

be simply  $\hat{y}(\mathbf{x}) = 1$ , from which we obtain (3.64). Note that the kernel function can be negative as well as positive, so although it satisfies a summation constraint, the corresponding predictions are not necessarily convex combinations of the training set target variables.

Finally, we note that the equivalent kernel (3.62) satisfies an important property shared by kernel functions in general, namely that it can be expressed in the form an inner product with respect to a vector  $\psi(\mathbf{x})$  of nonlinear functions, so that

$$k(\mathbf{x}, \mathbf{z}) = \psi(\mathbf{x})^T \psi(\mathbf{z}) \quad (3.65)$$

where  $\psi(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \phi(\mathbf{x})$ .

### 3.4. Bayesian Model Comparison

In Chapter 1, we highlighted the problem of over-fitting as well as the use of cross-validation as a technique for setting the values of regularization parameters or for choosing between alternative models. Here we consider the problem of model selection from a Bayesian perspective. In this section, our discussion will be very general, and then in Section 3.5 we shall see how these ideas can be applied to the determination of regularization parameters in linear regression.

As we shall see, the over-fitting associated with maximum likelihood can be avoided by marginalizing (summing or integrating) over the model parameters instead of making point estimates of their values. Models can then be compared directly on the training data, without the need for a validation set. This allows all available data to be used for training and avoids the multiple training runs for each model associated with cross-validation. It also allows multiple complexity parameters to be determined simultaneously as part of the training process. For example, in Chapter 7 we shall introduce the *relevance vector machine*, which is a Bayesian model having one complexity parameter for every training data point.

The Bayesian view of model comparison simply involves the use of probabilities to represent uncertainty in the choice of model, along with a consistent application of the sum and product rules of probability. Suppose we wish to compare a set of  $L$  models  $\{\mathcal{M}_i\}$  where  $i = 1, \dots, L$ . Here a model refers to a probability distribution over the observed data  $\mathcal{D}$ . In the case of the polynomial curve-fitting problem, the distribution is defined over the set of target values  $\mathbf{t}$ , while the set of input values  $\mathbf{X}$  is assumed to be known. Other types of model define a joint distributions over  $\mathbf{X}$  and  $\mathbf{t}$ . We shall suppose that the data is generated from one of these models but we are uncertain which one. Our uncertainty is expressed through a prior probability distribution  $p(\mathcal{M}_i)$ . Given a training set  $\mathcal{D}$ , we then wish to evaluate the posterior distribution

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i). \quad (3.66)$$

The prior allows us to express a preference for different models. Let us simply assume that all models are given equal prior probability. The interesting term is the *model evidence*  $p(\mathcal{D}|\mathcal{M}_i)$  which expresses the preference shown by the data for

different models, and we shall examine this term in more detail shortly. The model evidence is sometimes also called the *marginal likelihood* because it can be viewed as a likelihood function over the space of models, in which the parameters have been marginalized out. The ratio of model evidences  $p(\mathcal{D}|\mathcal{M}_i)/p(\mathcal{D}|\mathcal{M}_j)$  for two models is known as a *Bayes factor* (Kass and Raftery, 1995).

Once we know the posterior distribution over models, the predictive distribution is given, from the sum and product rules, by

$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^L p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D}). \quad (3.67)$$

This is an example of a *mixture distribution* in which the overall predictive distribution is obtained by averaging the predictive distributions  $p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})$  of individual models, weighted by the posterior probabilities  $p(\mathcal{M}_i|\mathcal{D})$  of those models. For instance, if we have two models that are a-posteriori equally likely and one predicts a narrow distribution around  $t = a$  while the other predicts a narrow distribution around  $t = b$ , the overall predictive distribution will be a bimodal distribution with modes at  $t = a$  and  $t = b$ , not a single model at  $t = (a + b)/2$ .

A simple approximation to model averaging is to use the single most probable model alone to make predictions. This is known as *model selection*.

For a model governed by a set of parameters  $\mathbf{w}$ , the model evidence is given, from the sum and product rules of probability, by

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}. \quad (3.68)$$

## Chapter 11

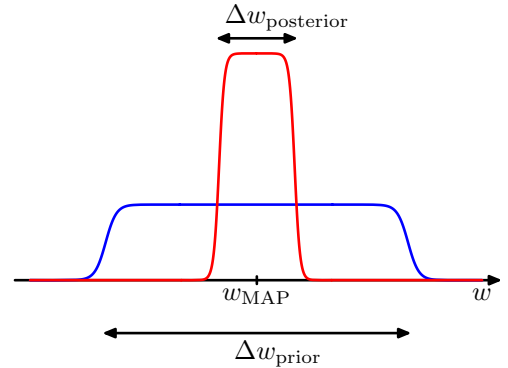
From a sampling perspective, the marginal likelihood can be viewed as the probability of generating the data set  $\mathcal{D}$  from a model whose parameters are sampled at random from the prior. It is also interesting to note that the evidence is precisely the normalizing term that appears in the denominator in Bayes' theorem when evaluating the posterior distribution over parameters because

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)}. \quad (3.69)$$

We can obtain some insight into the model evidence by making a simple approximation to the integral over parameters. Consider first the case of a model having a single parameter  $w$ . The posterior distribution over parameters is proportional to  $p(\mathcal{D}|w)p(w)$ , where we omit the dependence on the model  $\mathcal{M}_i$  to keep the notation uncluttered. If we assume that the posterior distribution is sharply peaked around the most probable value  $w_{\text{MAP}}$ , with width  $\Delta w_{\text{posterior}}$ , then we can approximate the integral by the value of the integrand at its maximum times the width of the peak. If we further assume that the prior is flat with width  $\Delta w_{\text{prior}}$  so that  $p(w) = 1/\Delta w_{\text{prior}}$ , then we have

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w) dw \simeq p(\mathcal{D}|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \quad (3.70)$$

**Figure 3.12** We can obtain a rough approximation to the model evidence if we assume that the posterior distribution over parameters is sharply peaked around its mode  $w_{\text{MAP}}$ .



and so taking logs we obtain

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right). \quad (3.71)$$

This approximation is illustrated in Figure 3.12. The first term represents the fit to the data given by the most probable parameter values, and for a flat prior this would correspond to the log likelihood. The second term penalizes the model according to its complexity. Because  $\Delta w_{\text{posterior}} < \Delta w_{\text{prior}}$  this term is negative, and it increases in magnitude as the ratio  $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$  gets smaller. Thus, if parameters are finely tuned to the data in the posterior distribution, then the penalty term is large.

For a model having a set of  $M$  parameters, we can make a similar approximation for each parameter in turn. Assuming that all parameters have the same ratio of  $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$ , we obtain

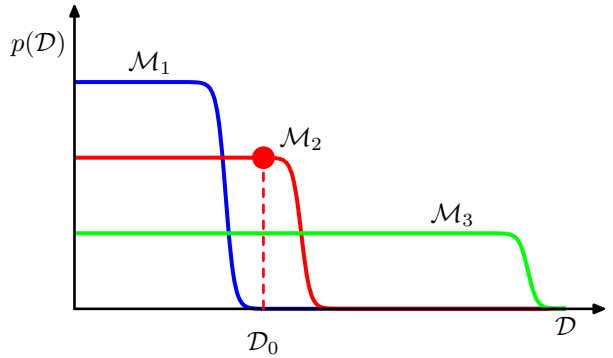
$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\text{MAP}}) + M \ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right). \quad (3.72)$$

Thus, in this very simple approximation, the size of the complexity penalty increases linearly with the number  $M$  of adaptive parameters in the model. As we increase the complexity of the model, the first term will typically decrease, because a more complex model is better able to fit the data, whereas the second term will increase due to the dependence on  $M$ . The optimal model complexity, as determined by the maximum evidence, will be given by a trade-off between these two competing terms. We shall later develop a more refined version of this approximation, based on a Gaussian approximation to the posterior distribution.

#### Section 4.4.1

We can gain further insight into Bayesian model comparison and understand how the marginal likelihood can favour models of intermediate complexity by considering Figure 3.13. Here the horizontal axis is a one-dimensional representation of the space of possible data sets, so that each point on this axis corresponds to a specific data set. We now consider three models  $\mathcal{M}_1$ ,  $\mathcal{M}_2$  and  $\mathcal{M}_3$  of successively increasing complexity. Imagine running these models generatively to produce example data sets, and then looking at the distribution of data sets that result. Any given

**Figure 3.13** Schematic illustration of the distribution of data sets for three models of different complexity, in which  $\mathcal{M}_1$  is the simplest and  $\mathcal{M}_3$  is the most complex. Note that the distributions are normalized. In this example, for the particular observed data set  $\mathcal{D}_0$ , the model  $\mathcal{M}_2$  with intermediate complexity has the largest evidence.



model can generate a variety of different data sets since the parameters are governed by a prior probability distribution, and for any choice of the parameters there may be random noise on the target variables. To generate a particular data set from a specific model, we first choose the values of the parameters from their prior distribution  $p(\mathbf{w})$ , and then for these parameter values we sample the data from  $p(\mathcal{D}|\mathbf{w})$ . A simple model (for example, based on a first order polynomial) has little variability and so will generate data sets that are fairly similar to each other. Its distribution  $p(\mathcal{D})$  is therefore confined to a relatively small region of the horizontal axis. By contrast, a complex model (such as a ninth order polynomial) can generate a great variety of different data sets, and so its distribution  $p(\mathcal{D})$  is spread over a large region of the space of data sets. Because the distributions  $p(\mathcal{D}|\mathcal{M}_i)$  are normalized, we see that the particular data set  $\mathcal{D}_0$  can have the highest value of the evidence for the model of intermediate complexity. Essentially, the simpler model cannot fit the data well, whereas the more complex model spreads its predictive probability over too broad a range of data sets and so assigns relatively small probability to any one of them.

Implicit in the Bayesian model comparison framework is the assumption that the true distribution from which the data are generated is contained within the set of models under consideration. Provided this is so, we can show that Bayesian model comparison will on average favour the correct model. To see this, consider two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  in which the truth corresponds to  $\mathcal{M}_1$ . For a given finite data set, it is possible for the Bayes factor to be larger for the incorrect model. However, if we average the Bayes factor over the distribution of data sets, we obtain the expected Bayes factor in the form

$$\int p(\mathcal{D}|\mathcal{M}_1) \ln \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} d\mathcal{D} \quad (3.73)$$

where the average has been taken with respect to the true distribution of the data. This quantity is an example of the *Kullback-Leibler* divergence and satisfies the property of always being positive unless the two distributions are equal in which case it is zero. Thus on average the Bayes factor will always favour the correct model.

We have seen that the Bayesian framework avoids the problem of over-fitting and allows models to be compared on the basis of the training data alone. However,

a Bayesian approach, like any approach to pattern recognition, needs to make assumptions about the form of the model, and if these are invalid then the results can be misleading. In particular, we see from Figure 3.12 that the model evidence can be sensitive to many aspects of the prior, such as the behaviour in the tails. Indeed, the evidence is not defined if the prior is improper, as can be seen by noting that an improper prior has an arbitrary scaling factor (in other words, the normalization coefficient is not defined because the distribution cannot be normalized). If we consider a proper prior and then take a suitable limit in order to obtain an improper prior (for example, a Gaussian prior in which we take the limit of infinite variance) then the evidence will go to zero, as can be seen from (3.70) and Figure 3.12. It may, however, be possible to consider the evidence ratio between two models first and then take a limit to obtain a meaningful answer.

In a practical application, therefore, it will be wise to keep aside an independent test set of data on which to evaluate the overall performance of the final system.

### 3.5. The Evidence Approximation

In a fully Bayesian treatment of the linear basis function model, we would introduce prior distributions over the hyperparameters  $\alpha$  and  $\beta$  and make predictions by marginalizing with respect to these hyperparameters as well as with respect to the parameters  $\mathbf{w}$ . However, although we can integrate analytically over either  $\mathbf{w}$  or over the hyperparameters, the complete marginalization over all of these variables is analytically intractable. Here we discuss an approximation in which we set the hyperparameters to specific values determined by maximizing the *marginal likelihood function* obtained by first integrating over the parameters  $\mathbf{w}$ . This framework is known in the statistics literature as *empirical Bayes* (Bernardo and Smith, 1994; Gelman *et al.*, 2004), or *type 2 maximum likelihood* (Berger, 1985), or *generalized maximum likelihood* (Wahba, 1975), and in the machine learning literature is also called the *evidence approximation* (Gull, 1989; MacKay, 1992a).

If we introduce hyperpriors over  $\alpha$  and  $\beta$ , the predictive distribution is obtained by marginalizing over  $\mathbf{w}$ ,  $\alpha$  and  $\beta$  so that

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) d\mathbf{w} d\alpha d\beta \quad (3.74)$$

where  $p(t|\mathbf{w}, \beta)$  is given by (3.8) and  $p(\mathbf{w}|\mathbf{t}, \alpha, \beta)$  is given by (3.49) with  $\mathbf{m}_N$  and  $\mathbf{S}_N$  defined by (3.53) and (3.54) respectively. Here we have omitted the dependence on the input variable  $\mathbf{x}$  to keep the notation uncluttered. If the posterior distribution  $p(\alpha, \beta|\mathbf{t})$  is sharply peaked around values  $\hat{\alpha}$  and  $\hat{\beta}$ , then the predictive distribution is obtained simply by marginalizing over  $\mathbf{w}$  in which  $\alpha$  and  $\beta$  are fixed to the values  $\hat{\alpha}$  and  $\hat{\beta}$ , so that

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \hat{\alpha}, \hat{\beta}) = \int p(t|\mathbf{w}, \hat{\beta}) p(\mathbf{w}|\mathbf{t}, \hat{\alpha}, \hat{\beta}) d\mathbf{w}. \quad (3.75)$$

From Bayes' theorem, the posterior distribution for  $\alpha$  and  $\beta$  is given by

$$p(\alpha, \beta | \mathbf{t}) \propto p(\mathbf{t} | \alpha, \beta) p(\alpha, \beta). \quad (3.76)$$

If the prior is relatively flat, then in the evidence framework the values of  $\hat{\alpha}$  and  $\hat{\beta}$  are obtained by maximizing the marginal likelihood function  $p(\mathbf{t} | \alpha, \beta)$ . We shall proceed by evaluating the marginal likelihood for the linear basis function model and then finding its maxima. This will allow us to determine values for these hyperparameters from the training data alone, without recourse to cross-validation. Recall that the ratio  $\alpha/\beta$  is analogous to a regularization parameter.

As an aside it is worth noting that, if we define conjugate (Gamma) prior distributions over  $\alpha$  and  $\beta$ , then the marginalization over these hyperparameters in (3.74) can be performed analytically to give a Student's t-distribution over  $\mathbf{w}$  (see Section 2.3.7). Although the resulting integral over  $\mathbf{w}$  is no longer analytically tractable, it might be thought that approximating this integral, for example using the Laplace approximation discussed (Section 4.4) which is based on a local Gaussian approximation centred on the mode of the posterior distribution, might provide a practical alternative to the evidence framework (Buntine and Weigend, 1991). However, the integrand as a function of  $\mathbf{w}$  typically has a strongly skewed mode so that the Laplace approximation fails to capture the bulk of the probability mass, leading to poorer results than those obtained by maximizing the evidence (MacKay, 1999).

Returning to the evidence framework, we note that there are two approaches that we can take to the maximization of the log evidence. We can evaluate the evidence function analytically and then set its derivative equal to zero to obtain re-estimation equations for  $\alpha$  and  $\beta$ , which we shall do in Section 3.5.2. Alternatively we use a technique called the expectation maximization (EM) algorithm, which will be discussed in Section 9.3.4 where we shall also show that these two approaches converge to the same solution.

### 3.5.1 Evaluation of the evidence function

The marginal likelihood function  $p(\mathbf{t} | \alpha, \beta)$  is obtained by integrating over the weight parameters  $\mathbf{w}$ , so that

$$p(\mathbf{t} | \alpha, \beta) = \int p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w} | \alpha) d\mathbf{w}. \quad (3.77)$$

One way to evaluate this integral is to make use once again of the result (2.115) for the conditional distribution in a linear-Gaussian model. Here we shall evaluate the integral instead by completing the square in the exponent and making use of the standard form for the normalization coefficient of a Gaussian.

From (3.11), (3.12), and (3.52), we can write the evidence function in the form

$$p(\mathbf{t} | \alpha, \beta) = \left( \frac{\beta}{2\pi} \right)^{N/2} \left( \frac{\alpha}{2\pi} \right)^{M/2} \int \exp \{ -E(\mathbf{w}) \} d\mathbf{w} \quad (3.78)$$

*Exercise 3.16*

*Exercise 3.17*

where  $M$  is the dimensionality of  $\mathbf{w}$ , and we have defined

$$\begin{aligned} E(\mathbf{w}) &= \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\ &= \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}. \end{aligned} \quad (3.79)$$

### Exercise 3.18

We recognize (3.79) as being equal, up to a constant of proportionality, to the regularized sum-of-squares error function (3.27). We now complete the square over  $\mathbf{w}$  giving

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{A} (\mathbf{w} - \mathbf{m}_N) \quad (3.80)$$

where we have introduced

$$\mathbf{A} = \alpha \mathbf{I} + \beta \Phi^T \Phi \quad (3.81)$$

together with

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^T \mathbf{m}_N. \quad (3.82)$$

Note that  $\mathbf{A}$  corresponds to the matrix of second derivatives of the error function

$$\mathbf{A} = \nabla \nabla E(\mathbf{w}) \quad (3.83)$$

and is known as the *Hessian matrix*. Here we have also defined  $\mathbf{m}_N$  given by

$$\mathbf{m}_N = \beta \mathbf{A}^{-1} \Phi^T \mathbf{t}. \quad (3.84)$$

Using (3.54), we see that  $\mathbf{A} = \mathbf{S}_N^{-1}$ , and hence (3.84) is equivalent to the previous definition (3.53), and therefore represents the mean of the posterior distribution.

### Exercise 3.19

The integral over  $\mathbf{w}$  can now be evaluated simply by appealing to the standard result for the normalization coefficient of a multivariate Gaussian, giving

$$\begin{aligned} & \int \exp \{-E(\mathbf{w})\} d\mathbf{w} \\ &= \exp \{-E(\mathbf{m}_N)\} \int \exp \left\{ -\frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{A} (\mathbf{w} - \mathbf{m}_N) \right\} d\mathbf{w} \\ &= \exp \{-E(\mathbf{m}_N)\} (2\pi)^{M/2} |\mathbf{A}|^{-1/2}. \end{aligned} \quad (3.85)$$

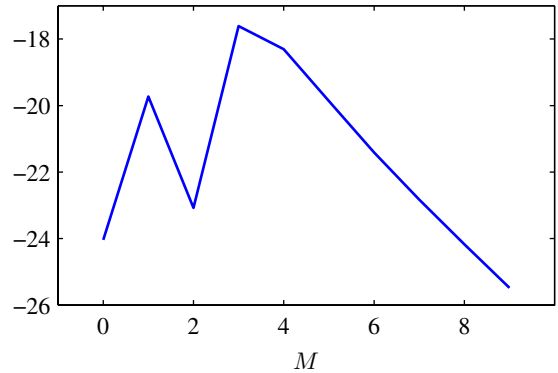
Using (3.78) we can then write the log of the marginal likelihood in the form

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\mathbf{m}_N) - \frac{1}{2} \ln |\mathbf{A}| - \frac{N}{2} \ln(2\pi) \quad (3.86)$$

which is the required expression for the evidence function.

Returning to the polynomial regression problem, we can plot the model evidence against the order of the polynomial, as shown in Figure 3.14. Here we have assumed a prior of the form (1.65) with the parameter  $\alpha$  fixed at  $\alpha = 5 \times 10^{-3}$ . The form of this plot is very instructive. Referring back to Figure 1.4, we see that the  $M = 0$  polynomial has very poor fit to the data and consequently gives a relatively low value

**Figure 3.14** Plot of the model evidence versus the order  $M$ , for the polynomial regression model, showing that the evidence favours the model with  $M = 3$ .



for the evidence. Going to the  $M = 1$  polynomial greatly improves the data fit, and hence the evidence is significantly higher. However, in going to  $M = 2$ , the data fit is improved only very marginally, due to the fact that the underlying sinusoidal function from which the data is generated is an odd function and so has no even terms in a polynomial expansion. Indeed, Figure 1.5 shows that the residual data error is reduced only slightly in going from  $M = 1$  to  $M = 2$ . Because this richer model suffers a greater complexity penalty, the evidence actually falls in going from  $M = 1$  to  $M = 2$ . When we go to  $M = 3$  we obtain a significant further improvement in data fit, as seen in Figure 1.4, and so the evidence is increased again, giving the highest overall evidence for any of the polynomials. Further increases in the value of  $M$  produce only small improvements in the fit to the data but suffer increasing complexity penalty, leading overall to a decrease in the evidence values. Looking again at Figure 1.5, we see that the generalization error is roughly constant between  $M = 3$  and  $M = 8$ , and it would be difficult to choose between these models on the basis of this plot alone. The evidence values, however, show a clear preference for  $M = 3$ , since this is the simplest model which gives a good explanation for the observed data.

### 3.5.2 Maximizing the evidence function

Let us first consider the maximization of  $p(\mathbf{t}|\alpha, \beta)$  with respect to  $\alpha$ . This can be done by first defining the following eigenvector equation

$$(\beta \Phi^T \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i. \quad (3.87)$$

From (3.81), it then follows that  $\mathbf{A}$  has eigenvalues  $\alpha + \lambda_i$ . Now consider the derivative of the term involving  $\ln |\mathbf{A}|$  in (3.86) with respect to  $\alpha$ . We have

$$\frac{d}{d\alpha} \ln |\mathbf{A}| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}. \quad (3.88)$$

Thus the stationary points of (3.86) with respect to  $\alpha$  satisfy

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^T \mathbf{m}_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha}. \quad (3.89)$$