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.* [59] on generalisation 'at non–zero temperature.'

If the suggested approximation to the moderated output and its derivative is found dissatisfactory, 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 approach [51] 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 5.7 with the contours of the most probable networks output in figure 5.2b 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.

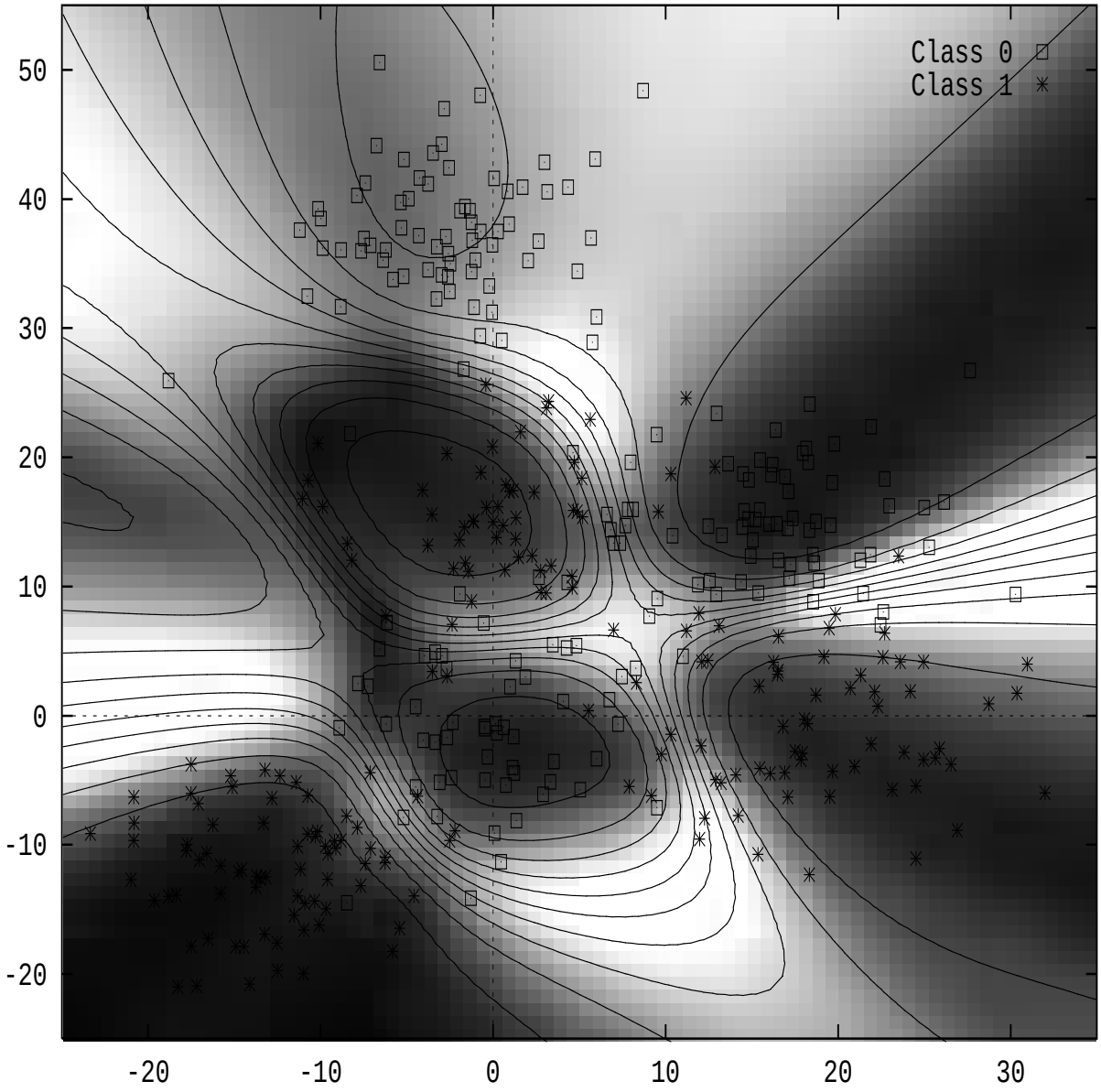

Figure 5.7: **Demonstration of expected mean marginal information gain**
The mean marginal information gain was computed for the network demonstrated in figures 5.2b,c. The region of interest was defined by 100 data points from the test set. The grey level represents the utility of a single observation 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 5.2c. The mean marginal information gain is quantified: the grey scale is linear from 0 to 0.0025 nats.

where ψ is the moderated probability defined in (5.8) and ψ' denotes $\partial\psi/\partial a$. This yields

$$c \simeq \rho \frac{\partial}{\partial \rho} P(t=1, t_u=1) = \mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)} \; \psi'(a^{\mathrm{MP}}, s^2) \psi'(a_u^{\mathrm{MP}}, s_u^2). \tag{5.20}$$

Substituting this into (5.19), we find

$$E(\Delta S_M^{(u)} | t) \simeq -\frac{(\mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)})^2 \; \psi'(a^{\mathrm{MP}}, s^2)^2 \; \psi'(a_u^{\mathrm{MP}}, s_u^2)^2}{2 \; P(t=1) P(t_u=1) P(t=0) P(t_u=0)}. \tag{5.21}$$

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

$$E(\Delta S_M^{(u)} | t) \simeq -\kappa(s)^2 \kappa(s_u)^2 \; f'(\kappa(s) a^{\mathrm{MP}}) f'(\kappa(s_u) a_u^{\mathrm{MP}}) \; (\mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)})^2 / 2. \tag{5.22}$$

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 5.7, for the network solving the toy problem described in figure 5.2. 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 evaluate the mean marginal information are at fault, in particular (5.20). Or as discussed in chapter 4, 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 (5.22) 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.

5.5 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

[5] This approximation becomes inaccurate where $a^{\mathrm{MP}} \gg s \gg 1$ (see figure 5.1c). Because of this it might be wise to use numerical integration then implement $\Delta S_M^{(u)}$ in look–up tables.

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 (5.7), and may be approximated by $\phi(a_u^{\rm MP}, s_u^2)$ (5.8).

The mean marginal entropy is

$$\bar{S}_M(P(\mathbf{w})) = \sum_u P_u S_M^{(u)}. \tag{5.14}$$

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 $\psi(a^{\rm MP}(\mathbf{x}), s^2(\mathbf{x})) \simeq \phi(a^{\rm MP}, s^2)$. 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:

$$P(t=1, t_u=1) = \int\!\!\int da\, da_u\, f(a) f(a_u) \frac{1}{Z} \exp\left(-\frac{1}{2}\Delta\mathbf{a}^{\rm T}\Sigma^{-1}\Delta\mathbf{a}\right), \text{ etc.} \tag{5.15}$$

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

$$\Sigma = \begin{pmatrix} \mathbf{g}^{\rm T}\mathbf{A}^{-1}\mathbf{g} & \mathbf{g}^{\rm T}\mathbf{A}^{-1}\mathbf{g}_{(u)} \\ \mathbf{g}^{\rm T}\mathbf{A}^{-1}\mathbf{g}_{(u)} & \mathbf{g}_{(u)}^{\rm T}\mathbf{A}^{-1}\mathbf{g}_{(u)} \end{pmatrix} \equiv \begin{pmatrix} s^2 & \rho s s_u \\ \rho s s_u & s_u^2 \end{pmatrix}. \tag{5.16}$$

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:

$$E(\Delta S_M^{(u)}|t) = S(P(t, t_u)) - S(P(t)) - S(P(t_u)). \tag{5.17}$$

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:

$$\begin{array}{ll} P(t=1,t_u=1) = P(t=1)P(t_u=1) + c & P(t=1,t_u=0) = P(t=1)P(t_u=0) - c \\ P(t=0,t_u=1) = P(t=0)P(t_u=1) - c & P(t=0,t_u=0) = P(t=0)P(t_u=0) + c. \end{array} \tag{5.18}$$

Taylor–expanding (5.17), we find

$$E(\Delta S_M^{(u)}|t) \simeq -\frac{1}{P(t=1)P(t_u=1)P(t=0)P(t_u=0)}\, c^2/2. \tag{5.19}$$

Finally, we Taylor–expand (5.15) 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

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

5.4 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.* [29] propose that samples should be made on and near the current decision boundaries. This strategy and that of Baum [5] 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 chapter, as in chapter 4, 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 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 chapter will study the 'mean marginal information.' This objective function was suggested in chapter 4, 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

$$S_M^{(u)} = y_u \log y_u + (1 - y_u) \log(1 - y_u), \tag{5.13}$$

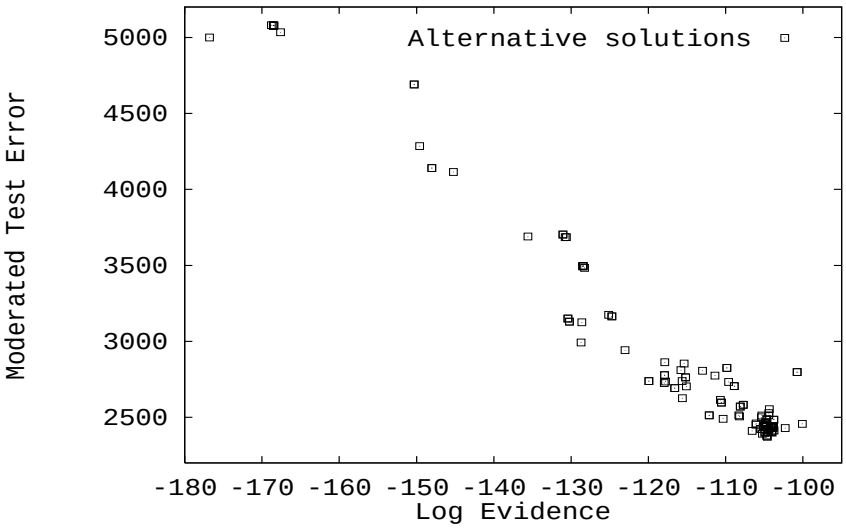

Figure 5.5: **Test error versus evidence**
Each solution was found using the same training set of $N = 300$ data points. All solutions in which a symmetry was detected among the hidden units were omitted from this graph because the evidence evaluation for such solutions is unreliable.

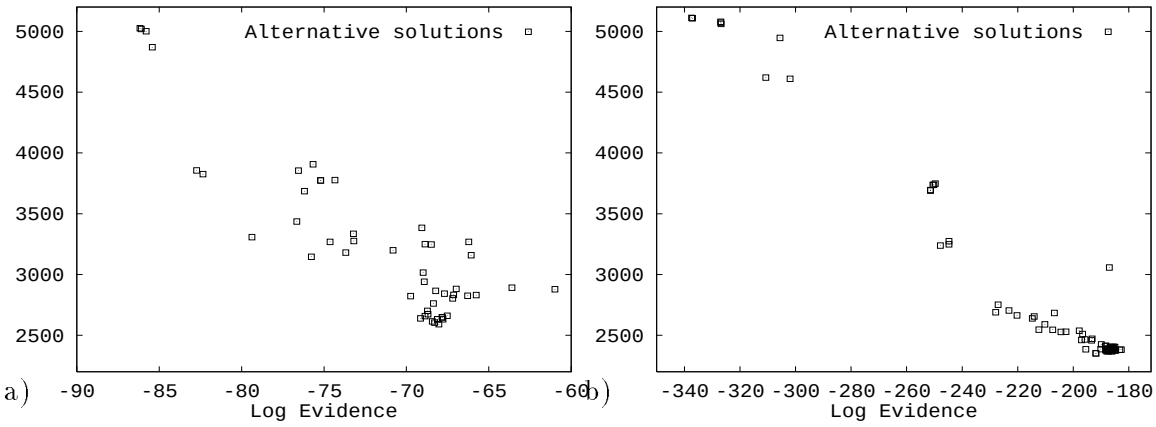

Figure 5.6: **Correlation between Test error and evidence as the amount of data varies.**
a) $N = 150$ data points. b) $N = 600$ data points. c.f. Figure 5.5, for which $N = 300$. For comparison, the number of parameters in a typical (10 hidden unit) network is 41. Note that only about 25% of the data 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.

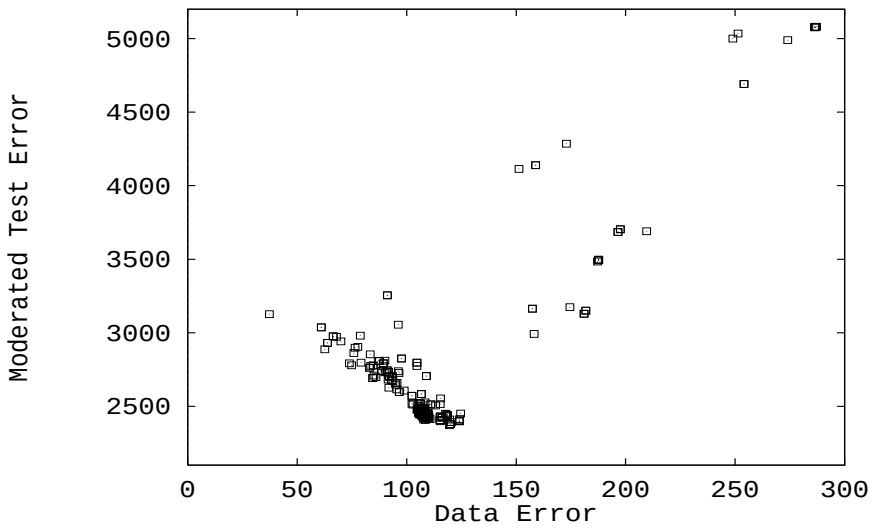

Figure 5.4: **Test error versus data error**
This figure illustrates that the task of ranking solutions to the classification problem requires Occam's razor; the solutions with smallest data error do not generalise best.

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

$$P(\{\alpha_c\}|D,\mathcal{H}) = \frac{P(D|\{\alpha_c\},\mathcal{H})P(\{\alpha_c\}|\mathcal{H})}{P(D|\mathcal{H})}. \tag{5.11}$$

The data–dependent term $P(D|\{\alpha_c\},\mathcal{H})$ is the 'evidence,' the normalising constant from equation (5.9). The evaluation of this quantity and the optimisation of the parameters $\{\alpha_c\}$ is accomplished using a framework due to Gull and Skilling, discussed in detail in chapters 2 and 3.

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

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}). \tag{5.12}$$

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 chapter 3, 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 chapter 2.

Figure 5.4 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 5.5 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 5.6 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.

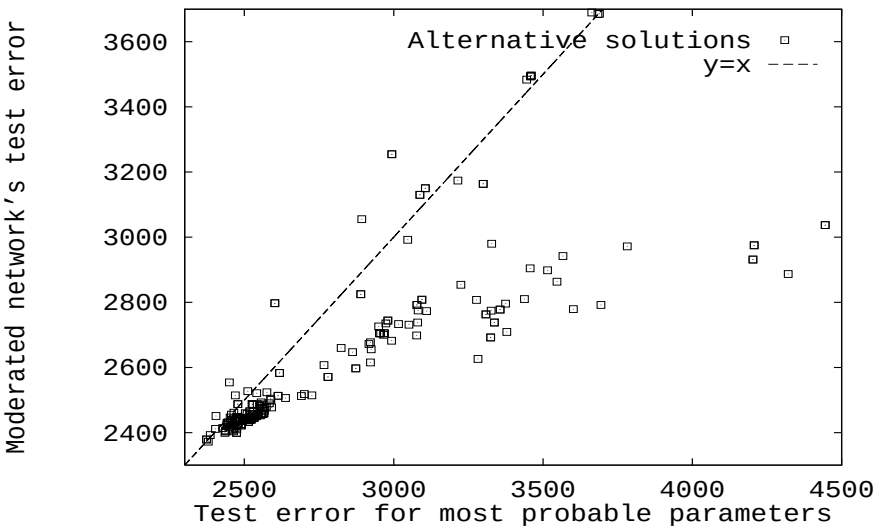

Figure 5.3: **Moderation is a good thing!**
The training set for all the networks contained 300 data points. For each network, the test error of the 'most probable' outputs and the 'moderated' outputs were evaluated on a test set of 5000 data points. The test error is the value of G. Note that for most solutions, the moderated outputs make better predictions.

the moderated outputs are found to have slightly but significantly inferior performance.

5.3 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 chapter 2, three levels of inference can be distinguished: parameter 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:

$$P(\mathbf{w}|D, \{\alpha_c\}, \mathcal{H}) = \frac{P(D|\mathbf{w}, \{\alpha_c\}, \mathcal{H}) P(\mathbf{w}|\{\alpha_c\}, \mathcal{H})}{P(D|\{\alpha_c\}, \mathcal{H})}. \tag{5.9}$$

Throughout this thesis this posterior is approximated locally by a gaussian:

$$P(\mathbf{w}|D, \{\alpha_c\}, \mathcal{H}) = \frac{\exp(-M(\mathbf{w}))}{Z_M} \simeq \frac{\exp(-M_{\rm MP} - \frac{1}{2}\Delta\mathbf{w}^{\rm T}\mathbf{A}^{-1}\Delta\mathbf{w})}{Z_M^*}, \tag{5.10}$$

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

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