Comparison of Approximate Methods for Handling Hyperparameters

David J.C. MacKay
Cavendish Laboratory,
Cambridge, CB3 0HE. United Kingdom.
mackay@mrao.cam.ac.uk

To appear in *Neural Computation*

Abstract

I examine two approximate methods for computational implementation of Bayesian hierarchical models, that is, models which include unknown hyperparameters such as regularization constants and noise levels. In the 'evidence framework' the model parameters are *integrated* over, and the resulting evidence is *maximized* over the hyperparameters. The optimized hyperparameters are used to define a Gaussian approximation to the posterior distribution. In the alternative 'MAP' method, the true posterior probability is found by *integrating* over the hyperparameters. The true posterior is then *maximized* over the model parameters, and a Gaussian approximation is made. The similarities of the two approaches, and their relative merits, are discussed, and comparisons are made with the ideal hierarchical Bayesian solution.

In moderately ill–posed problems, integration over hyperparameters yields a probability distribution with a skew peak which causes significant biases to arise in the MAP method. In contrast, the evidence framework is shown to introduce negligible predictive error, under straightforward conditions. General lessons are drawn concerning inference in many dimensions.

1 The overfitting problem and hyperparameters in neural networks

Feedforward neural networks are often trained to solve regression and classification problems using algorithms that minimize an *error function*, a measure of goodness of fit to the training data (Rumelhart *et al.* 1986). If nothing is done to control the complexity of the resulting neural network, an inevitable consequence of error–minimization will be *overfitting* — the neural network will learn a function which fits spurious details and noise in the data.

There are several approaches to the overfitting problem in neural networks. A crude technique known as 'early stopping' attempts to track a measure of generalization performance during optimization and halt the learning algorithm at the point where this generalization error appears to start to increase. However, most generalization measures are themselves noisy, so the turning point is not easy to identify. Furthermore, the outcome of early stopping will depend on the details of the optimizer chosen to perform the minimization and the initial conditions. And early stopping is unable to independently control *multiple dimensions of complexity*; if, as seems reasonable in the case of large models, there is more than one degree of freedom in the model's 'complexity', early stopping would seem too crude a method for complexity control, since it controls complexity using only one degree of freedom — the simulation time.

A more principled approach to overfitting, and one that is less implementation–dependent, is to change the objective function by adding one or more *regularizers* that penalize complex functions. There are various regularizers, the simplest and most popular being 'weight decay' (Hinton and Sejnowski 1986) (also known as 'ridge regression'). The regularizer in this case is αE_W where E_W is half the sum of the squares of the weights $\{w_i\}$ in the neural network,

$$E_W = \frac{1}{2} \sum_i w_i^2. \tag{1}$$

The motivation for this regularizer is that functions with a complex dependence on the inputs of a network require larger weights than simple functions, so this regularizer penalizes the more complex functions and favours smooth ones. This is known as a weight decay regularizer because its derivative with respect to w_i is $\partial(\alpha E_W)/\partial w_i = \alpha w_i$, a term which under gradient descent causes the weights to decay exponentially to zero with a 'weight decay rate' of α. When such a regularizer is used the overfitting problem reappears as the task of setting this complexity control 'hyperparameter' α. Too large a value of α will cause the interpolant to be too smooth so that genuine structure is neglected. Too small a value of α will also give poor generalization because of overfitting. Other regularization schemes have been suggested (Weigend *et al.* 1991), and the same problem of controlling the hyperparameters applies to those models too.

One way of describing the overfitting problem is to view the neural network as an approximation or estimation tool and describe the control of complexity as a trade–off between *bias* and *variance* (see Bishop (1995) for a review). This might be termed the sampling theory approach to the problem.

This paper is concerned with an alternative Bayesian viewpoint of neural network learning (MacKay 1991; Buntine and Weigend 1991; MacKay 1992c; Neal 1993a; Ripley 1996; Neal 1996) in which the data error is interpreted as defining a likelihood function, and the regularizer corresponds to a prior probability distribution over the weights. From this viewpoint the question of what value α should take can be thought of as a model comparison question, where the models being compared differ by assigning different priors to the parameters. In (MacKay 1991; MacKay 1992c) it was shown that it made theoretical sense, and could be practically beneficial, to use multiple hyperparameters $\{\alpha_c\}$, each one controlling a different aspect of the prior probability distribution. Methods for controlling these multiple hyperparameters were developed by MacKay (1991) using Gaussian approximations and by Neal (1993a) using Markov chain Monte Carlo methods. The approach to implementing Bayesian neural networks suggested by Buntine and Weigend (1991) was subtly different in its treatment of the hyperparameters. As in MacKay's (1991) approach, the use of Gaussian approximations was suggested, but the hyperparameters were 'integrated out' of the problem analytically *before* the Gaussian approximation.

In this paper I compare the approximate strategies of MacKay (1991) and Buntine and Weigend (1991) for handling hyperparameters, assuming a Bayesian approach to neural networks. This comparison is also relevant to other ill–posed problems such as image reconstruction (Gull 1989). For simplicity I will concentrate on the case of a single hyperparameter α, and I will assume that the prior is Gaussian over $\mathbf{w}$, and that the likelihood function is also a Gaussian function of $\mathbf{w}$. I believe that the insights obtained concerning the differences between the approximate methods also apply to models that have more complex likelihood functions and that have priors with multiple hyperparameters.

2 The model studied

In inference problems, a Bayesian model $\mathcal{H}$ commonly takes the form:

$$P(D, \mathbf{w}, \alpha, \beta|\mathcal{H}) = P(D|\mathbf{w}, \beta, \mathcal{H}) P(\mathbf{w}|\alpha, \mathcal{H}) P(\alpha, \beta|\mathcal{H}), \tag{2}$$

where D is the data, $\mathbf{w}$ is the parameter vector, β defines a noise variance $\sigma_\nu^2 = 1/\beta$, and α is a regularization constant. In a regression problem, for example, D might be a set of data points, $\{\mathbf{t}\}$, at given locations $\{\mathbf{x}\}$, and the vector $\mathbf{w}$ might parameterize a function $f(\mathbf{x}; \mathbf{w})$. The model $\mathcal{H}$ states that for some $\mathbf{w}$, the dependent variables $\{\mathbf{t}\}$ arise from the addition of noise to $\{f(\mathbf{x}; \mathbf{w})\}$; the likelihood function $P(D|\mathbf{w}, \beta, \mathcal{H})$ describes the assumed noise process, parameterized by a noise level $1/\beta$; the prior probability of the parameters $P(\mathbf{w}|\alpha, \mathcal{H})$ embodies assumptions about the spatial correlations and smoothness that the true function is expected to have, parameterized by a regularization constant α. The variables α and β are known as hyperparameters. Problems for which models can be written in the form (2) include linear interpolation with a fixed basis set (Gull 1988; MacKay 1992a), nonlinear regression with a neural network (MacKay 1992c), nonlinear classification (MacKay 1992b), and image deconvolution (Gull 1989).

In the simplest case (linear models, Gaussian noise), the first factor in (2), the likelihood, can be written in terms of a quadratic function of $\mathbf{w}$, $E_D(\mathbf{w})$:

$$P(D|\mathbf{w},\beta,\mathcal{H}) = \frac{1}{Z_D(\beta)}\exp(-\beta E_D(\mathbf{w})), \tag{3}$$

where $Z_D(\beta)$ is a normalization constant with no $\mathbf{w}$–dependence. In the case of 'ill–posed' problems, the hessian $\nabla\nabla E_D$ is ill–conditioned — some of its eigenvalues are very small, so that the maximum likelihood parameters depend undesirably on the noise in the data. The model is 'regularized' by the second factor in (2), the prior, which in the simplest case is a spherical Gaussian:

$$P(\mathbf{w}|\alpha,\mathcal{H}) = \frac{1}{Z_W(\alpha)}\exp(-\alpha\tfrac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}). \tag{4}$$

where $Z_W(\alpha) = \int d^k\mathbf{w}\,\exp(-\alpha\mathbf{w}^{\mathsf{T}}\mathbf{w}/2)$, with k denoting the dimensionality of the parameter vector $\mathbf{w}$. The regularization constant α defines the variance $\sigma_w^2 = 1/\alpha$ of the components w_i of $\mathbf{w}$ under the prior. This simple linear model will be studied in this paper because it provides a convenient test–bed for comparing approximate inference methods. If a method behaves pathologically in this simple case, how can we expect it to behave well when applied to more complex nonlinear models?

Much interest has centred on the question, for models like the one defined in equations (3–4), of how the constants α and β — or the ratio α/β — should be set, and Gull (1989) has derived an appealing Bayesian prescription for these constants (see also MacKay (1992a) for a review). This 'evidence framework' *integrates* over the *parameters* $\mathbf{w}$ to give the 'evidence' $P(D|\alpha,\beta,\mathcal{H})$. The evidence is then *maximized* over the *regularization constant* α *and noise level* β. A Gaussian approximation is then made with the hyperparameters fixed to their optimized values. This relates closely to the 'generalized maximum likelihood' or 'MLII' method in statistics (Wahba 1975). This method can be applied to nonlinear models by making appropriate local linearizations (so that the integral over the parameters is made approximately rather than exactly) and has been used successfully in image reconstruction (Gull 1989; Weir 1991) and in neural networks (MacKay 1992c; Thodberg 1996; MacKay 1996).

An alternative procedure for computing inferences under the same Bayesian model has been suggested by Buntine and Weigend (1991), Strauss *et al.* (1993) and Wolpert (1993). In this approach, one *integrates* over the *regularization constant* α first to obtain the 'true prior', and over the *noise level* β to obtain the 'true likelihood'; then *maximizes* the 'true posterior' (which is proportional to the product of the true prior and the true likelihood) over the *parameters* $\mathbf{w}$. A Gaussian approximation is then made around this true probability density maximum. I will call this the 'MAP' method (for *maximum a posteriori*) although this use of the term 'MAP' may not coincide precisely with its general usage. In the MAP method, the integrations over α can typically be performed exactly, and the posterior probability density maximum is found without any approximations being made. The MAP method is an approximation in that the Gaussian fitted at the posterior maximum is an approximation to the true posterior distribution.

The purpose of this paper is to examine the choice between these two Gaussian approximations, both of which might be used to approximate predictive inference for high–dimensional problems. Of course the *ideal* Bayesian approach would be to obtain predictions by integrating out *all* the parameters and hyperparameters, and this would certainly be preferred. The assumption here is that this is a challenging integral to perform, and that we are only able to analytically integrate over *either* the parameters (for fixed hyperparameters), as in the evidence framework, *or* over the hyperparameters (for fixed parameters) as in the MAP method.

It is assumed that predictive *distributions* are of interest, rather than point *estimates*. Estimation will only appear as a computational stepping stone in the process of approximating a predictive distribution. I concentrate on the simplest case of the linear model with Gaussian noise, but the insights obtained are expected to apply to more general nonlinear models and to models with multiple hyperparameters. When a nonlinear model has multiple local optima, one can approximate the posterior by a sum of Gaussians, one fitted at each optimum. There is then an analogous choice between either (a) optimizing α separately at each local optimum in $\mathbf{w}$ and using a Gaussian approximation conditioned on α (MacKay 1992c); or (b) fitting multiple Gaussians to local maxima of the true posterior with the hyperparameter α integrated out. The results of this paper shed light on this choice.

We will assume for simplicity that the noise level β is known precisely, so that only the regularization constant α is respectively optimized or integrated over. Comments about α can apply equally well to β.

3 Pictorial comparison of the two methods

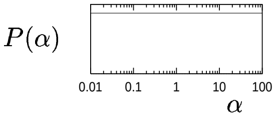
The two approximations are illustrated graphically for a simple two–parameter problem in figures 1 and 2. There are two unknown parameters w_1, w_2, with a prior distribution that is Gaussian with mean zero and variance $1/\alpha$,

$$P(w_1, w_2|\alpha) = \frac{\alpha}{2\pi} \exp\left(-\frac{\alpha}{2}(w_1^2 + w_2^2)\right), \tag{5}$$

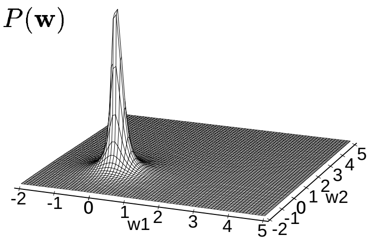
where α is an unknown hyperparameter whose prior distribution (figure 1(a)) is uniform over $\log \alpha$ from $\alpha = 0.01$ to $\alpha = 100$. This prior expresses a belief that w_1 and w_2 are likely to be similar in magnitude, and that their magnitudes might be about 0.1, 1.0, or 10. There are two data points d_1 and d_2 which differ from w_1 and w_2 by additive Gaussian noise of known variance $\sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$ respectively.

$$P(d_1, d_2|w_1, w_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(d_1 - w_1)^2}{2\sigma_1^2} + \frac{(d_2 - w_2)^2}{2\sigma_2^2}\right]. \tag{6}$$

(a) Prior over α

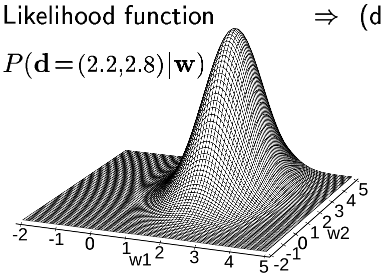


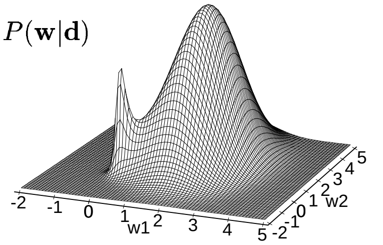
(b) True prior over $\mathbf{w}$



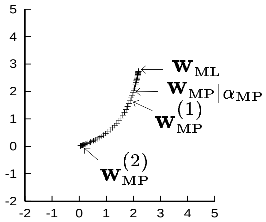
⇓

(c) Likelihood function ⇒ (d) True posterior

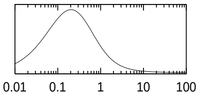




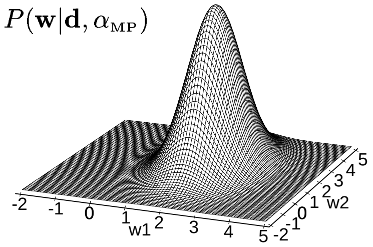
(e) Alpha trajectory



(f) $P(\alpha|\mathbf{d})$



(g) Evidence approximation



(h) MAP approximations:

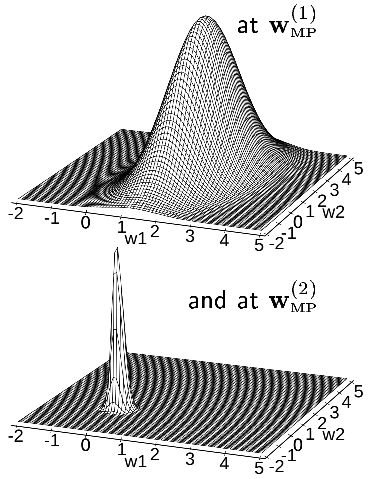
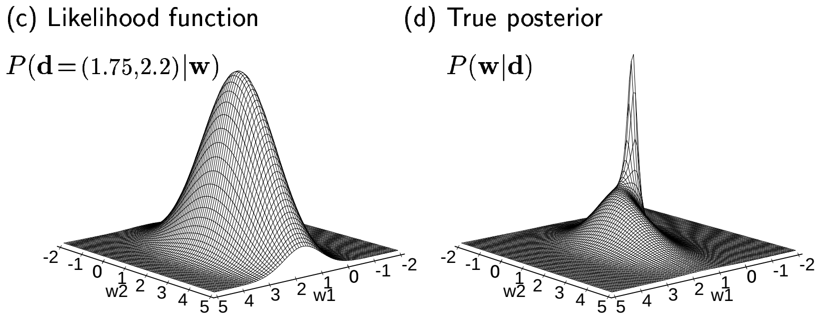
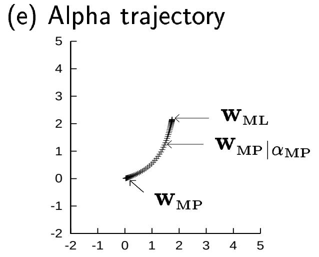


Figure 1: Comparison of the evidence approximation and the MAP approximation for a two–dimensional problem with data $\mathbf{d} = (2.2, 2.8)$.

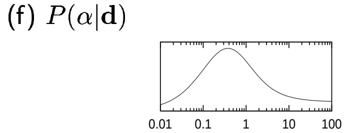


(c) Likelihood function

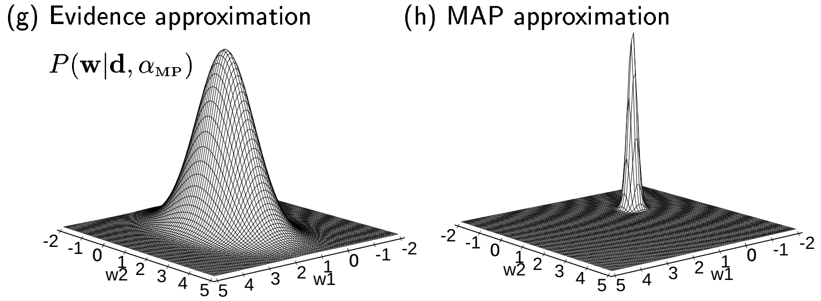
(d) True posterior



(e) Alpha trajectory



(f) $P(\alpha|\mathbf{d})$



(g) Evidence approximation

(h) MAP approximation

Figure 2: Comparison of the evidence approximation and the MAP approximation when $\mathbf{d}=(1.75,2.2)$.

(Or equivalently, there could be more than two data points, all having Gaussian distributions with equal variance, for example, if w_1 is measured independently sixteen times, and w_2 is measured once, with the measurements having variance $\sigma^2 = 2$.)

The 'true prior',

$$P(w_1, w_2) = \int d\alpha \, P(w_1, w_2|\alpha) P(\alpha), \tag{7}$$

is shown in figure 1(b). It is obtained by integrating the prior conditional on α (equation (5)) with respect to the prior on α,

$$P(\log \alpha) = \begin{cases} 1/\log \frac{100}{0.01} & \alpha \in (0.01, 100) \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

We are interested in the posterior distribution of w_1 and w_2 conditional on $\{d_1, d_2\}$; the 'true posterior' (as distinct from the posterior distribution conditional on some value of α) is:

$$P(w_1, w_2|d_1, d_2) \propto P(d_1, d_2|w_1, w_2) P(w_1, w_2). \tag{9}$$

3.1 Let the data be $\{d_1, d_2\} = \{2.2, 2.8\}$

The likelihood function for the case $\{d_1, d_2\} = \{2.2, 2.8\}$ is shown in figure 1(c). The 'true posterior' (which is proportional to the product of the likelihood and the prior) is shown in figure 1(d). At this point we notice that the true posterior has two maxima — one associated with a large peak that encompasses the maximum likelihood parameters, and one close to the origin which is associated with a very narrow peak. The 'alpha trajectory' is shown in figure 1(e). This is the path that is followed by the maximum of the posterior conditional on α, $P(\mathbf{w}|\mathbf{d}, \alpha)$, as α is varied from a large value (which puts the posterior maximum near the origin) to a small value (which puts it close to the maximum likelihood value, $\mathbf{w} = \mathbf{w}_{\text{ML}}$). We will see in section 4.4 that the maxima and saddle points of the true posterior happen to lie exactly on the alpha trajectory. The posterior probability of α, which is maximized in the evidence framework, is shown in figure 1(f). The evidence approximation, $P(\mathbf{w}|\mathbf{d}, \alpha_{\text{MP}})$, is shown in figure 1(g). The Gaussian approximations found by the MAP method (there are two, because the true posterior has two maxima) are shown in figure 1(h).

In this first example, it is not clear if one approximation is superior to the other. We note that whereas the true posterior (figure 1(d)) is multimodal, the posterior probability of α is unimodal in this case, and the posterior probability of $\mathbf{w}$ given α_{MP} is also unimodal.

Let us now study the situation for a slightly different data set.

3.2 Let the data be $\{d_1, d_2\} = \{1.75, 2.2\}$

The likelihood function for the case $\{d_1, d_2\} = \{1.75, 2.2\}$ is shown in figure 2(c). The true posterior is shown in figure 2(d). In this case, unlike figure 1(d), the true posterior has only one maximum. Both the maximum formerly associated with the large peak and the saddle point between the maxima have vanished. The sole maximum of the true posterior is a sharp peak close to the origin. The posterior probability of α is shown in figure 2(f). The evidence approximation $P(\mathbf{w}|\mathbf{d}, \alpha_{\rm MP})$, is shown in figure 2(g). The Gaussian approximation found by the MAP method is shown in figure 2(h).

In this case, it seems that the MAP method is being led astray by the tall but narrow and skew peak of the probability density. Although the *density* is maximized at this peak, most of the posterior probability *mass* is elsewhere. The Gaussian fitted by the method suggested by Buntine and Weigend (1991), Strauss *et al.* (1993) and Wolpert (1993) appears to be a poor representation of the true posterior. The evidence approximation is not a perfect approximation either — it fails to capture the narrow peak where the true posterior is maximized; but it appears to robustly capture most of the posterior probability mass.

Of course, we cannot judge between two approximate methods on the basis of a toy problem alone. The rest of this paper aims to fill out the picture, with an emphasis on what is expected to happen in high–dimensional problems in which there are ill–determined as well as well–determined parameters. What we will see is that figure 2 gives a good intuition for what happens in high dimensions. We will show that the true posterior distribution usually has a skew peak if there are ill–determined parameters, and the true posterior density's maximum is usually unrepresentative of the true posterior density.

4 The alternative methods in detail

Given the Bayesian model defined in (2), we might be interested in the following inferences.
Problem A: Infer the parameters, *i.e.*, obtain a compact representation of $P(\mathbf{w}|D, \mathcal{H})$ and the marginal distributions $P(w_i|D, \mathcal{H})$.
Problem B: Infer the relative model plausibility, which requires the 'evidence' $P(D|\mathcal{H})$.
Problem C: Make predictions, *i.e.*, obtain some representation of $P(D_2|D, \mathcal{H})$, where D_2, in the simplest case, is a single new datum.

4.1 The ideal approach

Ideally, if we were able to do all the necessary integrals, we would just generate the probability distributions $P(\mathbf{w}|D, \mathcal{H})$, $P(D|\mathcal{H})$, and $P(D_2|D, \mathcal{H})$ by direct integration over everything that we are not concerned with. The pioneering work of Box and Tiao (1973) used this approach to develop Bayesian robust statistics.

For real problems of interest, however, such exact integration methods are seldom available. A partial solution can still be obtained by using Monte Carlo methods to simulate the full probability distribution (see Neal (1993b) for an excellent review of Monte Carlo methods and Neal (1996) for the application of these methods to hierarchical models). Thus one can obtain (problem A) a set of samples $\{\mathbf{w}\}$ which represent the posterior $P(\mathbf{w}|D,\mathcal{H})$, and (problem C) a set of samples $\{D_2\}$ which represent the predictive distribution $P(D_2|D,\mathcal{H})$. Unfortunately, the evaluation of the evidence $P(D|\mathcal{H})$ with Monte Carlo methods (problem B) is a difficult undertaking. Recent developments (Neal 1993a; Skilling 1993) now make it possible to use gradient and curvature information so as to sample high dimensional spaces more effectively, even for highly non–Gaussian distributions. Let us come down from these clouds however, and turn attention to the two deterministic approximations under study.

4.2 The evidence framework

The evidence framework divides our inferences into distinct 'levels of inference':
Level 1: Infer the parameters $\mathbf{w}$ for a given value of α:

$$P(\mathbf{w}|D,\alpha,\mathcal{H}) = \frac{P(D|\mathbf{w},\alpha,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})}{P(D|\alpha,\mathcal{H})}. \tag{10}$$

Level 2: Infer α:

$$P(\alpha|D,\mathcal{H}) = \frac{P(D|\alpha,\mathcal{H})P(\alpha|\mathcal{H})}{P(D|\mathcal{H})}. \tag{11}$$

Level 3: Compare models:

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}). \tag{12}$$

There is a pattern in these three applications of Bayes' rule: at each of the higher levels 2 and 3, the data–dependent factor (*e.g.* in level 2, $P(D|\alpha,\mathcal{H})$) is the normalizing constant (the 'evidence') from the preceding level of inference.

The inference problems listed at the beginning of this section are solved approximately using the following procedure.

- The level 1 inference is approximated by making a quadratic expansion of $\log P(D|\mathbf{w},\alpha,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})$ around a maximum of $P(\mathbf{w}|D,\alpha,\mathcal{H})$; this expansion defines a Gaussian approximation to the posterior. The evidence $P(D|\alpha,\mathcal{H})$ is estimated by evaluating the appropriate determinant. For linear models the Gaussian approximation is exact.

- By maximizing the evidence $P(D|\alpha,\mathcal{H})$ at level 2, we find the most probable value of the regularization constant, α_{MP}, and by Taylor–expanding $\log P(D|\alpha,\mathcal{H})$ with respect to $\log\alpha$, we obtain error bars on $\log\alpha$, $\sigma_{\log\alpha|D}$. (Because α is a positive scale variable, it is natural to represent its uncertainty on a log scale.)

- The value of α_{MP} is substituted at level 1. This defines a probability distribution $P(\mathbf{w}|D,\alpha_{\text{MP}},\mathcal{H})$ which is intended to be a good approximation (in a sense we will clarify later) to the posterior $P(\mathbf{w}|D,\mathcal{H})$. The solution offered for problem A is a

Gaussian distribution around the maximum of this distribution, $\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$, with covariance matrix $\mathbf{\Sigma}$ defined by

$$\mathbf{\Sigma}^{-1} = -\nabla\nabla \log P(\mathbf{w}|D, \alpha_{\mathrm{MP}}, \mathcal{H}). \tag{13}$$

Marginals for the components of $\mathbf{w}$ are easily obtained from this distribution.

- The evidence for model $\mathcal{H}$ (problem B) is estimated using Laplace's approximation:

$$P(D|\mathcal{H}) \simeq P(D|\alpha_{\mathrm{MP}}, H)P(\log\alpha_{\mathrm{MP}}|\mathcal{H})\,\sqrt{2\pi}\,\sigma_{\log\alpha|D}. \tag{14}$$

- Problem C: The predictive distribution $P(D_2|D, \mathcal{H})$ is approximated by using the posterior distribution with $\alpha = \alpha_{\mathrm{MP}}$:

$$P(D_2|D, \alpha_{\mathrm{MP}}, \mathcal{H}) = \int d^k\mathbf{w}\;\; P(D_2|\mathbf{w}, \mathcal{H})P(\mathbf{w}|D, \alpha_{\mathrm{MP}}, \mathcal{H}), \tag{15}$$

where k is the dimensionality of the parameter vector $\mathbf{w}$. For a locally linear model with Gaussian noise, both the distributions inside the integral are Gaussian, and this integral is straightforward to perform.

As reviewed in MacKay (1992a), the most probable value of α satisfies a simple implicit equation,

$$\frac{1}{\alpha_{\mathrm{MP}}} = \frac{\sum_1^k w_i^2}{\gamma} \tag{16}$$

where w_i are the components of the vector $\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$ and γ is the *number of well–determined parameters*, which can be expressed in terms of the eigenvalues λ_a of the matrix $\beta\nabla\nabla E_D(\mathbf{w})$:

$$\gamma = k - \alpha \mathrm{Trace}\mathbf{\Sigma} = \sum_1^k \frac{\lambda_a}{\lambda_a + \alpha}. \tag{17}$$

This quantity is a number between 0 and k. Recalling that α can be interpreted as the variance σ_w^2 of the distribution from which the parameters w_i come, we see that equation (16) corresponds to an intuitive prescription for a variance estimator. The idea is that we are estimating the variance of the distribution of w_i from only γ well–determined parameters, the other $(k-\gamma)$ having been set roughly to zero by the regularizer and therefore not contributing to the sum in the numerator.

In principle, there may be multiple optima in α, but this is not the typical case for a model well matched to the data. Under general conditions, the error bars on $\log\alpha$ are $\sigma_{\log\alpha|D} \simeq \sqrt{2/\gamma}$ (MacKay 1992a) (see section 8). Thus $\log\alpha$ is well determined by the data if $\gamma \gg 1$.

The central computation can be summarised thus:

Evidence approximation: find a self–consistent solution $\{\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}, \alpha_{\mathrm{MP}}\}$ such that $\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$ maximizes $P(\mathbf{w}|D, \alpha_{\mathrm{MP}}, \mathcal{H})$ and α_{MP} satisfies equation (16).

If one is concerned that there may be multiple optima in α, then one may explicitly evaluate the evidence as a function of α.

Justification for the evidence approximation

The central approximation in this scheme can be stated as follows: when we integrate out a parameter α, the effect for most purposes is to estimate the parameter from the data, and then constrain the parameter to that value (Box and Tiao 1973; Bretthorst 1988). When we predict an observable D_2, the predictive distribution is dominated by the value $\alpha = \alpha_{\mathrm{MP}}$. In symbols,

$$P(D_2|D, \mathcal{H}) = \int P(D_2|D, \alpha, \mathcal{H}) P(\log\alpha|D, \mathcal{H})\, d\log\alpha \simeq P(D_2|D, \alpha_{\mathrm{MP}}, \mathcal{H}). \tag{18}$$

This approximation is accurate (in a sense that will be made more precise in section 8) as long as $P(D_2|D, \alpha, \mathcal{H})$ is insensitive to changes in $\log\alpha$ on a scale of $\sigma_{\log\alpha|D}$, so that the distribution $P(\log\alpha|D, \mathcal{H})$ is effectively a delta function.

This is a well–established idea. A similar equivalence of two probability distributions arises in statistical thermodynamics. The 'canonical ensemble' over all states r of a system,

$$P(r|\beta) = \exp(-\beta E_r)/Z, \tag{19}$$

describes equilibrium with a heat bath at temperature $1/\beta$. Although the energy of the system is not fixed, the probability distribution of the energy is usually sharply peaked about the mean energy $\bar{E}$. The corresponding 'microcanonical ensemble' describes the system when it is isolated and has fixed energy:

$$P(r|E = \bar{E}) = \begin{cases} 1/\Omega & E_r \in [\bar{E} \pm \delta E/2] \\ 0 & \text{otherwise} \end{cases} . \tag{20}$$

Under these two distributions, a particular microstate r may have numerical probabilities that are completely different. For example, the most probable microstate under the canonical ensemble is always the ground state, for any temperature $1/\beta \geq 0$; whereas its probability under the microcanonical ensemble is zero. But if the system has a large number of degrees of freedom, it is well known (Reif 1965) that for most macroscopic purposes, the two distributions are indistinguishable, because most of the probability *mass* of the canonical ensemble is concentrated in the states in a small interval around $\bar{E}$.

The same reasoning justifies the evidence approximation for ill–posed problems, with particular values of $\mathbf{w}$ corresponding to microstates. If the number of well–determined parameters is large, then α, like the energy above, is well determined. This does not imply that the two densities $P(\mathbf{w}|D, \mathcal{H})$ and $P(\mathbf{w}|D, \alpha_{\text{MP}}, \mathcal{H})$ are numerically close in value, but we have no interest in the probability of the high dimensional vector $\mathbf{w}$. For practical purposes, we only care about distributions of low–dimensional quantities (*e.g.*, an individual parameter w_i or a new datum); what matters, and what is asserted here, is that when we project the distributions down in order to predict low–dimensional quantities, the approximating distribution $P(\mathbf{w}|D, \alpha_{\text{MP}}, \mathcal{H})$ puts most of its probability mass in the right place. A more precise discussion of this approximation is given in section 8.

4.3 The MAP method

The alternative procedure studied in this paper is first to integrate out α to obtain the true prior:

$$P(\mathbf{w}|\mathcal{H}) = \int d\alpha \, P(\mathbf{w}|\alpha, \mathcal{H}) P(\alpha|\mathcal{H}). \tag{21}$$

We can then write down the true posterior directly (except for its normalizing constant):

$$P(\mathbf{w}|D, \mathcal{H}) \propto P(D|\mathbf{w}, \mathcal{H}) P(\mathbf{w}|\mathcal{H}). \tag{22}$$

This posterior can be maximized to find the MAP parameters, $\mathbf{w}_{\text{MP}}$. How does this relate to the desired inferences listed at the head of this section? Not all authors describe how they intend the true posterior to be used in practical problems (*e.g.*, Wolpert (1993)); here I describe a method based on the suggestions of Buntine and Weigend (1991).

Problem A: The posterior distribution $P(\mathbf{w}|D, \mathcal{H})$ is approximated by a Gaussian distribution, fitted around the most probable parameters, $\mathbf{w}_{\text{MP}}$; to find the Hessian of the log posterior, one needs the Hessian of the log prior, derived below. [A simple evaluation of the factors on the right hand side of (22) is not a satisfactory solution of problem A, since (a) the normalizing constant is missing; (b) even if the right hand side of (22) were normalized, the ability to evaluate the local value of this density would be of little use as a summary of the distribution in the high–dimensional space; for example, the marginal distribution over one parameter w_i can only be obtained from (22) by somehow performing the marginalization integral over the other parameters.]

Problem B: An estimate of the evidence is obtained from the determinant of the covariance matrix of this Gaussian distribution.

Problem C: The parameters $\mathbf{w}_{\text{MP}}$ with error bars are used to generate predictions as in (15).

A simple example will illustrate that this approach actually gives results qualitatively similar to the evidence framework. Let us consider the weight decay prior. If we apply the improper prior over α, $P_{\text{Imp}}(\log \alpha) = 1$, and evaluate the true prior over the parameters $\mathbf{w}$, we obtain a particularly simple result:[1]

$$P_{\text{Imp}}(\mathbf{w}|\mathcal{H}) = \int_{\alpha=0}^{\infty} \frac{e^{-\alpha \sum_{i=1}^{k} w_i^2/2}}{Z_W(\alpha)} \, d\log\alpha \;\propto\; \frac{1}{(\sum_i w_i^2)^{k/2}}. \tag{23}$$

The derivative of the true log prior with respect to $\mathbf{w}$ is $-(k/\sum_i w_i^2)\mathbf{w}$. This 'weight decay' term can be directly viewed in terms of an 'effective α',

$$\frac{1}{\alpha_{\text{eff}}(\mathbf{w})} = \frac{\sum_i w_i^2}{k}. \tag{24}$$

Any maximum of the true posterior $P(\mathbf{w}|D,\mathcal{H})$ is therefore also a maximum of the conditional posterior $P(\mathbf{w}|D,\alpha,\mathcal{H})$, with α set to α_{eff}. The similarity of equation (24) to equation (16) of the evidence framework is clear. We can therefore describe the MAP method thus:

> **MAP method** (improper prior over α): find a self–consistent solution $\{\mathbf{w}_{\text{MP}}, \alpha_{\text{eff}}\}$ such that $\mathbf{w}_{\text{MP}}$ maximizes $P(\mathbf{w}|D,\alpha_{\text{eff}},\mathcal{H})$ and α_{eff} satisfies equation (24).

This procedure is suggested in (MacKay 1992c) as a 'quick and dirty' approximation to the evidence framework. What the above result shows is that it is also an exact method for locating the weights that maximize the true posterior probability density.

4.4 The effective α and the curvature resulting from a general prior over α

We have just established that, when the improper prior over α (23) is used, the MAP solution lies exactly on the 'alpha trajectory' — the graph of $\mathbf{w}_{\text{MP}|\alpha}$ — for a particular value of $\alpha = \alpha_{\text{eff}}$. This result still holds when a proper prior over α is used to define the true prior over $\mathbf{w}$ (21). The derivative of $\log P(\mathbf{w}|\mathcal{H})$ with respect to $\mathbf{w}$ is

$$\frac{\partial}{\partial \mathbf{w}} \log P(\mathbf{w}|\mathcal{H}) = \frac{\int d\alpha\, (-\alpha\mathbf{w}) \exp(-\alpha\mathbf{w}^2/2)/Z_W(\alpha)P(\alpha|\mathcal{H})}{P(\mathbf{w}|\mathcal{H})} = -\alpha_{\text{eff}}(\mathbf{w})\mathbf{w} \tag{25}$$

where the effective $\alpha(\mathbf{w})$ is:

$$\alpha_{\text{eff}}(\mathbf{w}) = \int d\alpha\, \alpha\, P(\alpha|\mathbf{w},\mathcal{H}), \tag{26}$$

and

$$P(\alpha|\mathbf{w},\mathcal{H}) = \frac{P(\mathbf{w}|\alpha,\mathcal{H})P(\alpha|\mathcal{H})}{P(\mathbf{w}|\mathcal{H})}. \tag{27}$$

So at any stationary point of the true posterior, it must be the case that

$$-\beta\frac{\partial}{\partial \mathbf{w}} E_D(\mathbf{w}) - \alpha_{\text{eff}}(\mathbf{w})\mathbf{w} = 0, \tag{28}$$

---

[1] If a uniform prior over α from 0 to ∞ is used (instead of a uniform prior over $\log\alpha$) then the exponent in equation (23) changes from $k/2$ to $(k/2+1)$.

which shows that all maxima, minima and saddle points of the true posterior lie on the alpha trajectory. In summary, optima $\mathbf{w}_{\rm MP}$ found by the MAP method can be described thus:

> **MAP method** (proper prior over α): find the self–consistent solution $\{\mathbf{w}_{\rm MP}, \alpha_{\rm eff}\}$ such that $\mathbf{w}_{\rm MP}$ maximizes $P(\mathbf{w}|D, \alpha_{\rm eff}, \mathcal{H})$ and $\alpha_{\rm eff}$ satisfies equation (26).

The curvature of the true prior over $\mathbf{w}$ is needed for evaluation of the error bars on $\mathbf{w}$ in the MAP method. The true posterior probability maximum $\mathbf{w}_{\rm MP}$ coincides with the maximum of the distribution $P(\mathbf{w}|D, \alpha_{\rm eff}, \mathcal{H})$, but the curvature of the true log posterior is not equal to the curvature of $\log P(\mathbf{w}|D, \alpha_{\rm eff}, \mathcal{H})$. By direct differentiation of the true log prior (21), we find:

$$-\nabla\nabla \log P(\mathbf{w}|\mathcal{H}) = \alpha_{\rm eff}\mathbf{I} - \sigma^2_\alpha(\mathbf{w})\mathbf{w}\mathbf{w}^{\scriptscriptstyle T}, \qquad (29)$$

where $\alpha_{\rm eff}(\mathbf{w})$ is defined in (26), and the effective variance of α is:

$$\sigma^2_\alpha(\mathbf{w}) \equiv \overline{\alpha^2}(\mathbf{w}) - \alpha_{\rm eff}(\mathbf{w})^2 \equiv \int d\alpha\, \alpha^2\, P(\alpha|\mathbf{w}, \mathcal{H}) - \left(\int d\alpha\, \alpha\, P(\alpha|\mathbf{w}, \mathcal{H})\right)^2 . \qquad (30)$$

This is an intuitive result: if α were fixed to $\alpha_{\rm eff}$, then the curvature would just be the first term in (29), $\alpha_{\rm eff}\mathbf{I}$. The fact that α is uncertain depletes the curvature in the radial direction $\hat{\mathbf{w}} = \mathbf{w}/|\mathbf{w}|$. To obtain the Hessian for the MAP method's Gaussian approximation, the curvature of the log prior in equation (29) would be added to the curvature of the log likelihood $\log P(D|\mathbf{w}, \mathcal{H})$.

4.5 Condition satisfied by typical samples

The conditions (16) and (24), satisfied by the optima $(\alpha_{\rm MP}, \mathbf{w}_{{\rm MP}|\alpha_{\rm MP}})$ and $(\alpha_{\rm eff}, \mathbf{w}_{\rm MP})$ respectively, are complemented by an additional result concerning typical samples from posterior distributions conditioned on α. The maximum $\mathbf{w}_{{\rm MP}|\alpha}$ of a Gaussian distribution is not typical of that distribution: the maximum has an atypically small value of $\mathbf{w}^{\scriptscriptstyle T}\mathbf{w}$, because, as discussed in section 6, nearly all of the mass of a Gaussian is in a shell at some distance surrounding the maximum.

Consider samples $\{\mathbf{w}\}$ from the Gaussian posterior distribution with α fixed to $\alpha_{\rm MP}$, $P(\mathbf{w}|D, \alpha_{\rm MP}, \mathcal{H})$. The average value of $\mathbf{w}^{\scriptscriptstyle T}\mathbf{w} = \sum_i w_i^2$ for these samples satisfies:

$$\alpha_{\rm MP} = \frac{k}{\langle \sum_i w_i^2 \rangle_{|D,\alpha_{\rm MP}}}. \qquad (31)$$

Proof: The deviation $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{{\rm MP}|\alpha_{\rm MP}}$ is Gaussian distributed with $\Delta\mathbf{w}\Delta\mathbf{w}^{\scriptscriptstyle T} = \mathbf{\Sigma}$. So $\alpha_{\rm MP}\langle \sum_i w_i^2 \rangle_{|D,\alpha_{\rm MP}} = \alpha_{\rm MP}(\mathbf{w}_{{\rm MP}|\alpha_{\rm MP}} + \Delta\mathbf{w})^{\scriptscriptstyle T}(\mathbf{w}_{{\rm MP}|\alpha_{\rm MP}} + \Delta\mathbf{w}) = \alpha_{\rm MP}\mathbf{w}^2_{{\rm MP}|\alpha_{\rm MP}} + \alpha_{\rm MP}{\rm Trace}\mathbf{\Sigma} = k$, using equations (16) and (17).

Thus a typical sample from the evidence approximation prefers just the same value of α as does the evidence $P(D|\alpha, \mathcal{H})$, in the sense that if one were to draw samples $\{\mathbf{w}\}$ from

$P(\mathbf{w}|D, \alpha_{\mathrm{MP}}, \mathcal{H})$ and then estimate α so as to maximize the probability of those samples, α would be set to α_{MP}.

5 Pros and cons

The algorithms for finding the evidence framework's $\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$ and the MAP method's $\mathbf{w}_{\mathrm{MP}}$ have been seen to be very similar. Is there any significant distinction to be drawn between these two aproaches?

The MAP method has the advantage that it involves no approximations until after we have found the MAP parameters $\mathbf{w}_{\mathrm{MP}}$; in contrast, the evidence framework approximates an integral over α.

In the MAP method the integrals over α and β need only be performed once and can then be used repeatedly for different data sets; in the evidence framework, each new data set has to receive individual attention, with a sequence of (Gaussian) integrations being performed each time α and β are optimized.

So why not always integrate out hyperparameters whenever possible? Let us answer this question by magnifying the systematic differences between the two approaches. With sufficient magnification it will become evident to the intuition that the approximation of the evidence framework is superior to the MAP approximation.

The distinction between $\mathbf{w}_{\mathrm{MP}}$ and $\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$ is similar to that between the two estimators of standard deviation on a calculator, σ_N and σ_{N-1}, the former being the biased maximum likelihood estimator, whereas the latter is unbiased. The true posterior distribution has a skew peak, so that the MAP parameters are not representative of the whole posterior distribution. This is best illustrated by an example.

5.1 The widget example

A collection of widgets $i = 1..k$ have a property called 'wodge', w_i, which we measure, widget by widget, in noisy experiments with a known noise level $\sigma_\nu = 1.0$. Our model for these quantities is that they come from a Gaussian prior $P(w_i|\alpha, \mathcal{H})$, where $\alpha = 1/\sigma_w^2$ is not known. Our prior for this variance is flat over $\log \sigma_w$ from $\sigma_w = 0.1$ to $\sigma_w = 10$.

Scenario 1. Suppose four widgets have been measured and give the following data: $\{d_1, d_2, d_3, d_4\} = \{2.2,\ -2.2,\ 2.8,\ -2.8\}$. The task (problem A) is to infer the wodges of these four widgets, *i.e.*, to produce a representative $\mathbf{w}$ with error bars.

Evidence framework: using equation (16) iteratively we find $\alpha_{\mathrm{MP}} = 0.19$, $\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}} = \{1.9, -1.9, 2.4, -2.4\}$, each with error bars ± 0.9.

MAP method: We can identify maxima of the true posterior by finding attracting fixed points of equation (26) using a computer algebra system. For scenario 1, there are two attracting fixed points, corresponding to two maxima like those in figure 1(f): the fixed point with the smaller value of α_{eff} has $\alpha_{\text{eff}} = 0.25$, $\mathbf{w}_{\text{MP}} = \{1.8, -1.8, 2.2, -2.2\}$, each with error bars ± 0.9. The other maximum is located at $\mathbf{w}_{\text{MP}} = \{0.03, -0.03, 0.04, -0.04\}$ and is associated with $\alpha_{\text{eff}} = 65$; here, each parameter has error bars ± 0.1.

Concentrating our attention on the sensible maximum, we might note that $\mathbf{w}_{\text{MP}|\alpha_{\text{MP}}}$ is slightly less regularized than $\mathbf{w}_{\text{MP}}$, but there is not much disagreement between the two methods when all the parameters are well–determined.

Scenario 2. Suppose in addition to the four measurements above we are now informed that there are an additional four widgets that have been measured with a much less accurate instrument, having $\sigma'_\nu = 100.0$. Thus we now have both well–determined and ill–determined parameters, as in a typical ill–posed problem. The data from these measurements were a string of uninformative values, $\{d_5, d_6, d_7, d_8\} = \{100,\ -100,\ 100,\ -100\}$.

We are again asked to infer the wodges of the widgets. Intuitively, we would like our inferences about the well–measured widgets to be negligibly affected by this vacuous information about the poorly–measured widgets, just as the true Bayesian predictive distributions are unaffected. But clearly with $k = 8$, the difference between k and γ in equations (16) and (24) is going to become significant. The value of α_{eff} will be substantially greater than that of α_{MP}.

In the evidence framework the value of γ is almost exactly the same, since each of the ill–determined parameters has $\lambda_i \simeq 0$ and adds nothing to the number of well–determined parameters (17). So the value of α_{MP} and the predictive distributions are unchanged.

In contrast, the MAP solution changes drastically. The maximum associated with $\alpha_{\text{eff}} = 0.25$ vanishes, and the only maximum of the true posterior probability is the spike $\mathbf{w}_{\text{MP}}$ which is squashed close to zero. Solving equation (26) in a computer algebra system, we find: $\alpha_{\text{eff}} = 79.5$, $\mathbf{w}_{\text{MP}} = \{0.03, -0.03, 0.03, -0.03, 0.0001, -0.0001, 0.0001, -0.0001\}$, with marginal error bars on all eight parameters $\sigma_{w|D} = 0.11$.

Thus the MAP Gaussian approximation is terribly biased towards zero. The final disaster of this approach is that the error bars on the parameters are also very small.

This is not a contrived example. It contains the basic feature of ill–posed problems: that there are both well–determined and poorly–determined parameters. To aid comprehension, the two sets of parameters are separated. This example can be transformed into a typical ill–posed problem simply by rotating the basis to mix the parameters together. In neural networks, a pair of scenarios identical to those discussed above can arise if there are a large number of poorly determined parameters which have been set to zero by the regularizer, and we consider two scenarios. In scenario 1, the network is 'pruned', removing the ill–determined parameters. In scenario 2, the parameters are retained, and take on their most probable value, zero. In each case, what is the optimal setting of the weight decay rate α (assuming the traditional regularizer $\mathbf{w}^{\mathsf{T}}\mathbf{w}/2$)? We would expect the answer to be unchanged. Yet the MAP method effectively sets α to a much larger value in the second scenario.

The MAP method may locate the true posterior maximum, but it fails to capture most of the true probability mass. Figure 2 conveys in two dimensions this difference between the MAP Gaussian approximation and the Gaussian approximation given by evidence maximization. The larger the number of dimensions we are in, the higher the density in the skew peak becomes, and the more it dominates the maximization of the density. But the mass associated with the peak is not increasing.

If we maximize a probability density which is equal to a superposition of Gaussians, the location of the maximum will be chiefly determined by the locations of the Gaussians with *smallest standard deviation*, rather than the locations of the Gaussians with greatest probability mass.

6 Inference in many dimensions

In many dimensions, therefore, new intuitions are needed.

Nearly all of the volume of a k–dimensional hypersphere is in a thin shell near its surface. For example, in 1000 dimensions, 90% of a hypersphere of radius 1.0 is within a depth of 0.0023 of its surface. A central core of the hypersphere, with radius 0.5, contains less than $1/10^{300}$ of the volume.

This has an important effect on high–dimensional probability distributions. Consider a Gaussian distribution $P(\mathbf{w}) = (1/\sqrt{2\pi}\,\sigma_w)^k \exp(-\sum_1^k w_i^2/2\sigma_w^2)$. Nearly all of the probability mass of a Gaussian is in a thin shell of radius $r = \sqrt{k}\sigma_w$ and of thickness $\propto r/\sqrt{k}$. For example, in 1000 dimensions, 90% of the mass of a Gaussian with $\sigma_w = 1$ is in a shell of radius 31.6 and thickness 2.8. However, the probability *density* at the origin is $e^{k/2} \simeq 10^{217}$ times bigger than the density at this shell where most of the probability mass is.

Now consider two Gaussian densities in 1000 dimensions which differ in radius σ_w by just 1%, and which contain equal total probability mass. The maximum probability density is greater at the centre of the Gaussian with smaller σ_w by a factor of $\sim \exp(0.01k) \simeq 20{,}000$.

A typical true posterior distribution for an ill–posed problem is a weighted superposition of Gaussians with varying means and standard deviations, so the true posterior has a skew peak, with the maximum of the probability density located near the mean of the Gaussian distribution that has the smallest standard deviation, not the Gaussian with the greatest weight. Thus a Gaussian fitted at the MAP parameters is a bad approximation to the distribution: it is in the wrong place, and its error bars are far too small. In contrast, the evidence approximation is given by selecting from the superposition of Gaussians the Gaussian component which has the biggest weight, and which thus captures most of the probability mass of the true posterior.

In summary, probability density maxima often have very little associated probability mass — even though the value of the probability density there may be immense — because they have so little associated volume. If a distribution is composed of a mixture of Gaussians with different σ_w, the probability density maxima are strongly dominated by smaller values of σ_w. This is why the MAP method finds a silly solution in the widget example. Recall that in the case of a thermodynamic system in its canonical ensemble (section 4.2), the state of the system that has maximum probability density is the ground state, regardless of the temperature of the system.

Thus the locations of probability density maxima in many dimensions are generally misleading and irrelevant. Probability densities should only be maximized if there is good reason to believe that the location of the maximum conveys useful information about the whole distribution, *e.g.*, if the distribution is approximately Gaussian.

7 Relationship between evidence maximization and 'ensemble learning'

A novel approach to the approximation of Bayesian inference has recently been introduced by Hinton and van Camp (1993). I will first review the concept of *Ensemble Learning by Free Energy Minimization* for a simplified model with the hyperparameter α omitted.

In traditional approaches to neural networks, a *single* parameter vector $\mathbf{w}$ is optimized by maximum likelihood or penalized maximum likelihood. In the Bayesian interpretation, these optimized parameters are viewed as defining the mode of a posterior probability distribution $P(\mathbf{w}|D,\mathcal{H})$ (given data D and model assumptions $\mathcal{H}$), which can be approximated, with a Gaussian distribution for example, in order to obtain predictive distributions and optimize model control parameters.

The new concept introduced by Hinton and van Camp (1993) is to work 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 θ) so that it approximates the posterior distribution of the parameters $P(\mathbf{w}|D,\mathcal{H})$ as well as possible. The objective function chosen to measure the quality of the approximation is a *variational free energy* (Feynman 1972),

$$F(\theta) = -\int d^k\mathbf{w}\; Q(\mathbf{w};\theta)\log\frac{P(D|\mathbf{w},\mathcal{H})P(\mathbf{w}|\mathcal{H})}{Q(\mathbf{w};\theta)}. \tag{32}$$

The free energy $F(\theta)$ is bounded below by $-\log P(D|\mathcal{H})$ and only attains this value for $Q(\mathbf{w};\theta) = P(\mathbf{w}|D,\mathcal{H})$. $F(\theta)$ can be viewed as the sum of $-\log P(D|\mathcal{H})$ and the Kullback–Leibler divergence between $Q(\mathbf{w};\theta)$ and $P(\mathbf{w}|D,\mathcal{H})$. For certain models and certain approximating distributions, this free energy, and its derivatives with respect to the ensemble's parameters, can be evaluated. [This is the main reason for choosing the objective function $F(\theta)$ rather than some other measure of distance between $Q(\mathbf{w};\theta)$ and $P(\mathbf{w}|D,\mathcal{H})$.] A longer review of Ensemble Learning including references to applications may be found in (MacKay 1995).

In this section I demonstrate that a free energy approximation for the model studied in this paper reproduces the method of the evidence framework precisely. This result is not viewed as a justification for the evidence framework, but rather as giving insight into the nature of the approximations made by this framework.

7.1 Free energy approximation for a model with a hyperparameter

Let us assume, in addition to the likelihood function and prior over $\mathbf{w}$ of equations (3) and (4), that the prior over α is a gamma distribution, $P(\alpha|\mathcal{H}) = \Gamma(\alpha; b_\alpha, c_\alpha)$, where this notation means:

$$\Gamma(\alpha; b_\alpha, c_\alpha) = \frac{1}{\Gamma(c_\alpha)} \frac{\alpha^{c_\alpha - 1}}{b_\alpha^{c_\alpha}} \exp\left(-\frac{\alpha}{b_\alpha}\right), 0 \leq \alpha < \infty. \tag{33}$$

This distribution has mean $b_\alpha c_\alpha$ and variance $b_\alpha^2 c_\alpha$.

Let us consider approximating the joint distribution of $\mathbf{w}$ and α given the data,

$$P(\mathbf{w}, \alpha|D, \mathcal{H}) = \frac{P(D|\mathbf{w}, \mathcal{H})P(\mathbf{w}|\alpha, \mathcal{H})P(\alpha|\mathcal{H})}{P(D|\mathcal{H})}, \tag{34}$$

by a distribution $Q(\mathbf{w}, \alpha)$. I make one assumption only, namely we will use an approximating distribution that is constrained to have the separable form $Q(\mathbf{w}, \alpha) = Q_{\mathbf{w}}(\mathbf{w})Q_\alpha(\alpha)$. *No functional form for these distributions is assumed.* [The reason for choosing this separable form is that this is the most complex approximating distribution for which the computations are tractable — we don't necessarily believe the posterior density is approximately separable.] We write down a variational free energy,

$$F(Q) = -\int d\mathbf{w}\, d\alpha\, Q_{\mathbf{w}}(\mathbf{w})Q_\alpha(\alpha) \log \frac{P(D|\mathbf{w}, \mathcal{H})P(\mathbf{w}|\alpha, \mathcal{H})P(\alpha|\mathcal{H})}{Q_{\mathbf{w}}(\mathbf{w})Q_\alpha(\alpha)}. \tag{35}$$

This functional is bounded below by the evidence for the model thus: $F \geq -\log P(D|\mathcal{H})$, with equality if and only if $Q(\mathbf{w}, \alpha) = P(\mathbf{w}, \alpha|D, \mathcal{H})$. We can find the optimal separable distribution Q by considering separately the optimization of F over $Q_{\mathbf{w}}(\mathbf{w})$ for fixed $Q_\alpha(\alpha)$, and then the optimization of $Q_\alpha(\alpha)$ for fixed $Q_{\mathbf{w}}(\mathbf{w})$.

7.2 Optimization of $Q_{\mathbf{w}}(\mathbf{w})$

As a functional of $Q_{\mathbf{w}}(\mathbf{w})$, F is:

$$\begin{aligned} F &= -\int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \left[\int d\alpha\, Q_\alpha(\alpha) \log P(\mathbf{w}|\alpha) + \log P(D|\mathbf{w}, \mathcal{H}) - \log Q(\mathbf{w})\right] + \text{const} && (36) \\ &= \int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \left[\int d\alpha\, Q_\alpha(\alpha) \alpha \frac{1}{2}\mathbf{w}\mathbf{w}^{\mathsf{T}} + \beta E_D(\mathbf{w}) + \log Q(\mathbf{w})\right] + \text{const.}' && (37) \end{aligned}$$

The dependence on Q_α thus collapses down to a dependence simply on the mean value of α,

$$\bar{\alpha} \equiv \int d\alpha\, Q_\alpha(\alpha)\alpha. \tag{38}$$

$$F = \int d\mathbf{w}\, Q_\mathbf{w}(\mathbf{w}) \left[\bar{\alpha}\frac{1}{2}\mathbf{w}\mathbf{w}^{\mathsf{T}} + \beta E_D(\mathbf{w}) + \log Q(\mathbf{w}) \right] + \text{const.}' \tag{39}$$

Noting that the $\mathbf{w}$–dependent terms $-\bar{\alpha}\frac{1}{2}\mathbf{w}\mathbf{w}^{\mathsf{T}} - \beta E_D(\mathbf{w})$ are the log of a posterior distribution, and using the theorem that a divergence $\int Q \log(Q/P)$ is minimized by setting $Q = P$, we can immediately write down the distribution $Q_\mathbf{w}(\mathbf{w})$ that minimizes this expression. For given data D and Q_α, the optimizing distribution $Q_\mathbf{w}^{\text{opt}}(\mathbf{w})$ is a Gaussian identical to the posterior distribution for a particular value of $\alpha = \bar{\alpha}$.

$$Q_\mathbf{w}^{\text{opt}}(\mathbf{w}) = P(\mathbf{w}|D, \bar{\alpha}, \mathcal{H}) = \text{Normal}(\mathbf{w}_{\text{MP}|\bar{\alpha}}, \mathbf{\Sigma}). \tag{40}$$

7.3 Optimization of $Q_\alpha(\alpha)$

As a functional of $Q_\alpha(\alpha)$, F is:

$$\begin{aligned} F &= -\int d\alpha\, Q_\alpha(\alpha) \left[\int d\mathbf{w}\, Q_\mathbf{w}(\mathbf{w}) \log P(\mathbf{w}|\alpha, \mathcal{H}) + \log P(\alpha|\mathcal{H}) - \log Q_\alpha(\alpha) \right] + \text{const} & (41) \\ &= \int d\alpha\, Q_\alpha(\alpha) \left[\tfrac{\alpha}{2} \int d\mathbf{w}\, Q_\mathbf{w}(\mathbf{w}) \mathbf{w}^{\mathsf{T}}\mathbf{w} - \tfrac{k}{2}\log\alpha - (c_\alpha - 1)\log\alpha + \tfrac{\alpha}{b_\alpha} + \log Q_\alpha(\alpha) \right] & (42) \\ &= \int d\alpha\, Q_\alpha(\alpha) \left[\left(\frac{1}{2}\mathbf{w}_{\text{MP}|\bar{\alpha}}{}^{\mathsf{T}}\mathbf{w}_{\text{MP}|\bar{\alpha}} + \frac{1}{2}\text{Trace}\mathbf{\Sigma} + \frac{1}{b_\alpha} \right) \alpha \right. \\ &\qquad \left. - \left(\frac{k}{2} + c_\alpha - 1 \right) \log\alpha + \log Q_\alpha(\alpha) \right] + \text{const.}' & (43) \end{aligned}$$

where c_α, b_α are the parameters of the gamma prior on α. Here, the α–dependent expression in the brackets can be recognized as the log of a gamma distribution, giving as the optimal distribution that minimizes F for fixed $Q_\mathbf{w}$:

$$Q_\alpha^{\text{opt}}(\alpha) = \Gamma(\alpha; b', c') \tag{44}$$

where

$$\begin{aligned} 1/b' &= 1/b_\alpha + \tfrac{1}{2}\mathbf{w}_{\text{MP}|\bar{\alpha}}{}^{\mathsf{T}}\mathbf{w}_{\text{MP}|\bar{\alpha}} + \tfrac{1}{2}\text{Trace}\mathbf{\Sigma} \\ c' &= k/2 + c_\alpha \end{aligned} . \tag{45}$$

This completes our derivation of the free energy optimization. The optimal approximating distribution is given by finding the gamma distribution for α and the normal distribution for $\mathbf{w}$ that satisfy the simultaneous equations (38), (40) and (45).

7.4 Comparison with the evidence framework

To understand this result we complete the loop by evaluating the mean $\bar{\alpha}'$ for this optimized gamma distribution, which is:

$$\bar{\alpha}' = b'c' = \frac{\frac{k}{2} + c_\alpha}{\frac{1}{b_\alpha} + \frac{1}{2}\mathbf{w}_{\mathrm{MP}|\bar{\alpha}}{}^{\mathsf{T}}\mathbf{w}_{\mathrm{MP}|\bar{\alpha}} + \frac{1}{2}\mathrm{Trace}\boldsymbol{\Sigma}}. \tag{46}$$

In the special case of an uninformative prior on α ($c_\alpha \to 0$ and $\frac{1}{b_\alpha} \to 0$) we obtain:

$$\bar{\alpha}' = \frac{k}{\mathbf{w}_{\mathrm{MP}|\bar{\alpha}}{}^{\mathsf{T}}\mathbf{w}_{\mathrm{MP}|\bar{\alpha}} + \mathrm{Trace}\boldsymbol{\Sigma}}. \tag{47}$$

Is this the same optimal α as that found by evidence maximization?[2] The answer is yes. Substituting (equation 16) $\mathbf{w}^{\mathsf{T}}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}} = \gamma/\alpha_{\mathrm{MP}}$, and using $\gamma = k - \alpha\mathrm{Trace}\boldsymbol{\Sigma}$, we find that if we set $\alpha = \bar{\alpha} = \alpha_{\mathrm{MP}}$ on the right hand side we obtain

$$\bar{\alpha}' = \frac{k}{\gamma/\bar{\alpha} + (k-\gamma)/\bar{\alpha}} = \bar{\alpha}. \tag{48}$$

Thus any optimum of the evidence approximation also corresponds to a minimum of the free energy. This relationship is only exact in the case of the linear regression model studied in this paper. If the likelihood is non–Gaussian then $P(\mathbf{w}|D, \bar{\alpha}, \mathcal{H})$ is no longer a Gaussian, so the step at equation (40) does not follow.

Intuition for the relationship between evidence maximization and ensemble learning

These two approaches give complementary views of the task of inferring α given the data.

In the evidence framework we examine the optimized value of $\mathbf{w}$, $\mathbf{w}_{\mathrm{MP}|\alpha}$, and think of $(\mathbf{w}_{\mathrm{MP}|\alpha})^2$ as giving information about the variance σ_w^2 of the prior distribution of $\mathbf{w}$. The maximum likelihood estimator of σ_w^2 would be $\sigma^2_{w(\mathrm{ML})} = (\mathbf{w}_{\mathrm{MP}|\alpha})^2/k$, but the evidence framework modifies this estimator to take into account the fact that some of the k parameters have not been determined by the data, and have effectively been set to zero by the prior. Thus the evidence–maximizing estimate replaces k by the effective number of well determined parameters γ: $\sigma^2_{w(\mathrm{MP})} = (\mathbf{w}_{\mathrm{MP}|\alpha})^2/\gamma$.

The free energy minimization approach is like an EM algorithm (Dempster *et al.* 1977), in which we wish to find the most probable α and do this by introducing an E–step in which a distribution over $\mathbf{w}$ is obtained (Neal and Hinton 1998). This distribution takes into account the $k - \gamma$ ill–determined parameters by assigning each of them a variance of σ_w^2 in the matrix $\boldsymbol{\Sigma}$. Then when the M–step occurs, finding the optimal α, the maximum likelihood equation $\sigma^2_{w\,(\mathrm{ML})} = (\mathbf{w}_{\mathrm{MP}|\alpha})^2/k$ is modified by adding these variance terms to the numerator: $\sigma^2_{w\,(\mathrm{FE})} = \left[(\mathbf{w}_{\mathrm{MP}|\alpha})^2 + \mathrm{Trace}\boldsymbol{\Sigma}\right]/k$.

---

[2] Or 'are *these* the same as *those* found by evidence maximization?' if there are multiple optima.

Thus evidence maximization decrements the denominator of the equation $\sigma^2_{w\,(\mathrm{ML})} = (\mathbf{w}_{\mathrm{MP}|\alpha})^2/k$ to take into account the smallness of the ill–determined parameters, whereas free energy minimization increments the numerator to take into account their variability. As we have seen, the two formulae converge on the identical result.

Further comments

There are two small differences between evidence maximization and free energy minimization.

1. The variance of the optimized gamma distribution for α is, in the limit of the uninformative prior,
$$\mathrm{var}(\alpha) = {b'}^2 c' = 2k/(k/\bar{\alpha})^2 = \bar{\alpha}^2/k \tag{49}$$
so that $\log\alpha$ has standard error $\sqrt{2/k}$. This contrasts with the result $\sqrt{2/\gamma}$ from the evidence framework.
2. This free energy approximation for $Q_{\mathbf{w}}(\mathbf{w})$ fails to produce the small order correction terms to be identified in section 8.3, which arise because of the uncertainty in α. This failure is caused by the separability assumption in the ensemble approximation.

8 Conditions for the evidence approximation

We have observed in section 5.1 that the MAP method can lead to absurdly biased answers if there are many ill–determined parameters. In contrast, I now discuss conditions under which the evidence approximation works. I discuss again the case of linear models with Gaussian probability distributions.

What do we care about when we approximate a complex probability distribution by a simple one? My definition of a good approximation is a practical one, concerned with (A) estimating parameters; (B) estimating the evidence accurately; and (C) getting the predictive mass in the right place. Estimation of individual parameters (A) is a special case of prediction (C), so in the following I will address only problems (C) and (B).

For convenience let us work in the eigenvector basis where the prior over $\mathbf{w}$ (given α) and the likelihood are both diagonal Gaussian functions. The curvature of the log likelihood is represented by eigenvalues $\{\lambda_a\}$. For a typical ill–posed problem these eigenvalues vary in value by several orders of magnitude. Without loss of generality let us assume k data measurements $\{d_a\}$, such that $d_a = \sqrt{\lambda_a} w_a + \nu$, where the noise standard deviation is $\sigma_\nu = 1$.

We define the probability distribution of everything by the product of the distributions:

$$P(\log\alpha|\mathcal{H}) = \frac{1}{\log(\alpha_{\max}/\alpha_{\min})}, \quad P(\mathbf{w}|\alpha,\mathcal{H}) = \left(\frac{\alpha}{2\pi}\right)^{k/2} \exp\left(-\frac{1}{2}\alpha\sum_{1}^{k} w_a^2\right), \text{ and} \tag{50}$$

$$P(D|\mathbf{w},\mathcal{H}) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\sum_{1}^{k}\left(\sqrt{\lambda_a}w_a - d_a\right)^2\right\}. \tag{51}$$

The discussion proceeds in two steps. First, the posterior distribution over α must have a single sharp peak at α_{MP}. No general guarantee can be given for this to be the case, but various pointers are given. Second, given a sharp Gaussian posterior over $\log\alpha$, it is proved that the evidence approximation introduces negligible error.

8.1 Concentration of $P(\log\alpha|D,\mathcal{H})$ in a single maximum

Condition 1 *In the posterior distribution over* $\log\alpha$*, all the probability mass should be contained in a single sharp maximum.*

For this to hold, several sub–conditions are needed. If there is any doubt whether these conditions are sufficient, it is straightforward (at least in the case of a single hyperparameter) to iterate all the way down the α trajectory, explicitly evaluating $P(\log\alpha|D,\mathcal{H})$.

The prior over α must be such that the posterior has negligible mass at $\log\alpha \to \pm\infty$. In cases where the signal to noise ratio of the data is very low, there may be a significant tail in the evidence for large α. There may even be no maximum in the evidence, in which case the evidence framework gives singular behaviour, with α going to infinity. But often the tails of the evidence are small, and contain negligible mass if our prior over $\log\alpha$ has cutoffs at some $\alpha_{\min}$ and $\alpha_{\max}$ surrounding α_{MP}. For each data analysis problem, one may evaluate the critical $\alpha_{\max}$ above which the posterior would be measurably affected by the large α tail of the evidence (Gull 1989). Often, as Gull points out, this critical value of $\alpha_{\max}$ has bizarrely large magnitude.

Even if a flat prior between appropriate $\alpha_{\min}$ and $\alpha_{\max}$ is used, it is possible in principle for the posterior $P(\log\alpha|D,\mathcal{H})$ to be multi–modal. However this is not expected when the model space is well matched to the data. Examples of multi–modality only arise if the data are grossly at variance with the model. For example, if some large eigenvalue measurements give small $d_{a(l)}$, and some measurements with small eigenvalue give large $d_{a(s)}$, then the posterior over α can have two peaks, one at large α which nicely explains $d_{a(l)}$, but must attribute $d_{a(s)}$ to unusually large amounts of noise, and one at small α which nicely explains $d_{a(s)}$, but must attribute $d_{a(l)}$ to $w_{a(l)}$ being unexpectedly close to zero. This concept may be formalized into a quantitative test as follows.

If we accept the model, then we believe that there is a true value of $\alpha = \alpha_{\text{T}}$, and that given α_{T}, the data measurements d_a are the sum of two independent Gaussian variables $\sqrt{\lambda_a}w_a$

and ν_a, so that $P(d_a|\alpha_{\mathrm{T}},\mathcal{H}) = \mathrm{Normal}(0,\sigma^2_{a|\alpha_T})$, where $\sigma^2_{a|\alpha_T} = \frac{\lambda_a}{\alpha_T}+1$. The expectation of d_a^2 is $\langle d_a^2\rangle = \frac{\lambda_a}{\alpha_T}+1$. We therefore expect that there is an α_{T} such that the quantities $\{d_a^2/\sigma^2_{a|\alpha_T}\}$ are independently distributed like χ^2 with one degree of freedom.

Definition 1 *A data set* $\{d_a\}$ *is* grossly at variance *with the model for a given value of* α *at significance level* τ*, if any of the quantities* $j_a = d_a^2/(\frac{\lambda_a}{\alpha}+1)$ *is not in the interval* $[e^{-\tau}, 1+\tau]$.

It is conjectured that if we find a value of $\alpha = \alpha_{\mathrm{MP}}$ which locally maximizes the evidence, and with which the data are not grossly at variance, then there are no other maxima over α.

Conversely, if the data are grossly at variance with a local maximum α_{MP}, then there may be multiple maxima in α, and the evidence approximation may be inaccurate. In these circumstances one might also suspect that the entire model is inadequate in some way.

Assuming that $P(\log\alpha|D,\mathcal{H})$ has a single maximum over $\log\alpha$, how sharp is it expected to be? I now establish conditions under which the $P(\log\alpha|D,\mathcal{H})$ is locally Gaussian and sharp.

Definition 2 *The symbol* n_e *is defined by:*

$$n_e \equiv \sum_a \frac{4\lambda_a\alpha_{\mathrm{MP}}}{(\lambda_a+\alpha_{\mathrm{MP}})^2}. \tag{52}$$

This is a measure of the number of eigenvalues λ_a *within approximately e–fold of* α_{MP}.

In the following, I will assume that $n_e \ll \gamma$, but this condition is not essential for the evidence approximation to be valid. If $n_e \ll \gamma$, and the data are not grossly at variance with α_{MP}, then the Taylor expansion of $\log P(\alpha|D,\mathcal{H})$ about $\alpha = \alpha_{\mathrm{MP}}$ is:

$$\left.\frac{\partial \log P(D|\alpha,\mathcal{H})}{\partial \log\alpha}\right|_{\alpha_{\mathrm{MP}}} = \frac{1}{2}\left(\gamma - \alpha \mathbf{w}^2_{\mathrm{MP}|\alpha_{\mathrm{MP}}}\right) = 0 \tag{53}$$

$$\left.\frac{\partial^2 \log P(D|\alpha,\mathcal{H})}{\partial (\log\alpha)^2}\right|_{\alpha_{\mathrm{MP}}} \simeq -\alpha \mathbf{w}^2_{\mathrm{MP}|\alpha_{\mathrm{MP}}} = -\frac{\gamma}{2} \tag{54}$$

$$\left.\frac{\partial^3 \log P(D|\alpha,\mathcal{H})}{\partial (\log\alpha)^3}\right|_{\alpha_{\mathrm{MP}}} \simeq -\alpha \mathbf{w}^2_{\mathrm{MP}|\alpha_{\mathrm{MP}}} = -\frac{\gamma}{2}. \tag{55}$$

The first derivative is exact, assuming that the eigenvalues λ_a are independent of α, which is true in the case of a Gaussian prior on $\mathbf{w}$ (Bryan 1990). The second and third derivatives are approximate, with terms proportional to n_e being omitted. Now, if $\gamma \gg 1$, then the second derivative is relatively large, and the third derivative is relatively small (even though they are numerically equal), since in the expansion $P(l) = \exp(-\frac{c}{2}l^2 + \frac{d}{6}l^3 + \ldots)$, the second term gives a negligible perturbation for $l \sim c^{-1/2}$ if $d \ll c^{3/2}$. In this case, since $d \simeq c \simeq \gamma \gg 1$, the perturbation introduced by the higher order terms is $O(\gamma^{-1/2})$. Thus the posterior distribution over $\log\alpha$ has a maximum that is both locally Gaussian and sharp if $\gamma \gg 1$ and $n_e \ll \gamma$. The expression for the evidence (14) follows.

8.2 Error of low–dimensional predictive distributions

I will now assume that the posterior distribution $P(\log\alpha|D,\mathcal{H})$ is Gaussian with standard deviation $\sigma_{\log\alpha|D} = 1/\sqrt{\kappa\gamma}$, with $\kappa\gamma \gg 1$, and $\kappa = O(1)$.

Theorem 1 *Consider a scalar which depends linearly on* $\mathbf{w}$, $y = \mathbf{g}\cdot\mathbf{w}$. *The evidence approximation's predictive distribution for* y *is close to the exact predictive distribution, for nearly all projections* $\mathbf{g}$. *In the case* $\mathbf{g} = \mathbf{w}$, *the error (measured by a cross–entropy) is of order* $\sqrt{n_e/\kappa\gamma}$. *For all* $\mathbf{g}$ *perpendicular to this direction, the error is of order* $\sqrt{1/\kappa\gamma}$.

A similar result is expected still to hold when the dimensionality of y is greater than one, provided that it is much less than $\sqrt{\gamma}$.

Proof: At 'level 1', we infer $\mathbf{w}$ for a fixed value of α:

$$P(\mathbf{w}|D,\alpha,\mathcal{H}) \propto \exp\left\{-\frac{1}{2}\sum_a(\lambda_a+\alpha)\left(w_a - \frac{\sqrt{\lambda_a}d_a}{\lambda_a+\alpha}\right)^2\right\}. \tag{56}$$

The most probable $\mathbf{w}$ given this value of α is: $w_a^{\text{MP}|\alpha} = \sqrt{\lambda_a}d_a/(\lambda_a+\alpha)$. The posterior distribution is Gaussian about this most probable $\mathbf{w}$. We introduce a *typical* $\mathbf{w}$, that is, a sample from the posterior for a particular value of α:

$$w_a^{\text{TYP}|\alpha} = \frac{\sqrt{\lambda_a}d_a}{\lambda_a+\alpha} + \frac{r_a}{\sqrt{\lambda_a+\alpha}}, \tag{57}$$

where r_a is a sample from Normal(0,1).

Now, assuming that $\log\alpha$ has a Gaussian posterior distribution with standard deviation $1/\sqrt{\kappa\gamma}$, a typical α, *i.e.*, a sample from this posterior, is given to leading order by

$$\alpha^{\text{TYP}} = \alpha_{\text{MP}}\left(1 + \frac{s}{\sqrt{\kappa\gamma}}\right), \tag{58}$$

where s is a sample from Normal(0,1). We now substitute this α^{TYP} into (57) and obtain a typical $\mathbf{w}$ from the true posterior distribution, which depends on $k+1$ random variables $\{r_a\}, s$. We expand each component of this vector $\mathbf{w}^{\text{TYP}}$ in powers of $1/\gamma$:

$$\begin{aligned} w_a^{\text{TYP}} &= \frac{\sqrt{\lambda_a}d_a}{\lambda_a+\alpha_{\text{MP}}}\left(1 - \frac{s}{\sqrt{\kappa\gamma}}\frac{\alpha_{\text{MP}}}{\lambda_a+\alpha_{\text{MP}}} + \frac{s^2}{\kappa\gamma}\frac{\alpha_{\text{MP}}^2}{(\lambda_a+\alpha_{\text{MP}})^2} + \ldots\right) + \\ & \frac{r_a}{\sqrt{\lambda_a+\alpha_{\text{MP}}}}\left(1 - \frac{1}{2}\frac{s}{\sqrt{\kappa\gamma}}\frac{\alpha_{\text{MP}}}{\lambda_a+\alpha_{\text{MP}}} + \frac{3}{8}\frac{s^2}{\kappa\gamma}\frac{\alpha_{\text{MP}}^2}{(\lambda_a+\alpha_{\text{MP}})^2}\ldots\right). \end{aligned} \tag{59}$$

We now examine the mean and variance of $y^{\text{TYP}} = \sum_a g_a w_a^{\text{TYP}}$. Setting $\langle r_a^2 \rangle = \langle s^2 \rangle = 1$ and dropping terms of higher order than $1/\gamma$, we find that whereas the evidence approximation gives a Gaussian predictive distribution for y which has mean and variance:

$$\mu_0 = \sum_a g_a w_a^{\text{MP}|\alpha_{\text{MP}}}, \quad \sigma_0^2 = \sum_a \frac{g_a^2}{\lambda_a + \alpha_{\text{MP}}}, \tag{60}$$

the true predictive distribution is, to order $1/\gamma$, Gaussian with mean and variance:

$$\mu_1 = \mu_0 + \frac{1}{\kappa\gamma} \sum_a g_a w_a^{\text{MP}|\alpha_{\text{MP}}} \frac{\alpha_{\text{MP}}^2}{(\lambda_a + \alpha_{\text{MP}})^2}, \tag{61}$$

$$\sigma_1^2 = \sigma_0^2 + \frac{1}{\kappa\gamma} \left\{ \left(\sum_a g_a w_a^{\text{MP}|\alpha_{\text{MP}}} \frac{\alpha_{\text{MP}}}{(\lambda_a + \alpha_{\text{MP}})} \right)^2 + \sum_a \frac{g_a^2}{\lambda_a + \alpha_{\text{MP}}} \frac{\alpha_{\text{MP}}^2}{(\lambda_a + \alpha_{\text{MP}})^2} \right\}. \tag{62}$$

How wrong can the evidence approximation be? Since both distributions are Gaussian, it is simple to evaluate the Kullback–Leibler distance between them. The cross entropy between $p_0 = \text{Normal}(\mu_0, \sigma_0^2)$ and $p_1 = \text{Normal}(\mu_1, \sigma_1^2)$ is

$$H(p_0, p_1) \equiv \int p_1 \log \frac{p_1}{p_0} = \frac{1}{2} \frac{(\mu_1 - \mu_0)^2}{\sigma_0^2} + \frac{1}{4} \left(\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2} \right)^2 + O\left\{ \left(\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2} \right)^3 \right\}. \tag{63}$$

We consider the two dominant terms separately. The difference in means gives the term

$$\frac{(\mu_1 - \mu_0)^2}{\sigma_0^2} = \frac{1}{\kappa^2\gamma^2} \left(\sum_a h_a w_a^{\text{MP}|\alpha_{\text{MP}}} \frac{\alpha_{\text{MP}}^2}{(\lambda_a + \alpha_{\text{MP}})^{3/2}} \right)^2 \Big/ \sum h_a^2, \tag{64}$$

where $h_a = g_a / \sqrt{\lambda_a + \alpha_{\text{MP}}}$. The worst case is given by the direction $\mathbf{g}$ such that $h_a = w_a^{\text{MP}|\alpha_{\text{MP}}} \frac{\alpha_{\text{MP}}^2}{(\lambda_a + \alpha_{\text{MP}})^{3/2}}$. This worst case gives an upper bound to the contribution to the cross entropy:

$$\frac{(\mu_1 - \mu_0)^2}{\sigma_0^2} \leq \frac{1}{\kappa^2\gamma^2} \sum_a \frac{{w_a^{\text{MP}|\alpha_{\text{MP}}}}^2 \alpha_{\text{MP}}^4}{(\lambda_a + \alpha_{\text{MP}})^3} \tag{65}$$

$$\leq \frac{\alpha_{\text{MP}}}{\kappa^2\gamma^2} \sum_a {w_a^{\text{MP}|\alpha_{\text{MP}}}}^2 = \frac{1}{\kappa^2\gamma} \ll 1. \tag{66}$$

So the change in μ *never* has a significant effect.

The variance term can be split into two terms:

$$\left(\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2} \right)^2 = \frac{1}{\kappa\gamma} \left\{ \left(\sum_a \frac{h_a w_a^{\text{MP}|\alpha_{\text{MP}}} \alpha_{\text{MP}}}{\sqrt{\lambda_a + \alpha_{\text{MP}}}} \right)^2 + \sum_a h_a^2 \frac{\alpha_{\text{MP}}^2}{(\lambda_a + \alpha_{\text{MP}})^2} \right\} \Big/ \sum_a h_a^2, \tag{67}$$

where, as above, $h_a = g_a / \sqrt{\lambda_a + \alpha_{\text{MP}}}$.

For the first term, the worst case is the direction $h_a = w_a^{\text{MP}|\alpha_{\text{MP}}} \frac{\alpha_{\text{MP}}}{\sqrt{\lambda_a + \alpha_{\text{MP}}}}$, *i.e.*, the radial direction $\mathbf{g} = \alpha_{\text{MP}} \mathbf{w}_{\text{MP}|\alpha_{\text{MP}}}$. Substituting in this direction, we find:

$$\text{First term} \;\leq\; \frac{1}{\kappa\gamma} \sum_a {w_a^{\text{MP}|\alpha_{\text{MP}}}}^2 \frac{\alpha_{\text{MP}}^2}{\lambda_a + \alpha_{\text{MP}}} \tag{68}$$

$$\leq\; \frac{\alpha_{\text{MP}}}{\kappa\gamma} \sum_a {w_a^{\text{MP}|\alpha_{\text{MP}}}}^2 \;=\; \frac{1}{\kappa} \;=\; O(1). \tag{69}$$

We can improve this bound by substituting for $w_a^{\text{MP}|\alpha_{\text{MP}}}$ in terms of d_a and making use of the definition of n_e. Only n_e of the terms in the sum in equation (68) are significant. Thus

$$\text{First term} \lesssim \frac{n_e}{\kappa\gamma}. \tag{70}$$

So this term can give a significant effect, but only in one direction; for any direction orthogonal (in $\mathbf{h}$) to this radial direction, this term is zero.

Finally, we examine the second term:

$$\frac{1}{\kappa\gamma} \sum_a h_a^2 \frac{\alpha_{\text{MP}}^2}{(\lambda_a + \alpha_{\text{MP}})^2} \Bigg/ \sum_a h_a^2 \;<\; \frac{1}{\kappa\gamma} \;\ll\; 1. \tag{71}$$

So this term never has a significant effect.

Conclusion

The evidence approximation affects the mean and variance of properties y of $\mathbf{w}$, but only to within $O(\gamma^{-1/2})$ of the property's standard deviation; this error is insignificant, for large γ. The sole exception is the direction $\mathbf{g} = \mathbf{w}_{\text{MP}|\alpha_{\text{MP}}}$, along which the variance is erroneously small, with a cross–entropy error of order $O(n_e/\gamma)$.

8.3 A correction term

This result motivates a straightforward term which could be added to the inverse Hessian of the evidence approximation, to correct the predictive variance in this direction. The predictive variance for a general $y = \mathbf{g}^{\mathsf{T}}\mathbf{w}$ could be estimated by

$$\sigma_y^2 = \mathbf{g}^{\mathsf{T}} \left(\mathbf{\Sigma} + \sigma^2_{\log\alpha|D} \mathbf{w}'_{\text{MP}|\alpha} {\mathbf{w}'_{\text{MP}|\alpha}}^{\mathsf{T}} \right) \mathbf{g}, \tag{72}$$

where $\mathbf{w}'_{\text{MP}|\alpha} \equiv \partial \mathbf{w}_{\text{MP}|\alpha} / \partial(\log\alpha) = \alpha \mathbf{\Sigma} \mathbf{w}_{\text{MP}|\alpha}$, and $\sigma^2_{\log\alpha|D} = \frac{2}{\gamma}$. With this correction, the predictive distribution for any direction would be in error only by order $O(1/\gamma)$. If the noise variance $\sigma_\nu^2 = \beta^{-1}$ is also uncertain, then the factor $\sigma^2_{\log\alpha|D}$ is incremented by $\sigma^2_{\log\beta|D} = \frac{2}{N-\gamma}$.

9 Discussion

The MAP method, though it can give exact values for the relative probability densities of two weight vectors, is capable of giving a Gaussian approximation which is highly unrepresentative of the true posterior. In high dimensional spaces, maxima of densities are misleading. MAP estimates play no fundamental role in Bayesian inference, and they can change arbitrarily with arbitrary re–parameterizations. The problem with MAP estimates is that they maximize the probability *density*, without taking account of the complementary *volume* information. What matters is where the probability *mass* is, and mass is equal to density times volume. When there are many ill–determined parameters, the MAP method's integration over α yields a $\mathbf{w}_{\text{MP}}$ which is severely over–regularized. Integration over the noise level $1/\beta$ to give the true likelihood leads to a bias in the other direction. [These two biases may cancel: the evidence framework's $\mathbf{w}_{\text{MP}|\alpha_{\text{MP}},\beta_{\text{MP}}}$ coincides with $\mathbf{w}_{\text{MP}}$ if the number of well–determined parameters happens to obey the condition $\gamma/k = N/(N+k)$, where N is the number of data points.]

There are two general take–home messages.

(1) When one has a choice of which variables to integrate over and which to maximize over, one should integrate over as many variables as possible, in order to capture the relevant volume information. There are typically far fewer regularization constants and other hyperparameters than there are 'level 1' parameters.

(2) If practical Bayesian methods involve approximations such as fitting a Gaussian to a posterior distribution, then one should think twice before integrating out hyperparameters (Gull 1988). The probability density which results from such an integration typically has a skew peak; a Gaussian fitted at the peak may not approximate the distribution well. In contrast, optimization of the hyperparameters can give a Gaussian approximation which, for predictive purposes, puts most of the probability mass in the right place.

The evidence approximation, which sets hyperparameters so as to maximize the evidence, is not intended to produce an accurate approximation to the numerical value of the true posterior density over $\mathbf{w}$; and it does not. But what matters is whether low–dimensional properties of $\mathbf{w}$ (*i.e.*, predictions) are seriously mis–calculated as a result of the evidence approximation. The main conditions for the evidence approximation are that the data should not be grossly at variance with the model, and that the number of well–determined parameters γ should be large. How large depends on the problem, but often a value as small as $\gamma \simeq 3$ is sufficient, because this means that α is determined to within a factor of e (recall $\sigma_{\log \alpha|D} \simeq \sqrt{2/\gamma}$); predictive distributions are often insensitive to changes of α of this magnitude. Thus the approximation is usually good if we have enough data to determine a few parameters.

If satisfactory conditions do not hold for the evidence approximation (*e.g.*, if γ is too small), then it should be emphasized that this would not motivate integrating out α first.

The MAP approximation is systematically inferior to the evidence approximation. Practical alternative methods for dealing with hyperparameters include the deterministic method of Bryan (1990), who finds it most convenient numerically to retain α as an explicit variable, and integrate it out *last*, and the Markov chain Monte Carlo implementation of Neal (1996) which samples the hyperparameters and parameters from the joint distribution $P(\mathbf{w}, \alpha | D, \mathcal{H})$.

The relationship between evidence maximization and ensemble learning derived in section 7 gives a convergence proof (at least for linear models) for a re–estimation formula for α (equation 46) which previous work on the evidence framework had not provided. The steps of re–estimating $\bar{\alpha}$ and computing the new distribution $Q_{\mathbf{w}}(\mathbf{w})$ both decrease F, and F is bounded below, so the iterative procedure must converge.

A final point in favour of the evidence framework is that it can be naturally extended (at least approximately) to more elaborate priors such as mixture models; it would be difficult to integrate over the mixture hyperparameters in order to evaluate the true prior in these cases.

Acknowledgments

I thank Radford Neal, David R.T. Robinson, Steve Gull, Steve Waterhouse and Martin Oldfield for helpful discussions, and John Skilling for invaluable contributions to the proof in section 8. I am grateful to Mike Lewicki, Anton Garrett and Mark Gibbs for comments on the manuscript.

References

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

Box, G. E. P., and Tiao, G. C. (1973) *Bayesian inference in statistical analysis*. Addison–Wesley.

Bretthorst, G. (1988) *Bayesian spectrum analysis and parameter estimation*. Springer. Also available at `bayes.wustl.edu`.

Bryan, R. (1990) Solving oversampled data problems by Maximum Entropy. In *Maximum Entropy and Bayesian Methods, Dartmouth, U.S.A., 1989*, ed. by P. Fougere, pp. 221–232. Kluwer.

Buntine, W., and Weigend, A. (1991) Bayesian back–propagation. *Complex Systems* **5**: 603–643.

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

Feynman, R. P. (1972) *Statistical Mechanics*. W. A. Benjamin, Inc.

Gull, S. F. (1988) Bayesian inductive inference and maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering, vol. 1: Foundations*, ed. by G. Erickson and C. Smith, pp. 53–74, Dordrecht. Kluwer.

Gull, S. F. (1989) Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods, Cambridge 1988*, ed. by J. Skilling, pp. 53–71, Dordrecht. Kluwer.

Hinton, G. E., and Sejnowski, T. J. (1986) Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing*, ed. by D. E. Rumelhart and J. E. McClelland, pp. 282–317. Cambridge Mass.: MIT Press.

Hinton, G. E., and van Camp, D. (1993) Keeping neural networks simple by minimizing the description length of the weights. In *Proc. 6th Annu. Workshop on Comput. Learning Theory*, pp. 5–13. ACM Press, New York, NY.

MacKay, D. J. C. (1991) *Bayesian Methods for Adaptive Models*. California Institute of Technology dissertation.

MacKay, D. J. C. (1992a) Bayesian interpolation. *Neural Computation* **4** (3): 415–447.

MacKay, D. J. C. (1992b) The evidence framework applied to classification networks. *Neural Computation* **4** (5): 698–714.

MacKay, D. J. C. (1992c) A practical Bayesian framework for backpropagation networks. *Neural Computation* **4** (3): 448–472.

MacKay, D. J. C. (1995) Developments in probabilistic modelling with neural networks—ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands, 14-15 September 1995*, pp. 191–198, Berlin. Springer.

MacKay, D. J. C. (1996) Bayesian non-linear modelling for the 1993 energy prediction competition. In *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, ed. by G. Heidbreder, pp. 221–234, Dordrecht. Kluwer.

Neal, R. M. (1993a) Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 5*, ed. by C. L. Giles, S. J. Hanson, and J. D. Cowan, pp. 475–482, San Mateo, California. Morgan Kaufmann.

Neal, R. M. (1993b) Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG–TR–93–1, Dept. of Computer Science, University of Toronto.

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

Neal, R. M., and Hinton, G. E. (1998) A new view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, ed. by M. I. Jordan, NATO Science Series, pp. 355–368. Kluwer Academic Press.

Reif, F. (1965) *Fundamentals of Statistical and Thermal Physics*. McGraw–Hill.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986) Learning representations by back–propagating errors. *Nature* **323**: 533–536.

Skilling, J. (1993) Bayesian numerical analysis. In *Physics and Probability*, ed. by W. T. Grandy, Jr. and P. Milonni, Cambridge. C.U.P.

Strauss, C. E. M., Wolpert, D. H., and Wolf, D. R. (1993) Alpha, evidence, and the entropic prior. In *Maximum Entropy and Bayesian Methods, Paris 1992*, ed. by A. Mohammed-Djafari, Dordrecht. Kluwer.

Thodberg, H. H. (1996) Review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Networks* **7** (1): 56–72.

Wahba, G. (1975) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Numer. Math.* **24**: 383–393.

Weigend, A. S., Rumelhart, D. E., and Huberman, B. A. (1991) Generalization by weight–elimination with applications to forecasting. In *Advances in Neural Information Processing Systems 3*, ed. by R. P. L. et. al., pp. 875–882. Morgan Kaufmann.

Weir, N. (1991) Applications of maximum entropy techniques to HST data. In *Proceedings of the ESO/ST-ECF Data Analysis Workshop, April 1991*, ed. by P. Grosbol and R. Warmels, pp. 115–129, Garching. European Southern Observatory/Space Telescope – European Coordinating Facility.

Wolpert, D. H. (1993) On the use of evidence in neural networks. In *Advances in Neural Information Processing Systems 5*, ed. by C. L. Giles, S. J. Hanson, and J. D. Cowan, pp. 539–546, San Mateo, California. Morgan Kaufmann.

November 11, 1998 — Version 3.6. (Final version.)