literature [18], and generalisations to multiple measurements and multiple output variables can be found there, and also in [39].

In each case a function of $\mathbf{x}$ has been derived that predicts the information gain for a measurement at that $\mathbf{x}$. This function can be used to search for an optimal value of $\mathbf{x}$ (which in large–dimensional input spaces may not be a trivial task). This function could also serve as a way of reducing the size of a large data set by omitting the data points that are expected to be least informative. And this function could form the basis of a stopping rule, *i.e.* a rule for deciding whether to gather more data, given a desired exchange rate of information gain per measurement [37].

A possible weakness of these information–based approaches is that they estimate the utility of a measurement assuming that the model is correct. This might lead to undesirable results. The search for ideal measures of data utility is still open.

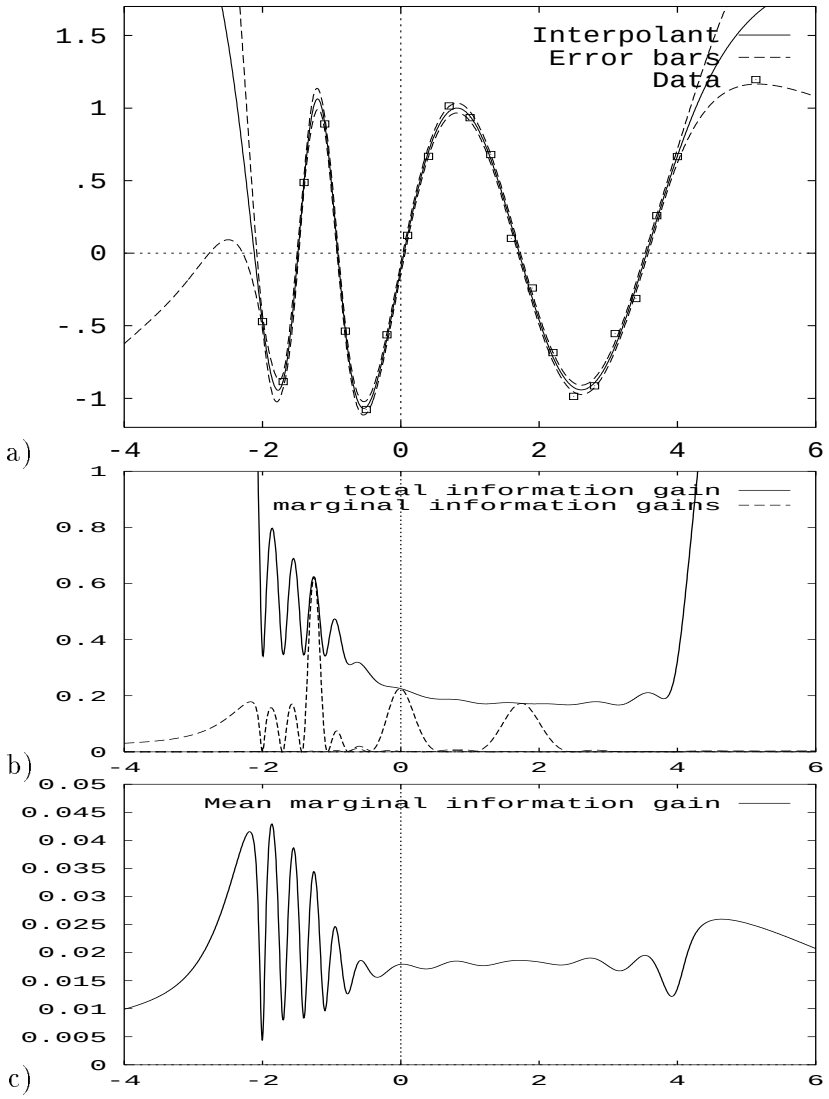

Figure 4.1: **Demonstration of total and marginal information gain**
a) The data set, the interpolant, and error bars. b) The expected total information gain and three marginal information gains. c) The mean marginal information gain, with the region of interest defined by 300 equally spaced points on the interval $[-2.1, 4.1]$. The information gains are shown on a scale of nats (1 nat $= \log_2 e$ bits).

The only surprising feature is that the estimated utility in that region is lower on the data points than the estimated utility in the smooth region towards the right.

The Achilles' heel of these methods

This approach has a potential weakness: there may be models for which, even though we have defined the region of interest by the points $\{x^{(u)}\}$, the expected marginal information gain for a measurement at x still blows up as $x \to \pm\infty$, like the error bars. This can occur because the information gain estimates the utility of a data point *assuming* that the model is correct; if we know that the model is actually an approximation tool that is incorrect, then it is possible that undesirable behaviour will result.

A simple example that illustrates this problem is obtained if we consider modelling data with a straight line $y = w_1 x$, where w_1 is the unknown parameter. Imagine that we want to select data so as to obtain a model that predicts accurately at $x^{(u)}$. Then if we assume that the model is right, clearly we gain most information if we sample at the largest possible $|x|$, since such points give the largest signal to noise ratio for determining w_1. If however we assume that the model is actually not correct, but only an approximation tool, then common sense tells us we should sample closer to $x^{(u)}$.

Thus if we are using models that we know are incorrect, the marginal information gain is really the right answer to the wrong question. It is a task for further research to formulate a new question whose answer is appropriate for any approximation model. Meanwhile, the mean marginal information gain seems a promising objective function to test further.

Computational complexity

The computation of the suggested objective functions is moderately cheap once the inverse Hessian $\mathbf{A}^{-1}$ has been obtained for the models concerned. This is a $O(Nk^2)+O(k^3)$ process, where N is the number of data points and k is the number of parameters; this process may already have been performed in order to evaluate error bars for the models, to evaluate the 'evidence,' to evaluate parameter 'saliencies,' and to enable efficient learning. This cost can be compared with the cost of locating a minimum of the objective function M, which in the worst case scales as $O(Nk^3)$ (taking the result for a quadratic function). Evaluation of the mean marginal information gain at C candidate points $\mathbf{x}$ then requires $O(Ck^2)+O(CVk)$ time, where V is the number of points of interest $\mathbf{x}^{(u)}$ ($O(k^2)$ to evaluate $\mathbf{A}^{-1}\mathbf{g}$ for each $\mathbf{x}$, and $O(Vk)$ to evaluate the dot product of this vector with each $\mathbf{g}_{(u)}$). So if $C = O(k)$ and $V = O(k)$, evaluation of the mean marginal information gain will be less computationally expensive than the inverse Hessian evaluation.

For contexts in which this is too expensive, work in progress is exploring the possibility of reducing these calculations to $O(k^2)$ or smaller time by statistical methods.

The question of how to efficiently search for the most informative $\mathbf{x}$ is not addressed here; gradient–based methods could be constructed, but figure 4.1c shows that the information gain is locally non–convex, on a scale defined by the inter–datum spacing.

4.7 Conclusion

For three specifications of the information to be maximised, a solution has been obtained. The solutions apply to linear and non–linear interpolation models, but depend on the validity of a local gaussian approximation. Each solution has a direct analog in the non–Bayesian

where the parameters μ_i, σ_i^2 are obtained for each interpolation model $\mathcal{H}_i$ from its own best fit parameters $\mathbf{w}_{\rm MP}(i)$, its own Hessian $\mathbf{A}$, and its own sensitivity vector $\mathbf{g}_i$:

$$\begin{aligned} \mu_i &= \mathbf{y}(\mathbf{x}; \mathbf{w}_{\rm MP}(i)) \\ \sigma_i^2 &= \mathbf{g}_i^{\rm T} \mathbf{A}_i^{-1} \mathbf{g}_i + 1/\beta. \end{aligned}$$

Intuitively, we expect that the most informative measurement will be at a value of $\mathbf{x}$ such that μ_1 and μ_2 are as separated as possible from each other on a scale defined by σ_1, σ_2. Further thought will also confirm that we expect to gain more information if σ_1^2 and σ_2^2 differ from each other significantly; at such points, the 'Occam factor' penalising the more powerful model becomes more significant.

Let us define the information gain to be $\Delta S = S_N - S_{N+1}$, where $S = -\sum_i P(\mathcal{H}_i) \log P(\mathcal{H}_i)$. Exact calculations of ΔS are not analytically possible, so I will assume that we are in the regime of small information gain, *i.e.* we expect measurement of t to give us a rather weak likelihood ratio $P(t|\mathcal{H}_1)/P(t|\mathcal{H}_2)$. This is the regime where $|\mu_1 - \mu_2| \ll \sigma_1, \sigma_2$.

Using this assumption we can take the expectation over t, and a page of algebra leads to the result:

$$E(\Delta S) \simeq \frac{P(\mathcal{H}_1)P(\mathcal{H}_2)}{2} \left[\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) (\mu_1 - \mu_2)^2 + \left(\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1 \sigma_2} \right)^2 \right]. \qquad (4.14)$$

These two terms correspond precisely to the two expectations stated above. The first term favours measurements where μ_1 and μ_2 are well separated; the second term favours places where σ_1^2 and σ_2^2 differ. Thus the third task has been solved.

Fedorov [18] makes a similar derivation but he uses a poor approximation which loses the second term.

4.6 Demonstration and Discussion

A data set consisting of 21 points from a one–dimensional interpolation problem was interpolated with an eight hidden unit neural network. The data were generated from a smooth function by adding noise with standard deviation $\sigma_\nu = 0.05$. The neural network was adapted to the data using weight decay terms α_c which were controlled using the methods of chapter 3 and noise level β fixed to $1/\sigma_\nu^2$. The data and the resulting interpolant, with error bars, are shown in figure 4.1a.

The expected total information gain, *i.e.* the change in entropy of the parameters, is shown as a function of x in figure 4.1b. This is just a monotonic function of the size of the error bars. The same figure also shows the expected marginal information gain about three points of interest, $\{x^{(u)}\} = \{-1.25, 0.0, 1.75\}$. Notice that the marginal information gain is in each case peaked near the point of interest, as we would expect. Note also that the height of this peak is greatest for $x^{(u)} = -1.25$, where the interpolant oscillates rapidly, and lower for $x^{(u)} = 1.75$, where the interpolant is smoother. At each $x = x^{(u)}$, the marginal information gain about $x^{(u)}$ and the total information gain are equal.

Figure 4.1c shows the mean marginal information gain, where the points of interest, $\{x^{(u)}\}$, were defined to be a set of equally spaced points on the interval $[-2.1, 4.1]$ (the same interval in which the training data lie). The mean marginal information gain gradually decreases to zero away from the region of interest, as hoped. In the region to the left where the characteristic period of the interpolant is similar to the data spacing, the expected utility oscillates as x passes through the existing data points, which also seems reasonable.

This criticism of the joint entropy is not restricted to the case $V = k$. The reason that this objective function does not achieve what we want is that the joint entropy is decreased by measurements which introduce *correlations* among predictions about $\{y^{(u)}\}$ as well as by measurements which reduce the individual uncertainties of predictions. However, we don't want the variables $\{y^{(u)}\}$ to be strongly correlated in some *arbitrary* way; rather we want each $y^{(u)}$ to have small variance, so that if we are subsequently asked to predict the value of y at any one of the u's, we will be able to make confident predictions.

Second objective function for multiple points

This motivates an alternative objective function: to maximise the average over u of the information gained about $y^{(u)}$ alone. Let us define the mean marginal entropy,

$$S^{\mathrm{M}} = \sum_u P_u \, S(P(y^{(u)})) = \frac{1}{2} \sum_u P_u \, \log \sigma_u^2 + \text{const},$$

where P_u is the probability that we will be asked to predict $y^{(u)}$, and $\sigma_u^2 = \mathbf{g}_{(u)}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)}$. For a measurement with sensitivity vector $\mathbf{g}$, we obtain from (4.10):

$$\text{Mean marginal information gain} = -\frac{1}{2} \sum_u P_u \, \log \left(1 - \frac{(\mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)})^2}{\sigma_u^2 (\sigma_\nu^2 + \sigma_{\mathbf{x}}^2)} \right). \tag{4.13}$$

The mean marginal information gain is demonstrated and criticised in section 4.6.

Two simple variations on this objective function can be derived. If instead of minimising the mean marginal entropy of our predictions $y^{(u)}$, we minimise the mean marginal entropy of the predicted noisy variables $t^{(u)}$, which are modelled as deviating from $y^{(u)}$ under additive noise of variance σ_ν^2, we obtain (4.13) with σ_u^2 replaced by $\sigma_u^2 + \sigma_\nu^2$. This alternative may lead to significantly different choices from (4.13) when any of the marginal variances σ_u^2 fall below the intrinsic variance σ_ν^2 of the predicted variable.

If instead we take an approach based on loss functions, and require that the datum we choose minimises the expectation of the mean squared error of our predictions $\{y^{(u)}\}$, which is $E^{\mathrm{M}} = \sum_u P_u \sigma_u^2$, then we obtain as our objective function, to leading order, $\Delta E^{\mathrm{M}} \simeq \sum_u P_u (\mathbf{g}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{g}_{(u)})^2 / (\sigma_\nu^2 + \sigma_{\mathbf{x}}^2)$; this increases the bias in favour of reducing the variance of the variables $y^{(u)}$ with largest σ_u^2. This is the same as the 'Q–optimal' design [18].

Comment on the case of linear models

It is interesting to note that for a linear model (one for which $\mathbf{y}(\mathbf{x}; \mathbf{w}) = \sum w_h \phi_h(\mathbf{x})$) with quadratic penalty functions, the solutions to the first and second tasks depend only on the $\mathbf{x}$ locations where data were previously gathered, not on the actual data gathered $\{\mathbf{t}\}$; this is because $\mathbf{g}(\mathbf{x}) = \phi(\mathbf{x})$ independent of $\mathbf{w}$, so $\mathbf{A} = \alpha \nabla\nabla E_W + \beta \sum_m \mathbf{g}\mathbf{g}^{\mathrm{T}}$ is independent of $\{\mathbf{t}\}$. A complete data–gathering plan can be drawn up before we start. It is only for a non–linear model that our decisions about what data to gather next are affected by our previous observations!

4.5 Maximising the discrimination between two models

Under the quadratic approximation, two models will make slightly different gaussian predictions about the value of any datum. If we measure a datum t at input value $\mathbf{x}$, then

$$P(t|\mathcal{H}_i) = \text{Normal}(\mu_i, \sigma_i^2),$$

Generalisation to multiple points

Now imagine that the objective function is defined to be the information gained about the interpolant at a set of points $\{\mathbf{x}^{(u)}\}$. These points should be thought of as representatives of the region of interest, for example, points in a test set. This case also includes the generalisation to more than one output variable y; however the full generalisation, to optimisation of an experiment in which many measurements are made, will not be made here (see Fedorov [18] and Luttrell [39]). The preceding objective function, the information about $y^{(u)}$, can be generalised in several ways, some of which lead to dissatisfactory results.

First objective function for multiple points

An obvious objective function is the *joint entropy* of the output variables that we are interested in. Let the set of output variables for which we want to minimise the uncertainty be $\{y^{(u)}\}$, where $u = 1 \ldots V$ runs either over a sequence of different input locations $\mathbf{x}^{(u)}$, or over a set of different scalar outputs, or both. Let the sensitivities of these outputs to the parameters be $\mathbf{g}_{(u)}$. Then the covariance matrix of the values $\{y^{(u)}\}$ is

$$\mathbf{Y} = \mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}, \tag{4.11}$$

where the matrix $\mathbf{G} = \left[\mathbf{g}_{(1)}\mathbf{g}_{(2)} \ldots \mathbf{g}_{(V)}\right]$. Disregarding the possibility that $\mathbf{Y}$ might not have full rank, which would necessitate a more complex treatment giving similar results, the joint entropy of our output variables $S(P(\{y^{(u)}\}))$ is related to $\log\det \mathbf{Y}^{-1}$. We can find the information gain for a measurement with sensitivity vector $\mathbf{g}$, under which $\mathbf{A} \to \mathbf{A} + \beta\mathbf{g}\mathbf{g}^{\mathrm{T}}$, using the identities (4.8).

$$\text{Joint information gain} = \frac{1}{2}\Delta \log\det \mathbf{Y}^{-1} = -\frac{1}{2}\log\left[1 - \frac{(\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G})\mathbf{Y}^{-1}(\mathbf{G}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g})}{\sigma_\nu^2 + \sigma_{\mathbf{x}}^2}\right] \tag{4.12}$$

The row vector $\mathbf{v} = \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{G}$ measures the correlations between the sensitivities $\mathbf{g}$ and $\mathbf{g}_{(u)}$. The quadratic form $\mathbf{v}\mathbf{Y}^{-1}\mathbf{v}^{\mathrm{T}}$ measures how effectively these correlations work together to reduce the joint uncertainty in $\{y^{(u)}\}$. The denominator $\sigma_\nu^2 + \sigma_{\mathbf{x}}^2$ moderates this term in favour of measurements with small uncertainty.

Criticism

I will now argue that actually the joint entropy $S(P(\{y^{(u)}\}))$ of the interpolant's values is *not* an appropriate objective function. A simple example will illustrate this.

Imagine that $V = k$, *i.e.* the number of points defining our region of interest is the same as the dimensionality of the parameter space $\mathbf{w}$. The resulting matrix $\mathbf{G} = \left[\mathbf{g}_{(1)}\mathbf{g}_{(2)} \ldots \mathbf{g}_{(V)}\right]$ may be almost singular if the points $\mathbf{x}^{(u)}$ are close together, but typically it will still have full rank. Then the parameter vector $\mathbf{w}$ and the values of the interpolant $\{y^{(u)}\}$ are in one to one (locally) linear correspondence with each other. This means that the change in entropy of $P(\{y^{(u)}\})$ is *identical* to the change in entropy of $P(\mathbf{w})$ [37]. This can be confirmed by substitution of $\mathbf{Y}^{-1} = \mathbf{G}^{-1}\mathbf{A}\mathbf{G}^{-1^{\mathrm{T}}}$ into (4.12), which yields (4.9). So if the datum is chosen in accordance with equation (4.12), so as to maximise the expected joint information gain about $\{y^{(u)}\}$, exactly the same choice will result as is obtained maximising the first criterion, the expected total information gain about $\mathbf{w}$ (section 4.3)! Clearly, this choice is independent of our choice of $\{y^{(u)}\}$, so it will have nothing to do with our region of interest.

Let us now see what property of a datum causes it to be maximally informative. The new entropy S_{N+1} is equal to $-\frac{1}{2}\log(m^2 \det \mathbf{A}_{N+1})$, neglecting additive constants. This determinant can be analytically evaluated [18], using the identities

$$[\mathbf{A} + \beta \mathbf{g}\mathbf{g}^{\mathrm{T}}]^{-1} = \mathbf{A}^{-1} - \frac{\beta \mathbf{A}^{-1}\mathbf{g}\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}}{1 + \beta \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}} \quad \text{and} \quad \det[\mathbf{A} + \beta \mathbf{g}\mathbf{g}^{\mathrm{T}}] = (\det \mathbf{A})(1 + \beta \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}), \quad (4.8)$$

from which we obtain:

$$\text{Total information gain} = \frac{1}{2}\Delta \log\left(m^2 \det \mathbf{A}\right) = \frac{1}{2}\log(1 + \beta \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}). \quad (4.9)$$

In the product $\beta \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}$, the first term β tells us that, not surprisingly, we learn more information if we make a low noise (high β) measurement. The second term $\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}$ is precisely the variance of the interpolant at the point where the datum is collected.

Thus we have our first result: to obtain maximal information about the interpolant, take the next datum at the point where the error bars on the interpolant are currently largest (assuming the noise σ_ν^2 on all measurements is the same). This rule is the same as that resulting from the 'D–optimal' and 'minimax' design criteria [18].

For many interpolation models, the error bars are largest beyond the most extreme points where data have been gathered. This first criterion would in those cases lead us to repeatedly gather data at the edges of the input space, which might be considered non–ideal behaviour; but we do not necessarily need to introduce an ad hoc procedure to avoid this. The reason we do not want repeated sampling at the edges is that we do not want to *know* what happens there. Accordingly, we can derive criteria from alternative objective functions which only value information acquired about the interpolant in a defined region of interest.

4.4 Maximising information about the interpolant in a region of interest

Thus we come to the second task. First assume we wish to gain maximal information about the value of the interpolant at a particular point $\mathbf{x}^{(u)}$. Under the quadratic approximation, our uncertainty about the interpolant $\mathbf{y}$ has a gaussian distribution, and the size of the error bars is given in terms of the Hessian of the parameters by

$$\sigma_u^2 = \mathbf{g}_{(u)}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)},$$

where $\mathbf{g}_{(u)}$ is $\partial y/\partial \mathbf{w}$ evaluated at $\mathbf{x}^{(u)}$. As above, the entropy of this gaussian distribution is $\frac{1}{2}\log \sigma_u^2 + \text{const}$. After a measurement t is made at $\mathbf{x}$ where the sensitivity is $\mathbf{g}$, these error bars are scaled down by a factor of $1-\rho^2$, where ρ is the correlation between the variables t and $y^{(u)}$, given by $\rho^2 = (\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)})^2/(\sigma_u^2(\sigma_\nu^2 + \sigma_{\mathbf{x}}^2))$, where $\sigma_{\mathbf{x}}^2 = \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}$. Thus the information gain about $y^{(u)}$ is:

$$\text{Marginal information gain} = \frac{1}{2}\Delta \log \sigma_u^2 = -\frac{1}{2}\log\left(1 - \frac{(\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)})^2}{\sigma_u^2(\sigma_\nu^2 + \sigma_{\mathbf{x}}^2)}\right). \quad (4.10)$$

The term $\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_{(u)}$ is maximised when the sensitivities $\mathbf{g}$ and $\mathbf{g}_{(u)}$ are maximally correlated, as measured by their inner product in the metric defined by $\mathbf{A}^{-1}$. The second task is thus solved for the case of extrapolation to a single point. This objective function is demonstrated and criticised in section 4.6.

4.3 Maximising total information gain

Let us now solve the first task: how to choose $\mathbf{x}^{N+1}$ so that the expected information gain about $\mathbf{w}$ is maximised. Intuitively we expect that we will learn most about the interpolant by gathering data at the $\mathbf{x}$ location where our error bars on the interpolant are currently greatest. Within the quadratic approximation, we will now confirm that intuition.

Notation

The likelihood of the data is defined in terms of a noise level $\sigma_\nu^2 = \beta^{-1}$ by $P(\{\mathbf{t}\}|\mathbf{w}, \beta, \mathcal{N}) = \exp(-\beta E_D(\mathbf{w}))/Z_D$, where $E_D(\mathbf{w}) = \sum_m \frac{1}{2}(\mathbf{t}^m - \mathbf{y}(\mathbf{x}^{(m)}; \mathbf{w}))^2$, and Z_D is the appropriate normalising constant. The likelihood could also be defined with an $\mathbf{x}$–dependent noise level $\beta^{-1}(\mathbf{x})$, or correlated noise in multiple outputs (in which case β^{-1} would be the covariance matrix of the noise). From here on $\mathbf{y}$ will be treated as a scalar y for simplicity. When the likelihood for the first N data is combined with a prior $P(\mathbf{w}|\alpha, \mathcal{R}) = \exp(-\alpha E_W(\mathbf{w}))/Z_W$, in which the regularising constant (or weight decay rate) α corresponds to the prior expected smoothness of the interpolant, we obtain our current probability distribution for $\mathbf{w}$, $P^N(\mathbf{w}) = \exp(-M(\mathbf{w}))/Z_M$, where $M(\mathbf{w}) = \alpha E_W + \beta E_D$. The objective function $M(\mathbf{w})$ can be quadratically approximated near to the most probable parameter vector, $\mathbf{w}_{\rm MP}$, by

$$M(\mathbf{w}) \simeq M^*(\mathbf{w}) = M(\mathbf{w}_{\rm MP}) + \frac{1}{2}\Delta\mathbf{w}^{\rm T}\mathbf{A}\Delta\mathbf{w}, \tag{4.4}$$

where $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\rm MP}$ and the Hessian $\mathbf{A} = \nabla\nabla M$ is evaluated at the minimum $\mathbf{w}_{\rm MP}$. We will use this quadratic approximation from here on. If M has other minima, those can be treated as distinct models as in chapter 3.

First we will need to know what the entropy of a gaussian distribution is. It is easy to confirm that if $P(\mathbf{w}) \propto e^{-M^*(\mathbf{w})}$, then for a flat measure $m(\mathbf{w}) = m$,

$$S = \frac{k}{2}(1 + \log 2\pi) + \frac{1}{2}\log\left(m^2 \det \mathbf{A}^{-1}\right). \tag{4.5}$$

Thus our aim in minimising S is to make the size of the joint error bars on the parameters, $\det \mathbf{A}^{-1}$, as small as possible.

Expanding $\mathbf{y}$ around $\mathbf{w}_{\rm MP}$, let

$$\mathbf{y}(\mathbf{x}) \simeq \mathbf{y}(\mathbf{x}; \mathbf{w}_{\rm MP}) + \mathbf{g}(\mathbf{x}) \cdot \Delta\mathbf{w}, \tag{4.6}$$

where $g_j = \frac{\partial y}{\partial w_j}$ is the ($\mathbf{x}$–dependent) sensitivity of the output variable to parameter w_j, evaluated at $\mathbf{w}_{\rm MP}$.

Now imagine that we choose a particular input $\mathbf{x}$ and collect a new datum. If the datum $\mathbf{t}$ falls in the region such that our quadratic approximation applies, the new Hessian $\mathbf{A}_{N+1}$ is:

$$\mathbf{A}_{N+1} \simeq \mathbf{A} + \beta\mathbf{g}\mathbf{g}^{\rm T}, \tag{4.7}$$

where we have used the approximation $\nabla\nabla\frac{1}{2}(\mathbf{t} - \mathbf{y}(\mathbf{x}; \mathbf{w}))^2 \simeq \mathbf{g}\mathbf{g}^{\rm T}$. This expression neglects terms in $\frac{\partial^2 y}{\partial w_j \partial w_k}$; those terms are exactly zero for the linear models discussed in chapter 2, but they are not necessarily negligible for non–linear models such as neural networks. Notice that this new Hessian is independent of the value that the datum $\mathbf{t}$ actually takes, so we can specify what the information gain ΔS will be for any datum, because we can evaluate $\mathbf{A}_{N+1}$ just by calculating $\mathbf{g}$.

An alternative information measure is the *cross entropy* between $P^N(\mathbf{w})$ and $P^{N+1}(\mathbf{w})$:

$$G = \int d^k\mathbf{w}\, P^{N+1}(\mathbf{w}) \log \frac{P^N(\mathbf{w})}{P^{N+1}(\mathbf{w})}. \tag{4.2}$$

Let us define $G' = -G$ so as to obtain a positive quantity; then G' is a measure of how much information we gain when we are informed that the true distribution of $\mathbf{w}$ is $P^{N+1}(\mathbf{w})$, rather than $P^N(\mathbf{w})$.

These two information measures are not equal. Intuitively they differ in that if the measure $m(\mathbf{w})$ is flat, ΔS only quantifies how much the probability 'bubble' of $P(\mathbf{w})$ shrinks when the new datum arrives; G' also incorporates a measure of how much the bubble *moves* because of the new datum. Thus according to G', even if the probability distribution does not shrink and become more certain, we have *learnt* something if the distribution moves from one region to another in $\mathbf{w}$–space.

The question of which information measure is appropriate is potentially complicated by the fact that G' is not a consistent additive measure of information: if we receive datum A then datum B, in general, $G'_{AB} \neq G'_A + G'_B$. This intriguing complication will not however hinder our task: we can only base our decisions on the *expectations* of ΔS and G'; we will now see that in expectation ΔS and G' are equal, so for our purposes there is no distinction between them. This result holds independent of the details of the models we study and independent of any gaussian approximation for $P(\mathbf{w})$.

Proof that $E(\Delta S) = E(G')$

To evaluate the expectation of these quantities, we have to assume a probability distribution from which the datum $\mathbf{t}^{N+1}$ (hence abbreviated as $\mathbf{t}$) comes. We will define this probability distribution by assuming that our current model, complete with its error bars, is correct. This means that the probability distribution of $\mathbf{t}$ is $P(\mathbf{t}|D_N, \mathcal{H})$, where $\mathcal{H}$ is the total specification of our model. The conditioning variables on the right will be omitted in the following proof.

We can now compare the expectations of ΔS and G'.

$$\begin{aligned} G' &= -\int d^k\mathbf{w}\, P(\mathbf{w}|\mathbf{t}) \log \frac{P(\mathbf{w})}{P(\mathbf{w}|\mathbf{t})} \\ &= -\int d^k\mathbf{w}\, P(\mathbf{w}|\mathbf{t}) \log \frac{m(\mathbf{w})}{P(\mathbf{w}|\mathbf{t})} + \int d^k\mathbf{w}\, P(\mathbf{w}|\mathbf{t}) \log \frac{m(\mathbf{w})}{P(\mathbf{w})}, \end{aligned} \tag{4.3}$$

where m is free to be any measure on $\mathbf{w}$; let us make it the same measure m as in (4.1). Then the first term in (4.3) is $-S_{N+1}$. So

$$\begin{aligned} E(G') &= -E(S_{N+1}) + \int d\mathbf{t}\, P(\mathbf{t}) \int d^k\mathbf{w}\, P(\mathbf{w}|\mathbf{t}) \log \frac{m(\mathbf{w})}{P(\mathbf{w})} \\ &= -E(S_{N+1}) + \int d^k\mathbf{w}\, P(\mathbf{w}) \log \frac{m(\mathbf{w})}{P(\mathbf{w})} \\ &= E(-S_{N+1} + S_N) = E(\Delta S). \end{aligned}$$

•

Thus the two candidate information measures are equivalent for our purposes. This proof also implicitly demonstrates that $E(\Delta S)$ is independent of the measure $m(\mathbf{w})$. Other properties of $E(\Delta S)$ are proved in [37]. The rest of this chapter will use ΔS as the information measure, with $m(\mathbf{w})$ set to a constant.

The first and third definitions of information gain have both been studied in the abstract by Lindley [37]. All three cases have been studied by Fedorov [18], mainly in non–Bayesian terms. In this chapter, solutions will be obtained for the interpolation problem by using a gaussian approximation and in some cases assuming that the new datum is a relatively weak piece of information. In common with most other work on active learning, the utility is evaluated assuming that the probability distributions defined by the interpolation model are correct. For some models, this assumption may be the Achilles' heel of this approach, as discussed in section 4.6.

Can our choice bias our inferences?

One might speculate that the way we choose to gather data might be able to bias our inferences systematically away from the truth. If this were the case we might need to make our inferences in a way which undoes such biases by taking into account how we gathered the data. In orthodox statistics many estimators and statistical tests do depend on the sampling strategy.

However, the *likelihood principle* states that our inferences should depend on the likelihood of the actual data received, not on other data that we might have gathered but didn't. Bayesian inference is consistent with this principle; there is no need to undo biases introduced by the data collecting strategy, because it is *not possible* for such biases to be introduced — as long as we perform inference using all the data gathered [7, 38]. When the models are concerned with estimating the distribution of output variables $\mathbf{t}$ given input variables $\mathbf{x}$, we are allowed to look at the $\mathbf{x}$ value of a datum, and decide whether or not to include the datum in the data set. This will not bias our inferences about the distribution $P(\mathbf{t}|\mathbf{x})$.

4.2 Choice of information measure

Before we can start, we need to select a measure of the information gained about an unknown variable when we receive the new datum $\mathbf{t}^{N+1}$. Having chosen such a measure we will then select the $\mathbf{x}^{N+1}$ for which the *expected* information gain is maximal. Two measures of information have been suggested, both based on Shannon's entropy, whose properties as a sensible information measure are well known. Let us explore this choice for the first task, where we want to gain maximal information about the parameters of the interpolant, $\mathbf{w}$.

Let the probability distributions of the parameters before and after we receive the datum $\mathbf{t}^{N+1}$ be $P^N(\mathbf{w})$ and $P^{N+1}(\mathbf{w})$. Then the *change in entropy* of the distribution is $\Delta S = S_N - S_{N+1}$, where:

$$S_N = \int d^k\mathbf{w}\, P^N(\mathbf{w}) \log \frac{m(\mathbf{w})}{P^N(\mathbf{w})}, \tag{4.1}$$

where m is the measure on $\mathbf{w}$ that makes the argument of the log dimensionless.[1] The greater ΔS is, the more information we have gained about $\mathbf{w}$. In the case of the quadratic models discussed in chapter 2, if we set the measure $m(\mathbf{w})$ equal to the prior $P^0(\mathbf{w})$, the quantity S_N is closely related to the log of the 'Occam factor.'[2]

[1] This measure m will be unimportant in what follows but is included to avoid committing dimensional crimes. Note that the sign of ΔS has been defined so that our information gain corresponds to positive ΔS.

[2] If the Occam factor is $O.F. = (2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}\, \exp(-\alpha E_W^{\rm MP})/Z_W(\alpha)$, then $S_N = \log O.F. + \gamma/2$, using notation from chapter 2.

given by John Skilling at Maxent 91 [66].

Recent work in the neural networks literature on active data selection, also known as 'query learning,' has concentrated on slightly different problems: The work of Baum [5] and Hwang *et. al.* [29] relates to perfectly separable classification problems only; in both these papers a sensible query–based learning algorithm is proposed, and empirical results of the algorithm are reported; Baum also gives a convergence proof. But since the algorithms are both human–designed, it is not clear what objective function their querying strategy optimises, nor how the algorithms could be improved. In contrast, this chapter (which discusses noisy interpolation problems) *derives* criteria from *defined* objective functions; each objective function leads to a different data selection criterion. Chapter 5 will discuss the application of the same ideas to classification problems.

Plutowski and White [51] study a different problem from the above, in the context of noise–free interpolation: they assume that a large amount of data has already been gathered, and work on principles for selecting a subset of that data for efficient training; the entire data set (inputs *and* targets) is consulted at each iteration to decide which example to add to the training subset, an option that is not permitted here.

Statement of the problem

Imagine that we are gathering data in the form of a set of input–output pairs $D_N = \{\mathbf{x}^{(m)}, \mathbf{t}^{(m)}\}$, where $m = 1 \ldots N$. This data is modelled with an interpolant $\mathbf{y}(\mathbf{x}; \mathbf{w}, \mathcal{A})$. An interpolation model $\mathcal{H}$ specifies the 'architecture' $\mathcal{A}$, which defines the functional dependence of the interpolant on the parameters w_i, $i = 1 \ldots k$. The model also specifies a regulariser, or prior on $\mathbf{w}$, and a cost function, or noise model $\mathcal{N}$ describing the expected relationship between $\mathbf{y}$ and $\mathbf{t}$. We may have more than one interpolation model, which may be linear or non–linear in $\mathbf{w}$. Chapters 2 and 3 described the Bayesian framework for fitting and comparing such models, assuming a fixed data set. This chapter discusses how the same framework for interpolation relates to the task of selecting *what data to gather next.*

Our criterion for how informative a new datum is will depend on what we are interested in. Several alternatives spring to mind:

1. If we have decided to use one particular interpolation model, we might wish to select new data points to be maximally informative about the values that that model's parameters $\mathbf{w}$ should take.

2. Alternatively, we might not be interested in getting a globally well–determined interpolant; we might only want to be able to predict the value of the interpolant accurately in a limited region, perhaps at a point in input space which we are not able to sample directly.

3. Lastly, we might be unsure which of two or more models is the best interpolation model, and we might want to select data so as to give us maximal information to discriminate between the models.

This chapter will study each of these tasks for the case where we wish to evaluate the utility as a function of $\mathbf{x}^{N+1}$, the input location at which a single measurement of a scalar t^{N+1} will be made. The more complex task of selecting *multiple* new data points will not be addressed here, but the methods used can be generalised to solve this task, as is discussed in [18, 39]. The similar problem of choosing the $\mathbf{x}^{N+1}$ at which a *vector* of outputs $\mathbf{t}^{N+1}$ is measured will not be addressed either.

Chapter 4

Information–based objective functions for active data selection

Abstract

Learning can be made more efficient if we can actively select particularly salient data points. Within a Bayesian learning framework, objective functions are discussed which measure the *expected informativeness* of candidate measurements. Three alternative specifications of what we want to gain information about lead to three different criteria for data selection. All these criteria depend on the assumption that the hypothesis space is correct, which may prove to be their main weakness.

4.1 Introduction

Theories for data modelling often assume that the data is provided by a source that we do not control. However, there are two scenarios in which we are able to actively select training data. In the first, data measurements are relatively expensive or slow, and we want to know where to look next so as to learn as much as possible. According to Jaynes [30], Bayesian reasoning was first applied to this problem two centuries ago by Laplace, who in consequence made more important discoveries in celestial mechanics than anyone else. In the second scenario, there is an immense amount of data and we wish to select a subset of data points that are most useful for our purposes. Both these scenarios will benefit if we have ways of objectively estimating the utility of candidate data points.

The problem of 'active learning' or 'sequential design' has been extensively studied in economic theory and statistics [16, 18]. Experimental design within a Bayesian framework using the Shannon information as an objective function has been studied by Lindley [37] and by Luttrell [39]. A distinctive feature of this approach is that it renders the optimisation of the experimental design independent of the 'tests' that are to be applied to the data and the loss functions associated with any decisions. This chapter uses similar information–based objective functions and discusses the problem of optimal data selection within the Bayesian framework for interpolation described in chapters 2 and 3. Most of the results in this chapter have direct analogs in Fedorov [18], though the quantities involved have different interpretations: for example, Fedorov's dispersion of an estimator becomes the Bayesian's posterior variance of the parameter. This work was directly stimulated by a presentation

[0]Chapter 4 of PhD thesis 'Bayesian Methods for Adaptive Models' by David MacKay, California Institute of Technology, submitted December 10 1991.