{\mathrm{T}} \left(\mathbf{x} - \mathbf{c}_i\right)\right\} \tag{40}$$

where $\mathbf{c}_i$ are the centres of the H basis functions, and u_i the weights. Both $\mathbf{c}_i$ and u_i are treated as adaptive parameters, while the scale parameters λ_i are taken to be fixed. The machinery presented previously for two-layer feed-forward neural networks can then be used in exactly the same manner on this model, including the framework for hyperparameter adaptation. However, the integrals required to compute the Kullback-Leibler divergence are significantly simpler, requiring only Gaussian integration, and so can be performed analytically.

Acknowledgements

We are grateful to thank Mehdi Azzouzi, Neil Lawrence, David MacKay, Bernhard Schottky and Chris Williams for helpful discussions. Also we would like to thank the Isaac Newton Institute for Mathematical Sciences in Cambridge for their hospitality.

References

Barber, D. and C. M. Bishop (1998). Ensemble learning for Multi-Layer Networks. In *Advances in Neural Information Processing Systems NIPS 10*. MIT Press.

Barber, D. and B. Schottky (1998). Radial basis functions : a Bayesian treatment. In *Advances in Neural Information Processing Systems NIPS 10*. MIT Press. In press.

Barber, D. and W. Wiegerinck (1998). Tractable Undirected Approximations for Graphical Models. In *ICANN'98: International Conference on Artificial Neural Networks, Skövde*.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (Second ed.). New York: Springer-Verlag.

Bishop, C. M. (1992). Exact calculation of the Hessian matrix for the multilayer perceptron. *Neural Computation 4*(4), 494–501.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C. M. (1998a). Latent variables, mixture distributions and topographic mappings. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Kluwer.

Bishop, C. M. (1998b). Variational learning in graphical models and neural networks. To appear in Proceedings ICANN'98.

Bishop, C. M., N. Lawrence, T. Jaakkola, and M. I. Jordan (1998). Approximating posterior distributions in belief networks using mixtures. In *Advances in Neural Information Processing Systems*, Volume 10. MIT Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B 39*(1), 1–38.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B 195*(2), 216–222.

Hinton, G. E. and D. van Camp (1993). Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pp. 5–13.

Jaakkola, T. and M. I. Jordan (1998). Approximating posteriors via mixture models. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Kluwer.

Jordan, M. I., Z. Gharamani, T. S. Jaakkola, and L. K. Saul (1998). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Kluwer.

Lawrence, N., C. M. Bishop, and M. Jordan (1998). Mixture representations for inference and learning in Boltzmann machines. In *Uncertainty in Artificial Intelligence*. Morgan Kaufmann.

MacKay, D. J. C. (1992). A practical Bayesian framework for back-propagation networks. *Neural Computation 4*(3), 448–472.

MacKay, D. J. C. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems 6*(3), 469–505.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer. Lecture Notes in Statistics 118.

Neal, R. M. and G. E. Hinton (1998). A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Kluwer.

Pearlmutter, B. A. (1994). Fast exact multiplication by the Hessian. *Neural Computation 6*(1), 147–160.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, B 31*(1), 80–88.

A Evaluation of the Lower Bound

In this Appendix we give a more detailed overview of the calculations needed to evaluate the lower bound on the log marginal likelihood as well as the derivatives of the bound with respect to the model parameters. We do not spell out every step explicitly, rather our aim is to provide enough detail to allow the results to be reconstructed without difficulty.

A.1 Reduction of Dimensionality

Although the evaluation of $\mathcal{F}[Q]$ involves integration over the multi-dimensional parameter space, the integrals can be straightforwardly reduced to one dimension. Here we consider a simple example, although the argument is quite general.

The integrals we need to evaluate involve the Gaussian expectation of some (in general) non-linear function of $\mathbf{w}$ of the form

$$\left\langle g\left(\mathbf{a}^{\mathrm{T}}\mathbf{w}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} = \frac{1}{(2\pi)^{k/2}|\mathbf{C}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{w}-\mathbf{m})\right\} g\left(\mathbf{a}^{\mathrm{T}}\mathbf{w}\right) d\mathbf{w} \tag{41}$$

for some vector $\mathbf{a}$. The key observation is that the argument of the expectation depends on $\mathbf{w}$ only through its projection onto a particular direction defined by $\mathbf{a}$. By making a linear change of variables we can transform the Gaussian to have zero mean and an isotropic covariance, which allows us to integrate trivially over all $k-1$ dimensions orthogonal to $\mathbf{a}$. Thus, if we define $\mathbf{s} = \mathbf{C}^{-\frac{1}{2}}(\mathbf{w}-\mathbf{m})$, we obtain

$$\left\langle g\left(\mathbf{a}^{\mathrm{T}}\mathbf{w}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} = \frac{1}{(2\pi)^{k/2}} \int \exp\left\{-\frac{1}{2}\mathbf{s}^{\mathrm{T}}\mathbf{s}\right\} g\left(\mathbf{a}^{\mathrm{T}}\mathbf{m} + \mathbf{a}^{\mathrm{T}}\mathbf{C}^{1/2}\mathbf{s}\right) d\mathbf{s}. \tag{42}$$

The central idea is now to rotate the coordinate system so that $\mathbf{s}$ can be decomposed into $\mathbf{s} = s_{\|}\mathbf{e} + \mathbf{s}_{\perp}$ where $\mathbf{e}$ is a unit vector parallel to $\mathbf{C}^{1/2}\mathbf{a}$, and $\mathbf{s}_{\perp}$ is orthogonal to $\mathbf{e}$. Then

$$\left\langle g\left(\mathbf{a}^{\mathrm{T}}\mathbf{w}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} = \frac{1}{(2\pi)^{k/2}} \iint \exp\left\{-\frac{1}{2}\mathbf{s}_{\perp}{}^{\mathrm{T}}\mathbf{s}_{\perp} - \frac{1}{2}s_{\|}^2\right\} g\left(s_{\|}\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{C}\mathbf{a}} + \mathbf{a}^{\mathrm{T}}\mathbf{m}\right) d\mathbf{s}_{\perp}\, ds_{\|}. \tag{43}$$

Since the components of $\mathbf{s}_{\perp}$ integrate to unity we obtain

$$\left\langle g\left(\mathbf{a}^{\mathrm{T}}\mathbf{w}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} = \left\langle g\left(w\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{C}\mathbf{a}} + \mathbf{a}^{\mathrm{T}}\mathbf{m}\right)\right\rangle_{\mathcal{N}(0,1)} \tag{44}$$

where we have used the notation $\langle \ldots \rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}$ to represent the expectation with respect to a Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{C}$. Thus the integration over the multi-dimensional vector $\mathbf{w}$ on the left hand side of (44) has been reduced to a one-dimensional integral over the scalar w on the right hand side.

A.2 Specific Integrals

The most complex integral required for computing the KL value is given by the Gaussian average of the product of a quadratic form and two sigmoidal functions, as in (23), of the form

$$\left\langle \left(c + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{d}^{\mathrm{T}}\mathbf{w}\right) \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a} + a_0\right) \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{b} + b_0\right)\right\rangle_{\mathcal{N}(\mathbf{0},\mathbf{I})} \tag{45}$$

where $\mathbf{w}$ is a k-dimensional Gaussian random variable. Without loss of generality, we have considered an expectation with respect to a zero-mean unit-covariance Gaussian, since the expectation with respect to a more general Gaussian distribution can always be reduced to the form (45) by a linear coordinate transformation of $\mathbf{w}$. We defer the derivation for this integral until later, and initially concentrate on a slightly simpler integral, from which we can later generate the result for (45).

Thus we first consider the following

$$I = \left\langle \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a} + a_0\right) \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{b} + b_0\right)\right\rangle_{\mathcal{N}(0,\mathbf{I})}. \tag{46}$$

Following our general discussion of dimensionality reduction in such integrals, we rotate the co-ordinate system so that $\mathbf{w} = w_1\hat{\mathbf{a}} + w_2\hat{\mathbf{c}} + \mathbf{w}_\perp$,where $\hat{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\|$ and $\mathbf{c} = \mathbf{c}/\|\mathbf{c}\|$ with

$$\mathbf{c} = \mathbf{b} - \frac{\mathbf{b}^{\mathrm{T}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\mathbf{a} \tag{47}$$

so that $\mathbf{c}$ is orthogonal to $\mathbf{a}$ and in the plane spanned by $\mathbf{a}$ and $\mathbf{b}$. Similarly, $\mathbf{w}_\perp$ is orthogonal to $\mathbf{a}$ and $\mathbf{c}$. With this choice of coordinate system, the arguments of the activation functions in (46) are independent of the components of $\mathbf{w}_\perp$, and the average over the Gaussian for these directions is unity. This leaves

$$\begin{aligned} I = \frac{1}{2\pi}\iint \exp\left\{-\frac{1}{2}w_1^2 - \frac{1}{2}w_2^2\right\} \sigma\left(w_1\|\mathbf{a}\| + a_0\right) \\ \sigma\left(w_1\hat{\mathbf{a}}^{\mathrm{T}}\mathbf{b} + w_2\hat{\mathbf{c}}^{\mathrm{T}}\mathbf{b} + b_0\right)\, dw_1\, dw_2. \end{aligned} \tag{48}$$

We then make use of the relation

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\sigma\left(cx + d\right)\, dx = \sigma\left(\frac{d}{\sqrt{1+c^2}}\right) \tag{49}$$

which is easily verified by differentiating the left hand side of (49) with respect to d. This enables us to integrate over w_2 and reduce the integral (48) to one dimension, so that

$$I = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}w^2}\sigma\left(\|\mathbf{a}\|w + a_0\right)\sigma\left(\frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}w + \|\mathbf{a}\|b_0}{\sqrt{\|\mathbf{a}\|^2\left(1+\|\mathbf{b}\|^2\right) - \left(\mathbf{a}^{\mathrm{T}}\mathbf{b}\right)^2}}\right) dw. \tag{50}$$

Note that we can also re-express (50) as an integral involving only one 'erf' function by using the identity

$$\begin{aligned} \left\langle \sigma\left(wa + a_0\right)\sigma\left(wb + b_0\right)\right\rangle_{\mathcal{N}(0,1)} = \\ \sigma\left(\frac{b_0}{\sqrt{1+b^2}}\right) - \sqrt{\frac{2}{\pi}}\int_{\alpha}^{\infty} e^{-\frac{1}{2}w^2}\sigma\left(\beta - \gamma w\right) dw \end{aligned} \tag{51}$$

where we have defined

$$\alpha = a_0(1+a^2)^{-1/2} \tag{52}$$

$$\beta = b_0\left(\frac{1+a^2}{1+a^2+b^2}\right)^{1/2} \tag{53}$$

$$\gamma = ab(1+a^2+b^2)^{-1/2}. \tag{54}$$

One potential benefit of such a representation is that the integral (46) can then be computed using a three dimensional lookup table. In our implementation, however, we have numerical integration to evaluate the required one-dimensional integrals.

Next we consider the evaluation of averages of the form

$$\left\langle \left(c + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{d}^{\mathrm{T}}\mathbf{w}\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a}+a_0\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{b}+b_0\right)\right\rangle_{\mathcal{N}(0,\mathbf{I})}. \tag{55}$$

A useful trick in deriving this and similar relations comes from writing the quantity we desire as the derivative of a calculable integral. Consider

$$-\frac{\partial}{\partial\lambda}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\|\mathbf{w}\|^2 - \lambda\left(\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}+\mathbf{d}^{\mathrm{T}}\mathbf{w}\right)\right)f(\mathbf{w})d\mathbf{w}\Big|_{\lambda=0}$$
$$= \int_{-\infty}^{\infty}\left(\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}+\mathbf{d}^{\mathrm{T}}\mathbf{w}\right)\exp\left(-\frac{1}{2}\|\mathbf{w}\|^2\right)f(\mathbf{w})\,d\mathbf{w}. \tag{56}$$

This can be used to evaluate (45) by setting

$$f(\mathbf{w}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a}+a_0\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{b}+b_0\right). \tag{57}$$

We then have to calculate the derivative (with respect to λ) of the (non-zero mean, non-unit covariance) Gaussian integral over the product of two sigmoid functions. We can write this as an integral over a zero mean, unit covariance Gaussian by applying the linear transformation $\mathbf{w}' = \mathbf{B}^{\frac{1}{2}}\left(\mathbf{w}+\lambda\mathbf{B}^{-1}\mathbf{d}\right)$, where $\mathbf{B} = \mathbf{I}+2\lambda\mathbf{A}$ (terms of order λ^2 can be ignored since their derivatives vanish at $\lambda = 0$). We then use (50) to compute this Gaussian integral. Finally, we take the derivative with respect to λ and set this to zero to arrive at

$$\left\langle \left(c_0 + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{d}^{\mathrm{T}}\mathbf{w}\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a}+a_0\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{b}+b_0\right)\right\rangle_{\mathcal{N}(\mathbf{0},\mathbf{I})}$$
$$= \left(c_0+\mathrm{Tr}\left(\mathbf{A}\right)\right)\left\langle\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a}+a_0\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{b}+b_0\right)\right\rangle_{\mathcal{N}(\mathbf{0},\mathbf{I})}$$
$$+ D\left(a_0,b_0,\mathbf{a},\mathbf{b},\mathbf{A},\mathbf{d}\right) + D\left(b_0,a_0,\mathbf{b},\mathbf{a},\mathbf{A},\mathbf{d}\right) \tag{58}$$

where the auxiliary function $D\left(a_0,b_0,\mathbf{a},\mathbf{b},\mathbf{A},\mathbf{d}\right)$ is given by

$$D = \left(\frac{2}{\pi(1+\|\mathbf{b}\|^2)}\right)^{1/2}\exp\left\{-\frac{b_0^2}{2(1+\|\mathbf{b}\|^2)}\right\}$$
$$\left\langle\left(\mathbf{d}^{\mathrm{T}}\mathbf{b}+\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{b}\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a}+a_0\right)\right\rangle_{\mathcal{N}\left(b_0(\mathbf{I}+\mathbf{b}\mathbf{b}^{\mathrm{T}})^{-1}\mathbf{b},(\mathbf{I}+\mathbf{b}\mathbf{b}^{\mathrm{T}})^{-1}\right)}. \tag{59}$$

To evaluate (59), we use the identity

$$\left\langle\left(c+\mathbf{d}^{\mathrm{T}}\mathbf{w}\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{a}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} = \sqrt{\frac{2}{\pi}}\frac{\mathbf{d}^{\mathrm{T}}\mathbf{C}\mathbf{a}}{\sqrt{1+\mathbf{a}^{\mathrm{T}}\mathbf{C}\mathbf{a}}}$$
$$\exp\left\{-\frac{\left(\mathbf{a}^{\mathrm{T}}\mathbf{m}\right)^2}{2(1+\mathbf{a}^{\mathrm{T}}\mathbf{C}\mathbf{a})}\right\} + \left(\mathbf{d}^{\mathrm{T}}\mathbf{m}+c\right)\sigma\left(\frac{\mathbf{a}^{\mathrm{T}}\mathbf{m}}{\sqrt{1+\mathbf{a}^{\mathrm{T}}\mathbf{C}\mathbf{a}}}\right) \tag{60}$$

which can be verified using the coordinate transformation and rotation method discussed earlier.

A.3 Derivatives

Evaluation of the derivatives of the KL divergence is straightforward. However, the resulting expressions tend to be rather lengthy and so, instead of writing out the results in full, we give an outline derivation of the more complex contributions.

In a numerical implementation, a powerful check on both the analysis and the software can be obtained by comparing the derivatives of the KL divergence found by evaluation of the analytic expressions with the same quantities calculated using central differences applied to the expression for the KL divergence itself (Bishop 1995). Note, however, that numerical differentiation is not appropriate for a run-time implementation due to its computational inefficiency.

As before, we concatenate all the parameters into a single vector $\mathbf{w} = (\{v_k\}, \{\mathbf{u}_k\})$. We can then write, for example, $\mathbf{x}^{\mathrm{T}}\mathbf{u}_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}^i$ where we have defined $\mathbf{x}^i$ to be a vector with the same number of components as $\mathbf{w}$ and with zeros everywhere except in the components identified with $\mathbf{u}_i$, which contain $\mathbf{x}$. Similarly, we write $v_i = \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^i$ where we have defined $\boldsymbol{\delta}^i$ to be a zero vector with a single 1 in the component corresponding to v_i. $\boldsymbol{\delta}^{ij}$ will denote the vector $\boldsymbol{\delta}^i$ if $i = j$ and the zero vector otherwise.

We illustrate the general approach by evaluating the derivatives of (21), which we can rewrite in the form

$$l\left(\mathbf{x}, t\right) = \frac{|C|^{1/2}}{(2\pi)^{k/2}} \int e^{-\frac{1}{2}(\mathbf{w}-\mathbf{m})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{w}-\mathbf{m})} \left(\sum_{i=1}^{H} \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^i \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^i\right) - t \right)^2 d\mathbf{w} \tag{61}$$

As we saw in section (A.2), the integral over a Gaussian of two sigmoid functions is the only integral that cannot be computed analytically and we therefore wish to avoid expressions containing such integrals. Our goal is to differentiate (61) with respect to the parameters $\mathbf{m}$ and $\mathbf{C}$ of the variational Gaussian Q distribution. If we simply differentiate (61) directly we will obtain several integrals involving the average of the product of two sigmoid functions. We therefore adopt an alternative approach and first make a linear transformation of the $\mathbf{w}$ variable so that the parameters $\mathbf{m}$ and $\mathbf{C}$ are transferred to the arguments of the sigmoid functions. Since the derivatives of the sigmoids are Gaussian, we obtain simpler expressions. We therefore consider a linear transformation $\mathbf{w}' = \mathbf{C}^{-1}\left(\mathbf{w} - \mathbf{m}\right)$ which gives

$$l\left(\mathbf{x}, t\right) = \frac{1}{(2\pi)^{k/2}} \int e^{-\frac{1}{2}\mathbf{w}'^{\mathrm{T}}\mathbf{w}'} \left[\sum_{i=1}^{H} \left(\mathbf{m} + \mathbf{C}\mathbf{w}'\right)^{\mathrm{T}} \boldsymbol{\delta}^i \sigma\left(\mathbf{m} + \mathbf{C}\mathbf{w}'\right)^{\mathrm{T}} \mathbf{x}^i - t \right]^2 d\mathbf{w}'. \tag{62}$$

If we now differentiate l with respect to some (as yet unspecified) parameter and then transform

back to the original coordinate system, we find

$$\begin{aligned}\partial l(\mathbf{x},t) &= \sqrt{\frac{2}{\pi}}\sum_{ij=1}^{H}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^{i}\boldsymbol{\delta}^{j\,\mathrm{T}}\mathbf{w}\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{i}\right)e^{-\frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{j}\right)^{2}}\left[\boldsymbol{\Omega}_{x}+\mathbf{w}^{\mathrm{T}}\boldsymbol{\theta}_{x}\right]\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} \\ &\quad -t\sqrt{\frac{2}{\pi}}\sum_{j=1}^{H}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^{j}e^{-\frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{j}\right)^{2}}\left[\boldsymbol{\Omega}_{x}+\mathbf{w}^{\mathrm{T}}\boldsymbol{\theta}_{x}\right]\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} \\ &\quad +\sum_{ij=1}^{H}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^{i}\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{i}\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{j}\right)\left[\boldsymbol{\Omega}_{\delta}+\mathbf{w}^{\mathrm{T}}\boldsymbol{\theta}_{\delta}\right]\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})} \\ &\quad -t\sum_{j}^{H}\left\langle \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{j}\right)\left[\boldsymbol{\Omega}_{\delta}+\mathbf{w}^{\mathrm{T}}\boldsymbol{\theta}_{\delta}\right]\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}. \end{aligned} \tag{63}$$

where we have defined

$$\boldsymbol{\Omega}_{x}=\mathbf{x}^{j\,\mathrm{T}}\partial\mathbf{m}-\mathbf{C}\partial\mathbf{C}\mathbf{C}^{-1}\mathbf{m} \qquad \boldsymbol{\theta}_{x}=\mathbf{C}^{-1}\partial\mathbf{C}\mathbf{x}^{i} \tag{64}$$

$$\boldsymbol{\Omega}_{\delta}=\boldsymbol{\delta}^{j\,\mathrm{T}}\partial\mathbf{m}-\mathbf{C}\partial C\mathbf{C}^{-1}\mathbf{m} \qquad \boldsymbol{\theta}_{\delta}=\mathbf{C}^{-1}\partial\mathbf{C}\boldsymbol{\delta}^{i}. \tag{65}$$

We now evaluate these expressions explicitly for specific parameters of the Q distribution. First we consider the components of the mean $\mathbf{m}$ corresponding to the hidden-to-output weights in the network. If we differentiate with respect to $\overline{v}_k$ we obtain

$$\boldsymbol{\Omega}_{\delta}=\boldsymbol{\delta}^{jk}, \qquad \boldsymbol{\Omega}_{x}=\boldsymbol{\theta}_{x}=\boldsymbol{\theta}_{\delta}=\mathbf{0} \tag{66}$$

where $\boldsymbol{\delta}^{jk}$ is as defined above. Substituting this into the expression for the general derivative, we find

$$\frac{1}{2}\frac{\partial l}{\partial\overline{v}_{k}}=\sum_{i=1}^{H}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^{i}\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{i}\right)\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{k}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}-t\left\langle \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{k}\right)\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}. \tag{67}$$

The first term is evaluated using (58), and the second follows straightforwardly by using (44) and then (49) after a linear transformation.

Next we consider the components of the mean $\mathbf{m}$ corresponding to the input-to-hidden weights in the network. If we differentiate with respect to the k-th component of the h-th input to hidden vector which we denote $\overline{\mathbf{u}}_{kh}$ we obtain

$$\boldsymbol{\Omega}_{x}=\mathbf{x}^{j}\delta_{jk}, \qquad \boldsymbol{\Omega}_{\delta}=\boldsymbol{\theta}_{x}=\boldsymbol{\theta}_{\delta}=\mathbf{0}. \tag{68}$$

This leads to the following formula for the derivative

$$\begin{aligned}\sqrt{\frac{\pi}{2}}\frac{\partial l}{\partial\overline{\mathbf{u}}_{kh}} &= x_{k}\left[\sum_{i=1}^{H}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^{i}\boldsymbol{\delta}^{h\,\mathrm{T}}\mathbf{w}\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{i}\right)e^{-\frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{h})^{2}}\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}\right. \\ &\quad \left.-t\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^{h}e^{-\frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{h}\right)^{2}}\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}\right]. \end{aligned} \tag{69}$$

The final term is calculated using the identity

$$\left\langle \mathbf{w}^{\mathrm{T}}\mathbf{d}\exp\left\{-\frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)^2\right\}\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}$$
$$= \frac{1}{\sqrt{1+\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x}}}\exp\left\{-\frac{\left(\mathbf{m}^{\mathrm{T}}\mathbf{x}\right)^2}{2(1+\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x})}\right\}\left(\mathbf{d}^{\mathrm{T}}\mathbf{m}-\frac{(\mathbf{d}^{\mathrm{T}}\mathbf{C}\mathbf{x})(\mathbf{x}^{\mathrm{T}}\mathbf{m})}{1+\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x}}\right). \tag{70}$$

Rearranging the exponent of the first term in brackets of (69), we get

$$\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^i\boldsymbol{\delta}^{h\mathrm{T}}\mathbf{w}\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^i\right)e^{-\frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^h)^2}\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}$$
$$= \frac{1}{\sqrt{1+\mathbf{x}^{h\mathrm{T}}\mathbf{C}\mathbf{x}^h}}\exp\left\{-\frac{1}{2}\frac{\left(\mathbf{m}^{\mathrm{T}}\mathbf{x}^h\right)^2}{\left(1+\mathbf{x}^{h\mathrm{T}}\mathbf{C}\mathbf{x}^h\right)}\right\}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^i\mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^h\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^i\right)\right\rangle_{\mathcal{N}(\mathbf{b},\mathbf{D})} \tag{71}$$

where

$$\mathbf{D} = \mathbf{C} - \frac{\mathbf{C}\mathbf{x}^h\mathbf{x}^{h\mathrm{T}}\mathbf{C}}{1+\mathbf{x}^{h\mathrm{T}}\mathbf{C}\mathbf{x}^h}, \qquad \mathbf{b} = \mathbf{m} - \frac{\mathbf{x}^{h\mathrm{T}}\mathbf{m}\mathbf{C}\mathbf{x}^h}{1+\mathbf{x}^{h\mathrm{T}}\mathbf{C}\mathbf{x}^h}. \tag{72}$$

This average can then be computed using the results to be presented in (79).

Finally we consider the derivatives with respect to the elements of the covariance matrix $\mathbf{C}$. The form of these derivatives depends on how we represent the covariance matrix. Here we write the covariance matrix in Cholesky factorized form

$$\mathbf{C} = \widetilde{\mathbf{C}}^{\mathrm{T}}\widetilde{\mathbf{C}} \tag{73}$$

where $\widetilde{\mathbf{C}}$ is an upper triangular matrix. If we consider the derivative with respect to the α,β element of $\widetilde{\mathbf{C}}$ we have

$$\boldsymbol{\Omega}_\delta + \mathbf{w}^{\mathrm{T}}\boldsymbol{\theta}_\delta = (\mathbf{w}-\mathbf{m})^{\mathrm{T}}\mathbf{c}^\alpha\boldsymbol{\delta}^{j\beta} \tag{74}$$
$$\boldsymbol{\Omega}_x + \mathbf{w}^{\mathrm{T}}\boldsymbol{\theta}_x = (\mathbf{w}-\mathbf{m})^{\mathrm{T}}\mathbf{c}^\alpha\mathbf{x}^{j\beta} \tag{75}$$

where $\mathbf{x}^{j\beta}$ is the β component of $\mathbf{x}^j$. Since $\mathbf{x}^j$ is zero except for the components corresponding to hidden unit j, $\mathbf{x}^{j\beta}$ is zero unless β also refers to the same hidden unit. In (74) and (75) we have used the notation, $[\mathbf{c}^\alpha]_i = \left[\widetilde{\mathbf{C}}^{-1}\right]_{i,j=\alpha}$, that is $\mathbf{c}^\alpha$ is the α-th column of the inverse of the Cholesky factor.

The first term on the right hand side of (63) is the only one that we have not described how to evaluate and for this reason, we shall present here only the results needed for this term. This involves the Gaussian integral of a cubic weight term with a sigmoid of the form

$$\sqrt{\frac{2}{\pi}}\sum_{i,j=1}^{H}\left\langle \mathbf{w}^{\mathrm{T}}\boldsymbol{\delta}^i\boldsymbol{\delta}^{j\mathrm{T}}\mathbf{w}\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^i\right)e^{-\frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^j\right)^2}(\mathbf{w}-\mathbf{m})^{\mathrm{T}}\mathbf{c}^\alpha\mathbf{x}^{j\beta}\right\rangle_{\mathcal{N}(\mathbf{m},\mathbf{C})}. \tag{76}$$

Note that the sum over j only extends over those values which correspond to the same hidden unit as β (as a result of the factor $\mathbf{x}^{j\beta}$). Completing the square in the exponential and rearranging, we are required to compute, for various settings of i,j,k, terms of the form $h = \hat{h}\tilde{h}$ where

$$\hat{h} = \frac{1}{\sqrt{1+\mathbf{x}^{i\mathrm{T}}\mathbf{C}\mathbf{x}^i}}\exp\left\{-\frac{1}{2}\frac{\left(\mathbf{m}^{\mathrm{T}}\mathbf{x}^i\right)^2}{1+\mathbf{x}^{i\mathrm{T}}\mathbf{C}\mathbf{x}^i}\right\}\left\langle w_iw_jw_k\sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^i\right)\right\rangle_{\mathcal{N}\left(-\frac{1}{2}\mathbf{D}^{-1}\mathbf{d},\mathbf{D}^{-1}\right)} \tag{77}$$

and

$$\tilde{h} = \frac{1}{\sqrt{1 + {\mathbf{x}^i}^{\mathrm{T}} \mathbf{C} \mathbf{x}^i}} \exp\left\{ -\frac{1}{2} \frac{\left(\mathbf{m}^{\mathrm{T}} \mathbf{x}^i\right)^2}{1 + {\mathbf{x}^i}^{\mathrm{T}} \mathbf{C} \mathbf{x}^i} \right\} \bar{w}_k \left\langle w_i w_j \sigma\left(\mathbf{w}^{\mathrm{T}} \mathbf{x}^i\right) \right\rangle_{\mathcal{N}\left(-\frac{1}{2}\mathbf{D}^{-1}\mathbf{d}, \mathbf{D}^{-1}\right)} \tag{78}$$

in which $\mathbf{d} = 2\mathbf{C}^{-1}\mathbf{m}$ and $\mathbf{D} = \mathbf{C}^{-1} + \mathbf{x}^i {\mathbf{x}^i}^{\mathrm{T}}$.

The integrals $\hat{h}$ and $\tilde{h}$ can be computed using the results:

$$\begin{aligned} &\left\langle w_a w_b \sigma\left(\mathbf{w}^{\mathrm{T}} \mathbf{a}\right) \right\rangle_{\mathcal{N}(\mathbf{m}, \mathbf{C})} \\ &\qquad = -4\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\phi^2} \left[c_a c_b \phi + \frac{1}{2}\left(c_a m_b + c_b m_a\right) \right] + \sigma\left(\phi\right)\left(\mathbf{C}_{ab} + m_a m_b\right) \end{aligned} \tag{79}$$

where

$$\mathbf{c} = -\frac{1}{2} \frac{\mathbf{C}\mathbf{a}}{\sqrt{1 + \mathbf{a}^{\mathrm{T}} \mathbf{C} \mathbf{a}}}, \qquad \phi = \frac{\mathbf{m}^{\mathrm{T}} \mathbf{a}}{\sqrt{1 + \mathbf{a}^{\mathrm{T}} \mathbf{C} \mathbf{a}}}. \tag{80}$$

With these same definitions we finally have

$$\begin{aligned} &\left\langle w_a w_b w_c \sigma\left(\mathbf{w}^{\mathrm{T}} \mathbf{a}\right) \right\rangle_{\mathcal{N}(\mathbf{m}, \mathbf{C})} \\ &\qquad = \sigma\left(\phi\right)\left(m_a \mathbf{C}_{bc} + m_b \mathbf{C}_{ac} + m_c \mathbf{C}_{ab} + m_a m_b m_c\right) \\ &\quad - 8\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\phi^2} \left\{ c_a c_b c_c \left[\phi^2 - 1\right] + \frac{1}{4}\left(c_c m_a m_b + c_b m_a m_c + c_a m_c m_b\right) \right. \\ &\qquad \left. + \frac{1}{2}\phi\left(c_a c_b m_c + c_a c_c m_b + c_c c_b m_a\right) + \frac{1}{4}\left(c_a \mathbf{C}_{bc} + c_b \mathbf{C}_{ac} + c_c \mathbf{C}_{ab}\right) \right\}. \end{aligned} \tag{81}$$