In: Fourth International Conference on Artificial Neural Networks, Churchill College, University of Cambridge, UK. IEE Conference Publication No. 409 pp 160-165, (1995)

ON THE RELATIONSHIP BETWEEN BAYESIAN ERROR BARS AND THE INPUT DATA DENSITY

C K I Williams, C Qazaz, C M Bishop and H Zhu

Neural Computing Research Group, Aston University, UK.

ABSTRACT

We investigate the dependence of Bayesian error bars on the distribution of data in input space. For generalized linear regression models we derive an upper bound on the error bars which shows that, in the neighbourhood of the data points, the error bars are substantially reduced from their prior values. For regions of high data density we also show that the contribution to the output variance due to the uncertainty in the weights can exhibit an approximate inverse proportionality to the probability density. Empirical results support these conclusions.

1 INTRODUCTION

When given a prediction, it is also very useful to be given some idea of the "error bars" associated with that prediction. Error bars arise naturally in a Bayesian treatment of neural networks and are made up of two terms, one due to the posterior weight uncertainty, and the other due to the intrinsic noise in the data[1]. As the two contributions are independent, we have

$$\sigma_y^2(\boldsymbol{x}) = \sigma_w^2(\boldsymbol{x}) + \sigma_\nu^2(\boldsymbol{x}) \quad (1)$$

where $\sigma_w^2(\boldsymbol{x})$ is the variance of the output due to weight uncertainty and $\sigma_\nu^2(\boldsymbol{x})$ is the variance of the intrinsic noise.

Under the assumption that the posterior in weight space can be approximated by a Gaussian (MacKay (1)), we have

$$\sigma_w^2(\boldsymbol{x}) = \boldsymbol{g}^T(\boldsymbol{x})\boldsymbol{A}^{-1}\boldsymbol{g}(\boldsymbol{x}) \quad (2)$$

where $\boldsymbol{A}$ is the Hessian matrix of the model and $\boldsymbol{g}(\boldsymbol{x}) = \partial y(\boldsymbol{x};\boldsymbol{w})/\partial \boldsymbol{w}$ is the vector of the derivatives of the output with respect to the weight parameters in the network. $\boldsymbol{A}$ contains contributions from both the prior distribution on the weights and the effect of the training data.

Although the weight uncertainty component of the error bar is given by equation 2, the dependence of this quantity on the location of the training points is not at all obvious. Intuitively we would expect the error bars from the prior (i.e. before any data is seen) to be quite large, and that the effect of the training data would be to reduce the magnitude of the error bars for those regions of the input space close to the data points, while leaving large error bars further away. The purpose of this paper is to provide theoretical insights to support this intuition. In particular, our analysis focusses on generalized linear regression (such as radial basis function networks with fixed basis function parameters) and allows us to quantify the extent of the reduction and the length scale over which it occurs.

We also show that the relationship $\sigma_w^2(\boldsymbol{x}) \simeq \sigma_\nu^2[Np(\boldsymbol{x})V(\boldsymbol{x})]^{-1}$ holds approximately, where $p(\boldsymbol{x})$ is the density of the data in the input space, N is the number of data points in the training set and $V(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ that measures a volume in the input space. This relationship pertains to the "high-data" limit where the effect of the data overwhelms the prior in the Hessian.

2 GENERALIZED LINEAR REGRESSION

Consider a generalized linear regression (GLR) model of the form

$$y(\boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{w} = \sum_{j=1}^{m} w_j\phi_j(\boldsymbol{x}) \quad (3)$$

where $j = 1,...,m$ labels the basis functions $\{\phi_j\}$ of the model. Given a data set $D = ((\boldsymbol{x}_1,t_1),(\boldsymbol{x}_2,t_2),\ldots,(\boldsymbol{x}_N,t_N))$, a squared error function with noise variance[2] $\sigma_\nu^2 = \beta^{-1}$ and a regularizer of the form $\alpha\boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}/2$, the posterior mean value of the weights $\hat{\boldsymbol{w}}$ is the choice of $\boldsymbol{w}$ that minimizes the quadratic form

$$E = \frac{1}{2\sigma_\nu^2}\sum_i^N\left(t_i - \sum_j w_j\phi_j(\boldsymbol{x}_i)\right)^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w} \quad (4)$$

so that $\hat{\boldsymbol{w}}$ is the solution of

$$(\beta\boldsymbol{B} + \alpha\boldsymbol{S})\hat{\boldsymbol{w}} = \beta\Phi^T\boldsymbol{t} \quad (5)$$

[1] If the network used is not the correct generative model for the data there will be a third component due to model mis-specification; we do not discuss this further in this paper.

[2] In this section we assume that σ_ν^2 is independent of x. This assumption can be easily relaxed, but at the expense of somewhat more complicated notation.

where Φ is the $n \times m$ design matrix

$$\Phi = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_1) & \cdots & \phi_m(\boldsymbol{x}_1) \\ \phi_1(\boldsymbol{x}_2) & \phi_2(\boldsymbol{x}_2) & \cdots & \phi_m(\boldsymbol{x}_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \phi_2(\boldsymbol{x}_n) & \cdots & \phi_m(\boldsymbol{x}_n) \end{pmatrix} \tag{6}$$

$\boldsymbol{B} = \Phi^T \Phi$ and $\boldsymbol{t}$ is the vector of targets. Writing $\boldsymbol{A} = \beta \boldsymbol{B} + \alpha \boldsymbol{S}$, we find

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x})\hat{\boldsymbol{w}} = \beta \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{A}^{-1}\Phi^T \boldsymbol{t} \overset{def}{=} \boldsymbol{k}^T(\boldsymbol{x})\boldsymbol{t} \tag{7}$$

where $\hat{y}(\boldsymbol{x})$ is the function obtained from equation 3 using $\hat{\boldsymbol{w}}$ as the weight vector. Equation 7 defines the effective kernel $\boldsymbol{k}(\boldsymbol{x})$ and makes it clear that $\hat{y}(\boldsymbol{x})$ can be written as a linear combination of the target values, i.e. it is a *linear smoother* (see, e.g. Hastie and Tibshirani (2)).

The contribution of the uncertainty of the weights to the variance of the prediction is given from equation 2 by

$$\sigma_w^2(\boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{A}^{-1}\boldsymbol{\phi}(\boldsymbol{x}) \tag{8}$$

Note that for generalized linear regression this expression is exact, and that the error bars (given σ_ν^2) are independent of the targets.

3 ERROR BARS FOR GLR

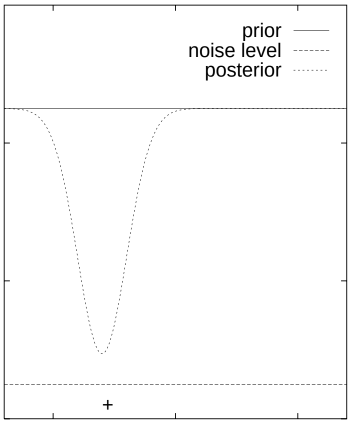

Figure 1: A schematic illustration of the effect of one data point on $\sigma_y^2(x)$. The posterior variance is reduced from its prior level in the neighbourhood of the data point (+), but remains above the noise level.

In this section we analyze the response of the prior variance to the addition of the data points. In particular we show that the effect of a single data point is to pull the $\sigma_y^2(\boldsymbol{x})$ surface down to a value less than $2\sigma_\nu^2(\boldsymbol{x})$[3] at and nearby to the data point, and that the length scale over which this effect operates is determined by the prior covariance function

$$C(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{A}_0^{-1}\boldsymbol{\phi}(\boldsymbol{x}') \tag{9}$$

where $\boldsymbol{A}_0 = \alpha \boldsymbol{S}$.

The main tool used in this analysis is the effect of adding just one data point. A schematic illustration of this effect is shown in Figure 1. The variance due to the prior is quite large (and roughly constant over $\boldsymbol{x}$-space). Adding a single data point pulls down the variance in its neighbourhood (but not as far as the σ_ν^2 limit).

Figure 1 is relevant because we can show (see Appendix A.1) that $\sigma_y^2(\boldsymbol{x})$, when all data points are used to compute the Hessian, is never greater than $\sigma_y^2(\boldsymbol{x})$ when any subset of the data points are used, and hence the surface pertaining to any particular data point is an upper bound on the overall surface.

To obtain a bound on the depth of the dip, consider the case when there is only one data point (at $\boldsymbol{x} = \boldsymbol{x}_i$), so that the Hessian is given by $\boldsymbol{A} = \boldsymbol{A}_0 + \beta(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}^T(\boldsymbol{x}_i)$. Using the identity

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}} \tag{10}$$

it is easy to show that

$$\sigma_{w|\boldsymbol{x}_i}^2(\boldsymbol{x}_i) = \sigma_\nu^2(\boldsymbol{x}_i)\frac{r_i}{1 + r_i} \tag{11}$$

where $\sigma_{w|\boldsymbol{x}_i}^2$ denotes the posterior weight uncertainty surface due to a data point at $\boldsymbol{x}_i$ and

$$r_i = \frac{\boldsymbol{\phi}^T(\boldsymbol{x}_i)\boldsymbol{A}_0^{-1}\boldsymbol{\phi}(\boldsymbol{x}_i)}{\sigma_\nu^2(\boldsymbol{x}_i)} \tag{12}$$

i.e. r_i is the ratio of the prior to noise variances at the point $\boldsymbol{x}_i$. For any positive value of z, the function $z/(1+z)$ lies between 0 and 1, hence we see that the σ_w^2 contribution to the error bars must always be less than $\sigma_\nu^2(\boldsymbol{x})$ at a data point. Typically the noise variance is much smaller than the prior variance, so $r_i \gg 1$.

Further evidence that σ_w^2 at any data point is of the order of $\sigma_\nu^2(\boldsymbol{x})$ is provided by the calculation in appendix A.2 which shows that the average of $\sigma_w^2(\boldsymbol{x}_i)$ at the data points is less than $\sigma_\nu^2 m/N$, where m is the number of weights in the model and N is the number of data points.

For a single data point at $\boldsymbol{x}_i$, we can use equation 10 to show that

$$\sigma_{w|\boldsymbol{x}_i}^2(\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{x}) - \frac{(C(\boldsymbol{x}, \boldsymbol{x}_i))^2}{1 + C(\boldsymbol{x}_i, \boldsymbol{x}_i)} \tag{13}$$

Hence the width of the depression in the variance surface is related to the characteristic length scale of the

[3] The analysis in this section permits the noise level to vary as a function of x.

prior covariance function $C(\boldsymbol{x}, \boldsymbol{x}')$. It is also possible to show that if a test point $\boldsymbol{x}$ has zero covariance $C(\boldsymbol{x}, \boldsymbol{x}_i)$ with all of the training points $\{\boldsymbol{x}_i\}$, then its posterior variance will be equal its prior variance. We are currently exploring the properties of $C(\boldsymbol{x}, \boldsymbol{x}')$ for different weight priors and choices of basis functions. However, we note that a simple diagonal prior $\boldsymbol{S} = \boldsymbol{I}$ as used by some authors is not in general a very sensible prior, because if the type of basis functions used (e.g. Gaussians, tanh functions etc.) is changed, then the covariance structure of the prior also changes. More sensibly, the weight prior should be chosen so as to approximate some desired prior covariance function $C(\boldsymbol{x}, \boldsymbol{x}')$.

4 DENSITY DEPENDENCE OF $\sigma_w^2(\boldsymbol{x})$

As we have already noted, error bars on network predictions would be expected to be relatively large in regions of input space for which there is little data, and smaller in regions of high data density. In this section, we establish an approximate proportionality between the variance due to weight uncertainty and the inverse of the probability density of training data, valid in regions of high data density. A relationship of this kind was conjectured in Bishop (3).

We first consider a special case of the class of generalized linear models where the basis functions are non-overlapping bin (or "top-hat") activation functions[4]. Let the i^{th} basis function have height h_i and a d-dimensional "base area" of V_i, where d is the dimensionality of $\boldsymbol{x}$. If we choose a diagonal prior ($\boldsymbol{S} = \boldsymbol{I}$) then the Hessian is diagonal and thus easy to invert.

$$\boldsymbol{B}_{ij} = \sum_{q=1}^{N} \phi_i(\boldsymbol{x}_q)\phi_j(\boldsymbol{x}_q) = \begin{cases} n_i h_i^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where n_i is the number of data points falling in bin i. From equation (8) the error bars associated with a point $\boldsymbol{x}$ which falls into the i^{th} bin are given by

$$\sigma_w^2(\boldsymbol{x}) = \frac{1}{\alpha/h_i^2 + \beta n_i} \tag{15}$$

As usual, the effect of the prior is to reduce the size of error bar compared to the case where it is not present. In the limit of $\alpha \to 0$ we have

$$\sigma_w^2(\boldsymbol{x}) = \frac{\sigma_\nu^2}{n_i} = \frac{\sigma_\nu^2}{N V_i \, \hat{p}(\boldsymbol{x})} \tag{16}$$

where N is the total number of data points and $\hat{p}(\boldsymbol{x})$ is the histogram estimate of the density inside the bin containing $\boldsymbol{x}$. Equation (16) demonstrates that for this kind of model the error bars are inversely proportional to the input density and to the volume factor V_i. It also shows that we can understand the reduction in the variance $\sigma_y^2(\boldsymbol{x})$ in regions of high density as the $1/n_i$ effect for the variance of the mean of n_i (iid) Gaussian variables each of which has variance σ_ν^2.

The aim of the remainder of this section is to show how results similar to those for the bin basis functions can be obtained, in certain circumstances, for generalized linear regression models, i.e. that the error bars will be inversely proportional to $p(\boldsymbol{x})$ and an area factor $V(\boldsymbol{x})$. The key idea needed is that of an *effective kernel*, which we now describe.

As noted in equation 7, we can write $\hat{y}(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x})\boldsymbol{t}$, where $\boldsymbol{k}(\boldsymbol{x})$ is the effective kernel. To take this analysis further it is helpful to think of $\hat{y}(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x})\boldsymbol{t} = \sum_i k_i t_i$ as an approximation to the integral $\hat{y}(\boldsymbol{x}) = \int K(\boldsymbol{z}; \boldsymbol{x}) t(\boldsymbol{z}) d\boldsymbol{z}$, where $K(\boldsymbol{z}; \boldsymbol{x})$ (regarded as a function of $\boldsymbol{z}$) is the effective kernel for the point $\boldsymbol{x}$ and $t(\boldsymbol{z})$ is a "target function".

Following similar reasoning we obtain

$$\boldsymbol{B} = \sum_{q=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_q)\boldsymbol{\phi}^T(\boldsymbol{x}_q) \simeq N \int p(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^T(\boldsymbol{x}) \, d\boldsymbol{x} \tag{17}$$

If the original basis functions $\boldsymbol{\phi}$ are linearly combined to produce a new set $\tilde{\boldsymbol{\phi}} = \boldsymbol{C}\boldsymbol{\phi}$, then the matrix $\boldsymbol{C}$ can be chosen so that $\int p(\boldsymbol{x})\tilde{\phi}_i(\boldsymbol{x})\tilde{\phi}_j(\boldsymbol{x}) d\boldsymbol{x} = \delta_{ij}$, where δ_{ij} is the Kronecker delta. From now on it is assumed that we are working with the orthonormal basis functions (i.e. the tildes are omitted) and that $\boldsymbol{B} = N\boldsymbol{I}$. Ignoring the weight prior we obtain

$$\hat{\boldsymbol{w}} = \beta \boldsymbol{A}^{-1}\Phi^T \boldsymbol{t} = \frac{1}{N}\Phi^T \boldsymbol{t} \tag{18}$$

However,

$$(\Phi^T \boldsymbol{t})_i = \sum_{q=1}^{N} \phi_i(\boldsymbol{x}^q) t(\boldsymbol{x}^q) \simeq N \int \phi_i(\boldsymbol{z}) p(\boldsymbol{z}) t(\boldsymbol{z}) d\boldsymbol{z} \tag{19}$$

and so

$$\begin{aligned} \hat{y}(\boldsymbol{x}) &= \frac{1}{N}\boldsymbol{\phi}^T(\boldsymbol{x})\Phi^T \boldsymbol{t} && (20) \\ &= \int \left\{ \sum_i \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{z})p(\boldsymbol{z}) \right\} t(\boldsymbol{z}) d\boldsymbol{z} && (21) \\ &= \int K(\boldsymbol{z}; \boldsymbol{x}) t(\boldsymbol{z}) d\boldsymbol{z} && (22) \end{aligned}$$

We can also show that $K(\boldsymbol{z}; \boldsymbol{x})$ is the projection of the delta function onto the basis space $\{\psi_i\}$, where $\psi_i(\boldsymbol{x}) = \phi_i(\boldsymbol{x})p(\boldsymbol{x})$, and that if a constant (bias) function is one of the original basis functions (before orthonormalization), then $\int K(\boldsymbol{z}; \boldsymbol{x}) d\boldsymbol{z} = 1$. The fact that $K(\boldsymbol{z}; \boldsymbol{x})$ is an approximation to the delta

[4] This analysis can easily be extended to arbitrary non-constant basis functions as long as they do not overlap.

function suggests that as the number of basis functions increases the effective kernel should become more tightly peaked and concentrated around $\boldsymbol{x}$.

We now turn to the variance of the generalized linear model. Using orthonormal basis functions, the error bar at $\boldsymbol{x}$ is given by $\sigma_y^2(\boldsymbol{x}) = \frac{\sigma_\nu^2}{N}\boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})$[5]. However, this can be rewritten in terms of the effective kernel

$$\sigma_w^2(\boldsymbol{x}) = \frac{\sigma_\nu^2}{N}\int \frac{K^2(\boldsymbol{z};\boldsymbol{x})}{p(\boldsymbol{z})}d\boldsymbol{z} \tag{23}$$

using the orthonormality properties. If $K(\boldsymbol{z};\boldsymbol{x})$ is sharply peaked around $\boldsymbol{x}$ (i.e. it looks something like a Gaussian) then the $p(\boldsymbol{z})$ in the denominator can be pulled through the integral sign as $p(\boldsymbol{x})$. Also, $\int K^2(\boldsymbol{z};\boldsymbol{x})d\boldsymbol{z}$ measures the inverse base area of $K(\boldsymbol{z};\boldsymbol{x})$; for example, for a one dimensional Gaussian with standard deviation σ centered at $\boldsymbol{x}$ we find that $\int K^2(\boldsymbol{z};\boldsymbol{x})d\boldsymbol{z} = 1/(2\sigma\sqrt{\pi})$. Defining

$$\int K^2(\boldsymbol{z};\boldsymbol{x})d\boldsymbol{z} \stackrel{def}{=} \frac{1}{V(\boldsymbol{x})} \tag{24}$$

we can write

$$\sigma_w^2(\boldsymbol{x}) \simeq \frac{\sigma_\nu^2}{Np(\boldsymbol{x})V(\boldsymbol{x})} \tag{25}$$

By extending the analysis of appendix A.2 to the continuous case we obtain

$$\int \sigma_w^2(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \frac{\sigma_\nu^2}{N}\gamma \tag{26}$$

where γ is the effective number of parameters in the model (1), showing that we would expect $\sigma_w^2(\boldsymbol{x})$ to be larger for a model with more parameters.

Under the assumption that $K(\boldsymbol{z};\boldsymbol{x})$ is sharply peaked about $\boldsymbol{x}$ we have obtained a result in equation 25 similar to equation 16 for the bin basis functions. We will now present evidence to show that this relationship holds experimentally.

The first experiment has a one dimensional input space. The probability density form which the data was drawn is shown in Figure 2(A). Figure 2(B) shows that for a range of GLR models (and for a two-layer perceptron) there is a close relationship between $1/\sigma_w^2(\boldsymbol{x})$ and the density, indicating that $V(\boldsymbol{x})$ is roughly constant in the high density regions for these models. This conclusion is backed up by Figure 4, which plots $1/V(\boldsymbol{x}) = \int K^2(\boldsymbol{z};\boldsymbol{x})d\boldsymbol{z}$ against $\boldsymbol{x}$. The log-log plot in Figure 3 also indicates that the relationship $\sigma_w^2(\boldsymbol{x}) \propto p^{-1}(\boldsymbol{x})$ holds quite reliably, especially for areas with high data density.

[5] It is interesting to note that the error bar $\sigma_w^2(x)$ can also be obtained from the finite-dimensional effective kernel defined by $\hat{y}(x) = k^T(x)t$. Using the assumption that each t_i has independent, zero-mean noise of variance σ_ν^2, we find that the variance of the linear combination $\hat{y}(x)$ is $\sigma_w^2(x) = \sigma_\nu^2 k^T k$, which can easily be shown to be equivalent to $\sigma_w^2(x) = \sigma_\nu^2 \phi^T\phi/N$ for $\alpha = 0$.

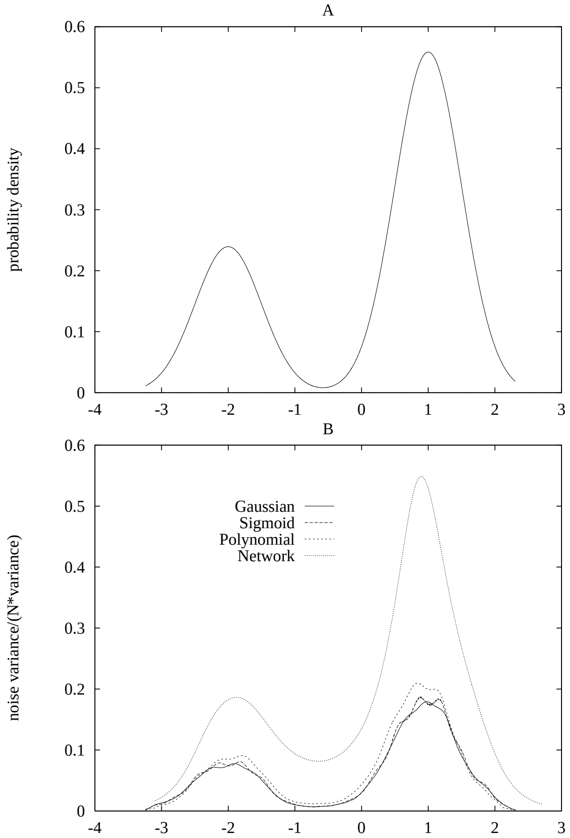

Figure 2: (A) A mixture of two Gaussian densities, from which data points were drawn for the experiments. (B) shows the (scaled) inverse variance against x for three generalized linear regression (GLR) models and a neural network. The GLR models used Gaussian, sigmoid and polynomial basis functions respectively, and each model consisted of 16 basis functions and a bias and was trained on 1000 data points. (B) also shows the inverse variance for a two layer perceptron with two hidden units. The net was trained on a data set consisting of 200 data points with inputs drawn from the density shown in panel (A) and targets generated from $\sin(x)$ with added zero-mean Gaussian noise of standard deviation 0.1. For all four models the similarity between the inverse variance for these models and the plot of the density is striking.

Figure 2(B) also shows that the dependence of the overall magnitude of σ_w^2 on the number of effective parameters described in equation 26 holds; the two-layer perceptron, which has only seven weights compared to the 16 in the GLR models, has a correspondingly larger inverse variance.

Some effective kernels for the GLR model with a bias and 16 Gaussian basis functions of standard deviation 0.5, spaced equally between -5.0 and 4.0 are shown in Figures 5 and 6[6]. The kernels in Figure 5 correspond to areas of high density and show a strong, narrow single peak. For regions of low density Figure 6 shows that the kernels are much wider and more oscillatory, indicating that target values from a wide range of x values are used to compute $\hat{y}(x)$. As the widths of the kernels in the low density regions

[6] Similar $\{\phi_i\}$ and kernels are obtained for sigmoidal and polynomial basis functions.

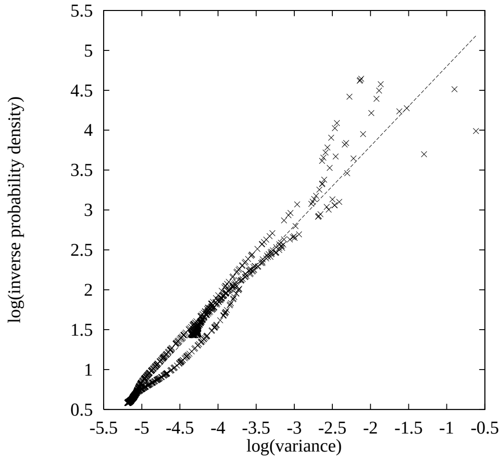

Figure 3: Plot of the log inverse density of the input data against the log of $\sigma_w^2(x)$ for a generalized linear model with 16 Gaussian basis functions. Note that the points lie close to the line with slope 1, indicating that $\sigma_w^2(x) \propto p^{-1}(x)$.

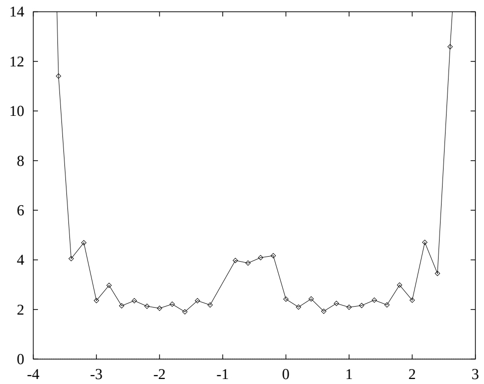

Figure 4: Plot of $1/V(x) = \int K^2(z; x)dz$ against x for a GLR model with 16 Gaussian basis functions spaced equally between -5.0 and 4.0, and a bias. Note that the plot is roughly constant in regions of high density.

are greater than the length scale of the variation of the density, we would expect the approximation used in equation 25 to break down at this point.

We have conducted several other experiments with one and two dimensional input spaces which produce similar results to those shown in the log-log plot, Figure 3, including a two-layer perceptron which learned to approximate a function of two inputs.

While this relationship between $\sigma_w^2(\boldsymbol{x})$ and the input data density is interesting, it should be noted that its validity is limited at best to regions of high data density. Furthermore, in such regions the contribution to the error bars from $\sigma_w^2(\boldsymbol{x})$ is dwarfed by that from the noise term σ_ν^2. This can be seen in the case of non-overlapping basis functions from equation 16. More generally we can consider the extension of the result 11 to the case of n_i data points all located at $\boldsymbol{x}_i$. This leads to

$$\sigma^2_{w|n_i \boldsymbol{x}_i}(\boldsymbol{x}_i) = \sigma_\nu^2 \frac{r_i}{1 + n_i r_i} \simeq \sigma_\nu^2 / n_i \qquad (27)$$

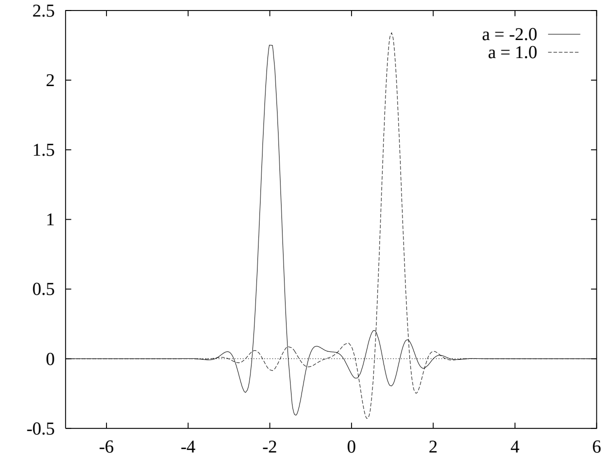

Figure 5: Effective kernels at $x = 1.0$ and -2.0, corresponding to high density regions, as shown in figure 2(A). See text for further discussion.

again indicating that for regions of high data density the noise term will dominate.

5 DISCUSSION

In this paper we have analyzed the behaviour of the Bayesian error bars for generalized linear regression models. For the case of a single isolated data point we have shown that the error bar is pulled down close to the noise level, and that the length scale over which this effect occurs is characterized by the prior covariance function. We have also shown theoretically that, in regions of high data density, the contribution to the output variance due to the uncertainty in the weights can exhibit an approximate inverse proportionality to the data density. These findings have been supported by numerical simulation. Also, we have noted that, in such high-density regions, this contribution to the variance will be insignificant compared to the contribution arising from the noise term.

Although much of the theoretical analysis has been performed for generalized linear regression models, there is empirical evidence that similar results hold also for multi-layer networks. Furthermore, if the outputs of the network have linear activation functions, then under least-squares training it is effectively a generalized linear regression model with adaptive basis functions. It is therefore a linear smoother with $\hat{y}(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x})\boldsymbol{t}$, and hence the result that $\sigma_w^2 = \boldsymbol{k}^T\boldsymbol{k}\sigma_\nu^2$ will still hold. Other results, including the expression 36 derived in Appendix A.2, also hold for general non-linear networks, provided we make the usual Gaussian approximation for the posterior weight distribution, and the outer-product approximation to the Hessian.

One potentially important limitation of the models

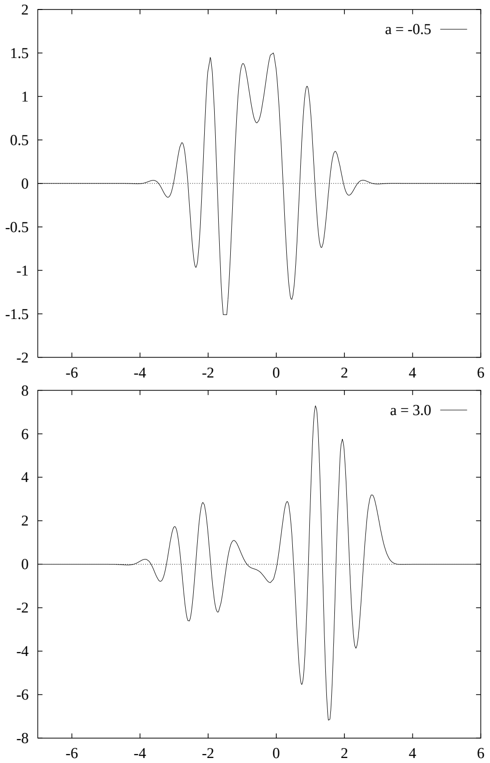

Figure 6: Effective kernels at $x = -0.5$ and 3.0 corresponding to low density regions of the input space, as shown in figure 2(A). Note that the density function seems to define an "envelope" for the lower kernel; even though x may be in a low density region, the magnitude of $K(z;x)$ is largest in the high density regions. See text for further discussion.

considered in this paper (and indeed of the models considered by most authors) is that the noise variance σ_ν^2 is assumed to be a constant, independent of $\boldsymbol{x}$. To understand why this assumption may be particularly restrictive, consider the situation in which there is a lot of data in one region of input space and a single data point in another region. The estimate of the noise variance, which we shall assume to be relatively small, will be dominated by the high density region. However, as we have seen, the error bar will be pulled down to less than $2\sigma_\nu^2$ in the neighbourhood of the isolated data point. The model is therefore highly confident of the regression function (i.e. the most probable interpolant) in this region even though there is only a single data point present! If, however, we relax the assumption of a constant σ_ν^2 then we see that there in the neighbourhood of the isolated data point there is little evidence to suggest a small value of σ_ν^2 and so we would expect much larger error bars. We are currently investigating models in which $\sigma_\nu^2(\boldsymbol{x})$ is adapted to the data.

Acknowledgements

This work was mainly supported by EPSRC grant GR/J75425 (CW) and by an EPSRC post-graduate scholarship to CQ.

6 APPENDICES

A.1

In this appendix we show that for generalized linear regression, $\sigma^2_{y|D}(x) \leq \sigma^2_{y|T}(x)$, where D is the full data set $((x_1, t_1), \ldots, (x_N, t_N))$ and T is a subset of this data set.

We first note that as $\sigma_\nu^2(x)$ is equal in both cases, we are only concerned about the relative contributions from the weight uncertainty to the overall variance. The key to the proof is to decompose the Hessian A into two parts, A_1 and A_2, where

$$A_1 = A_0 + \sum_{q \in T} \beta_q \phi_q \phi_q^T \quad A_2 = \sum_{q \notin T} \beta_q \phi_q \phi_q^T \tag{28}$$

and $A_0 = \alpha S$. Note that A_1 and A_2 are symmetric non-negative definite, and hence A_1^{-1} and A_2^{-1} are also (using the Moore-Penrose pseudo-inverse if necessary). The matrix identity

$$(A_1 + A_2)^{-1} = A_1^{-1} - A_1^{-1}(A_1^{-1} + A_2^{-1})^{-1} A_1^{-1} \tag{29}$$

implies that for any vector v

$$\begin{aligned} &v^T (A_1 + A_2)^{-1} v = \\ &v^T A_1^{-1} v - (A_1^{-1} v)^T (A_1^{-1} + A_2^{-1})^{-1} (A_1^{-1} v) \end{aligned} \tag{30}$$

From non-negative definite condition we see that the second term in equation 31 is always non-negative, and hence

$$v^T A_1^{-1} v \geq v^T (A_1 + A_2)^{-1} v \tag{31}$$

Substituting $\phi(x)$ for v completes the proof.

A.2

In this appendix we show that $\langle \sigma_w^2 \rangle$, the average value of $\sigma_w^2(x)$ evaluated at the data points, is equal to $\sigma_\nu^2 \gamma / N$, where γ ($\leq m$) is the *effective number of parameters* in the model (1).

$$\begin{aligned} \langle \sigma_w^2 \rangle &= \frac{1}{N} \sum_i^N \sigma_w^2(x_i) &(32)\\ &= \frac{1}{N} \sum_i \phi_i^T A^{-1} \phi_i &(33)\\ &= \frac{1}{N} tr[(\sum_i \phi_i \phi_i^T) A^{-1}] &(34)\\ &= \frac{\sigma_\nu^2}{N} tr(\beta B A^{-1}) &(35)\\ &= \frac{\sigma_\nu^2}{N} \gamma &(36) \end{aligned}$$

References

1. MacKay D. J. C., 1992, "Bayesian Interpolation", *Neural Computation*, **4(3)**, 415–447.
2. Hastie T. J. and Tibshirani R. J., 1990, "Generalized Additive Models", Chapman and Hall.
3. Bishop C. M., 1994, "Novelty detection and neural network validation", *IEE Proceedings: Vision, Image and Signal Processing*, **141**, 217–222.