The Evidence Framework applied to Classification Networks

David J.C. MacKay*
Computation and Neural Systems
California Institute of Technology 139–74
Pasadena CA 91125
mackay@ras.phy.cam.ac.uk

Abstract

Three Bayesian ideas are presented for supervised adaptive classifiers. First, it is argued that the output of a classifier should be obtained by marginalising over the posterior distribution of the parameters; a simple approximation to this integral is proposed and demonstrated. This involves a 'moderation' of the most probable classifier's outputs, and yields improved performance. Second, it is demonstrated that the Bayesian framework for model comparison described for regression models in (MacKay, 1992a, 1992b) can also be applied to classification problems. This framework successfully chooses the magnitude of weight decay terms, and ranks solutions found using different numbers of hidden units. Third, an information–based data selection criterion is derived and demonstrated within this framework.

1 Introduction

A quantitative Bayesian framework has been described for learning of mappings in feedforward networks (MacKay, 1992a, 1992b). It was demonstrated that this 'evidence' framework could successfully choose the magnitude and type of weight decay terms, and could choose between solutions using different numbers of hidden units. The framework also gives quantified error bars expressing the uncertainty in the network's outputs and its parameters. In (MacKay, 1992c) information–based objective functions for active learning were discussed within the same framework.

These three papers concentrated on interpolation (regression) problems. Neural networks can also be trained to perform classification tasks.[1] This paper will show that the Bayesian framework for model comparison can be applied to these problems too.

Assume that a set of candidate classification models are fitted to a data set, using standard methods. Three aspects of the use of classifiers can then be distinguished:

*Current address: Darwin College, Cambridge CB3 9EU, U.K.

[1] In regression the target variables are real numbers, assumed to include additive errors; in classification the target variables are discrete class labels.

1. The individual classification models are used to make predictions about new targets.

2. The alternative models are ranked in the light of the data.

3. The expected utility of alternative new data points is estimated for the purpose of 'query learning' or 'active data selection'.

This paper will present Bayesian ideas for these three tasks. Other aspects of classifiers use such as prediction of generalisation ability are not addressed.

First let us review the framework for supervised adaptive classification.

Derivation of the objective function $G = \sum t \ln p$

The same notation and conventions will be used as in (MacKay, 1992a, 1992b). Let the data set be $D = \{\mathbf{x}^{(m)}, t_m\}$, $m = 1 \ldots N$. In a classification problem, each target t_m is a binary (0/1) variable (more than two classes can also be handled (Bridle, 1989)), and the activity of the output of a classifier is viewed as an estimate of the probability that $t=1$. It is assumed that the classification problem is noisy, that is, repeated sampling at the same $\mathbf{x}$ would produce different values of t with certain probabilities; those probabilities, as a function of $\mathbf{x}$, are the quantities that a discriminative classifier is intended to model. It is well known that the natural objective function in this case is an information–based distance measure, rather than the sum of squared errors (Bridle, 1989, Hinton and Sejnowski, 1986, Hopfield, 1987, Solla *et al.*, 1988).

A classification model $\mathcal{H}$ consists of a specification of its architecture $\mathcal{A}$ and the regulariser $\mathcal{R}$ for its parameters $\mathbf{w}$. When a classification model's parameters are set to a particular value, the model produces an output $y(\mathbf{x}; \mathbf{w}, \mathcal{A})$ between 0 and 1, which is viewed as the probability $P(t = 1|\mathbf{x}, \mathbf{w}, \mathcal{A})$. The likelihood, *i.e.* the probability of the data[2] as a function of $\mathbf{w}$, is then:

$$\begin{aligned} P(D\,|\mathbf{w}, \mathcal{A}) &= \prod_m y^{t_m}(1-y)^{1-t_m} \\ &= \exp G(D\,|\mathbf{w}, \mathcal{A}), \end{aligned}$$

where

$$G(D\,|\mathbf{w}, \mathcal{A}) = \sum_m t_m \log y + (1 - t_m)\log(1 - y). \tag{1}$$

This is the probabilistic motivation for the cross–entropy objective function $\sum p \log \frac{q}{p}$. Now if we assign a prior over alternative parameter vectors $\mathbf{w}$,

$$P(\mathbf{w}|\{\alpha_c\}, \mathcal{A}, \mathcal{R}) = \frac{\exp(-\sum_c \alpha_c E_W^{(c)})}{Z_W}, \tag{2}$$

[2]Strictly this is the probability of $\{t_m\}$ given $\{\mathbf{x}^{(m)}\}, \mathbf{w}, \mathcal{A}$; the density over $\{\mathbf{x}\}$ is not modelled by the 'discriminative' classifiers discussed in this paper.

where $E_W^{(c)}$ is a cost function for a subset (c) of the weights and α_c is the associated regularisation constant (see MacKay, 1992b), we obtain a posterior:

$$P(\mathbf{w}|D, \{\alpha_c\}, \mathcal{A}, \mathcal{R}) = \frac{\exp(-\sum_c \alpha_c E_W^{(c)} + G)}{Z_M}, \tag{3}$$

where Z_W and Z_M are the appropriate normalising constants. Thus the identical framework is obtained to that in (MacKay, 1992b) with $-G$ replacing the term βE_D. Note that in contrast to the framework for regression in (MacKay, 1992b) there is now no free parameter β and no $Z_D(\beta)$. If however a teacher were to supply probability estimates t instead of binary targets, then a constant equivalent to β would appear, expressing the precision of the teacher's estimates. This constant would correspond to the effective number of observations on which the teacher's opinion is based.

The calculation of the gradient and Hessian of G is as easy as for a quadratic E_D, if the output unit's activation function is the traditional logistic $f(a) = 1/(1 + e^{-a})$, or the generalised 'softmax' in the case of more than two classes (Bridle, 1989). The appropriateness of a logisitic output function for a classifier is well known: it is the function that converts a log probability ratio a into a probability $f(a)$.

Gradient: If $y(\mathbf{x}^{(m)}) = f(a(\mathbf{x}^{(m)}))$ as defined above, the gradient of G with respect to the parameters $\mathbf{w}$ is

$$\nabla G = \sum_m (t_m - y)\mathbf{g}_{(m)}, \tag{4}$$

where $\mathbf{g}_{(m)} = \partial a / \partial \mathbf{w}|_{\mathbf{x}=\mathbf{x}^{(m)}}$.

Hessian: The Hessian can be analytically evaluated (Bishop, 1991), but a useful approximation neglecting terms in $\partial^2 a / \partial^2 \mathbf{w}$ is:

$$\nabla\nabla G \simeq -\sum_m f' \mathbf{g}_{(m)} \mathbf{g}_{(m)}^{\mathrm{T}}, \tag{5}$$

where $f' = \partial f / \partial a$. This approximation is expected to be adequate for the evaluation of error bars, for use in data selection and for the evaluation of the number of well–determined parameters γ. A more accurate evaluation of the Hessian is probably needed for estimation of the evidence. In this paper's demonstrations, the Hessian is evaluated using second differences, *i.e.*, numerical differentiation of ∇G with respect to $\mathbf{w}$.

Validity of approximations

On account of the central limit theorem, we expect the posterior distribution to converge to a set of locally Gaussian peaks with increasing quantities of data. However, the quadratic approximation to G is expected to converge more slowly than the quadratic approximation to E_D, the error function for regression models, because (a) G is not a quadratic function even for a linear model (a model for which $a = \sum w_h \phi_h(\mathbf{x})$): each term in G has the large scale form of a ramp function; and (b) only inputs which fall in the 'bend' of the

ramp contribute curvature to G. If we have the opportunity for active data selection we could improve the convergence of this quadratic approximation by selecting inputs that are expected to contribute maximal curvature. A related data selection criterion is derived in section 4.

2 Every classifier should have two sets of outputs

Consider a classifier with output $y(\mathbf{x};\mathbf{w}) = f(a(\mathbf{x};\mathbf{w}))$. Assume that we receive data D and infer the posterior probability of the parameters $\mathbf{w}$ (*i.e.* we perform 'learning'). Now if we are asked to make predictions with this classifier, it is common for the most probable parameter vector $\mathbf{w}_{\rm MP}$ to be used as the sole representative of the posterior distribution. This strategy seems unwise, however, since there may be regions in input space where the posterior ensemble is very uncertain about what the class is; in such regions the output of the network should be $y \simeq 0.5$ (assuming equiprobable classes *a priori*), whereas typically the network with parameters $\mathbf{w}_{\rm MP}$ will give a more extreme, unrepresentative and overconfident output. *The error bars on the parameters should be taken into account when predictions are made.*

In regression problems, it is also important to calculate error bars on outputs, but the problem is more acute in the case of classification because, on account of the non–linear output, the mean output over the posterior distribution is not equal to the most probable network's output. To obtain an output representative of the posterior ensemble of networks around $\mathbf{w}_{\rm MP}$, we need to *moderate* the output of the most probable network in relation to the error bars on $\mathbf{w}_{\rm MP}$.

Of course this idea of averaging over the hidden parameters is not new: marginalisation goes back to Laplace. More recently, and in a context closer to the present one, the same message can be found for example in (Spiegelhalter and Lauritzen, 1990). But it seems that most practitioners of adaptive classification do not currently use marginalisation.

I suggest that any classifier should have two sets of outputs. The first set would give the usual class probabilities corresponding to $\mathbf{w}_{\rm MP}$, $y(\mathbf{x};\mathbf{w}_{\rm MP})$; these outputs would be used for learning, *i.e.* for calculating the error signals for optimisation of $\mathbf{w}_{\rm MP}$. The second set would be the moderated outputs $y(\mathbf{x};P(\mathbf{w}|D)) = \int d^k\mathbf{w}\, y(\mathbf{x};\mathbf{w})P(\mathbf{w}|D)$; these outputs would be used for all other applications, *e.g.* prediction, evaluation of test error, and for evaluating the utility of candidate data points (section 4). Let us now discuss how to calculate the moderated outputs. It will then be demonstrated that these outputs can indeed provide better estimates of class probabilities.

Calculating the moderated outputs

If we assume a locally Gaussian posterior probability distribution[3] over $\mathbf{w} = \mathbf{w}_{\rm MP} + \Delta\mathbf{w}$, $P(\mathbf{w}|D) \simeq P(\mathbf{w}_{\rm MP}) \exp(-\frac{1}{2}\Delta\mathbf{w}^{\rm T}\mathbf{A}\Delta\mathbf{w})$, and if we assume that the activation $a(\mathbf{x};\mathbf{w})$ is a locally linear function of $\mathbf{w}$ with $\partial a/\partial \mathbf{w} = \mathbf{g}$, then for any $\mathbf{x}$, the activation a is approximately Gaussian distributed:

$$P(a(\mathbf{x})|D) = \text{Normal}(a^{\rm MP}, s^2) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(a-a^{\rm MP})^2}{2s^2}\right), \tag{6}$$

where $a^{\rm MP} = a(\mathbf{x};\mathbf{w}_{\rm MP})$ and $s^2 = \mathbf{g}^{\rm T}\mathbf{A}^{-1}\mathbf{g}$. This means that the moderated output is:

$$P(t=1|\mathbf{x},D) = \psi(a^{\rm MP}, s^2) \equiv \int da\, f(a)\, \text{Normal}(a^{\rm MP}, s^2). \tag{7}$$

This is to be contrasted with the most probable network's output, $y(\mathbf{x};\mathbf{w}_{\rm MP}) = f(a^{\rm MP})$. The integral of a sigmoid times a Gaussian cannot be solved analytically; here I suggest a simple numerical approximation to it:

$$\psi(a^{\rm MP}, s^2) \simeq \phi(a^{\rm MP}, s^2) \equiv f(\kappa(s) a^{\rm MP}) \tag{8}$$

with $\kappa = 1/\sqrt{1+\pi s^2/8}$. This approximation is not globally accurate over $(a^{\rm MP}, s^2)$, (for large $s^2 > a$ the function should tend to an error function, not a logisitic) but it breaks down gracefully. The value of κ was chosen so that the approximation has the correct gain at $a^{\rm MP} = 0$, as $s^2 \to \infty$. A representative of this approximation is given in figure 1 which compares ϕ and ϕ' with numerical evaluations of ψ and ψ'. A similar approximation in terms of the error function is suggested in (Spiegelhalter and Lauritzen, 1990).

If the output is immediately used to make a (0/1) decision, then the use of moderated outputs will make no difference to the performance of the classifier (unless the costs associated with error are assymetrical), since both functions pass through 0.5 at $a^{\rm MP} = 0$. But moderated outputs will make a difference if a more sophisticated penalty function is involved. In the following demonstration the performance of a classifier's outputs is measured by the value of G achieved on a test set.

A model classification problem with two input variables and two possible classes is shown in figure 2a. Figure 2b illustrates the output of a typical trained network, using its *most probable* parameter values. Figure 2c shows the *moderated* outputs of the same network. Notice how the moderated output is similar to the most probable output in regions where the data are dense. In contrast, where the data are sparse, the moderated output becomes significantly less certain than the most probable output; this can be seen by the widening of the contours. Figure 2d shows the correct posterior probability for this problem given the knowledge of the true class densities.

Several hundred neural networks having two inputs, one hidden layer of sigmoid units and one sigmoid output unit were trained on this problem. During optimisation, the second

[3] Conditioning variables such as $\mathcal{A}, \mathcal{R}, \{\alpha_c\}$ will be omitted in this section, since the emphasis is not on model comparison.

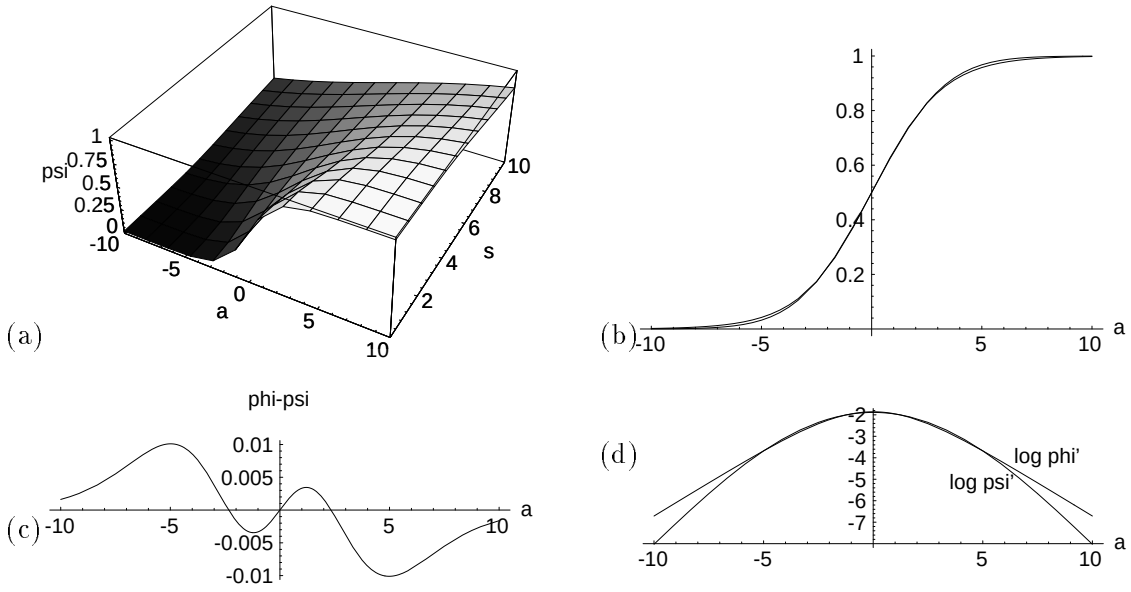

Figure 1: **Approximation to the moderated probability**

(a) The function $\psi(a, s^2)$, evaluated numerically. In (b) the functions $\psi(a, s^2)$ and $\phi(a, s^2)$ defined in the text are shown as a function of a for $s^2 = 4$. In (c), the difference $\phi - \psi$ is shown for the same parameter values. In (d), the breakdown of the approximation is emphasised by showing $\log \phi'$ and $\log \psi'$ (derivatives with respect to a). The errors become significant when $a \gg s$.

weight decay scheme of (MacKay, 1992b) was used, using independent decay rates for each of three weight classes: hidden weights, hidden unit biases, and output weights and biases. This corresponds to the prior which models the weights in each class as coming from a Gaussian; the scale of the Gaussians for different classes are independent and are specified by regularising constants α_c. Each regularising constant is optimised on line by intermittently updating it to its *most probable* value as estimated within the 'evidence' framework.

The prediction abilities of a hundred networks using their 'most probable' outputs and using the moderated outputs suggested above are compared in figure 3. It can be seen that the predictions given by the moderated outputs are in nearly all cases superior. The improvement is most substantial for underdetermined networks with relatively poor performance. In a small fraction of the solutions however, especially among the best solutions, the moderated outputs are found to have slightly but significantly inferior performance.

3 Evaluating the evidence

Having established how to use a particular model $\mathcal{H} = \{\mathcal{A}, \mathcal{R}\}$ with given regularising constants $\{\alpha_c\}$ to make predictions, we now turn to the question of model comparison. As discussed in (MacKay, 1992a), three levels of inference can be distinguished: parameter

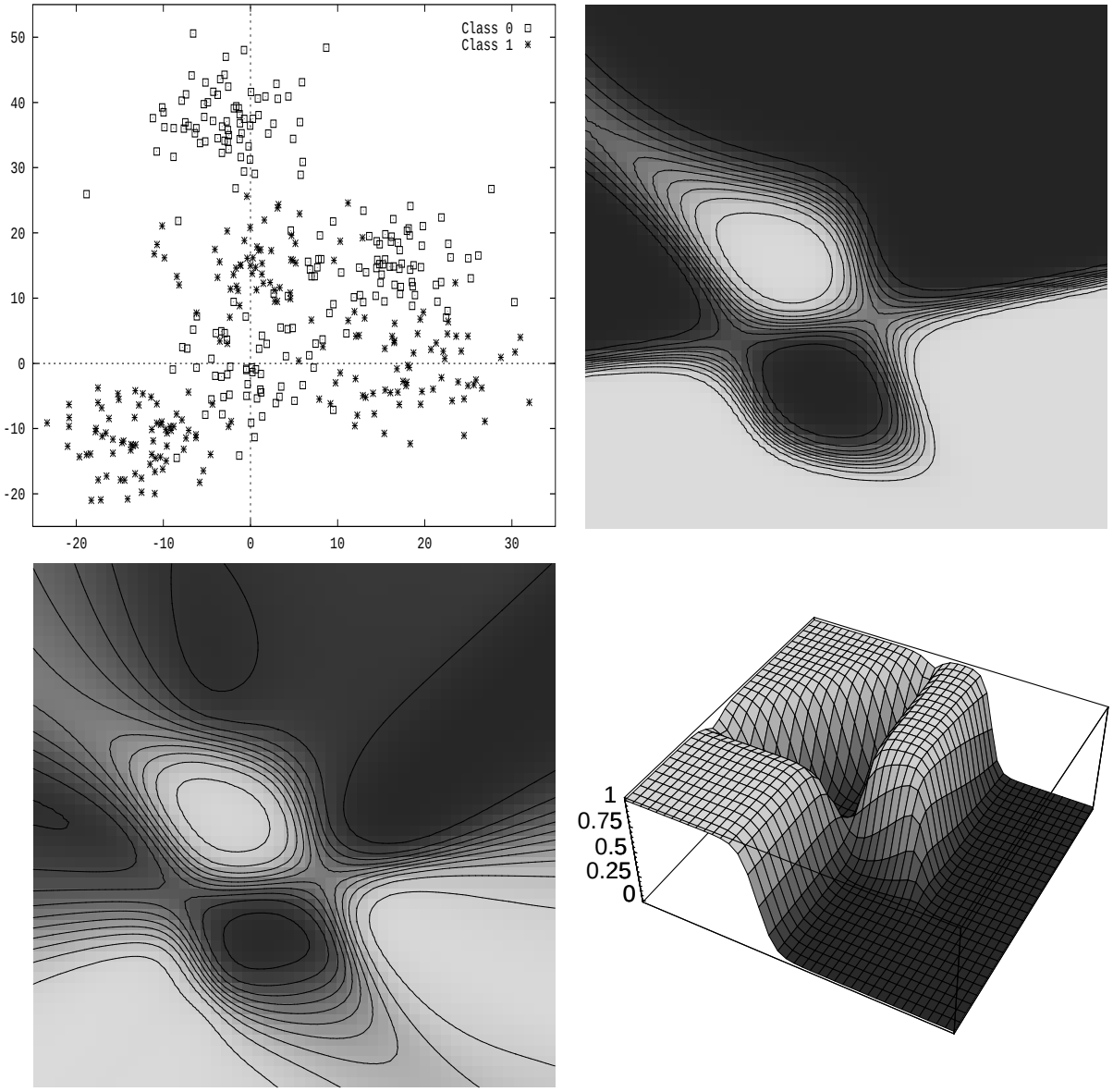

Figure 2: **Comparison of most probable outputs and moderated outputs for a toy problem**

a) The data set. The data were generated from six circular Gaussian distributions, three Gaussians for each class. The training sets for the demonstrations use between 100 and 1000 data points drawn from this distribution. b) 'Most probable' output of an eight hidden unit network trained on 100 data points. The contours are equally spaced between 0.0 and 1.0. c) 'Moderated' output of the network. Notice that the output becomes less certain compared with the most probable output as the input moves away from regions of high training data density. d) The true posterior probability, given the class densities that generated the data. The viewpoint is from the upper right corner of (a). In (b,c,d) a common grey scale is used, linear from 0 (dark grey) to 1 (light grey).

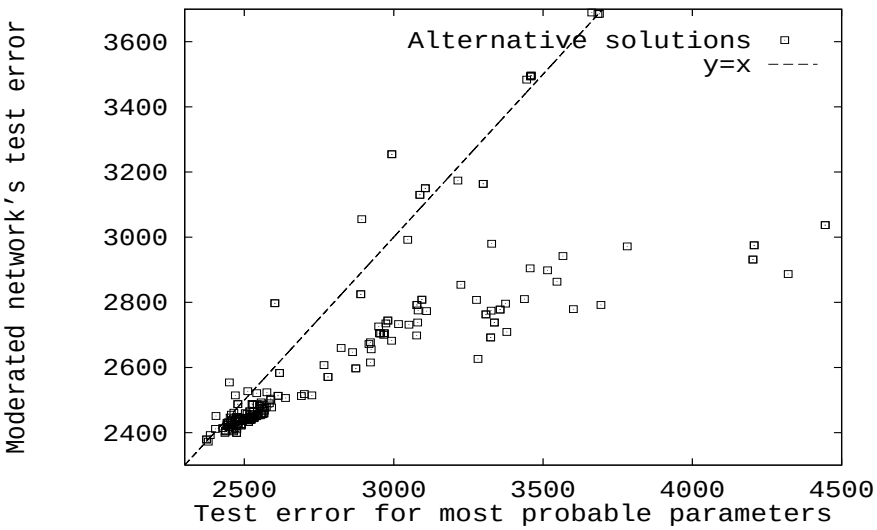

Figure 3: **Moderation is a good thing!**

The training set for all the networks contained 300 data points. For each network, the test error of the 'most probable' outputs and the 'moderated' outputs were evaluated on a test set of 5000 data points. The test error is the value of G. Note that for most solutions, the moderated outputs make better predictions.

estimation, regularisation constant determination, and model comparison.[4] The second two levels of inference both require 'Occam's razor'; that is, the solution that best fits the data is not the most plausible model, and we need a way to balance goodness of fit against complexity. Bayesian inference embodies such an Occam's razor automatically.

At the first level, a model $\mathcal{H}$, with given regularising constants $\{\alpha_c\}$ is fitted to the data D. This involves inferring what value the parameters $\mathbf{w}$ should probably have. Bayes' rule for this level of inference has the form:

$$P(\mathbf{w}|D,\{\alpha_c\},\mathcal{H}) = \frac{P(D|\mathbf{w},\{\alpha_c\},\mathcal{H})P(\mathbf{w}|\{\alpha_c\},\mathcal{H})}{P(D|\{\alpha_c\},\mathcal{H})}. \tag{9}$$

Throughout this paper this posterior is approximated locally by a Gaussian:

$$P(\mathbf{w}|D,\{\alpha_c\},\mathcal{H}) = \frac{\exp(-M(\mathbf{w}))}{Z_M} \simeq \frac{\exp(-M_{\rm MP} - \frac{1}{2}\Delta\mathbf{w}^{\rm T}\mathbf{A}^{-1}\Delta\mathbf{w})}{Z^*_M}, \tag{10}$$

where $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\rm MP}$, $M(\mathbf{w}) = \sum_c \alpha_c E_W^{(c)} - G$, and $\mathbf{A} = \nabla\nabla M$.

At the second level of inference, the regularising constants are optimised:

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

The data–dependent term $P(D|\{\alpha_c\},\mathcal{H})$ is the 'evidence', the normalising constant from equation (9). The evaluation of this quantity and the optimisation of the parameters $\{\alpha_c\}$

[4] The use of a specified model to predict the class of a datum can be viewed as the zeroeth level of inference.

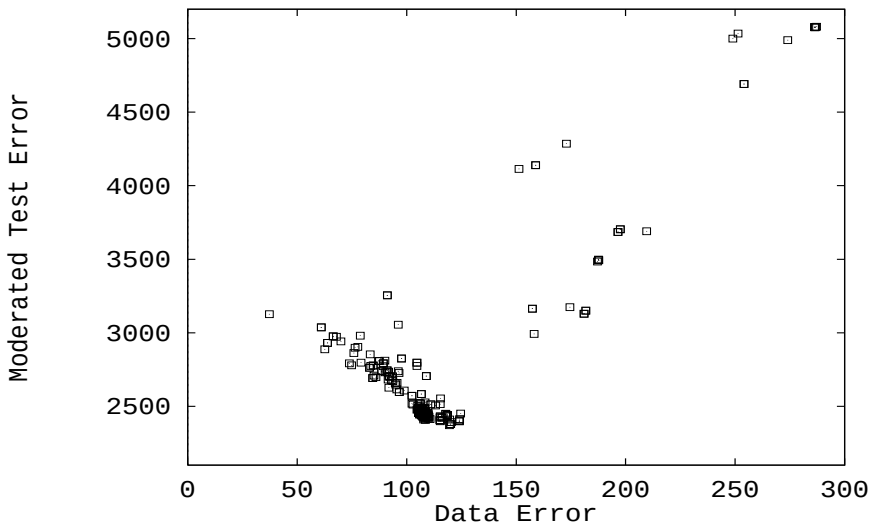

Figure 4: **Test error versus data error**

This figure illustrates that the task of ranking solutions to the classification problem requires Occam's razor; the solutions with smallest data error do not generalise best.

is accomplished using a framework due to Gull and Skilling, discussed in detail in (MacKay, 1992a, 1992b).

Finally, at the third level of inference, the alternative models are compared:

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

Again, the data's opinion about the alternatives is given by the evidence from the previous level, in this case $P(D|\mathcal{H})$.

Omitting the details of the second level of inference, since they are identical to the methods in (MacKay, 1992b), this demonstration presents the final inferences, the evidence for alternative solutions. The evidence is evaluated within the Gaussian approximation from the properties of the 'most probable' fit $\mathbf{w}_{\rm MP}$, and the error bars $\mathbf{A}^{-1}$, as described in (MacKay, 1992a).

Figure 4 shows the test error (calculated using the moderated outputs) of the solutions against the data error, and the 'Occam's razor' problem can be seen: the solutions with smallest data error do not generalise best. Figure 5 shows the log evidence for the solutions against the test error, and it can be seen that a moderately good correlation is obtained. The correlation is not perfect. It is speculated that the discrepancy is mainly due to inaccurate evaluation of the evidence under the quadratic approximation, but further study is needed here. Finally, figure 6 explores the dependence of the correlation between evidence and generalisation on the amount of data. It can be seen that the correlation improves as the number of data points in the test set increases.

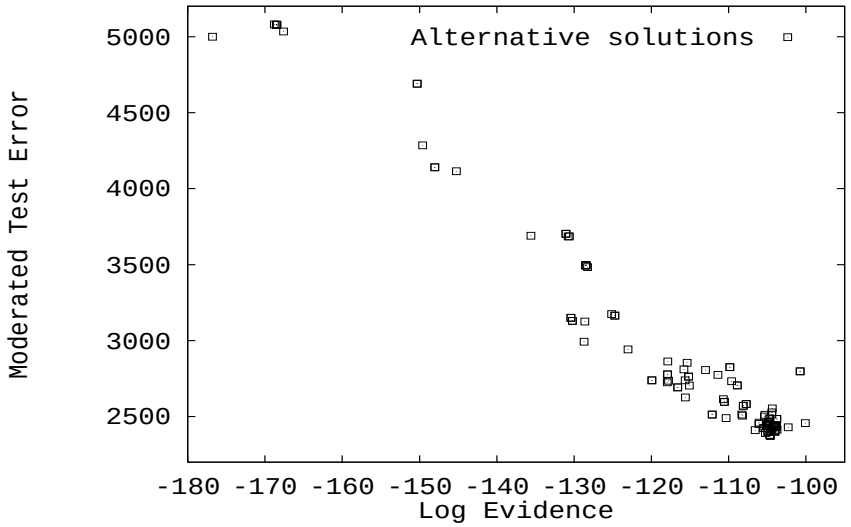

Figure 5: **Test error versus evidence**

Each solution was found using the same training set of $N = 300$ data points. All solutions in which a symmetry was detected among the hidden units were omitted from this graph because the evidence evaluation for such solutions is unreliable.

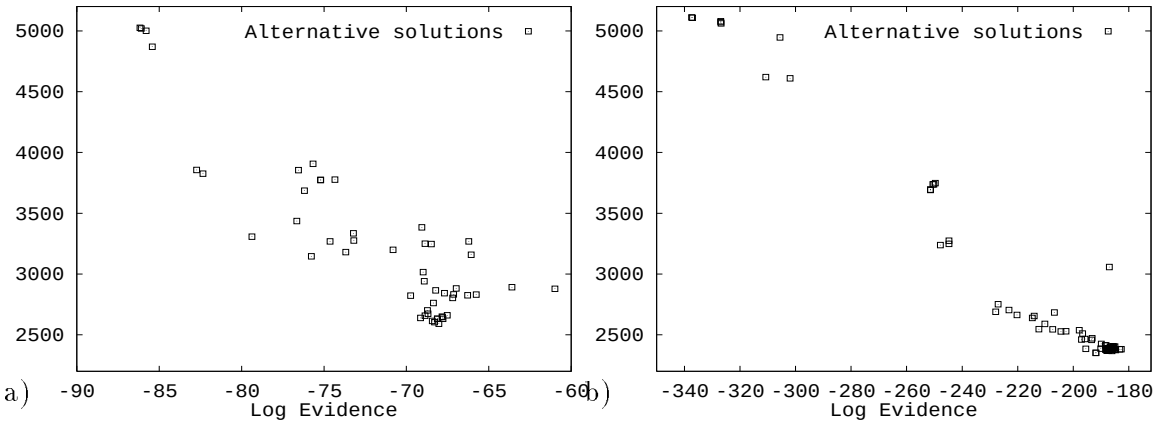

Figure 6: **Correlation between test error and evidence as the amount of data varies.**

a) $N = 150$ data points. b) $N = 600$ data points. c.f. Figure 5, for which $N = 300$. For comparison, the number of parameters in a typical (10 hidden unit) network is 41. Note that only about 25% of the data points fall in informative decision regions; so the effective number of data points is smaller in each case; bear in mind also that each data point only consists of one bit. All solutions in which a symmetry was detected among the hidden units were omitted because the evidence evaluation for such solutions is unreliable.

4 Active learning

Assume now that we have the opportunity to select the input $\mathbf{x}$ where a future datum will be gathered ('query learning'). Several papers have suggested strategies for this active learning problem, for example Hwang *et al.* (1991) propose that samples should be made on and near the current decision boundaries. This strategy and that of Baum (1991) are both human–designed strategies and it is not clear what objective function if any they optimise, nor is it clear how the strategies could be improved. In this paper, as in (MacKay, 1992c), the philosophy will be to *derive* a criterion from a *defined* sensible objective function that measures how useful a datum is expected to be. This criterion may then be used as a guide for query learning, or for the alternative scenario of pruning uninformative data points from a large data set.

Desiderata

Let us criticise Hwang *et al.*'s strategy to try to establish a reasonable objective function. The strategy of sampling on decision boundaries is motivated by the argument that we are unlikely to gain information by sampling in a region where we are already confident of the correct classification. But similarly, if we have already sampled a great deal on one particular boundary then we don't gain useful information by repeatedly sampling there either, because the location of the boundary has been established! Repeated sampling at such locations generates data with large entropy that are 'informative' in the same way that white noise is informative. There must be more to the utility of a sample than its distance from a decision boundary. We would prefer to sample near boundaries whose location has not been well determined, because this would probably enable us to make more precise predictions there. Thus we are interested in measurements which convey *mutual* information about the unknowns that we are interested in.

A second criticism is that a strategy that only samples near existing boundaries is not likely to make new discoveries; a strategy that also samples near *potential* boundaries is expected to be more informative. A final criticism is that to be efficient, a strategy should take into account how influential a datum will be: some data may convey information about the discriminant over a larger region than others. So we want an objective function that measures the global expected informativeness of a datum.

Objective function

This paper will study the 'mean marginal information.' This objective function was suggested in (MacKay, 1992c), and a discussion of why it is probably more desirable than the joint information is given there. To define this objective function, we first have to define a region of interest. (The objective of maximal information gain about the model's parameters without a region of interest would lead us to sample at unsampled extremes of the input space.) Here this region of interest will be defined by a set of representative points

$\mathbf{x}^{(u)}$, $u = 1 \ldots V$, with a normalised distribution P_u on them. P_u can be interpreted as the probability that we will be asked to make a prediction at $\mathbf{x}^{(u)}$. (The theory could be worked out for the case of a continuous region defined by a density $\rho(\mathbf{x})$, but the discrete case is preferred since it relates directly to practical implementation.) The marginal entropy of a distribution over $\mathbf{w}$, $P(\mathbf{w})$, at one point $\mathbf{x}^{(u)}$ is defined to be

$$S_M^{(u)} = y_u \log y_u + (1 - y_u) \log(1 - y_u), \tag{13}$$

where $y_u = y(\mathbf{x}^{(u)}; P(\mathbf{w}))$ is the average output of the classifier over the ensemble $P(\mathbf{w})$. Under the Gaussian approximation for $P(\mathbf{w})$, y_u is given by the moderated output (7), and may be approximated by $\phi(a_u^{\mathrm{MP}}, s_u^2)$ (8).

The mean marginal entropy is

$$\bar{S}_M(P(\mathbf{w})) = \sum_u P_u S_M^{(u)}. \tag{14}$$

The sampling strategy studied here is to maximise the expected change in mean marginal entropy. (Note that our information gain is *minus* the change in entropy.)

Estimating marginal entropy changes

Let a measurement be made at $\mathbf{x}$. The result of this measurement is either $t = 1$ or $t = 0$. Assuming that our current model, complete with Gaussian error bars, is correct, the probability of $t = 1$ is $\psi(a^{\mathrm{MP}}(\mathbf{x}), s^2(\mathbf{x})) \simeq \phi(a^{\mathrm{MP}}, s^2)$. We wish to estimate the average change in marginal entropy of t_u at $\mathbf{x}^{(u)}$ when this measurement is made.

This problem can be solved by calculating the joint probability distribution $P(t, t_u)$ of t and t_u, then finding the mutual information between the two variables. The four values of $P(t, t_u)$ have the form:

$$P(t=1, t_u=1) = \iint da\, da_u\, f(a) f(a_u) \frac{1}{Z} \exp\left(-\frac{1}{2} \Delta\mathbf{a}^{\mathrm{T}} \Sigma^{-1} \Delta\mathbf{a}\right), \text{ etc.,} \tag{15}$$

where $\Delta\mathbf{a}^{\mathrm{T}} = (\Delta a, \Delta a_u)$ and the activations $a = a^{\mathrm{MP}} + \Delta a$ and $a_u = a_u^{\mathrm{MP}} + \Delta a_u$ are assumed to have a Gaussian distribution with covariance matrix

$$\Sigma = \begin{pmatrix} \mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g} & \mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)} \\ \mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)} & \mathbf{g}_{(u)}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)} \end{pmatrix} \equiv \begin{pmatrix} s^2 & \rho s s_u \\ \rho s s_u & s_u^2 \end{pmatrix}. \tag{16}$$

The normalising constant is $Z = 2\pi s s_u (1 - \rho^2)^{\frac{1}{2}}$. The expected change in entropy of t_u is:

$$E(\Delta S_M^{(u)} | t) = S(P(t, t_u)) - S(P(t)) - S(P(t_u)). \tag{17}$$

Notice that this mutual information is symmetric in t and t_u. We can approximate $E(\Delta S_M^{(u)} | t)$ by Taylor–expanding $P(t, t_u)$ about independence ($\rho = 0$). The first order perturbation to $P(t, t_u)$ introduced by ρ can be written in terms of a single variable c:

$$\begin{aligned} P(t=1, t_u=1) &= P(t=1)P(t_u=1) + c \quad & P(t=1, t_u=0) &= P(t=1)P(t_u=0) - c \\ P(t=0, t_u=1) &= P(t=0)P(t_u=1) - c \quad & P(t=0, t_u=0) &= P(t=0)P(t_u=0) + c. \end{aligned} \tag{18}$$

Taylor–expanding (17), we find

$$E(\Delta S_M^{(u)}|t) \simeq -\frac{1}{P(t=1)P(t_u=1)P(t=0)P(t_u=0)}\, c^2/2. \tag{19}$$

Finally, we Taylor–expand (15) so as to obtain the dependence of c on the correlation between the activations. The derivative of $P(t=1,t_u=1)$ with respect to ρ at $\rho=0$ is

$$\begin{aligned}\frac{\partial}{\partial\rho}P(t=1,t_u=1) &= \iint da\, da_u\, f(a)f(a_u)\frac{\Delta a\,\Delta a_u}{s s_u}\frac{1}{Z}\exp\left(-\frac{1}{2}\Delta\mathbf{a}^{\mathrm{T}}\Sigma^{-1}\Delta\mathbf{a}\right)\\ &= s\psi'(a^{\mathrm{MP}},s^2)\, s_u\psi'(a_u^{\mathrm{MP}},s_u^2),\end{aligned}$$

where ψ is the moderated probability defined in (8) and ψ' denotes $\partial\psi/\partial a$. This yields

$$c \simeq \rho\,\frac{\partial}{\partial\rho}P(t=1,t_u=1) = \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)}\ \psi'(a^{\mathrm{MP}},s^2)\psi'(a_u^{\mathrm{MP}},s_u^2). \tag{20}$$

Substituting this into (19), we find

$$E(\Delta S_M^{(u)}|t) \simeq -\frac{(\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)})^2\ \psi'(a^{\mathrm{MP}},s^2)^2\ \psi'(a_u^{\mathrm{MP}},s_u^2)^2}{2\,P(t=1)P(t_u=1)P(t=0)P(t_u=0)}. \tag{21}$$

Assuming that the approximation $\psi \simeq \phi \equiv f(\kappa(s)a^{\mathrm{MP}})$ is good, we can numerically approximate $\partial\psi(a^{\mathrm{MP}},s^2)/\partial a$ by $\kappa(s)f'(\kappa(s)a^{\mathrm{MP}})$.[5] Using $f'=f(1-f)$ we obtain

$$E(\Delta S_M^{(u)}|t) \simeq -\kappa(s)^2\kappa(s_u)^2\ f'(\kappa(s)a^{\mathrm{MP}})f'(\kappa(s_u)a_u^{\mathrm{MP}})\,(\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)})^2/2. \tag{22}$$

The two f' terms in this expression correspond to the two intuitions that sampling near decision boundaries is informative, and that we are able to gain more information about points of interest if they are near boundaries. The term $(\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)})^2$ modifies this tendency in accordance with the desiderata.

The expected mean marginal information gain is computed by adding up the $\Delta S_M^{(u)}$s over the representative points $\mathbf{x}^{(u)}$. The resulting function is plotted on a grey scale in figure 7, for the network solving the toy problem described in figure 2. For this demonstration the points of interest $\mathbf{x}^{(u)}$ were defined by drawing 100 input points at random from the test set. A striking correlation can be seen between the regions in which the moderated output is uncertain and regions of high expected information gain. In addition the expected information gain tends to increase in regions where the training data were sparse.

Now to the negative aspect of these results. The regions of greatest expected information gain lie *outside* the region of interest to the right and left; these regions extend in long straight ridges hundreds of units away from the data. This estimation of utility, which reveals the 'hyperplanes' underlying the model, seems unreasonable. The utility of points so far from the region of interest, if they occurred, could not really be so high. There are two plausible explanations of this. It may be that the Taylor approximations used to

[5] This approximation becomes inaccurate where $a^{\mathrm{MP}} \gg s \gg 1$ (see figure 1c). Because of this it might be wise to use numerical integration then implement $\Delta S_M^{(u)}$ in look–up tables.

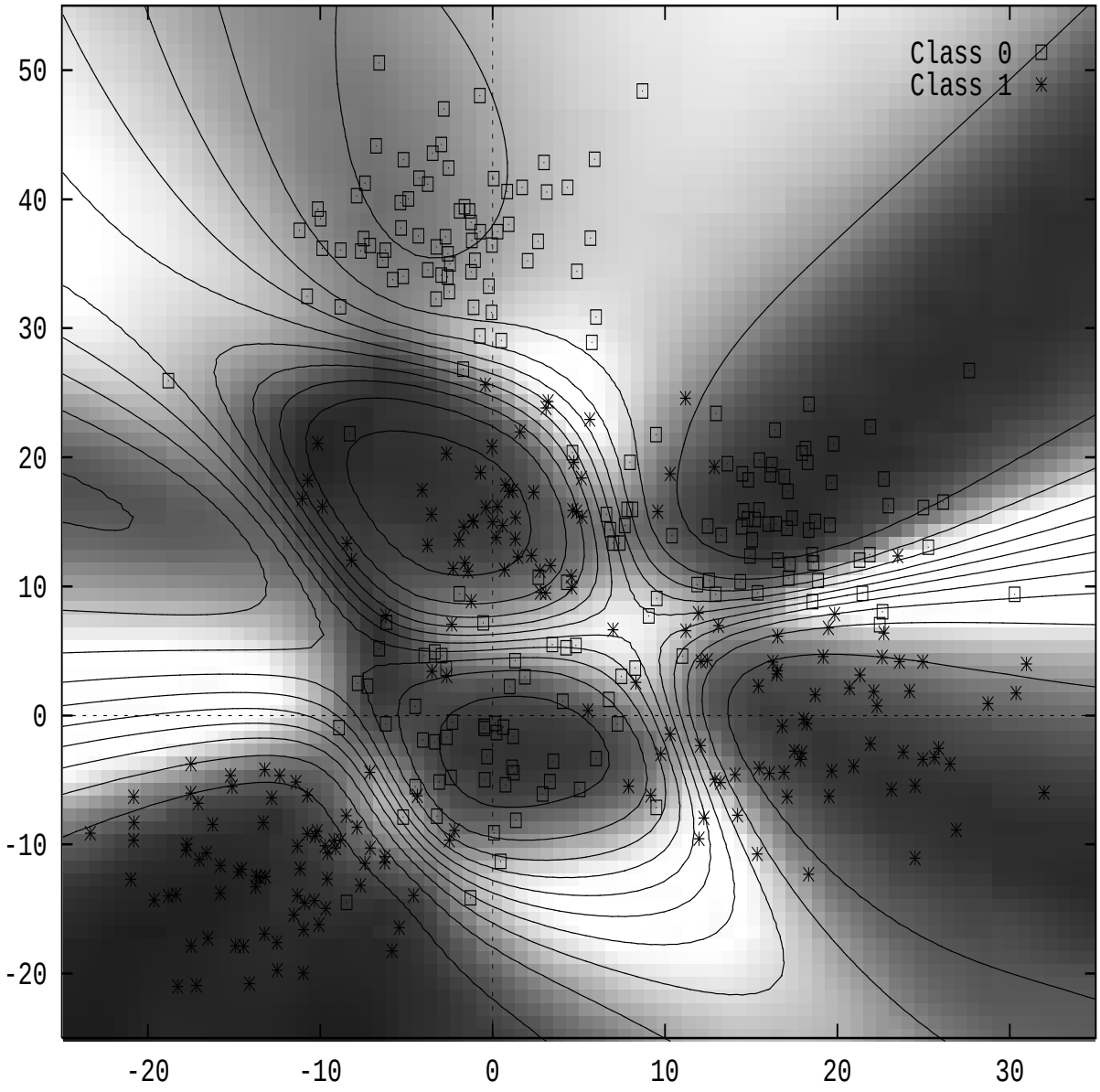

Figure 7: **Demonstration of expected mean marginal information gain**

The mean marginal information gain was computed for the network demonstrated in figures 2b,c. The region of interest was defined by 100 data points from the test set. The grey level represents the utility of a single observation as a function of where it is made. The darkest regions are expected to yield little information, and white corresponds to large expected information gain. The contours that are superposed represent the moderated output of the network, as shown in figure 2c. The mean marginal information gain is quantified: the grey scale is linear from 0 to 0.0025 nats.

evaluate the mean marginal information are at fault, in particular (20). Or as discussed in (MacKay, 1992c), the problem might arise because the mean marginal information estimates the utility of a point assuming that the model is true; if we assume that the classification surface really can be described in terms of hyperplanes in the input space, then it may be that the greatest torque on those planes can be obtained by sampling away from the core of the data. Comparison of the approximation (22) with numerical evaluations of $\Delta S_M^{(u)}$ indicate that the approximation is never more than a factor of two wrong. Thus the latter explanation is favoured, and we must tentatively conclude that the mean marginal information gain is likely to be most useful only for models well matched to the real world.

5 Discussion

Moderated outputs: The idea of moderating the outputs of a classifier in accordance with the uncertainty of its parameters should have wide applicability, for example to hidden Markov models for speech recognition. Moderation should be especially important where a classifier is expected to extrapolate to points outside the training region. There is presumably a relationship of this concept to the work of Seung *et al.* (1991) on generalisation 'at non–zero temperature.'

If the suggested approximation to the moderated output and its derivative is found unsatisfactory, a simple brute force solution would be to set up a look–up table of values of $\psi(a, s^2)$ and $\psi'(a, s^2)$.

Evidence: The evidence has been found to be well correlated with generalisation ability. This depends on having a sufficiently large amount of data. There remain open questions, including what the theoretical relationship between the evidence and generalisation ability is, and how large the data set must be for the two to be well correlated; how well these calculations will scale up to larger problems; and when the quadratic approximation for the evidence breaks down.

Mean marginal information gain: This objective function was derived with active learning in mind. It could also be used for selection of a subset of a large quantity of data, as a filter to weed out fractions of the data which are unlikely to be informative. Unlike Plutowski and White's (1991) approach this filter only depends on the *input* variables in the candidate data. A strategy that selectively omits data on the basis of their *output* values would violate the likelihood principle and risk leading to inconsistent inferences.

A comparison of the mean marginal information gain in figure 7 with the contours of the most probable networks output in figure 2b indicates that this proposed data selection criterion offers some improvements over the simple strategy of just sampling on and near decision boundaries: the mean marginal information gain shows a plausible preference for samples in regions where the decision boundary is uncertain. On the other hand, this criterion may give artefacts when applied to models that are poorly matched to the real

world. How useful the mean marginal information gain will be for real applications remains an open question.

References

E.B. Baum (1991). 'Neural net algorithms that learn in polynomial time from examples and queries', *IEEE Trans. on neural networks* **2** 1, 5–19.

C.M. Bishop (1992). 'Exact calculation of the Hessian matrix for the multilayer perceptron', *Neural Computation* **4**.

J.S. Bridle (1989). 'Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition', in *Neuro–computing: algorithms, architectures and applications*, F. Fougelman–Soulie and J. Hérault, editors, Springer–Verlag.

G.E. Hinton and T.J. Sejnowski (1986). 'Learning and relearning in Boltzmann machines', in *Parallel Distributed Processing*, Rumelhart *et al.*, MIT Press.

J.J. Hopfield (1987). 'Learning algorithms and probability distributions in feed–forward and feed–back networks', *Proc. Natl. Acad. Sci. USA* **84**, 8429–8433.

J-N. Hwang, J.J. Choi, S. Oh, and R.J. Marks II (1991). 'Query–based learning applied to partially trained multilayer perceptrons', *IEEE Trans. on neural networks* **2** 1, 131–136.

D.J.C. MacKay (1991). 'Bayesian methods for adaptive models', Ph.D. Thesis, California Institute of Technology.

D.J.C. MacKay (1992a). 'Bayesian interpolation', *Neural Computation* **4** 3. Also in (MacKay, 1991).

D.J.C. MacKay (1992b). 'A practical Bayesian framework for backprop networks', *Neural Computation* **4** 3. Also in (MacKay, 1991).

D.J.C. MacKay (1992c). 'Information–based objective functions for active data selection', to appear in *Neural Computation*. Also in (MacKay, 1991).

M. Plutowski and H. White (1991). 'Active selection of training examples for network learning in noiseless environments', Dept. Computer Science, UCSD, TR 90-011.

H.S. Seung, H. Sompolinsky and N. Tishby (1991). 'Statistical mechanics of learning from examples', preprint.

S.A. Solla, E. Levin and M. Fleisher (1988). 'Accelerated learning in layered neural networks', *Complex systems* **2**, 625–640.

D.J. Spiegelhalter and S.L. Lauritzen (1990). 'Sequential updating of conditional probabilities on directed graphical structures', *Networks* **20**, 579–605.

This work was supported by a Caltech Fellowship and a Studentship from SERC, UK.