to obtain the second derivatives, neglecting terms in f'', where f is the activation function of a neuron. The Hessian is built up as a sum of outer products of gradient vectors:

$$\nabla\nabla E_D \simeq \sum_{i,m} \mathbf{g}_i^m \mathbf{g}_i^{m\mathrm{T}} \tag{3.17}$$

where $\mathbf{g}_i^m = \frac{dy_i(\mathbf{x}^m)}{d\mathbf{w}}$. Unlike Denker *et. al.*, I did not ignore the off–diagonal terms; the diagonal approximation is not good enough! For the evaluation of γ the two methods gave similar results, and either approach seemed satisfactory. However, for the evaluation of the evidence, the approximate analytic method failed to give satisfactory results. The 'Occam factors' are very weak, scaling only as $\log N$, and the above approximation apparently introduces systematic errors greater than these. The reason that the evidence evaluation is more sensitive to errors than the γ evaluation is because γ is related to the sum of eigenvalues, whereas the evidence is related to the product; errors in small eigenvalues jeopardise the product more than the sum. I expect an exact analytic evaluation of the second derivatives [8] would resolve this. To save programming effort I instead used second differences, which is computationally more demanding ($\sim kN$ backprops) than the analytic approach ($\sim N$ backprops). There were still problems with errors in small eigenvalues, but it was possible to correct these errors, by detecting eigenvalues which were smaller than theoretically permitted.

Demonstrations

The demonstrations were performed as follows:
Initial weight configuration: random weights drawn from a gaussian with $\sigma_W = 0.3$.
Optimisation algorithm for $M(\mathbf{w})$: variable metric methods, using code from [54], used several times in sequence with values of the fractional tolerance decreasing from 10^{-4} to 10^{-8}. Every other loop, the regularising constants α_c were allowed to adapt in accordance with the re–estimation formula:

$$\alpha_c := \gamma_c / 2E_W^c \tag{3.18}$$

Precaution

When evaluating the evidence, care must be taken to verify that the permutation term g is appropriately set. It may be the case (probably mainly in toy problems) that the regulariser makes two or more hidden units in a network adopt identical connection values; alternatively some hidden units might switch off, with all weights set to zero; in these cases the permutation term should be smaller. Also in these cases, it is likely that the quadratic approximation will perform badly (quartic rather than quadratic minima are likely), so it is preferable to automate the deletion of such redundant units.

demonstrations show that careful objective choice of regulariser and α is essential for the best solutions to be obtained.

Worst case analysis has a complementary role alongside Bayesian methods. Neither can substitute for the other.

Future tasks

Further work is needed to formalise the relationship of this framework to the pragmatic model comparison technique of cross–validation. Moody's work on 'generalised prediction error' (GPE) is an interesting contribution in this direction [45]. His sampling theory approach predicts that the generalisation error, in χ^2 units, will be $\frac{1}{N}(\chi^2_D + 2\gamma)$. However, I have evaluated the GPE for the interpolation models in this chapter's demonstration, and found the correlation between GPE and the actual test error was poor. More work is needed to understand this.

The gaussian approximation used to evaluate the evidence breaks down when the number of data points is small compared to the number of parameters. For the model problems I have studied so far, the gaussian approximation seemed to break down significantly for $N/k < 3 \pm 1$. It is a matter for further research to characterise this failure and investigate techniques for improving the evaluation of the integral Z^*_M, for example the use of random walks on M in the neighbourhood of a solution.

It is expected that evaluation of the evidence should provide an objective rule for deciding whether a network pruning or growing procedure should be stopped, but a careful study of this idea has yet to be performed.

It will be interesting to see the results of evaluating the evidence for networks applied to larger real–world problems.

Appendix: Numerical methods

Quick and dirty version

The three numerical tasks are automatic optimisation of α_c and β, calculation of error bars, and evaluation of the evidence. I will describe a cheap approximation for solving the first of these tasks without evaluating the Hessian. If we neglect the distinction between well–determined and poorly–determined parameters, we obtain the following update rules for α and β:

$$\begin{aligned} \alpha_c &:= k_c/2E^c_W \\ \beta &:= N/2E_D \end{aligned}$$

If you want an easy–to–program taste of what a Bayesian framework can offer, try using this procedure to update your decay terms.

Hessian evaluation

The Hessian of M, $\mathbf{A}$, is needed to evaluate γ (which relates to Trace $\mathbf{A^{-1}}$), to evaluate the evidence (which relates to det $\mathbf{A}$), and to assign error bars to network outputs (using $\mathbf{A^{-1}}$).

I used two methods for evaluating $\mathbf{A}$: a) an approximate analytic method and b) second differences. The approximate analytic method was, following Denker *et. al.*, to use backprop

3.5 Discussion

The Bayesian method that has been presented is well–founded theoretically, and it works practically, though it remains to be seen how this approach will scale to larger problems. For a particular data set, the evaluation of the **evidence** has led us objectively from an inconsistent regulariser to a more probable one. The evidence is maximised by networks which generalise best, showing that Occam's razor has been successfully embodied with no ad hoc terms. Furthermore the solutions with greatest evidence perform better on a test set than any other solutions found. I believe there is currently no other technique that could reliably find and identify better solutions using only the training set. Essential to this success was the simultaneous Bayesian optimisation of the three regularising constants (decay terms) α_c. Optimisation of these parameters by any orthodox search technique such as cross–validation would be laborious; if there were many more than three regularising constants, as could easily be the case in larger problems, it is hard to imagine any such search being possible.[5]

This brings up the question of how these Bayesian calculations scale with problem size. In terms of the number of parameters k, calculation of the determinant and inverse of the Hessian scales as k^3. Note that this is a computation that needs to be carried out only a small number of times compared with the immense number of derivative calculations involved in a typical learning session. However, for large problems it may be too demanding to evaluate the determinant of the Hessian. If this is the case, numerical methods are available to approximate the determinant or trace of a matrix in k^2 time [62].

Application to classification problems

This chapter has thus far discussed the evaluation of the evidence for backprop networks trained on interpolation problems. Neural networks can also be trained to perform classification tasks. A future publication [43] will demonstrate that the Bayesian framework for model comparison can be applied to these problems too.

Relation to V–C dimension

Some papers advocate the use of V–C dimension [1] as a criterion for penalising over–complex models [2, 35]. V–C dimension is most often applied to classification problems; the evidence, on the other hand, can be evaluated equally easily for both interpolation and classification problems. V–C dimension is a worst case measure, so it yields different results from Bayesian analysis [26]. For example, V–C dimension is indifferent to the use of regularisers like (3.2), and to the value of α, because the use of such regularisers does not rule out absolutely any particular network parameters. Thus V–C dimension assigns the same complexity to a model whether or not it is regularised.[6] So it cannot be used to set regularising constants α or to compare alternative regularisers. In contrast, the preceding

[5] Radford Neal (personal communication) has pointed out that it is possible to evaluate the gradient of a validation error with respect to parameters such as $\{\alpha_c\}$, using $\partial E_{\rm val}/\partial \alpha_c = \partial E_{\rm val}/\partial \mathbf{w}_{\rm MP} \cdot \partial \mathbf{w}_{\rm MP}/\partial \alpha_c$. The first quantity could be evaluated by backprop, and the second term could be found within the quadratic approximation which gives $\partial \mathbf{w}_{\rm MP}/\partial \alpha_c = \mathbf{A}^{-1}\mathbf{I}_c\mathbf{w}_{\rm MP}$, where $\mathbf{I}_c$ is the identity matrix for the weights regularised by α_c, and zero elsewhere. Alternatively, Radford Neal has suggested that the gradients $\partial E_{\rm val}/\partial \alpha_c$ could be more efficiently calculated using 'recurrent backpropagation' [52], viewing $\mathbf{w}$ as the vector of activities of a recurrent network, and $\mathbf{w}_{\rm MP}$ as the fixed point whose error $E_{\rm val}$ we wish to minimise.

[6] However, E. Levin (personal communication) has mentioned that a measure of 'effective VC dimension' of a regularised model is being developed. It is speculated that this measure may be identical to γ, equation (3.12).

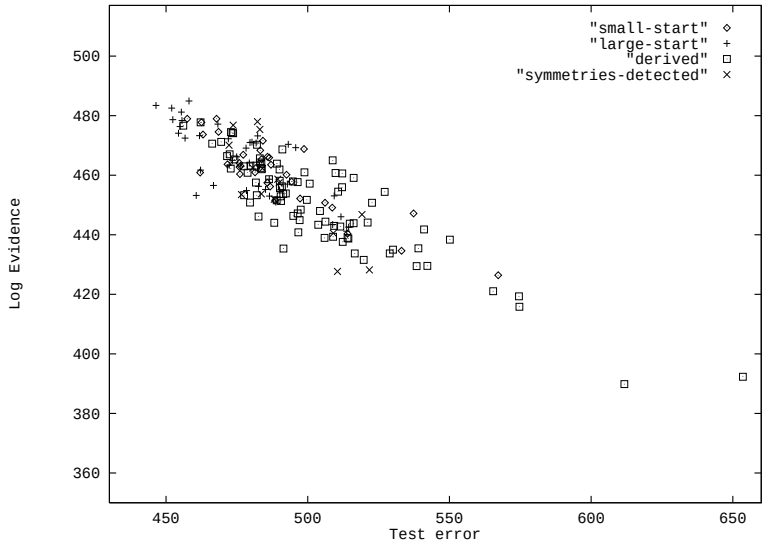

Figure 3.12: **Log Evidence for the second prior versus test error.**
The correlation between the evidence and the test error for the second prior is very good. Note that the largest value of evidence has increased relative to figure 3.8, and the smallest test error has also decreased.

a similar prior modelling weights as coming from a gaussian mixture, and using Bayesian re–estimation techniques to update the mixture parameters; they found such a model was good at discovering elegant solutions to problems with translation invariances. This model also achieves better performance on the task of sunspot time series prediction than any published model [48].

Using the second prior, each regularising constant is independently adapted to its most probable value by evaluating the number of well–measured parameters γ_c associated with each regularising function, and finding the optimum where $2\alpha_c E_W^c = \gamma_c$. The increased complexity of this prior model is penalised by an Occam factor for each new parameter α_c (see chapter 2). Let me preempt questions along the lines of 'why didn't you use four weight classes, or non–zero means?' — any other way of assigning weight decays is just another model, and you can try as many as you like; by evaluating the evidence you can then find out what preference the data have for the alternative decay schemes.

New solutions have been found using this second prior, and the evidence evaluated. The evidence for these new solutions with the new prior is shown in figure 3.11. Notice that the evidence has increased compared to the evidence for the first prior. For some solutions the new prior is more probable by a factor of 10^{30}.

Now the crunch: does this more probable model make good predictions? The evidence for the second prior is shown against the test error in figure 3.12. The correlation between the two is greatly improved. Notice furthermore that not only is the second prior more probable, the best test error achieved by solutions found using the second prior is slightly better than any achieved using the first prior, and the number of good solutions has increased substantially. Thus the Bayesian evidence is a good predictor of generalisation ability, and the Bayesian choice of regularisers has enabled the best solutions to be found.

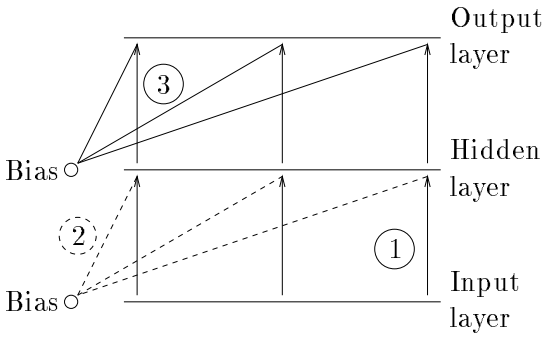

Figure 3.10: **The three classes of weights under the second prior**
1: Hidden unit weights. 2: Hidden unit biases. 3: Output unit weights and biases. The weights in one class c share the same decay constant α_c.

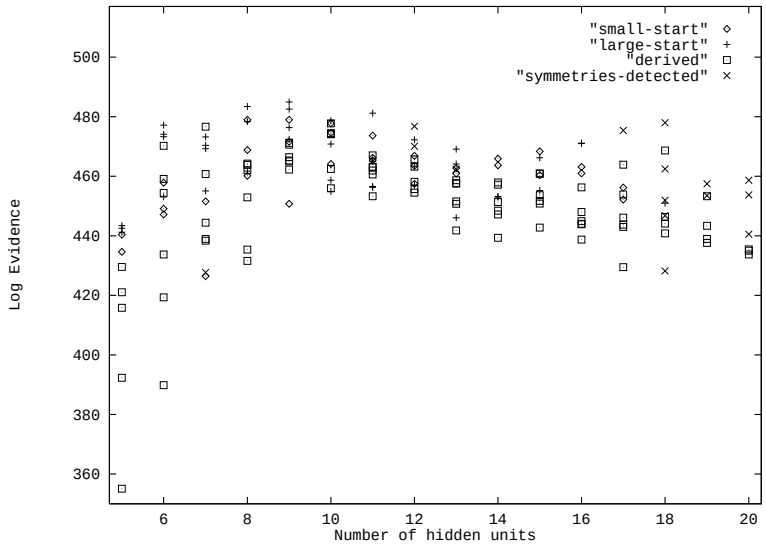

Figure 3.11: **Log Evidence versus number of hidden units for the second prior**
The different point styles correspond to networks with learning initialised with small and large values of σ_W; networks previously trained using the first regulariser and subsequently trained on the second regulariser; and networks in which a weight symmetry was detected (in such cases the evidence evaluation is possibly less reliable).

their two most probable interpolants are completely identical. In this case, the generalisation error for the two solutions must be the same, but the evidence will not in general be the same: typically, the model that was *a priori* more complex will suffer a larger Occam factor and will have smaller evidence. Also, the evidence is a measure of plausibility of the whole ensemble of networks about the optimum, not just the optimal network. Thus there is more to the evidence than there is to the generalisation error.

What if the Bayesian method fails?

I do not want to dismiss the utility of the generalisation error: it can be important for detecting failures of the model being used. For example, if we obtain a poor correlation between the evidence and the generalisation error, such that Bayes fails to assign a strong preference to solutions which actually perform well on test data, then we are able to detect and attempt to correct such failures.

A failure indicates one of two things, and in either case we are able to learn and improve: either numerical inaccuracies in the evaluation of the probabilities caused the failure; or else the alternative models which were offered to Bayes were a poor selection, ill–matched to the real world (for example, using inappropriate regularisers). When such a failure is detected, it prompts us to examine our models and try to discover the implicit assumptions in the model which the data didn't agree with; alternative models can be tried until one is found that makes the data more probable.

We have just met exactly such a failure. Let us now establish what assumption in our model caused this failure and *learn* from it. Note that this mechanism for human learning is not available to those who just use the test error as their performance criterion. Going by the test error alone, there would have been no indication that there was a serious mismatch between the model and the data.

Back to the demonstration: comparing different regularisers

The demonstrations thus far used the regulariser (3.2). This is equivalent to a prior that expects all the weights to have the same characteristic size. This is actually an inconsistent prior: the input and output variables and hidden unit activities could all be arbitrarily rescaled; if the same mapping is to be performed (a simple consistency requirement), such transformations of the variables would imply *independent* rescaling of the weights to the hidden layer and to the output layer. Thus the scales of the two layers of weights are unrelated, and it is inconsistent to force the characteristic decay rates of these different classes of weights to be the same. This inconsistency is the major cause of the failure illustrated in figure 3.8. *All the networks deviating substantially from the desired trend have weights to the output layer far larger than the weights to the input layer;* this poor match to the model implicit in the regulariser causes the evidence for those solutions to be small.

This failure enables us to progress with insight to new regularisers. The alternative that I now present is a prior which is not inconsistent in the way explained above, so there are theoretical reasons to expect it to be 'better.' However, we will allow the data to choose, by evaluating the evidence for solutions using the new prior; we will find that the new prior is indeed *more probable*.

The second prior has three independent regularising constants, corresponding to the characteristic magnitudes of the weights in three different classes c, namely hidden unit weights, hidden unit biases, and output weights and biases (see figure 3.10). The term αE_W is replaced by $\sum_c \alpha_c E_W^c$, where $E_W^c = \sum_{i \in c} w_i^2/2$. Hinton and Nowlan [47] have used

neural nets used had one hidden layer of sigmoid units and linear output units. During optimisation, the regulariser (3.2) was used initially, and an alternative regulariser was introduced later; β was fixed to its true value (to enable demonstration of the properties of the quantity γ), and α was allowed to adapt to its locally most probable value.

Figure 3.1 illustrates the performance of a typical neural network trained in this way. Each output is accompanied by error bars evaluated using Denker *et. al.*'s method, *including off–diagonal Hessian terms.* If β had not been known in advance, it could have been inferred from the data using equation (3.11). For the solution displayed, the model's estimate of β in fact differed negligibly from the true value, so the displayed error bars are the same as if β had been inferred from the data.

Figure 3.2 shows the data misfit versus the number of hidden units. Notice that, as expected, the data error tends to decrease monotonically with increasing number of parameters. Figure 3.3 shows the error of these same solutions on an unseen test set, which does not show the same trend as the data error. This Occam problem is illustrated by figure 3.4, which compares the test error with the data error. The data misfit cannot serve as a criterion for choosing between solutions.

Figure 3.5 shows the evidence for about 100 different solutions using different numbers of hidden units. Notice how the evidence maximum has the characteristic shape of an 'Occam hill' — steep on the side with too few parameters, and shallow on the side with too many parameters. The quadratic approximations break down when the number of parameters becomes too big compared with the number of data points.

The next figures introduce the quantity γ, discussed in chapter 2, the number of well–measured parameters. In cases where the evaluation of the evidence proves difficult, it may be that γ will serve as a useful tool. For example, frequentist theory predicts that the addition of redundant parameters to a model should reduce χ^2_D by one unit per well–measured parameter; a stopping criterion could detect the point at which, as parameters are deleted, χ^2_D started to increase faster than with gradient 1 with decreasing γ (figure 3.7).[4] This use of γ requires prior knowledge of the noise level β; that is why β was fixed to its known value for these demonstrations.

Now the question is how good a predictor of network quality the evidence is. The fact that the evidence has a maximum at a reasonable number of hidden units is promising. A comparison with figure 3.3 shows that the performance of the solutions on an unseen test set has similar overall structure to the evidence. However, figure 3.8 shows the evidence against the performance on a test set, and it can be seen that a significant number of solutions with poor evidence actually perform well on the test set. Something is wrong! Let us discuss the relationship between the evidence and generalisation ability. We will then return to the failure in figure 3.8 and see that it is rectified by the development of new, more probable regularisers.

Relation to 'generalisation error'

What is the relationship between the evidence and the generalisation error (or its close relative, cross–validation)? A correlation between the two is certainly expected. But the evidence is not necessarily a good predictor of generalisation error (see discussion in chapter 2). First, as illustrated in figure 3.9, the error on a test set is a noisy quantity, and a lot of data has to be devoted to the test set to get an acceptable signal to noise ratio. Furthermore, imagine that two models have generated solutions to an interpolation problem, and that

[4] This suggestion is closely related to Moody's 'generalised prediction error', GPE $= \frac{1}{N}(\chi^2_D + 2\gamma)$ [45].

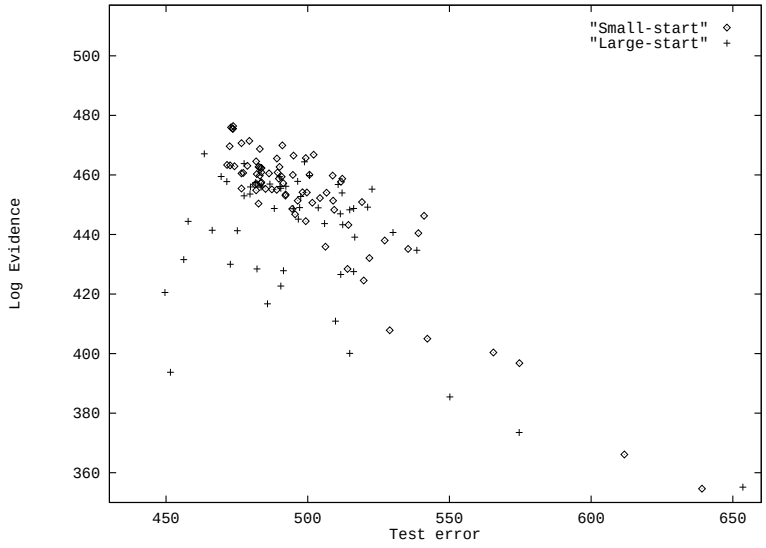

Figure 3.8: **Log Evidence versus Test error for the first regulariser**
The desired correlation between the evidence and the test error has negative slope. A significant number of points on the lower left violate this desired trend, so we have a failure of Bayesian prediction. The points which violate the trend are networks in which there is a significant difference in typical weight magnitude between the two layers. They are all networks whose learning was initialised with a large value of σ_W. The first regulariser is ill–matched to such networks, and the low evidence is a reflection of this poor prior hypothesis.

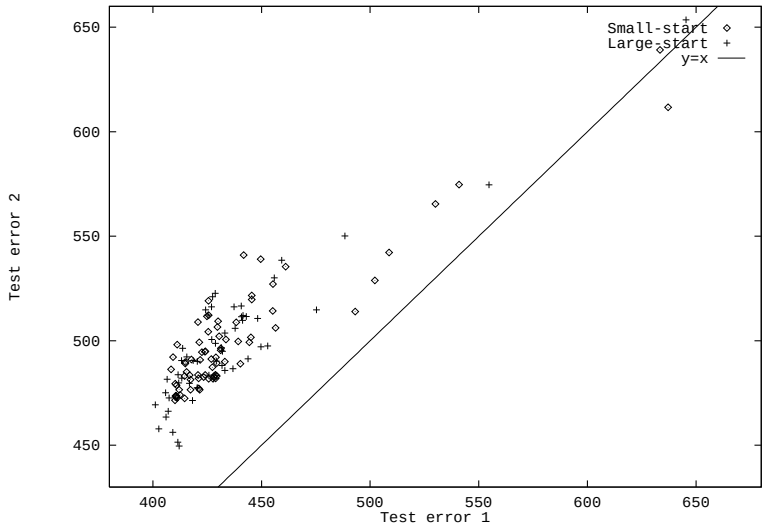

Figure 3.9: **Comparison of two test errors.**
This figure illustrates how noisy a performance measure the test error is. Each point compares the error of a trained network on two different test sets. Both test sets consist of 200 data points from the same distribution as the training set.

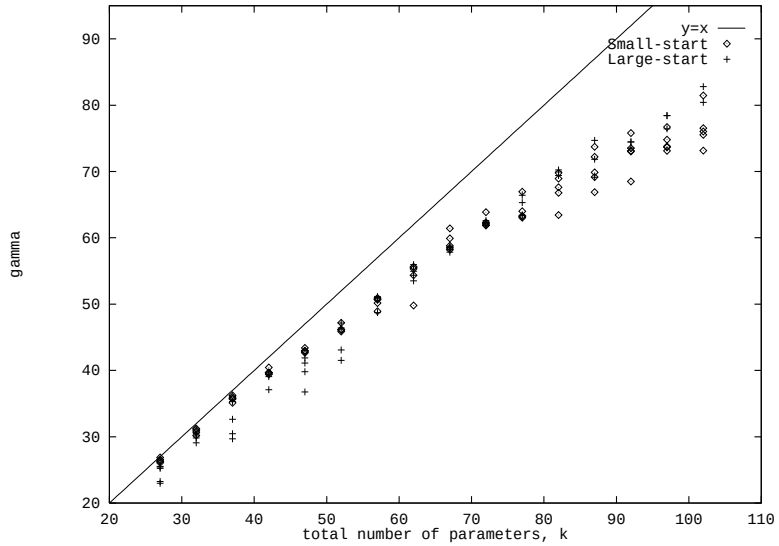

Figure 3.6: **The number of well–determined parameters.**
This figure displays γ as a function of k, for the same network solutions as in figure 3.5.

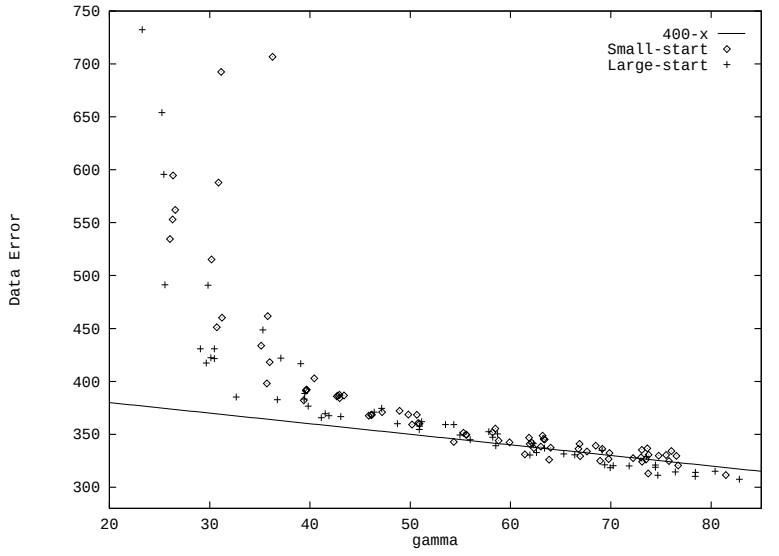

Figure 3.7: **Data misfit versus γ.**
This figure shows χ^2_D against γ, and a line of gradient -1. Towards the right, the data's misfit χ^2_D is reduced by 1 for every well–measured parameter. When the model has too few parameters however (towards the left), the misfit gets worse at a greater rate.