HYPERPARAMETERS: OPTIMIZE, OR INTEGRATE OUT?

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

ABSTRACT. I examine two approximate methods for computational implementation of Bayesian hierarchical models, that is, models which include unknown hyperparameters such as regularization constants. 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 the distinctive properties of inference in many dimensions.

"Integrating over a nuisance parameter is very much like estimating the parameter from the data, and then using that estimate in our equations." *G.L. Bretthorst*

"This integration would be counter-productive as far as practical manipulation is concerned." *S.F. Gull*

1 Outline

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

$$P(D, \mathbf{w}, \alpha, \beta|\mathcal{H}) = P(D|\mathbf{w}, \beta, \mathcal{H})P(\mathbf{w}|\alpha, \mathcal{H})P(\alpha, \beta|\mathcal{H}), \tag{1}$$

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, $\{\mathbf{x}, \mathbf{t}\}$, 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 $\{\mathbf{t}\}$ are given by adding 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 $P(\mathbf{w}|\alpha, \mathcal{H})$ 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 (1) include linear interpolation with a fixed basis set

(Gull 1988; MacKay 1992a), non-linear regression with a neural network (MacKay 1992b), non-linear classification (MacKay 1992c), and image deconvolution (Gull 1989).

In the simplest case (linear models, Gaussian noise), the first factor in (1), the likelihood, can be written in terms of a quadratic function of $\mathbf{w}$, $E_D(\mathbf{w})$:

$$P(D|\mathbf{w},\beta,\mathcal{H}) = \frac{1}{Z_D(\beta)} \exp(-\beta E_D(\mathbf{w})). \tag{2}$$

What makes the problem 'ill-posed' is that 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 (1), the prior, which in the simplest case is a spherical Gaussian:

$$P(\mathbf{w}|\alpha,\mathcal{H}) = \frac{1}{Z_W(\alpha)} \exp(-\alpha \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}). \tag{3}$$

The regularization constant α defines the variance $\sigma_w^2 = 1/\alpha$ of the prior for the components w_i of $\mathbf{w}$.

Much interest has centred on the question 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' method in statistics (Wahba 1975). This method can be applied to non-linear models by making appropriate local linearizations, and has been used successfully in image reconstruction (Gull 1989; Weir 1991) and in neural networks (MacKay 1992b; Thodberg 1993; MacKay 1994).

Recently 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' 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*); this use of the term 'MAP' may not coincide precisely with its general usage.

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. It is assumed that it is predictive *distributions* that 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 non-linear models and to models with multiple hyperparameters. When a non-linear 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 1992b); or (b) fitting Gaussians to local maxima of the true posterior with the hyperparameter α integrated out.

2 The Alternative Methods

Given the Bayesian model defined in (1), 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.

Let us 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.

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). 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.

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:

$$P(\mathbf{w}|D,\alpha,\mathcal{H}) = \frac{P(D|\mathbf{w},\alpha,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})}{P(D|\alpha,\mathcal{H})}. \tag{4}$$

Level 2: Infer α:

$$P(\alpha|D,\mathcal{H}) = \frac{P(D|\alpha,\mathcal{H})P(\alpha|\mathcal{H})}{P(D|\mathcal{H})}. \tag{5}$$

Level 3: Compare models:

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}). \tag{6}$$

There is a pattern in these three applications of Bayes' rule: at each of 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, around a maximum of $P(\mathbf{w}|D,\alpha,\mathcal{H})$, of $\log P(D|\mathbf{w},\alpha,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})$; 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, $\alpha_{\rm MP}$, and error bars on it, $\sigma_{\log\alpha|D}$. (Because α is a positive scale variable, it is natural to represent its uncertainty on a log scale.)

- The value of $\alpha_{\rm MP}$ is substituted at level 1. This defines a probability distribution $P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H})$ which is intended as a 'good approximation' 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, $\mathbf{w}_{\rm MP|\alpha_{\rm MP}}$, with covariance matrix $\mathbf{\Sigma}$ defined by $\mathbf{\Sigma}^{-1} = -\nabla\nabla\log P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H})$. Marginals for the components of $\mathbf{w}$ are easily obtained from this distribution.

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

$$P(D|\mathcal{H}) \simeq P(D|\alpha_{\rm MP},H)P(\log\alpha_{\rm MP}|\mathcal{H})\sqrt{2\pi}\sigma_{\log\alpha|D}. \tag{7}$$

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

$$P(D_2|D,\alpha_{\rm MP},\mathcal{H}) = \int d^k\mathbf{w}\ \ P(D_2|\mathbf{w},\mathcal{H})P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H}). \tag{8}$$

  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 and intuitive implicit equation,

$$\frac{1}{\alpha_{\rm MP}} = \frac{\sum_1^k w_i^2}{\gamma} \tag{9}$$

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

$$\gamma = k - \alpha{\rm Trace}\mathbf{\Sigma} = \sum_1^k \frac{\lambda_a}{\lambda_a+\alpha}. \tag{10}$$

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 (9) 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 5). Thus $\log\alpha$ is well-determined by the data if $\gamma \gg 1$.

The central computation can be summarised thus:

Evidence approximation: find the 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 (9).

*Justification for the Evidence Approximation*The central approximation in this scheme can be stated as follows: when we integrate out a parameter, 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,

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

This approximation is accurate 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,

$$P(r) = \exp(-\beta E_r)/Z, \tag{11}$$

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:

$$P(r) = \begin{cases} 1/\Omega & E_r \in [\bar{E} \pm \delta E/2] \\ 0 & \text{otherwise} \end{cases} . \tag{12}$$

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 it is well known (Reif 1965) that for most macroscopic purposes, if the system has a large number of degrees of freedom, 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 $P(\mathbf{w}|D, \alpha_{\mathrm{MP}}, \mathcal{H})$ 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 $P(\mathbf{w}|D, \alpha_{\mathrm{MP}}, \mathcal{H})$ puts most of its probability mass in the right place. A more precise discussion of this approximation is given in section 5.

The MAP Method

The alternative procedure is first to integrate out α to obtain the true prior:

$$P(\mathbf{w}|\mathcal{H}) = \int d\alpha \, P(\mathbf{w}|\alpha, \mathcal{H}) P(\alpha|\mathcal{H}). \tag{13}$$

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

$$P(\mathbf{w}|D, \mathcal{H}) \propto P(D|\mathbf{w}, \mathcal{H}) P(\mathbf{w}|\mathcal{H}). \tag{14}$$

This posterior can be maximized to find the MAP parameters, $\mathbf{w}_{\rm MP}$. 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, $\mathbf{w}_{\rm MP}$; to find the Hessian of the posterior, one needs the Hessian of the prior, derived below. A simple evaluation of the factors on the right hand side of (14) is not a satisfactory solution of problem A, since (a) the normalizing constant is missing; (b) even if the r.h.s. of (14) 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; how, for example, is one to obtain marginal distributions over w_i from (14)?

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

Problem C: The parameters $\mathbf{w}_{\rm MP}$ with error bars are used to generate predictions as in (8).

A simple example will illustrate that this approach actually gives results qualitatively very similar to the evidence framework. If we apply the improper prior $P_{\rm Imp}(\log \alpha) = \text{const}$ and evaluate the true prior, we obtain:[1]

$$P_{\rm Imp}(\mathbf{w}|\mathcal{H}) = \int_{\alpha=0}^{\infty} \frac{e^{-\alpha \sum_{i=1}^{k} w_i^2/2}}{Z_W(\alpha)} \, d\log\alpha \;\propto\; \frac{1}{(\sum_i w_i^2)^{k/2}}. \tag{15}$$

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',

$$\frac{1}{\alpha_{\rm eff}(\mathbf{w})} = \frac{\sum_i w_i^2}{k}. \tag{16}$$

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 $\alpha_{\rm eff}$. The similarity of equation (16) to equation (9) of the evidence framework is clear. We can therefore describe the MAP method thus:

> **MAP method** (improper prior): 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 (16).

This procedure is suggested in (MacKay 1992b) as a 'quick and dirty' approximation to the evidence framework.

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

The Effective α and the Curvature of a General Prior

We have just established that, when the improper prior (15) is used, the MAP solution lies on the 'alpha trajectory' — the graph of $\mathbf{w}_{\mathrm{MP}|\alpha}$ — for a particular value of $\alpha = \alpha_{\mathrm{eff}}$. This result still holds when a proper prior over α is used to define the true prior (13). The effective $\alpha(\mathbf{w})$, found by differentiation of $\log P(\mathbf{w}|\mathcal{H})$, is:

$$\alpha_{\mathrm{eff}}(\mathbf{w}) = \int d\alpha \; \alpha \; P(\alpha|\mathbf{w}, \mathcal{H}). \tag{17}$$

In general there may be multiple local probability maxima, all of which lie on the alpha trajectory. In summary, optima $\mathbf{w}_{\mathrm{MP}}$ found by the MAP method can be described thus:

MAP method (proper prior): find the self-consistent solution $\{\mathbf{w}_{\mathrm{MP}}, \alpha_{\mathrm{eff}}\}$ such that $\mathbf{w}_{\mathrm{MP}}$ maximizes $P(\mathbf{w}|D, \alpha_{\mathrm{eff}}, \mathcal{H})$ and α_{eff} satisfies equation (17).

The curvature of the true prior is needed for evaluation of the error bars on $\mathbf{w}$ in the MAP method. By direct differentiation of the true prior (13), we find:

$$-\nabla\nabla \log P(\mathbf{w}|\mathcal{H}) = \alpha_{\mathrm{eff}}\mathbf{I} - \sigma_\alpha^2(\mathbf{w})\mathbf{w}\mathbf{w}^{\mathrm{T}}, \tag{18}$$

where $\alpha_{\mathrm{eff}}(\mathbf{w})$ is defined in (16), and the effective variance of α is:

$$\sigma_\alpha^2(\mathbf{w}) = \overline{\alpha^2}(\mathbf{w}) - \alpha_{\mathrm{eff}}(\mathbf{w})^2 = \int d\alpha \; \alpha^2 \; P(\alpha|\mathbf{w}, \mathcal{H}) - \left(\int d\alpha \; \alpha \; P(\alpha|\mathbf{w}, \mathcal{H}) \right)^2 . \tag{19}$$

This is an intuitive result: if α were fixed to α_{eff}, then the curvature would just be the first term in (18), $\alpha_{\mathrm{eff}}\mathbf{I}$. The fact that α is uncertain depletes the curvature in the radial direction $\hat{\mathbf{w}} = \mathbf{w}/|\mathbf{w}|$.

3 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 real 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.

The Widget Example

A collection of widgets $i = 1..k$ have a property called 'wibble', 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: $\{d_1, d_2, d_3, d_4\} = \{3.2, -3.2, 2.8, -2.8\}$. The task is (problem A) to infer the wibbles of these four widgets, *i.e.* to produce a representative $\mathbf{w}$ with error bars. On the back of an envelope, or in a computer algebra system, we find the following answers using equations (9) and (16/17):

Evidence framework: $\alpha_{\rm MP} = 0.124$, $\mathbf{w}_{\rm MP|\alpha_{\rm MP}} = \{2.8, -2.8, 2.5, -2.5\}$, each with error bars ± 0.9.

MAP method: $\alpha_{\rm eff} = 0.145$, $\mathbf{w}_{\rm MP} = \{2.8, -2.8, 2.4, -2.4\}$, each with error bars ± 0.9. These answers are insensitive to the details of the prior over σ_w.

So far so good: $\mathbf{w}_{\rm MP|\alpha_{\rm MP}}$ is slightly less regularized than $\mathbf{w}_{\rm MP}$, but there is not much disagreement 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 unmeasured widgets in a box next door. Thus we now have both well-determined and ill-determined parameters, as in an ill-posed problem. Intuitively, we would like our inferences about the well-measured widgets to be negligibly affected by this vacuous information about the unmeasured widgets, just as the true Bayesian predictive distributions are unaffected. But clearly with $k = 8$, the difference between k and γ in equations (9) and (16) is going to become significant. The value of $\alpha_{\rm eff}$ will be substantially greater than that of $\alpha_{\rm MP}$.

In the evidence framework the value of γ is exactly the same, since each of the ill-determined parameters has $\lambda = 0$ and adds nothing to the sum in (10). So the value of $\alpha_{\rm MP}$ and the predictive distributions are unchanged.

In contrast, the MAP parameter vector $\mathbf{w}_{\rm MP}$ is squashed close to zero. The precise value of $\mathbf{w}_{\rm MP}$ is sensitive to the prior over α. Solving equation (17) in a computer algebra system, we find: $\alpha_{\rm eff} = 79.2$, $\mathbf{w}_{\rm MP} = \{0.040, -0.040, 0.035, -0.035, 0, 0, 0, 0\}$, 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 correspondingly 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 'pruning' these parameters. In scenario 1, the network is pruned, removing the ill-determined parameters. In scenario 2, the parameters are retained, and assume their most probable value, zero. In each case, what is the optimal setting of the weight decay rate α (assuming the traditional regularizer $\mathbf{w}^{\rm T}\mathbf{w}/2$)? 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.

4 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.

Consider two Gaussian densities in 1000 dimensions which differ in σ_w by 1%, and which contain equal total probability mass. In each case 90% of the mass is located in a shell which differs in radius by only 1% between the two distributions. The maximum probability density, however, is greater at the centre of the Gaussian with smaller σ_w, by a factor of $\sim \exp(0.01k) \simeq 20{,}000$.

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.

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.

Condition Satisfied by Typical Samples

The conditions (9) and (16), 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 the posterior: the maximum has an atypically small value of $\mathbf{w}^{\rm T}\mathbf{w}$, because, as discussed above, 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 $\mathbf{w}^{\rm T}\mathbf{w} = \sum_i w_i^2$ for these samples satisfies:

$$\alpha_{\rm MP} = \frac{k}{\langle \sum_i w_i^2 \rangle_{|D,\alpha_{\rm MP}}}. \tag{20}$$

Proof: The deviation $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$ is Gaussian distributed with $\Delta\mathbf{w}\Delta\mathbf{w}^{\mathrm{T}} = \mathbf{\Sigma}$. So $\alpha_{\mathrm{MP}} \langle \sum_i w_i^2 \rangle_{|D,\alpha_{\mathrm{MP}}} = \alpha_{\mathrm{MP}} (\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}} + \Delta\mathbf{w})^{\mathrm{T}} (\mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}} + \Delta\mathbf{w}) = \alpha_{\mathrm{MP}} \mathbf{w}^2_{\mathrm{MP}|\alpha_{\mathrm{MP}}} + \alpha_{\mathrm{MP}} \mathrm{Trace}\mathbf{\Sigma} = k$, using equations (9) and (10).

Thus a typical sample from the evidence approximation prefers just the same value of α as does the evidence.

5 Conditions for the Evidence Approximation

We have observed 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 the case of linear models with Gaussian probability distributions. This includes the case of image reconstruction problems that have separable Gaussian distributions in the Fourier domain.

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 (C) and (B).

For convenience let us work in the eigenvector basis where the prior (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 cover several orders of magnitude in value. 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:

$$P(\log\alpha|\mathcal{H}) = \frac{1}{\log(\alpha_{\max}/\alpha_{\min})}, \quad P(\mathbf{w}|\alpha,\mathcal{H}) = \left(\frac{\alpha}{2\pi}\right)^{k/2} \exp\left(-\frac{1}{2}\alpha \sum_1^k w_a^2\right), \text{ and}$$

$$P(D|\mathbf{w},\mathcal{H}) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\sum_1^k \left(\sqrt{\lambda_a} w_a - d_a\right)^2\right\}.$$

In the case of a deconvolution problem the eigenvectors are the Fourier set and the point spread function in Fourier space is given by $\sqrt{\lambda_a}$.

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.

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 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 $\alpha_{\rm MP}$). For each data analysis problem, one may evaluate the critical $\alpha_{\max}$ above which the posterior is measurably affected by the large α tail of the evidence (Gull 1989). Often, as Gull points out, this critical value of $\alpha_{\max}$ has bizarre 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 likelihood and the prior. 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. I now suggest a way of formalizing this concept into a quantitative test.

If we accept the model, then we believe that there is a true value of $\alpha = \alpha_{\rm T}$, and that given $\alpha_{\rm T}$, the data measurements d_a are the sum of two independent Gaussian variables $\sqrt{\lambda_a} w_a$ and ν_a, so that $P(d_a | \alpha_{\rm T}, \mathcal{H}) = \mathrm{Normal}(0, \sigma^2_{a|\alpha_T})$, 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 $\alpha_{\rm T}$ 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, if any of the quantities $j_a = d_a^2 / (\frac{\lambda_a}{\alpha} + 1)$ is not in the interval $[e^{-\tau}, 1 + \tau]$; where τ is the significance level of this test.*

It is conjectured that if we find a value of $\alpha = \alpha_{\rm MP}$ 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 $\alpha_{\rm MP}$, 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:*

$$n_e \equiv \sum_a \frac{4 \lambda_a \alpha_{\rm MP}}{(\lambda_a + \alpha_{\rm MP})^2}. \tag{21}$$

This is a measure of the number of eigenvalues λ_a within approximately e-fold of $\alpha_{\rm MP}$.

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 $\alpha_{\rm MP}$, then we find on Taylor-expanding $\log P(\alpha|D,\mathcal{H})$ about $\alpha=\alpha_{\rm MP}$, that the second derivative is large, and that the third derivative is relatively small:

$$\begin{aligned}
\left.\frac{\partial \log P(D|\alpha,\mathcal{H})}{\partial \log \alpha}\right|_{\alpha_{\rm MP}} &= \frac{1}{2}\left(\gamma - \alpha \mathbf{w}^2_{{\rm MP}|\alpha_{\rm MP}}\right) = 0 \\
\left.\frac{\partial^2 \log P(D|\alpha,\mathcal{H})}{\partial (\log \alpha)^2}\right|_{\alpha_{\rm MP}} &\simeq -\alpha \mathbf{w}^2_{{\rm MP}|\alpha_{\rm MP}} = -\frac{\gamma}{2} \\
\left.\frac{\partial^3 \log P(D|\alpha,\mathcal{H})}{\partial (\log \alpha)^3}\right|_{\alpha_{\rm MP}} &\simeq -\alpha \mathbf{w}^2_{{\rm MP}|\alpha_{\rm MP}} = -\frac{\gamma}{2}.
\end{aligned}$$

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. The third derivative *is* relatively small (even though it is equal to the second derivative), 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 $d=c=\gamma \gg 1$, 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 (7) follows.

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:

$$P(\mathbf{w}|D,\alpha,\mathcal{H}) \propto \exp\left\{-\frac{1}{2}\sum_a (\lambda_a+\alpha)\left(w_a - \frac{\sqrt{\lambda_a}d_a}{\lambda_a+\alpha}\right)^2\right\}. \tag{22}$$

The most probable $\mathbf{w}$ given this value of α is: $w_a^{{\rm MP}|\alpha} = \sqrt{\lambda_a}d_a/(\lambda_a+\alpha)$. 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:

$$w_a^{{\rm TYP}|\alpha} = \frac{\sqrt{\lambda_a}d_a}{\lambda_a+\alpha} + \frac{r_a}{\sqrt{\lambda_a+\alpha}}, \tag{23}$$

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

$$\alpha^{\rm TYP} = \alpha_{\rm MP}\left(1+\frac{s}{\sqrt{\kappa\gamma}}\right), \tag{24}$$

where s is a sample from Normal(0,1). We now substitute this $\alpha^{\rm TYP}$ into (23) 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 $\mathbf{w}^{\rm TYP}$ in powers of $1/\gamma$:

$$\begin{aligned} w_a^{\rm TYP} &= \frac{\sqrt{\lambda_a}d_a}{\lambda_a+\alpha_{\rm MP}}\left(1-\frac{s}{\sqrt{\kappa\gamma}}\frac{\alpha_{\rm MP}}{\lambda_a+\alpha_{\rm MP}}+\frac{s^2}{\kappa\gamma}\frac{\alpha_{\rm MP}^2}{(\lambda_a+\alpha_{\rm MP})^2}+\ldots\right)+ \\ &\quad \frac{r_a}{\sqrt{\lambda_a+\alpha_{\rm MP}}}\left(1-\frac{1}{2}\frac{s}{\sqrt{\kappa\gamma}}\frac{\alpha_{\rm MP}}{\lambda_a+\alpha_{\rm MP}}+\frac{3}{8}\frac{s^2}{\kappa\gamma}\frac{\alpha_{\rm MP}^2}{(\lambda_a+\alpha_{\rm MP})^2}\ldots\right) \end{aligned} \tag{25}$$

We now examine the mean and variance of $y^{\rm TYP}=\sum_a g_a w_a^{\rm TYP}$. 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:

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

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

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

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

How wrong can the evidence approximation be? Since both distributions are Gaussian, it is simple to evaluate various distances between them. The cross entropy between p_0 =Normal(μ_0,σ_0^2) and p_1 =Normal(μ_1,σ_1^2) is

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

We consider the two dominant terms separately.

$$\frac{(\mu_1-\mu_0)^2}{\sigma_0^2} = \frac{1}{\kappa^2\gamma^2}\left(\sum_a h_a\, w_a^{{\rm MP}|\alpha_{\rm MP}}\frac{\alpha_{\rm MP}^2}{(\lambda_a+\alpha_{\rm MP})^{3/2}}\right)^2 \Big/ \sum h_a^2, \tag{26}$$

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

$$\frac{(\mu_1-\mu_0)^2}{\sigma_0^2} \leq \frac{1}{\kappa^2\gamma^2}\sum_a \frac{{w_a^{{\rm MP}|\alpha_{\rm MP}}}^2\alpha_{\rm MP}^4}{(\lambda_a+\alpha_{\rm MP})^3} \tag{27}$$

$$< \frac{\alpha_{\rm MP}}{\kappa^2\gamma^2}\sum {w_a^{{\rm MP}|\alpha_{\rm MP}}}^2 = \frac{1}{\kappa^2\gamma} \ll 1 \tag{28}$$

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

The variance term can be split into two terms:

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

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

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

$$\begin{aligned} \text{First term} &\leq \frac{1}{\kappa\gamma}\sum_a {w_a^{\mathrm{MP}|\alpha_{\mathrm{MP}}}}^2 \frac{\alpha_{\mathrm{MP}}^2}{\lambda_a+\alpha_{\mathrm{MP}}} &(29)\\ &< \frac{\alpha_{\mathrm{MP}}}{\kappa\gamma}\sum_a {w_a^{\mathrm{MP}|\alpha_{\mathrm{MP}}}}^2 = \frac{1}{\kappa} = O(1) &(30)\end{aligned}$$

We can improve this bound by substituting for $w_a^{\mathrm{MP}|\alpha_{\mathrm{MP}}}$ 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 (29) are significant. Thus

$$\text{First term} \lesssim \frac{n_e}{\kappa\gamma}. \tag{31}$$

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:

$$\frac{1}{\kappa\gamma}\sum_a h_a^2 \frac{\alpha_{\mathrm{MP}}^2}{(\lambda_a+\alpha_{\mathrm{MP}})^2} \Bigg/ \sum_a h_a^2 < \frac{1}{\kappa\gamma} \ll 1. \tag{32}$$

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 $\mathbf{g} = \mathbf{w}_{\mathrm{MP}|\alpha_{\mathrm{MP}}}$, along which the variance is erroneously small, with a cross-entropy error of order $O(n_e/\gamma)$.

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 $y = \mathbf{g}^{\mathrm{T}}\mathbf{w}$ could be estimated by

$$\sigma_y^2 = \mathbf{g}^{\mathrm{T}}\left(\mathbf{\Sigma} + \sigma^2_{\log\alpha|D}\mathbf{w}'_{\mathrm{MP}|\alpha}{\mathbf{w}'_{\mathrm{MP}|\alpha}}^{\mathrm{T}}\right)\mathbf{g}, \tag{33}$$

where $\mathbf{w}'_{\mathrm{MP}|\alpha} \equiv \partial\mathbf{w}_{\mathrm{MP}|\alpha}/\partial(\log\alpha) = \alpha\mathbf{\Sigma}\mathbf{w}_{\mathrm{MP}|\alpha}$, 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 $\sigma^2_{\log\beta|D} = \frac{2}{N-\gamma}$.

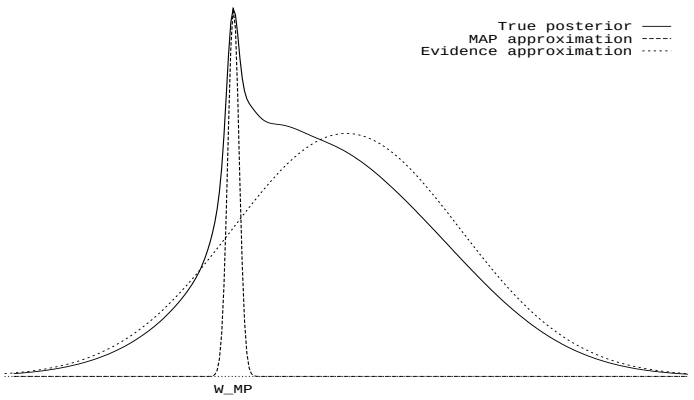

Figure 1: Approximating complicated distributions with a Gaussian
This is a schematic illustration of the properties of a multi-dimensional distribution. A typical posterior distribution for an ill-posed problem has a skew peak. A Gaussian fitted at the MAP parameters is a bad approximation to the distribution: it is in the wrong place, and its error bars are far too small. Additional features of the true posterior distribution not illustrated here are that it typically has spikes of high probability density at the origin $\mathbf{w} = 0$ and at the maximum likelihood parameters $\mathbf{w} = \mathbf{w}_{\rm ML}$. The evidence approximation gives a Gaussian distribution which captures most of the probability mass of the true posterior.

6 Discussion

The MAP method, though exact, is capable of giving MAP parameters which are unrepresentative of the true posterior. In high dimensional spaces, maxima 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. Figure 1 attempts, in one dimension, to convey this difference between the two Gaussian approximations.

When there are many ill-determined parameters, the MAP method's integration over α yields a $\mathbf{w}_{\rm MP}$ which is over-regularized.[2]

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

[2]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 $\mathbf{w}_{\rm MP|\alpha_{MP},\beta_{MP}}$ coincides with $\mathbf{w}_{\rm MP}$ if the number of well-determined parameters happens to obey the condition $\gamma/k = N/(N+k)$.

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 numerical approximation to the true posterior distribution 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 likelihood and the prior, 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 then motivate integrating out α first. The MAP approximation is systematically inferior to the evidence approximation. It would probably be most convenient numerically to retain α as an explicit variable, and integrate it out *last* (Bryan 1990).

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, and Martin Oldfield for helpful discussions, and John Skilling for invaluable contributions to the proof in section 5. I am grateful to Anton Garrett for comments on the manuscript.

References

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.

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.

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.

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

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

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

MACKAY, D. J. C. (1994) Bayesian non-linear modelling for the 1993 energy prediction competition. In *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, ed. by G. Heidbreder, 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.

REIF, F. (1965) *Fundamentals of statistical and thermal physics*. McGraw–Hill.

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. (1993) Ace of Bayes: application of neural networks with pruning. Technical Report 1132 E, Danish meat research institute.

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.

WEIR, N. (1991) Applications of maxmimum entropy techniques to HST data. In *Proceedings of the ESO/ST–ECF Data Analysis Workshop, April 1991*.

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.