Analysis of Sparse Bayesian Learning

Anita C. Faul **Michael E. Tipping**
Microsoft Research
St George House, 1 Guildhall St
Cambridge CB2 3NH, U.K.

Abstract

The recent introduction of the 'relevance vector machine' has effectively demonstrated how sparsity may be obtained in generalised linear models within a Bayesian framework. Using a particular form of Gaussian parameter prior, 'learning' is the maximisation, with respect to hyperparameters, of the *marginal likelihood* of the data. This paper studies the properties of that objective function, and demonstrates that conditioned on an individual hyperparameter, the marginal likelihood has a unique maximum which is computable in closed form. It is further shown that if a derived 'sparsity criterion' is satisfied, this maximum is exactly equivalent to 'pruning' the corresponding parameter from the model.

1 Introduction

We consider the approximation, from a training sample, of real-valued functions, a task variously referred to as prediction, regression, interpolation or function approximation. Given a set of data $\{\mathbf{x}_n, t_n\}_{n=1}^N$ the 'target' samples $t_n = f(\mathbf{x}_n) + \epsilon_n$ are conventionally considered to be realisations of a deterministic function f, potentially corrupted by some additive noise process. This function f will be modelled by a linearly-weighted sum of M fixed basis functions $\{\boldsymbol{\phi}_m(\mathbf{x})\}_{m=1}^M$:

$$\widehat{f}(\mathbf{x}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}), \tag{1}$$

and the objective is to infer values of the parameters/weights $\{w_m\}_{m=1}^M$ such that $\widehat{f}$ is a 'good' approximation of f.

While accuracy in function approximation is generally universally valued, there has been significant recent interest [2, 9, 3, 5]) in the notion of *sparsity*, a consequence of learning algorithms which set significant numbers of the parameters w_m to zero.

A methodology which effectively combines both these measures of merit is that of 'sparse Bayesian learning', briefly reviewed in Section 2, and which was the basis for the recent introduction of the *relevance vector machine* (RVM) and related models [6, 1, 7]. This model exhibits some very compelling properties, in particular a dramatic degree of sparseness even in the case of highly overcomplete basis sets

($M \gg N$). The sparse Bayesian learning algorithm essentially involves the maximisation of a marginalised likelihood function with respect to hyperparameters in the model prior. In the RVM, this was achieved through re-estimation equations, the behaviour of which was not fully understood. In this paper we present further relevant theoretical analysis of the properties of the marginal likelihood which gives a much fuller picture of the nature of the model and its associated learning procedure. This is detailed in Section 3, and we close with a summary of our findings and discussion of their implications in Section 4 (and which, to avoid repetition here, the reader may wish to preview at this point).

2 Sparse Bayesian Learning

We now very briefly review the methodology of sparse Bayesian learning, more comprehensively described elsewhere [6]. To simplify and generalise the exposition, we omit to notate any functional dependence on the inputs $\mathbf{x}$ and combine quantities defined over the training set and basis set within N- and M-vectors respectively. Using this representation, we first write the generative model as:

$$\mathbf{t} = \mathbf{f} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$, $\mathbf{f} = (f_1, \ldots, f_N)^{\mathrm{T}}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^{\mathrm{T}}$. The approximator is then written as:

$$\widehat{\mathbf{f}} = \mathbf{\Phi}\mathbf{w}, \tag{3}$$

where $\mathbf{\Phi} = [\boldsymbol{\phi}_1 \ldots \boldsymbol{\phi}_M]$ is a general $N \times M$ design matrix with column vectors $\boldsymbol{\phi}_m$ and $\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$. Recall that in the context of (1), $\mathbf{\Phi}_{nm} = \phi_m(\mathbf{x}_n)$ and $\widehat{\mathbf{f}} = \{\widehat{f}(\mathbf{x}_1), \ldots \widehat{f}(\mathbf{x}_N)\}^{\mathrm{T}}$.

The sparse Bayesian framework assumes an independent zero-mean Gaussian noise model, with variance σ^2, giving a multivariate Gaussian likelihood of the target vector $\mathbf{t}$:

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi)^{-N/2}\sigma^{-N} \exp\left\{-\frac{\|\mathbf{t} - \widehat{\mathbf{f}}\|^2}{2\sigma^2}\right\}. \tag{4}$$

The prior over the parameters is mean-zero Gaussian:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = (2\pi)^{-M/2} \prod_{m=1}^{M} \alpha_m^{1/2} \exp\left(-\frac{\alpha_m w_m^2}{2}\right), \tag{5}$$

where the key to the model sparsity is the use of M independent hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^{\mathrm{T}}$, one per weight (or basis vector), which moderate the strength of the prior. Given $\boldsymbol{\alpha}$, the *posterior* parameter distribution is Gaussian and given via Bayes' rule as $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{\Sigma})$ with

$$\mathbf{\Sigma} = (\mathbf{A} + \sigma^{-2}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1} \qquad \boldsymbol{\mu} = \sigma^{-2}\mathbf{\Sigma}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}, \tag{6}$$

and $\mathbf{A}$ defined as $\mathrm{diag}\,(\alpha_1, \ldots, \alpha_M)$. Sparse Bayesian learning can then be formulated as a *type-II maximum likelihood* procedure, in that objective is to maximise the *marginal likelihood*, or equivalently, its logarithm $\mathcal{L}(\boldsymbol{\alpha})$ with respect to the hyperparameters $\boldsymbol{\alpha}$:

$$\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \log \int_{-\infty}^{\infty} p(\mathbf{t}|\mathbf{w}, \sigma^2)\; p(\mathbf{w}|\boldsymbol{\alpha})\; d\mathbf{w}, \tag{7}$$

$$= -\frac{1}{2}\left[N \log 2\pi + \log|\mathbf{C}| + \mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}\right], \tag{8}$$

with $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}$.

Once most-probable values $\boldsymbol{\alpha}_{\mathrm{MP}}$ have been found[1], in practice they can be plugged into (6) to give a posterior mean (most probable) point estimate for the parameters $\boldsymbol{\mu}_{\mathrm{MP}}$ and from that a mean final approximator: $\widehat{\mathbf{f}}_{\mathrm{MP}} = \boldsymbol{\Phi}\boldsymbol{\mu}_{\mathrm{MP}}$.

Empirically, the local maximisation of the marginal likelihood (8) with respect to $\boldsymbol{\alpha}$ has been seen to work highly effectively [6, 1, 7]. Accurate predictors may be realised, which are typically highly sparse as a result of the maximising values of many hyperparameters being infinite. From (6) this leads to a parameter posterior infinitely peaked at zero for many weights w_m with the consequence that $\boldsymbol{\mu}_{\mathrm{MP}}$ correspondingly comprises very few non-zero elements.

However, the learning procedure in [6] relied upon heuristic re-estimation equations for the hyperparameters, the behaviour of which was not well characterised. Also, little was known regarding the properties of (8), the validity of the local maximisation thereof and importantly, and perhaps most interestingly, the conditions under which α-values would become infinite. We now give, through a judicious re-writing of (8), a more detailed analysis of the sparse Bayesian learning procedure.

3 Properties of the Marginal Likelihood $\mathcal{L}(\boldsymbol{\alpha})$

3.1 A convenient re-writing

We re-write $\mathbf{C}$ from (8) in a convenient form to analyse the dependence on a single hyperparameter α_i:

$$\begin{aligned}\mathbf{C} &= \sigma^2\mathbf{I} + \sum_m \alpha_m \boldsymbol{\phi}_m \boldsymbol{\phi}_m^{\mathrm{T}}, = \sigma^2\mathbf{I} + \sum_{m\neq i} \alpha_m^{-1}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathrm{T}} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathrm{T}},\\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathrm{T}},\end{aligned}\tag{9}$$

where we have defined $\mathbf{C}_{-i} = \sigma^2\mathbf{I} + \sum_{m\neq i}\alpha_m^{-1}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathrm{T}}$ as the covariance matrix with the influence of basis vector $\boldsymbol{\phi}_i$ removed, equivalent also to $\alpha_i = \infty$.

Using established matrix determinant and inverse identities, (9) allows us to write the terms of interest in $\mathcal{L}(\boldsymbol{\alpha})$ as:

$$|\mathbf{C}| = |\mathbf{C}_{-i}|\,|1 + \alpha_i^{-1}\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i|,\tag{10}$$

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i},\tag{11}$$

which gives

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}) &= -\frac{1}{2}\Big[N\log(2\pi) + \log|\mathbf{C}_{-i}| + \mathbf{t}^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t}\\ &\qquad - \log\alpha_i + \log(\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i) - \frac{(\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t})^2}{\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i}\Big],\\ &= \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \frac{1}{2}\Big[\log\alpha_i - \log(\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i) + \frac{(\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t})^2}{\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i}\Big],\\ &= \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i),\end{aligned}\tag{12}$$

where $\mathcal{L}(\boldsymbol{\alpha}_{-i})$ is the log marginal likelihood with α_i (and thus w_i and $\boldsymbol{\phi}_i$) removed from the model and we have now isolated the terms in α_i in the function $\ell(\alpha_i)$.

[1] The most-probable noise variance σ^2_{MP} can also be directly and successfully estimated from the data [6], but for clarity in this paper, we assume without prejudice to our results that its value is fixed.

3.2 First derivatives of $\mathcal{L}(\boldsymbol{\alpha})$

Previous results. In [6], based on earlier results from [4], the gradient of the marginal likelihood was computed as:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{1}{2}\left[\frac{1}{\alpha_i} - \mu_i^2 - \Sigma_{ii}\right], \tag{13}$$

with μ_i the i-th element of $\boldsymbol{\mu}$ and Σ_{ii} the i-th diagonal element of $\boldsymbol{\Sigma}$. This then leads to re-estimation updates for α_i in terms of μ_i and Σ_{ii} where, disadvantageously, these latter terms are themselves functions of α_i.

A new, simplified, expression. In fact, by instead differentiating (12) directly, (13) can be seen to be equivalent to:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{\partial \ell(\alpha_i)}{\partial \alpha_i} = \frac{1}{2}\left[\frac{1}{\alpha_i} - \frac{1}{\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i} - \frac{(\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t})^2}{(\alpha_i + \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i)^2}\right], \tag{14}$$

where advantageously, α_i now occurs only explicitly since $\mathbf{C}_{-i}$ is independent of α_i. For convenience, we combine terms and re-write (14) as:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{\alpha_i^{-1}S_i^2 - (Q_i^2 - S_i)}{2(\alpha_i + S_i)^2}, \tag{15}$$

where, for simplification of this and forthcoming expressions, we have defined:

$$Q_i \triangleq \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t}, \qquad S_i \triangleq \boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i. \tag{16}$$

The term Q_i can be interpreted as a 'quality' factor: a measure of how well $\boldsymbol{\phi}_i$ increases $\mathcal{L}(\boldsymbol{\alpha})$ by helping to explain the data, while S_i is a 'sparsity' factor which measures how much the inclusion of $\boldsymbol{\phi}_i$ serves to decrease $\mathcal{L}(\boldsymbol{\alpha})$ through 'inflating' $\mathbf{C}$ (*i.e.* adding to the normalising factor).

3.3 Stationary points of $\mathcal{L}(\boldsymbol{\alpha})$

Equating (15) to zero indicates that stationary points of the marginal likelihood occur both at $\alpha_i = +\infty$ (note that, being an inverse variance, α_i must be positive) and for:

$$\alpha_i = \frac{S_i^2}{Q_i^2 - S_i}, \tag{17}$$

subject to $Q_i^2 > S_i$ as a consequence again of $\alpha_i > 0$.

Since the right-hand-side of (17) is independent of α_i, we may find the stationary points of $\ell(\alpha_i)$ analytically without iterative re-estimation. To find the nature of those stationary points, we consider the second derivatives.

3.4 Second derivatives of $\mathcal{L}(\boldsymbol{\alpha})$

3.4.1 With respect to α_i

Differentiating (15) a second time with respect to α_i gives:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i^2} = \frac{-\alpha_i^{-2}S_i^2(\alpha_i + S_i)^2 - 2(\alpha_i + S_i)\left[\alpha_i^{-1}S_i^2 - (Q_i^2 - S_i)\right]}{2(\alpha_i + S_i)^4}, \tag{18}$$

and we now consider (18) for both finite- and infinite-α_i stationary points.

Finite α. In this case, for stationary points given by (17), we note that the second term in the numerator in (18) is zero, giving:

$$\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i^2}\right|_{\alpha_i = \frac{S_i^2}{Q_i^2 - S_i}} = \frac{-S_i^2}{2\alpha_i^2(\alpha_i + S_i)^2}. \tag{19}$$

We see that (19) is always negative, and therefore $\ell(\alpha_i)$ has a *maximum*, which must be unique, for $Q_i^2 - S_i > 0$ and α_i given by (17).

Infinite α. For this case, (18) and indeed, all further derivatives, are uninformatively zero at $\alpha_i = \infty$, but from (15) we can see that as $\alpha_i \to \infty$, the sign of the gradient is given by the sign of $-(Q_i^2 - S_i)$.

If $Q_i^2 - S_i > 0$, then the gradient at $\alpha_i = \infty$ is negative so as α_i decreases $\ell(\alpha_i)$ must increase to its unique maximum given by (17). It follows that $\alpha_i = \infty$ is thus a minimum. Conversely, if $Q_i^2 - S_i < 0$, $\alpha_i = \infty$ is the unique maximum of $\ell(\alpha_i)$. If $Q_i^2 - S_i = 0$, then this maximum and that given by (17) coincide.

We now have a full characterisation of the marginal likelihood as a function of a single hyperparameter, which is illustrated in Figure 1.

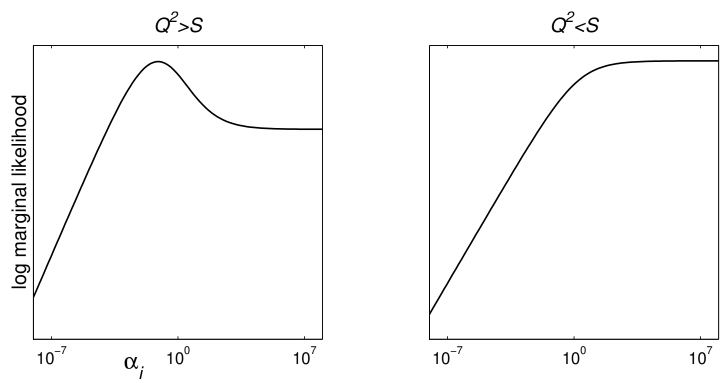

Figure 1: Example plots of $\ell(\alpha_i)$ against α_i (on a log scale) for $Q_i^2 > S_i$ (left), showing the single maximum at finite α_i, and $Q_i^2 < S_i$ (right), showing the maximum at $\alpha_i = \infty$.

3.4.2 With respect to α_j, $j \neq i$

To obtain the off-diagonal terms of the second derivative (Hessian) matrix, it is convenient to manipulate (15) to express it in terms of $\mathbf{C}$. From (11) we see that

$$\boldsymbol{\phi}_i^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{t} = \frac{\alpha_i Q_i}{\alpha_i + S_i}, \quad \text{and} \quad \boldsymbol{\phi}_i^{\mathrm{T}} \mathbf{C}^{-1} \boldsymbol{\phi}_i = \frac{\alpha_i S_i}{\alpha_i + S_i}. \tag{20}$$

Utilising these identities in (15) gives:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{1}{2\alpha_i^2} \left[\boldsymbol{\phi}_i^{\mathrm{T}} \mathbf{C}^{-1} \boldsymbol{\phi}_i - (\boldsymbol{\phi}_i^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{t})^2 \right]. \tag{21}$$

We now write:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_i \partial \alpha_j} = \nabla_{ij}^2 + \delta_{ij} \nabla_{ii}^2, \tag{22}$$

where δ_{ij} is the Kronecker 'delta' function, allowing us to separate out the additional (diagonal) term that appears only when $i = j$.

Writing, similarly to (9) earlier, $\mathbf{C} = \mathbf{C}_{-j} + \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^{\mathrm{T}}$, substituting into (21) and differentiating with respect to α_j gives:

$$\begin{aligned}
\nabla_{ij}^2 &= \frac{1}{2\alpha_i^2}\left[\frac{(\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j)^2}{(\alpha_j + \boldsymbol{\phi}_j^{\mathrm{T}}\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j)^2} - 2(\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t})\frac{(\boldsymbol{\phi}_i^{\mathrm{T}}\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j)(\boldsymbol{\phi}_j^{\mathrm{T}}\mathbf{C}_{-j}^{-1}\mathbf{t})}{(\alpha_j + \boldsymbol{\phi}_j^{\mathrm{T}}\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j)^2}\right], \\
&= \frac{\boldsymbol{\phi}_i^T\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j}{2\alpha_i^2(\alpha_j + \boldsymbol{\phi}_j^T\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j)^2}\Big[\boldsymbol{\phi}_i^T\mathbf{C}_{-j}^{-1}\boldsymbol{\phi}_j - 2(\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\mathbf{t})(\boldsymbol{\phi}_j^T\mathbf{C}_{-j}^{-1}\mathbf{t})\Big], \\
&= \frac{\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\boldsymbol{\phi}_j}{2\alpha_i^2\alpha_j^2}\Big[\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\boldsymbol{\phi}_j - 2(\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\mathbf{t})(\boldsymbol{\phi}_j^T\mathbf{C}^{-1}\mathbf{t})\Big], \qquad (23)
\end{aligned}$$

while we have

$$\nabla_{ii}^2 = -\frac{1}{\alpha_i^3}\left[\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\boldsymbol{\phi}_i - (\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\mathbf{t})^2\right]. \qquad (24)$$

If all hyperparameters α_i are individually set to their maximising values, *i.e.* $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{MP}}$ such that all $\partial\mathcal{L}(\boldsymbol{\alpha})/\partial\alpha_i = 0$, then even if all $\partial^2\mathcal{L}(\boldsymbol{\alpha})/\partial\alpha_i^2$ are negative, there may still be a non-axial direction in which the likelihood could be increasing. We now rule out this possibility by showing that the Hessian is negative semi-definite.

First, we note from (21) that if $\partial\mathcal{L}(\boldsymbol{\alpha})/\partial\alpha_i = 0$, $\nabla_{ii}^2 = 0$. Then, if $\mathbf{v}$ is a generic nonzero direction vector:

$$\begin{aligned}
\mathbf{v}^{\mathrm{T}}\left[\frac{\partial^2\mathcal{L}(\boldsymbol{\alpha})}{\partial\alpha_i\partial\alpha_j}\right]\mathbf{v} &= \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{v_i v_j}{\alpha_i^2\alpha_j^2}\left[(\mathbf{C}^{-1/2}\boldsymbol{\phi}_i)^{\mathrm{T}}(\mathbf{C}^{-1/2}\boldsymbol{\phi}_j)\right]^2 \\
&\quad - \left(\sum_{i=1}^{M}\frac{|v_i|}{\alpha_i^2}|\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\mathbf{t}|\,\|\mathbf{C}^{-1/2}\boldsymbol{\phi}_i\|\right)^2, \\
&\leq \frac{1}{2}\left(\sum_{i=1}^{M}\frac{|v_i|}{\alpha_i^2}\|\mathbf{C}^{-1/2}\boldsymbol{\phi}_i\|^2\right)^2 \\
&\quad - \left(\sum_{i=1}^{M}\frac{|v_i|}{\alpha_i^2}|\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\mathbf{t}|\,\|\mathbf{C}^{-1/2}\boldsymbol{\phi}_i\|\right)^2, \qquad (25)
\end{aligned}$$

where we use the Cauchy-Schwarz inequality. If the gradient vanishes, then for all $i = 1, \ldots, M$ either $\alpha_i = \infty$, or from (21), $\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\boldsymbol{\phi}_i = (\boldsymbol{\phi}_i^T\mathbf{C}^{-1}\mathbf{t})^2$. It follows directly from (25) that the Hessian is negative semi-definite, with (25) only zero where $\mathbf{v}$ is orthogonal to all finite α values.

4 Summary

Sparse Bayesian learning proposes the iterative maximisation of the marginal likelihood function $\mathcal{L}(\boldsymbol{\alpha})$ with respect to the hyperparameters $\boldsymbol{\alpha}$. Our analysis has shown the following:

I. As a function of an individual hyperparameter α_i, $\mathcal{L}(\boldsymbol{\alpha})$ has a unique maximum computable in closed-form. (This maximum is, of course, dependent on the values of all other hyperparameters.)

II. If the criterion $Q_i^2 - S_i$ (defined in Section 3.2) is negative, this maximum occurs at $\alpha_i = \infty$, equivalent to the removal of basis function i from the model.

III. The point where all individual marginal likelihood functions $\ell(\alpha_i)$ are maximised is a joint maximum (not necessarily unique) over all α_i.

These results imply the following consequences.

- From I, we see that if we update, in any arbitrary order, the α_i parameters using (17), we are guaranteed to increase the marginal likelihood at each step, unless already at a maximum. Furthermore, we would expect these updates to be more efficient than those given in [6], which individually only increase, not maximise, $\ell(\alpha_i)$.
- Result III indicates that sequential optimisation of individual α_i cannot lead to a stationary point from which a joint maximisation over all $\boldsymbol{\alpha}$ may have escaped. (*i.e.* the stationary point is not a saddle point.)
- The result II confirms the qualitative argument and empirical observation that many $\alpha_i \to \infty$ as a result of the optimisation procedure in [6]. The inevitable implication of finite numerical precision prevented the genuine sparsity of the model being verified in those earlier simulations.
- We conclude by noting that the maximising hyperparameter solution (17) remains valid if α_i is already infinite. This means that *basis functions not even in the model* can be assessed and their corresponding hyperparameters updated if desired. So as well as the facility to increase $\mathcal{L}(\boldsymbol{\alpha})$ through the 'pruning' of basis functions if $Q_i^2 - S_i \leq 0$, new basis functions can be introduced if $\alpha_i = \infty$ but $Q_i^2 - S_i > 0$. This has highly desirable computational consequences which we are exploiting to obtain a powerful 'constructive' approximation algorithm [8].

References

[1] C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In C. Boutilier and M. Goldszmidt, editors, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53. Morgan Kaufmann, 2000.

[2] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, 1995.

[3] Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalisation. In L. Niklasson, M. Bodén, and T. Ziemske, editors, *Proceedings of the Eighth International Conference on Artificial Neural Networks (ICANN98)*, pages 201–206. Springer, 1998.

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

[5] A. J. Smola, B. Schölkopf, and G. Rätsch. Linear programs for automatic accuracy control in regression. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, pages 575–580, 1999.

[6] M. E. Tipping. The Relevance Vector Machine. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 652–658. MIT Press, 2000.

[7] M. E. Tipping. Sparse kernel principal component analysis. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[8] M. E. Tipping and A. C. Faul. Bayesian pursuit. Submitted to NIPS*01.

[9] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.