determined more by the other measurements and by the regulariser than by t_m, and the error E_{Dm} does convey information about the noise level.

The similarity of γ_m to the concept of the number of well–determined parameters, γ, will be obvious, and in fact there is a mathematical relationship too. The number of *bad* noise measurements is identical to the number of well–determined parameters, *i.e.* $\sum_m (1 - \gamma_m) = \gamma$.

An implementation of this framework will depend on two further issues. First, a prior should of course be placed on the parameters of the network that produces $\log \beta(\mathbf{x})$; this prior might contain unknown regularisation constants which can be controlled using the methods of chapter 3. Secondly, the management of these different levels of optimisation will not be trivial. A suggested procedure is to start optimising the noise model only once the interpolant is fitting the data quite well; then the three levels (fitting the interpolant, inferring the noise level, and setting regularisation constants of the noise model) could be optimised cyclically.

in the same place as the centre of mass of the likelihood. When we marginalise over μ, with a flat prior, we find that the most probable value of σ^2 is $\sigma^2_{N-1} = \sum(x-\bar{x})^2/(N-1)$. The subtraction of one from the denominator represents the fact that one parameter μ has been determined by the data, which typically consumes one unit (χ^2) of noise. It is well known that σ^2_N and σ^2_{N-1} are respectively 'biased' and 'unbiased' estimators of variance.

The generalisation of this distinction has already been given. When we fit a regularised interpolation model, there are distinct levels of inference. At the first level, we assume a particular noise level β and regularisation constant α, and find the most probable parameters $\mathbf{w}$ with error bars. Then at the second level of inference, we compare alternative values of α, and alternative values of β. When this separation is made, we find that the most probable noise level is given by $\sigma^2 = 2E_D/(N-\gamma)$, where γ is the number of well–determined parameters. This quantity will be significantly less than the total number of parameters k if the regulariser (prior) is playing a significant role in determining the interpolant. The separation of the two levels of inference and the use of marginalisation thus leads to an unbiased estimator for β and an automatic Occam's razor for the choice of α.

Let us now see how this should work in the case of an $\mathbf{x}$–dependent noise level.

At the first level of inference, the gradients with respect to model parameters will be calculated in the obvious way: $\partial M/\partial \mathbf{w} = \alpha \partial E_W/\partial \mathbf{w} + \sum_m \beta(\mathbf{x}_m)\partial E_{Dm}/\partial \mathbf{w}$.

The second level of inference centres on the log evidence for α and β, which can be written (neglecting additive constants):

$$\log P(D|\alpha, \beta(\mathbf{x}), \mathcal{H}) = -\alpha E_W^{\mathrm{MP}} - \sum_m \beta(\mathbf{x}_m) E_{Dm}^{\mathrm{MP}} - \frac{1}{2}\log\det \mathbf{A} - \log Z_W(\alpha) + \sum_m \frac{1}{2}\log\beta(\mathbf{x}_m) \tag{6.1}$$

The most probable values of α and β, if we have a flat prior, are obtained by maximising the log evidence. Of course if we infer an $\mathbf{x}$–dependent noise level, we typically *will* be imposing a prior on $\beta(\mathbf{x})$ by the choice of model; in this case we will still need the gradient of the evidence, which at this level serves as the likelihood driving the learning. When we differentiate the evidence with respect to $\log\beta(\mathbf{x}_m)$, we obtain:

$$\frac{\partial \log P(D|\alpha,\beta,\mathcal{A},\mathcal{R})}{\partial \log\beta(\mathbf{x}_m)} = -\beta(\mathbf{x}_m)E_{Dm}^{\mathrm{MP}} - \frac{1}{2}\beta(\mathbf{x}_m)\,\mathrm{Trace}\left[\mathbf{A}^{-1}\mathbf{B}_m\right] + \frac{1}{2} \tag{6.2}$$

where $\beta(\mathbf{x}_m)\mathbf{B}_m$ is the contribution to $\mathbf{A}$ made by the mth datum, which in the case of a linear model is $\beta(\mathbf{x}_m)\mathbf{B}_m = \beta(\mathbf{x}_m)\mathbf{g}_m\mathbf{g}_m^{\mathrm{T}}$, with $\mathbf{g}_m = \partial y(\mathbf{x}_m)/\partial \mathbf{w}$. The quantity $\mathrm{Trace}\left[\mathbf{A}^{-1}\mathbf{B}_m\right]$ is precisely the magnitude of the error bars on the interpolant at $\mathbf{x}_m$, $\mathbf{g}_m^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_m$. Thus the term $\beta(\mathbf{x}_m)\mathrm{Trace}\left[\mathbf{A}^{-1}\mathbf{B}_m\right] = \mathbf{g}_m^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_m/\sigma^2(\mathbf{x}_m)$ is the ratio of the error bars on the interpolant to the presumed variance of the measurement at $\mathbf{x}_m$. The quantity $\gamma_m \equiv 1 - \mathbf{g}_m^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}_m/\sigma^2(\mathbf{x}_m)$ is a measure of *how good a noise measurement* the datum at $\mathbf{x}_m$ contributed.

The gradient can then be written:

$$\frac{\partial \log P(D|\alpha,\beta,\mathcal{A},\mathcal{R})}{\partial \log\beta(\mathbf{x}_m)} = -\beta(\mathbf{x}_m)E_{Dm}^{\mathrm{MP}} + \gamma_m/2 \tag{6.3}$$

The terms γ_m are easy to evaluate from the error bars at $\mathbf{x}_m$. It can be noted that in the case of an isolated data point, the contribution to the gradient is well–behaved, because E_{Dm} and γ both go to zero. So there is no singular behaviour involving β blowing up, as occurs if the likelihood is maximised.

If $\gamma_m \simeq 0$, the interpolant is determined locally only by the datum at $\mathbf{x}_m$, and the measurement gives no information about the noise. If $\gamma_m \simeq 1$, the interpolant is locally

Chapter 6

Inferring an input–dependent noise level

Abstract

Assume that when interpolating a data set, we wish to model an input–dependent noise level. This short chapter shows how to calculate an unbiased gradient.

Given a data set $D = \{\mathbf{x}_m, t_m\}$ modelled with an interpolant–plus–noise model $t_m = y(\mathbf{x}_m; \mathbf{w}) + \nu_m$, chapter 2 described the Bayesian framework for regularisation and model comparison assuming a single global noise level $\beta^{-1} = \sigma_\nu^2$. It is also possible to invent models in which β is $\mathbf{x}$–dependent. For example we might use two coupled neural networks, the first of which predicts $y(\mathbf{x})$, and the second of which predicts $\log \beta(\mathbf{x})$. These networks would be coupled in that the gradient of the objective function for each network would be calculated by consulting the output of the other. Intuitively, if the first network's errors in the neighbourhood of $\mathbf{x}$ are large, then we encourage the second network to give a large value of $\beta^{-1}(\mathbf{x}) = \sigma_\nu^2(\mathbf{x})$ there; and similarly the error signals that teach the first network need to be scaled up and down by the second network — where there is a small value of $\sigma_\nu^2(\mathbf{x})$, errors are penalised more strongly.

What should the gradients for optimisation of such a model be? A simple approach would be to maximise the likelihood of the data. For the traditional quadratic model, the log likelihood of the data is $-\sum_m \beta(\mathbf{x}_m) E_{Dm} + \sum_m \frac{1}{2} \log \beta(\mathbf{x}_m)$, where $E_{Dm} = \frac{1}{2}(t_m - y(\mathbf{x}_m; \mathbf{w}))^2$. However maximisation of this function would lead to *biased* estimates of $\beta(\mathbf{x})$. As was discussed earlier, the maximum likelihood noise estimate is not the most probable value of the noise. This distinction is not caused by the use of priors; rather it is a result of *marginalisation*. The worst symptom of maximising the likelihood would be that in regions in which data is sparse, such that the best fit interpolant passes very close to the data, the maximum likelihood estimate of β blows up: the estimated noise level goes to zero.

Separating the two levels of inference

Let us consider the case of a single noise level σ^2 for a moment, and imagine that we are estimating a single parameter μ corresponding to the mean of a gaussian cluster. We already examined this problem in chapter 2. The likelihood, as a function of μ and σ^2, has a skew peak. The maximum is located at $(\bar{x}, \sigma_N^2)$, where $\sigma_N^2 = \sum (x - \bar{x})^2 / N$, but this peak is not

[0] Chapter 6 of PhD thesis 'Bayesian Methods for Adaptive Models' by David MacKay, California Institute of Technology, submitted December 10 1991.