The Statistical Mechanics of Bayesian Model Selection

Glenn Marion

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the
University of Edinburgh
1996

EX BIBL UNIV. EDINBURGEN.

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by me, unless otherwise stated.

Acknowledgements

I am deeply indebted to my supervisors. To David Saad, without whom this thesis would be much the poorer and whose unflagging support, commitment and enthusiasm are an inspiration. To David Wallace for many ideas and much appreciated encouragement which set me on the right course and to David Willshaw for helpful advice and his alternative perspective. Thank you David, David and David!

I have also learned a great deal from many others and particularly wish to thank Alastair Bruce, Peter Sollich, Ronny Meir, David Wolpert, David Mackay, Chris Bishop, Ton Coolen, and Micheal Kearns. I would like to gratefully acknowledge the studentship provided by the Engineering and Physical Sciences Research Council. My work has also greatly benefited from attending a number of conferences made possible by grants from the Neural Information Processing Systems Foundation and the European Union (grant No. ERB CHRX-CT92-0063).

That I have thoroughly enjoyed my time in neural networks is due in no small part to the other members of the group. Collectively, they created both an academically and socially lively atmosphere. I owe a great debt to Pete whose insight and calming influence saw me through many a mathematical crisis! Many thanks are also due to my other office mates; David for his puns and many enjoyable runs, Ansgar for *enforced* tea breaks, Alan for skiing and table footie, Jason for philosophy, David Harris who took the David density to implausible heights, to Fortunato for his 'mister' description of the group and to Martin, Jonathan and Hon for getting me settled in. I wish them all happiness and good luck for the future. Good luck also, to the new incumbents of rm 4408, David and Chris.

I would also like to thank all my friends for making the last few years such great fun. Special mentions go to Ruth and Simon (for reading chapter 1), Angus, Pete, Eirian, everyone at 42, Yta for picking on me, Anna R for geese and chocolate cake, the 78 mob (wherever they are) everyone at *VTO* (keep up the good work!), Claire, Bobby (for all the quick breaks) and anyone I forgot to mention!

More than thanks are due to my parents and Alex for all their support and encouragement over the years. Finally, the biggest THANK YOU! to you, Anna, for all your love and for always being there.

Abstract

In this thesis we examine the question of model selection in systems which learn input-output mappings from a data set of examples. The models we consider are inspired by feed-forward architectures used within the artificial neural networks community. The approach taken here is to elucidate the properties of various *model selection* criteria by calculation of relevant quantities derived in a Bayesian framework. These calculations make the assumption that examples are generated from some underlying rule or *teacher* by randomly sampling the input space and are performed using techniques borrowed from statistical mechanics. Such an approach allows for the comparison of different approaches on the basis of the resultant ability of the system to *generalize* to novel examples. Broadly stated, the model selection problem is the following. Given only a limited set of examples, which model, or *student*, should one choose from a set of candidates in order to achieve the highest level of generalization? We consider four model selection criteria. A penalty based method utilising a quantity derived from Bayesian statistics termed the *evidence*, and two methods based on estimates of the generalization performance namely, the *test error* and the *cross validation error*. The fourth method, less widely used, is based on the *noise sensitivity* of the models.

In a simple scenario we demonstrate that model selection based on the evidence is susceptible to misspecification of the student. Our analysis is conducted in the *thermodynamic limit* where the system size is taken to be arbitrarily large. In particular we examine the *evidence procedure* assignments of the *hyper-parameters* which control the learning algorithm. We find that, where the student is not sufficiently powerful to fully model the teacher, despite being sub-optimal this procedure is remarkably robust towards such misspecifications. In a scenario in which the student is more than able to represent the teacher we find the evidence procedure is optimal.

In the learnable linear setting we explore the relevance of the thermodynamic limit to real systems through the calculation of *finite size corrections* which reveal a rich behaviour. In particular, we focus on the hyper-parameter known as the *weight decay* showing that in finite sized systems the evidence assignment is *inconsistent* that is, it

is not optimal even in the limit of large data sets. Consideration of model selection based on the test and cross validated errors in finite sized systems shows that they too are inconsistent in terms of weight decay assignment. The performance resulting from the test error is shown to be an order of magnitude worse than that associated with the evidence. However, whilst of the same order the performance resulting from leave one out cross validation is shown to be generally superior to that associated with the evidence.

Finally, returning to the thermodynamic limit we explore the problem of architecture selection for a simple model. We find that the evidence, cross validation and the noise sensitivity can be used to select the number of segments of a piece-wise linear student learning a linear teacher at least when the models are appropriately regularized.

Publications

Some of the material in this thesis has been published, or is to be submitted for publication as follows.

Marion G and Saad D 1995 A statistical mechanical analysis of a Bayesian inference scheme for an unrealisable rule. *J. of Phys A: Math. Gen.* **28**:2159-2171.

Marion G and Saad D 1996 Finite size effects in Bayesian model selection and generalization. *J. of Phys A: Math. Gen.* **29**:5387-5404.

Marion G and Saad D 1996 The statistical mechanics of cross-validation. *In preparation.*

Marion G and Saad D 1995 Hyper-parameters, evidence and generalization for an unrealisable rule. In *Advances in Neural Information Processing Systems* **7**:232-Tesauro G, Touretzky D S and Leen T K (Eds.). Cambridge, Massachusetts: The MIT Press.

Marion G and Saad D 1995 Data dependent hyper-parameter assignment. *Annals of Mathematics and Artificial Intelligence.* In press.

Contents

Chapter 1

Introduction

1.1 Background

The history of science has witnessed the proposal, testing and rejection of competing models to account for various physical phenomena. In rare cases, for a time, the available data supports one of the suggested models and no alternatives are sought, until new data reveal inadequacies in the chosen model. Models are sometimes based on deep theoretical insights, whilst others are based on observation; the so called empirical laws. Often, empirical models have proved a useful spring board to deeper theoretical understanding. For example, the empirical laws of Kepler were at least a vindication of the revolutionary ideas of Newton. In recent times, due to the increasing availability of affordable computing power, remote sensing and data logging techniques, vast quantities of data have been produced and empirical models have come into their own.

In this thesis we examine empirical modelling in a limited sense. In particular, the work presented here is grounded in the ideas, literature and ethos of the artificial neural networks community. Broadly speaking, the approach we take can be stated as follows. Through the examination of some simple cases we seek a clear understanding of some of the issues involved in the construction and comparison of empirical models inspired by the neural network paradigm. In particular we aim, where possible, to shed light on the issue of *model selection* through exact calculations of relevant quantities. In general these quantities are derived from Bayesian statistics and the calculations performed using tools from statistical mechanics.

In the following pages we introduce the type of neural network models we will be concerned with, describing how they can learn from examples and introducing some terminology along the way. We briefly review the recent developments in the field of artificial neural networks mentioning some notable successes in the use of these networks

as applied to real problems. We then introduce the central concepts of *generalization* and *model selection*, the study of which will occupy us throughout this thesis. In section 1.2 we will introduce the Bayesian formalism within which we conduct our study and review some of the existing approaches to the analysis of these problems, with particular reference to the statistical mechanical approach. Finally, in section 1.3 we will outline the remainder of the thesis. However, firstly we introduce some preliminary concepts surrounding the use of artificial neural networks as empirical models.

When constructing empirical models of a given system or process one has essentially two choices. Firstly, one can attempt to utilise the limited, problem specific, mechanistic (theoretical) understanding available. The resulting model is then completed through incorporation of the experimental (empirical) data to hand. Such models are also referred to as parametric. This, semi-mechanistic, approach can be regarded as a half way house between theoretical models and the fully empirical ones which are the focus of this thesis. In this latter approach the original model can be regarded as essentially *unstructured* or non-parametric. On presentation of the data we must then, to some extent, mould the model to fit the data. Nonetheless, as we shall see the incorporation of some information of a general nature, into the models *prior* structure is crucial to the success of this endeavour. Thus, the semi-mechanistic and the unstructured modelling approaches are subtly linked.

One of the principal motivations behind the scientific endeavour, and thus the modelling enterprise, is the desire to make predictions. When constructing empirical models we accept that our understanding of the process being modelled is limited but, nevertheless hope to make the best possible predictions or inferences in light of the available data. In the neural networks community this process is known as *learning from examples*. As the name suggests work in this field is based on the notion that human beings, and indeed other animals are able to learn from experience. A child learns to speak through experimenting and listening to others, a lion cub learns to hunt through trial and error. Neither is born with either ability but learns through experience. Despite great efforts, humanity has not yet been able to mimic these natural learning systems using conventional computational approaches and hence the idea behind artificial neural networks is to borrow from nature's elegant solutions. Thus, we construct models based on a network of simple neuron-like elements connected by synaptic-like weights and adapt these connections in the light of experience, that is mould the model to the data. Indeed, in the attempt to recreate some rudimentary features of natural learning systems, this approach has proved useful and, as we shall see, its novelty has lead to many developments in recent years. However, it should be stressed that the models examined in this thesis bear only passing resemblance to real

neural systems.

Learning from examples

The problem of learning from examples is and has been studied in many disciplines. For example, in statistics where it is variously known as statistical inference, or non-parametric inference (in the case of neural network type models), regression, interpolation and classification (see, for example, Ripely (92) on the relation between neural networks and statistics). In the field of artificial intelligence, one refers to machine learning (*e.g.* Valiant (84)), whilst the problem has also been studied in the areas of speech recognition and image restoration (*e.g.* Geman and Geman (84)). Indeed, in comparison to these techniques, neural networks are relative new comers and should be considered as one of many possible approaches to the problem at hand. In general, the concept of learning from examples can be applied to many diverse problems but, in this thesis we focus on the problem of learning a rule which is an input-output mapping. In this problem we are supplied with a set of data by a friendly experimentalist who has gathered the data in one, or a series of experiments. In general, we note that the data will not be wholly reliable and we must take this into account in the modelling process. The observations are in the form of data pairs consisting of a set of *attributes* (inputs) and a set of *properties* (outputs) of the particular system under scrutiny. Our task is to construct an empirical model using this data which allows us to predict future values of the system properties from the attributes. That is, we wish to infer the rule relating the two; the mapping from the attributes to the properties, the inputs to the outputs. If no such rule exists, then we are wasting our time, but experience and indeed the history of science, shows that very often it does. If we are successful, in future when we measure the system attributes we will be able to predict the associated properties. This assumes that the relationship between attributes and properties does not change with time, or at least varies slowly, an assumption we will make throughout this thesis.

The more general case of prediction in, so-called, non-autonomous systems is however, a growing field tackled in the artificial neural networks community by *recurrent nets* (see *e.g.* Williams and Zipser (89)). Nonetheless, the case studied here does not preclude application to dynamical systems where the attributes could represent the system at some time, t, and the properties represent the system after some fixed interval, δt; only the mapping between states at time t and $t + \delta t$ must remain constant. In this latter case the time interval, δt, will clearly affect the mapping of the attributes to the properties but could also be included as an attribute. To summarise then, in this thesis we will be concerned with the problem of learning an input-output mapping which

does not vary with time. Furthermore, here we focus on the narrow definition where the system attributes and properties take on, or can be coded in terms of, numerical values.

Mappings such as these can be realised by feed forward neural networks known as Multi-Layer Perceptrons, **MLPs** for short. Figure 1.1 shows a schematic diagram of an **MLP**, in which the information feeds-forward (up the page), from the input layer, through the hidden layers (only one being shown) to the output. A simple perceptron has no hidden layer. The hidden layers themselves consist of a number of *hidden units* which sum the inputs received from the preceding layer. This sum, known as the activation and denoted h in figure 1.1, is then passed through a *transfer function*, f, before being passed on to the next layer. Thus, each layer feeds forward its output to the next layer. The connections between each node are known as *weights*, the larger the weight the more influence a given input will have, with a weight of zero equivalent to disconnection. It is through adapting these weights that we alter the mapping represented by a given perceptron. In the context studied here a learning rule, or training algorithm, is a process by which, given a set of examples, one can modify the weights in the *student* network so as to better model the data supplied by a *teacher* (*i.e.* the underlying process generating the data). That is we mould or *fit* the model to the data. This is also referred to as supervised learning and, broadly speaking, training algorithms within this paradigm fall into two categories. In *batch* learning the student trains on all the available examples concurrently whilst in *on-line* learning one example is presented at a time. The potential of multi-layer perceptrons was demonstrated by, Cybenko (89) and Hornik *et al.*(89) who showed that **MLPs** using sigmoidal transfer functions were *Universal Approximators.* That is, given only one hidden layer, with sufficient hidden units and appropriate weights, they can represent any mapping of the inputs (attributes) to the outputs (properties) from very wide class of functions. Thus, **MLPs** have found a very broad range of applications. For example, we may wish to predict the possibility of a patient developing a particular malady given the outcome of certain tests, the price of oil in six months given key economic indicators, or the temperature in a blast furnace based on the running parameters.

History

Interest in artificial neural networks as an alternative computational paradigm dates back to the mid-forties (McCulloch and Pitts (43)). However, the current resurgence resulted from two developments in the early to mid 1980s. A thorough account of this history is to be found in the excellent introductory text by Hertz *et al.*(91) and we

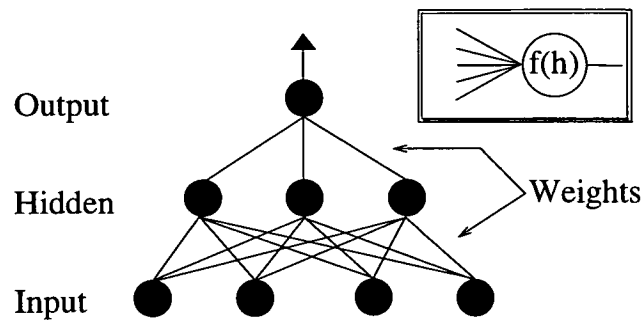


Figure 1.1. Schematic of a two layer Multi-layer Perceptron (**MLP**): The flow of information passes up the page, from the input layer through the hidden layer. The inset shows a close up of nodes in the hidden and output layer, the output of each being a function, f, known as the transfer function, of the activations, h. The activations are weighted sums of the inputs to each node. It is adaptation of these weights that allows us to modify the network during training.

will only briefly discuss these developments here. The first of these breakthroughs did not relate directly to feed forward networks but to symmetric networks in which, for each connection, from one node to another, there exists a connection in the opposite direction. In relating the behaviour of such networks to an energy function Hopfield (82) stirred the interest of statistical physicists leading to a substantial body of work in which this thesis is grounded. We will discuss some of the more relevant contributions later. The second important development was, in fact, the (re)-discovery of a learning rule, known as back-propagation, for general feed forward networks by Rumelhart *et al.*(86). This was perhaps the more significant discovery for it led to the successful application of such networks to a number of real problems.

One of the most most famous of these applications was the NETtalk project (Senjowski and Rosenberg (87)). This project trained an **MLP**, from a data base of examples, to recode written text into the appropriate phonemes which could then be 'spoken' by a speech synthesiser. In fact, this network learned the task to a reasonable degree but does not perform as well as the commercially available package, DECtalk, which is based on linguistic rules painstakingly elucidated over many years. However, in comparison, NETtalk achieves remarkable performance for the relatively small effort it required. We can perhaps begin to see why neural networks have been termed the *second best way to do anything*; the best way involving a detailed and perhaps elusive

understanding of the system being studied. Amongst, other applications of feed forward neural networks are the recognition of hand written ZIP codes (US numerical postal codes) (Le Cun *et al.*(89)), voice recognition (Lippmann (89)) and steering a car (Pomerleau (89)). A common feature to all these applications is that no principled method exists to solve the problem. That is, we can not write down a set of mathematically rigorous rules which would solve them (*i.e.* drive a car), at least not in generality. Thus, a useful approach has been to attempt to infer or learn underlying rules from examples.

We briefly note here that this is not the only area in which artificial neural networks can be useful. Bishop *et al.*(95) successfully used an **MLP** to implement a well understood rule. In this case there was a need to control the magnetic flux inside an experimental fusion chamber. The physics of this problem is well understood but conventional computational approaches could not provide answers on the small time scales required by the experiment. In Bishop *et al.*(95) an **MLP** was able to learn the rule required to control the flux from examples generated by conventional simulation. Then, by implementing this network in hardware it was used to control the experiment in real time. This application demonstrates that the parallel architecture of artificial neural networks can lead to significant increases in speed as compared to conventional serial approaches. Other advantages often claimed for the neural network paradigm include robustness to noise and a graceful degradation in performance as individual components fail (see *e.g.* Hertz *et al.*(91)). This latter quality contrasts strongly with the catastrophic failure of conventional computers in such circumstances. However, here we do not consider these issues further.

Generalization

The preceding discussion glossed over many of the complications involved in the implementation of neural networks. These include the learning rule, or training algorithm and the architecture of our student, that is, the number of hidden layers and hidden units, to be used. Linked to these issues is the crucial question of generalization. That is, given a data base of examples we want to train our student such that it will be able to generalize to situations not included in these examples, since it is of little benefit to simply reproduce the training examples themselves. Indeed, as Wolpert (92) has pointed out a simple look up table would suffice. We formally define measures of generalization performance later. In fact, in the examples discussed above the training algorithms and architectures used do critically affect the generalization ability. An intuitive understanding of why this is so can be gained from considering the common

problem of fitting a curve to a set of data points.

This is illustrated in figure 1.2 where the data points are shown by crosses. When these data points are unreliable, as in any real experiment they will be, we all *know* that drawing a curve passing through all the points is not the best thing to do in terms of predicting the *true* curve, the structure of which is generally supposed to be somewhat simpler. However, Wolpert (92) has noted there are no *apriori* reasons to reject the curve actually drawn, indeed, in some circumstances it might be the true curve. Nonetheless, experience shows that in the universe in which we live all possibilities are not equally likely. In fact, if they were it would not be possible to learn or generalize at all. Thus, in figure 1.2 we would choose a simpler curve because our prior conviction and experience point towards smoother curves. However, eventually, as we gathered more data points which supported the curve drawn, we would revise our opinion.

Thus, given that we want our trained model to generalize, the training process must involve a careful balance between fitting the model to the examples in our data base whilst remaining true to our prior beliefs. In other words, during training we seek to avoid *over-fitting* the data. In general, complex models with many adjustable parameters (our example curve may have been generated by such a model) will be able to fit the data points more closely than simpler models. In fact, the principle, known as *Occam's Razor*, that a simpler model is preferable to a more complex one, subject to the data, is widespread in science. The question of what constitutes an overly complex model is a difficult one and ultimately will depend on the underlying teacher one is trying to learn. Thus, problem specific knowledge (mechanistic understanding) should be employed where available. However, in this thesis we will be concerned with methods which attempt avoid over-fitting of the data in a more general sense and as we shall see the role of the training algorithm, in addition to the model architecture, plays a crucial role in this.

Model selection

In general, we will have a number of competing model architectures and training algorithms and could, thus, generate a number of students all of which would generalize in different ways, some poorly, others better. In practice we only have access to our data base of examples and we must decide how best to use this data, we can not cheat by looking at the answer! The key question we seek to answer in this thesis is which models should be chosen in order to obtain the best generalization performance (*i.e.* the best predictions). In other words how do we select the best model from amongst a

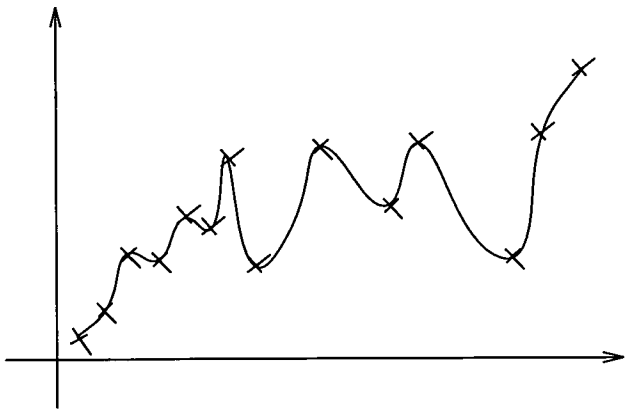


Figure 1.2. Fitting a curve: The diagram shows a set of noisy data points marked by the crosses. The curve drawn is a rather unlikely fit and demonstrates the danger of allowing ones model (student) too much flexibility

set of candidates. A number of different model selection methods have been proposed in the literature and here we will compare and contrast some of them, in particular with expressed regard towards the generalization performance we can expect from each of them. In the next section we will introduce the Bayesian formalism in which we will conduct our study, describe the model selection criteria we will examine and formally define the notion of generalization.

1.2 General framework

In this section we outline the general framework which we will use throughout this thesis, introducing some of the notation we will later use. This, section is thus, somewhat more formal than the preceding pages but hopefully, still of general interest. We high-light the areas on which we will focus, in later chapters, and briefly describe the connection between various approaches to the analysis of learning from examples, high-lighting work relevant to our study.

Data generation

Throughout this thesis we consider the case of modelling a mapping from an input space $\mathbf{x}$, of N dimensions to an output space $\mathbf{y}$. In fact, in all the examples we consider in subsequent chapters the output space is simply one dimensional and both inputs and outputs are real. Nonetheless, until otherwise stated, the formalism can be applied in the more general case.

In order to perform our analyses we must make some assumptions about the way the

data is generated. Here we assume that the outputs, for a given input, are generated by an unknown teacher mapping described by the distribution $P(\mathbf{y}_t \mid \mathbf{x})$. We shall continue to use this notation, where the arguments denote the distribution of $\mathbf{y}_t$ given $\mathbf{x}$. This assumption accommodates, for example, deterministic teachers whose output is corrupted by noise. Implicit in this assumption is that the unreliability of the data can be modelled as a random process. Furthermore, we assume that during the data collection the input space is sampled with probability $P(\mathbf{x})$. We describe this as our sampling assumption. Following its collection we have a data set, $\mathcal{D}$, available consisting of p input-output pairs $(\mathbf{y}_t(\mathbf{x}), \mathbf{x})$, drawn independently from $(P(\mathbf{y}_t \mid \mathbf{x})P(\mathbf{x}), P(\mathbf{x}))$. The data pairs are said to be independently but identically distributed (*i.i.d*). Thus, the data set or data base, which is also referred to as the training or learning set, is denoted, $\mathcal{D} = \{(\mathbf{y}_t(\mathbf{x}^\mu), \mathbf{x}^\mu) : \mu = 1..p\}$. Moreover, we also assume that after training, when we measure the resulting generalization ability, the test data will be drawn from this same distribution as the training data. Thus, the sampling assumption remains constant, in addition to the constancy of the mapping, $P(\mathbf{y}_t \mid \mathbf{x})$, discussed earlier.

Finally, we remark that the methods of experimental design, query learning or active data selection offer alternatives to this randomised data collection procedure. For reviews of this approach see Plutowski (94) and Sollich (95b). Indeed, it is intuitively clear that random examples contain redundant information as the possibility exists of selecting the same example twice. In general, the methods of experimental design listed above involve an active collection of data in an attempt optimise the useful information contained in the data set. That, is we design our experiment to provide us with the maximum information possible. Such an approach can be incorporated within the framework adopted here by adjustment of the sampling distribution, $P(\mathbf{x})$ (see *e.g.* Sollich (94b) and (95))

1.2.1 Bayesian modelling

A key feature of the Bayesian scheme for model construction and comparison, outlined below, is the assumption that the model under consideration is actually capable of generating the data. In other words that the model, or student, is powerful enough to mimic the data generation process, $P(y_t \mid \mathbf{x})$, *i.e.* the teacher. Nevertheless, in general the models being entertained will not coincide with the teacher. In the framework, adopted here we are free to consider scenarios in which this is, indeed, the case. We will speak of an unrealizable scenario when the student is unable to model the teacher fully, over realizable when it is more than capable of doing so and realisable or learnable when the student architecture matches that of the teacher. This will enable us to examine the

effect of a violation of the assumptions implicit in the Bayesian approach. In this sense, perhaps we should talk of a pseudo-Bayesian framework (Meir and Merhav (94)). We also note that this goes some way towards the extended Bayesian formalism of Wolpert (92). However, here we do not consider a distribution of teachers but, rather, a fixed teacher which the student may or may not be able to model.

Model selection using the evidence

As discussed earlier, when considering modelling the data, $\mathcal{D}$, provided to us we will often have a number of candidate models, $\mathcal{M}_i$, in mind. The notation, $\mathcal{M}_i$, is short hand for the complete specification of the model which will include the model architecture $\mathcal{A}_i$ and the learning algorithm $\mathcal{L}_i$ used for the i^{th} model. In general, the models or students we consider have a vector of adjustable parameters $\mathbf{w}_i$, these determine the output of the student, given the input, in a way dependent on the architecture, $\mathcal{A}_i$. Accordingly, when the architecture is linear, the model output, in one dimension, for an input x, is $\mathbf{y}_s(x) = w_1 x + w_0$. For an **MLP** the architecture is more complex as we must specify the number of hidden units and layers and the transfer function. As we have seen it is by tuning the parameters, $\mathbf{w}_i$, that we adapt the model to the data. The learning algorithm, $\mathcal{L}_i$, determines how this adjustment takes place, in other words how the student learns from the examples. The complete specification of model i is, then, denoted by $\mathcal{M}_i = \{\mathcal{A}_i, \mathcal{L}_i\}$. As we shall see below, in Bayesian terminology the learning algorithm $\mathcal{L}_i$ is defined by the the noise model and the prior distribution of model parameters.

Since in general the data will be corrupted by noise the Bayesian scheme incorporates a noise model which we write as the density,

$$P(\mathbf{y}_s \mid \mathbf{x}, \mathbf{w}, \mathcal{M}_i). \tag{1.1}$$

In isolation this gives the probability of the model $\mathcal{M}_i$, with parameters $\mathbf{w}_i$, generating any given output $\mathbf{y}_s$, given an input $\mathbf{x}$. When we have an observed data set, $\mathcal{D}$, available then we can write the probability of the model having generated the data as,

$$P(\mathcal{D} \mid \mathbf{w}_i, \mathcal{M}_i) \tag{1.2}$$

where, as we will continue to do, we have dropped the explicit dependence on the inputs $\mathbf{x}$ when we have more than one predicted value. Standard maximum likelihood estimation would have us infer the value of the adjustable parameters $\mathbf{w}_i$ by maximising the expression in equation (1.2) which is then referred to as a likelihood function.

However, as we shall see in section 2.2.1 this method of parameter estimation can lead to the problem of over-fitting discussed earlier.

It is for this reason that the Bayesian approach to model specification augments the noise model with a prior distribution (see *e.g.* Box (80), Mackay (92a), Gelfand and Dey (94)),

$$P(\mathbf{w}_i \mid \mathcal{M}_i). \tag{1.3}$$

This is, then, the *a priori* belief in a particular assignment of the model parameters $\mathbf{w}_i$. However, as we shall see this can often be interpreted as some kind of regularization procedure (Mackay 92a). Often, the latter interpretation is easier to comprehend than the somewhat abstract notion of a prior distribution over student parameters. In fact, the framework outlined here encompasses a rather broad range of approaches to the problems of avoiding the over fitting of data. This includes both the weight decay approach to training which we examine throughout this thesis and Rissanen's minimum description length principle (Rissanen (86)).

Combining the noise model and the prior gives us the full Bayesian specification of the model parameters, the, so called, posterior distribution of $\mathbf{w}$ conditioned on the data $\mathcal{D}$,

$$P(\mathbf{w}_i \mid \mathcal{D}, \mathcal{M}_i) \propto P(\mathcal{D} \mid \mathbf{w}_i, \mathcal{M}_i) P(\mathbf{w}_i \mid \mathcal{M}_i). \tag{1.4}$$

Thus, we see that, in the Bayesian framework, the noise model and the prior specification determine how the model parameters depend on the data, $\mathcal{D}$. Thus, they depend on the learning algorithm as is implied by their dependence on the model specification $\mathcal{M}_i = \{\mathcal{A}_i, \mathcal{L}_i\}$. We will see this relationship demonstrated in a more concrete fashion in chapter 2. Since the specification of the noise model and prior can often be difficult, and to an extent arbitrary, an important question which we also discus in chapter 2 is the sensitivity of the posterior with respect to variations in these prior specifications. Such questions are considered under the title *Bayesian robustness* (Berger (84)).

It is clear that whilst the posterior may be useful in determining the model parameters it can not be directly used to compare different models. This is because the parameters of two models with different architectures have no common interpretation. Thus, Box (80) advocates the use of the posterior to determine the model parameters whilst he argues that models themselves must be compared on the basis of the

predictions they make. Indeed, the predictive distribution of model $\mathcal{M}_i$ is given by,

$$P(\mathbf{y}_s \mid \mathbf{x}, \mathcal{M}_i) = \int d\mathbf{w}_i P(\mathbf{y}_s \mid \mathbf{x}, \mathbf{w}_i, \mathcal{M}_i) P(\mathbf{w}_i \mid \mathcal{M}_i) \tag{1.5}$$

The predictive density associated with a particular data set, $\mathcal{D}$, is then the missing normalisation constant in equation (1.4) namely,

$$P(\mathcal{D} \mid \mathcal{M}_i) = \int d\mathbf{w}_i P(\mathcal{D} \mid, \mathbf{w}_i, \mathcal{M}_i) P(\mathbf{w}_i \mid \mathcal{M}_i). \tag{1.6}$$

This has been dubbed the evidence by Gull (88) and Skilling (93), a practice adopted by Mackay (see *e.g.* Mackay (92a)) and continued here. In fact, this term refers to the evidence *for a given model* provided by the data, $\mathcal{D}$ and derives from the common practice (see *e.g.* Gelfand and Dey (94)) of assigning equal prior probabilities $P(\mathcal{M}_i)$ to the competing models $\mathcal{M}_i$, in which case $P(\mathcal{D} \mid \mathcal{M}_i) \propto P(\mathcal{M}_i \mid \mathcal{D})$. Indeed, then in pairwise comparison of models $i = 1, 2$ the celebrated Bayes' factor is the ratio of the evidence of model 1 to model 2. That is,

$$\mathbf{B}_{\mathcal{F}} = \frac{P(\mathcal{D} \mid \mathcal{M}_1)}{P(\mathcal{D} \mid \mathcal{M}_2)} \tag{1.7}$$

The Bayes' factor, $\mathbf{B}_{\mathcal{F}}$, thus provides us with our first model selection criterion; model $\mathcal{M}_1$ being preferred if $\mathbf{B}_{\mathcal{F}} > 1$ and model $\mathcal{M}_2$ otherwise. That is, the model with the larger evidence is preferred. Mackay (92a) demonstrated that if the data was, in fact, generated by one of the models under scrutiny then the Bayes factor, or to be precise $\ln \mathbf{B}_{\mathcal{F}}$, would on average not favour any other model over this true model. However, we note that in general this average over all possible data sets can be misleading and this issue is explored in some detail in chapter 4. Furthermore, it is very often the case that none of the entertained models coincides with the true teacher. In this latter case the model with the largest evidence need not be the best generalizer, a fact shown for an explicit example in chapter 2.

Once we have chosen a model we then wish to make predictions from it. One approach might be to take the maximum a posteriori, **MAP**, estimate, that is take the parameters that maximise the posterior and base our predictions on this. One advantage of such an approach is the reduction in computational cost gained by only having to deal with one student from the posterior. However, it has been noted that averaging over the posterior can actually improve performance (see *e.g.* Pryce and Bruce (95)). Indeed, intuitively, by taking only the maxima we are discarding potentially valuable information contained in the remainder of the distribution. In addition the **MAP** estimate is not invariate under non-linear re-parameterisation of the parameters $\mathbf{w}$ (see

Bishop (95)). Furthermore, as argued above, one can not make meaningful comparisons between models using the posterior, thus we focus on the predictive distribution as a means to make such comparisons.

The predictive distribution of equation (1.5) is, however, independent of the training data, $\mathcal{D}$. Clearly we seek to make predictions conditioned on the data set since this reflects the effect of the incorporation of the data into our model; the effect of learning from the examples in our data set. The predictive density for an output, $\mathbf{y}_s$, at $\mathbf{x}$ conditioned on the training data is then dependent on the noise model and the posterior,

$$P(\mathbf{y}_s \mid \mathbf{x}, \mathcal{D}, \mathcal{M}_i) = \int d\mathbf{w}_i P(\mathbf{y}_s \mid \mathbf{x}, \mathbf{w}_i, \mathcal{M}_i) P(\mathbf{w}_i \mid \mathcal{D}, \mathcal{M}_i). \tag{1.8}$$

In fact, following the argument above we will make predictions based on the average over this distribution. In other words, the output predicted by model $\mathcal{M}_i$ for an input, $\mathbf{x}$, when trained on the data, $\mathcal{D}$, will be

$$y_p = < \mathbf{y}_s(x) >_{P(\mathbf{y}_s|\mathbf{x},\mathcal{D},\mathcal{M}_i)} . \tag{1.9}$$

Here we have introduced the notation $\langle f(z) \rangle_{P(z|h)}$ to denote the average of the quantity $f(z)$ over the distribution $P(z \mid h)$. We now turn to the questions regarding the quality of this prediction.

Generalization

In order to quantify the quality of predictions made by our model we must choose an error measure and throughout this thesis we will concentrate on the squared error measure. However, this is not the only possible form of error measure (Levin *et al.*(89)). Indeed, Wahba (85) makes this comment but notes that often proximity in a squared error sense implies closeness in alternate error measures. Furthermore, there is more than one choice of definition of squared error measure (See *e.g.* Hansen (93) and Krogh and Hertz (92)). For example, for a student output y_p, of one-dimension an example of such a measure is

$$\epsilon(\mathbf{x}, y_p) = (y_p - y_t(\mathbf{x}))^2 , \tag{1.10}$$

which is the square difference between the prediction, y_p, of some model and a sample of the output of the true teacher, drawn from $P(\mathbf{y}_t \mid \mathbf{x})$. If the model prediction y_p was simply a sample from the predictive distribution $P(\mathbf{y}_s \mid \mathbf{x}, \mathcal{D}, \mathcal{M}_i)$, namely $\mathbf{y}_s(x)$, then the error measure (1.10) would be equivalent to that defined by Hansen (93). However, in this thesis we will take the model output y_p to be defined by equation (1.9). An

alternative measure which we also consider is

$$\epsilon(\mathbf{x}, y_p) = \left(y_p - < y_t(\mathbf{x}) >_{P(y_t|\mathbf{x})}\right)^2, \tag{1.11}$$

that is the squared difference between a prediction y_p and the expected teacher output at $\mathbf{x}$. This corresponds to the error measure used by Krogh and Hertz (92). The difference between these error measures (equations 1.10 and 1.11) is, on average, the variance over the conditional teacher distribution, $P(y_t \mid \mathbf{x})$. In chapter 2, where the teacher is a mixture model we will use the former error measure. Whilst, in chapter 3 onwards where the teacher is a deterministic function corrupted by noise, we will employ the second measure, equation (1.11). As stated above, we will employ the average of the model output, over the conditional predictive density (equation 1.8), as the model prediction, y_p (see equation 1.9).

Given a particular error measure, $\epsilon(\mathbf{x}, y_p)$, defined with respect to the underlying data generating process, the data dependent generalization error is defined as the expectation of $\epsilon(\mathbf{x}, y_p)$ with respect to the data generating process, $P(y_t \mid \mathbf{x})P(\mathbf{x})$. That is,

$$\epsilon_g(\mathcal{D}) = < \epsilon(\mathbf{x}, y_p) >_{P(y_t|\mathbf{x})P(\mathbf{x})} . \tag{1.12}$$

This data dependent generalization error, based on squared error measures, is our principal measure of performance and we will use it to compare the various model selection criteria we investigate in the subsequent chapters. The data dependency enters through the student output y_p which is conditioned on the data (see equation 1.9).

Other model selection criteria

To date we have introduced only one model selection method based on ranking models according to the *evidence* afforded to each by the data. However, as Mackay himself notes there is no *apriori* reason why the evidence optimal model should coincide with the model of optimal generalization ability (Mackay (92a)). In fact, we shall see an example in chapter 2 where this is, indeed the case. Given this, one might naturally seek alternatives which are closer in spirit to the ultimate goal of optimising the generalization error. We examine two such methods in this thesis, both are based on the idea of minimising some estimate of the generalization error over all the candidate models. We define these formally in chapter 5 where we examine model selection based on the *test error* and on the *cross-validation error*. Furthermore, we compare both of these

methods with that of the evidence. Finally, in chapter 6 we explore a model selection method based on the *noise sensitivity signature* of the model.

1.2.2 Review of existing work

In a real experiment one can never fully answer the question of which model is the optimal, rather one can only make statements in light of the available data. However, throughout this thesis we adopt a rather privileged position. That is we set up learning scenarios in which we know the true rule (the teacher), in the words of Brieman (94) we have a 'crystal ball'. We are then free to calculate the behaviour of various models (students) when presented with data sampled from this underlying rule. For example, we can calculate the model with the largest evidence. However, we can also calculate the generalization error for each model. As stated our ultimate goal is to achieve the best generalization performance. Thus, using this approach we can examine different model selection criteria in terms of their expected generalization performance and thus make some objective statements on the relative merits of each.

Much theoretical work has been carried out concerning the problem of learning from examples. Broadly speaking, we can classify this work into three areas. That with the longest lineage comes from statistics, where much work has been conducted in the asymptotic regime, where the number of examples is arbitrarily large, $p \to \infty$. In particular, a number of results from this approach are relevant to us here, some relate to the performance of the cross-validation error (see *e.g.* Stone (77a) and (77b), Shao (93), Plutowski (94) and Wahba (85)) whilst Gelfand and Dey (94) have unified a number of asymptotic analyses relating to the Bayes' factor and its variants. In addition, Meir and Merhav (94) studied Rissanen's stochastic complexity (Rissanen (86)) in the asymptotic regime. Clearly, the asymptotic approach is hardly relevant to the situation, more usual in practical applications, where the number of examples is of the order of the number of model parameters N_s. Indeed, this is recognised in the statistics community where increasing use is being made Monte Carlo techniques to evaluate quantities like the Bayes' factor for more realistic sample sizes, p (see *e.g.* Gelfand and Dey (94), Gelfand *et al.*(94), Neal (92) and Neal (93)).

The second approach, known as Probably Almost Correct, PAC for short, derives from the computer science and artificial intelligence communities (see Engel (94) and Anthony (95) for reviews). The PAC approach allows one to bound the generalization error of a particular model in terms of the *sample complexity*, p, and a model dependent quantity known as the Vapnik-Chervonenkis (V-C) dimension (see *e.g.* Valiant (84), Baum and Haussler (89), Vapnik and Chervonenkis (71)). In particular, results in this

framework are said to be distribution free, that is no assumption is made concerning the input distribution, $P(\mathbf{x})$. As a consequence the bounds on the generalization error take into account the worst case scenario and can thus be rather weak. Finally, PAC deals, in the main, with mappings between discrete spaces and thus has little direct relevance to the real valued inputs and outputs we consider. Nonetheless, within this approach Kearns *et al.*(95) investigated a broad class of learning algorithms, which they termed *penalty based*, comparing them to cross-validatory schemes. We will see in chapter 2 that the prior and noise model we consider fall into this class.

The third approach, and one to which we alluded earlier, is rooted in the statistical physics community. This applies to the regime in which the input dimension, or system size, N of the mappings to be learned is very large. In fact, we must also allow the number of examples to grow with N and we define this *thermodynamic limit* more rigorously in the next chapter. The principal advantage of this statistical mechanics approach as compared to the asymptotic regime is that it allows one to consider cases where the ratio of sample size, p to model parameters is finite. In this approach one can analyse specific learning scenarios in which the sampling assumption, $P(\mathbf{x})$, along with the student and teacher is defined. In contrast with the PAC approach more representative, *typical case*, generalization curves are calculated. Within the statistical mechanics approach generalization curves have been calculated for networks of varying complexity, an excellent review by Watkin *et al.*(93) covers much of this work. Note worthy are the contributions of Györgyi and Tishby (90), Sueng *et al.*(92), Krogh and Hertz (92) and Bös *et al.*(93). The majority of these calculations relate to the batch learning process where, as we shall see in chapter 2, training results in a Gibbs form for the posterior distribution. In the batch learning framework the most complicated network for which generalization curves have been calculated is the *committee machine* (see *e.g.* Schwarze (93)). The committee machine is a special case of an **MLP**, but similar calculations for the general case seem somewhat intractable at present. However, recently some advances have been made in the case of on-line learning where generalization curves can be calculated for general **MLP**s (see *e.g.* Saad and Solla (95a)(95b)) Although, it should be noted that this has not yet been achieved for arbitrary numbers of hidden units so that the universal approximation theorems, we mentioned in section 1.1, do not apply. Unfortunately, these online algorithms do not conveniently generate a posterior distribution and thus these exciting results are not relevant to our study.

The calculations presented in this thesis are performed within this statistical mechanics framework and in particular develop the work of Hertz *et al.*(89), Bruce and Saad (94) and Sollich (94a). Of particular relevance to model selection problems are

the works of Bruce and Saad (94) who examined the use of evidence in a learnable linear scenario and Meir and Fontanari's (93) study, within a binary system, of model selection based on the stochastic complexity (Rissanen (86)). We note here that the algorithms examined by both publications can be classed as penalty based approaches. Finally, we comment that the thermodynamic limit does compromise the applicability of the statistical mechanical approach to real world problems. However, recent work on finite size corrections has sought to rectify this problem. In particular, the work of Sollich (94a) demonstrates that thermodynamic results can be remarkably accurate for surprisingly small systems and recently, Barber (95) has studied finite size effects in cross-validatory errors. Furthermore, we calculate finite size effects in model selection problems in chapters 4 and 5.

1.3 Outline

We now, briefly, outline the remainder of this thesis. In chapter 2 we introduce a particular form for the noise model and the prior distribution. In particular, we make the assumption of Gaussian noise and a prior based on a complexity cost. We show how this Bayesian interpretation relates to the training process which can be described as a penalty based algorithm. We introduce two performance measures relating to generalization ability and consider a simple model selection problem namely that of selecting appropriate regularization parameters using the evidence criterion. The data generating process (teacher) we introduce allows us to interpolate between the learnable linear case and an unlearnable one in which the linear student is unable to model the teacher. The calculations presented are valid in the thermodynamic limit and are thus average case analyses. We briefly examine the robustness of the procedure with respect to the prior. In addition, we examine the evidence procedure in terms of the performance measures introduced.

Chapter 3 picks up where the previous chapter left off asking what would happen if the student was more than able to represent the teacher. We consider the case of a piece-wise linear student learning a linear teacher. The scenario considered is a case of nested models where the reduced model is a subset of the full (student) model. The true model, in other words the teacher, is the reduced model. Again, the calculations are performed in the thermodynamic limit.

In chapter 4 we explore finite size effects relating to the evidence procedure and selection of regularization parameters. We thus emphasise problems associated with average case analyses, showing that for finite sized systems such an approach can be highly misleading. We explore evidence assignments of the regularization parameters

and using simulations we corroborate these results showing that some of the qualitative features found in our first order finite size corrections are present in very low dimensional systems. Returning to the thermodynamic region we quantify the degradation in performance of these evidence assignments and find that in the noiseless case a phase transition exists, mirroring that found by Bruce and Saad (94). Finally, making use of numerical simulation we compare the evidence assignments with those of a cross-validatory assignment in a 1-dimensional system finding that the cross-validatory assignments are generally superior in terms of performance.

We pursue this result in chapter 5 analysing not only the leave-one-out cross-validation error (**CV(1)**) but, also the test error, in terms of their efficacy in model selection. In particular, once again, we focus on the simple problem of choosing regularization parameters. We adopt the noise model and priors used in the previous chapters. In so doing we find that both **CV(1)** and the test error are optimal in the thermodynamic limit and, indeed, in an average case sense. Thus, the focus of this chapter is, once again, finite sized systems. In the case of the test error we focus on the question of optimal partitioning of the data base into test and training sets examining two criterion on which to base this partition. We compare the performance obtained for these partitions with that found for the evidence in chapter 4. Turning to the cross-validatory method **CV(1)** we examine its assignment of the regularization parameters and the effect of these on performance. We then compare performance of all three approaches, evidence, test error and **CV(1)**.

Finally, in chapter 6 we consider the problem of selecting a model architecture rather than the hyper-parameter assignment examined in the earlier chapters. We introduce model selection based on the noise sensitivity signature of Grossman and Lapedes (95) and compare it with cross-validation and the evidence in a simple architecture selection problem. We conclude by summarising our main results and considering open questions and topics for future research.

Chapter 2

A Statistical Mechanical Analysis of a Bayesian Inference Scheme for an Unrealizable Rule

Abstract

Within the Bayesian framework outlined in the previous chapter we consider a system that learns from examples. In particular, using statistical mechanical methods in the thermodynamic limit, we calculate the evidence and two performance measures, namely the generalization error and the consistency measure, for a linear perceptron trained and tested on a set of examples generated by a non-linear teacher. The learning task is said to be unrealizable because the student can never model the teacher without error even for noiseless examples. In fact, our model allows us to interpolate between the known linear case and an unrealizable, non-linear, case. A comparison of the hyper-parameters which maximize the evidence with those that optimize the performance measures reveals that, when the student and teacher are fundamentally mismatched, the evidence procedure is a misleading guide to optimizing the performance measures considered. However, consideration of the degradation in performance invoked by the evidence assignments, as compared with the optimal, demonstrates that the procedure is nonetheless remarkably robust.

2.1 Introduction

In this chapter we seek to explore the effect of a misspecification of our model, with respect to the teacher, in terms of the reliability of the evidence as a criterion for model selection. In particular, we examine the case where the student is not sufficiently

powerful to model the teacher. As discussed in chapter 1 we will use techniques from statistical physics to explore this issue in a simple learning scenario. Firstly, however, we consider training from a different perspective to the Bayesian view discussed in the previous chapter. In the next section we relate the two approaches.

In general, one has a model mapping (*a student*) parameterized by some N_s-dimensional vector $\mathbf{w}$ and some possibly noisy examples, $\mathcal{D}$, generated by the true mapping (*the teacher*). During the training process one attempts to optimize the student parameters with respect to the underlying teacher. This task is said to be unrealizable when the optimal student does not model the teacher without error. The training error $E_{\mathbf{w}}(\mathcal{D})$ is some measure of the difference between the student and the teacher outputs over the set $\mathcal{D}$. Here this will be based on a squared error measure as will the generalization error (see section 1.2.1). Clearly, $E_{\mathbf{w}}(\mathcal{D})$ is an unsatisfactory measure of performance since it is limited to the training examples and very often we are interested in the students performance on a random example potentially, but not necessarily, in the training data; one measure of this performance is the generalization error itself (see *e.g.* Krogh and Hertz (92)).

The problem of over-fitting, discussed in section 1.1, occurs when we train our student so as to reproduce the noisy training data too closely. Thus, it can be seen that minimization of the training energy, with respect to the weights $\mathbf{w}$, can lead to over fitting. In fact, this observation has lead to the procedure known as *early stopping* where one stops training the student when it has reached some finite residual training error which is above the minimum possible. In fact, in can be shown that early stopping is equivalent to the regularization procedure known as weight decay when using a quadratic training error (see Bishop (95)). In this chapter we do not examine early stopping explicitly but do consider regularization by weight decay.

As noted earlier, in order to make successful predictions out with the set $\mathcal{D}$ (*i.e.* generalize) it is essential to have some prior preference for particular rules (Wolpert (92)). Occam's razor is an expression of our preference for the simplest rules which account for the data. Thus, in the learning process one can attempt to minimize $\beta E_{\mathbf{w}}(\mathcal{D}) + \gamma C(\mathbf{w})$, combining a measure of the performance on the data set and some *complexity cost* $C(\mathbf{w})$ of the model. The inclusion of the complexity cost penalizes complex models which, in general, will be able to over-fit the data to a greater degree than simpler ones. Kearns *et al.*(95) refer to such algorithms as *penalty based* for this reason. In discrete systems this is also known as the *minimum description length* principle (Rissanen (78)), where one minimizes the length of the code needed to describe the model itself and the errors it makes on the training set. In other words, as in the present case, one is trading off fidelity to the data set with model complexity. Rissanen's

stochastic complexity is one approach to this problem in discrete systems (Rissanen (86)). We also note, that in statistics minimization of a penalised cost function is known as *regularization* with the complexity cost termed the regularizer and γ and β regularization parameters (see *e.g.* Craven and Wahba (79) and Wahba (85) for examples applied to splines). In the neural networks community when $C(\mathbf{w}) = \mathbf{w}.\mathbf{w}$ this regularization procedure is known as *weight decay*. The setting of the regularization parameters, β and γ, also known as *hyper-parameters*, controls the learning algorithm. In this chapter we will concern ourselves with the question of how to set these hyper-parameters.

As we saw earlier, one can also consider the supervised learning paradigm within the context of Bayesian inference (see section 1.2.1). We describe how the above view of training, based on a penalized cost function, relates to this Bayesian framework in the next section. For now, however, we note that in this situation MacKay (92a) advocates a method based on the evidence (a quantity introduced in section 1.2.1) as a 'principled' method of setting hyper-parameters. Moreover, this *evidence procedure* is also a practical method since the evidence can be calculated from the data alone. Recently, there has been some debate as to the validity of this procedure (see *e.g.* Wolpert (93), MacKay (93) and Wolpert and Strauss (94)). However, most of this debate has focused on the validity of the evidence procedure as an approximation to a 'hierarchical' Bayesian calculation as opposed to its effects on student performance. We briefly comment on this debate later but focus on performance. In fact, in some situations the evidence procedure does seem to improve performance (Thodberg (93)) whilst in others, as MacKay points out, it can be misleading (MacKay (92 b)). We seek to explore these issues in a limited sense.

In particular we ask two questions; which performance measures do we seek to optimize and under what conditions will the evidence procedure optimize them? Performance measures, like the generalization error, are in some sense *objective* in that they indicate the extent to which the student has learned the underlying teacher. In order to investigate performance we consider particular classes of teacher and student. To date theoretical results have been obtained for a linear perceptron trained and tested on data produced by a linear perceptron (Bruce and Saad (94)). They suggest that the evidence procedure is a useful guide to optimizing the learning algorithms performance in this learnable case. In addition, also in an average case setting, Meir and Merhav (94) have investigated hyper-parameter assignment via minimization of the stochastic complexity of Rissanen (86).

In the remainder of this chapter we examine the evidence procedure hyper-parameter assignments, in relation to performance, for the case of a linear perceptron learning a

non-linear teacher. In the next section we review the Bayesian scheme, as it applies to the current problem, introducing the evidence and the relevant performance measures. In sections 2.3 and 2.4 we calculate these quantities in the case where the data is generated by a nonlinear mapping and the student is linear. We then proceed, in section 2.5, to examine the effects of the resultant unrealizability on the hyper-parameters derived from the evidence procedure. Finally, in section 2.6, we consider the effects of these assignments on performance, asking how robust is the evidence procedure to misspecification of the underlying problem?

2.2 Bayesian Formalism

2.2.1 The evidence

In this section we introduce the specific forms for the noise model and prior distribution relevant to us here. In particular, we note that since we are concerned with the setting of the hyper-parameters we consider the model architecture, $\mathcal{A}$ to be held fixed. Furthermore, we adopt the training algorithm outlined earlier which is defined through the setting of β and γ. We thus, have continuous spectrum of models to choose from (*i.e.* choice of γ and β). Therefore, we drop the subscript in $\mathcal{M}_i$ and simply write the model specification as $\mathcal{M} = \{\gamma, \beta\}$, with the dependence on the architecture and penalty based algorithm implicit.

As noted in section 1.2.1 we will concentrate on the squared error measure and thus, the training error, $E_{\mathbf{w}}(\mathcal{D})$, is the commonly used sum squared error. If our noise model assumes that the data is corrupted by Gaussian noise with variance $1/2\beta$ then the probability, or *likelihood* of the data($\mathcal{D}$) being produced given the parameters $\mathbf{w}$ and β is,

$$P(\mathcal{D} \mid \mathbf{w}, \beta) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D})}. \tag{2.1}$$

which is analogous to equation (1.2). However, in this case the dependence of the likelihood on the model, $\mathcal{M}$ is expressed solely in terms of the hyper-parameter, β. We note here that maximum likelihood specification of the model parameters is tantamount to minimization of the training error, $E_{\mathbf{w}}(\mathcal{D})$, and can thus lead to over-fitting.

In order to incorporate Occam's razor we also assume a prior distribution on our models parameters. That is, we believe *a priori* in some parameter assignments more strongly than others. Specifically we believe that,

$$P(\mathbf{w} \mid \gamma) \propto e^{-\gamma C(\mathbf{w})}. \tag{2.2}$$

Similarly, this is analogous to equation (1.3) and is dependent only on γ, although the full model specification, as noted above, is $\mathcal{M} = \{\beta, \gamma\}$.

Multiplying these together, as before, we obtain the posterior distribution, also known as the post training or student distribution,

$$P(\mathbf{w} \mid \mathcal{D}, \mathcal{M}) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D}) - \gamma C(\mathbf{w})}. \tag{2.3}$$

It is clear that the most probable model parameter vector, $\mathbf{w}^*$, is given by minimizing the composite cost function $\beta E_{\mathbf{w}}(\mathcal{D}) + \gamma C(\mathbf{w})$ with respect to $\mathbf{w}$. In this sense the Bayesian viewpoint coincides with minimization of this composite cost function by gradient descent (*e.g. backpropagation*). In fact, it should be noted that a stochastic learning algorithm based on a Langevin dynamics on the composite cost function can also give rise to the post training distribution, equation (2.3) (Seung *et al* (92)). Indeed, Buntine and Weigend (91) refer to this process as *Bayesian Backpropagation.*

The evidence itself is the missing normalisation constant in (2.3),

$$P(\mathcal{D} \mid \gamma, \beta) = \int \prod_j dw_j P(\mathcal{D} \mid \beta, \mathbf{w}) P(\mathbf{w} \mid \gamma). \tag{2.4}$$

This is, the probability of (or evidence for) the data set ($\mathcal{D}$) given the hyper-parameters β and γ. The *evidence procedure* fixes the hyper-parameters to the values that *simultaneously* maximize this probability for a given data set (MacKay (92a)). It is, therefore, analogous to use of the Bayes' factor discussed in section 1.2.1.

ML II and Hierarchical Bayes

We now comment briefly, on the specific approach taken here in relation to more general issues in Bayesian analysis. In the current context the evidence is used to determine the hyper-parameters. In particular, γ parameterises the prior distribution of equation (2.2) and β can also be regarded as a prior assumption on the noise model. In using the evidence procedure, therefore we are selecting a prior belief, from a class of priors, using the data. Broadly, speaking such an approach has come to be known as *empirical Bayes.* Furthermore, in the approach used here, namely maximization of the evidence, the resultant prior is referred to as a type II maximum likelihood, or **ML II**, prior (Berger and Berliner (83)). In this case the evidence can be regarded as a likelihood function (see equation 1.2) for the hyper-parameters. As noted in section 1.2.1 a further issue, to be considered when one has a number of possible priors, is the question of robustness, or sensitivity with respect to changes in prior (Berger (85)). We discuss, these issues within the current context in section 2.5.2 and 2.6

In cases, such as that considered here, where one has a parameterised prior *Hierarchical Bayesian* analysis offers an alternative to identification of a single most likely prior. In this case a prior is introduced on the hyper-parameter (*e.g.* $P(\gamma)$), a so called *hyper-prior*. One then proceeds by integrating out the hyper-parameter over this prior. Very often the choice of hyper-prior is difficult and uniform, sometimes improper (unnormalisable), distributions are chosen. In the case considered here if we denote the hyper-prior on both our hyper-parameters by $P(\beta, \gamma)$, then the posterior resulting from the hierarchical approach would be,

$$P(\mathbf{w} \mid \mathcal{D}) \propto \int P(\mathbf{w} \mid \mathcal{D}, \beta, \gamma) P(\mathcal{D} \mid \beta, \gamma) P(\beta, \gamma) d\beta d\gamma \tag{2.5}$$

At the heart of the debate, mentioned earlier, concerning the validity of the evidence procedure is Mackay's claim that often the evidence, $P(\mathcal{D} \mid \beta, \gamma)$, is rather peaked around the evidence procedure assignments, β_{ev} and γ_{ev} (see Mackay (92a) and (93)). In such cases Mackay, goes on to argue that $P(\mathbf{w} \mid \mathcal{D}, \beta_{ev}, \gamma_{ev})$ is then a good approximation to the hierarchical posterior $P(\mathbf{w} \mid \mathcal{D})$. Wolpert (93) and Wolpert and Strauss (94) have pointed out that in many cases this is indeed not true, and the posterior resulting from evidence procedure assignments is a poor approximation to the hierarchical posterior. In section 2.4 we comment on this debate in relation to the learning scenario considered in this chapter.

2.2.2 The performance measures

Before defining our performance measures let us clarify our notation (introduced in section 1.2). In general we consider a teacher with real one dimensional output, $y_t(\mathbf{x})$, described by the conditional density $P(y_t \mid \mathbf{x})$. This accommodates, for example, deterministic teachers whose output is corrupted by noise. Furthermore, the inputs $\mathbf{x}$ are N dimensional vectors sampled with probability $P(\mathbf{x})$. Thus a data set $\mathcal{D}$= $\{(y_t(\mathbf{x}^\mu), \mathbf{x}^\mu) : \mu = 1..p\}$ is generated with probability $P(\mathcal{D}) = \prod_{\mu=1}^{p} P(y_t \mid \mathbf{x}^\mu) P(\mathbf{x}^\mu)$. Also we will, in general continue to use the notation $\langle f(z) \rangle_{P(z|h)}$ to denote the average of the quantity $f(z)$ over the distribution $P(z \mid h)$. However, we will use the short hand $\langle . \rangle_{\mathbf{w}}$ to mean the average over the posterior distribution $P(\mathbf{w} \mid \mathcal{D}, \gamma, \beta)$.

As mentioned earlier, we will base our predictions, for a given input $\mathbf{x}$, on the average over the conditional predictive distribution, equation (1.8). That is, for the input $\mathbf{x}$ our student will output, $\langle y_s(\mathbf{x}) \rangle_{P(y_s|\mathbf{x},\mathcal{D},\mathcal{M})}$. The notation implies that the output is conditioned on the input, $\mathbf{x}$, for which a response is required, the data, $\mathcal{D}$, on which the student has been trained, and on the form of the model, $\mathcal{M}$, from which the student has been generated (see section 1.2.1). However, given our noise model

the average student output is equivalent to the average over the posterior, namely, $\langle y_s(\mathbf{x})\rangle_{\mathbf{w}}$.

Many performance measures have been introduced in the literature (See *e.g.* Hansen (93), Levin *et al.*(89) and Krogh and Hertz (92)). Throughout this thesis we consider squared error measures and here we adopt the definition of equation (1.10). The data dependent generalization error (equation 1.12) averaged over all possible data sets of size p is then,

$$\epsilon_g = \langle \epsilon_g(\mathcal{D})\rangle_{P(\mathcal{D})} = \langle (y_t(\mathbf{x}) - \langle y_s(\mathbf{x})\rangle_{\mathbf{w}})^2\rangle_{P(y_t|\mathbf{x})P(\mathbf{x})P(\mathcal{D})}. \tag{2.6}$$

This average generalization error is equivalent to that given by Krogh and Hertz (92) up to an additive teacher dependent constant, namely the variance in the teacher output over $P(y_t \mid \mathbf{x})$.

Another feature we can consider is the variance of the student output, $y_s(\mathbf{x})$, over the posterior and input distributions, $\langle \{y_s(\mathbf{x}) - \langle y_s(\mathbf{x})\rangle_{\mathbf{w}}\}^2\rangle_{\mathbf{w},P(\mathbf{x})}$. This gives us a measure of the confidence we should have in our post training distribution and could be estimated if we could estimate the input distribution $P(\mathbf{x})$. Bruce and Saad define the consistency measure as the difference between this variance and the generalization error (Bruce and Saad (94)). Here we extend this definition to include the case of unlearnable rules, by adding the asymptotic value of the generalization error (*i.e.* adding $\epsilon_g^\infty = \lim_{\alpha\to\infty}\epsilon_g$, where $\alpha = p/N$.). Thus, ϵ_g^∞ is the minimum possible error achieved by the student considered. The consistency measure δ_c is now defined by,

$$\delta_c = \langle \delta_c\rangle_{P(\mathcal{D})} = \langle \{y_s(\mathbf{x}) - \langle y_s(\mathbf{x})\rangle_{\mathbf{w}}\}^2\rangle_{\mathbf{w},P(\mathbf{x}),P(\mathcal{D})} - \left(\epsilon_g - \epsilon_g^\infty\right). \tag{2.7}$$

In the case considered here we will see that the variance of the student output tends to zero in the limit of large data sets. Thus, δ_c also tends to zero as $\alpha \to \infty$, even though the generalization error may not be zero. We regard $\delta_c = 0$ as optimal since then we can estimate the decaying part of the expected error, $\epsilon_g - \epsilon_g^\infty$, from the variance of our student output.

The fact that the quantities (2.6) and (2.7) are averages over the data is just analytical artifice. For example, in an experiment we would wish to make predictions based on a single data set. In other words we would be interested in the data dependent generalization error $\epsilon_g(\mathcal{D})$ and consistency measure $\delta_c(\mathcal{D})$. In this chapter we conduct an average case analysis and thus, concentrate on the average generalization and consistency, however, in chapters 4 and 5 we will focus on the data dependent measures. For now though when we refer to the generalization error or the consistency measure the data average is implied. Unfortunately these performance measures (averaged or not)

can only be calculated if we assume we know more about the teacher than simply the training examples. However, the evidence can in principal be calculated exactly from the data alone, although it does embody our *assumptions* about the noise process and prior distribution. Arguably, minimization of $\epsilon_g(\mathcal{D})$ is the ultimate goal of supervised learning. It is, therefore, desirable to know when the evidence procedure minimizes this quantity. We now set up a specific learning scenario in which we can examine this questions analytically.

2.3 Learning scenario

In the scenario considered here the student is simply a linear perceptron and the input dimension N equals the model dimension N_s. The output for an input vector $\mathbf{x}^\mu$ is given by

$$y_s^\mu = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j x_j^\mu \tag{2.8}$$

In contrast, our teacher is a non-linear mapping which we refer to as an n-teacher because it is a mixture of n linear component teachers. The Ωth component teacher is is corrupted by Gaussian noise of mean zero and variance σ_Ω^2. The resulting conditional output distribution for the n-teacher is,

$$P(y_t \mid \mathbf{x}) = \sum_{\Omega=1}^{n} P(y_t \mid \mathbf{x}, \Omega) P(\Omega \mid \mathbf{x}). \tag{2.9}$$

where $P(y_t \mid \mathbf{x}, \Omega) \propto \exp([y_t - \mathbf{w}^\Omega.\mathbf{x}/\sqrt{N}]^2/2\sigma_\Omega^2)$ accounts for the corrupting output noise, and using Bayes' theorem one can write

$$P(\Omega \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Omega) P_\Omega^t}{P(\mathbf{x})}. \tag{2.10}$$

Here the input distribution $P(\mathbf{x}) = \sum_{\Omega=1}^{n} P_\Omega^t P(\mathbf{x} \mid \Omega)$ with $P(\mathbf{x} \mid \Omega) \sim \mathrm{N}(\bar{a}_\Omega, \sigma_{x_\Omega}^2)$ [1] and P_Ω^t, which denotes the weight given to each component Ω, is chosen such that $\sum_{\Omega=1}^{n} P_\Omega^t = 1$. Mixture models, of which equation (2.9) is an example, have recently received attention in the neural network community where one refers to a *mixture of experts* (see, for example Jacobs (95) and Jordan and Jacobs (94)).

One way of visualising the n-teacher mapping is as the average over the conditional distribution $P(y_t|\mathbf{x})$. Figure 2.1 displays some examples of a 2-teacher with one dimensional input vector. Figure 2.1(i) shows the linear case whilst (ii) shows the average

[1] Here we are using $\mathrm{N}(\bar{\mathrm{x}}, \sigma^2)$ to denote a normal distribution with mean $\bar{\mathrm{x}}$ and variance σ^2.

of two linear teachers when the distributions $P(\mathbf{x} \mid \Omega)$ are the same, again the average output is a linear function of the input. Finally in Figure 2.1(iii) the distributions $P(\mathbf{x} \mid \Omega = 1)$ and $P(\mathbf{x} \mid \Omega = 2)$ are both centred on the origin but have different variances; the average output is a non-linear function of the input. In fact, for the general case, where the distributions, $P(\mathbf{x} \mid \Omega)$, have different means and variances, in the large N limit the input space is divided between the component teachers with each one representing a linear section of the mapping. In this way a non-linear teacher is constructed, in a piece-wise linear fashion, with n segments. As n grows we can steadily improve our approximation of arbitrary piece-wise linear functions.

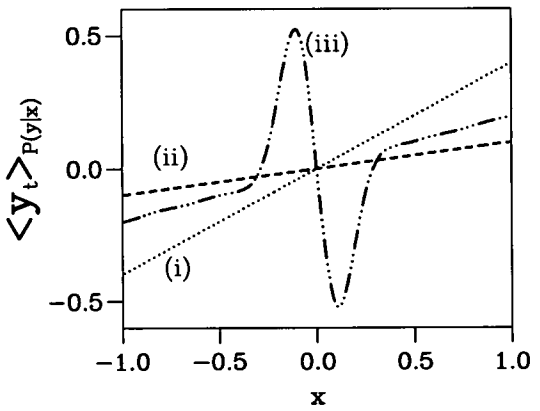


Figure 2.1. A 2-teacher in 1D : The average output $\langle y_t \rangle_{P(y|x)}$ (i) when the component teacher vectors are aligned , (ii) when they are misaligned but $\sigma_{x_1} = \sigma_{x_2}$ and (iii) with $\sigma_{x_1} \neq \sigma_{x_2}$ and with the teachers misaligned.

Given this model a data set, of p examples, is given by $\mathcal{D} = \prod_{\Omega=1}^{n} \{(\mathbf{w}^{\Omega}.\mathbf{x}_{\Omega}^{\mu_\Omega}/\sqrt{N} + \eta_{\Omega}^{\mu_\Omega}, \mathbf{x}_{\Omega}^{\mu_\Omega}) : \mu_\Omega = 1..p_\Omega\}$. The variables $\eta_{\Omega}^{\mu_\Omega}$ are drawn independently from a Gaussian distribution with zero mean and variance σ_{Ω}^2 whilst the $\mathbf{x}_{\Omega}^{\mu_\Omega}$ are drawn independently from $P(\mathbf{x} \mid \Omega)$. The range of the index μ_Ω is from 1 to p_Ω where on average $p_\Omega = p \times P_{\Omega}^{t}$.

Adopting weight decay as our regularization scheme, that is $C(\mathbf{w}) = \mathbf{w}.\mathbf{w}$, we can now explicitly write the evidence in terms of these random variables and then perform the integration over the student parameters (*over weights*). Taking the logarithm of the resulting expression leads to $\ln P(\mathcal{D} \mid \lambda, \beta) = -N f(\mathcal{D})$, where we have introduced the *weight decay* parameter, $\lambda = \gamma/\beta$. The quantity $f(\mathcal{D})$ is analogous to a free energy density in statistical physics. This analogy has been noted by others, for example Neal

(92). The expression for the free energy is of the form,

$$-f(\mathcal{D}) = \frac{1}{2}\ln\frac{\lambda}{\pi} + \frac{\alpha}{2}\ln\frac{\beta}{\pi} + \frac{1}{2}\ln 2\pi + \frac{1}{2N}\ln \det g + \frac{1}{N}\rho_j g_{jk}\rho_k - \kappa \tag{2.11}$$

where,

$$\rho_j = 2\beta\left\{(A_\Omega)_{jk}w_k^\Omega + \frac{1}{\sqrt{N}}\eta_\Omega^{\mu_\Omega}(x_\Omega^{\mu_\Omega})_j\right\}$$
$$\kappa = \frac{\beta}{N}\left\{(A_\Omega)_{jk}w_j^\Omega w_k^\Omega + \frac{2}{\sqrt{N}}\eta_\Omega^{\mu_\Omega}(x_\Omega^{\mu_\Omega})_j w_j^\Omega + \eta_\Omega^{\mu_\Omega}\eta_\Omega^{\mu_\Omega}\right\}$$
$$g_{jk}^{-1} = \sum_{\Omega=1}^{n}(A_\Omega)_{jk} + \lambda\delta_{jk} \qquad (A_\Omega)_{jk} = \frac{1}{N}(x_\Omega^{\mu_\Omega})_j(x_\Omega^{\mu_\Omega})_k \qquad \alpha = \frac{p}{N}$$

Here we are using the convention that summations are implied where repeated indices occur.

The performance measures can be calculated from data averages of appropriate quantities derived from the evidence. Thus we find

$$\epsilon_g = \langle\frac{\sigma_{x_\Omega}^2}{N}P_\Omega^t\left\{w_j^\Omega w_j^\Omega - 2w_j^\Omega\langle w_j\rangle_{\mathbf{w}} + \langle w_j\rangle_{\mathbf{w}}^2\right\}\rangle_{P(\mathcal{D})} + P_\Omega^t\sigma_\Omega^2, \tag{2.12}$$

$$\delta_c = \frac{\sigma_{x_{\text{eff}}}^2}{2N\beta}\langle\text{tr } \mathbf{g}\rangle_{P(\mathcal{D})} - \left(\epsilon_g - \epsilon_g^\infty\right), \tag{2.13}$$

where $\langle w_j\rangle_{\mathbf{w}} = \rho_k g_{kj}$ and $\sigma_{x_{\text{eff}}}^2 = P_\Omega^t\sigma_{x_\Omega}^2$. The calculation of the free energy and these performance measures is outlined in the appendix 4.8.1.

Due to the sampling assumptions in our learning scenario all these quantities are functions of random variables, that is of random data sets. To proceed analytically we must perform an average over these data sets (*i.e.* over the distribution $P(\mathcal{D})$).

2.4 Thermodynamic averages

In order to perform these averages we are forced to consider a particular n-teacher. We choose the 2-teacher ($n = 2$) with an input distribution with zero mean, $\bar{a}_\Omega = 0$. The method used to calculate the average is an extension to that used by Hertz and his co-workers (Hertz *et al.*(89)). Using this we can calculate the data average of the free energy, f, in the thermodynamic limit. That is, as $N, p \to \infty$ with $\alpha = p/N =$ constant.

As we discussed earlier, considering the average over all possible sets of data is somewhat artificial in that we could calculate $f(\mathcal{D})$ and would be interested in the

generalization error for our learning algorithm given a particular instance of the data. However, in the thermodynamic limit, due to our sampling assumptions these quantities, as functions of a particular set of data $\mathcal{D}$, coincide with their averages over all data sets. We will discuss the effects, on the evidence procedure, of the thermodynamic approximation in more detail in chapter 4. However, the essential point is that the variances, over data sets, of quantities like the free energy or the generalization error are of the order $\mathcal{O}(1/N)$. Thus, in the thermodynamic limit, the fluctuations, from one data set to the next, vanish and the behaviour associated with any particular data set corresponds to the average case.

The fact that the free energy is self averaging has consequences for the validity of the evidence procedure as an approximation to the full hierarchical Bayesian calculation discussed in section 2.2.1. Sollich (95c) has noted that, since the evidence itself is given by $P(\mathcal{D} \mid \gamma, \beta) = \exp(-Nf(\mathcal{D}, \gamma, \beta))$, then in the thermodynamic limit this is dominated by the saddle point of the left hand expression, namely by minimum of the free energy, $f(\mathcal{D}, \gamma, \beta)$. In other words, for large N the distribution, $P(\mathcal{D} \mid \gamma, \beta)$ concentrates its mass around $(\gamma_{ev}, \beta_{ev})$, which minimise the free energy. Thus, in the case considered here the posterior resulting from the evidence procedure is a good approximation to that derived from the hierarchical calculation (see equation 2.5).

We now calculate these thermodynamic averages. Following the average over the noise variables we are left with the average over the input distribution. In particular, we need to calculate $\ll \mathbf{g} \gg$, $\ll \mathbf{A}_\Omega\, \mathbf{g} \gg$ and $\ll \mathbf{A}_2\, \mathbf{g}\mathbf{A}_1 \gg$, where the double brackets refer to averages, in the thermodynamic limit, over the input distribution. The details are relegated to appendix 2.8, where equation (2.32) defines $NG = \mathrm{tr} \ll \mathbf{g} \gg$ and $\ll \mathbf{A}_\Omega\, \mathbf{g} \gg$ and $\ll \mathbf{A}_2\, \mathbf{g}\mathbf{A}_1 \gg$ are defined by equations (2.29) and (2.30) respectively.

The averaged free energy f can now be written

$$\begin{aligned} f = -\frac{1}{2}\ln\frac{\lambda}{\pi} - \frac{\alpha}{2}\ln\frac{\beta}{\pi} - \frac{1}{2}\ln 2\pi - \frac{1}{2N} \ll \ln\det g \gg + \beta\sigma_1^2(P_1^t\alpha - \Psi_1 G) \\ + \beta\sigma_2^2(P_2^t\alpha - \Psi_2 G) + \beta G(\sigma_{w_1}^2\lambda\Psi_1 + \sigma_{w_2}^2\lambda\Psi_2 + \Psi_1\Psi_2 D_w), \end{aligned} \tag{2.14}$$

where $\Psi_\Omega = \ll \Sigma_\Omega \gg$ and Σ_Ω is defined in the appendix (see equation 2.26). Similarly, the generalization error is re-expressed in the form,

$$\begin{aligned} \epsilon_g = \quad & P_1^t\sigma_{w_1}^2\sigma_{x_1}^2 + P_2^t\sigma_{w_2}^2\sigma_{x_2}^2 + P_1^t\sigma_1^2 + P_2^t\sigma_2^2 \\ & -2P_1^t\sigma_{x_1}^2(\Psi_1\sigma_{w_1}^2 + \Psi_2\theta_w)G - 2P_2^t\sigma_{x_2}^2(\Psi_2\sigma_{w_2}^2 + \Psi_1\theta_w)G \\ & +\sigma_{x_{\mathrm{eff}}}^2\frac{\partial}{\partial\lambda}\left[G\left\{\Psi_1\Psi_2 D_w + \Psi_1(\lambda\sigma_{w_1}^2 - \sigma_1^2) + \Psi_2(\lambda\sigma_{w_2}^2 - \sigma_2^2)\right\}\right], \end{aligned} \tag{2.15}$$

whilst the consistency measure becomes,

$$\delta_c = \frac{\sigma_{x_{\text{eff}}}^2}{2\beta} G - \left(\epsilon_g - \epsilon_g^\infty\right), \tag{2.16}$$

and we have defined,

$$D_w = \frac{1}{N} |\mathbf{w}^{o1} - \mathbf{w}^{o2}|^2 \quad \sigma_{w_\Omega}^2 = \frac{1}{N} \mathbf{w}^{o\Omega} \cdot \mathbf{w}^{o\Omega} \quad \text{and} \quad \theta_w = \frac{1}{N} \mathbf{w}^{o1} \cdot \mathbf{w}^{o2}.$$

The variable D_w is a measure of the Euclidean distance in weight space between the two components of the teacher whilst, $\sigma_{w_\Omega}^2$ measures the magnitude of component Ω and θ_w is the overlap between the two linear components of which the teacher is comprised. We also note that in two limits we recover the learnable, linear teacher, case. Firstly, if the probability, P_Ω^t, of picking one of the components of the teacher is zero then the additional terms in the free energy, equation (2.14), vanish and the response function, appendix 2.8, collapses to the quadratic linear solution. Similarly, if we assume that D_w and $\sigma_{x_1} - \sigma_{x_2}$ are small then response function and the free energy (and hence the generalization error and consistency) can be expressed as a Taylor expansion around the linear case. We now examine the evidence and the performance measures for the unlearnable problem.

2.5 Results and discussion

2.5.1 The performance measures

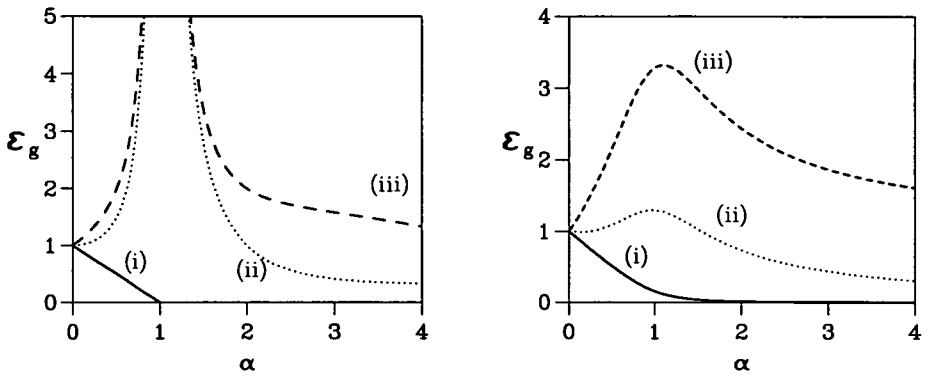
Firstly let us consider the performance measures. The asymptotic value of ϵ_g for large α is

$$\epsilon_g \approx \frac{P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w}{\sigma_{x_{\text{eff}}}^2} + P_\Omega^t \sigma_\Omega^2 + \frac{P_1^t P_2^t \sigma_{x_1}^4 \sigma_{x_2}^4 D_w}{\sigma_{x_{\text{eff}}}^6} \frac{1}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right), \tag{2.17}$$

where repeated indices imply summation. Similarly, also for large α,

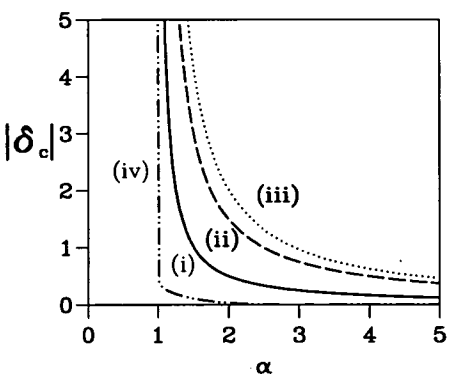
$$|\delta_c| \approx \frac{1}{\alpha} \left(\frac{P_1^t P_2^t \sigma_{x_1}^4 \sigma_{x_2}^4 D_w}{\sigma_{x_{\text{eff}}}^6} + \frac{1}{\beta \sigma_{x_{\text{eff}}}^4} \right) + \mathcal{O}\left(\frac{1}{\alpha^2}\right). \tag{2.18}$$

In the limit of infinite α, $|\delta_c|$ tends to zero and $\epsilon_g^\infty = P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w / \sigma_{x_{\text{eff}}}^2 + P_\Omega^t \sigma_\Omega^2$. This is the minimum generalization error attainable and reflects the effective noise level with a component due to the mismatch between student and teacher which vanishes when the two component teacher vectors are aligned ($D_w = 0$). This minimum error corresponds to a student weight vector with components $w_k = (P_1^t \sigma_{x_1}^2 w_k^1 + P_2^t \sigma_{x_2}^2 w_k^2)/\sigma_{x_{\text{eff}}}^2$, which is simply an appropriate mixing of the component teacher weights.

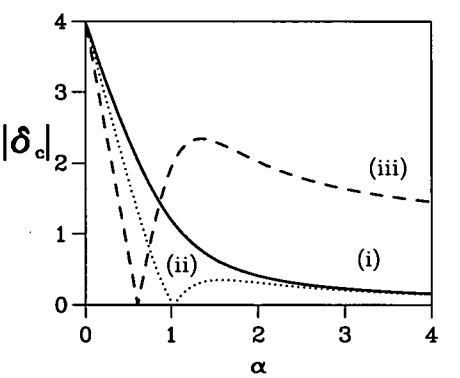


(a) ϵ_g for zero γ

(b) ϵ_g for finite γ



(c) δ_c for zero γ



(d) δ_c for finite γ

Figure 2.2. The performance measures: Graph a shows the generalization error v's α for zero γ. a(i) and (ii) are the realizable case without noise and with noise respectively. Curve a(iii) is an unlearnable case where we can see that the unlearnability qualitatively acts in the same manner as noise (*i.e.* causes a divergence). Graph b shows ϵ_g for finite γ. b(i) and (ii) are learnable scenarios in the latter case with noise. b(iii) shows that the effect of *adding* unlearnability is qualitatively the same as adding noise. Graph c shows $|\delta_c|$ for $\gamma \to 0$, note that for $\alpha < 1$ the consistency measure diverges. Graph c(i) shows the learnable linear case. c(ii) shows the unlearnable but linear case and c(iii) is the non-linear case. Curve c(iv) shows the effect of setting the learning temperature to $T_{ev} \equiv 1/\beta_{ev}$; the evidence optimal temperature. In this latter case the optimal value of the consistency measure is $|\delta_c| = 0$, for $\alpha > 1$, and thus, the evidence assignment is sub-optimal in the non-linear regime. Graph d shows the modulus of the consistency error v's α for finite γ. Curves d(i) and (ii) are the learnable case without and with noise respectively. Curve d(iii) is an unlearnable case with the same noise level.

Another limit which we can examine is the case of an unregularized student distribution ($\gamma \to 0$). In this case we must be careful as the response function G is ill defined for $\alpha < 1$. In fact, the consistency measure diverges in this region. However, in this limit for $\alpha < 1$ the generalization error is,

$$\epsilon_g = \tau\left(\alpha, P_\Omega^t, \sigma_{x_\Omega}^2, \sigma_\Omega^2, \sigma_{w_\Omega}^2\right) + \frac{\alpha\sigma_{x_{\text{eff}}}^2}{1-\alpha}\left(\frac{P_1^t\sigma_1^2}{\sigma_{x_1}^2} + \frac{P_2^t\sigma_2^2}{\sigma_{x_2}^2} + \alpha P_1^t P_2^t D_w\right), \tag{2.19}$$

which clearly shows a divergence, as α approaches unity from below, if we have noise on the examples and/or the component teacher vectors are not aligned ($D_w > 0$). The function τ represents the remaining, non-diverging, component. This divergence is also seen as α approaches 1 from above. If we expand ϵ_g about $\alpha = 1$ the first term is

$$\epsilon_g \sim \left(\frac{\sigma_1^2\sigma_{x_2}^2 P_1^t + \sigma_2^2\sigma_{x_1}^2 P_2^t + P_1^t P_2^t \sigma_{x_2}^2\sigma_{x_1}^2 D_w}{\sigma_{x_{\text{eff}}}^2\sigma_{x_1}^2\sigma_{x_2}^2}\right)(\alpha-1)^{-1}. \tag{2.20}$$

In accord with standard results (*e.g.* Krogh and Hertz (92) and Dunmur and Wallace (93)), if there is no noise and $D_w = 0$ (*i.e.* noiseless learnable linear case), the generalization error is proportional to $1-\alpha$ for $\alpha < 1$ and zero for $\alpha > 1$. Figure 2.2(a) shows the generalization error in the zero γ limit. The case of noiseless linear teacher is included for reference. In this case, the addition of noise causes ϵ_g to diverge at $\alpha = 1$. We also observe the same effect when we have a nonlinear teacher. As the scalar product between the component teachers reduces (D_w increases) the divergence becomes more rapid. Thus, the unlearnability of the teacher acts as an effective noise on the examples.

We also see this effect in Figure 2.2(b) which shows the generalization error for finite γ plotted against α. In this case also, the addition of unlearnability has a similar effect to the addition of noise on the examples. The peak in the generalization error, for small but finite γ, can be regarded as the precursor to the divergence at $\alpha = 1$ as $\gamma \to 0$ discussed above. The appearance of this maximum can be easily understood; If there is no noise or γ is large enough then there is a steady reduction in ϵ_g (2.2(b), curve (i)), however, if this is not so then for small α the student learns the effective noise and the generalization error increases with α. As the student gets more examples the effects of the noise begin to average out and the student starts to learn the rule. The point at which the generalization error starts to decrease is influenced by the effective noise level and the prior constraint. We note here that the idea that unlearnability acts as an effective noise is not new (see for example Sollich (95a)).

Figure 2.2(c) shows the consistency measure for $\gamma \to 0$, for $\alpha < 1$ this diverges even

in the learnable noiseless limit. Again unlearnability acts as an effective noise. As we shall see in section 2.5.2, in this limit, the consistency is optimized by the evidence procedure for the linear case only. A non-linear case is shown in figure 2.2(c), curve (iv), where the temperature, defined by $T \equiv 1/\beta$, is set by the evidence procedure. Since, the optimal consistency measure is actually zero the evidence assignment is seen to be suboptimal in this case.

Finally Figure 2.2(d) shows the absolute value of the consistency measure versus α for finite γ. Again we see that unlearnability acts as an effective noise. The post training distribution variance reduces as α increases. For a few examples with γ small or with large effective noise the student distribution narrows until δ_c is zero. However, the generalization error is non-optimal since the students have simply learned the effective noise. The position of the zero of the consistency measure is a reflection of the trade-off between the effective noise and the weight decay described above (curves (ii) and (iii) in figure 2.2(d) show the result of varying the effective noise). As α increases further $|\delta_c|$ begins to increases to a local maximum, it then asymptotically tends to zero. If there is no noise or γ is large enough then $|\delta_c|$ steadily reduces as the number of examples increases (as shown in curve (i) of 2.2(d)), since then both $\epsilon_g - \epsilon_g^\infty$ and the variance of the posterior distribution decay monotonically with α.

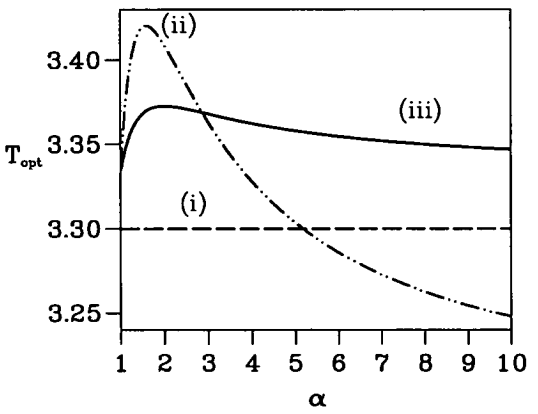


Figure 2.3. Optimal temperatures in the $\gamma \to 0$ model: (i) The evidence procedure estimate T_{ev} and that which optimizes the consistency measure T_{δ_c} coincide in the linear regime. In the non-linear regime (ii) shows the dependence of T_{ev} on α and (iii) shows that of T_{δ_c}.

2.5.2 The evidence procedure

We now turn to the evidence and, in particular, to the assignments of the hyper-parameters we can make from it. We define $\beta_{ev}(\gamma)$ and $\gamma_{ev}(\beta)$ to be the hyper-parameters which maximise the evidence with respect to fixed γ and β respectively. The evidence procedure picks the point in hyper-parameter space where these curves coincide. Furthermore, we define β_{ev}^{∞} and γ_{ev}^{∞} to be the solutions to $\lim_{\alpha\to\infty} \frac{\partial f}{\partial \beta} \mid_{\gamma=const.} = 0$ and $\lim_{\alpha\to\infty} \frac{\partial f}{\partial \gamma} \mid_{\beta=const.} = 0$ respectively. In what follows we shall refer to the *linear regime* as the case when $\sigma_{x_1} = \sigma_{x_2}$ or when $D_w = 0$ and $\sigma_1 = \sigma_2$. This is because the average teacher output is then linear. In contrast, when $D_w > 0$ and $\sigma_{x_1} \neq \sigma_{x_2}$ and the average teacher output is an N dimensional analogue of curve (iii) in figure 2.1, we shall speak of the *non-linear regime.* We also note that if $\sigma_{x_1} \neq \sigma_{x_2}$ and $\sigma_1 \neq \sigma_2$ then the noise is not constant across input space.

The $\gamma \to 0$ limit

The simplest case is the unregularized limit where we have only one hyper-parameter (β) to optimize. Before proceeding we note that the hyper-parameter β can be written in terms of the temperature $T \equiv 1/\beta$. In the limit $\gamma \to 0$, for $\alpha < 1$, the evidence is independent of β whereas, for $\alpha > 1$, the evidence optimal temperature (T_{ev}) is finite as shown in figure 2.3, curves (i) and (ii). In fact, for $\alpha > 1$ and increasing we can make steadily better estimates of the noise on the examples.

This transition in behaviour is analogous to the phase transition found by Bruce and Saad (94). In the regime $\alpha < 1$ there is not even enough data to specify the perceptron weights let alone the hyper-parameter, β. This is demonstrated in the linear case where the variance in T_{ev} diverges as α approaches unity from above. In fact, for $\alpha > 1$, $\mathrm{Var}(T_{ev}) = 8\sigma^4/(N(\alpha-1))$. Thus, the variance in the evidence estimate of the learning temperature is an order $\mathcal{O}(1/N)$ quantity for $\alpha > 1$ and the evidence optimal temperature itself is well defined in the thermodynamic limit. However, as α approaches unity from above this variance becomes order $\mathcal{O}(1)$ revealing a breakdown in the self averaging assumptions of the thermodynamic limit. In the regularized case (see below) this phase transition does not occur because our prior belief provides the additional information required to estimate the noise from the data even for $\alpha < 1$.

Let us contrast the evidence procedure assignments with those that optimize the consistency. We note that $|\delta_c|$ is also independent of the learning temperature for $\alpha < 1$ and that the generalization error is a function of the weight decay, $\lambda = \gamma/\beta$, only and

so, in the limit $\gamma \to 0$, is independent of β. In the large α limit $T_{ev} \to T_{ev}^{\infty}$ where,

$$T_{ev}^{\infty} = 2(P_1^t \sigma_1^2 + P_2^t \sigma_2^2) + \frac{2P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w}{\sigma_{x_{\text{eff}}}^2}. \tag{2.21}$$

In the linear regime T_{ev} is constant ($\forall \alpha > 1$) as shown in figure 2.3, curve (i), whereas in the non-linear regime figure 2.3, curve (ii), shows that there are finite α effects. Furthermore, it can be shown that T_{ev} optimizes the consistency measure in the linear regime only. That is the evidence procedure optimizes the consistency measure if $\sigma_{x_1} = \sigma_{x_2}$ or if $D_w = 0$ and $\sigma_1 = \sigma_2$. The effect, on $|\delta_c|$, of setting the learning temperature to T_{ev} in the non-linear case is shown in curve (iv) of figure 2.2(c) where the optimal $|\delta_c|$ is actually zero. The learning temperature which minimizes the consistency (T_{δ_c}) is shown for this case in curve (iii) figure 2.3 (this is the same case as curve (ii) of the same figure which shows T_{ev}). In the limit $\alpha \to \infty$, T_{δ_c} becomes,

$$T_{\delta_c}^{\infty} = \frac{2(\sigma_1^2 \sigma_{x_1}^2 P_1^t + \sigma_2^2 \sigma_{x_2}^2 P_2^t)}{\sigma_{x_{\text{eff}}}^2} + \frac{2P_1^t P_2^t \sigma_{x_1}^4 \sigma_{x_2}^4 D_w}{\sigma_{x_{\text{eff}}}^6}. \tag{2.22}$$

Contrasting this with equation (2.21) above we note that $T_{\delta_c}^{\infty}$ and T_{ev}^{∞} are the same only in the linear regime.

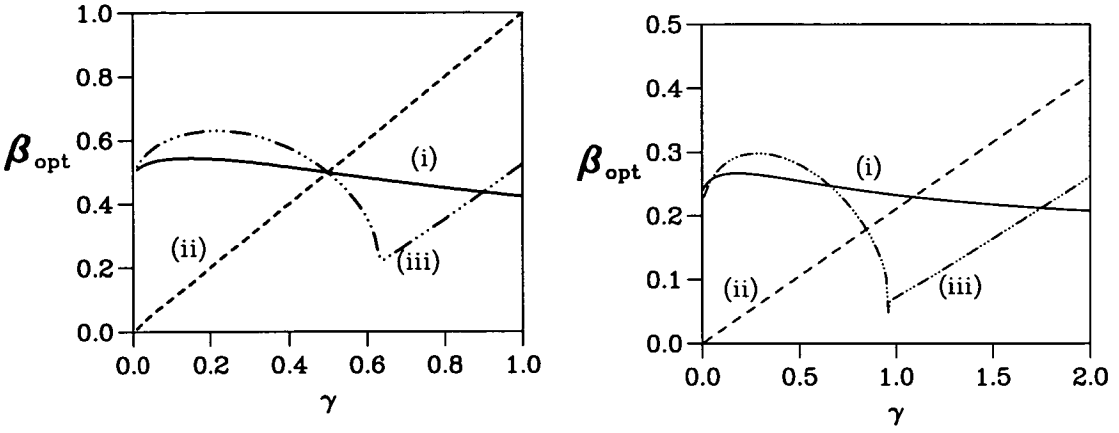
Thus, in summary, for $\gamma \to 0$ in the linear case the evidence procedure optimizes the consistency measure ($T_{ev} = T_{\delta_c} = \text{const. } \forall \alpha \text{ s.t. } \alpha > 1$). However, for a non-linear teacher or noise that varies across the input space, even in the large α limit, it does not.

The $\gamma > 0$ case

We now turn to the regularized case. In this instance in the large α limit T_{ev}^{∞} is still given by equation (2.21) above, whilst,

$$\frac{1}{\gamma_{ev}^{\infty}} = \frac{2(\sigma_{w_1}^2 \sigma_{x_1}^2 P_1^t + \sigma_{w_2}^2 \sigma_{x_2}^2 P_2^t)}{\sigma_{x_{\text{eff}}}^2} - \frac{2P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w}{\sigma_{x_{\text{eff}}}^4}. \tag{2.23}$$

These asymptotic assignments can be understood intuitively. The setting of T_{ev}^{∞} reflects the average noise on the examples ($P_1^t \sigma_1^2 + P_2^t \sigma_2^2$) and the noise due to the unlearnability, $P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w / \sigma_{x_{\text{eff}}}^2$, discussed earlier. The weight decay term is not as easy to interpret. However, in the linear regime we have $N/2\gamma_{ev}^{\infty} = |\mathbf{w}^{o1} P_1^t + \mathbf{w}^{o2} P_2^t|^2$; the variance of the prior is set to be the square of the normalized average teacher vector magnitude. Both these assignments can be considered optimal in the sense that they are the evidence estimates in the limit of infinite data.



(a) Linear Case

(b) Non-Linear Case

Figure 2.4. The evidence procedure: Optimal β v's γ. In both graphs the β which optimizes evidence, $\beta_{ev}(\gamma)$, is curve (i), that which optimizes the generalization error, $\beta_{\epsilon_g}(\gamma)$, curve (ii) and that which optimizes the consistency measure, $\beta_{\delta_c}(\gamma)$, curve (iii). The evidence procedure is the point in the γ - β plane where the evidence is maximal with respect to both β and γ. In graph (a) the evidence procedure picks the point, in the γ - β plane, where all three curves coincide. However, in the non-linear case shown in graph (b) the evidence procedure point coincides only with curve (i). In other words the evidence procedure does not optimise either the generalization error or the consistency measure.

In order to asses the evidence procedure for finite γ and α we are forced to optimize the free energy and the performance measures numerically. In addition to $\beta_{ev}(\gamma)$ we define $\beta_{\epsilon_g}(\gamma)$ and $\beta_{\delta_c}(\gamma)$ to be those assignments which optimize, for a given γ, ϵ_g and δ_c respectively.

In the learnable linear case ($D_w = 0$ and $\sigma_1 = \sigma_2$) the evidence procedure assignments of the hyper-parameters (for finite α) coincide with β_{ev}^{∞} and γ_{ev}^{∞} and also optimize ϵ_g and δ_c in agreement with Bruce and Saad (94). This is shown in figure 2.4(a) where we plot $\beta_{ev}(\gamma)$, $\beta_{\epsilon_g}(\gamma)$ and $\beta_{\delta_c}(\gamma)$. The point at which the three curves coincide is the point (in the β - γ plane) which the evidence procedure picks.

However, Bruce and Saad (94) noted that if the hyper-parameter γ is held fixed, at some sub-optimal value, then the evidence assignment of β is also sub-optimal. Similarly, performance is not optimised in this case. In fact, these observations can be considered in terms of Bayesian robustness (Berger (85)). When we fix the parameter γ we define the prior chosen. Thus, by varying γ we can examine, in a rather limited

sense, the effect of changes in the prior. Let us consider the identification of the optimal inverse noise level, β, as a reduced problem. Then we can ask how robust is the estimate of β from the evidence maxima to changes in the prior (changes in γ). Figure 2.4(a) shows this evidence estimate, $\beta_{ev}(\gamma)$, for a range of γ when the variance of the noise corrupting the teacher outputs is 0.5; the evidence estimate seems relatively robust in that it changes little for a wide range of γ [2]. Indeed, if we consider a non-informative prior on β itself, then this tells us something of the robustness of the posterior distribution of β. For the nonlinear case we will consider the robustness of the true posterior, $P(\mathbf{w} \mid \mathcal{D}, \beta, \lambda)$, in the next section.

The results for an unrealizable rule in the linear regime ($D_w > 0$ and $\sigma_{x_1} = \sigma_{x_2}$) are qualitatively the same as in figure 2.4(a), but with an increased effective noise level due to the variance of the teacher output. The evidence procedure sets $\beta = \beta_{ev}^{\infty}$, which takes into account this effective noise, and sets $\gamma = \gamma_{ev}^{\infty}$ which reflects the effective size of the weights. The evidence assignments still optimize the generalization error and the consistency measure.

The situation in the non-linear regime is shown in figure 2.4(b) where the point picked by the evidence procedure coincides only with curve (i) in this instance. That is, the parameters picked by *the evidence procedure neither minimize* ϵ_g *or* δ_c nor do they set β and γ to their asymptotic values. In fact, in analogy to the unregularized limit the evidence procedure assignments are α dependent in this regime.

Discussion

Meir and Merhav (94) found behaviour in a study of the stochastic complexity which mirrors that which we find here. In an average case setting, in the limit of large data sets, they found that minimization of the stochastic complexity resulted in optimal generalization performance in a realizable case but not in an unrealizable case. As we saw in section 2.2.1 the stochastic complexity approach to over-fitting avoidance is also a penalty based method.

As we noted in section 1.2 any Bayesian scheme must make assumptions concerning the process generating the data (*i.e.* assumptions concerning the teacher) and in general such assumptions will be at best approximations to the truth. In this chapter, in the non-linear regime, we have explicitly violated the linearity assumption of our Bayesian scheme and so perhaps it is not surprising that the evidence procedure breaks down. In fact, in the non-linear regime, if we have different real noise levels associated with

[2] We note that the evidence optimal $\gamma(\beta)$ is not shown in figure 2.4 but that it is relatively robust to changes in β.

each teacher ($\sigma_1 \neq \sigma_2$) this mismatch, between the evidence procedure assignments and those which optimize performance, increases. In this case we have not only violated the assumption that the teacher is linear but also that of our single Gaussian noise model. However, when $\sigma_{x_1} = \sigma_{x_2}$ and $D_w > 0$ then the evidence procedure is optimal despite the fact that the data is produced by a mixture of linear rules which our student can not model. In general then, it is not easy to assess the effects of the inadequacies of the modelling procedure. As we have already noted an important question is that of robustness; given that the evidence procedure does not optimise performance in the non-linear regime how far from optimality is it? We touched on this issue above in a simple case and we now pursue this by exploring the effects on performance of these sub-optimal hyper-parameter assignments.

2.6 Robustness of evidence procedure

As argued previously, the importance of the posterior distribution is ultimately its effect on performance (see section 1.2.1). Thus here, in assessing the procedure's robustness we focus on the effect of its hyper-parameter assignments on performance. Secondly, we note that in exploring robustness it is more usual, as we did above, to vary the prior. However, here we consider variation of the underlying *reality* (*i.e.* the teacher) whilst holding the prior assumptions fixed; nonetheless the results are illuminating.

Since we are interested in how far the evidence assignments are from the optimal we examine the fractional degradation in average generalization performance defined by

$$\kappa_{\epsilon_g} \equiv \frac{\epsilon_g(\lambda_{ev}) - \epsilon_g(\lambda_{opt})}{\epsilon_g(\lambda_{opt})}. \tag{2.24}$$

Where λ_{ev} is the evidence procedure assignment and λ_{opt} is the optimal weight decay. Thus, $\epsilon_g(\lambda_{opt})$ is the best possible generalization error achievable by our linear student, for a given number of examples α. The percentage degradation, $\kappa = 100\kappa_{\epsilon_g}$ is plotted in Figure 2.5(a) for different noise levels and degrees of unlearnability. We also define

$$\kappa_{\delta_c} = \frac{|\delta_c(\lambda_{ev}, \beta_{ev})|}{\epsilon_g(\lambda_{ev})}. \tag{2.25}$$

This measures the error in using the variance of the post training distribution to estimate the decaying part of the generalization error, as a percentage of the generalization

error itself, when we use the evidence procedure to set the hyper-parameters. The percentage error in estimating the generalization in this way, $\kappa_\delta = 100\kappa_{\delta_c}$, is plotted in Figure 2.5(b). There are three important points to note concerning κ and κ_δ . Firstly, the larger the deviation from a linear rule the greater is the error. Secondly, that it is the magnitude of the effective noise due to unlearnability relative to the real noise which determines this error. In other words, if the real noise is large enough to swamp the non-linearity of the rule then the evidence procedure will not be very misleading. In general, the fractional error associated with predicting the generalization performance, κ_δ , is larger than that in the error itself, κ . In fact, when the teacher deviates significantly from linearity the former becomes rather large for small α. However, the performance of the evidence assignments as expressed by κ seems remarkably robust to this misspecification of the prior assumptions. Indeed, the magnitude of, κ , for relatively large deviations from linearity is only a few percent and thus the evidence procedure might well be a reasonable, if not optimal, method for setting the hyper-parameters in this non-linear regime.

Whether one should, indeed, use the evidence procedure depends upon the alternatives to hand. Clearly, we do not have direct access to the optimal weight decay but there are alternative methods to the evidence, for instance those which attempt to estimate the generalization error, from which hyper-parameter assignments can be made. As already noted one such method is cross-validation and we will explore this and its relation to the evidence procedure in more detail in chapter 5. However, we note here that in the thermodynamic limit the cross-validation estimate of the generalization error and the generalization error itself are the same, at least when the number of test examples is not extensive (*i.e.* is lower than order $\mathcal{O}(N)$). This is because the more examples we leave-out for testing the fewer are available for training, and as we shall see in chapter 5 the cross-validation error is an unbiased estimate of the generalization error based on the number examples left for training. However, as we have seen, in the thermodynamic limit the generalization error depends on the ratio $\alpha = p/N$ where the number of examples p is $\mathcal{O}(N)$. It is clear that if we leave-out $p_0 < \mathcal{O}(N)$ examples for testing then $\alpha \to p/N - p_0/N$ is unchanged in the thermodynamic limit (*i.e.* $N \to \infty$). Therefore, even in this unlearnable scenario the hyper-parameter assignment from the cross-validation error will be optimal in the thermodynamic limit whilst that of the evidence is not. This demonstrates an important difference between penalty based methods, such as the evidence, and methods such as cross validation noted by Kearns *et al.*(95), namely that the former tend to be more problem specific; recall that the failure of the evidence that we have witnessed in this chapter resulted from a break down of the assumptions on which it was based. Finally, we note that depending on

the level of error one is prepared to tolerate it would be preferable to improve our student space to enable it to model the teacher more fully rather than attempt to squeeze the most from a sub-optimal model. Thus, as Mackay (92b) states, the failure of the evidence procedure to optimise performance is an opportunity to learn since it suggests our model is in conflict with the the truth. Furthermore, this is an opportunity not afforded by model selection based on the cross-validation error.

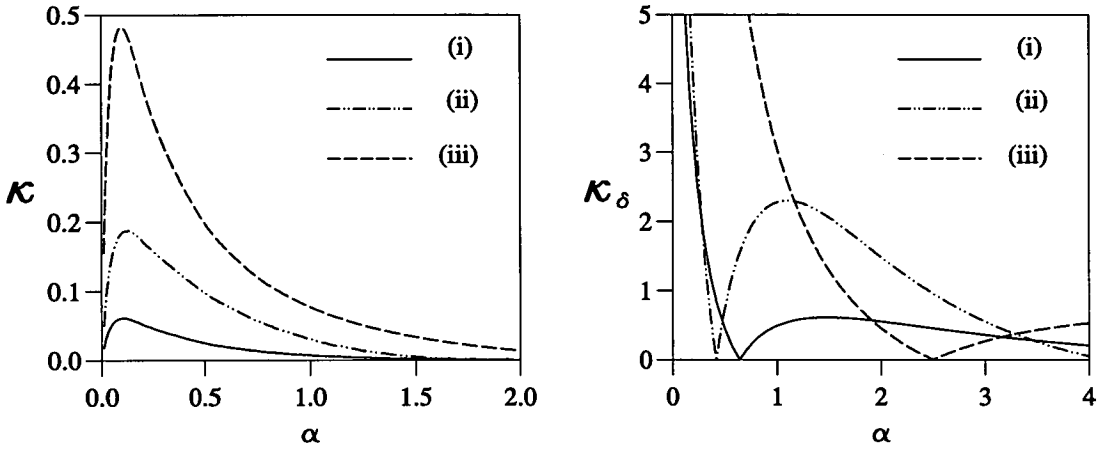


Figure 2.5. The relative degradation in performance compared to the optimal when using the evidence procedure to set the hyper-parameters. Graph (a) shows the percentage degradation in generalization performance κ . a(i) has $D_w = 1$ with the real noise level $\sigma = 1$. a(ii) has this noise level reduced to $\sigma = 0.1$ and a(iii) has increased non-linearity, $D_w = 3$, and $\sigma = 1$. Graph (b) shows the error made in predicting the decaying part of the generalization error, $\epsilon_g - \epsilon_g^{\infty}$, from the variance of the post training distribution as a percentage of the generalization error itself, κ_δ . b(i) and b(ii) have the same parameter values as a(i) and a(ii), whilst b(iii) has $D_w = 3$ and $\sigma = 0.1$

2.7 Conclusion

In this chapter we have analysed a simple system which enabled us to examine the efficacy of the evidence procedure for the case when the student was not sufficiently powerful to model the teacher. Such a situation may well arise in a real world application since we rarely know the form of the teacher and, as discussed in the introduction, learning is a trade-off between minimizing student complexity and modelling the teacher on the data set.

In particular, we have examined the generalization error, the consistency measure and the evidence procedure within a model which allows us to interpolate between a learnable scenario and an unlearnable one in which our model serves as the basis for a general piecewise-linear teacher. We have seen that the unlearnability acts as an effective noise on the examples. Furthermore, we have seen that the evidence procedure optimizes performance, even in the unlearnable case, if the average teacher output is a linear function of the input. In the case of a non-linear teacher (and a linear student) the evidence procedure breaks down in that it fails to optimize the performance measures. Furthermore, we noted that, at least in the thermodynamic limit, the hyper-parameter assignments from cross-validation were optimal. However, by examining the resulting performance we discovered that, even for quite severe mismatch between student and teacher, the evidence procedure is close to optimal, especially in terms of generalization. In other words, the evidence procedure was seen to be relatively robust against misspecifications of the prior assumptions. Whether or not such a breakdown of the evidence procedure, as witnessed here, is a generic feature of a mismatch between the hypothesis (student) and teacher, along with the consequent impact on performance are both matters for further study.

2.8 Appendix: response function for unrealisable rule

In this appendix we calculate the averages over the input distribution, $P(\mathbf{x})$, required in section 2.4.

We note that $P(\mathbf{x}) = P(\mathbf{x} \mid \Omega = 1)P_1^t + P(\mathbf{x} \mid \Omega = 2)P_2^t$ and in what follows $<< \mathbf{g} >_1>_2 = << \mathbf{g} >_2>_1 = \ll \mathbf{g} \gg$, where $< .. >_1$ and $< .. >_2$ refer to averages over the distributions $P(\mathbf{x} \mid \Omega = 1)$ and $P(\mathbf{x} \mid \Omega = 2)$ respectively.

Firstly lets rewrite $\mathbf{g}^{-1}$ as $\mathbf{g}^{-1} = \mathbf{A}_1 + \mathbf{\Gamma}$ where $\mathbf{\Gamma} = \mathbf{A}_2 + \lambda \mathbf{I}$ and $\mathbf{I}$ is the identity matrix. Now we can average over the distribution $P(\mathbf{x} \mid \Omega = 1)$. This step is similar to the calculation, in Hertz *et al.*(89), of the average of the matrix $(\mathbf{A}_1 + \lambda \mathbf{I})$ with $\lambda \mathbf{I}$ replaced by the matrix $\mathbf{\Gamma}$.

In the thermodynamic limit we obtain,

$$< \mathbf{g} >_1 = (\mathbf{\Gamma} + \Sigma_1 \mathbf{I})^{-1} \qquad \text{where,} \quad \Sigma_\Omega = \frac{\alpha P_\Omega^t \sigma_{x_\Omega}^2}{1 + \sigma_{x_\Omega}^2 \frac{1}{N} < \text{tr } \mathbf{g} >_\Omega}. \tag{2.26}$$

We can then rewrite (2.26) as

$$< \mathbf{g} >_1 \mathbf{A}_2 = \mathbf{I} - \lambda < \mathbf{g} >_1 - < \mathbf{g} >_1 \Sigma_1. \tag{2.27}$$

Now we wish to perform the average over the second distribution $P(\mathbf{x} \mid \Omega = 2)$ but the last term in the expression (2.27) is potentially problematic. However, if following the diagrammatic method of Hertz *et al.*(89) we examine the diagrams for this term we see that the 'crossings', or *interactions*, between $< \mathbf{g} >_1$ and Σ_1 are $\mathcal{O}(1/N^2)$ and can be ignored in the thermodynamic limit. Thus, we can average the two factors independently, ignoring any interaction between them. This leads to

$$\ll \mathbf{g}\mathbf{A}_2 \gg = \mathbf{I} - \lambda \ll \mathbf{g} \gg - \ll \mathbf{g} \gg \ll \Sigma_1 \gg . \tag{2.28}$$

Using the matrix identity $\mathbf{g}\mathbf{A}_2 = \mathbf{I} - \lambda\mathbf{g} - \mathbf{A}_1\mathbf{g}$ and defining $\Psi_1 =< \Sigma_1 >_2$ we obtain $\ll \mathbf{g}\mathbf{A}_1 \gg = \Psi_1 \ll \mathbf{g} \gg$. If we perform these averages the other way around and define $\Psi_2 =< \Sigma_2 >_1$ we find the analogous expression. Thus, in general we have

$$\ll \mathbf{g}\mathbf{A}_\Omega \gg = \Psi_\Omega \ll \mathbf{g} \gg . \tag{2.29}$$

Now by multiplying equation (2.27) by $\mathbf{A}_2$, averaging over the distribution $P(\mathbf{x} \mid \Omega = 2)$ and using the matrix identity $\mathbf{A}_2\,\mathbf{g}\mathbf{A}_2 = \mathbf{A}_2 - \lambda\mathbf{A}_2\,\mathbf{g} - \mathbf{A}_2\,\mathbf{g}\mathbf{A}_1$ we obtain

$$\ll \mathbf{A}_2\,\mathbf{g}\mathbf{A}_1 \gg = \Psi_1\Psi_2 \ll \mathbf{g} \gg . \tag{2.30}$$

We now have all the averages we require in terms of the average $\ll \mathbf{g} \gg$. To evaluate this quantity, firstly, we average the matrix identity $\mathbf{g}\mathbf{A}_1 = \mathbf{I} - \lambda\mathbf{g} - \mathbf{A}_2\,\mathbf{g}$ which gives us,

$$(\Psi_1 + \Psi_2 + \lambda) \ll \mathbf{g} \gg = \mathbf{I}. \tag{2.31}$$

This shows that $\ll \mathbf{g} \gg$ is diagonal, in this case, where the distributions $P(\mathbf{x} \mid \Omega)$ are normal and have zero mean. Taking the trace gives us an implicit equation for the response function $G = \frac{1}{N} \ll \mathrm{tr}\ \mathbf{g} \gg$. Namely,

$$G^{-1} = \lambda + \frac{\alpha P_1^t \sigma_{x_1}^2}{1 + \sigma_{x_1}^2 G} + \frac{\alpha P_2^t \sigma_{x_2}^2}{1 + \sigma_{x_2}^2 G}, \tag{2.32}$$

which resolves into a cubic in G. Now since the variance of the student output over the post training distribution is $\sigma_{x_{\mathrm{eff}}}^2 G/2\beta$ then G must be positive. Fortunately, we can show that only one of the three solutions, to the cubic, is positive. We also note here that G could be calculated using the more general method of Sollich (94a).

Chapter 3

Over-Realisable - the case of clever students

Abstract

In this chapter we consider the validity of the evidence procedure for the case where the student is more powerful than the teacher. To this end we explore a simple model of over-realizability, namely a piece-wise linear student trained and tested on examples generated by a linear teacher. Examination, in the thermodynamic limit, of this model reveals that although the generalization error is increased in comparison with a simple linear student, the evidence procedure remains optimal. This is in stark contrast to the effect of unrealizability examined in the previous chapter. Further, our results suggest that the evidence assignments will also be optimal, in the thermodynamic limit, for a linear teacher and a student analogue of the n-teacher introduced in chapter 2.

3.1 Introduction

In the previous chapter we saw that when the assumptions of our Bayesian scheme were violated the optimality of the evidence procedure was questionable. However, in the cases examined there the student was in some sense less powerful than the teacher and this unrealizability was seen to account for the sub-optimality of the evidence procedure. Given this state of affairs one might attempt to learn the unknown teacher with a very powerful student so as to avoid the unlearnable case. In the case of neural networks and probably all semi-parametric methods this strategy is likely to prove far too computationally expensive in general due to the large number of basis functions (*e.g.* number of hidden units in an **MLP**) required and may also result in degraded generalization performance due to over-fitting. However, some work has been done

to allow the application of powerful networks (Neal (94)) provided that appropriate priors are applied. This begs the question as to whether maximum-likelihood methods such as the evidence procedure are optimal in the over-realisable case.

In this chapter we attempt to answer this question by studying a simple case of over-realizability in relation to the optimality of the evidence procedure. Perhaps the simplest possible case of over-realizability would be to take account of more inputs than necessary. In an experimental setting this would correspond to measuring attributes of the system which had no impact on the quantity one is attempting to predict. For a linear student and teacher this means having more student weights (and thus, inputs) than teacher weights. In this case it is straight-forward to show that, for *i.i.d* inputs and noise in the thermodynamic limit, this over-realizability has no ill-effects on the validity of the evidence procedure or on the generalization performance.

In fact, for the analogous (linear) case of *unrealizability*, namely having too few student parameters, the optimality of the evidence procedure is also not compromised, although the generalization ability was degraded. This is because, in this case, the extra inputs in the teacher simply appear as noise to the student. However, in chapter 2, for a more serious form of unrealizability we found the evidence procedure to be sub-optimal. The question is then, will a similar form of over-realizability degrade performance?

3.2 Piece-wise linear student

In this section we introduce a piece-wise linear student in which each linear component takes *responsibility* for modelling a region of the input space. Generically we write the output, $y_s(\mathbf{x})$, of this student given an N dimensional input $\mathbf{x}$ as;

$$y_s(\mathbf{x}) = \sum_{k=1}^{n} \frac{1}{\sqrt{N}} \mathbf{w}^k \cdot \mathbf{x}\ \delta_k(\mathbf{x}) \tag{3.1}$$

where the functions $\delta_k(\mathbf{x}) = 1$ on some region (or zone), $\mathcal{Z}_k$, of the input space and are zero everywhere else. In what follows we assume that these regions are non-overlapping. The parameters $\mathbf{w}^k$ are the weights of the k^{th} component student. The model, thus, defines a piece-wise linear function on the regions $\mathcal{Z}_k$.

An alternative model we might have adopted for our student is the n-'teacher' model we examined in the previous chapter. Here we would have weighted the k^{th} linear component with a Gaussian probability on input space, with mean a_k and variance σ_k^2. As before, in the thermodynamic limit we would obtain a piece-wise linear function with each of the n components covering a hyper-sphere. The 'hard' constraints used in

equation (3.1) ensure a piece-wise linear function for finite input dimension N.

In the remainder of this chapter we will consider the student defined in equation (3.1) learning from and being tested on examples generated by a teacher which is assumed to be a linear function corrupted by zero mean Gaussian noise of variance σ^2. We, therefore, represent the teacher by the conditional probability,

$$p(y_t \mid \mathbf{x}) \propto \exp -[y_t - \mathbf{w}^o \cdot \mathbf{x}/\sqrt{N}]^2/2\sigma^2. \tag{3.2}$$

As before we are supplied only with a data set $\mathcal{D}$ consisting of p pairs of inputs and outputs sampled from the distribution $p(y_t \mid \mathbf{x})p(\mathbf{x})$, where $p(\mathbf{x})$ is our sampling assumption.

3.3 Hyper-parameters and priors

We now have n student weight vectors (each of N- dimensions) which we wish to regularize and we will again examine the case of weight decay. We also adopt the noise model of chapter 2 and thus, the composite cost function of section 2.2.1 now becomes $\beta E_{\mathbf{w}}(\mathcal{D}) + \sum_k \gamma_k C(\mathbf{w}^k)$, where $E_{\mathbf{w}}(\mathcal{D})$ is the quadratic sum of errors and the cost function $C(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$. In Bayesian terms this straightforwardly translates into priors on the student weights,

$$P(\mathbf{w}^k \mid \gamma_k) \propto e^{-\gamma_k \mathbf{w}^k \cdot \mathbf{w}^k}. \tag{3.3}$$

and a Gaussian noise model $P(\mathcal{D} \mid \beta, \mathbf{w}) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D})}$, as before. The evidence then becomes,

$$P(\mathcal{D} \mid \mathcal{M}_n) = \int \prod_k d\mathbf{w}^k P(\mathbf{w}^k \mid \gamma_k) P(\mathcal{D} \mid \beta, \{\mathbf{w}^k\}) \tag{3.4}$$

Recall that the notation $\mathcal{M}$ denotes the model architecture and learning algorithm. In particular, here we have $\mathcal{M}_n = \{n, \{\gamma_k : k = 1..n\}, \beta\}$. As before (see section 2.3) we define the free energy $f(\mathcal{D}) = -\frac{1}{N} \ln P(\mathcal{D} \mid \mathcal{M}_n)$. However, since here the teacher is essentially deterministic we adopt the error function of equation (1.11) and use the average of the generalization error, thus defined (see equation 1.12), as our performance measure,

$$\epsilon_g = \langle (\langle y_t(\mathbf{x}) \rangle_{P(y_t|\mathbf{x})} - \langle y_s(\mathbf{x}) \rangle_{\mathbf{w}})^2 \rangle_{P(\mathbf{x})P(\mathcal{D})} \ . \tag{3.5}$$

The consistency measure, δ_c (see equation 2.7) remains unchanged.

3.4 Calculating average behaviour

Using the rationale of chapter 2, we expect the system to be self averaging, and will now focus on the average case behaviour in the thermodynamic limit deferring, once again, a discussion of the data dependent behaviour until chapter 4. We can now calculate the average free energy explicitly.

Performing the Gaussian integrals and taking the logarithm is straightforward, since the integrals over the weights, $\mathbf{w}^k$, of the different components factorize. The integration of these individual factors can then be performed as shown in appendix 4.8.1. This factorization occurs because the regions $\mathcal{Z}_k$ are non-overlapping. We are then free to average over the additive Gaussian noise on the examples introduced in equation (3.2). Also, since the problem is isotropic, at least as far as the examples, $\mathcal{D}$, are concerned, we are free to average over the possible directions of the teacher vector $\mathbf{w}^o$. In other words we can average $\mathbf{w}^o$ over a spherical distribution since only its length $\sigma_w^2 = \mathbf{w}\cdot\mathbf{w}/N$ will be relevant. Neglecting an additive constant, these steps lead to the noise averaged free energy,

$$f(\{\mathbf{x}^\mu\}) = -\frac{\alpha}{2}\ln\beta + \alpha\beta\sigma^2 - \sum_{k=1}^{n}\left(\frac{1}{2}\ln\lambda_k + \frac{1}{2N}\ln\det(\mathbf{g}_k) + \Xi_k\right) \tag{3.6}$$

Where,

$$\Xi_k = \beta\sigma_w^2\left(\lambda_k^2\frac{1}{N}\mathrm{tr}\ \mathbf{g}_k - \lambda_k\right) + \beta\sigma^2\left(1 - \lambda_k\frac{1}{N}\mathrm{tr}\ \mathbf{g}_k\right)$$

and in analogy to $\mathbf{g}$ and λ defined in section 2.3 we have defined,

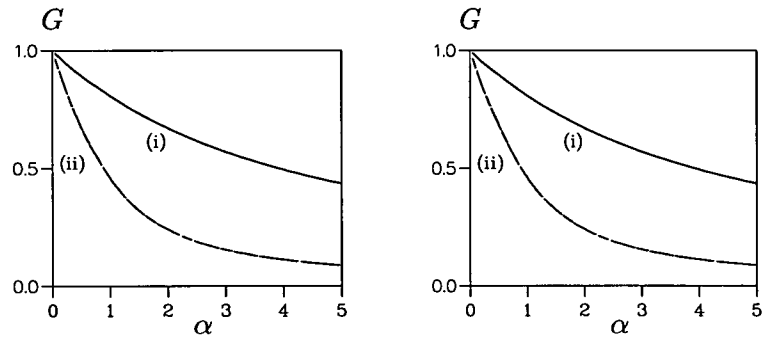
$$\mathbf{g}_k = (\mathbf{A}_k + \lambda_k \mathrm{I})^{-1}$$

$$\text{with,} \quad \lambda_k = \frac{\gamma_k}{\beta} \quad \text{and} \quad A_k = \frac{1}{N}\sum_{\mu=1}^{p}(\mathbf{x}^\mu)^T\mathbf{x}^\mu\delta_k(\mathbf{x}^\mu) \tag{3.7}$$

whilst the definition $\alpha = p/N$ is unchanged. The important point to note concerning equation (3.6) is that since the regions $\mathcal{Z}_k$ are non-overlapping, the last term is a linear sum of terms attributable to each component k of the student. The individual terms in this sum are analogous to the corresponding term in the free energy for the learnable linear case (equation 4.A8). Note also that if we have only one component we then recover the linear case (*i.e.* , the free energy of equation (2.11) if the teacher is linear).

We will demonstrate the main characteristics of this learning scenario by examining

a simplified version. Firstly, we will assume that our student is made up of only 2 linear components ($n = 2$) and secondly that the regions $\mathcal{Z}_k$, $k = 1, 2$, on which they are defined are as follows. The *inner* region $\mathcal{Z}_1$ is a hyper-sphere centred on the origin with nominal radius c. The *outer* region $\mathcal{Z}_2$ is the remaining volume in the N dimensional input space. Thus, the first component $\mathbf{w}^1$ is responsible for modelling the inner region where $\mid \mathbf{x} \mid < c$, similarly, the second component $\mathbf{w}^2$ models the outer region where $\mid \mathbf{x} \mid \geq c$. Finally we take our sampling assumption, $p(\mathbf{x})$, to be normally distributed with variance σ_x^2 and zero mean. In fact, as we will concentrate on the thermodynamic limit we could equally well assume that the inputs are drawn independently at random from any distribution $p(\mathbf{x})$ with zero mean and variance σ_x^2 (*i.e.* , i.i.d. inputs).



(a) $\mathcal{V}_1 = 0.2$ (b) $\mathcal{V}_1 = 0.75$

Figure 3.1. Response functions versus α: The thermodynamic value plotted with the average value calculated from simulations for a system of size $N = 20$ and the fractional volumes occupied by the inner linear segment, $\mathcal{V}_1$, as indicated. In both (a) and (b): (i) is the *inner* response function G_1 and (ii) the *outer* response G_2. In all cases shown the experimental and theoretical curves are indistinguishable with standard error bars too small to be visible on the scale of the figure.

3.4.1 Average response function

We can now calculate the average in the thermodynamic limit, over the sampling distribution $P(\mathbf{x})$, of the quantities tr $\mathbf{g}_k$. We define these averages as the response functions $NG_k = \langle \text{tr } \mathbf{g}_k \rangle_{P(\mathbf{x})}$, calculation of which will enable us to explicitly examine the performance of our student.

We calculate these averages following the method of (Sollich 94), details of which are presented in appendix 3.8. The crucial point to note is that our decomposition of the space into shells centred on the origin allows us to calculate the average response functions analytically and show they are self averaging. For the case considered here

with a two component student defined on $\mathcal{Z}_k$, the inner and outer response functions in the thermodynamic limit are given by,

$$G_k = \frac{-\lambda_k - (\alpha-1)\sigma_x^2\mathcal{V}_k + \sqrt{(\lambda_k + (\alpha-1)\sigma_x^2\mathcal{V}_k)^2 + 4\lambda_k\sigma_x^2\mathcal{V}_k}}{2\lambda_k\sigma_x^2\mathcal{V}_k} \tag{3.8}$$

where we have defined the fractional volume occupied by component k as,

$$\begin{aligned} \mathcal{V}_1 &= \sigma_x^{-(N+2)}(2\pi)^{-N/2}\int_{\mathcal{Z}_k} \mathbf{x}^T\mathbf{x}\exp\left(\frac{-\mathbf{x}^T\mathbf{x}}{2\sigma_x^2}\right)d\mathbf{x} \\ \mathcal{V}_2 &= 1 - \mathcal{V}_1 \end{aligned} \tag{3.9}$$

Examples of these response functions are plotted in figure 3.1. To demonstrate the validity of the results for finite system size we have also plotted the results of Monte Carlo averages of the response functions for a system with $N = 20$. As indicated in figure 3.1, the difference between these *experimental* values and the thermodynamic results is not distinguishable on the scale of the graphs. Thus, in terms of the response function, which is an average quantity, the thermodynamic limit is good approximation for relatively small systems. Armed with these results we can examine the performance of our student.

3.5 Generalization performance

Following appendix 4.8.1 (see equations 4.A9 and 4.A10) it can be shown that the average generalization error of our piece-wise linear student is given by,

$$\epsilon_g = \sum_{k=1}^{n}\sigma_x^2\mathcal{V}_k\left[\frac{\sigma^2}{N}\frac{\partial}{\partial\lambda_k}(\lambda_k\langle\text{tr } \mathbf{g}_k\rangle_{P(\mathbf{x})}) - \frac{\sigma_w^2}{N}\lambda_k^2\frac{\partial}{\partial\lambda_k}\langle\text{tr } \mathbf{g}_k\rangle_{P(\mathbf{x})}\right] \tag{3.10}$$

where we have explicitly averaged over the noise and the teacher direction as before. Here we have written the generalization error for an arbitrary number of components, n, and the fractional volumes $\mathcal{V}_k$ will not, in general, be those defined in equation (3.9). Similarly, (see equation 4.A11) one can write the consistency measure as,

$$\delta_c = \sum_{k=1}^{n}\frac{\sigma_x^2\mathcal{V}_k}{2N\beta}\langle\text{tr } \mathbf{g}_k\rangle_{P(\mathbf{x})} - \epsilon_g \tag{3.11}$$

Calculating the optimal hyper-parameters from equations (3.10) and (3.11), we find that the generalization error is optimised by $\lambda_k = \sigma^2/\sigma_w^2$ whilst the consistency is zero for these weight decay assignments if $\beta = 1/2\sigma^2$. We emphasise that these results are

average case results valid for any N, p and n.

Setting $n = 2$ and replacing $\langle \text{tr } \mathbf{g}_k \rangle_{P(\mathbf{x})}$ in equation (3.10) by the thermodynamic average value G_k, as defined in equation (3.8), allows one to explore the generalization performance in the large N limit. In figure 3.2(a) we compare this average optimal[1] generalization for our piece-wise linear student learning a linear teacher against that for a linear student learning the same teacher. As one might expect, we find that the generalization error is larger for the piece-wise linear student. This is wholly understandable as there are more parameters to fit for this model. However, in the limit of noise free examples and optimal weight decay, this difference vanishes and the generalization error decays according to $1 - \alpha$ for both models. This is because, although in the piece-wise linear case, to fix the weights, we have $2N$ equations to solve, when the examples are noise free we have two equivalent sets of N equations (and thus enough information to fix the weights) for $\alpha = 1$. A more intuitive explanation is that the difference, in terms of generalization error, between the $n = 1$ and $n = 2$ models is a reflection of the over-fitting of the noise by the more powerful model. Thus, in the noiseless case they are equivalent because there is no noise to over-fit. In the asymptotic limit we find that,

$$\begin{aligned} \lim_{\alpha \to \infty} \epsilon_g(n = 1) &= \frac{\sigma^2}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right) \\ \lim_{\alpha \to \infty} \epsilon_g(n = 2) &= \frac{2\sigma^2}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right). \end{aligned} \tag{3.12}$$

Thus, although asymptotically both students achieve zero generalization error, to second order in α, that of the $n = 2$ student is twice that of the linear student.

Intuitively one might think that an optimal solution would be to set all but one of the weight decays to infinity, effectively 'killing off' the associated student components. However, this is not the case as each component is responsible for only a fraction of the input space. In the $n = 2$ case it is easy to show that

$$\begin{aligned} \lim_{\lambda_2 \to \infty} \epsilon_g(\lambda_0, \lambda_2) &= \sigma_x^2(\sigma^2 \mathcal{V}_1 G_1 + \sigma_w^2 \mathcal{V}_2) \\ \epsilon_g(\lambda_0, \lambda_0) &= \Sigma_{k=1}^2 \sigma_x^2 \sigma^2 \mathcal{V}_k G_k. \end{aligned} \tag{3.13}$$

In fact, the solution $\lambda_1 = \lambda_0, \lambda_2 \to \infty$ is a saddle point of the generalization error, whilst the solution $\lambda_k = \lambda_0$ is a true minima. In addition, using the fact that $\lambda G_k(\lambda) < 1$, it can be shown that, $\lim_{\lambda_2 \to \infty} \epsilon_g(\lambda_0, \lambda_2) > \epsilon_g(\lambda_0, \lambda_0)$ for all α. Thus, the true optimal

[1] That is when the weight decay has been optimised

true optimal weight decay setting is λ_0 for both linear segments of the student.

Again in the $n = 2$ case, for $\alpha = 1$ and signal to noise ratio of unity figure 3.2(b) shows, as a function of $\mathcal{V}_1$, the difference between the generalization error of our piece-wise linear student and that of a linear student as a fraction of the latter. The weight decays in each model have been set optimally. The figure reveals that the deviation from the 'optimal' student (*i.e.* the linear one) is maximal when $\mathcal{V}_1 = \mathcal{V}_2 = 1/2$ and vanishes when either component is responsible for the whole input space (*i.e.* when the piece-wise linear student is effectively linear).

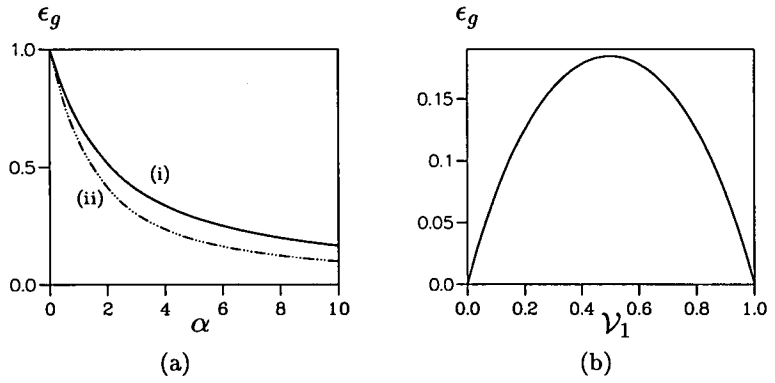


Figure 3.2. Comparison of generalization errors: for linear and piece-wise linear student learning a linear teacher. Graph (a) shows the case where the noise to signal ratio is one and the weight decay, λ, has been set optimally; (i) for a piece-wise linear student with $n = 2$ and a cut-off $c = 1$ as described in the text. (ii) a linear student. As one would expect the generalization error of the piece-wise linear student is worse than that of the linear student. Graph (b) shows the difference between these two generalization errors as a fraction of the linear students error versus the fractional volume $\mathcal{V}_1$ occupied by the inner region for $\alpha = 0.5$. The fractional increase in generalization error is maximal when the two components are responsible for equal volumes of input space ($\mathcal{V}_1 = 0.5$) and the linear case is recovered for $\mathcal{V}_k \to 1$.

3.6 Optimality of evidence assignments

Finally, we consider the evidence procedure assignments. Using the obvious extension to our definition of the evidence procedure point (*i.e.* simultaneously minimize f *w.r.t.* $\lambda_k : k = 1...n$ and β), from equation (3.6), it is straight forward to show that the evidence optimal assignments are $\lambda_k = \sigma^2/\sigma_w^2$ and $\beta = 1/2\sigma^2$. Thus, as in the case of a linear student, the evidence assignments correspond to those that optimise the generalization error and the consistency measure. Again, this statement is valid for all N,

p and n but, as we shall see in chapter 4, such average case results can be misleading. However, if self-averaging holds, as it should for any finite number of partitions n then this average case result will be *representative in the thermodynamic limit.* Furthermore, since the crucial step in the calculation is that the n components of the student decouple, we note here that this result will also hold, in the thermodynamic limit, for any student represented by the Gaussian mixture model discussed in section 3.2. That is, the evidence assignments will be optimal, and although the learning curves will not be identical, they should be qualitatively similar.

As we have seen, the $n = 2$ model allows explicit calculation of the generalization performance. Similarly, using the result of appendix 6.6, we can also calculate the free energy in this case. Doing so we find that the solution $\lambda_k = \lambda_0$ is a true minima of the free energy. In contrast, the symmetry breaking solution ($\lambda_1 = \lambda_0, \lambda_2 \to \infty$) is not a true minima, but a saddle point of the free energy; recall that the generalization error is minimised by the symmetric solution. In addition numerical exploration of the free energy shows that, of the two, the symmetric solution is associated with the lowest free energy. Thus, we note that in the $n = 2$ case, we have found that the generalization ability of the over-parametrised model is worse than the correct model even when both are optimally regularized (see figure 3.2). This result suggests that the application of overly powerful networks to a learning task (*e.g.* Neal (94)), even when optimally regularized, may result in sub-optimal generalization performance.

3.7 Summary

We have considered an over-realisable learning scenario in which the student is more than able to mimic its teacher. As an example we examined a piece-wise linear student trained and tested on examples generated by a linear teacher. For a particular instance of this student ($n = 2$) we found that in the thermodynamic limit this over-realizability caused over-fitting leading to an increase in the generalization error, at least for noisy data, but that nonetheless, the evidence procedure was found to optimise the generalization error and the consistency measure. The former result may have relevance to the work of Neal (94). In fact, the evidence assignments were found to be optimal in an average sense for any finite number of components n. In addition, we deduced that in the thermodynamic limit this statement would also hold if our student was the mixture model described in section 3.2. Thus, whilst we have not shown in general that the evidence procedure will be optimal in an over-realisable scenario our results suggest that un-realizability has a greater impact in this regard.

EX BIBL. UNIV. EDINBURGEN.

3.8 Appendix: Response function

In this appendix we outline the calculation of the response function given in equation (3.8). This derivation follows the method of Sollich (94) and as described there we can write the matrix $\mathbf{g}_k(p+1) = (\mathbf{A}_k + \lambda_k \mathbf{I})^{-1}$ based on $p+1$ examples (see equation 3.6) in terms of the matrix $\mathbf{g}_k(p)$,

$$\mathbf{g}_k(p+1) = \mathbf{g}_k(p) - \frac{1}{N}\frac{\delta_k(\mathbf{x})\mathbf{g}_k(p)\mathbf{x}\mathbf{x}^T\mathbf{g}_k(p)}{1+\frac{1}{N}\delta_k(\mathbf{x})\mathbf{x}^T\mathbf{g}_k(p)\mathbf{x}} \quad . \tag{3.14}$$

Here the $(p+1)^{\text{th}}$ input example is denoted by $\mathbf{x}$. Since the matrix $\mathbf{g}_k(p)$ is independent of this example we are then free to average over $\mathbf{x}$. Given our choice of constraints $\delta_k(\mathbf{x})$ such that we have two concentric regions $\mathcal{Z}_k$ centred on the origin we can write the average $\langle \delta_k(\mathbf{x})\mathbf{x}^T\mathbf{x}\rangle_{P(\mathbf{x})} = \sigma_x^2\mathcal{V}_k\mathbf{I}$ where the $\mathcal{V}_k$ are defined in equation (3.9). Taking the trace of equation (3.14) and averaging over the example $\mathbf{x}$ it is then straight forward to show that,

$$\frac{1}{N}\text{tr } \mathbf{g}_k(p+1) = \frac{1}{N}\text{tr } \mathbf{g}_k(p) - \frac{1}{N}\frac{\frac{\sigma_x^2\mathcal{V}_k}{N}\text{tr } \mathbf{g}_k^2(p)}{1+\frac{\sigma_x^2\mathcal{V}_k}{N}\text{tr } \mathbf{g}_k(p)\mathbf{x}} + \mathcal{O}\left(\frac{1}{N^{3/2}}\right) \quad . \tag{3.15}$$

The response functions G_k are thus self averaging. Recalling that $\alpha = p/N$ we can re-write equation (3.15) in the thermodynamic limit as

$$\frac{\partial G_k}{\partial \alpha} - \frac{\partial G_k}{\partial \lambda_k}\frac{\sigma_x^2\mathcal{V}_k}{1+\sigma_x^2\mathcal{V}_k G_k} \quad . \tag{3.16}$$

Given the initial condition $G_k \mid_{\alpha=0} = 1/\lambda_k$ this partial differential equation may then be solved using the method of characteristic curves (see *e.g.* John (78)). The resulting solution then leads to the form of the response function given in equation (3.8).

Chapter 4

Finite Size Effects in Bayesian Model Selection and Generalization

Abstract

In this chapter we show that in supervised learning from a supplied data set Bayesian model selection, based on the evidence, does not optimise generalization performance even for a learnable linear problem. This is demonstrated by examining the finite size effects in hyper-parameter assignment from the evidence procedure and the resultant generalization performance. Our approach demonstrates the weakness of average case and asymptotic analyses. Using simulations we corroborate our analytic results and examine an alternative model selection criterion, namely cross-validation. This numerical study shows that the cross-validation hyper-parameter estimates correlate more strongly than those of the evidence with optimal performance. However, we show that for a sufficiently large input dimension the evidence procedure could provide a reliable alternative to the more computationally expensive cross-validation.

4.1 Introduction

As noted previously a major advantage of the statistical physics studies of learning and generalization over the usual approach in the statistics community is that one can examine the situation where the fraction (α) of the number of examples (p) to the number of free parameters (N) is finite (see chapters 1 and 2 or for example, Krogh and Hertz (92), Seung *et al* (92), Watkin *et al.*(93)). This contrasts with the asymptotic (in α) treatments found in the statistics literature (see *e.g.* Plutowski *et*

al.(94),Stone (77a),(77b), Shao (93), Gelfand and Dey (94)). However, one draw-back of the statistical physics approach is that it is based on the thermodynamic limit where one allows N and p to approach infinity whilst keeping α constant. Naturally, as noted before, this limits the applicability of these theoretical results to the real world. In this chapter we address this problem by calculating the first order correction to the thermodynamic limit, that is we explore the finite size effects. Finite size effects in supervised learning have been studied previously by Sollich (94) and Barber *et al.*(95). We show that in the problem studied here, namely Bayesian model selection based on the evidence, conclusions drawn from the thermodynamic results are qualitatively at odds with the finite size behaviour. In addition, we also show that average case and asymptotic ($p \to \infty$) results are also misleading if applied to particular instances of finite sized data sets.

The supervised learning scenario we consider here, is that introduced in chapter 1. Thus, we are presented with a set of data $\mathcal{D}$= $\{(y_t(\mathbf{x}^\mu), \mathbf{x}^\mu) : \mu = 1..p\}$ consisting of p examples of an otherwise unknown *teacher* mapping denoted by the conditional distribution, $P(y_t \mid \mathbf{x})$, of its one dimensional output, y_t. Furthermore, we assume that the N dimensional input space is sampled with probability $P(\mathbf{x})$ and thus, the data set is generated with probability $P(\mathcal{D}) = \prod_{\mu=1}^{p} P(y_t \mid \mathbf{x}^\mu)P(\mathbf{x}^\mu)$. The learning task is to use the data base, $\mathcal{D}$, to set the N_s parameters $\mathbf{w}$ of some model (or student), with output $y_t(\mathbf{x})$, such that it *learns* to mimic the underlying mapping as closely as possible, a popular measure of this performance is the *generalization error*. Often, and that which we consider here, the training process consists of minimising a weighted sum, $\beta E_{\mathbf{w}}(\mathcal{D}) + \gamma C(\mathbf{w})$ of the quadratic error of the student on the examples, $E_{\mathbf{w}}(\mathcal{D})$, and some *cost function*, $C(\mathbf{w})$, which penalises over complex models. As we saw in chapter 2, provided γ is non-zero this serves to alleviate the problem of *over-fitting*. Once again it is the setting of the hyper-parameters β and γ which we will examine in this chapter.

As we have noted, Langevin type dynamics on the gradient of the penalized cost function, $\beta E_{\mathbf{w}}(\mathcal{D})$
$+\gamma C(\mathbf{w})$, results in a Gibbs form for the posterior distribution of student parameters, (*i.e.* the post training distribution of section 2.2.1). If we wish to make a prediction at a novel input using the average, or the maximum, of this distribution then this prediction depends solely on the hyper-parameters. Thus, as we saw in section 2.2.1 the selection of β and γ can be regarded as a model selection. In terms of practical methods for hyper-parameter assignment there are essentially two choices. Firstly one can attempt to estimate the generalization error (*e.g.* by cross-validation) and then optimise this measure with respect to the hyper-parameters. However, such an

approach can be computationally expensive. Secondly, one can optimise some other measure and hope that the resulting assignments produce low generalization error. In particular, as advocated by MacKay (92) and others, we have and will continue to use the evidence as such a measure.

In chapters 2 and 3, building on the study of Bruce and Saad (94), we have explored the evidence procedure in relation to performance for both under and over realizable cases. However, these results pertain to the thermodynamic limit and are average case analyses. In section 2.4 we argued that this approach was valid because the relevant quantities were self averaging. Thus, in the thermodynamic limit average case results apply to a particular (that is every) data set. In this chapter we seek to explore the difference between the average case analysis and the data dependent behaviour.

In the learnable linear case considered by Bruce and Saad (94) it was found that optimising the average, over all possible data sets $\mathcal{D}$, of the log evidence with respect to the hyper-parameters optimises the average generalization error in the thermodynamic limit. In addition, Meir and Merhav (94) have studied the stochastic complexity in an average case setting in the asymptotic limit ($p \to \infty$). These authors demonstrated, in a realisable scenario, that minimization of the stochastic complexity optimised hyper-parameter assignment. Here we examine hyper-parameter assignment using the evidence *based on an individual data set*, in the learnable linear case for a finite system size. That is we avoid the extremes of both infinite system size and infinite data set.

Our standpoint can be summarised as follows. In any real experiment a single set of data is available for training and one seeks to optimise performance based on this data set alone. The optimal policy (*e.g.* those hyper-parameter assignments which minimise the generalization error) will fluctuate from data set to data set, as will policies based on the evidence and the cross-validation error. What is of interest is how close our chosen strategy is to the optimal for the particular set of data in question. It is clear that average case analyses and measures of average performance do not reveal this. Thus, in section 4.2.2 we define data dependent measures of performance and then subsequently explore the performance of the evidence assignments in relation to them. In addition, we also briefly consider the average case showing that such an analysis is in general highly misleading. However, we note that in the thermodynamic limit, if self averaging holds, then both approaches are equivalent.

In the next section we review the evidence framework and the performance measures we will deal with. In section 4.3, we show that the evidence procedure is unbiased, mean square consistent and, employing some of the results of Sollich (94), we demonstrate that for large N the variances over data sets of the evidence and generalization error are $\mathcal{O}(1/N)$, that is the system is self averaging. In sections 4.4 and 4.5 we examine

hyper-parameter assignment from the evidence based on a particular data set. We calculate the variances of these assignments and those made from the generalization error showing them to be $\mathcal{O}(1/N)$. In addition, we numerically explore, in some detail, the hyper-parameter assignments in very small systems ($N = 1$, $N = 2$). First order corrections to the performance measures show that in general the evidence procedure does not lead to optimal performance. From these corrections we estimate a lower bound on the system size necessary for the evidence procedure to give reliable results. Also in terms of performance, we explore the relative importance of fluctuations in the optimal and in the evidence procedure assignments. Finally, in section 4.6 we corroborate these conclusions using a numerical study which, furthermore, reveals that cross-validation is a superior model selection criterion to the evidence for small linear systems.

4.2 Objective functions

4.2.1 The evidence

Following section 2.2.1 we denote the model specification solely in terms of the hyper-parameters, $\mathcal{M} = \{\beta, \gamma\}$. Then since $E_{\mathbf{w}}(\mathcal{D})$ is the sum squared error, if we assume that our data is corrupted by Gaussian noise with variance $1/2\beta$, the probability, or *likelihood* of the data($\mathcal{D}$) being produced given the model parameters $\mathbf{w}$ and β is $P(\mathcal{D} \mid \beta, \mathbf{w}) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D})}$. Also following section 2.2.1 the complexity cost can also be incorporated into this Bayesian scheme by assuming the *a priori* probability of a rule is weighted against 'complex' rules, $P(\mathbf{w} \mid \gamma) \propto e^{-\gamma C(\mathbf{w})}$. Multiplying the likelihood and the prior together we obtain the posterior density of student parameters, $P(\mathbf{w} \mid \mathcal{D}, \gamma, \beta) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D}) - \gamma C(\mathbf{w})}$.

The evidence itself is the normalisation constant for the post training distribution

$$P(\mathcal{D} \mid \gamma, \beta) = \int \prod_j dw_j P(\mathcal{D} \mid \beta, \mathbf{w}) P(\mathbf{w} \mid \gamma), \tag{4.1}$$

that is, the likelihood function for the hyper-parameters β and γ. We continue to refer to the *evidence procedure* as the process of fixing the hyper-parameters to the values that simultaneously maximize the evidence for a given data set. Thus, although the Bayesian framework outlined here envisages the hyper-parameters as defining the whole distribution of input-output pairs, the assignments from the evidence procedure will depend on the data set at hand. Indeed, as we have seen one could regard this

procedure as *empirical Bayes* (see *e.g.* Berger (85)) where, to some extent, the data is allowed to influence the choice of prior. In addition, we emphasize that this is the way in which the evidence procedure is used in practice (Mackay (92)).

4.2.2 The performance measures

Here we briefly review the two performance measures with which we will concern ourselves. As before the short hand $\langle . \rangle_{\mathbf{w}}$ denotes the average over the posterior distribution $P(\mathbf{w} \mid \mathcal{D}, \gamma, \beta)$.

As the principal performance measure we choose the expected squared difference over the input dimension $P(\mathbf{x})$ between the average student and the average teacher. This is the data dependent error of equation (1.12) with the error measure defined in equation (1.11). Thus, we have the data dependent generalization error,

$$\epsilon_g(\mathcal{D}) = \langle (\langle y_t(\mathbf{x}) \rangle_{P(y_t|\mathbf{x})} - \langle y_s(\mathbf{x}) \rangle_{\mathbf{w}})^2 \rangle_{P(\mathbf{x})}. \tag{4.2}$$

If we were to *average over all possible data sets* of fixed size then this would correspond to the generalization error studied by Bruce and Saad (94) and Krogh and Hertz (92). Indeed, the key difference between this performance measure and that used in chapter 2 is the explicit data dependency of the former. A second difference is the choice of error measure discussed in section 1.2.1 which means that the average of $\epsilon_g(\mathcal{D})$, in equation (4.2), is equivalent to the error defined in chapter 2 up to an additive constant; the variance in teacher output. In the linear case studied in this chapter, this variance is simply that of the noise corrupting the teacher outputs. The question arises as to what one means by optimal procedure. As noted previously, in the context of a real supervised learning experiment we are concerned with the performance based on the actual data set available and not on the average performance. Thus, the optimal policy is that which minimises the data dependent generalization error and our focus will be on the performance of the evidence procedure in relation to this. However, in section 4.3.1 we will consider an average case approach. Further, in section 4.5 we will also consider the effect of defining the optimal hyper-parameter assignment in terms of the average $\langle \epsilon_g(\mathcal{D}) \rangle_{P(\mathcal{D})}$ whilst using the data dependent evidence assignments. This will enable us to asses the relative importance of fluctuations in the optimal and the evidence assignments.

We again consider the variance of the student output, $y_s(\mathbf{x})$, over the student distribution $\langle \{y_s(\mathbf{x}) - \langle y_s(\mathbf{x}) \rangle_{\mathbf{w}}\}^2 \rangle_{\mathbf{w},P(\mathbf{x})}$. Adapting the definition of Bruce and Saad (94)

we define the *data dependent* consistency measure as,

$$\delta_c(\mathcal{D}) = \langle \{ y_s(\mathbf{x}) - \langle y_s(\mathbf{x}) \rangle_{\mathbf{w}} \}^2 \rangle_{\mathbf{w},P(\mathbf{x})} - \epsilon_g(\mathcal{D}). \tag{4.3}$$

For a linear student and teacher in the limit $\alpha \to \infty$ δ_c tends to zero. As before, we regard $\delta_c(\mathcal{D}) = 0$ as optimal since then we can estimate our expected error, $\epsilon_g(\mathcal{D})$, from the variance of our student output; which in principle we can calculate if we could estimate the input distribution. Indeed, Krogh and Vedelsby (95) suggest using unlabelled data to estimate the variance over the ensemble of students, albeit in slightly different context. Again note that we are principally concerned with the optimal procedure based on the training data available and not on the average over all such sets.

4.3 Finite system size

In this section we consider a finite system size N examining the large p limit and showing that in the learnable linear case under consideration in this chapter the evidence procedure is unbiased in a particular sense. Here, since the student is linear with output $y(\mathbf{x}) = \mathbf{w}.\mathbf{x}/\sqrt{N}$ we have $N_s = N$. We also assume that the teacher mapping is linear, parameterised by the weight vector $\mathbf{w}^o$, and corrupted by zero mean Gaussian noise of variance σ^2. Thus, $P(y_t \mid \mathbf{x}_\mu) \propto \exp[-(y_t^\mu - \mathbf{w}^o.\mathbf{x}_\mu/\sqrt{N})^2/2\sigma^2]$. Further, we assume $P(\mathbf{x})$ is $\mathcal{N}(0, \sigma_x)$ and adopt weight decay as our regularization procedure, that is $C(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$. In this case we can explicitly calculate the evidence, or rather the normalised log of the evidence $f(\mathcal{D}) = -1/N \ln P(\mathcal{D} \mid \lambda, \beta)$, analogous to a free energy, where we have introduced the weight decay parameter $\lambda = \gamma/(\beta\sigma_x^2)$. We note that this definition is a scaled version of that given in section 2.3; in the linear case it is more convenient to normalise by the factor σ_x^2. The calculation of the free energy in this case is shown in appendix 4.8.1 and equations (4.A7) and (4.A8) there lead to,

$$f(\mathcal{D}) = -\frac{1}{2}\ln\frac{\lambda}{\pi} - \frac{\alpha}{2}\ln\frac{\beta}{\pi} + \frac{1}{2}\ln 2 + \beta\lambda\sigma_x^2\sigma_w^2 - \frac{1}{2N}\ln\det \mathbf{g} + \beta\mathbf{n}^T\Gamma\mathbf{n} + \mathbf{y}.\mathbf{n} + \beta a_{ev} \tag{4.4}$$

where,

$$\Gamma_{\mu\nu} = -\frac{(\mathbf{x}_\mu)^T\mathbf{g}\mathbf{x}_\nu}{N^2\sigma_x^2} + \frac{\delta_{\mu\nu}}{N}, \quad y_\mu = \frac{2\lambda(\mathbf{w}^o)^T\mathbf{g}\mathbf{x}_\mu}{N\sqrt{N}},$$

$$a_{ev} = -\frac{\sigma_x^2\lambda^2(\mathbf{w}^o)^T\mathbf{g}\mathbf{w}^o}{N}$$

$$\text{and} \quad \mathbf{g} = (\mathbf{A} + \lambda \mathrm{I})^{-1} \quad \text{with} \quad \mathbf{A} = \frac{1}{N\sigma_x^2}\Sigma_{\mu=1}^p(\mathbf{x}_\mu)^T\mathbf{x}_\mu \ .$$

Here μ and ν index the p patterns, I is the identity matrix in N dimensions, $N\sigma_w^2 = \mathbf{w}^o \cdot \mathbf{w}^o$ and the p dimensional noise vector $\mathbf{n}$ has components drawn independently from $\mathcal{N}(0, \sigma)$. The term a_{ev} does not fluctuate with the noise but only with the inputs $\mathbf{x}^\mu$. The normalization of the input correlation matrix $\mathbf{A}$ results in the rescaling of the weight decay noted earlier. Also as outlined in appendix 4.8.1 the generalization error and the consistency can be calculated from $f(\mathcal{D})$ by averaging appropriate expressions over the input distribution $P(\mathbf{x})$.

The generalization error is given by,

$$\epsilon_g(\mathcal{D}) = \mathbf{n}^T \Delta \mathbf{n} + \mathbf{z}.\mathbf{n} + a_{\epsilon_g} \tag{4.5}$$

where,

$$\Delta_{\mu\nu} = -\frac{1}{N^2\sigma_x^2}(\mathbf{x}_\mu)^T \frac{\partial \mathbf{g}}{\partial \lambda}\mathbf{x}_\nu \qquad z_\mu = \frac{2\lambda}{N\sqrt{N}}(\mathbf{w}^o)^T \frac{\partial \mathbf{g}}{\partial \lambda}\mathbf{x}_\nu$$
$$\text{and} \qquad a_{\epsilon_g} = -\frac{\lambda^2\sigma_x^2}{N}(\mathbf{w}^o)^T \frac{\partial \mathbf{g}}{\partial \lambda}\mathbf{w}^o \ .$$

Finally, the consistency is,

$$\delta_c(\mathcal{D}) = \frac{1}{2\beta N}\text{tr } \mathbf{g} - \epsilon_g(\mathcal{D}) \ . \tag{4.6}$$

We note here that the generalization error depends only on the weight decay, λ, thus in the remainder of this chapter we refer to the optimal weight decay $\lambda_{opt}(\mathcal{D})$ as that which minimizes $\epsilon_g(\mathcal{D})$. Similarly, for fixed weight decay the optimal inverse temperature, $\beta_{opt}(\mathcal{D})$, ensures that $\delta_c(\mathcal{D}) = 0$ and thus that the variance of the student distribution is equal to the generalization error. We denote the hyper-parameters that simultaneously maximize the evidence as $\lambda_{ev}(\mathcal{D})$ and $\beta_{ev}(\mathcal{D})$. Thus, the term *optimal* refers to the optimization of, or with respect to, the performance measures whilst *evidence optimal* refers to maximisation of the evidence.

4.3.1 Consistency and unbiasedness

Firstly we consider the question of consistency, that is, we examine the free energy, $f(\mathcal{D})$, and the generalization error in the limit of large amounts of data (*i.e.* as $p \to \infty$ with N fixed). Using the fact, shown in appendix 4.8.2, that, for large p, $g_{ij} = \delta_{ij}N/p + \mathcal{O}(1/p^{3/2})$ we can find the asymptotic evidence optimal hyper-parameter assignments,

namely,

$$\lim_{p\to\infty} \lambda_{ev}(\mathcal{D}) = \lambda_0 + \mathcal{O}\left(\frac{1}{\sqrt{p}}\right) \text{ and } \quad \lim_{p\to\infty} \beta_{ev}(\mathcal{D}) = \beta_0 + \mathcal{O}\left(\frac{1}{\sqrt{p}}\right) \tag{4.7}$$

where, $\lambda_0 = \sigma^2/(\sigma_x^2\sigma_w^2)$ and $\beta_0 = 1/(2\sigma^2)$. In addition, it can be shown that, to first order in $1/p$, the generalization error is independent of λ. As we shall see later in the context of large N this insensitivity of the generalization error to the value of the weight decay is associated with a divergence in the variance of the optimal weight decay as the number of examples grows large.

That the generalization error is independent of the weight decay for large p implies that any scheme, and in particular the evidence assignments, will achieve optimal performance asymptotically (*i.e.* the generalization error tends to zero irrespective of λ). However, as we shall see in section 4.4 this does not imply that the evidence assignments correspond to the optimal hyper-parameters. Rather, it is a reflection of the fact that, for any weight decay setting, our linear student is *mean square consistent* (see *e.g.* Stone (77b)) when the teacher is also linear.

For this reason, instead of looking directly at the generalization error when assessing the performance of the evidence assignments we will focus on the fractional increase, κ_{ϵ_g}, in generalization error above the optimal incurred by their use. That is, on

$$\kappa_{\epsilon_g}(\lambda_{ev}, \mathcal{D}) \equiv \frac{\epsilon_g(\lambda_{ev}, \mathcal{D}) - \epsilon_g(\lambda_{opt}, \mathcal{D})}{\epsilon_g(\lambda_{opt}, \mathcal{D})}. \tag{4.8}$$

Similarly the fractional error in estimating the generalization error from the variance of the student distribution is,

$$\kappa_{\delta_c}(\lambda_{ev}, \beta_{ev}, \mathcal{D}) \equiv \frac{\delta_c(\lambda_{ev}, \beta_{ev}, \mathcal{D})}{\epsilon_g(\lambda_{ev}, \mathcal{D})}. \tag{4.9}$$

These data dependent measures of the performance of the evidence assignments are analogous to those average case quantities defined in section 2.6. In section 4.5 we examine the asymptotic behaviour of both $\kappa_{\epsilon_g}(\mathcal{D})$ and $\kappa_{\delta_c}(\mathcal{D})$ in the thermodynamic regime.

Secondly, we consider average case behaviour. Using the result of appendix 4.8.3 it can be shown that,

$$\langle \epsilon_g(\mathcal{D}) \rangle_{P(\mathcal{D})} = \sigma^2 G_{\text{av}} + \lambda \partial_\lambda G_{\text{av}} (\sigma^2 - \lambda \sigma_x^2 \sigma_w^2) \tag{4.10}$$

where the response function $G_{\text{av}} = \langle \text{tr } \mathbf{g} \rangle_{P(\mathcal{D})}$ is unknown in general. The average

generalization error is clearly optimised by $\lambda = \lambda_0$. Similarly, it can be shown that the average consistency is optimised by $\beta = \beta_0$ whilst the resulting average free energy, $f = \langle f(\mathcal{D}) \rangle_{P(\mathcal{D})}$ is extremised by $\lambda = \lambda_0$ and $\beta = \beta_0$. This corresponds to the average case result obtained for the thermodynamic limit by Bruce and Saad (94) but is valid *for all N and p*. However, we are not able to explore the behaviour in more detail in this regime since we can only calculate $G_{\rm av}$ explicitly in the region of the thermodynamic limit.

Thus, the particular conclusion, of the thermodynamic average case analysis of Bruce and Saad (94), that the evidence procedure optimises average performance is valid for all N and p and in this sense the evidence procedure is unbiased. Indeed, this is a reflection of the fact, noted by Mackay (92a), that on average the evidence will always favour the true model (see also section 1.2.1). However, we now show that the fluctuations around this average optimum performance become increasingly important as N gets smaller.

4.3.2 Self averaging

Using the result of Sollich (94) [1] that the variance of tr $\mathbf{g}/N$ is $\mathcal{O}(1/N^2)$ one can calculate the variance, over possible realisations of the data set, of the free energy, $f(\mathcal{D})$, obtaining

$$\begin{aligned} \mathrm{Var}(f(\mathcal{D})) = 2\sigma^4 \langle \mathrm{tr}\ (\Gamma\Gamma) \rangle_{P(\{\mathbf{x}^\mu : \mu = 1..p\})} + \sigma^2 \langle \mathrm{tr}\ (\mathbf{y}^T\mathbf{y}) \rangle_{P(\{\mathbf{x}^\mu : \mu = 1..p\})} \\ + \beta^2 \langle a_{ev}^2 \rangle_{P(\{\mathbf{x}^\mu : \mu = 1..p\})} - \beta^2 \langle a_{ev} \rangle^2_{P(\{\mathbf{x}^\mu : \mu = 1..p\})}. \end{aligned} \tag{4.11}$$

Here we have explicitly performed the noise average, the remaining average over the input points is with respect to $P(\{\mathbf{x}^\mu : \mu = 1..p\})$. As shown in appendix 4.8.4, it is readily verified that $\langle \mathrm{tr}\ (\Gamma\Gamma) \rangle_{P(\{\mathbf{x}^\mu : \mu = 1..p\})}$, $\langle \mathrm{tr}\ (\mathbf{y}^T\mathbf{y}) \rangle_{P(\{\mathbf{x}^\mu : \mu = 1..p\})}$ and the variance of a_{ev} are $\mathcal{O}(1/N)$ as we approach the thermodynamic limit. Thus, the variance of the free energy is $\mathcal{O}(1/N)$, *i.e.* it is self averaging. Similarly, it can also be shown that the generalization error and consistency measure are also self averaging. This means that in the thermodynamic limit the behaviour exhibited by the system for every data set will correspond to the average case behaviour, that is the fluctuations around the average vanish. Thus, we see that the average case analysis of Bruce and Saad (94) corresponds to the case of any *particular data set* because their results were obtained in the thermodynamic limit. In fact, although we do not do so here the same can be shown, in a similar fashion, for the models examined in the preceding chapters.

[1] Alternatively one can show this result using diagrammatic methods.

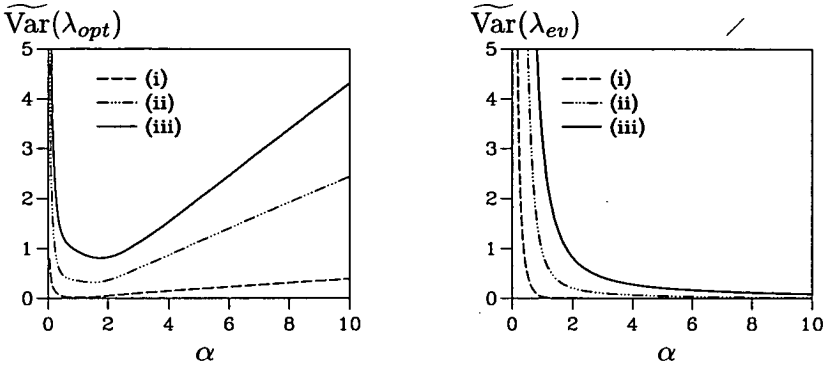
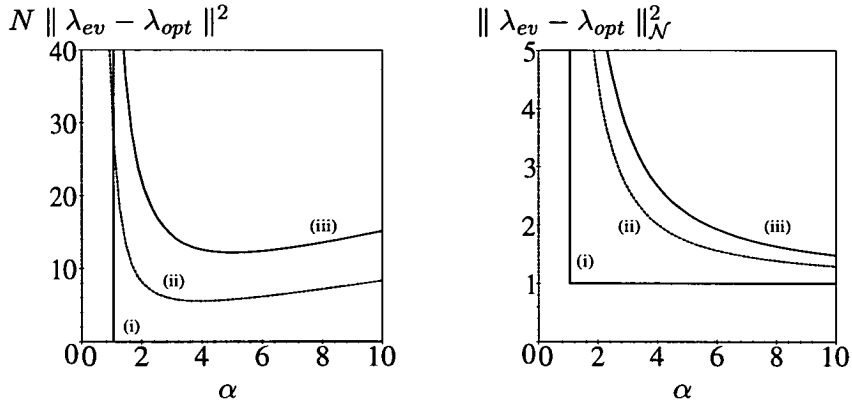


Figure 4.1. The scaled variance in the optimal weight decay, $\widetilde{\mathrm{Var}}(\lambda_{opt})$, for various noise levels, (i) $\lambda_0 = 0.04$, (ii) $\lambda_0 = 0.25$ and (iii) $\lambda_0 = 0.44$ is shown in the left-hand graph. Notice the linear divergence in α which corresponds to our result in section 4.3.1 that, for sufficiently large p, the generalization error is independent of λ. The variance in the evidence optimal weight decay, $\widetilde{\mathrm{Var}}(\lambda_{ev})$, is shown, in the right-hand graph, for the same noise levels. The $\mathcal{O}(1/\alpha)$ decay of this quantity is a reflection of the fact that for large p the evidence optimal weight decay $\lambda_{ev}(\mathcal{D}) = \lambda_0$.

The asymptotic $\mathcal{O}(1/\alpha)$ decay of the former reflects the fact that, as discussed in section 4.3.1, $\lim_{\alpha\to\infty} \lambda_{ev}(\mathcal{D}) = \lambda_0$. Similarly, the divergence of the latter is indicative of the insensitivity of the generalization error to the weight decay for large α. The divergence of both curves for small α is order $\mathcal{O}(1/(N\alpha))$ indicating a break down of the thermodynamic limit as $p/N \to 0$ and in fact, for $p = 1$ it can be shown *analytically* that these quantities are $\mathcal{O}(1)$. In the limit of zero noise we find that the variance of λ_{ev} diverges for $\alpha \leq 1$ and is zero for $\alpha > 1$. However, in this limit of zero noise the variance of the optimal weight decay tends to zero irrespective of α. Since, at least to first order, the average of $\Delta\lambda_{opt}$ is zero this means that optimal weight decay ($\lambda_{opt} = \lambda_0 + \Delta\lambda_{opt}$) is zero in the limit of no noise. In this case the evidence procedure can only set the weight decay with confidence for $\alpha > 1$, whilst the optimal policy is to accept the data completely (zero weight decay) for all α. Thus, in the noiseless limit there is a phase transition at $\alpha = 1$ below which the evidence weight decay assignment is ill defined and above which it is optimal. This is analogous to the phase transition discussed in section 2.5.2.



(a) Unnormalised (b) Normalised

Figure 4.2. The distance between the optimal and the evidence optimal weight decay: The left-hand graph shows the unnormalised separation, $\parallel \lambda_{ev} - \lambda_{opt} \parallel^2$, (scaled by N) for various noise levels; (i) zero noise, (ii) $\lambda_0 = 0.8$, and (iii) $\lambda_0 = 1$. In large α limit the unnormalised separation diverges linearly, whilst for $\alpha < 1$ it diverges as $\lambda_0 \to 0$ and is zero otherwise. The right-hand graph shows the normalised distance, $\parallel \lambda_{ev} - \lambda_{opt} \parallel^2_{\mathcal{N}}$ for the same noise levels. In this case asymptotically we find a separation of 1. In the limit of zero noise $\parallel \lambda_{ev} - \lambda_{opt} \parallel^2_{\mathcal{N}} \to 1$ for $\alpha > 1$ whilst it too diverges for $\alpha < 1$.

A second feature we consider is the average separation between the evidence assignment of the weight decay and the optimal,

$$\parallel \lambda_{ev} - \lambda_{opt} \parallel^2 \equiv < (\lambda_{ev}(\mathcal{D}) - \lambda_{opt}(\mathcal{D}))^2 >_{P(\mathcal{D})} . \tag{4.16}$$

This quantity, scaled by N, is shown in the left-hand graph of figure 4.2 for a number of different noise levels. We see that as the noise increases the separation also increases. However, in the limit of zero noise whilst $\parallel \lambda_{ev} - \lambda_{opt} \parallel^2$ is zero for $\alpha > 1$ we find that it diverges for $\alpha < 1$. This divergence is linked to the divergence in the evidence assignment of the weight decay discussed in the preceding paragraph. In the limit of large data sets we find that the average distance between the optimal weight decay and the evidence assignment diverges linearly, indeed for large α we find that,

$$\parallel \lambda_{ev} - \lambda_{opt} \parallel^2 \approx \mathrm{Var}(\lambda_{opt}). \tag{4.17}$$

Thus, we see that this divergence is caused by the fact that, whilst the evidence assignment becomes ever closer to λ_0, the variance, over data sets, of the optimal regularization parameter diverges. For this reason we define the normalised averaged separation

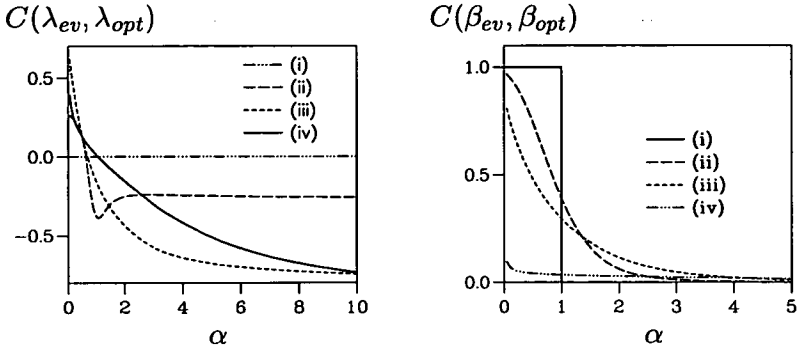


Figure 4.3. The correlation between the optimal weight decay and the evidence optimal weight decay $C(\lambda_{ev}, \lambda_{opt})$ is shown, in the left-hand graph, for (i) $\lambda_0 \to 0.0$, (ii) $\lambda_0 = 0.01$, (iii) $\lambda_0 = 1$ and (iv) $\lambda_0 = 4$. The right-hand graph shows the correlation between the optimal inverse temperature β_{opt} and the evidence optimal β_{ev} for (i) $\lambda_0 \to 0.0$, (ii) $\lambda_0 = 0.025$, (iii) $\lambda_0 = 1$ and (iv) $\lambda_0 = 16$.

between the evidence and the optimal weight decay assignments as,

$$\| \lambda_{ev} - \lambda_{opt} \|^2_{\mathcal{N}} \equiv \frac{\| \lambda_{ev} - \lambda_{opt} \|^2}{\mathrm{Var}(\lambda_{opt})}, \tag{4.18}$$

which gives us a measure of the distance of the evidence assignments from the optimal, as a fraction of the uncertainty in that optimal. We note that it is an order $\mathcal{O}(1)$ quantity. In the right-hand graph of figure 4.2 we see that in the small noise limit the normalised separation tends to unity for $\alpha > 1$ but still diverges for $\alpha < 1$. Asymptotically, for large α, $\| \lambda_{ev} - \lambda_{opt} \|^2_{\mathcal{N}}$ tends to 1 irrespective of the noise level, λ_0, reflecting the fact that the variance in the optimal weight decay diverges whilst that in the evidence estimate tends to zero (see equation 4.15).

Similar calculations can be carried out for the optimal inverse temperature, β_{opt}, and the evidence optimal, β_{ev}. For example in the data dominated regime ($\alpha \to \infty$) we find that

$$Var(\beta_{ev}) \approx \frac{1}{2\sigma^4 \alpha N} \quad \text{and} \quad Var(\beta_{opt}) \approx \frac{2}{\sigma^4 N}. \tag{4.19}$$

Thus, asymptotically there is uncertainty in the optimal assignment, whilst the evidence assignment is well defined in this regime, reflecting the consistency result of equation (4.7). This discrepancy between the optimal and the evidence assignments of the inverse

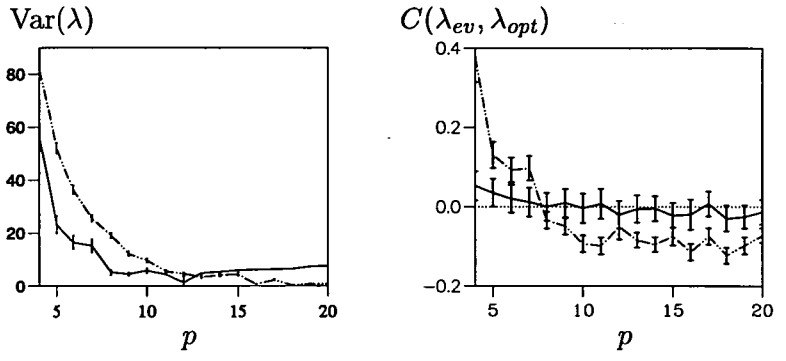


Figure 4.4. 1-D simulation results: The left-hand graph shows the variance in the optimal weight decay λ_{opt} (solid curve) and that in evidence optimal λ_{ev} (dot-dashed curve) both for $\lambda_0 = 1.0$. The latter curve has been scaled by a factor of 0.01 for ease of presentation and standard error bars are shown. Qualitatively, both curves show similar characteristics to the theoretical curves of figure 4.1. For larger p the variance of λ_{opt} continues to diverge linearly. In the right hand graph, the correlation between the optimal weight decay and the evidence optimal weight decay $C(\lambda_{ev}, \lambda_{opt})$ is shown, for $\lambda_0 = 0.01$ (full curve) and $\lambda_0 = 1$ (dot-dashed curve).

temperature is reflected in their average separation

$$\lim_{\alpha \to \infty} < (\beta_{ev}(\mathcal{D}) - \beta_{opt}(\mathcal{D}))^2 >_{P(\mathcal{D})} = \frac{2}{\sigma^4 N}. \tag{4.20}$$

We shall see in the next section that this result has implications for the estimation of the generalization error from the variance of the post training distribution of students.

Finally, we examine the normalised correlation between $\lambda_{ev}(\mathcal{D})$ and $\lambda_{opt}(\mathcal{D})$, $C(\lambda_{ev}, \lambda_{opt})$ and that between $\beta_{ev}(\mathcal{D})$ and $\beta_{opt}(\mathcal{D})$, $C(\beta_{ev}, \beta_{opt})$ to order $\mathcal{O}(1)$ as shown in figure 4.3. The normalised correlation between two fluctuating quantities $h(\mathcal{D})$ and $k(\mathcal{D})$ is $C(h(\mathcal{D}), k(\mathcal{D})) = (\langle hk \rangle_{P(\mathcal{D})} - \langle h \rangle_{P(\mathcal{D})} \langle k \rangle_{P(\mathcal{D})}) / (\mathrm{Var}(h)\mathrm{Var}(k))^{1/2}$. For small α the non-monotonic behaviour of $C(\lambda_{ev}, \lambda_{opt})$ with the noise level λ_0, is a reflection of the fact, discussed above, that the variance in the evidence assignment diverges for small noise whilst that of the optimal tends to zero. As the noise level increases $\mathrm{Var}(\lambda_{ev})$ reduces and $\mathrm{Var}(\lambda_{opt})$ increases causing the correlation to first increase and then decrease as a function of λ_0. For zero noise $C(\lambda_{ev}, \lambda_{opt})$ tends to zero for all α, since the fluctuations in λ_{opt} tend to zero in the noiseless limit. The behaviour of $C(\beta_{ev}, \beta_{opt})$ is more straight forward. For small α this correlation reduces monotonically with increasing λ_0. In the limit of zero noise $C(\beta_{ev}, \beta_{opt}) = 1$ for $\alpha < 1$ and is zero otherwise. The behaviour in the region $\alpha < 1$, where the variances of both β_{opt} and

β_{ev} diverge for small noise level is indicative of the fact that, for this case, in the thermodynamic limit neither the consistency nor the evidence are dependent on the inverse temperature, β. For $\alpha > 1$ both $\beta_{ev}(\mathcal{D})$ and $\beta_{opt}(\mathcal{D})$ are uniquely determined in the noiseless case.

Finally, in the large α limit we find

$$\lim_{\alpha\to\infty} C(\lambda_{ev}, \lambda_{opt}) = -\frac{\sqrt{2\lambda_0}}{\sqrt{2\lambda_0 + 1}}, \tag{4.21}$$

and

$$\lim_{\alpha\to\infty} C(\beta_{ev}, \beta_{opt}) \approx 4\lambda_0^2 \alpha^{-7/2}. \tag{4.22}$$

Thus, for large noise the asymptotic correlation between the evidence and the optimal weight decays tends to -1 whilst for small noise it tends to zero. In contrast $C(\beta_{ev}, \beta_{opt})$ invariably tends to zero. In general then, to order $\mathcal{O}(1/N)$ the evidence assignments correlate rather poorly with the optimal assignments.

As noted earlier, when defining the evidence procedure, we can choose whether to optimise the evidence with respect to each of the hyper-parameters whilst holding the other fixed or simultaneously *w.r.t.* both. In the thermodynamic limit, in the linear case, we find that the evidence assignments are optimal only in the case where we simultaneously minimise the free energy *w.r.t.* to both hyper-parameters (Bruce and Saad 94). This was the motivation for studying the later case here. However, we briefly note that if we fix $\beta_{ev} = \beta_0$ and optimise the evidence *w.r.t.* the weight decay only we are free to expand $\lambda_{ev}(\mathcal{D})$ about λ_0 as before. In this case we find that, in analogy to the thermodynamic limit, this assignment is less correlated with the optimal than in the situation we have been discussing where we optimise the evidence simultaneously with respect to both hyper-parameters.

4.4.1 Simulation results

To corroborate these results qualitatively, we performed Monte Carlo simulations of one and two dimensional perceptron students and teachers. In these simulations we generated random data sets and found the evidence assignments of the hyper-parameters and the true optimal assignments. Then by averaging over many such data sets we calculated the variances and correlations of these assignments. Some results from the one dimensional simulations are shown in figure 4.4. The left-hand graph shows the variance of λ_{opt} and of λ_{ev} versus the number of examples, p, in this case. They show qualitative agreement with the large N results of figure 4.1, with the variance of λ_{opt} diverging linearly for large p whilst that of λ_{ev} falls off with p. The right-hand graph of

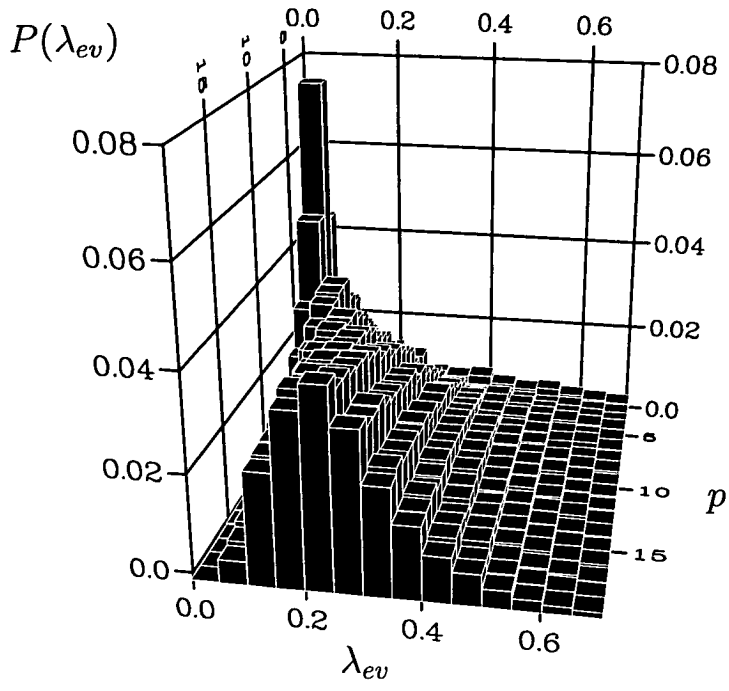


Figure 4.5. Probability distribution of evidence weight decay assignment: The figure shows histogrammed samples from this distribution for the 2D linear model; 10000 data sets were sampled for a noise level of $\lambda_0 = 0.25$. For a small number of examples the distribution is similar to that of the optimal but as p increases we find that the mass of the distribution concentrates around λ_0 (*i.e.* 0.25) in agreement with our asymptotic results.

figure 4.4 shows the correlation between λ_{opt} and λ_{ev}. These simulation results demonstrate that there is a region of positive correlation for a small number of examples and that as the noise reduces so does the level of the (anti)- correlation.

A better understanding of this behaviour is to be had by examining the histogrammed samples of the probability distributions, of the evidence and the optimal weight decay assignments. These are shown, for a two dimensional linear student and teacher with $\lambda_0 = 0.25$, in figure 4.5 for the evidence and figure 4.6 for the optimal assignments. In both graphs the number of examples, p runs from 2 to 20. For small p the distribution of evidence assignments looks qualitatively similar to that of the optimal assignment, namely a distribution peaked at zero but with a long tail. Thus, there are many occasions where λ_{ev} and λ_{opt} are coincident and the correlation between them is positive although as we expect from figure 4.1 the variances in the assignments

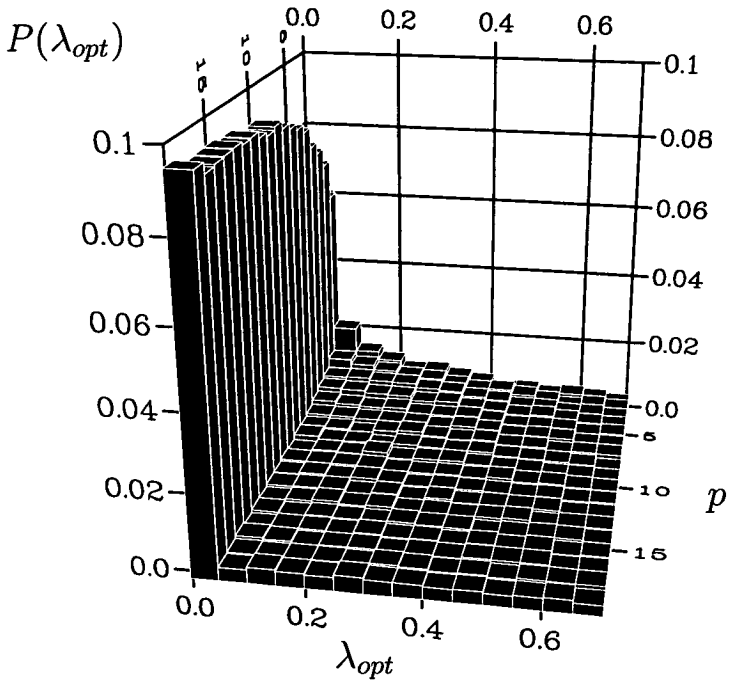


Figure 4.6. Probability distribution of optimal weight decay assignment: The figure shows histogrammed samples from this distribution for the 2D linear model; 10000 data sets were sampled for a noise level of $\lambda_0 = 0.25$. This distribution, skewed towards zero, but with a long tail becomes more accentuated as p grows, explaining the growth in the variance of λ_{opt}.

are large. The fact that the correlation is positive is confirmed by figure 4.7 where we plot a histogrammed sample of the co-occurrence of the optimal and the evidence weight decay assignments for $p = 2$.

As p grows the evidence assignments begin to cluster around λ_0 (see figure 4.5) as by our consistency results they must for large p. The mean of λ_{ev} thus tends to λ_0 and its variance decays in accord with our thermodynamic results. In contrast, as p grows the distribution of the optimal assignment remains similar to its small p form but becomes more accentuated; the peak at zero and the tail both grow. Thus, the variance in λ_{opt} becomes larger in accord with our theoretical results (see equation 4.15). Given the differences between these two distributions it is hardly surprising that the correlation between the two corresponding hyper-parameter assignments is not positive in this region. This is shown by the coincidence (or lack of) of the two assignments in figure 4.8 for $p = 10$ and $\lambda_0 = 0.25$. Finally, we note that the increasing dissimilarity of the

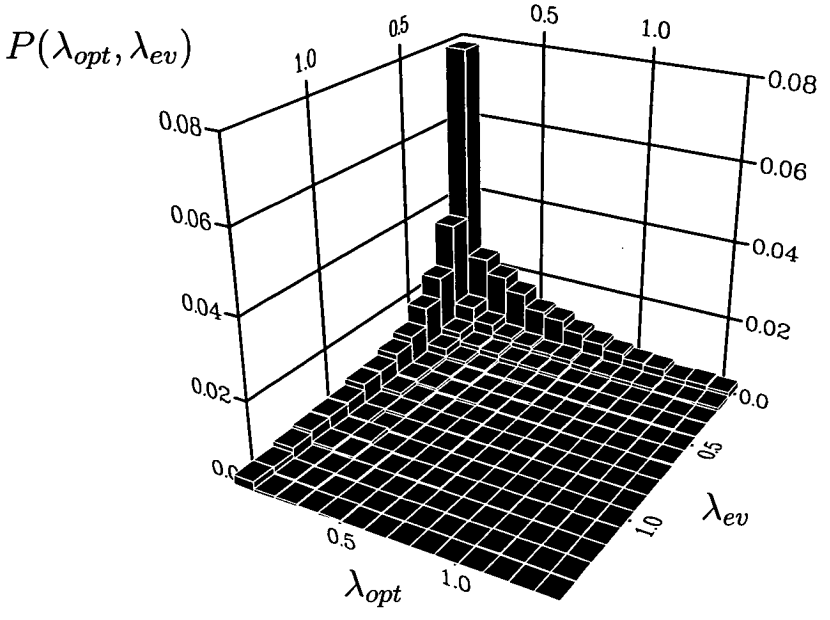


Figure 4.7. Correlation of evidence assignments with the optimal: Histogrammed samples of the probability of co-occurrence of $\lambda_{ev}(\mathcal{D})$ versus $\lambda_{opt}(\mathcal{D})$ for 10000 data sets of $p = 2$ examples in the 2 dimensional linear model with noise level $\lambda_0 = 0.25$. The mode of this distribution is seen to be where the two assignments are the same (*i.e.* zero) suggesting a positive correlation.

two distributions, as p increases, is reflected in the fact that the distance, $\| \lambda_{ev} - \lambda_{opt} \|^2$, we calculated in the thermodynamic limit diverges as α increases.

A word or two should be said of the relation of these results to the thermodynamic regime with which our analysis is concerned. Firstly, the simulations were performed to demonstrate that, at least qualitatively, the features found in our analysis were evident in a system of finite size, N. Indeed, the one and two dimensional systems studied are extreme cases (*i.e.* very small N) and thus we would expect differences to our theoretical results to become smaller as the system size increases. In particular, the evidence and the optimal assignments of the weight decay must cluster around λ_0 as the system size increases, at least when p increases correspondingly (as we have seen the thermodynamic limit breaks down for $\alpha = p/N \to 0$. see *e.g.* figure 4.1). Given this, the results for the optimal weight decay assignment of the $N = 2$ system, shown in figure 4.6, seem particularly strange. The peak at zero is not readily understandable

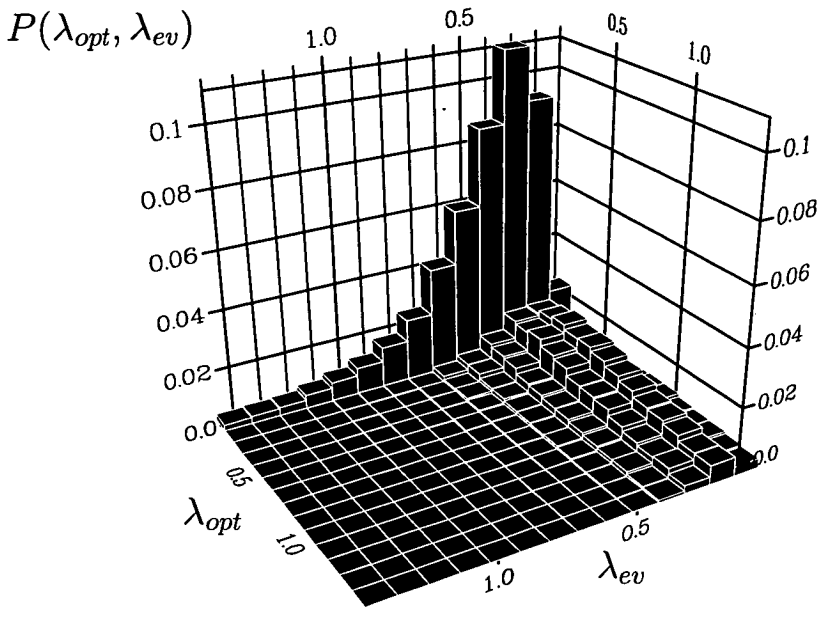


Figure 4.8. Correlation of evidence assignments with the optimal: Histogrammed samples of the probability of the co-occurrence of $\lambda_{ev}(\mathcal{D})$ versus $\lambda_{opt}(\mathcal{D})$ for 10000 data sets of $p = 10$ examples in the 2 dimensional linear model with $\lambda_0 = 0.25$. In this case the mode of the distribution does not occur where the two assignments are coincident revealing that they are negatively correlated. Indeed, the shape of this distribution reflects the fact that the optimal assignments are concentrated around zero whilst the evidence assignments cluster around $\lambda_{ev} = \lambda_0$.

and is in disagreement with our finite size corrections, although it should be noted that $N = 2$ is far from the thermodynamic limit. However, preliminary numerical results suggest that in a slightly bigger system this problem is largely resolved; for $N = 10$ and $p \approx 10...16$ the distribution of the optimal assignments, $\lambda_{opt}(\mathcal{D})$, does indeed start to cluster around λ_0. However, for any finite sized system we would expect our analytic results to be unreliable for very small or very large p. The reason for this is that, as stated above, for $\alpha = p/N \to 0$ the thermodynamic limit breaks down whilst for $\alpha \to \infty$ our small fluctuations ansatz for λ_{opt} breaks down (see equation 4.15).

4.5 Effects on performance

We now examine the effects on performance of these sub-optimal hyper-parameter assignments. Firstly, for the generalization error to order $\mathcal{O}(1/\sqrt{N})$ the optimal performance, $\epsilon_g(\lambda_{opt}, \mathcal{D})$, and that resulting from use of the evidence procedure, $\epsilon_g(\lambda_{ev}, \mathcal{D})$ are the same. However, to order $\mathcal{O}(1/N)$ they differ, thus we can write the correlation between them, somewhat suggestively, as $1 - \mathcal{O}(1/N)$. Unfortunately, we are unable to calculate this correlation to $\mathcal{O}(1/N)$. Therefore, we examine the increase in error invoked by use of the evidence procedure

$$\begin{aligned} \Delta\epsilon(\mathcal{D}) &\equiv \epsilon_g(\lambda_{ev}, \mathcal{D}) - \epsilon_g(\lambda_{opt}, \mathcal{D}) \\ &= \Delta\lambda_{ev}\partial_\lambda\epsilon_g + \frac{1}{2}\Delta\lambda_{ev}^2\partial_\lambda^2\epsilon_g + \frac{1}{2}\Delta\lambda_{opt}^2\partial_\lambda^2\epsilon_g + \mathcal{O}\left(\frac{1}{N^2}\right), \end{aligned} \tag{4.23}$$

where the quantities in the second line are evaluated at λ_0. The degradation in performance, $\Delta\epsilon(\mathcal{D})$, is a fluctuating quantity (over data sets) and in order to estimate its typical magnitude we calculate its average and variance. The average degradation in performance can be written in terms of the average separation of the evidence weight decay assignment and the optimal, as defined in equation (4.16). Thus, we find that,

$$< \Delta\epsilon(\mathcal{D}) >_{P(\mathcal{D})} = \frac{1}{2}(\partial_\lambda^2\epsilon_g)_0 \parallel \lambda_{ev} - \lambda_{opt} \parallel^2 + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{4.24}$$

Whilst, the calculation of this average is then straight forward that of the variance is more tricky. The variance is $\mathcal{O}(1/N^2)$ and thus we would have to calculate the variance of the response function tr $\mathbf{g}/N$ to this order. Instead, we simply calculate the variance over the noise ignoring that over the inputs. Clearly, this will give a *lower bound* on the true variance. We also expect this to become increasingly tight as α grows since for zero noise the fluctuations generated by the input variables vanish for $\alpha > 1$. Thus, to $\mathcal{O}(1/N)$, a lower bound on the *typical* error invoked by use of the evidence procedure is the average degradation of equation (4.24) plus the square root of its variance over the noise.

In figure 4.9, to first order, we plot this typical error, $\langle\Delta\epsilon\rangle_{P(\mathcal{D})} + (\mathrm{Var}(\Delta\epsilon))^{1/2}$, scaled by N as a fraction of the optimal generalization error. This quantity is denoted, $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev})$. As before the notation $\widetilde{h}$ denotes the function h scaled by N. Figure 4.9 shows that use of the evidence procedure results in a fractional degradation of significant magnitude for finite system size, N, and number of examples, α. This is

true of the degradation itself and clearly demonstrates the failings of the average case approach which, as we have seen, suggests the evidence assignments are optimal in this case. Figure 4.9 allows one to determine a lower bound on the typical fractional degradation for any system size. For example, for $N = 100$, we see that the fractional errors shown in figure 4.9 will range between 0.01 and 0.29 and for a larger sized system the evidence procedure results in closer to optimal behaviour. In fact, in the large α limit we find that, for the *average* fractional degradation,

$$\lim_{\alpha\to\infty} < \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})} = \frac{1}{N} + \frac{2(\lambda_0+1)}{N\alpha} + \mathcal{O}(\frac{1}{N\alpha^2}) \tag{4.25}$$

Note that the average relative degradation, $< \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})}$, does not decay with α despite the fact that the degradation in performance, $< \Delta\epsilon(\lambda_{ev}) >_{P(\mathcal{D})}$, is itself $\mathcal{O}(1/\alpha N)$. Thus, although the evidence assignments are consistent in a mean square sense they are never optimal even asymptotically. Furthermore, given the large fractional degradation associated with the evidence for finite α and N (shown in figure 4.9) even this mean square consistency is of questionable relevance in practice. If we consider the fluctuations, induced by the noise, in the relative degradation we find that asymptotically they do not contribute being order $\mathcal{O}(1/\alpha N)$. Indeed, in general the importance of the fluctuation term can be seen in figure 4.10. Graph 4.10(a) shows that the fluctuations do not qualitatively change the behaviour of the relative degradation, $< \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})}$. The relative size of the fluctuation term as a fraction of the typical error, $< \Delta\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev}) > \equiv (\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev}) - < \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})})/\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev})$, is shown in graph 4.10(b) where we see that the fluctuations are most important for a mid range α.

As the noise level increases so does κ_{ϵ_g} which is a reflection of the increasing uncertainty in λ_{ev} as shown in the right-hand graph of figure 4.1. In the zero noise limit, since we consider only the variance induced by the noise, the fluctuation term vanishes in both the degradation and the fractional degradation, for all α. However, whilst the average degradation, $< \Delta\epsilon(\lambda_{ev}) >_{P(\mathcal{D})}$ vanishes for $\alpha > 1$ it diverges for $\alpha < 1$. Thus, for zero noise the evidence procedure gives optimal performance for $\alpha > 1$ but very poor performance for $\alpha < 1$. The fractional degradation is more revealing in this limit, as we find that $< \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})}$ diverges when the normalised number of examples, α is less than one, but for $\alpha > 1$,

$$\lim_{\lambda_0\to 0} < \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})} = \frac{1}{N}\frac{\alpha+1}{\alpha-1}, \tag{4.26}$$

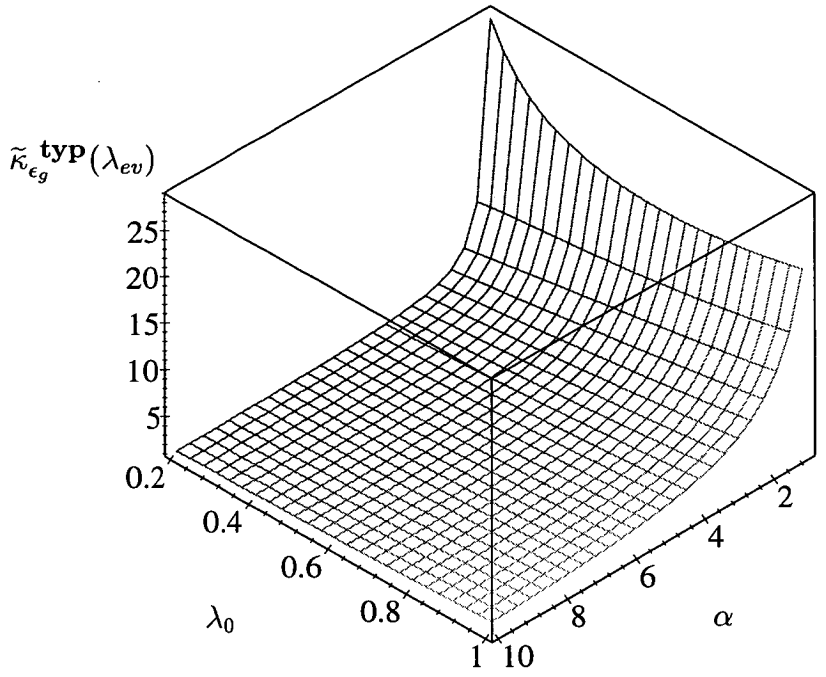
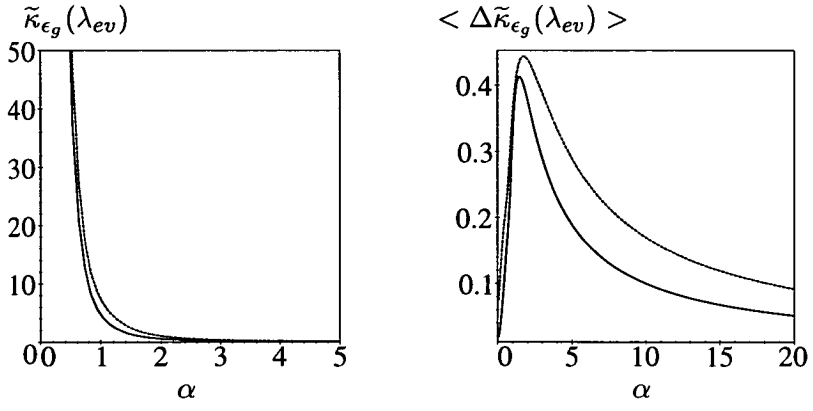


Figure 4.9. Scaled estimate of the fractional error κ_{ϵ_g}: for a system size of N dividing $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev})$ by N gives the an estimate of the true fractional increase in error above the optimal incurred by using the evidence procedure. $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev})$ diverges as $\lambda_0 \to \infty$ and as $\alpha \to 0$. For large α $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev})$ tends to $1/N$ and for small noise it diverges for $\alpha < 1$ and is a finite function of α for $\alpha > 1$ (see text).



(a) Average and typical degradation

(b) Fractional fluctuation.

Figure 4.10. Importance of the fluctuations around the average fractional degradation in performance: Graph (a) compares the average degradation in performance, $< \widetilde{\kappa}_{\epsilon_g} >_{P(\mathcal{D})}$, shown by the solid curve, with the larger (dotted) curve showing the *typical* degradation for λ_0. The latter includes an estimate of the variance of the typical error (see text). Graph (b) shows this fluctuation as a fraction of the typical error (*i.e.* average plus fluctuation), $< \Delta\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev}) > \equiv (\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev}) - < \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})})/\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}(\lambda_{ev})$, for $\lambda_0 = 0.5$ in the upper (dotted) curve and $\lambda_0 = 0.1$ in the lower (full) curve. The importance of the fluctuations diminishes as the number of examples grows.

showing that, for small noise, the evidence does not give optimal performance. We can understand this behaviour if we consider the evidence weight decay assignments in the case of zero noise. In the region $\alpha < 1$ the variance of $\lambda_{ev}(\mathcal{D})$ diverges as $\lambda_0 \to 0$ and thus $\lambda_{ev}(\mathcal{D})$ is ill defined. This mirrors the behaviour we found in the thermodynamic limit for the $\gamma \to 0$ model discussed in section 2.5.2 where β_{ev} is ill defined for $\alpha < 1$. Furthermore, as we noted in the previous section, in the current scenario we find that for $\alpha > 1$ the variance, $\widetilde{\mathrm{Var}}(\lambda_{ev}) \to 0$ in the limit of no noise and thus the evidence weight decay assignment is zero (*i.e.* $\lambda_{ev} = \lambda_0 + \Delta\lambda_{ev} \to \lambda_0 \to 0$). When there is no noise on the examples the optimal weight decay, λ_{opt}, is zero for all α since there is no danger of over-fitting. Thus, the average degradation, $< \Delta\epsilon >_{P(\mathcal{D})}$ and the separation between the evidence and optimal weight decays diverge for $\alpha < 1$ and are zero otherwise. This reflects the fact that for $\alpha < 1$ we do not even have enough examples to fix all the weights and certainly do not have enough to set the weight decay. However, for $\alpha > 1$ the evidence optimal assignment is well determined. Thus, in the noiseless limit the performance of the evidence is optimal for $\alpha > 1$. However, this is not reflected in the average fractional degradation, equation (4.26), because the optimal error approaches zero at the same rate as the degradation in performance. In other words for small noise

level and $\alpha > 1$ the evidence assignments are still sub optimal. We also note here that this difference between the degradation in performance and the fractional degradation is mirrored by that between the separation and the normalised separation. In the limit of no noise the former tends to zero, for $\alpha > 1$, whilst the latter tends to unity (see figure 4.2).

We have argued that the optimal policy is a function of the actual data set available and to date we have largely focussed on this definition. However, we now briefly discuss the effect of re-defining the optimal policy as that which minimises the *average* generalization error. As we saw in section 4.3.1 this is achieved by choosing the weight decay $\lambda = \lambda_0$. Thus, in this case the optimal weight decay does not fluctuate over data sets and the error associated with the evidence assignments will be due to fluctuations in $\lambda_{ev}(\mathcal{D})$ alone. Furthermore, we have already seen that asymptotically the evidence assignment tends to λ_0. It is thus not surprising that we find the average relative degradation associated with the evidence assignment when compared with the new 'optimal' generalization error, $\langle \epsilon_g(\lambda_0, \mathcal{D}) \rangle_{P(\mathcal{D})}$, is to first order in α^{-1} $\mathcal{O}(1/N\alpha)$ and in fact, $< \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})} \approx 4\lambda_0/(N\alpha)$. Thus, in this case the evidence assignment is asymptotically optimal and it is clear that the fluctuations in the optimal weight decay caused the asymptotic inconsistency reflected in equation (4.25). In contrast, for this new optimal, at small α we find qualitatively similar behaviour in the fractional degradation to that displayed in figure 4.9. Moreover, fluctuations in the optimal are relatively unimportant, in terms of performance loss, for small α but grow rapidly with the number of examples; dominating in the asymptotic regime as we have seen. These results show that an average case definition of optimal is misleading especially in the data dominated regime.

Finally, we consider the error incurred in estimating the generalization error from the variance of the post training distribution of students. If we use the evidence assignment of the inverse temperature, $\beta_{ev}(\mathcal{D})$, then our error will be $\mathcal{O}(1/\sqrt{N})$; an order of magnitude larger than the degradation, $\Delta\epsilon(\lambda_{ev}, \mathcal{D})$, itself. On average this vanishes but we can estimate the typical size of the fluctuation by calculating the square root of its variance. Dividing this by the true generalization error gives an estimate of the fractional error, κ_{δ_c} defined in equation (4.9). To first order this quantity, scaled by $\sqrt{N}$ and denoted by $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$, is plotted in figure 4.11. In general, $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$ is much larger than $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}$. In the limit of large α we find that there is some residual error associated with this procedure when using the evidence assignments, $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}} \approx \sqrt{2}$. This result reflects the asymptotic discrepancy between the evidence and the optimal assignments of the inverse temperature as shown in equation (4.20). For $\lambda_0 \to 0$ $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$ diverges whereas $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}} \to 0$ as λ_0 increases. That is, as the noise level increases the generalization error

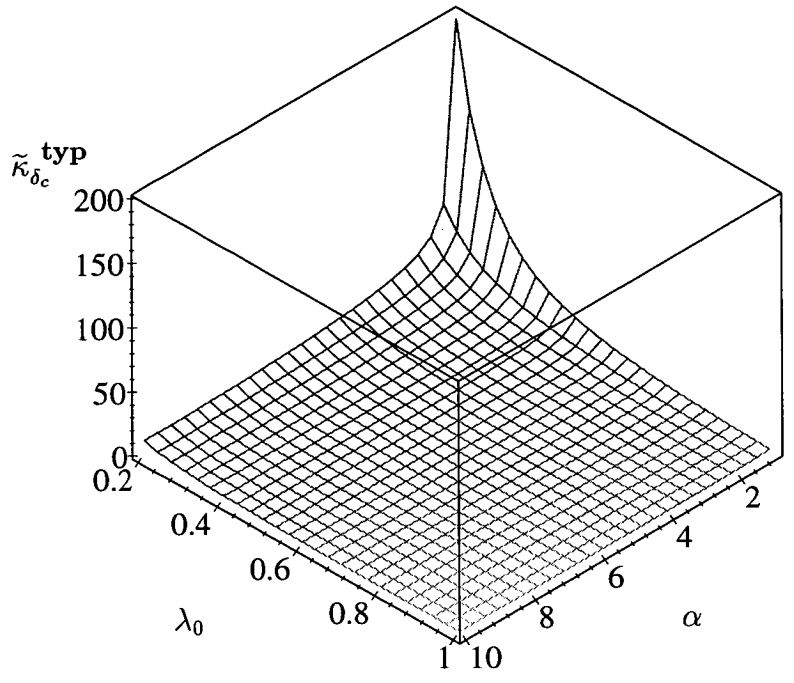


Figure 4.11. Scaled estimate of the fractional error κ_{δ_c}: for a system size of N dividing $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$ by $N^{1/2}$ gives an estimate of the true fractional error in estimating the generalization error from the variance of the student distribution. $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$ diverges as $\alpha \to 0$ and as $\lambda_0 \to 0$ whilst as $\alpha \to \infty$ we find that $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}} \approx \sqrt{2}$.

becomes larger and we are able to estimate it, using the consistency criterion, to a greater degree of accuracy when it is larger.

4.6 Comparison with cross-validation

Given, that the evidence procedure is sub-optimal it is natural to ask if another model selection criteria could do better. Here we compare the evidence procedure with leave-one-out cross-validation, **CV(1)** (Stone 74), using simulations of our 1-dimensional system. That is we set the weight decay using the cross-validatory estimate and the evidence estimate and compare the resulting generalization error to the optimal. The

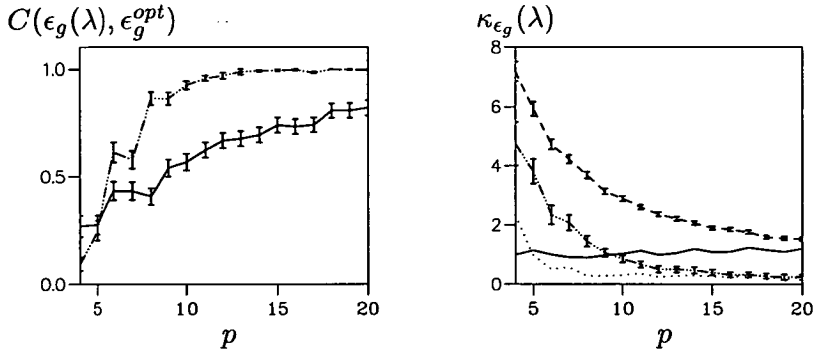


Figure 4.12. 1-D simulation results: The left-hand graph shows the correlation between the optimal generalization error and those obtained using the evidence (solid) and cross-validation (chain) with $\lambda_0 = 1.0$. The right-hand graph shows the fractional increase in generalization error $\kappa_{\epsilon_g}(\lambda) = (\epsilon_g(\lambda) - \epsilon_g(\lambda_{opt}))/\epsilon_g(\lambda_{opt})$. λ is set by the evidence (dashed) and by cross-validation (chain) for $\lambda_0 = 1.0$. For $\lambda_0 = 0.01$ the evidence case is the solid curve cross-validation the dotted curve. In the latter case the error bars are not shown for the sake of clarity but are of a similar magnitude.

results, averaged over 1000 realisations of the data set for each value of p, are plotted in figure 4.12. These results corroborate the results of the previous section in that they show the evidence procedure to be sub-optimal. They also show that cross-validation produces closer to optimal performance. The left-hand graph in figure 4.12 shows that the resulting error from the cross-validatory estimate correlates more strongly with the optimal generalization error than does that resulting from the evidence estimate. That this correlation for both methods increases with the number of examples, p, is a reflection of the mean square consistency of the linear student in this case. In addition, the right-hand graph shows that the fractional increase in the generalization error, $\kappa_{\epsilon_g}(\lambda)$, is considerably larger for the evidence procedure than for cross-validation.

We anticipate that the evidence and cross-validation assignments will become closer, or at least differences in performance will lessen, as the system size increases. However, due to the computational intensity of cross-validation, simulations become very expensive for larger systems. We thus defer further investigation of cross-validation until the next chapter where we calculate the finite size corrections to the **CV(1)** assignments in the thermodynamic limit.

4.7 Conclusion

By considering the fluctuations around the average case we have shown that in general, even in the learnable linear case the evidence assignments do not result in optimal performance, despite thermodynamic, asymptotic and average case results to the contrary. We have explored the evidence hyper-parameter assignments in terms of first order corrections to the thermodynamic limit and found qualitatively the same features in simulations of low dimensional systems. In particular, we found the evidence assignment of the weight decay became ever further from the optimal as the number of training examples increased and as the system size reduced. This is in stark contrast to the optimality of these assignments suggested by the average case approach. Consideration of the generalization performance reflected this sub-optimality, although it should be noted, that for large data sets, performance improves as more data become available, even if the weight decay is sub-optimal. Furthermore, we found that the inconsistency of the evidence weight decay assignment was due to asymptotically diverging fluctuations in the optimal for large data sets. The performance witnessed for finite normalised number of examples, α, showed that the asymptotic results are of little relevance to the data impoverished regime. We noted earlier the average case results of Meir and Merhav (94) on the consistency of hyper-parameter assignment via minimization of the stochastic complexity for a realizable case. Given our results it would be interesting to examine finite size effects in the stochastic complexity framework.

In addition, our numerical studies indicate that for small learnable linear systems cross-validation is closer, than the evidence procedure, to producing optimal performance. This is perhaps not surprising as cross-validation attempts directly to estimate the generalization error. However, we have found lower bounds on the system size required to make the evidence procedure reliable and in such instances it might still be a reasonable alternative to the computationally expensive cross-validation. In the next chapter we go on to examine model selection by methods, like cross-validation, which estimate the generalization error, but we will make comparisons with the results from this chapter throughout.

4.8 Appendices

4.8.1 Appendix A: calculation of the free energy

In this appendix we show how to calculate the evidence and the free energy in the linear case. Furthermore, we show how certain useful quantities, such as the average student weight vector, are calculated from it. We also indicate how to modify this to

obtain these quantities in the 2-teacher case of chapter 2 and in the scenario of chapter 3 where the student is a piece-wise linear function.

The noise model, as introduced in section 2.2.1, is normalised by the quantity,

$$\int \prod_{\mu=1}^{p} dn_\mu e^{-\beta(n_\mu)^2} = \left(\frac{\pi}{\beta}\right)^{\frac{p}{2}} \tag{4.A1}$$

when the training energy is a quadratic sum of the noise induced errors in each of the p examples represented by n_μ. Similarly, the prior distribution of equation (2.2) is normalised by the factor $(\gamma/\pi)^{N/2}$. The evidence as given in equation (2.4) can then be written,

$$P(\mathcal{D} \mid \gamma, \beta) = \left(\frac{\beta}{\pi}\right)^{\frac{p}{2}} \left(\frac{\gamma}{\pi}\right)^{\frac{N}{2}} \int \prod_{j} dw_j e^{-\mathcal{H}(\mathbf{w},\mathcal{D})} \tag{4.A2}$$

Where $\mathcal{H}(\mathbf{w}, \mathcal{D})$ is given by

$$\mathcal{H}(\mathbf{w}, \mathcal{D}) \equiv \beta E_w(\mathcal{D}) + \gamma \mathbf{w}^T \mathbf{w}. \tag{4.A3}$$

Recall that the weights $\mathbf{w}$ parameterise the student. In the case of a linear student we can write the quadratic training error, given a linear teacher parameterised by weight vector $\mathbf{w}^o$, as,

$$E_w(\mathcal{D}) = \sum_{\mu=1}^{p} \left(\frac{1}{\sqrt{N}}(\mathbf{w} - \mathbf{w}^o) \cdot \mathbf{x}_\mu - \mathbf{n}_\mu\right)^2. \tag{4.A4}$$

Here the variables n_μ represent instantiations, in the μ^{th} example, of the noise process corrupting the teacher output. The data set, $\mathcal{D}$, here expressed in terms of the examples generated by a linear teacher and corrupted by this noise process is written $\mathcal{D}$= $\{(\frac{1}{\sqrt{N}}\mathbf{w}^o \cdot \mathbf{x}_\mu + n_\mu, \mathbf{x}_\mu) : \mu = 1..p\}$ and the sum in the above expression is over the p elements of this set.

Upon introducing the vector difference between student and teacher weights, $\mathcal{R} = \mathbf{w} - \mathbf{w}^o$, following some rearrangement, we can express equation (4.A3) as;

$$\mathcal{H}(\mathcal{R}, \mathcal{D}) = \frac{1}{2}\mathcal{R}^T \Lambda^{-1} \mathcal{R} - \rho^T \mathcal{R} + \gamma(\mathbf{w}^o)^T \mathbf{w}^o \tag{4.A5}$$

Where, here we have defined the matrix Λ and the vector ρ by,

$$\Lambda^{-1} \;=\; 2(\beta\sigma_x^2 \mathbf{A} + \gamma \mathrm{I}) \qquad \text{with} \qquad \mathbf{A} = \frac{1}{N\sigma_x^2} \sum_{\mu=1}^{p} (\mathbf{x}_\mu)^T \mathbf{x}_\mu$$

$$\rho = \frac{2\beta}{\sqrt{N}} \sum_{\mu=1}^{p} n_\mu \mathbf{x}_\mu - 2\gamma \mathbf{w}^o. \tag{4.A6}$$

The matrix $\mathbf{A}$ is the normalised correlation matrix of the inputs, with σ_x^2 the variance of the of the sampling distribution from which the inputs are drawn (see *e.g.* section 1.2). The index k runs from one to the N dimensions of the input space. The response matrix $\mathbf{g}$ in this linear case is given by $\mathbf{g} = \frac{1}{2\beta\sigma_x^2}\Lambda$, that is $g = (\mathbf{A} + \lambda \mathrm{I})^{-1}$ where $\lambda = \gamma/\beta\sigma_x^2$.

The integral over the student weights is thus Gaussian and following integration we find that,

$$P(\mathcal{D} \mid \gamma, \beta) = \left(\frac{\beta}{\pi}\right)^{\frac{p}{2}} \left(\frac{\gamma}{\pi}\right)^{\frac{N}{2}} [\det \Lambda]^{1/2} \exp\left(\frac{1}{2}\rho^T \Lambda \rho - \gamma (\mathbf{w}^o)^T \mathbf{w}^o\right) \tag{4.A7}$$

In order to calculate the quenched averages over the training data we consider the free energy (see for example sections 2.4 or 3.4),

$$f(\mathcal{D}) = -\frac{1}{N} \ln P(\mathcal{D} \mid \gamma, \beta) \tag{4.A8}$$

Average student weight vectors

In addition to the evidence, another interesting quantity is the generalization error. For example, as defined in equation (4.2), when the teacher and student are linear this is given by,

$$\epsilon_g(\mathcal{D}) = \frac{\sigma_x^2}{N} \sum_{k=1}^{N} \langle \mathcal{R}_k \rangle_{\mathbf{w}} \langle \mathcal{R}_k \rangle_{\mathbf{w}}. \tag{4.A9}$$

Following, the notation introduced in section 2.2.2, the angle brackets with the subscript $\mathbf{w}$ denote the integration over the posterior distribution $P(\mathbf{w} \mid \mathcal{D}, \gamma, \beta)$. Thus, the generalization error depends on the difference between the student and teacher weight vectors averaged over the posterior distribution. These averages can thus be calculated from the logarithm of the evidence,

$$\frac{\partial}{\partial \rho_k} \ln P(\mathcal{D} \mid \gamma, \beta) = \langle \mathcal{R}_k \rangle_{\mathbf{w}} = \sum_{j=1}^{N} \rho_j \Lambda_{jk} \tag{4.A10}$$

Similarly, the variance of the k^{th} component of the vector $\mathcal{R}$ over the posterior is given by the second derivative of the logarithm of the evidence,

$$\frac{\partial^2}{\partial \rho_k^2} \ln P(\mathcal{D} \mid \gamma, \beta) = \langle \mathcal{R}_k^2 \rangle_{\mathbf{w}} - \langle \mathcal{R}_k \rangle_{\mathbf{w}}^2 = \Lambda_{kk} \tag{4.A11}$$

This last quantity is related to the variance of the student output over the posterior distribution and thus to the consistency measure (see *e.g.* section 2.4). Indeed, in the linear case, this variance is given by $\frac{1}{N}\sigma_x^2 \mathrm{tr}\ \Lambda$.

Determinant of response function matrix

One further quantity we will require is the logarithm of the determinant of a matrix with the form of $\mathbf{g}$. Indeed, this appears in sections 2.3, 3.4 and 4.3. Here we show that this quantity can be written in terms of the matrix itself when it is of the form $(\mathbf{A} + \lambda \mathrm{I})^{-1}$. Starting with the well known identity $\ln \det \mathbf{g}^{-1} = \mathrm{tr}\ \ln \mathbf{g}^{-1}$ and differentiating with respect to λ we find that,

$$\frac{\partial}{\partial \lambda} \ln \det \mathbf{g}(\lambda) = -\mathrm{tr}\ \mathbf{g}(\lambda). \tag{4.A12}$$

In fact, mostly we require only this derivative with respect to the weight decay since we are generally interested in the evidence assignments of this parameter (see section 2.5.2). However, in chapter 6 we require $\ln \det \mathbf{g}$ itself, which can be written,

$$\frac{1}{N} \ln \det \mathbf{g}(\lambda) = \int_{\lambda}^{\infty} \frac{1}{N} \mathrm{tr}\ \mathbf{g}(\lambda') d\lambda' - \int_{1}^{\infty} \frac{1}{\lambda'} d\lambda' \tag{4.A13}$$

In fact, for the learning scenarios examined in this thesis this integral diverges but nevertheless it is used in appendix 6.6 to find the ratio of two free energies.

Extensions from linear scenario

To date we have discussed only the linear case since this is the scenario we are concerned with in this chapter. However, in chapter 2 we calculate the free energy and other quantities such as the generalization error in the case where the data is generated by an n-teacher (see section 2.3). In this case the above derivations are somewhat similar but the initial training error is now written as,

$$E_w(\mathcal{D}) = \sum_{\Omega=1}^{n} \sum_{\mu=1}^{p_\Omega} \left(\frac{1}{\sqrt{N}} (\mathbf{w} - \mathbf{w}^\Omega) \cdot \mathbf{x}_\Omega^{\mu_\Omega} - \eta_\Omega^{\mu_\Omega} \right)^2 . \tag{4.A14}$$

4.8.3 Appendix C: average case.

Here we show that $\langle g_{ij} \rangle_{P(\mathcal{D})} = G_{\rm av}\delta_{ij}$. Firstly, we can expand $\mathbf{g}$ as,

$$g_{ij} = \lambda^{-1} - \lambda^{-2}A_{ij} + \lambda^{-3}A_{ik}A_{kj}.... \qquad \text{(4.C1)}$$
$$\text{where, } A_{ij} = \frac{1}{N\sigma_x^2}\Sigma_{\mu=1}^{p} x_i^\mu x_j^\mu$$

A typical term is then,

$$\lambda^{-(n+2)}\left(\frac{1}{N\sigma_x^2}\right)^{n+1} x_i^{\mu_1}x_{k_1}^{\mu_1}x_{k_1}^{\mu_2}x_{k_2}^{\mu_2}.....x_{k_{n-1}}^{\mu_{n-1}}x_{k_n}^{\mu_{n-1}}x_{k_n}^{\mu_n}x_j^{\mu_{n+1}} \qquad \text{(4.C2)}$$

In order to perform the average over the inputs we must pair all the indices. Ignoring, the pattern indices μ it is easy to see that any pairings of the lower indices, $i, k_1..k_n, j$, will lead to $i = j$. In order to have $i \neq j$ one index must remain unpaired and the resulting average will vanish. Thus, on average the matrix g_{ij} is diagonal.

4.8.4 Appendix D: self averaging.

In this appendix we show that quantities in equation (4.12) are $\mathcal{O}(1/N)$. Firstly, tr $\Gamma\Gamma$,

$$\text{tr } \Gamma\Gamma = \left(\frac{(\mathbf{x}_\mu)^T\mathbf{g}\mathbf{x}_\nu}{N^2\sigma_x^2} + \frac{\delta_{\mu\nu}}{N}\right)\left(\frac{(\mathbf{x}_\nu)^T\mathbf{g}\mathbf{x}_\mu}{N^2\sigma_x^2} + \frac{\delta_{\nu\mu}}{N}\right), \qquad \text{(4.D1)}$$

where repeated indices imply summation. Now the average of this, over the distribution $P(\{\mathbf{x}^\mu : \mu = 1..p\})$, can be re-expressed in terms of the average response function $G = \langle \text{tr } \mathbf{g}/N \rangle_{P(\{\mathbf{x}^\mu:\mu=1..p\})}$, which can be calculated using the method of Sollich (94) or the diagrammatic methods of Hertz *et al.*(89). Thus, we can write,

$$\langle \text{tr } \Gamma\Gamma \rangle_{P(\{\mathbf{x}^\mu:\mu=1..p\})} = \frac{1}{N}\left(\alpha - 1 + \lambda^2\partial_\lambda G\right) \qquad \text{(4.D2)}$$

Since G is $\mathcal{O}(1)$ then it is clear that $\langle \text{tr } \Gamma\Gamma \rangle_{P(\{\mathbf{x}^\mu:\mu=1..p\})}$ is $\mathcal{O}(1/N)$. Similarly $\langle \text{tr } \mathbf{y}^T\mathbf{y} \rangle_{P(\{\mathbf{x}^\mu:\mu=1..p\})}$ can also be shown to be $\mathcal{O}(1/N)$.

Finally we turn to the variance of a_{ev} over $P(\{\mathbf{x}^\mu : \mu = 1..p\})$. It is clear that,

$$\text{Var}(a_{ev}) = \sigma_x^4\lambda^4 Var\left(\frac{1}{N}(\mathbf{w}^o)^T\mathbf{g}\mathbf{w}^o\right) \qquad \text{(4.D3)}$$

Now, due to the isotropic nature of the inputs it is clear that only the magnitude of the teacher vector $\mathbf{w}^o$ is important since one could always transform the inputs to rotate

the teacher to any particular direction. Thus, we can evaluate the variance of a_{ev} by calculating the variance of $(\mathbf{w}^o)^T\mathbf{g}\mathbf{w}^o/N$ over a spherical distribution of weight vectors $\mathbf{w}^o$ constrained to be σ_w in length. We then obtain,

$$Var\left(\frac{1}{N}(\mathbf{w}^o)^T\mathbf{g}\mathbf{w}^o\right) = \frac{2\sigma_w^4}{N}((\partial_\lambda G)_0 - (G)_0^2) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{4.D4}$$

Where, again, $(h)_0$ denotes the value of h in the thermodynamic limit.

4.8.5 Appendix E: calculation of covariances

Here, as an example we calculate the correlation between λ_{ev} and λ_{opt}. From equation (4.12) we find,

$$\Delta\lambda_{ev} = -\frac{1}{\det \mathsf{M}}\{\partial_\beta^2 f\ \partial_\lambda f \ - \ \partial_\beta\partial_\lambda f\ \partial_\beta f\}_{\lambda_0,\beta_0} \tag{4.E1}$$

Where we have defined,

$$\mathsf{M} = \begin{pmatrix} \partial_\lambda^2 f & \partial_\beta\partial_\lambda f \\ \partial_\lambda\partial_\beta f & \partial_\beta^2 f \end{pmatrix} \tag{4.E2}$$

Now, we are expanding about the thermodynamic limit, that is around λ_0 and β_0. Since these are the evidence optimal assignments in this limit $\partial_\lambda f$ and $\partial_\beta f$ are of the order $\mathcal{O}(1/\sqrt{N})$. However, the second derivatives do not vanish at this point and so $\partial_\beta^2 f$ and $\partial_\beta\partial_\lambda f$ are $\mathcal{O}(1)$. Thus, expanding up to first order we obtain,

$$\Delta\lambda_{ev} = -\frac{1}{(\det \mathsf{M})_0}\{(\partial_\beta^2 f)_0\ \partial_\lambda f \ - \ (\partial_\beta\partial_\lambda f)_0\ \partial_\beta f\}_{\lambda_0,\beta_0} + \mathcal{O}\left(\frac{1}{N}\right) \tag{4.E3}$$

Similarly, from equation (4.13), we can write,

$$\Delta\lambda_{opt} = \left(-\frac{\partial_\lambda \epsilon_g}{(\partial_\lambda^2\epsilon_g)_0}\right)_{\lambda_0,\beta_0} + \mathcal{O}\left(\frac{1}{N}\right) \tag{4.E4}$$

Thus, the covariance of λ_{ev} and λ_{opt} is given by,

$$\begin{aligned}\langle\lambda_{opt}\lambda_{ev}\rangle_{P(\mathcal{D})} = -\frac{1}{(\det \mathsf{M})_0(\partial_\lambda^2\epsilon_g)_0}\{&(\partial_\beta^2 f)_0\ \langle\partial_\lambda f\ \partial_\lambda\epsilon_g\rangle_{P(\mathcal{D})} \\ &- (\partial_\beta\partial_\lambda f)_0\ \langle\partial_\beta f\ \partial_\lambda\epsilon_g\rangle_{P(\mathcal{D})}\}_{\lambda_0,\beta_0} + \mathcal{O}\left(\frac{1}{N^{3/2}}\right)\end{aligned} \tag{4.E5}$$

Now let us focus on one of these averages, namely $\langle\partial_\lambda f\ \partial_\lambda\epsilon_g\rangle_{P(\mathcal{D})}$. Firstly, using the

fact that $\langle \partial_\lambda f \mid_{\lambda_0} \rangle_{P(\mathcal{D})} = 0$ and $\langle \partial_\lambda \epsilon_g \mid_{\lambda_0} \rangle_{P(\mathcal{D})} = 0$ we can write this as the following,

$$\langle \partial_\lambda f \ \partial_\lambda \epsilon_g \rangle_{P(\mathcal{D})} = \mathrm{Cov}(\mathbf{n}^T \Gamma' \mathbf{n}, \mathbf{n}^T \Delta' \mathbf{n}) \tag{4.E6}$$
$$+\mathrm{Cov}(\mathbf{n}.\mathbf{y}', \mathbf{n}.\mathbf{z}') + \beta_0 \mathrm{Cov}(a'_{ev}, a'_{\epsilon_g}) + \mathcal{O}\left(\frac{1}{N}\right).$$

Here $h' = \partial_\lambda h$ and $\mathrm{Cov}(h(\mathcal{D}), k(\mathcal{D})) = \langle hk \rangle_{P(\mathcal{D})} - \langle h \rangle_{P(\mathcal{D})} \langle k \rangle_{P(\mathcal{D})}$, whilst the individual terms, Γ, Δ *etc...* are defined in equations (4.4) and (4.5). Equation (4.E6) can then expressed in terms of the response function as we saw in appendix 4.8.4. The second term, $\langle \partial_\beta f \ \partial_\lambda \epsilon_g \rangle_{P(\mathcal{D})}$, is similar.

Chapter 5

Model Selection by Estimating the Expected Error

Abstract

In this chapter we consider setting the weight decay parameter in a linear student learning from examples generated by a noisy linear teacher using empirical estimates of the generalization error. In the context of finite size corrections to the thermodynamic limit we examine two methods for achieving this, one based on the test error on an independent set, the other on a cross-validatory estimate of the generalization error. In the former we consider a scenario in which we must choose how to partition our data base into test and training sub-sets. We consider two partitioning methods and find that one, the minimal variance partition, requires that the fraction of data used for testing approaches unity as the data base grows in size. The other, optimal partitioning, results in a degradation in generalization performance, compared to the true optimal, which is an order $\mathcal{O}(\sqrt{N})$ worse than that associated with the evidence procedure. The particular version of cross-validation we consider is leave-one-out cross-validation (**CV(1)**). In contrast to the test error we find that, at added computational cost, **CV(1)** makes good use of the test data points resulting in a degradation in performance of the same order as the evidence. In fact, comparison of the evidence and leave-one-out cross-validation reveals the performance of the latter to be superior except in the noiseless case, when the number of examples exceeds the number of model parameters, and in the asymptotic regime where they are equivalent. However, it should be noted that the enhanced performance of **CV(1)** is achieved through greater computational effort as compared with the evidence procedure.

5.1 Introduction

In this chapter we analyse the performance of model selection methods based on the idea of estimating the generalization error directly. The process of model selection then proceeds by choosing the model, or models, with the lowest estimated error. In particular, we examine the test error and leave-one-out cross-validation, **CV(1)** (Stone 74). The test error is evaluated on a set of data independent from that used to train the student networks and is, thus somewhat wasteful of the available data. Cross-validation methods in general, and **CV(1)** in particular, are an attempt to overcome this drawback without encountering the attendant problem of underestimating the true error. Such an underestimation of the generalization error is caused by testing the students on data also used in the training set. Indeed, in the limiting case where all the data is used both for training and testing the test error is simply the training error and must be a biased estimate of the generalization error; recall that minimizing the training error can result in overfitting.

In particular, we will focus on the noisy, but learnable, linear case where weight decay is employed as a regularizer (see chapter 4). We will explore the behaviour of weight decay assignments based on the test and cross-validated errors, evaluate the performance of each method and make comparisons with that of the evidence procedure. As we shall shortly see, all three methods assign the optimal value to the weight decay parameter in the thermodynamic limit. Similarly, in the sense used in section 4.3.1, all three methods are unbiased. Thus, we will be chiefly concerned with finite size effects as in the previous chapter. Once again, focusing on this regime will reveal important behaviour not apparent in the thermodynamic limit. In particular, it will enable us to meaningfully compare all three algorithms in terms of performance, identifying their strengths and weaknesses.

In the next section we calculate the test error assignment of the weight decay and evaluate its variance and correlation with the optimal regularizer in two different scenarios. The first, in which the test set is separate from the training set, is used to explore the properties of the test error assignment without undue complication. The second is a more realistic situation in which we must decide how best to partition a data base into training and testing sub-sets. In the later case we base this partition on two criteria, namely minimising the variance of the weight decay assignment and optimising the performance. We compare this optimal performance to that found for the evidence in chapter 4.

In section 5.3 we calculate the popular leave-one-out cross-validation error, again for the learnable linear case. We show that this quantity is self averaging, find the

weight decay assignments made from it and calculate their correlation with the optimal assignments. We explore the implications of these assignments on the performance of **CV(1)** and finally, in section 5.4, we compare this performance with that associated with the evidence assignments calculated in chapter 4.

5.2 Model Selection from the test error

As before, our student, denoted by its 1-D output $y_s(\mathbf{x})$ for an input $\mathbf{x}$ in N dimensional space, is distributed according to the predictive distribution, $P(y_s(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}_{trn}, \mathcal{M})$. Here, $\mathcal{D}_{trn} = \{(y_t(\mathbf{x}^\nu), \mathbf{x}^\nu) : \nu = 1...l\}$ is the data set used to train the student and $\mathcal{M}$ denotes the form of the student model, including hyper-parameters such as the weight decay (see section 1.2.1). In particular, here we will adopt the noise model and prior outlined in section 2.2.1. Since the architecture will remain fixed, as previously, the model is specified by the inverse temperature, β, and the weight decay, λ, thus $\mathcal{M} = \{\beta, \lambda\}$. In Bayesian tradition and as we have done through out this thesis, we choose to make predictions based on the average over this predictive distribution. We also comment that, if we believe our model is correct, as is implicit in the Bayesian paradigm, then for a least squares loss criterion, such as those defined in equations (1.10) and (1.11), this average is also the optimal predictor. The teacher, with 1-D output $y_t(\mathbf{x})$, is drawn from the distribution $P(y_t \mid \mathbf{x})$ and we consider the case of random examples where the inputs are sampled from a distribution $P(\mathbf{x})$. The test error is then given by,

$$\epsilon_{test} = \frac{1}{m} \sum_{\mu=1}^{m} \left(< y_s(\mathbf{x}^\mu) >_{P(y_s(\mathbf{x})|\mathbf{x},\mathcal{D}_{trn},\mathcal{M})} - y_t(\mathbf{x}^\mu) \right)^2, \tag{5.1}$$

Where m is the number of test points available in a finite test set $\mathcal{D}_{tst} = \{(y_t(\mathbf{x}^\mu), \mathbf{x}^\mu) : \mu = 1...m\}$, which is drawn independently of the training set, $\mathcal{D}_{trn}$. Although one could consider alternative scenarios the most realistic case, and that which we examine here, is when the test points are drawn *i.i.d* from the same distribution as the training set. The total data base consisting of p examples, with m from the test set and l in the training set, is denoted $\mathcal{D}$.

In the linear case, with weight decay, the test error translates to,

$$\epsilon_{test} = \frac{1}{m} \sum_{\mu=1}^{m} \left(\frac{1}{\sqrt{N}} (< \mathbf{w} >_{P(\mathbf{w}|\mathcal{D}_{trn},\lambda,\beta)} - \mathbf{w}^o) \cdot \mathbf{x}^\mu - \eta^\mu \right)^2. \tag{5.2}$$

Here the average over the predictive distribution translates into an average of the, N dimensional, student parameters, $\mathbf{w}$, over the posterior $P(\mathbf{w} \mid \mathcal{D}_{trn}, \lambda, \beta)$. Furthermore, we have assumed that the linear teacher, with weights $\mathbf{w}^o$, is corrupted by uncorrelated Gaussian noise of mean zero and variance σ^2. The random variables η^μ denote an instantiation of this noise in example μ and are thus distributed accordingly. Similarly, in what follows, we assume that $P(\mathbf{x})$ is normal with mean zero and variance σ_x^2. If we average over the *test points* we find that the test error is a noisy estimate of the data dependent generalization error,

$$\langle \epsilon_{test}(\mathcal{D}) \rangle_{P(\mathcal{D}_{tst})} = \epsilon_g(\mathcal{D}_{trn}) + \sigma^2. \tag{5.3}$$

Thus, as we saw in section 4.3.1 for the evidence, the test error assignments will be unbiased in the sense that the average test error is minimised by the same weight decay assignment as the average generalization error. Once again this result is valid for any system size N. Also, as we did in section 4.4 for the evidence, we can examine the parameter assignments made from the test error based on a particular data set. In particular, we consider the thermodynamic limit where the number of training examples, $l = \alpha_{trn} N$, scales linearly with the system size. Recall that in chapter 4 we found that if the number of training examples was not extensive the thermodynamic limit broke down with key quantities, such as the variance in the optimal weight decay, diverging (see sections 4.4 and 4.5). Then, if the number of test examples, m, also scales with the system size N where $m \propto N^s$, with $0 < s \leq 1$, is the scaling law, it can be shown that the test error is self averaging (Barber *et al.*(94)). Thus, in the thermodynamic limit we find that the average case corresponds to any particular case and the test error assignments coincide with the optimal. We allow a wider range of scaling behaviour for the test set size than for the training set size since, as just stated the training set size must be order $\mathcal{O}(N)$ for the thermodynamic limit to make sense and latterly we will investigate optimal test set sizes.

Furthermore, for large N, we are justified in Taylor expanding the equation defining the test error optimal weight decay, $\lambda_{tst}(\mathcal{D})$,

$$\partial_\lambda \epsilon_{tst}(\mathcal{D}) \approx \partial_\lambda \epsilon_{tst}(\mathcal{D}) \mid_{\lambda_0} + \Delta\lambda \partial_\lambda^2 \epsilon_{tst}(\mathcal{D}) \mid_{\lambda_0} = 0. \tag{5.4}$$

Then, up to $\mathcal{O}(1/\sqrt{m})$ the test error optimal weight decay can be written as $\lambda_{tst} \approx$

$\lambda_0 + \Delta\lambda_{tst}(\mathcal{D})$ where

$$\Delta\lambda_{tst}(\mathcal{D}) = -\frac{\partial_\lambda \epsilon_{tst}(\mathcal{D}) \mid_{\lambda_0}}{\partial_\lambda^2 \epsilon_{tst}(\mathcal{D}) \mid_{\lambda_0}} \tag{5.5}$$

Using the methods introduced in chapter 4 we can then examine the variance of this fluctuation as well as its correlation with other quantities, such as the optimal weight decay $\lambda_{opt}(\mathcal{D})$. We now go on to explore weight decay assignments from the test error in two different scenarios. The first is somewhat unrealistic assuming, as it does, that we have a data set containing $l = \alpha_{trn}N$ examples available for training our student and independent of this we have a data set of m examples which *may only be used for testing.* Latterly, we will consider the more usual situation in which we have a single data base, of $p = \alpha N$ examples, which we must choose how to partition into training and testing sets.

5.2.1 Separate testing and training sets

As stated, here we will examine the situation in which we have a totally separate set of data which can only be used for testing. This will allow us to explore the properties of the test set estimate of the weight decay, $\lambda_{tst}(\mathcal{D})$ without complicating considerations concerning the optimal partitioning of the data base.

The calculations of the variance of $\lambda_{tst}(\mathcal{D})$ and its covariance with the optimal weight decay, when we have $l = \alpha_{trn}N$ examples available for training, are similar to those of chapter 4. In particular, we find that the variance in the test error weight decay assignment is of the form,

$$\mathrm{Var}(\lambda_{tst}) = \mathrm{Var}(\lambda_{opt}) + \frac{1}{m}(\mathrm{f}_v(\lambda_0, \alpha_{trn}))^2 \tag{5.6}$$

where $\mathrm{f}_v(\lambda_0, \alpha_{trn})$ is positive order $\mathcal{O}(1)$ quantity in the thermodynamic limit. Thus, for any finite test set, $\lambda_{tst}(\mathcal{D})$ is a noisy estimate of the optimal weight decay. In the limit of large training sets we find that,

$$\lim_{\alpha_{trn}\to\infty} \mathrm{Var}(\lambda_{tst}) = \frac{\lambda_0(2+\lambda_0)\alpha_{trn}}{2Nc} + \frac{\lambda_0\alpha_{trn}}{N}, \tag{5.7}$$

when the test set size, m, scales linearly with the training set size, l, (*i.e.* , $m = \alpha_{tst}N$ and $\alpha_{tst} = \alpha_{trn}c$). The second term is the variance of the optimal weight decay in the limit of large training set (see equation 4.15). Thus, the variance in λ_{tst} diverges linearly with α_{trn} and even in this limit the test error assignment remains a noisy

estimate of the optimal weight decay unless the test set is infinite (*i.e.* c diverges; $\alpha_{tst} \gg \alpha_{trn}$). In contrast, when the test set size $m \propto N^s$ does not scale with the number of training examples we find that the variance of the test error optimal weight decay is order $\mathcal{O}(N^{-s})$ and diverges even faster with α_{trn},

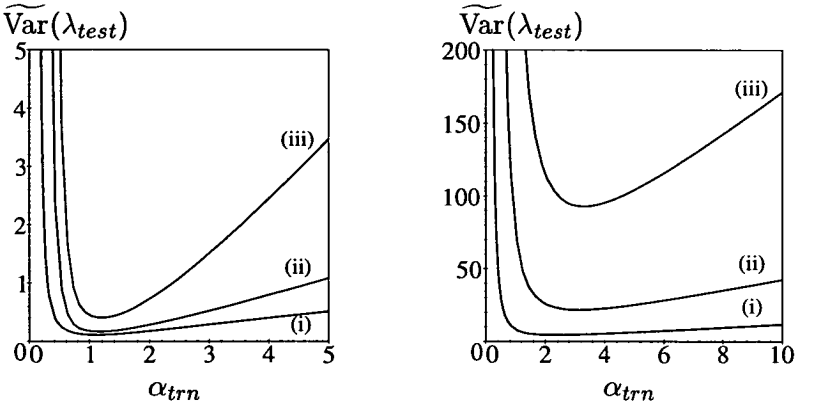
$$\lim_{\alpha_{trn}\to\infty} \mathrm{Var}(\lambda_{tst}) = \frac{\lambda_0(2+\lambda_0)\alpha_{trn}^2}{2m}. \tag{5.8}$$

Note that we return to the linear divergence with α_{trn} found in equation (5.7) if we substitute $m = \alpha_{trn}Nc$ into the above expression. Thus, as the number of training examples increases we must use more and more test points to achieve the same accuracy in estimating the weight decay, although as the generalization error improves for large and increasing α_{trn} the gains in identifying the optimal weight decay reduce. Furthermore, equation (5.7) in particular implies that the test error is even less sensitive to the weight decay than is the generalization error. Indeed, in chapter 4 we saw similar behaviour where, asymptotically, the evidence weight decay assignments became increasingly distant from the optimal as the number of training examples increased.

The variance in λ_{tst} scaled by N, $\widetilde{\mathrm{Var}}(\lambda_{tst})$, is shown in figure 5.1 for finite α_{trn} for two different noise levels and a number of different test set sizes, all of which scale linearly with α_{trn} and the system size. As we saw in chapter 4, for the optimal variance, as the number of training examples decreases towards zero the scaled variance diverges signalling a breakdown of the thermodynamic limit. Similarly, the linear divergence, for large α_{trn} is clearly evident in all the curves and naturally, the variances are larger when the noise levels are greater. This reflects the increase in the variance of the optimal assignment as well as that in the estimation of the generalization error by the test error. In addition, we note that, for a given noise level, the variance in $\lambda_{tst}(\mathcal{D})$ reduces as the size, m, of the test set increases. Thus, as can be seen from equation (5.6), to estimate the optimal weight decay as accurately as is possible one needs an arbitrarily large test set.

Now we turn to the normalised correlation, $C(\lambda_{tst}, \lambda_{opt})$, of the assignment of the weight decay from the test error with that of the optimal. In the case, under consideration here, where we have a completely separate test set the covariance between λ_{tst} and λ_{opt} takes on the simple form,

$$\mathrm{Cov}(\lambda_{tst}, \lambda_{opt}) = \mathrm{Var}(\lambda_{opt}). \tag{5.9}$$



(a) $\lambda_0 = 0.1$ (b) $\lambda_0 = 1$

Figure 5.1. Variance of λ_{test}. In graph (a) the noise to signal ratio is $\lambda_0 = 0.1$ and the test set size $m = \alpha_{trn}Nc$, with (i) $c = 10$, (ii) $c = 1$ and (iii) $c = 0.1$. In graph (b) the noise to signal ratio is unity with the test set sizes as before

Thus the correlation function is,

$$C(\lambda_{tst}, \lambda_{opt}) = \left[\frac{\mathrm{Var}(\lambda_{opt})}{\mathrm{Var}(\lambda_{tst})}\right]^{1/2}, \tag{5.10}$$

and minimising the variance, $\mathrm{Var}(\lambda_{tst}(\mathcal{D}))$, maximises the correlation. In the limit of large α_{trn} we find,

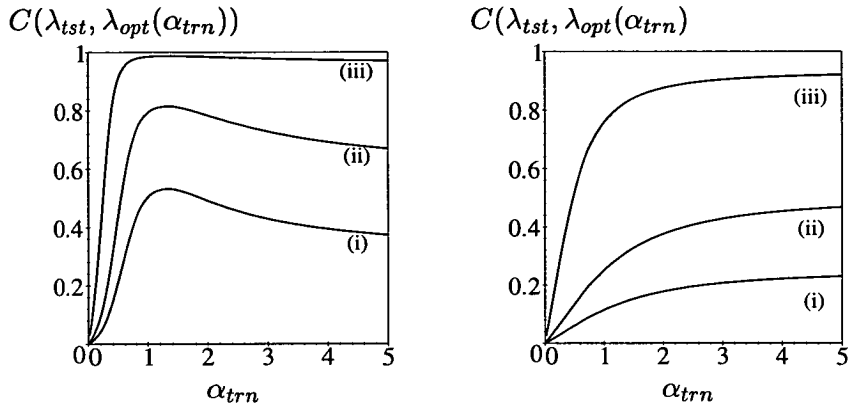
$$\lim_{\alpha_{trn}\to\infty} C(\lambda_{tst}, \lambda_{opt}) = \left[\frac{2c}{\lambda_0 + 2(1+c)}\right]^{1/2}, \tag{5.11}$$

when $m = \alpha_{trn}Nc$. However, if m does not scale with the number of training examples then we find that correlation decays as $\mathcal{O}(1/\sqrt{\alpha_{trn}})$,

$$\lim_{\alpha_{trn}\to\infty} C(\lambda_{tst}, \lambda_{opt}) = \left[\frac{2m}{N\alpha_{trn}(\lambda_0 + 2)}\right]^{1/2}. \tag{5.12}$$

Also, since $m \approx \mathcal{O}(N^s)$, for $s < 1$ the correlation is lower than order $\mathcal{O}(1)$.

The normalised correlation $C(\lambda_{tst}, \lambda_{opt})$ is shown, for the case where the test set size scales linearly with that of the training set, $l = \alpha_{trn}N$, in figure 5.2. It is apparent that, as the the test set size increases relative to, but not at the expense of, the training set (c increases) the correlation increases. Furthermore, as the noise level increases, for fixed test set size, the degree of correlation reduces, In fact, in the limit of large c, $C(\lambda_{tst}, \lambda_{opt}) \to 1$ and for the case of zero noise in the large training set limit ($\alpha_{trn} \to \infty$)



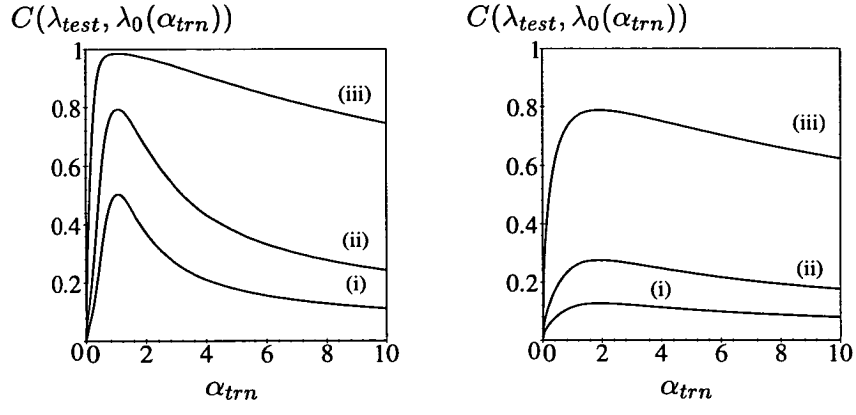
(a) $\lambda_0 = 0.1$ (b) $\lambda_0 = 1$

Figure 5.2. Normalised correlation of λ_{test} with the optimal weight decay versus the number of training examples. The noise levels are shown in the captions. In both (a) and (b) the size of the test set scales with α_{trn}, $m = \alpha_{trn} Nc$ and the fraction of test examples to training examples is (i) $c = 0.1$, (ii) $c = 1$ and (iii) $c = 10$ respectively. The explicit dependence of λ_{opt} on α_{trn} emphazises the fact that here the optimal is defined in terms of the number of training examples and not the total number of examples.

we find $C(\lambda_{tst}, \lambda_{opt}) \to \sqrt{c/(1+c)}$. Finally, for the finite values of c shown, we can see that the correlation tends towards non-zero asymptotic values as α_{trn} increases in accord with equation (5.11).

In figure 5.3 we see again the normalised correlation but now in the case where the size of the test set does not scale with the number of training examples. We witness broadly similar behaviour to the previous case, with the correlation increasing, for fixed α_{trn} as the size of the test set increases and reducing as the noise level increases. However, as expected from equation (5.12), as α_{trn} increases the correlation decays rather than asymptoting to some non-zero value.

An alternative measure of the similarity between the test error weight decay assignment and the optimal is the average squared distance between them, $< (\lambda_{opt}(\mathcal{D}) - \lambda_{tst}(\mathcal{D}))^2 >_{P(\mathcal{D})}$. A similar measure was introduced in section 4.4, however there we normalised this distance by the variance in the optimal weight decay (see equation 4.18). The rationale behind this being that this distance diverged for large training sets due to the divergence in the optimal weight decay; by normalising we obtained a measure of the distance from the optimal as a fraction of the uncertainty in the optimal, and avoided this divergence. Since $< (\lambda_{opt}(\mathcal{D}) - \lambda_{tst}(\mathcal{D}))^2 >_{P(\mathcal{D})}$ also diverges for large



(a) $\lambda_0 = 0.1$ (b) $\lambda_0 = 1$

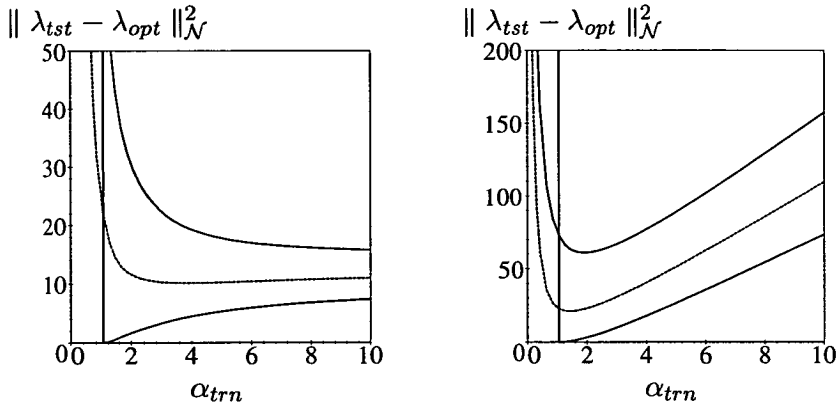
Figure 5.3. The normalised correlation of λ_{test} with the optimal weight decay when the test set size, m, does not scale with the number of training examples, l (in this case $m = \alpha_{tst}N$, with α_{tst} const. as $\alpha_{trn} \to \infty$). The noise levels are shown in the captions and in both (a) and (b) the size of the test set is (i) $\alpha_{tst} = 0.1$, (ii) $\alpha_{tst} = 1$ and (iii) $\alpha_{tst} = 10$.

data sets we adopt the normalised distance measure as before,

$$\| \lambda_{tst} - \lambda_{opt} \|^2_{\mathcal{N}} = \frac{< (\lambda_{opt}(\mathcal{D}) - \lambda_{tst}(\mathcal{D}))^2 >_{P(\mathcal{D})}}{\mathrm{Var}(\lambda_{opt})}. \tag{5.13}$$

$$= \frac{\mathrm{Var}(\lambda_{tst}) - \mathrm{Var}(\lambda_{opt})}{\mathrm{Var}(\lambda_{opt})}.$$

Figure 5.4 shows the normalised distance measure for the cases when the test set size does and does not scale with the training set size. In the latter case even the normalised distance diverges whilst in the former $\| \lambda_{tst} - \lambda_{opt} \|^2_{\mathcal{N}} \to (2 + \lambda_0)/2c$ in the large α_{trn} limit. Comparing with the evidence case (see equations 4.17 & 4.18) we see that for large enough test set, $m > \alpha_{trn} N c^{crit}$, where $c^{crit} = 1 + \lambda_0/2$, the test set assignment is asymptotically closer to the optimal than that of the evidence. We also note that in the limit of infinite test set size the distance, $\| \lambda_{tst} - \lambda_{opt} \|^2_{\mathcal{N}} \to 0$ for all α_{trn}. For finite test sets, the *un-normalised* distance from the optimal diverges for zero noise at $\alpha_{trn} < 1$ and is zero for larger training set size. That this behaviour is not fully reflected in figure 5.4 is due to the rates at which $< (\lambda_{opt}(\mathcal{D}) - \lambda_{tst}(\mathcal{D}))^2 >_{P(\mathcal{D})}$ and $\mathrm{Var}(\lambda_{opt})$ approach zero (for $\alpha_{trn} > 1$) as $\lambda_0 \to 0$. Indeed, asymptotically, for zero noise, the normalised distance equals c^{-1} as opposed to the un-normalised distance which is zero for $\alpha_{trn} > 1$ in the noiseless case.



(a) $m \propto l$ (b) m constant

Figure 5.4. Average normalised distance between $\lambda_{tst}(\mathcal{D})$ and $\lambda_{opt}(\mathcal{D})$: In graph (a) the test set size $m = \alpha_{trn}Nc$, scales with the training set size $l = \alpha_{trn}N$ with $c = 0.1$ and the three curves shown are as follows; The upper full curve is for $\lambda_0 = 1$ whilst the lower solid curve is for zero noise level and the dotted (middle) curve has $\lambda_0 = 0.25$. In graph (b) the test set size $m = Nc$ does not scale with the training set size. The example shown has $c = 0.1$ and the noise levels as in graph (a). Clearly, unless the test set scales with the training set the normalised distance of the test error weight decay assignment from the optimal diverges with α_{trn}. In both cases, for zero noise and $\alpha_{trn} < 1$, this distance diverges, whilst it is finite for $\alpha_{trn} > 1$. It should be noted, however, that the true (un-normalised) distance between the test error assignment and the optimal does tend to zero, for $\alpha_{trn} > 1$, in the noiseless limit but at the same rate as $\mathrm{Var}(\lambda_{opt})$.

Performance

Finally, we look at the effect of the test set weight decay assignment on performance. Recall that our measure of a particular algorithms performance, $\kappa_{\epsilon_g}(\lambda_{alg}, \mathcal{D})$ has been the degradation in generalization ability incurred in using it, as a fraction of the optimal generalization error (see *e.g.* section 4.3.1);

$$\Delta\epsilon(\lambda_{alg}, \mathcal{D}) = \epsilon_g(\lambda_{alg}, \mathcal{D}) - \epsilon_g(\lambda_0, \mathcal{D}) \text{ and } \kappa_{\epsilon_g}(\lambda_{alg}, \mathcal{D}) = \frac{\Delta\epsilon(\lambda_{alg}, \mathcal{D})}{\epsilon_g(\lambda_0, \mathcal{D})}. \tag{5.14}$$

In the case under current consideration $\lambda_{alg} = \lambda_{tst}$. Furthermore, in this chapter we will focus on the average performance, rather than concern ourselves with fluctuations around it. That is, we will focus on $< \kappa_{\epsilon_g}(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})}$ and not take into account the variance of this quantity as we did in chapter 4. Indeed, for the evidence, we found that for a small amount of training data these fluctuations were small compared

with the average fractional degradation (see *e.g.* figure 4.10). In addition, although they were seen to become relatively important for mid-sized training sets they did not qualitatively effect the behaviour of $\kappa_{\epsilon_g}(\lambda_{ev})$. Furthermore, asymptotically, as the number of training examples grew, the fluctuations were seen to become unimportant as we would expect.

Similarly to equation (4.24), in chapter 4, the average degradation in performance, to first order in N, is given by,

$$< \Delta\epsilon(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})} = \frac{1}{2}\partial_\lambda^2 \epsilon_g(\lambda_0) \left(\mathrm{Var}(\lambda_{tst}) - \mathrm{Var}(\lambda_{opt})\right). \tag{5.15}$$

Thus, as the test set becomes larger and $\mathrm{Var}(\lambda_{tst})$ approaches $\mathrm{Var}(\lambda_{opt})$ (from above) the performance approaches the optimal. However, as we found for the evidence procedure assignments (see section 4.5), in general, even in the limit of large α_{trn}, the average fractional degradation, $\kappa_{\epsilon_g}(\lambda_{tst})$ does not decay to zero. In fact, for the limit of large training set we find that when the test set size scales with that of the training set (*i.e.* $m = \alpha_{trn} N c$),

$$\lim_{\alpha_{trn}\to\infty} < \kappa_{\epsilon_g}(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})} = \frac{2 + \lambda_0}{2Nc}. \tag{5.16}$$

In comparison with our results of section 4.5, for the evidence procedure, we find that the performance resulting from this, rather idealized, use of test error is superior for large enough test sets. In particular, this method out performs the evidence if the test set size $m > \alpha_{trn} N c^{crit}$, where $c^{crit} = 1 + \lambda_0/2 \geq 1$ (see equation 4.25); the result revealing the dependence of the fractional degradation on the distance measure, $\| \lambda_{tst} - \lambda_{opt} \|^2$. We note that, since the test set is separate from the training set, this implies that the test error assignment of the weight decay will only out-perform that of the evidence by using more than twice the data. Indeed, when the test set size does not scale with the training set size, l, the performance is worse than associated with the evidence procedure, with the fractional degradation diverging linearly with α_{trn},

$$\lim_{\alpha_{trn}\to\infty} < \kappa_{\epsilon_g}(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})} = \frac{\alpha_{trn}(2 + \lambda_0)}{2m}. \tag{5.17}$$

In summary then, we have found that, in the learnable linear case, the test error allows us to make a noisy prediction of the true optimal weight decay, the degree of error diminishing as the size of the test set increases. Furthermore, we have seen that as the training set grows in size it becomes increasingly difficult to estimate the optimal weight decay requiring an ever larger test set to achieve the same accuracy, but of course the generalization improves and becomes ever more insensitive to the assignment of the

weight decay. Finally, both these features were seen to be evident in the performance resulting from using the test error estimate for the regularization parameter.

5.2.2 Partitioning the data base

We now turn to the problem of partitioning a data base, of $p = \alpha N$ examples, into testing and training sets. In the previous section we were forbidden from using the test set data for training, but in practice no such such proscription exists. However, as discussed in the introduction, in order not to produce a biased estimate of the true error we should not use training data in the test set, or vice versa. As we saw earlier, for a given training set size, the accuracy of the test error assignment of the weight decay increases with the size of the test set, but now increasing the test set reduces the data available for training. Similarly, devoting more data to training reduces the accuracy of the test error. Thus, there is a trade off between these two activities. In this section we look at two criteria one might use to identify the optimal partition of the data base into test and training sets.

Minimal variance

The first criterion we consider is that of minimising the variance of the test error assignment of the weight decay. That is, we want to identify the optimal weight decay as accurately as possible. As a practical option we should note that this policy might be difficult to implement. Nonetheless, this variance is dependent on the student and training data only and so one might imagine estimating it using separate data bases, however, we do not deal with this problem here. In mitigation we remark that, from a practical view point, it is more useful as method of partitioning the data base than our ultimate goal of minimising generalization error to which we do not have direct access.

Firstly, let us consider the form of the variance in equation (5.6) recalling that $\mathrm{Var}(\lambda_{opt})$ is an order $\mathcal{O}(1/N)$ quantity when the training set size is extensive. If the test set size m is of the order of N, (*i.e.* s=1 and $m \approx \mathcal{O}(N)$) then both terms in the variance (see equation 5.6) are also order $\mathcal{O}(1/N)$ whilst, if m is $\mathcal{O}(1)$ then the variance is also $\mathcal{O}(1)$. Likewise, for $s < 1$ the variance in λ_{tst} is larger than $\mathcal{O}(1/N)$ (*e.g.* $s = 0.5 \rightarrow \mathrm{Var}(\lambda_{tst}) \approx \mathcal{O}(N^{-1/2})$). Thus, since we have fixed the total number of examples to be $\mathcal{O}(N)$, the minimal variance will be achieved for test set of size $m = \alpha_{tst}N$. Thus, in this case we have $\alpha = \alpha_{trn} + \alpha_{tst}$ as the number of training examples, $l = \alpha_{trn}N$, must also be of the order of N for a variance of $\mathcal{O}(1/N)$ (see divergence as $\alpha_{trn} \rightarrow 0$ in figure 5.1).

Figure 5.5(a) shows the minimal variance fraction of test examples $\Delta m = \alpha_{tst}/\alpha$

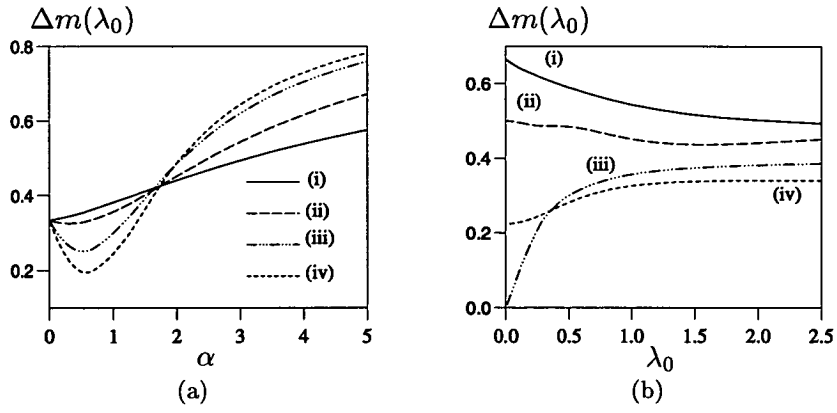


Figure 5.5. Minimal variance fraction: Partition of data set into training and testing subsets resulting in minimal variance of $\lambda_{tst}(\mathcal{D})$: The left-hand figure shows the fraction that should be devoted to testing as a function of the total number of examples available, α for (i) $\lambda_0 = 0.01$, (ii) $\lambda_0 = 0.1$, (iii) $\lambda_0 = 0.25$ and (iv) $\lambda_0 = 1$. The right-hand graph shows the same optimal fraction versus λ_0 for (i) $\alpha = 0.5$, (ii) $\alpha = 1$, (iii) $\alpha = 1.2$ and (iv) $\alpha = 1.5$. Note, that as α increases we should devote a progressively larger fraction to testing.

versus the normalised total number of examples α for different noise levels (λ_0). For $\alpha \to 0$ minimal variance partition is when $\Delta m = 1/3$ which corresponds to using half as much data for testing as for training. In contrast, as α increases, Δm tends to unity. Thus, when only a small amount of data is available a relatively small test set should be used, but where a large data base is used most of the data should be used for testing. The fact, that $\Delta m \to 1$ still means that $\alpha_{trn} \to \infty$ but that it is a progressively smaller fraction of α. At small values of α the minimal variance fraction is a decreasing function of the noise level. Figure 5.5(b) shows the same fraction versus the noise level, λ_0, for various values of α. For large noise level the minimal variance fraction is a monotonically increasing function of α which tends to one.

The result that $\Delta m \to 1$ as $\alpha \to \infty$ is similar to that obtained by Shao (93) in the context of cross-validation. Shao's results suggest that cross-validation will only be consistent if, as the total size of the data base (p) increases, a relatively large fraction (m/p) of data should be used in the validation set (similar to our test set), compared to the training set, such that $m/p \to 1$ as $p \to \infty$. Thus, as we find, in order to hit the target, as more data becomes available, a greater fraction is required for testing (validation).

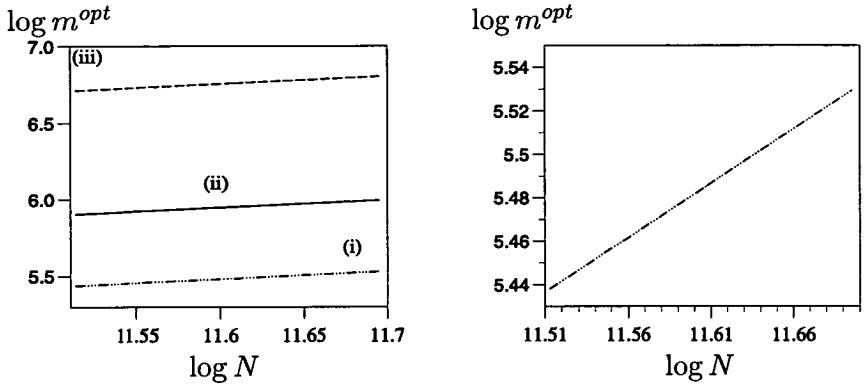


Figure 5.6. Optimal scaling: The logarithm of optimal test set size, which minimises the average degradation in performance, versus the log of the system size N. The left-hand figure shows the optimal test size for noise to signal levels (i)$\lambda_0 = 0.01$, (ii)$\lambda_0 = 0.25$ and (iii)$\lambda_0 = 1$ all for $\alpha = 2$. Other values of α show similar behaviour. The right-hand graph is curve (i) on a smaller scale, revealing that the scaling law is $m^{opt} \approx \mathcal{O}(N^{1/2})$.

Optimal partition

As noted in the last section our ultimate goal is the minimization of the generalization error. What partition of the data base will achieve this? Clearly, this is not something we can do in practice directly, but nevertheless it would be interesting to know the optimal policy.

Firstly, let us identify the optimal scaling of the test set size m with the system size N (*i.e.* find s_{opt}). Examining equation (5.14), for the average degradation in performance ($\Delta\epsilon$), we note that the optimal generalization is now achieved by setting the weight decay optimally and using all the p examples available as the training set. Thus, if we devote $m \approx \mathcal{O}(N)$ examples to testing the average degradation will be $\mathcal{O}(1)$. However, if $m \approx \mathcal{O}(N^s)$ with $0 < s < 1$ then then for large system size N we can expand, to order $\mathcal{O}(1/N)$, the average degradation as,

$$\begin{aligned}
< \Delta\epsilon(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})} &= < \epsilon_g(\lambda_0 + \Delta\lambda_{tst}, \alpha - m/N) >_{P(\mathcal{D})} - < \epsilon_g(\lambda_0 + \Delta\lambda_{opt}, \alpha) >_{P(\mathcal{D})} \\
&\approx (\epsilon_g(\lambda_0, \alpha))_0 - \frac{m}{N} (\partial_\alpha \epsilon_g(\lambda_0, \alpha))_0 + \frac{1}{2} \left(\partial_\lambda^2 \epsilon_g(\lambda_0)\right)_0 (\mathrm{Var}(\lambda_{tst}) - \mathrm{Var}(\lambda_{opt})) \\
&\quad - (\epsilon_g(\lambda_0, \alpha))_0 + \mathcal{O}\left(\frac{1}{N^{3/2}}\right) \\
&\approx -\frac{m}{N} (\partial_\alpha \epsilon_g(\lambda_0, \alpha))_0 + \frac{1}{2} \left(\partial_\lambda^2 \epsilon_g(\lambda_0)\right)_0 (\mathrm{Var}(\lambda_{tst}) - \mathrm{Var}(\lambda_{opt})) + \mathcal{O}\left(\frac{1}{N^{3/2}}\right) \quad (5.18)
\end{aligned}$$

Here we adopt the notation of chapter 4 whereby, $(h)_0$ denotes the value of the function h in the thermodynamic limit. Since, $\mathrm{Var}(\lambda_{opt})$ is order $\mathcal{O}(1/N)$ and $\mathrm{Var}(\lambda_{tst})$ is order $\mathcal{O}(1/m)$ then $m \approx \mathcal{O}(N^s)$ with $s < 0.5$ will give an average degradation of $\mathcal{O}(1/N^s)$ whilst for $s > 0.5$ we find $\Delta\epsilon \approx \mathcal{O}(1/N^{1-s})$. Thus, the minimal average degradation attainable is $\mathcal{O}(N^{-1/2})$ and occurs for a test set of size $m \approx \mathcal{O}(N^{1/2})$. We note that Barber *et al.*(94) found alternative scaling laws for the optimal test set size when, at fixed weight decay, trading off generalization performance with the ability to predict that performance. In fact, they found two scaling laws depending on whether the regularization was too strong or weak. That only one scaling law exists here is due to the fact that we are optimising the weight decay and thus, to first order in N, the generalization error is a monotonically decreasing function of α_{trn}.

Figure 5.6 shows the log of the optimal test set size, found numerically, versus the log of the system size for large systems. The left hand graph shows that, as we would expect, the scaling law does not vary over a range noise to signal levels. The right hand graph shows more clearly that the scaling law is $s = 0.5$.

Given that the optimal test set size scales with $N^{1/2}$ we can now calculate, to first order, the optimal test set size. That is, writing $m = bN^{1/2}$, we find b^{opt} which minimises,

$$\begin{aligned} &< \Delta\epsilon(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})} \\ &= -\frac{b}{N^{1/2}} \left(\partial_\alpha \epsilon_g(\lambda_0, \alpha) - \frac{\partial_\lambda^2 \epsilon_g(\lambda_0)_0 (\mathrm{f}_v(\lambda_0, \alpha_{trn}))^2}{2b^2} \right). \end{aligned} \tag{5.19}$$

The above equation is obtained by combining equations (5.6) and (5.18). The optimal partition b^{opt}, which has a relatively simple form, is shown in 5.7. As α tends to zero b^{opt} diverges like $\mathcal{O}(1/\sqrt{\alpha})$. Similarly, for a large number of examples, but $\alpha << N$, we find that the optimal test set size scales with α,

$$\lim_{\alpha \to \infty} b^{opt} = \frac{\alpha\sqrt{2 + \lambda_0}}{\sqrt{2}} \tag{5.20}$$

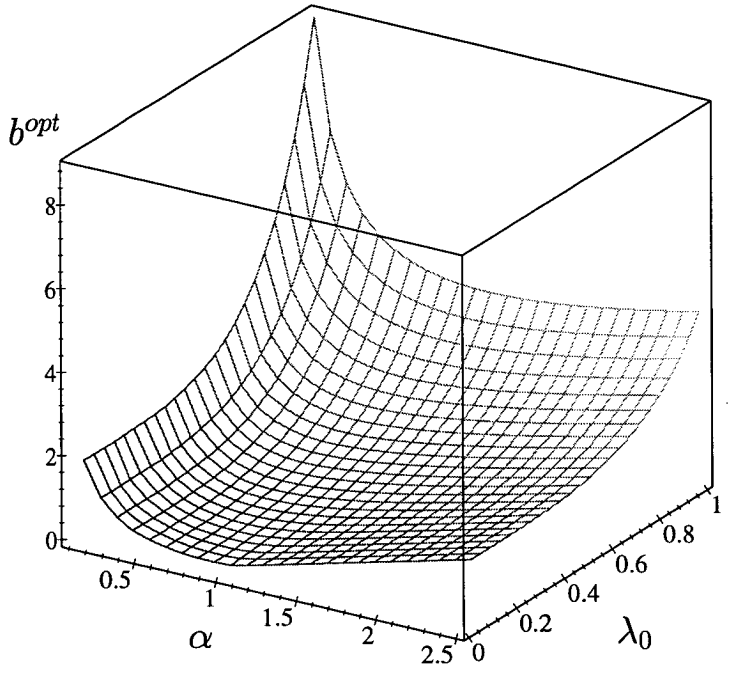


Figure 5.7. Optimal test set size scaled by $N^{1/2}$: The optimal number of test examples diverges linearly with α. For a small data base, $\alpha \to 0$, the optimal fraction of test examples tends to order $\mathcal{O}(1)$ at a rate of $p^{-1/2}$. The precursor to this divergence is evident for small α. Also, as the noise level increases more data must be devoted to the test set.

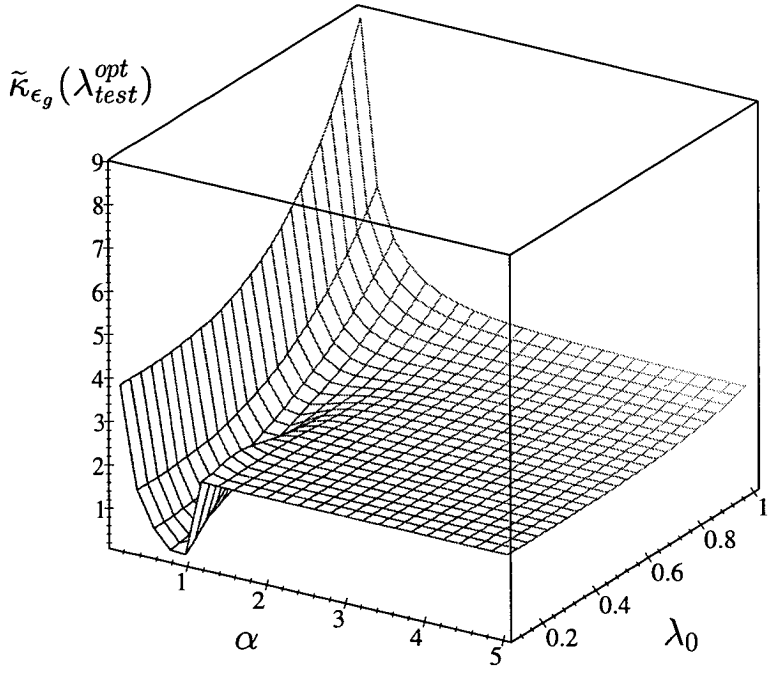


Figure 5.8. Scaled degradation in performance, from the optimal, when using the optimal partition of data set into training and testing subsets, $\widetilde{\kappa}_{\epsilon_g}(\lambda_{test}^{opt})$: Note in this case the scaling is $\mathcal{O}(N^{-1/2})$ (*i.e.* the fractional error is an order $\mathcal{O}(N^{1/2})$ larger than that associated with the evidence procedure shown in figure 4.9).

As can be seen in figure 5.7 as the noise increases so does the size of the optimal test set. In fact, as $\lambda_0 \to \infty$, b^{opt} diverges. As witnessed by figure 5.7 it also diverges as α increases. However, it should be noted that the optimal fraction of test examples to the total number of examples is $\mathcal{O}(1/N^{1/2})$ except where b^{opt} diverges in which case the test set size approaches $\mathcal{O}(1)$. If we compare the optimal partition in this section with the minimal variance partition we find that variance in the resulting weight decay assignment is much larger (*i.e.* order $\mathcal{O}(1/N^{1/2})$) in the optimal partitioning case. Similarly, as can be seen in equation (5.12), the resulting correlation in the present case is much smaller than that obtained by the minimal variance partition, being of the order $\mathcal{O}(1/N^{1/4})$ for the optimal partition. However, the minimal variance partition results in a degradation in performance which is order $\mathcal{O}(1)$. As we now discover this is much larger than the performance loss associated with the optimal partition.

We denote the weight decay assignment resulting from the optimal partition by λ_{tst}^{opt}. As can be seen in equation (5.19), the fractional degradation in performance resulting from this optimal partitioning is $\kappa_{\epsilon_g}(\lambda_{tst}^{opt}) \approx \mathcal{O}(1/\sqrt{N})$ which in the thermodynamic limit is much larger (by a factor of $\mathcal{O}(\sqrt{N})$) than the fractional degradation associated with the evidence procedure (see section 4.5) and, as we will see later, with **CV(1)**. The average fractional degradation, $\kappa_{\epsilon_g}(\lambda_{tst}^{opt})$, scaled by $\mathcal{O}(\sqrt{N})$, is shown in figure 5.8. As the noise increases we find that fractional error also increases. The minimum we find in this surface along $\alpha \approx 1$ is simply due to the different rates at which the degradation in performance and the optimal generalization error decay with the number of examples. Initially, as α increases the average degradation improves quickly but latterly this decay slows relative to that in the optimal error. Figure 5.9 shows this effect more clearly. In the limit of a small data base we find that $\widetilde{\kappa}_{\epsilon_g}(\lambda_{tst})$ diverges like $\mathcal{O}(1/\sqrt{\alpha})$[1]. However, as the size of the data base grows the average fractional degradation tends to a constant;

$$\lim_{\alpha \to \infty} \widetilde{\kappa}_{\epsilon_g}(\lambda_{tst}^{opt}) = \sqrt{2(2+\lambda_0)}. \tag{5.21}$$

Again, this disappointing performance is attributable to the increasing difficulty in identifying the optimal weight decay as the number of training examples grows and in this optimal partitioning scheme we are not able to devote sufficient resources to the test set. In other words, using a large test set to accurately identify the optimal weight decay is wasteful of data, in that it would be better used as training data. The evidence procedure does not suffer from this draw back in that all the data can be devoted to training whilst the method still provides an estimate for the weight decay. Similarly,

[1]Here we continue our convention that a quantity, h, which scales like N^{-s} can also be written, in rescaled form, $\widetilde{h} = N^s h$.

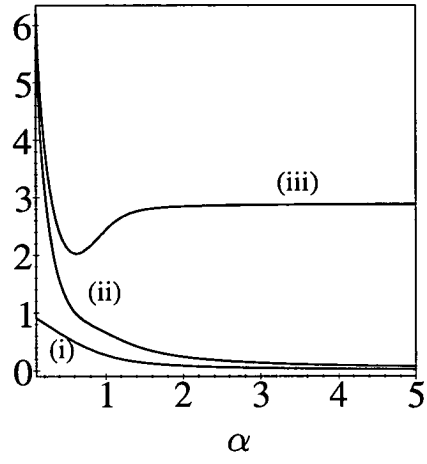


Figure 5.9. A closer look at the scaled degradation in performance: Curve (i) shows the optimal generalization error, curve (ii) the average degradation, $\langle \Delta\epsilon(\lambda_{tst}) \rangle_{P(\mathcal{D})}$, scaled by $N^{1/2}$ and curve (iii) shows, $\widetilde{\kappa}_{\epsilon_g}(\lambda_{tst})$, the ratio of (ii) to (i) The minimum in the fractional degradation is the due to the different rates at which (i) and (ii) decay as α increases.

as we shall see, leave-one-out cross-validation uses test data more efficiently.

Here we briefly comment on the work of Kearns (96) who has recently studied this question of the optimal training-test set split in the context of mappings with binary outputs. Building on the work of Vapnik (82) and Barron and Cover (91) he was able to bound the generalization error obtained when model choice is based on minimization of the error on an independent test set. Although the details of this problem differ from that which we have studied here, there are some broad similarities between our findings and those of Kearns (96). In particular, he noted that the optimal fraction of the data which should be devoted to the test set increases as the total amount of data (α) increases. This observation agrees very well with the result (5.20), which shows the test set size increasing with α. In addition, Kearns (96) showed that as α increases the generalization performance becomes ever more insensitive to the actual choice of training-test set split. Again this agrees well with our findings, with figure 5.10 showing the average degradation in performance, $< \Delta\epsilon(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})}$, for a range of values of the training set size, $m = bN^{1/2}$ and two values of α. The dotted straight lines show the optimal performance for each α (with the lowest being for the largest α). For small α the performance is rather sensitive to the choice of b, but this sensitivity reduces markedly as α increases.

In summary then, for large systems we have seen that the use of a test error to determine the weight decay assignment is considerably less powerful than the evidence procedure in terms of the performance achieved. Indeed, we should remember that figure 5.8 depicts the optimal performance possible, but that it is not clear how to find the optimal partition in practice. Thus, in reality we would expect even worse performance when using a test set to set the level of regularization. In the next section we will examine leave-one-out cross-validation and compare our results to those obtained for the test error in this. We will see that the more sophisticated use of a test set in the cross-validatory approach rectifies some of the short comings we found in the naive approach considered so far.

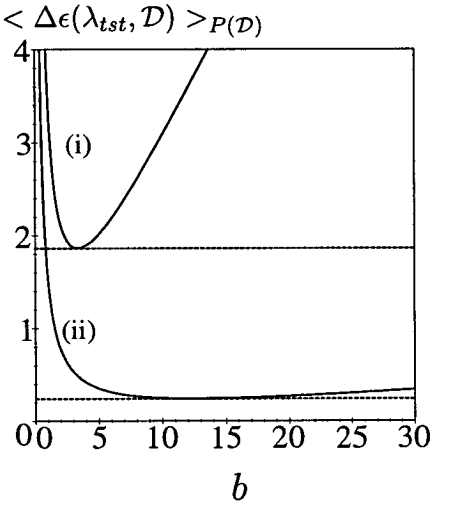


Figure 5.10. Sensitivity to the training-test set split: The average degradation $< \Delta\epsilon(\lambda_{tst}, \mathcal{D}) >_{P(\mathcal{D})}$ versus the normalised test set size b (*i.e.* the test set size $m = bN^{1/2}$). The noise to signal ratio λ_0 is unity, with the upper dotted line showing, for reference, the optimal performance ($b = b^{opt}$) for $\alpha = 1$ and the lower dotted line that for $\alpha = 10$. Curve (i) shows the degradation in performance versus b for $\alpha = 1$ and curve (ii) the corresponding values for $\alpha = 10$.

5.3 Leave-one-out cross-validation

As we have already implied the basic motivation behind cross-validation is to gain some of the benefits of estimating ones expected error using a test set whilst diverting as little data as possible from the training set. This is achieved by taking advantage of the fact that there are a large number of possible ways to divide a data base into l training

examples and m testing examples; $\binom{m+l}{m}$ to be precise. When practising cross-validation one tests and trains a student based on a particular partition of the data set. One then generates a number of such students and corresponding test errors based on different partitions of the data set. The cross-validation error is simply the average of each individual student's test set error over all the students generated. Adopting the notation of the previous section we write

$$\epsilon_{val}(\mathcal{D}) = \frac{1}{\mathcal{S}} \sum_{a=1}^{\mathcal{S}} \left[\frac{1}{m} \sum_{a_i=1}^{m} \left(< y_s(\mathbf{x}^{a_i}) >_{P(y_s|\mathbf{x}^{a_i},\mathcal{D}^{[a]},\mathcal{M})} - y_t(\mathbf{x}^{a_i}) \right)^2 \right]. \tag{5.22}$$

Here the outer sum is over the $\mathcal{S}$ partitions of the data set (*i.e.* over all the students generated) and the inner sum is over the m test points in each partition. That the training set and testing sets are non-overlapping is emphasized by denoting the testing set by a and the training set by $[a]$. [2] Thus, for each partition the student output is the average over the predictive distribution based on the data set $\mathcal{D}^{[a]}$ and it is tested on the m elements, $(y_t(\mathbf{x}^{a_i}), \mathbf{x}^{a_i})$, of $\mathcal{D}^a$. As before, the p elements $(y, \mathbf{x})$ of the data set, $\mathcal{D} = \mathcal{D}^a + \mathcal{D}^{[a]}$, are sampled *i.i.d* from $(P(y_t \mid \mathbf{x})P(\mathbf{x}), P(\mathbf{x}))$.

At this point, to avoid future confusion, we should clarify two terms. The cross-validation error, of equation (5.22), is used to estimate the expected error (generalization error) of various competing *models.* In order to calculate ϵ_{val} a number of *students* must be generated for each model from the different partitions of the data set. In other words students are generated by training the model in question on the different partitions of the data set. Model selection, in the cross-validatory sense, consists of picking the model with the lowest cross-validation error. Thus, in the particular case we consider in the next section the models are simply different values of the weight decay and in order to calculate the leave-one-out cross-validation error we must train p students for each λ value we consider.

Since, the number of data base partitions grows exponentially with the size of the test set, m is often kept low. Indeed, perhaps the most widely used form of cross-validation is 'leave-one-out', **CV(1)**, where only one data point is left out for testing, $m = 1$. However, a number of approaches have been adopted to overcome the computational cost incurred by the use of larger test sets. These include Monte Carlo Cross-Validation, **MCCV** and Balanced Incomplete Cross-Validation, **BICV** (see *e.g.* Shao (93) whose abbreviations we have adopted). In **MCCV** the cross-validation error

[2] Although in principle one could consider a degree of overlap between the two we do not do so here.

(equation 5.22) is estimated by randomly sampling a relatively small number of partitions rather than training students on all possible partitions. **BICV** also aims to reduce the computational effort through reducing the number of partitions considered. This is achieved by dividing the data base into, say $n < p$, blocks of example pairs. Students are then trained on the data with one of the blocks left out. The cross-validation error in the **BICV** scheme is then the test error on this left out set averaged over all the block partitions. In fact, **BICV** might be more appropriately named block cross-validation; the terminology **BICV** apparently derives from the field of experimental design. As we shall see below, there may be a number of good reasons for adopting a cross-validation scheme which leaves out a large proportion of examples for testing. However, leave-one-out cross-validation is widely used because it is straightforward to implement and has a relatively low computational cost in comparison to other cross-validation schemes.

A number of authors have investigated cross-validatory schemes from an analytical point of view. Much of this work has focussed on the asymptotic regime, that is, on the limit of infinite data sets. Indeed, Stone (77a) showed that leave-one-out cross-validation and Akaike's criterion were asymptotically equivalent when maximum likelihood parameter estimation is used within the models under scrutiny and the teacher is realizable by the student. In the same year Stone (77b) also explored the asymptotic consistency of cross-validation and found that for mean square consistent predictors, of which our linear student is an example when the teacher is also linear, **CV(1)** is consistent in that it estimates the generalization error correctly asymptotically. However, as Stone also noted this is a rather weak condition. In our terminology a model which is not mean square consistent would be described as being unable to represent the teacher; the task of learning the teacher is unrealisable. In fact, for nested linear models **CV(1)** was shown, asymptotically, to be able to distinguish between mean square consistent predictors and those which were not. That is between students powerful enough to model the teacher and those which are not. However, in this latter work Stone also demonstrated, for a simple univariate estimation problem, that asymptotically **CV(1)** is unable to identify the best model, from amongst a set of mean square consistent candidates. In other words, leave-one-out cross-validation is inconsistent in terms of *model choice*, except in the crudest sense. Indeed, shortly we will see this form of inconsistency displayed in the **CV(1)** selection of the weight decay parameter in the learnable linear case. As we noted in section 5.2.2, Shao (93) showed, in general, that, in the context of nested linear models, leave-one-out cross-validation was inconsistent, in that it picked the wrong model with finite probability in the limit of large data bases (large p). Furthermore, Shao demonstrated that the consistency of model choice could only be restored if the ratio of test set size, m, to data base size, p, approached unity

as $p \to \infty$. Thus, a considerable price must be paid for consistency in model choice.

Other work has focussed on the estimation of the generalization error itself. For example, Plutowski *et al.*(94) show that leave-out m cross-validation gives an unbiased and consistent estimate of the average generalization error for a student trained on $p - m$ examples, $< \epsilon_g(\mathcal{D}, p - m) >_{P(\mathcal{D})}$. That is, when we leave-out m data points for testing, the cross-validation error is an unbiased estimate of the average generalization error obtained for the student under consideration trained on $p - m$ examples. Thus, as Burman (89) used in support of leave-one-out cross-validation the larger the test set the further any given student is from its optimal performance (based on training on the whole data base). Although, we note here, that *correction* terms were suggested by the author to remedy this defect of hold out m cross-validation. In the statistical mechanical framework Barber (95) has investigated different cross-validatory schemes (*e.g.* **MCCV**, **BICV** and different partitioning schemes), ranking them from the point of view of minimising the variance of the cross-validation error itself. The motivation for this was to estimate the true model error as accurately as possible with the assumption that good model choice would follow suit.

As we have noted before, the principal novelty of the statistical mechanics approach is the ability to focus on the regime where the number of examples is of the order of the number of model parameters. This is in stark contrast to the limit of infinite data sets with which most of the preceding discussion was concerned. Thus, in this section we will explore the leave-one-out cross-validation assignment of the weight decay in a linear scenario and examine its effects on performance. As we have done throughout this thesis, we will rely on the thermodynamic limit, once again this will enable us to study effectively finite data sets, but at the expense of studying a system in which the dimension of the input space diverges. However, in mitigation of this we will focus on the first order finite size corrections to this limit, as we did in the previous section for the test error and in chapter 4 for the evidence. As we have already stated we will be concerned more with model choice, at least in the limited sense implied by the setting of a regularization parameter, and with its effect on performance than on the estimation of the expected error itself. We focus on **CV(1)** because it is the most widely used form of cross-validation and because of its relative simplicity (*e.g.* we will not have to consider optimal partitions of the data base).

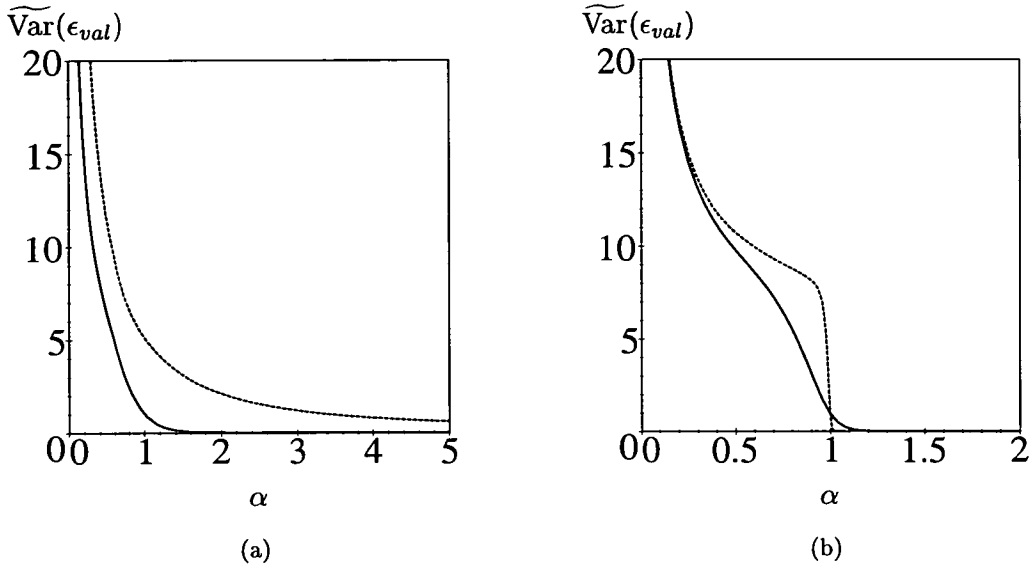


Figure 5.11. Variance of the leave-one-out cross-validation error: Both graphs show the scaled variance, $\widetilde{\mathrm{Var}}(\epsilon_{val})$ with the weight decay set to λ_0. In graph (a) the dotted upper curve is for $\lambda_0 = 1$, whilst the lower solid curve is for $\lambda_0 = 0.1$. In graph (b) the solid curve has $\lambda_0 = 0.01$ whilst the dotted curve is for zero noise.

5.3.1 Finite size effects

In the linear case, for leave-one-out cross-validation, equation (5.22) becomes,

$$\epsilon_{val}(\mathcal{D}, \lambda) = \frac{1}{p} \sum_{a=1}^{p} \left(\frac{1}{\sqrt{N}} (< \mathbf{w} >_{P(\mathbf{w}|\mathcal{D}^{[a]}, \lambda, \beta)} - \mathbf{w}^{o}) \cdot \mathbf{x}^{a} - \eta^{a} \right)^{2}. \tag{5.23}$$

As before, the random variables η^a represent instantiations, in the a^{th} example, of the Gaussian noise, of zero mean and variance σ^2, corrupting the output of the linear teacher parameterised by weights $\mathbf{w}^o$. The posterior, over which the student weights are averaged, is as used in chapter 4 (see section 4.2.1) but now it is conditioned on the reduced data set, $\mathcal{D}^{[a]}$, where the a^{th} example has been deleted.

It is straight forward to show that, since the test points are independent of the training sets, the average **CV(1)** error is,

$$< \epsilon_{val}(\mathcal{D}, \lambda) >_{P(\mathcal{D})} = < \epsilon_g(p-1, \lambda) >_{P(\mathcal{D})} + \sigma^2, \tag{5.24}$$

where, $< \epsilon_g(p-1, \lambda) >_{P(\mathcal{D})}$ denotes the average generalization error based on $p-1$ examples with a weight decay λ. Indeed, this is a special case of the results of Plutowski *et al.*(94). Thus, as with the test error, the average leave-one-out cross-validation error

is optimised by the same weight decay assignment as the average generalization error, namely λ_0. In this sense **CV(1)** can be said to be unbiased. In the remainder of this section, following, what is by now, a familiar route we calculate the variance of the leave-one-out cross-validated error showing it to be an $\mathcal{O}(1/N)$ quantity and then examine the fluctuations in the weight decay assignments for finite system size. In so doing, we find that this unbiasedness is somewhat irrelevant in terms of the performance to be expected on a typical data set for a finite sized system.

Since, we have calculated the average cross-validation error, in order to calculate its variance, we must now calculate the following quantity,

$$< \epsilon_{val}(\mathcal{D}, \lambda)^2 >_{P(\mathcal{D})} = \frac{1}{p^2} \sum_{a,b=1}^{p} < \epsilon^a \epsilon^b >_{P(\mathcal{D})} \tag{5.25}$$

Where we have defined,

$$\epsilon^a = \left(\frac{1}{\sqrt{N}} < \mathbf{w} - \mathbf{w}^o >_{P(\mathbf{w}|\mathcal{D}^{[a]},\lambda,\beta)} \cdot \mathbf{x}^a - \eta^a \right)^2 \tag{5.26}$$

The quantities ϵ^a are simply the test errors of individual students based on the a^{th} partition of the data set; that is students trained on data set $\mathcal{D}^{[a]}$ and tested on example $(\mathbf{x}^a, y_t(\mathbf{x}^a))$. In fact we can write equation (5.25) as the sum of two components, a *variance like* term and a *covariance like* term,

$$< \epsilon_{val}(\mathcal{D}, \lambda)^2 >_{P(\mathcal{D})} = \frac{1}{p^2} \sum_{a=1}^{p} < (\epsilon^a)^2 >_{P(\mathcal{D})} + \frac{1}{p^2} \sum_{a=1}^{p} \sum_{b \neq a}^{p} < \epsilon^a \epsilon^b >_{P(\mathcal{D})} \quad , \tag{5.27}$$

where the last summation in the second term is over $b = 1...p$ such that $b \neq a$ for each value of a. The first term in equation (5.27) is related (in an obvious way) to the variance in the test error of a student trained on $p - 1$ examples and its calculation is very similar to the calculations we performed in the previous section. Indeed, this term is $\mathcal{O}(1/p)$ due to the $1/p^2$ pre-factor and we need only calculate $< (\epsilon^a)^2 >_{P(\mathcal{D})}$ to order $\mathcal{O}(1)$, nonetheless a representative calculation is shown in appendix 5.6. The covariance like term, however, poses more problems as it is rather unlike anything we have had to calculate before, either in this chapter or in chapter 4. The key feature is that now the training data of one student contains the data used to test the other and vice versa. Thus, we can not, as we did in the case of the test error, simply average over the test points.

In fact, because we have split off the variance component, within the covariance term itself we can write the data sets used to train students a and b as $\mathcal{D}^{b+C} \equiv \mathcal{D}^{[a]} =$

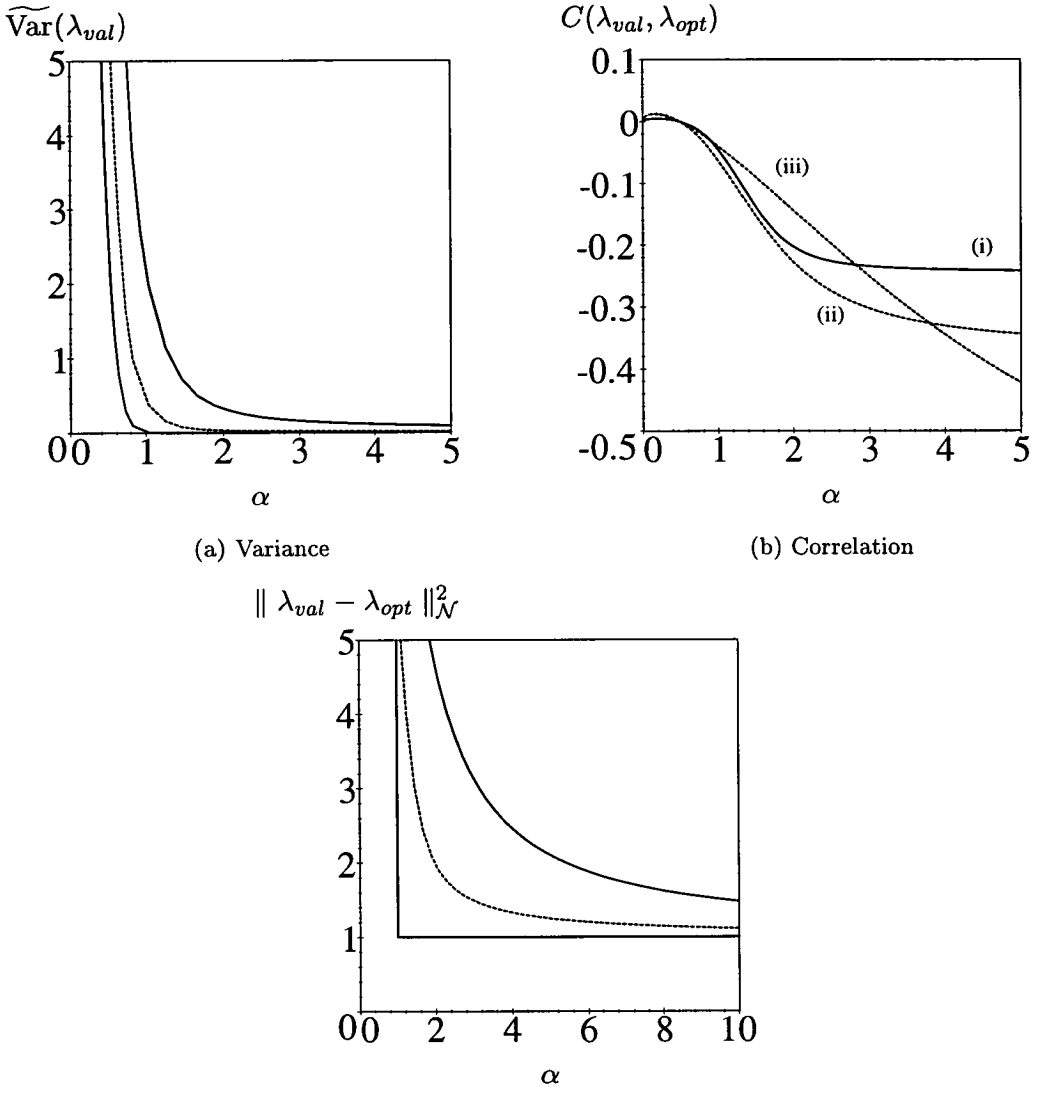
$\mathcal{D}^C + (\mathbf{x}^b, y_t(\mathbf{x}^b))$ and $\mathcal{D}^{a+C} \equiv \mathcal{D}^{[b]} = \mathcal{D}^C + (\mathbf{x}^a, y_t(\mathbf{x}^a))$ respectively. Thus, each training set contains a block common to each student and a block (just one example pair for **CV(1)**) corresponding to the test set of the other student. As explained in appendix 5.6 we can separate and average out this non-overlapping component, leaving us with the average over the common block, $\mathcal{D}^C$ to perform. However, this last average is very similar to the problems we dealt with in chapter 4 and presents no problems. The details are relegated to the appendix 5.6.

As expected we find that the variance is an $\mathcal{O}(1/N)$ quantity, with the self averaging properties of the thermodynamic limit breaking down, with the scaled variance diverging, for $\alpha \to 0$. The scaled variance of the cross-validation error, $\widetilde{\mathrm{Var}}(\epsilon_{val})$, thus calculated, is shown in figure 5.11 for the various noise levels indicated there [3]. The variance is an increasing function of the noise level, whilst in the limit of zero noise we find that $\widetilde{\mathrm{Var}}(\epsilon_{val})$ is zero for $\alpha > 1$ and some finite function of the normalised number of examples for $\alpha < 1$. As we shall see this is indicative of the fact that, for $\alpha > 1$, in the case of no noise on the examples, the cross-validatory weight decay assignment is zero (at least to the order we have calculated) whilst, for $\alpha < 1$ it is greater than zero but well defined. The $\alpha < 1$ behaviour is in contrast to that of the evidence assignment. Returning to the variance of the cross-validation error itself, in the limit of a large number of examples, we find that,

$$\lim_{\alpha \to \infty} \mathrm{Var}(\epsilon_{val}(\mathcal{D})) = \frac{2\sigma^4}{\alpha N} + \mathcal{O}\left(\frac{1}{N\alpha^2}\right). \tag{5.28}$$

Thus, since ϵ_{val} is unbiased, asymptotically the leave-one-out cross-validation error accurately estimates the generalization error, based on $p - 1$ examples, of our linear model. This result is in agreement with the consistency results of Plutowski *et al.*(94).

[3] The scaled variance $\widetilde{\mathrm{Var}}(\epsilon_{val}) = N\mathrm{Var}(\epsilon_{val})$.



(a) Variance

(b) Correlation

(c) Distance from optimal

Figure 5.12. Cross-validation weight decay assignments: Graph (a) shows the scaled variance of the leave-one-out cross-validation weight decay assignment, $\widetilde{\mathrm{Var}}(\lambda_{val})$. The upper curve is for $\lambda_0 = 1$, the dotted curve for $\lambda_0 = 0.1$ and the lower curve for $\lambda_0 = 0.01$. Graph (b) shows the normalised correlation between the cross-validatory assignment of the weight decay and the optimal for (i) $\lambda_0 = 0.1$, (ii) $\lambda_0 = 0.25$ and (iii) $\lambda_0 = 4$. Graph (c) shows the normalised square distance of λ_{val} from the optimal. The lowest, right angled, curve is for zero noise, the middle (dotted) curve for $\lambda_0 = 0.25$ whilst the upper most curve corresponds to a noise level of $\lambda_0 = 1$.

Parameter assignment

We now turn our attention to weight decay assignments from the cross-validation error. Since the cross-validation error is intended to be a good estimate of the generalization error itself we should pick models with the lowest cross-validation error. In the context under discussion here this simply means finding the weight decay with the lowest **CV(1)** error. In practice this will require considerable computational effort as one must train and test p students for each value of the weight decay considered. However, analytically we can determine the optimal leave-one-out cross-validatory weight assignment by minimising the **CV(1)** error of equation (5.23) *w.r.t.* the weight decay parameter λ. As before, because the cross-validation error is self averaging, in the thermodynamic limit the assignments from particular data sets correspond to the assignment from the average cross-validation error and thus, as we saw above, to the optimal. However, for a finite sized system we can explore the cross-validatory assignments by expanding around the average case, as we did earlier for the evidence and the test error assignments. Thus, we write $\lambda_{val} = \lambda_0 + \Delta\lambda_{val}$ where $\Delta\lambda_{val}$ vanishes in the thermodynamic limit.

In doing this we find that the variance in the cross-validatory assignments is given by,

$$\mathrm{Var}(\lambda_{val}(\mathcal{D})) = \frac{< (\partial_\lambda \epsilon_{val})^2 >_{P(\mathcal{D})}}{(\partial_\lambda^2 \epsilon_{val})_0^2} \Bigg|_{\lambda=\lambda_0} + \mathcal{O}\left(\frac{1}{N}\right), \tag{5.29}$$

Where, as before, the subscript zero indicates that the function in brackets is to be evaluated in the thermodynamic limit. The average required here can be evaluated from $< \epsilon_{val}(\lambda_1)\epsilon_{val}(\lambda_2) >_{P(\mathcal{D})}$, the calculation of which is outlined in appendix 5.6, by taking consecutive derivatives *w.r.t.* λ_1 and λ_2 and subsequently setting $\lambda_1 = \lambda_2 = 0$. In a similar manner we may obtain the covariance of the cross-validatory assignment with that of the optimal weight decay from $< \epsilon_{val}(\lambda_1)\epsilon_g(\lambda_2) >_{P(\mathcal{D})}$. The calculation of this latter quantity is similar to that shown in appendix 5.6.

Graph (a) of figure 5.12 shows the scaled variance in the cross-validatory assignment, $\widetilde{\mathrm{Var}}(\lambda_{val})$. We see that the variance increases with the noise level, λ_0. In fact in the limit of zero noise we find that the variance vanishes for $\alpha > 1$ whilst for a smaller number of examples it is an order $\mathcal{O}(1/N)$ function of α. This is in stark contrast to the evidence assignment whose scaled variance diverges (*i.e.* $\mathrm{Var}(\lambda_{ev}) = \mathcal{O}(1)$) for $\alpha < 1$ in this limit. Thus, we find that in the case of no noise where the evidence weight decay assignments are poorly defined the cross-validatory assignments are well defined. However, since the optimal weight decay in this region is zero we note that the

cross-validatory assignments are not optimal since the fluctuations in **CV(1)** around $\lambda_0 = 0$ are non zero for $\alpha < 1$. We also note that for $\alpha > 1$ both the evidence and cross-validatory assignments are optimal (*i.e.* zero) in the noiseless limit.

In the limit of large data sets we find that

$$\lim_{\alpha\to\infty} \mathrm{Var}(\lambda_{val}) = \frac{2\lambda_0^2(2\lambda_0+3)}{N\alpha} + \mathcal{O}\left(\frac{1}{N\alpha^2}\right). \tag{5.30}$$

Thus, the cross-validatory weight decay estimate is asymptotically equivalent to λ_0 to order $\mathcal{O}(1/\sqrt{N})$. However, this is at odds with the optimal assignment whose variance diverges in the large α limit (see section 4.4). Thus, the model selected by **CV(1)** must be sub-optimal asymptotically as, indeed, Stone (74b) showed for a univariate estimation problem. Nonetheless, given that the cross-validated error is a sum of test errors, the result (5.30) is somewhat surprising given the divergence in the variance of the test set assignment of the weight decay obtained earlier (see equation 5.8). The implication is that the correlations between test and training points reduce the variance in the cross-validation weight decay assignments. In a moment we will discuss some simulation results in support of this idea, however before that we discuss some other analytical results.

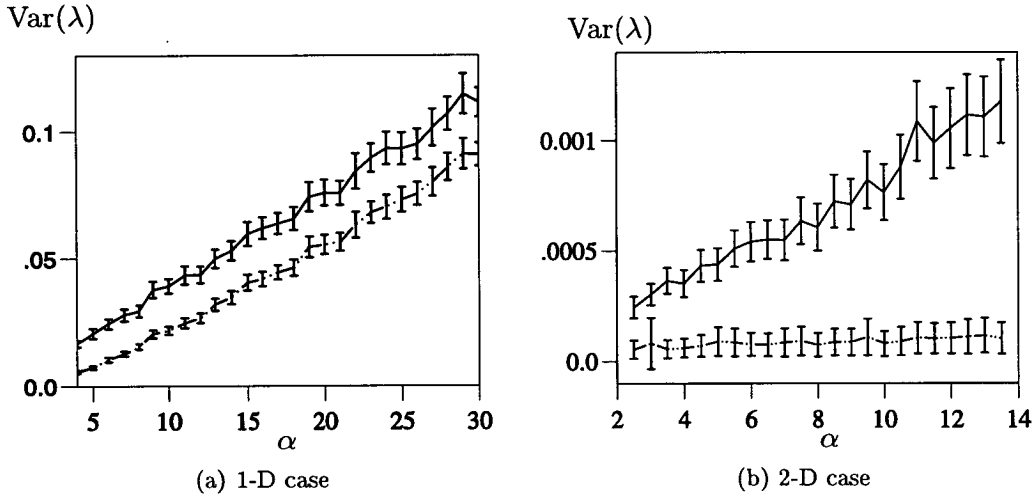
That the cross-validatory weight decay assignment is sub-optimal is confirmed by graphs (b) and (c) of figure 5.12 which show the cross-validatory assignment's normalised correlation with the optimal weight decay and its average normalised squared distance from the optimal. Considering the correlation we see that the smaller the noise level the larger (less negative) the asymptotic correlation. Indeed, we find,

$$\lim_{\alpha\to\infty} C(\lambda_{val}, \lambda_{opt}) = -\frac{\sqrt{2\lambda_0}}{\sqrt{2\lambda_0+3}} \tag{5.31}$$

Thus, asymptotically, the **CV(1)** estimates are more strongly correlated with the optimal than are those of the evidence (see section 4.4) except in two cases where they are equally correlated; for large noise or in the limit of zero noise. Indeed, for large noise we find, asymptotically, that both evidence and cross-validation assignments are fully anti-correlated with the optimal whilst, in the zero noise limit we find that $C(\lambda_{val}, \lambda_{opt}) \to 0$ for all α as for the evidence assignments. This latter phenomenon can be understood from the fact that the optimal weight decay does not fluctuate over data sets in the noiseless limit.

The normalised distance measure, $\| \lambda_{val} - \lambda_{opt} \|^2_{\mathcal{N}}$, defined in analogy to equation (4.18) but for the **CV(1)** weight decay assignment, has a more straightforward behaviour. It is a monotonically increasing function of the noise to signal ratio, λ_0, and

asymptotically, for large α, it tends to 1. In the limit of zero noise it is the same as $\| \lambda_{ev} - \lambda_{opt} \|^2_{\mathcal{N}}$ (see section 4.4), diverging for $\alpha < 1$ and unity otherwise. Examination of the un-normalised distance, $\| \lambda_{val} - \lambda_{opt} \|^2$, reveals that the cross-validatory, **CV(1)**, assignment of the weight decay is in fact, optimal in the noiseless case for $\alpha > 1$ but otherwise never coincides with the optimal. In section 4.4 we found the same to be true of the evidence assignment. However, in the noiseless regime for $\alpha < 1$ the **CV(1)** weight decay assignment is well defined in contrast to that of the evidence, that is $\| \lambda_{val} - \lambda_{opt} \|^2$ is finite in contrast to $\| \lambda_{ev} - \lambda_{opt} \|^2$.



(a) 1-D case (b) 2-D case

Figure 5.13. Simulations: Variance of the cross-validatory, **CV(1)**, weight decay assignment and the optimal assignment. Simulations were performed for a linear student and teacher the results averaged over 1000 data sets in each case. As indicated graph (a) shows the results for a system size $N = 1$ whilst in graph (b) $N = 2$. The noise to signal ratio in both cases is $\lambda_0 = 0.01$ and the variance of λ_{val} is the lower (dot - dash) curve whilst, $\mathrm{Var}(\lambda_{opt})$ is the solid curve. The standard error bars show that in both cases for a relatively large number of examples the variance of the cross-validatory assignment is lower than that of the optimal.

Simulations

At this point we should draw attention to some discrepancies between our finite size corrections to the thermodynamic limit and our simulation results for low dimensional systems. Our simulation results suggest that the variance of the cross-validatory weight decay assignment diverges linearly with p (see figure 5.13) whereas our finite size corrections reveal that asymptotically this variance decays to zero (equation 5.30). Thus,

these simulations demonstrate important differences in behaviour between our finite size corrections to the thermodynamic limit and small systems. Nonetheless, as we shall see, in terms of performance our simulations and finite size corrections are broadly in agreement (see figure 4.12).

The fact that the variance in the **CV(1)** assignment for a small system diverges seems reasonable when we recall that the variance of the test error assignments of the weight decay also diverge (even close to the thermodynamic limit, see section 5.2.1). In fact, the test error can be thought of as a noisy estimate of the generalization error and as we saw earlier the variance in the weight decay assigned from it is, indeed, larger than the variance in the optimal assignment (see equation 5.6). Since the cross-validation error is an average of the test errors from a number of students then we might think that the same should hold for the variance in the cross-validatory assignments. However, our simulations reveal that this is, in fact not true; the variance in the optimal assignment is larger than that in the **CV(1)** assignment (see figure 5.13). Furthermore, comparison of figures 5.13(a) and (b) shows that the divergence in the variance of the cross-validatory assignment is less marked in the larger system ($N = 2$) as compared with the $N = 1$ case. We expect that as the system size increases this trend will continue with simulation results approaching our finite size corrections. However, simulations become prohibitively expensive for larger system sizes.

5.3.2 Cross-validatory performance

Having established, in the first order finite size corrections, that the cross-validatory weight decay assignments are similar in character, asymptotically, to those of the evidence, we now explore their effect on performance. Once again we will focus on the relative degradation in performance but firstly let us consider the degradation in performance itself,

$$\Delta\epsilon(\mathcal{D}, \lambda_{val}) = \Delta\lambda_{val}\partial_\lambda\epsilon_g + \frac{1}{2}\Delta\lambda_{val}^2\partial_\lambda^2\epsilon_g + \frac{1}{2}\Delta\lambda_{opt}^2\partial_\lambda^2\epsilon_g + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{5.32}$$

The degradation in performance associated with **CV(1)** is thus the same order as that of the evidence assignments, that is an order $\mathcal{O}(1/N)$ quantity on average but with a variance of order $\mathcal{O}(1/N^2)$. However, as explained in section 5.2.1 we will not consider these fluctuations here, but concentrate our efforts on the average degradation. The most important point to notice concerning equation (5.32) is that it is an order $\mathcal{O}(\sqrt{N})$ improvement over the performance degradation achieved by the optimal partition of the test set found in section 5.2.2. Thus, since the cross-validatory approach involves the training and testing of p students, as compared with one in the case of the test

error, it improves the performance by a factor of $\sqrt{N}$, as compared with the naive use of a test set, but at the cost of increasing the computational effort by a factor of $p \approx \mathcal{O}(N)$.

The average degradation, $< \Delta\epsilon(\lambda_{val}) >_{P(\mathcal{D})}$, can be expressed in terms of the average separation between the cross-validatory assignment and the optimal weight decay, $\| \lambda_{val} - \lambda_{opt} \|^2$, defined in analogy to equation (4.16). In the large α limit we find that, as with the degradation associated with the evidence procedure, $< \Delta\epsilon(\lambda_{val}) >_{P(\mathcal{D})}$ is order $\mathcal{O}(1/\alpha N)$. This then, as we have noted, is an example of Stone's (77b) result that leave-one-out cross-validation produces asymptotically optimal performance for mean square consistent predictors. Similarly, in the zero noise limit we find that, when the number of examples exceeds the number of parameters in our model, the average degradation, $< \Delta\epsilon(\lambda_{val}) >_{P(\mathcal{D})}$, is the same as that associated with the evidence, namely zero. However, for $\alpha < 1$ the performance of the cross-validatory choice of regularizer is considerably better than that of the evidence since, $< \Delta\epsilon(\lambda_{val}) >_{P(\mathcal{D})}$ is order $\mathcal{O}(1/N)$ in this region whilst $< \Delta\epsilon(\lambda_{ev}) >_{P(\mathcal{D})}$ is order $\mathcal{O}(1)$ (see section 4.5).

We now turn to the average relative degradation in performance, $< \kappa_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})}$ (see equation 5.14). This scaled by N, $< \widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})}$, is plotted in figure 5.14. The divergence in the variance of the cross-validatory weight decay assignment as $\alpha \to 0$ (see *e.g.* figure 5.12(a)) is reflected in the divergence of $< \widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})}$ for small numbers of examples. Indeed, by now we are familiar with this breakdown of the thermodynamic limit when the number of training examples is not extensive. In contrast, in the limit of large data bases we find that,

$$\lim_{\alpha \to \infty} < \widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})} = \frac{1}{N} + \frac{2}{N\alpha}(\lambda_0 + 1) + \mathcal{O}\left(\frac{1}{N\alpha^2}\right). \tag{5.33}$$

So, asymptotically the cross-validatory assignments are not optimal and indeed, result in the same performance as do the evidence assignments (see equation 4.25). Thus, we find that model choice based on the leave-one-out cross-validatory error is asymptotically inconsistent, as Stone (77b) showed in the case of univariate estimation. In fact, as we have mentioned Shao (93) demonstrated that, for nested linear models, this inconsistency could only be remedied by holding out larger and larger test sets as the data base increased in size. It would be interesting to explore this issue in future work within the current framework.

The asymptotic result, equation (5.33), reveals that the relative degradation increases with the noise to signal ratio. In the limit of zero noise we find that $< \widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})}$ diverges at $\alpha = 1$ due to the rate at which the optimal generalization error approaches zero. This divergence can be seen in figure 5.15 which shows

the average relative degradation, $< \widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})}$, for a range of noise levels, λ_0, approaching zero. We see that, in the zero noise limit for $\alpha < 1$, $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val})$ does not diverge, in stark contrast to the evidence case. However, for $\alpha > 1$ we find that both the evidence and the cross-validatory assignments achieve the same performance (see equation 4.26) since,

$$\lim_{\lambda_0 \to 0} < \widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) >_{P(\mathcal{D})} = \frac{\alpha + 1}{\alpha - 1}. \tag{5.34}$$

Naturally, this performance is a reflection of the cross-validatory weight decay assignments themselves. In the zero noise limit, for $\alpha > 1$, we find that λ_{val} coincides with the optimal assignment, $\lambda_{opt} = 0$, at least to the order we have calculated. Indeed, we expect this to be generally true since in this regime one has access to all the information required to solve the problem fully. Similarly, for $\alpha < 1$ and zero noise, the cross-validatory assignments are also well defined, but sub-optimal. This contrasts strongly with the evidence weight decay assignment which is ill defined in this regime.

5.4 Comparison of model selection by cross-validation and evidence

Having explored the performance of weight decay assignment by leave-one-out cross-validation we now seek to directly compare it to the performance of the evidence procedure. Recall that, even with optimal partitioning, the naive application of a test set to model selection, as explored in section 5.2.2, is an order $\mathcal{O}(\sqrt{N})$ worse, in terms of the performance degradation suffered, than either **CV(1)** or the evidence procedure. Thus, it is somewhat meaningless to compare $\widetilde{\kappa}_{\epsilon_g}(\lambda_{tst})$ directly with $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val})$ or $\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev})$. In the previous section we were able to compare the evidence and **CV(1)** , algebraically, in certain limits, namely zero noise and asymptotically. In this section we look at the more realistic regimes of non-zero noise and finite data bases.

Figure 5.16 shows the average relative degradation associated with the evidence procedure and with leave-one-out cross-validation for three different noise levels at finite α. The top most graphs, (a) and (b), show that for $\lambda_0 = 0.1$ the performance of **CV(1)** is appreciably better than that of the evidence procedure for small sample sizes. However, for larger samples, say $\alpha > 2$, they are barely distinguishable, a reflection of the fact that asymptotically to order $\mathcal{O}(1/N\alpha)$ they are the same (see equations 4.25

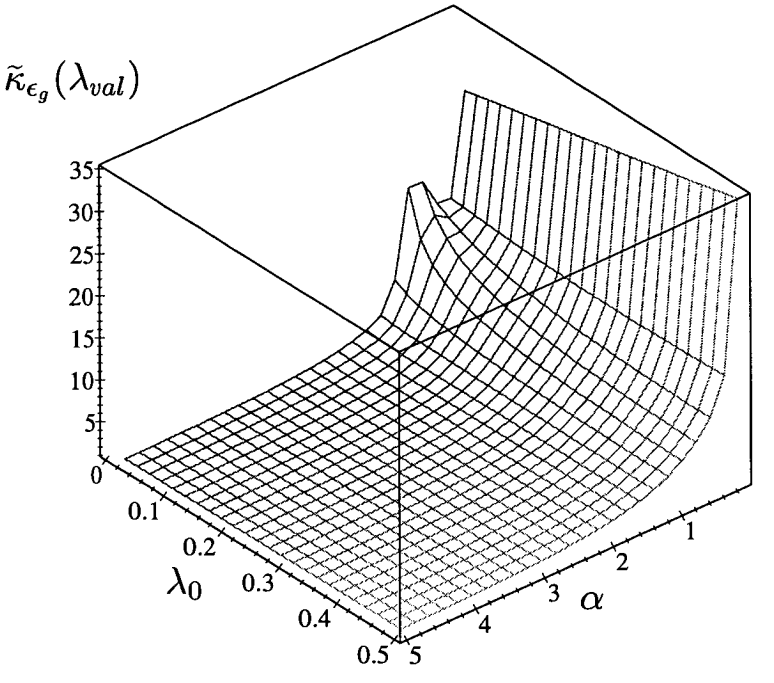
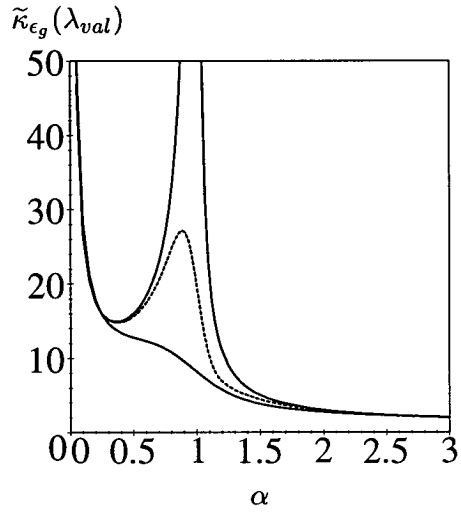
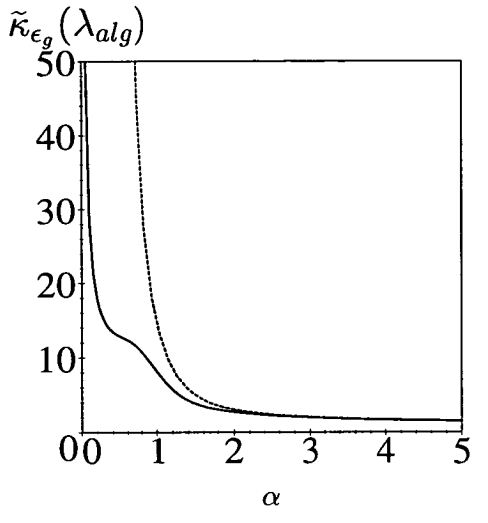


Figure 5.14. Scaled fractional degradation in performance, from the optimal, when using the leave-one-out cross-validatory assignment of the weight decay, $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val})$

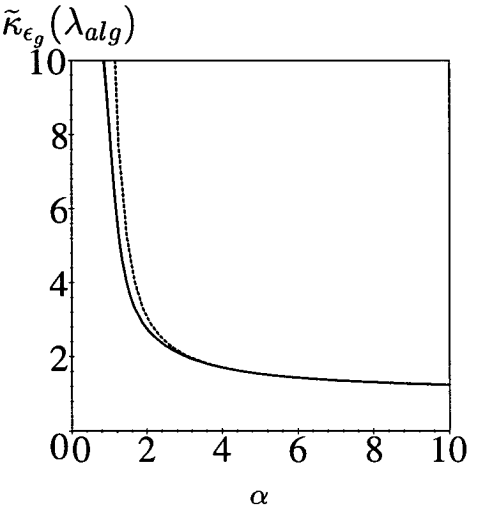


(a)

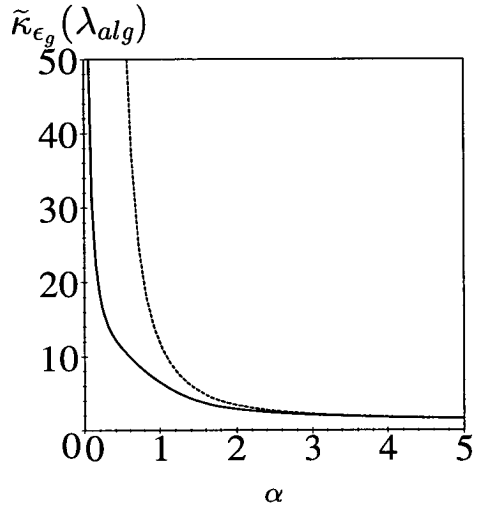
Figure 5.15. Relative degradation in performance when using **CV(1)** for low noise levels. The lowest curve is for $\lambda_0 = 0.1$, in the middle curve $\lambda_0 = 0.01$ and the top curve is the zero noise limit. In this limit $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val})$ diverges at $\alpha = 1$, whilst for $\alpha > 1$ we find that $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val}) = (\alpha + 1)/(\alpha - 1)$ and is finite for $0 < \alpha < 1$.



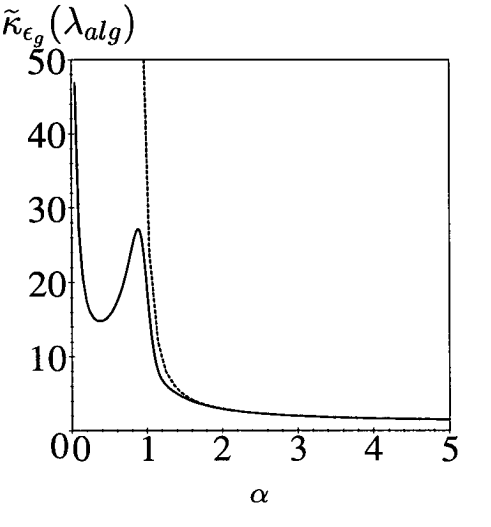
(a) $\lambda_0 = 0.1$



(b) $\lambda_0 = 0.1$: Close up



(c) $\lambda_0 = 0.25$



(d) $\lambda_0 = 0.01$

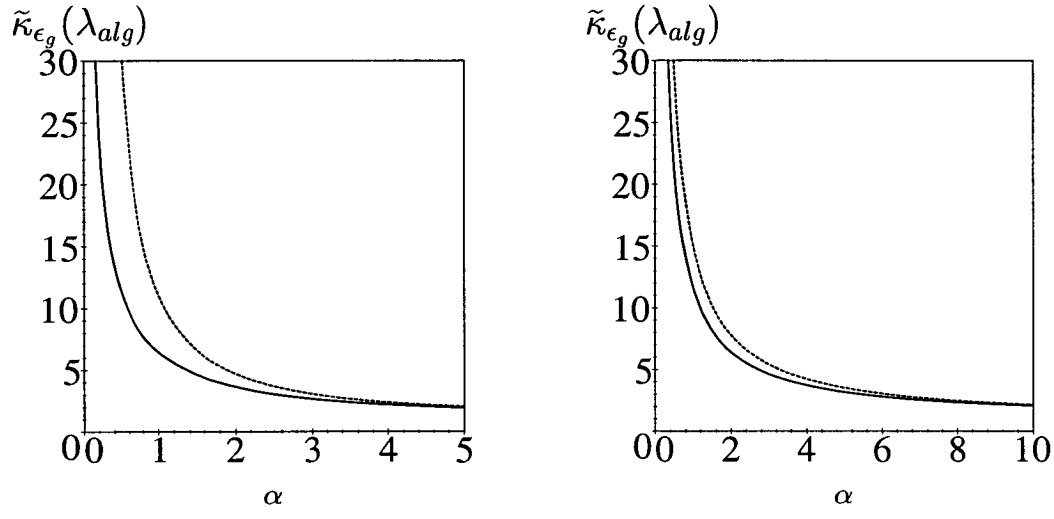
Figure 5.16. Comparison of the fractional degradation in performance of the evidence procedure with that associated with cross-validation. In both graphs the performance of the evidence $\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev})$ is shown in the upper dashed curve whilst that of cross-validation, $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val})$ is the solid lower curve. Both graphs (a) and (b) are for a noise level of $\lambda_0 = 0.1$ with graph (b) showing a close up of graph (a). In the former we see that leave-one-out cross-validation is superior to the evidence for small to moderate sample sizes. However, the latter shows that for larger data sets the two methods are barely distinguishable; a reflection of the fact that they are asymptotically equivalent to order $\mathcal{O}(1/\alpha)$ and indeed, in the limit, both tend to 1. Graph (c) shows broadly the same behaviour for a slightly larger noise level whilst graph (d) shows the case of relatively small noise to signal ratio. In the latter case we see precursors to the zero noise limit behaviour discussed in the text.

and 5.33). Figure 5.16(c) reveals that the situation is broadly similar for slightly a larger noise level whilst graph (d) reveals some of the small noise behaviour discussed in the preceding section. In particular, we see that for this rather small noise level, $\lambda_0 = 0.01$, the average fractional degradation associated with the evidence procedure is very large for $\alpha < 1$ whilst that of **CV(1)** is large close to $\alpha = 1$ but is much smaller than $\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev})$ in the $\alpha < 1$ regime. For a larger number of examples we find that the performance of both algorithms is very similar. Recall that in the zero noise limit the performance of both methods is identical for $\alpha > 1$ (see equations 4.26 and 5.34), whilst only $\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev})$ diverges for $\alpha < 1$. Figure 5.17 shows that as the noise level increases, whilst cross-validation is still superior, the difference in performance between the two diminishes; note that the scale on these graphs is smaller than that in figure 5.16. The broad conclusion that **CV(1)** generally achieves better performance than the evidence procedure is also supported by our simulation results for low dimensional systems (see figure 4.12).

Finally, figure 5.18 compares the normalised distance $\| \lambda_{alg} - \lambda_{opt} \|^2_{\mathcal{N}}$, between the optimal weight decay and the assignments from leave-one-out cross-validation and the evidence procedure. Both graphs show, for reference, the case of zero noise where the two methods are indistinguishable in terms of this measure [4]. Broadly speaking, for non-zero noise the differences in performance are reflected in this average distance from the optimal weight decay, with the evidence assignments being further away.

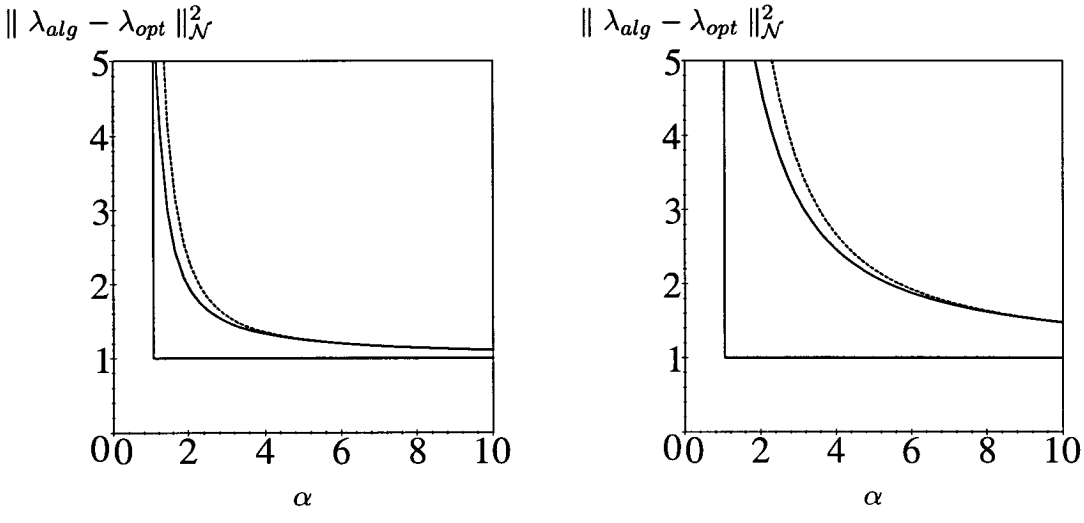
We briefly comment on our results in relation to those of Wahba (85) who investigated a similar problem to that with which we have busied ourselves here. In particular, she studied the problem of setting the regularization parameter in smoothing splines on the basis of a cross-validatory method and a maximum likelihood method analogous to the evidence. The details of the problem are somewhat different, for instance the student and teacher functions being defined on a finite region of input space, making absolute comparisons difficult. Nonetheless, qualitatively our results are in agreement with her findings. For function classes including our linear problem Wahba found that, asymptotically, the performance associated with the two algorithms was indistinguishable, as we find. In addition, Monte Carlo simulations revealed that, for moderate sample sizes, the cross-validatory approach was superior for small noise level whilst the difference was less significant for larger noise levels. Although, this later result is the less relevant of the two as it was obtained for a different class of functions from which the linear case is excluded. Secondly, Kearns *et al.*(95) compared model selection algorithms such as the evidence and cross-validation from a fairly general point of view

[4]Note, however, that they *are* distinguishable in terms of performance.



(a) $\lambda_0 = 1$ (b) $\lambda_0 = 4$

Figure 5.17. Comparison of the scaled fractional degradation in performance of the evidence procedure with that of cross-validation. In both graphs the performance of the evidence $\widetilde{\kappa}_{\epsilon_g}(\lambda_{ev})$ is shown in the upper dashed curve whilst that of cross-validation, $\widetilde{\kappa}_{\epsilon_g}(\lambda_{val})$ is the solid lower curve. In both graphs the noise level is rather high with $\lambda_0 = 1$ in (a) and $\lambda_0 = 4$ in (b). Comparing to figure 5.16 we see that for large noise level there is less to distinguish between the evidence and **CV(1)** in terms of performance.



(a) $\lambda_0 = 1$ (b) $\lambda_0 = 4$

Figure 5.18. Normalised distance between optimal weight decay and assignments from **CV(1)** and the evidence: In both graphs the right angled curve, zero noise, case is shown for reference. The upper curve (dotted) is $\| \lambda_{ev} - \lambda_{opt} \|^2_{\mathcal{N}}$ whilst the middle curve is $\| \lambda_{val} - \lambda_{opt} \|^2_{\mathcal{N}}$ for the noise levels indicated in the captions. All curves tend toward unity as α increases. Likewise the evidence and the cross-validatory assignments are indistinguishable in the zero noise limit. In general, however, the evidence assignments are further from the optimal than are those of **CV(1)**. In fact, in the zero noise limit for $\alpha < 1$, this normalised distance measure is misleading in that only the un-normalised scaled distance (see equation 4.16) between the evidence assignments diverges. Thus, even though the cross-validatory assignments are sub-optimal they are much closer, to the optimal, than are those of the evidence.

using the **PAC** approach. As is natural within that framework, they considered students and teacher mappings in discrete space in contrast to the linear case with real inputs and outputs studied here. These authors concluded that cross-validatory choice of model was to be generally preferred in situations where the generalization error obeys a power law decay (*i.e.* where phase transitions in generalization ability are absent). The learnable linear case is an example of such a system and thus, our results support this suggestion.

In summary then, we have seen that, in the learnable linear context, leave-one-out cross-validation, generally, achieves superior performance as compared to the evidence procedure. However, asymptotically the two procedures are equivalent in performance, at least to the first order finite size corrections we have considered. Similarly, for zero noise and $\alpha > 1$ they are indistinguishable. However, let us consider the computational effort required by each method. Since we choose to make predictions based on the average over the predictive distribution we must, in fact, calculate the evidence as a minimum requirement. Thus, we see that model selection based on **CV(1)** is considerably more computationally expensive than model selection from the evidence since in the former we must train and test p students for each of the models (*e.g.* values of the weight decay) to be compared. Thus, we must leave it to the practitioner to decide how much she is willing to pay for the improved performance of the **CV(1)** weight decay estimates. Nonetheless, we can suggest that in situations where large amounts of data are available or where the data is noiseless (if $p > N$) there is little to be gained in using the cross-validatory approach over the evidence procedure.

5.5 Summary

In this chapter we have examined two methods of model selection based on different estimates of the generalization error. In both cases we have focused on the problem of setting the weight decay parameter in the learnable linear case. As in chapter 4 our principal tools were statistical mechanical using which we have investigated the first order finite size corrections to the thermodynamic limit.

In the first model selection method investigated, the estimate of the expected error was the error on an independent test set. We, thus, had to decide how best to partition the data into training and testing sets. We examined two criteria on which to base such a partition, and found that in the first, the minimal variance partition, the fraction of data required for testing approached unity as the size of the data base increased. This is a reflection of the fact that the optimal model, from amongst a set of mean square consistent candidates, becomes harder to identify as the amount of training data

increases. The second criterion, namely minimisation of the resulting generalization error, revealed that when the data base contains an extensive number of examples, $p = \alpha N$, the optimal scaling of the test set size is $m \approx \mathcal{O}(N^{1/2})$ for large system size, N. Furthermore, the optimal degradation in performance associated with such a partition is an order $\mathcal{O}(\sqrt{N})$ larger than that associated with the evidence procedure calculated in chapter 4.

The second model selection method examined in this chapter was leave-one-out cross-validation, **CV(1)**. We investigated the parameter assignments of this method and their effects on performance. We found that the cross-validatory approach made better use of the test set, resulting in a degradation in performance of the same order as the evidence and thus, an order $\mathcal{O}(\sqrt{N})$ improvement over the naive use of a test set. In fact, the performance of **CV(1)** was also seen to be superior to that of the evidence procedure excepting the zero noise case when the number of examples exceeds the number of parameters and the asymptotic ($\alpha \to \infty$) regime. However, we noted that the enhanced performance of leave-one-out cross-validation is achieved at increased computational effort as compared with the evidence procedure.

5.6 Appendix: calculating (co)-variances of the cross-validation error.

In this appendix we outline the calculations required to obtain the variance of the cross-validation error itself (see equation 5.23) and the (co)-variance of weight decay assignments made from it. To this end we show how we calculate a representative term from these quantities but do not show the full calculations which are somewhat lengthy. As mentioned in section 5.3.1 we can calculate the variance of the weight decay assignments from the cross-validation error by differentiating $< \epsilon_{val}(\lambda_1)\epsilon_{val}(\lambda_2) >_{P(\mathcal{D})}$ with respect to λ_1 and λ_2. Since we also need this quantity (when $\lambda_1 = \lambda_2$) for the variance of the error itself, for the sake of generality, we will focus on its calculation here. Furthermore, we note that the co-variance of the cross-validation assignment, λ_{val}, with optimal weight decay can be found from $< \epsilon_{val}(\lambda_1)\epsilon_g(\lambda_2) >_{P(\mathcal{D})}$ calculation of which follows along similar lines to those presented here.

As in equation (5.27) we can write $< \epsilon_{val}(\lambda_1)\epsilon_{val}(\lambda_2) >_{P(\mathcal{D})}$ as the sum of variance and covariance like terms,

$$\frac{1}{p^2}\sum_{a=1}^{p} < \epsilon^a(\lambda_1)\epsilon^a(\lambda_2) >_{P(\mathcal{D})} + \frac{1}{p^2}\sum_{a\neq b} < \epsilon^a(\lambda_1)\epsilon^b(\lambda_2) >_{P(\mathcal{D})}, \tag{5.35}$$

Variance terms

Initially, let us deal with the first of these; the variance like term which is, in any case, the simpler of the two. Introducing the vector $\mathcal{R} = \mathbf{w} - \mathbf{w}^o$, and expanding this variance like term (see equation 5.26 & 5.27) an example term is found to be

$$\frac{4}{p^2 N} \sum_a^p < \langle \mathcal{R}(\mathcal{D}^{[a]}, \lambda_1) \rangle_{\mathbf{w}} \cdot \mathbf{x}^a \eta^a \eta^a \mathbf{x}^a \cdot \langle \mathcal{R}(\mathcal{D}^{[a]}, \lambda_2) \rangle_{\mathbf{w}} >_{P(\mathcal{D})} . \tag{5.36}$$

Here we are using the notation $\langle \rangle_{\mathbf{w}}$ to denote the average over the posterior distribution. Since, the training data (on which the $\mathcal{R}$ vectors depend) is independent of the test points one can simply average over these separately. On averaging out the test data we find that

$$\frac{4\sigma^2 \sigma_x^2}{p^2 N} \sum_a^p < \langle \mathcal{R}(\mathcal{D}^{[a]}, \lambda_1) \rangle_{\mathbf{w}} \cdot \langle \mathcal{R}(\mathcal{D}^{[a]}, \lambda_2) \rangle_{\mathbf{w}} >_{P(\mathcal{D}^{[a]})} . \tag{5.37}$$

The average vectors $\langle \mathcal{R} \rangle_{\mathbf{w}}$, which are the difference between the student and teacher weight vectors averaged over the posterior, can be expressed in terms of the response function matrix as described in Appendix 4.8.1 (see equation 4.A10). Substituting for these and averaging over the training data presents us with no problems which we have not already dealt with, in this chapter or earlier, and in so doing we obtain,

$$\frac{4\sigma^2}{p} \left[\sigma^2 (\widetilde{\mathrm{tr}}\, (\mathbf{g}_1 \mathsf{C} \mathbf{g}_2)) + \lambda_1 \lambda_2 \sigma_w^2 \sigma_x^2 \widetilde{\mathrm{tr}}\, (\mathbf{g}_1 \mathbf{g}_2) \right] \tag{5.38}$$

The matrix C is the normalised correlation matrix of the inputs in the common block of examples (*i.e.* those not used for testing). In addition, $\mathbf{g}_1$ and $\mathbf{g}_2$ are the response function matrices, based on the correlation matrix for this common block (see appendix 4.8.1), where we have introduced the subscript to denote the dependence on λ_1 or λ_2 (*e.g.* $\mathbf{g}_1 = \mathbf{g}(\lambda_1) = (\mathsf{C} + \lambda_1 \mathsf{I})^{-1}$) and $\widetilde{\mathrm{tr}} = \mathrm{tr}\, /N$ is the scaled trace. We discuss the evaluation of $\widetilde{\mathrm{tr}}\, (\mathbf{g}_1 \mathbf{g}_2)$ and $\widetilde{\mathrm{tr}}\, (\mathbf{g}_1 \mathsf{C} \mathbf{g}_2)$ at the end of this appendix.

Covariance terms

Now let us expand the covariance term and calculate the component corresponding to equation (5.36), namely,

$$\frac{4}{p^2 N}\sum_{a\neq b}^{p} < \langle \mathcal{R}(\mathcal{D}^{[a]},\lambda_1)\rangle_{\mathbf{w}} \cdot \mathbf{x}^a\eta^a\eta^b\mathbf{x}^b \cdot \langle \mathcal{R}(\mathcal{D}^{[b]},\lambda_2)\rangle_{\mathbf{w}} >_{P(\mathcal{D})} \tag{5.39}$$

Now since $a \neq b$ in this sum we can write $\mathcal{D}^{[b]} = \mathcal{D}^{a+C}$ where $\mathcal{D}^{a+C} \equiv \mathcal{D}^C + (\mathbf{x}^a, y_t(\mathbf{x}^a))$ as described in section 5.3.1 Substituting for the posterior averaged $\mathcal{R}$ vectors, our demonstration term, equation (5.39), becomes

$$\begin{aligned}\frac{4}{p^2N^2\sigma_x^4}\sum_{a\neq b}^{p}\langle &\left[\sum_{\mu^b=1}^{p-1}\eta^{\mu^b}(\mathbf{x}^{\mu^b})^T\mathbf{g}(\mathcal{D}^{b+C},\lambda_1) - \lambda_1(\mathbf{w}^o)^T\mathbf{g}(\mathcal{D}^{b+C},\lambda_1)\right]\cdot\mathbf{x}^a\eta^a \\ \times\ \eta^b\mathbf{x}^b\cdot &\left[\sum_{\nu^a=1}^{p-1}\eta^{\nu^a}(\mathbf{x}^{\nu^a})^T\mathbf{g}(\mathcal{D}^{a+C},\lambda_2) - \lambda_1(\mathbf{w}^o)^T\mathbf{g}(\mathcal{D}^{a+C},\lambda_2)\right]\rangle_{P(\mathcal{D})}.\end{aligned} \tag{5.40}$$

The sums involving ν^a and μ^b are over the examples

The response functions depend only on the input variables and not on the noise. Thus, we are free to average over the noise variables, the ηs, considerably simplifying the expression,

$$\frac{4\sigma^4}{p^2}\sum_{a\neq b}^{p} < \frac{1}{N^2\sigma_x^4}(\mathbf{x}^a)^T\mathbf{g}(\mathcal{D}^{b+C},\lambda_1)\mathbf{x}^b(\mathbf{x}^b)^T\mathbf{g}(\mathcal{D}^{a+C},\lambda_2)\mathbf{x}^a >_{P(\mathcal{D})}. \tag{5.41}$$

Here we should apologise for this slight abuse of notation, where we have integrated out the noise but have left the notation for the averaging, $<>_{P(\mathcal{D})}$, unchanged. We must now begin to average out the input variables. In order to do this we make use of the following identity used in Sollich (94a),

$$\begin{aligned}\mathbf{g}(\mathcal{D}^{a+C},\lambda_1) &= \mathbf{g}(\mathcal{D}^C,\lambda_1) - \frac{1}{N\sigma_x^2}\frac{\mathbf{g}(\mathcal{D}^C,\lambda_1)\mathbf{x}^a(\mathbf{x}^a)^T\mathbf{g}(\mathcal{D}^C,\lambda_1)}{1+\frac{1}{N\sigma_x^2}(\mathbf{x}^a)^T\mathbf{g}(\mathcal{D}^C,\lambda_1)\mathbf{x}^a} \\ &= \mathbf{g}(\mathcal{D}^C,\lambda_1) - \Delta_1^a \\ &= \mathbf{g}_1 - \Delta_1^a.\end{aligned} \tag{5.42}$$

This allows us to separate off the *extra* example, a, from the response function matrix, $\mathbf{g}$, such that we will be left with an expression in which the response functions depend

only on the data of the common blocks. Since, this is the case we have dropped the explicit dependence of $\mathbf{g}$ on $\mathcal{D}^C$ and furthermore have reintroduced the subscript to denote the dependence on λ_1 or λ_2. In the component Δ_1^a, the subscript again denotes a dependence on λ_1 whilst the superscript denotes the dependence on input, $\mathbf{x}^a$, the dependence on the common block being suppressed. We also note that, the terms in which the Δ_1^a, appear are of the order $\mathcal{O}(1/N)$ and since we only seek to calculate to this order we can take,

$$\frac{\Delta_1^a}{N} = \frac{1}{N}\frac{\mathbf{g}(\mathcal{D}^C,\lambda_1)\mathbf{x}^a(\mathbf{x}^a)^T\mathbf{g}(\mathcal{D}^C,\lambda_1)}{N\sigma_x^2(1+G_1)} + \mathcal{O}\left(\frac{1}{N^{3/2}}\right), \tag{5.43}$$

where in the thermodynamic limit $NG_1 = \text{tr } \mathbf{g}_1$. Thus, re-writing equation (5.41) using the identity (5.42) we obtain,

$$\begin{aligned} &= \frac{4\sigma^4}{p^2}\sum_{a\neq b}^{p} < \frac{1}{N^2\sigma_x^4}(\mathbf{x}^a)^T[\mathbf{g}_1 - \Delta_1^b]\mathbf{x}^b(\mathbf{x}^b)^T\left[\mathbf{g}_2 - \Delta_2^a\right]\mathbf{x}^a >_{P(\mathcal{D})} \qquad (5.44)\\ &= \frac{4\sigma^4}{p^2}\sum_{a\neq b}^{p} < \frac{1}{N\sigma_x^2}(\mathbf{x}^a)^T\left[\mathcal{F}^{b_0}_{\mathbf{g}_1,\mathbf{g}_2} - \mathcal{F}^{b_0,a_2}_{\mathbf{g}_1,\mathbf{g}_2,\mathbf{g}_2} - \mathcal{F}^{b_1,b_0}_{\mathbf{g}_1,\mathbf{g}_1,\mathbf{g}_2} + \mathcal{F}^{b_1,b_0,a_2}_{\mathbf{g}_1,\mathbf{g}_1,\mathbf{g}_2,\mathbf{g}_2}\right]\mathbf{x}^a >_{P(\mathcal{D})} \end{aligned}$$

Where we have introduced,

$$\begin{aligned} &\mathcal{F}^{\rho^1_{\sigma_1},\rho^2_{\sigma_2},\ldots\ldots\rho^{n-1}_{\sigma_{n-1}}}_{\mathbf{A},\ \mathbf{B},\ \mathbf{C},\ldots\ldots\ldots\ldots\mathbf{Z}}\\ &\quad\equiv \frac{m_{\sigma_1}m_{\sigma_2}\ldots\ldots m_{\sigma_{n-1}}}{(N\sigma_x^2)^{n-1}}\mathbf{A}\mathbf{x}^{\rho^1}(\mathbf{x}^{\rho^1})^T\mathbf{B}\mathbf{x}^{\rho^2}(\mathbf{x}^{\rho^2})^T\mathbf{C}\ldots\ldots\ldots\mathbf{x}^{\rho^{n-1}}(\mathbf{x}^{\rho^{n-1}})^T\mathbf{Z}, \end{aligned}$$

representing the scalar functions, of the response function matrices, to be averaged. Here the ρ^i take on the values a or b whilst σ_i is either, 0, 1 or 2 and the notation can be extended to any number, n, of square matrices $\mathbf{A}$, $\mathbf{B}$ *etc....* Finally, $m_0 = 1$, $m_1 = (1+G_1)^{-1}$ and $m_2 = (1+G_2)^{-1}$. We can now integrate out the examples not contained in the common block, $\mathcal{D}^C$ (*i.e.* $\mathbf{x}^a$ and $\mathbf{x}^b$). For example the first term in equation (5.44) averaged over these test points is

$$\begin{aligned} \frac{4\sigma^4}{p^2}\sum_{a\neq b}^{p} < \frac{1}{N\sigma_x^2}(\mathbf{x}^a)^T\mathcal{F}^{b_0}_{\mathbf{g}_1,\mathbf{g}_2}\mathbf{x}^a >_{P(\mathbf{x}^b,\mathbf{x}^a)} &= \frac{4\sigma^4 m_0}{N^2\sigma_x^4} < (\mathbf{x}^a)^T\mathbf{g}_1\mathbf{x}^b(\mathbf{x}^b)^T\mathbf{g}_2\,\mathbf{x}^a >_{P(\mathbf{x}^b,\mathbf{x}^a)}\\ &= \frac{4\sigma^4}{N}\widetilde{\text{tr}}\,(\mathbf{g}_1\mathbf{g}_2) + \mathcal{O}\left(\frac{1}{N^2}\right). \end{aligned}$$

Thus, retaining only terms to order $\mathcal{O}(1/N)$ we find that equation (5.39) can finally be

written,

$$\frac{4\sigma^4}{N}\left[\widetilde{\text{tr}}\,(\mathbf{g}_1\mathbf{g}_2) - m_2\widetilde{\text{tr}}\,(\mathbf{g}_1\mathbf{g}_2)\widetilde{\text{tr}}\,\mathbf{g}_2 - m_1\widetilde{\text{tr}}\,(\mathbf{g}_1\mathbf{g}_2)\widetilde{\text{tr}}\,\mathbf{g}_1 + m_1m_2\widetilde{\text{tr}}\,(\mathbf{g}_1\mathbf{g}_2)\widetilde{\text{tr}}\,\mathbf{g}_1\widetilde{\text{tr}}\,\mathbf{g}_2\right]$$

Thus, we see that this term is order $\mathcal{O}(1/N)$ as are all the (non vanishing) terms in the variance and co-variance calculations of section 5.3.

Response function

Finally, we note that the final forms, for both the variance and covariance like terms, are written in terms of the trace of the product of $\mathbf{g}_1$ and $\mathbf{g}_2$, $G_{12} = \widetilde{\text{tr}}\,(\mathbf{g}_1\mathbf{g}_2)$ which to date we have not calculated. Fortunately, however, we can obtain this quantity from the linear response function $G_1 = \widetilde{\text{tr}}\,\mathbf{g}_1$ which we have used extensively. Recall that this response function can be calculated as a special case of the response function calculated in Appendix 2.8 and indeed was calculated by Hertz *et al.*(89). We remind the reader that $\mathbf{g}_1 = (\mathsf{C} + \lambda_1\mathsf{I})^{-1}$ where C is the normalised correlation matrix of the inputs in the common block of data, $\mathcal{D}^C$. Thus we can write,

$$\begin{aligned}\text{tr}\,(\mathbf{g}_1\mathsf{C}\mathbf{g}_2) &= \text{tr}\,(\mathbf{g}_1(\mathsf{I} - \lambda_1\mathbf{g}_2)) \\ &= \text{tr}\,\mathbf{g}_1 - \lambda_1\text{tr}\,(\mathbf{g}_1\mathbf{g}_2),\end{aligned} \tag{5.45}$$

and similarly,

$$\begin{aligned}\text{tr}\,(\mathbf{g}_2\mathsf{C}\mathbf{g}_1) &= \text{tr}\,(\mathbf{g}_2(\mathsf{I} - \lambda_2\mathbf{g}_1)) \\ &= \text{tr}\,\mathbf{g}_2 - \lambda_2\text{tr}\,(\mathbf{g}_1\mathbf{g}_2).\end{aligned} \tag{5.46}$$

Since matrix multiplication commutes under the trace, equations (5.45) and (5.46) are equivalent and combining the last line in each we find that,

$$G_{12} = \frac{1}{N}\frac{\text{tr}\,\mathbf{g}_2 - \text{tr}\,\mathbf{g}_1}{\lambda_1 - \lambda_2}, \tag{5.47}$$

Which in the limit of $\lambda_1, \lambda_2 \to \lambda$ recovers the known result that $\text{tr}\,(\mathbf{gg}) = -\partial_\lambda \text{tr}\,\mathbf{g}$.

Chapter 6

Discrete Model Selection

Abstract

In this chapter we consider the problem of selecting model architectures from a discrete set of alternatives. In particular, we consider a method based on the noise sensitivity signature, **NSS**, of a model in addition to methods based the evidence and the cross-validation error. We consider this problem in the context of a simple example, namely the selection of the architecture of a piece-wise linear student trained and tested on data generated by a linear teacher whose outputs are corrupted by noise. Our analysis is average case and conducted in the thermodynamic limit. We establish that, in this context, model selection based on the **NSS** can, indeed, be used to select the architecture. However, we note that such methods fail to identify the optimal weight decay. Consideration of the cross-validation error shows that it too can be successfully employed to select the architecture in this scenario. Finally, we see that the evidence procedure leads to choice of the *true* architecture (*i.e.* linear), in the case where the linear model is optimally regularized. However, we note that a potential pitfall for all three methods is the case where the competing models are not optimally regularized. Indeed, we point out that such observations could be interpreted as support for the hierarchical Bayesian approach.

6.1 Introduction

To date we have considered various aspects of the model selection problem within a scenario in which the candidate models are distinguished by a number of continuous parameters (*i.e.* the hyper-parameters). Thus, we have had a continuum of possible models from which to choose. However, often model selection involves choices within a discrete set of models. For example, in the case of multi-layer perceptrons, discussed in section 1.1, we would like to select the appropriate number of hidden layers and units.

In this brief chapter we consider a simple example of such a model selection task in which we must choose between discrete model characteristics which specify the model architecture. In addition to cross-validation, **CV(1)**, and the Bayes factor (see section 1.2.1), we apply in this case a model selection criteria which we shall introduce shortly. This latter method is based on the *noise sensitivity signature* (**NSS**) of Grossman and Lapedes (94). We note that Meir and Fontanari (93) have made an analytical study of discrete model selection based on the stochastic complexity. Furthermore, Gelfand and Dey (94) present asymptotic results for the Bayes factor in a case of nested models.

As we have stated the main objective of this chapter is to gain some understanding of model choice from a discrete set of alternatives. Our focus to date has been selection of continuous regularization or hyper-parameters and in this chapter we compare and contrast the two problems. In the approach adopted here we shall, as we have done before, consider the hyper-parameters as part of the model specification. We shall see that the results of model choice as applied to architecture selection can be dependent on these parameters, which control the learning algorithm, as well as on the discrete model characteristics. A possible objection to this approach is based on the interpretation of the regularization parameters as hyper-parameters. As discussed in section 2.2.1, the hierarchical Bayes scheme would have us integrate out these hyper-parameters and thus no such dependence would be apparent, although, even in this scheme, different noise models and prior specifications would affect the selection of discrete characteristics. Here we do not consider this effect, examining only one form for the likelihood and prior namely that adopted in section 2.2.1 and used throughout this thesis. We rebut the foregoing criticism in two ways. Firstly, we argue that in practical situations one may not wish to integrate out the regularization parameters from the point of view of computational cost. Thus, a number of students may be trained, each based on a different architecture and possibly regularization scheme. We then seek to compare the merits of each architecture on the basis of these students, as reflected in our analysis. The second argument relies on the validity of the approximation of hierarchical Bayes by the **MLII** (evidence) procedure. Thus, as we did in section 2.4, we argue that, in the scenario considered, the setting of the hyper-parameters by the evidence is equivalent to the full hierarchical calculation in the thermodynamic limit. Interestingly, however, we will see that this presents some support for the use of the hierarchical Bayesian approach. In summary, then, we will sometimes consider the hyper-parameters as, in part, specifying the model, whilst when we employ the evidence procedure assignments argue that we are doing the full hierarchical calculation, since we conduct our analysis in the thermodynamic limit.

Learning scenario

In accord with our general approach we seek insight into this discrete model selection problem by performing an illustrative calculation in a relatively simple learning scenario which we now describe. In general we adopt the notation of the earlier chapters (see *e.g.* section 2.2.2). In, particular we consider a teacher with real one dimensional output, $y_t(\mathbf{x})$, described by the conditional density $P(y_t \mid \mathbf{x})$. Our sampling assumption is $P(\mathbf{x})$ and therefore a data set $\mathcal{D}$= $\{(y_t(\mathbf{x}^\mu), \mathbf{x}^\mu) : \mu = 1..p\}$ is generated with probability $P(\mathcal{D})$ = $\prod_{\mu=1}^{p} P(y_t \mid \mathbf{x}^\mu)P(\mathbf{x}^\mu)$. In fact, we will consider only a linear teacher corrupted by Gaussian noise of variance σ^2 in this chapter and thus write, $P(y_t \mid \mathbf{x}^\mu) \propto \exp[-(y_t^\mu - \mathbf{w}^o.\mathbf{x}^\mu/\sqrt{N})^2/2\sigma^2]$. Furthermore, our sampling assumption is also Gaussian, with mean zero and variance σ_x^2. Therefore, where $\sigma_w^2 = \frac{1}{N}(\mathbf{w}^o)^T\mathbf{w}^o$, the noise to signal ratio is $\lambda_0 = \sigma^2/\sigma_w^2$ as before.

The learning algorithm (*i.e.* noise model and prior) we consider here is that introduced in section 2.2.2 and used throughout this thesis. In other words, the posterior distribution of student parameters is based on a Gaussian noise model and a prior which corresponds to regularization by weight decay. Also, as before we will make predictions, at input $\mathbf{x}$, using the average over the predictive distribution conditioned on the training data which is written $\langle y_s(\mathbf{x})\rangle_{P(y_s|\mathbf{x},\mathcal{D},\mathcal{M})}$. In the case under consideration in this chapter the architecture will not be fixed as it was previously and therefore the model specification, $\mathcal{M}$, is written, $\mathcal{M} = \{\beta, \lambda, \mathcal{A}\}$, where we now define the architecture, $\mathcal{A}$, in the current context.

The students we study here are the piece-wise linear students introduced in chapter 3. We recall equation (3.1) which defined the student output, $y_s(\mathbf{x})$, at input $\mathbf{x}$ as;

$$y_s(\mathbf{x}) = \sum_{k=1}^{n} \frac{1}{\sqrt{N}} \mathbf{w}^k \cdot \mathbf{x}\delta_k(\mathbf{x}) \tag{6.1}$$

where the functions $\delta_k(\mathbf{x}) = 1$ on some region $\mathcal{Z}_k$ of the input space and are zero everywhere else, here we also assume these regions are non-overlapping. The parameters $\mathbf{w}^k$ are the weights of the k^{th} component student and the learning algorithm introduced in section 3.3 has n weight decay parameters λ_k. The discrete model characteristics are thus, the number of linear pieces, n, making up the student. In terms of notation the architecture, $\mathcal{A}$, is then specified by $\mathcal{A} = n$ and our model selection criteria must select the appropriate architecture. In what follows we consider only a simplified problem, namely the choice between the *true* model, $n = 1$, and the model $n = 2$. In the statistics literature such a scenario is referred to as a case of nested models since the *full* model, $n = 2$, can represent the *reduced*, $n = 1$, model. We adopt the constraints,

$\delta_k(\mathbf{x})$, of section 3.4 and the main elements of the necessary calculations are given in section 3.4.1 to which we refer the reader. Nonetheless, results directly relevant to the current study will be presented in this chapter. Here the reduced model is denoted by $\mathcal{M}_1 = \{\beta, \lambda, n=1\}$ and the full model by, $\mathcal{M}_2 = \{\beta, \lambda_1, \lambda_2, n=2\}$. However, in this chapter, since we will only be concerned with the generalization error (see equation 3.10) and since we make predictions using $\langle y_s(\mathbf{x}) \rangle_{P(y_s|\mathbf{x},\mathcal{D},\mathcal{M})}$ we will not be concerned with the inverse temperature β which controls the variance of this distribution. Indeed, all the quantities considered here are independent of β.

In the next section we introduce the noise sensitivity signature and a method of model selection based on it. We then briefly review the Bayes factor and the cross-validation error in this case. Finally, we will consider the relative strengths and weaknesses of these three model selection criteria as applied to the architecture selection problem outlined above.

6.2 Noise sensitivity signature

The basic idea behind the noise sensitivity signature, introduced by Grossman and Lapedes (94), is that complex, overly parameterised, models will over-fit the data more than those of an *appropriate* complexity. Indeed, it is the determination of what appropriate means in this context which has preoccupied us throughout this thesis. In order to determine whether a particular model is over-parameterised these authors suggest adding noise to the examples in our data base, $\mathcal{D}$, to create new data sets. Various student models are then trained on these noisy data sets in order to assess to what extent each of them fits the added noise. We describe below, in more detail, how this is achieved.

The advantages of this approach, as spelled out by these authors, are that unlike penalty based approaches it is not problem dependent (see *e.g.* chapter 2) and in contrast to cross-validation does not require that any of the data be *left out* for testing. Furthermore, Grossman and Lapedes (94) suggest that the **NSS** method can be used to set *any* complexity parameter in a model, in other words the architecture in addition to the regularization parameters. We note that these proposals were made in the context of classification algorithms; in our language, discrete input-output mappings. Indeed, they demonstrated the utility of this approach experimentally, by selecting the number of hidden units in an **MLP**, which was a binary classifier [1], in a case where the teacher was known. Nonetheless, the criteria used to judge the appropriate number of hidden

[1] Referring the reader to section 1.1 in this case the transfer functions are $\mathrm{sgn}(h)$

units was somewhat heuristic. One motivation for the study presented here was to reduce this degree of subjectivity. As noted above the problem which we explore in this chapter is a mapping between real inputs and outputs, in contrast to the discrete case considered by Grossman and Lapedes. Indeed, in the case studied here the very rich behaviour revealed by these authors is much simplified. Nevertheless, we show that the **NSS** method can be applied with some success in this case and our results perhaps point to its strengths and weaknesses in a more general setting.

The noise sensitivity signature of a model is based on a number of quantities which we now discuss, noting that the definitions given here pertain to mappings between continuous spaces. As stated above we produce a number of new data sets by adding noise to the existing data base, $\mathcal{D}$. In the current context we add zero mean Gaussian noise, of variance θ^2, to the outputs of our data base to produce corrupted data sets, $\mathcal{D}(\theta, d) = \{(y_t(\mathbf{x}^\mu, \theta_d^\mu), \mathbf{x}^\mu) : \mu = 1..p\}$. Here $d = 1...n_d$ indexes such sets and $y_t(\mathbf{x}^\mu, \theta_d^\mu)$ denotes that, in the d^{th} data set, the output of the μ^{th} example has been corrupted by zero mean Gaussian noise, of variance θ^2. In other words each example of each data set is independently corrupted by noise of the same variance. Using the learning algorithm of chapters 2 and 3, we train students on these corrupted data sets. The response of such a student at a novel input, $\mathbf{x}$, is the average over the predictive density, $P(y_s \mid \mathbf{x}, \mathcal{D}(\theta, d), \mathcal{M})$. The first quantity we consider is the average error of students, thus generated, when tested on the original data set. That is,

$$E_{\text{clean}}(\mathcal{D}, \theta) = \frac{1}{pn_d} \sum_{d=1}^{n_d} \sum_{\mu=1}^{p} \left(\langle y_s \rangle_{P(y_s|\mathbf{x}^\mu, \mathcal{D}(\theta,d), \mathcal{M})} - y_t(\mathbf{x}^\mu) \right)^2 \tag{6.2}$$

Similarly, when tested on their training set, the average error of these students is,

$$E_{\text{noise}}(\mathcal{D}, \theta) = \frac{1}{pn_d} \sum_{d=1}^{n_d} \sum_{\mu=1}^{p} \left(\langle y_s \rangle_{P(y_s|\mathbf{x}^\mu, \mathcal{D}(\theta,d), \mathcal{M})} - y_t(\mathbf{x}^\mu, \theta_d^\mu) \right)^2 . \tag{6.3}$$

Using these two quantities we can gain some insight into the appropriate model complexity. Specifically, we expect the performance on the clean data set to be superior to that on the corrupted sets. That is, we expect $E_{\text{clean}}(\mathcal{D}, \theta) < E_{\text{noise}}(\mathcal{D}, \theta)$ if this is not so then the model under consideration is (over) fitting the added noise and, in the **NSS** scheme is considered to be overly complex. In the original study a third quantity was introduced to complete the characterisation of the noise sensitivity signature of a model, namely the average functional distance between classifiers. In the current context we can examine the test set estimate of the average distance between the outputs

of students trained on different realizations of the corrupted data sets, that is,

$$\text{Dist}(\theta) = \frac{1}{m}\sum_{\mu=1}^{m}\frac{1}{n_d(n_d-1)}\sum_{d_1=1}^{n_d}\sum_{d_2=d_1}^{n_d}\left(\langle y_s\rangle_{P(y_s|\mathbf{x}^\mu,\mathcal{D}(\theta,d_1),\mathcal{M})} - \langle y_s\rangle_{P(y_s|\mathbf{x}^\mu,\mathcal{D}(\theta,d_2),\mathcal{M})}\right)^2. \tag{6.4}$$

Here m is the number of test points used to evaluate the functional distance, but we note that these can be unlabelled data, the indices d_k denote the different data sets. Thus, Dist(θ), measures the sensitivity of the model to the noise in the training data. In particular, we expect this measure to be relatively small for models of appropriate complexity since in this case the changes from one data set to the next will not affect the students output much, at least when the added noise level is low. However, we expect that over-parameterised models will learn this noise and thus the outputs of different students will vary considerably from one data set to the next. Similarly, an under-parameterised model will be unable to learn even the uncorrupted data and thus, we can expect significant variation when trained on the corrupted data sets. We note that in the case we consider here we are unable to explore a significantly under-parameterised model.

6.2.1 Average case

In contrast to chapters 4 and 5 here we investigate the average case in the thermodynamic limit. However, although we do not explore this issue here we expect the system to be self averaging. In fact, for the linear, $n = 1$, model we have explored this issue in chapter 4. In addition, we saw in chapter 3 that the free energy of the $n = 2$ model is simply a sum of terms from the free energies corresponding to suitably chosen linear systems.

Regarding the errors on the uncorrupted and corrupted data sets (equations 6.2 and 6.3) we shall consider the average difference,

$$\begin{aligned}\Delta E_{\text{NSS}}(\theta) &\equiv \langle E_{\text{noise}}(\mathcal{D},\theta) - E_{\text{clean}}(\mathcal{D},\theta)\rangle_{P(\mathcal{D},\theta)} \\ &= \theta^2\left[1 - \frac{2}{\alpha}\sum_{k=1}^{2}(1-\lambda_k G_k)\right] + \mathcal{O}\left(\frac{1}{N}\right)\end{aligned} \tag{6.5}$$

Where we have conducted the average, in the thermodynamic limit, over the supplied data set and over the extra noise added to the examples. Since we are considering the learning scenario of chapter 3 the response functions G_k are as given in equation (3.8).

A similarly straightforward calculation shows that the average distance between

two students trained on different corrupted data sets is,

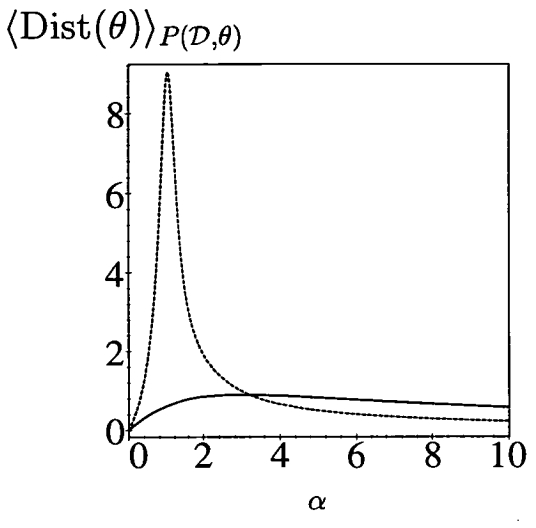
$$\langle \text{Dist}(\theta) \rangle_{P(\mathcal{D},\theta)} = 2\theta^2 \sum_{k=1}^{2} (G_k + \lambda_k \partial_\lambda G_k) + \mathcal{O}\left(\frac{1}{N}\right) \tag{6.6}$$

where ∂_y denotes the partial derivative *w.r.t.* y.

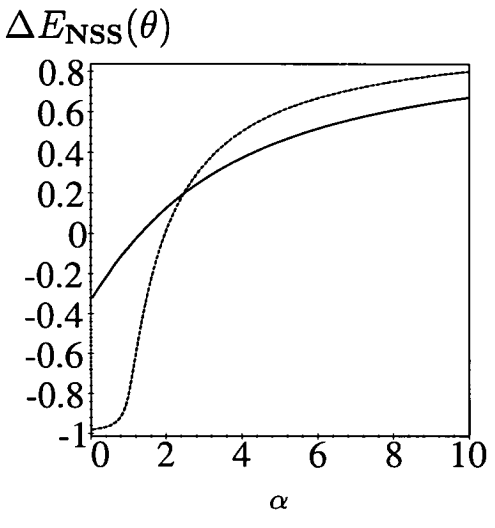
As already noted, the behaviour witnessed here is considerably simpler than that to which Grossman and Lapedes (94) testify. Immediately obvious is the linear dependence of both quantities (equations 6.5 and 6.6) on the added noise level, θ. This is in stark contrast to the non-linear behaviour seen by Grossman and Lapedes. However, in general this simplification is no surprise since we are conducting an average case analysis of a linear system in contrast to their experimental study of a highly non-linear binary system. Nonetheless some qualitative features are preserved here.

In particular, the noise sensitivity signature of the two models, as revealed by $\Delta E_{\text{NSS}}(\theta)$ and $\langle \text{Dist}(\theta) \rangle_{P(\mathcal{D},\theta)}$, allows us to choose between the two competing models. Figure 6.1 illustrates the situation for a case where both candidate models are optimally regularized. Graph (a) in figure 6.1 shows that the average distance, over the corrupted data sets, between two students generated from the $n = 2$ model is larger than that for students with the $n = 1$ architecture. This suggests that the $n = 2$ model is over parameterised. A similar conclusion can be drawn from $\Delta E_{\text{NSS}}(\theta)$ (see figure 6.1 (b)) which, shows that the $n = 2$ model fits the added noise to a greater extent than does the $n = 1$ model. In this case both models over fit the added noise with $\Delta E_{\text{NSS}}(\theta) < 0$ for small α but model $\mathcal{M}_2$ is the worse offender and we would expect its performance to suffer as a result. This is confirmed by the generalization performance of each model, in this case shown in figure 6.1 (c). In the cases where both models are over or under regularized to the same degree we find that the **NSS** prescription as used above identifies the optimal model. That is, model $\mathcal{M}_1$ ($n = 1$) is chosen and this model has the lowest of the generalization errors.

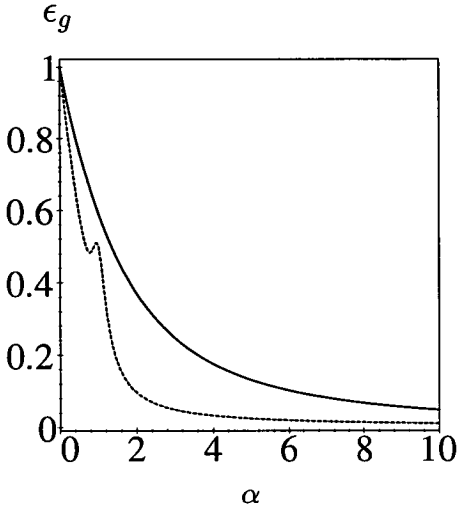
However, in the case where the models are not optimally regularized and the $n = 1$ model is under-regularized to a greater degree than the $n = 2$ model we find that, in general, the **NSS** prescription breaks down. This is shown in figure 6.2 where in graph (a) we see that, for a small number of examples (α), the average distance, $\langle \text{Dist}(\theta) \rangle_{P(\mathcal{D},\theta)}$, is larger for the $n = 1$ model. In this case because the regularization is weak the model can over-fit the corrupted data; asymptotically this is not the case and the $n = 2$ model is more sensitive to the added noise. A similar effect is indicated by $\Delta E_{\text{NSS}}(\theta)$, in figure 6.2 (b), where, again for small α, the $n = 2$ model over-fits the corrupted data less. Nonetheless, asymptotically, in the data dominated region, the



(a) **NSS**: distance



(b) **NSS**: errors



(c) Generalization

Figure 6.2. Noise sensitivity signature and generalization error: noise to signal ratio $\lambda_0 = 0.1$ with the $n = 1$ students under-regularized model $\mathcal{M}_1 = \{\beta, \lambda = 0.01, n = 1\}$ whilst the $n = 2$ students are over-regularized such that, $\mathcal{M}_2 = \{\beta, \lambda_1 = \lambda_2 = 1, n = 2\}$. In all three graphs quantities associated with $\mathcal{M}_1$ are shown by the dashed curve whilst solid curves relate to $\mathcal{M}_2$. In graph (a) for a small number of examples the average distance between students of model $\mathcal{M}_1$ is larger than that between students of $\mathcal{M}_2$. For large α the $n = 2$ model is more sensitive to the added noise. Graph (b) shows that the under-regularized $n = 1$ model over-fits the added noise to a greater extent than the over-regularized $n = 2$ model for small α. Again for larger numbers of examples we find that $\mathcal{M}_1$ does not over-fit as badly as $\mathcal{M}_2$. This is linked to the fact that regularization (prior knowledge) plays an ever decreasing role as the amount of data increases. Based on graphs (a) and (b), given a small number of examples, we would choose the $n = 2$ model whereas graph (c) reveals that the $n = 1$ model achieves better generalization performance for all α.

true model ($n = 1$) is seen to win out as we would expect. Thus, in the example shown in figure 6.2, the **NSS** prescription would have us choose the $n = 2$ model for small α and the $n = 1$ model as the number of examples grows. However, figure 6.2 (c) reveals that the generalization error is, in fact, smaller for the $n = 1$ model irrespective of α. The noise sensitivity signature has been misleading in this case, both in terms of generalization performance and architecture selection. Thus when comparing model architectures, at least if using model selection based on the **NSS**, one should be careful to set the regularization parameters optimally, or at the very least, in a way which does not put one of the architectures at a significant disadvantage.

This raises the question of whether the regularization parameter, the weight decay, can be set using the **NSS** approach. In fact, when we consider that in general the optimal weight decay is related to the true noise level inherent in the original data set, it is difficult to see how this could be achieved by adding further noise. Indeed, in the case under consideration the average distance, $\langle \mathrm{Dist}(\theta) \rangle_{P(\mathcal{D},\theta)}$, is minimised, and the difference between the corrupted and uncorrupted data set errors, $\Delta E_{\mathbf{NSS}}(\theta)$, maximised, by infinite weight decay independent of the true noise level, λ_0.

We therefore suggest that the noise sensitivity method proposed by Grossman and Lapedes (94) is not suited to choosing regularization parameters. However, provided that the models are optimally regularized it does provide a novel approach to architecture selection.

6.3 Cross-validation, CV(1)

As we saw in section 5.3 the leave-one-out cross-validation error is equivalent to the generalization error on average. Thus, in the thermodynamic limit, if self averaging holds, the cross-validatory error coincides with the generalization error and thus cross-validation will always optimise generalization performance. In terms of selecting a model architecture, however, one should still be cautious. This is because, if inappropriately regularized, the *true* model may well perform worse than the alternatives and thus may not be chosen by **CV(1)**. This was noted by Plutowski *et al.*(94) who pointed out that **CV(1)** may pick the more complex model if it performed better in terms of the cross-validated error.

In particular, this will occur in the data limited regime (small α), if the true model is significantly under-regularized with respect to the overly parameterised models. Indeed, this is what we found above when considering the noise sensitivity signature. Furthermore, this is likely to be the case for any model selection criterion based on measures of prediction accuracy, since in general even the *true* model will perform

badly if poorly regularized and presented with only a few noisy examples.

Given the need to use the optimal regularizer, in the current example, one could either use the evidence procedure (see section 3.6) or **CV(1)**, which both choose the optimal weight decay in the thermodynamic limit. Moreover, we argued in section 2.4 that fixing the hyper-parameters using the evidence procedure is equivalent to the full hierarchical Bayesian calculation in the thermodynamic limit. If this is so then the advice to integrate out the hyper-parameters seems sound in light of the preceding discussion. When they are not equivalent, it is an open question as to whether **ML II** or the full hierarchical Bayes approach is to be preferred.

6.4 Evidence

Finally, we ask whether we can identify the true model architecture by comparing the evidence for the competing models, $P(\mathcal{D} \mid \mathcal{M}_i)$. That is, we consider the Bayes factor (see equation 1.7) for the models, $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$\mathbf{B}_{\mathcal{F}} = \frac{P(\mathcal{D} \mid \mathcal{M}_1)}{P(\mathcal{D} \mid \mathcal{M}_2)} \quad . \tag{6.7}$$

Recall that the model specification $\mathcal{M}_i$ includes the hyper-parameters. In particular, we want to know which of these probabilities is the greater. In section 3.6, for the system of current concern, we calculated the average of the free energy $f = -\frac{1}{N} \ln P(\mathcal{D} \mid \mathcal{M}_i)$ (see equation 3.6) and thus we concentrate on the normalised logarithm of the Bayes factor, $\frac{1}{N} \ln \mathbf{B}_{\mathcal{F}}$. We note that model $\mathcal{M}_1$ is preferred if this quantity is positive and model $\mathcal{M}_2$ if it is negative. The chief difficulty here is the calculation of the determinants of the response function matrices $\mathbf{g}_k$ which appear in the free energy. However, as outlined in appendix 6.6 the appropriate quantities can at least be calculated numerically.

Applying the Bayes factor (equation 6.7) to the the determination of the number of segments of our piece-wise linear student we find that as with the noise sensitivity method and **CV(1)** the results depend on the regularization parameters of each model. Evaluation of $\frac{1}{N} \ln \mathbf{B}_{\mathcal{F}}$ in the case where the linear model $\mathcal{M}_1$ is optimally regularized shows that the evidence favours the linear model irrespective of the regularization of the piece-wise linear model $\mathcal{M}_2$. In this case the linear model also achieves the lowest generalization error and thus the evidence not only favours the correct architecture but also the best generalizer. However, when the linear model is sub-optimally regularized the situation is more complicated. Figure 6.3(a) shows for $\alpha = 0.5$ the normalised logarithm of Bayes' factor versus the weight decay setting of the linear model, λ, when the piece-wise linear model is optimally regularized. In the instance shown the optimal

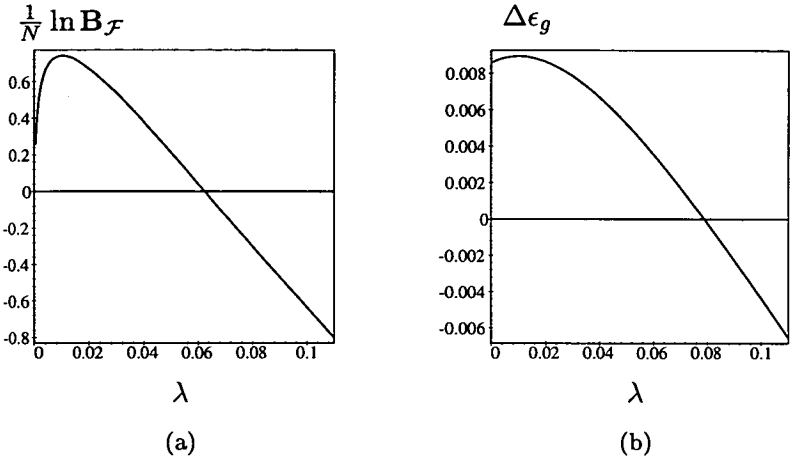


Figure 6.3. Model determination from the evidence: Graph (a) shows the normalised logarithm of the Bayes factor (equation 6.7) for fixed α versus the weight decay of the linear model, λ, in the case where the piece-wise linear model ($\mathcal{M}_2$) is optimally regularized. In the scenario depicted $\alpha = 0.5$ and the noise to signal ratio $\lambda_0 = 0.01$. The evidence favours model $\mathcal{M}_1$ when $\frac{1}{N}\ln \mathbf{B}_{\mathcal{F}} > 0$ and model $\mathcal{M}_2$ when $\frac{1}{N}\ln \mathbf{B}_{\mathcal{F}} < 0$. Graph (b) shows the difference in generalization errors, $\Delta\epsilon_g = \epsilon_g(\mathcal{M}_2) - \epsilon_g(\mathcal{M}_1)$ versus the model $\mathcal{M}_1$ weight decay (λ) in the same case. We see that for small and large λ the evidence favours the best generalizer whilst the for an intermediate range the evidence favours model $\mathcal{M}_2$ whilst model $\mathcal{M}_1$ has lower generalization error.

weight decay is $\lambda_0 = 0.01$, but the evidence favours the linear model for a range of values around this. However, for weight decay parameters $\lambda >\approx 0.06$ the evidence favours the piece-wise linear model. Figure 6.3(b) shows the difference in the generalization error of model $\mathcal{M}_2$ less that of model $\mathcal{M}_1$, $\Delta\epsilon_g = \epsilon_g(\mathcal{M}_2) - \epsilon_g(\mathcal{M}_1)$ in the same case. We see that for small values of the weight decay the linear student has a lower generalization error than the piece-wise linear model. Thus, the evidence favours the correct architecture and the best generalizer for small λ. However, when the linear weight decay λ is between ≈ 0.06 and ≈ 0.08 the evidence favours the piece-wise linear model whilst the linear model is the best generalizer. For larger values of the linear student weight decay, λ, the evidence favours the model $\mathcal{M}_2$ which is also the superior generalizer. Furthermore, although we have shown only the $\alpha = 0.5$ case, this picture holds for a wide range of α investigated.

Thus, we have seen in the case of the evidence that the regularization of the models under comparison is of crucial importance. Indeed, as we found with the **NSS** and with **CV(1)** unless the models are optimally regularized the evidence is not guaranteed to choose the correct model. We note that although we were unable to investigate the large α limit, Gelfand and Dey (94) demonstrated that in the case of nested models the

evidence will asymptotically favour the correct model architecture and may also fail to optimise generalization performance. Furthermore, Meir and Fontanari (93) reveal, in the thermodynamic limit, that minimization of the stochastic complexity reliably chooses the correct model from a nested class only asymptotically.

6.5 Comparison and summary

To summarise, we applied the noise sensitivity signature approach to the case of linear and piece-wise linear models. In particular, we found that in the case of a piece-wise linear student learning a linear teacher this method can be used to select the appropriate student architecture when the competing models are optimally regularized. However, we argued that, in general, one could not set regularization parameters optimally through consideration of the **NSS** and indeed, in the example considered here this was found to be the case. In the case of optimally regularized models both **CV(1)** and the Bayes factor can also be used to select the optimal architecture.

Furthermore, in the learning scenario studied here we found that all three model selection methods were not guaranteed to select the *true* architecture when the competing models were sub-optimally regularized. In addition, we found that model selection based on the **NSS** and the evidence could, in this case, not only pick the *full* model, $\mathcal{M}_2$, but also the model with the lowest generalization ability. In contrast, since we are working in the thermodynamic limit, leave-one-out cross-validation always picks the optimal generalizer. Thus, our results suggest that when comparing models one should optimise the regularization of each model first. Then through comparison of these regularized models one can contrast different architectures. In fact, one might interpret this as an argument for hierarchical Bayes, in that one should integrate out the hyper-parameters in the models before one can compare the different architectures.

6.6 Appendix: Bayes' factor

In this appendix we discuss the calculation of the Bayes factor (equation 6.7) or rather, as discussed in section 6.4 we will examine its normalised logarithm,

$$\frac{1}{N}\ln \mathbf{B}_{\mathcal{F}} = \frac{1}{N}\ln P(\mathcal{D} \mid \mathcal{M}_1) - \frac{1}{N}\ln P(\mathcal{D} \mid \mathcal{M}_2) \ . \tag{6.8}$$

Thus, since we are conducting our analysis in the thermodynamic limit we need to calculate the data average of the free energy given in equation (3.6) for the models $\mathcal{M}_1$ and $\mathcal{M}_2$. With the exception of the determinants of the matrices $\mathbf{g}_k$ all the

terms in the resulting expression can be written as functions of the response functions $G_k = \frac{1}{N} < \text{tr } \mathbf{g}_k >_{P(\mathcal{D})}$. The determinants themselves are given by applying the identity (4.A13) given in appendix 4.8.1. Individually, these integrals diverge, but they can be re-written as follows,

$$\begin{aligned} \frac{1}{N} \ln \det \mathbf{g}_k(\lambda_k) &= \int_{\lambda_k}^{\infty} \frac{1}{N} \text{tr } \mathbf{g}_k(\lambda') d\lambda' - \int_1^{\infty} \frac{1}{\lambda'} d\lambda' \\ &= \int_{\lambda_k}^{\infty} \left[\frac{1}{N} \text{tr } \mathbf{g}_k(\lambda') - \frac{1}{\lambda'} \right] d\lambda' - \int_1^{\lambda_k} \frac{1}{\lambda'} d\lambda' \quad . \end{aligned} \tag{6.9}$$

Since the first integral in the second line does not diverge and the last of these terms is simply $\ln \lambda$ then the average normalised logarithm of the determinant also depends on the response function G_k and can thus be calculated, at least numerically.

Chapter 7

Summary and outlook

We now briefly summarize our results drawing attention to connections between them. Finally we conclude by discussing future topics of research which suggest themselves.

We began in chapter 2 by defining the learning scenario via a cost function penalizing complexity. Langevin dynamics based on this cost function were interpreted within the Bayesian framework outlined in chapter 1. The model selection problem was thus cast in terms of the setting of hyper-parameters parametrising the prior and the noise model. Since we choose to make predictions based on the conditional predictive distribution, performance was controlled by selection of these hyper-parameters. This learning algorithm was the main focus of attention throughout the thesis.

In chapters 2 and 3 we were primarily concerned with the effects of a mismatch between Bayesian model specifications and the underlying data generating process. Indeed, despite our assumptions, in real problems the data is highly unlikely to have been generated by an artificial neural network. In both chapters the principal method of model selection investigated was the evidence procedure.

In chapter 2 we considered the case in which the student was not sufficiently powerful to model the teacher. Given that one is constantly attempting to avoid over-fitting this is very likely to be the case in practice. The learning scenario considered allowed us to interpolate between the learnable linear case and an unrealizable case in which the teacher was a non-linear function. Our analysis was conducted in the thermodynamic limit and average case; if self averaging holds then fluctuations around this average vanish in this limit. Comparison of the hyper-parameters derived from the evidence with the optimal assignments revealed that the evidence procedure was sub-optimal in the unrealizable non-linear case. Furthermore, we noted that the $\mathbf{CV(1)}$ assignments were optimal in the thermodynamic limit. However, as Mackay has suggested, the failure of

the evidence to identify the optimal rule is indicative of the fact that our model assumptions are inadequate. The evidence procedure thus offers us the opportunity to discover this and improve our model. The robustness of a Bayesian procedure to changes in the prior assumptions is generally considered desirable. In terms of performance we investigated the robustness of the evidence procedure to changes in the validity of the prior assumptions, once again in the thermodynamic limit. We concluded that in terms of expected generalization performance the evidence assignments are remarkably robust. Thus, even in the unrealizable case the evidence procedure might be considered as an alternative to the computationally expensive cross-validation.

In chapter 3 we investigated the performance of the evidence procedure assignments in the case where the student is more than able to represent the teacher. We considered, in the thermodynamic limit, a piece-wise linear student learning a linear teacher. We found that the evidence procedure identified the optimal hyper-parameters for the piece-wise linear student, despite the fact that the generalization error of this student was higher than that of the linear student. Thus, our results suggest that the evidence procedure is more sensitive to unrealizability than to over-realizability. However, they also suggest that overly powerful models, even when optimally regularized, perform worse than the *true* model and as we noted this may have some relevance to the work of Neal (94).

To this point we had studied model selection in supervised learning using average case analyses in the thermodynamic limit. In chapters 4 and 5 we sought to understand how one could expect model selection procedures to perform in practice. To this end we considered the first order finite size corrections to the thermodynamic limit and as in earlier chapters we considered the problem of setting the regularization parameters in our penalty based algorithm.

Initially, we explored these issues in terms of the evidence procedure assignments revealing that even in the learnable linear case these assignments were not optimal in general. We found the evidence procedure to be inconsistent in terms of weight decay assignment and that the degradation in performance associated with the evidence procedure was an order $\mathcal{O}(1/N)$ quantity for large systems. However, in the noiseless limit we found a phase transition in behaviour as the number of training examples, α, increased, above which the evidence assignments were seen to be optimal. Nonetheless, even for small noise the degradation in performance resulting from the use of the evidence procedure was seen to be considerable, particularly as $\alpha \to 0$. Using numerical simulations of small systems we found qualitatively the same behaviour as that suggested by our finite size corrections.

In chapter 5 we switched attention to model selection based on test set and cross-validatory estimates of the generalization error. In the former we considered how best to partition the data base of examples into training and testing subsets. We found that the optimal partition resulted in a degradation in performance which was an order of magnitude, $\mathcal{O}(\sqrt{N})$, larger than that associated with the evidence assignments. Furthermore, due to the difficulty in identifying this optimal partition, we noted that in practice the performance associated with the test set assignments was likely to be even worse. Examination of the leave-one-out cross-validatory assignments revealed a performance degradation of the same order as the evidence procedure. Thus, **CV(1)** makes better use of the test set than the naive approach, albeit at added computational cost. In fact, we saw that, in general, the performance associated with **CV(1)** was also superior to that obtained by the evidence procedure although, again, this was achieved through greater computational effort; simulation results for low dimensional systems supported this assertion. However, our finite size corrections showed that the evidence procedure and **CV(1)** were indistinguishable, in performance terms, both asymptotically (*i.e.* $\alpha \to \infty$) and in the noiseless limit, for $\alpha > 1$. Thus, in such instances the less computationally intense evidence procedure should be used.

In chapter 6 we focused our attention on the problem of architecture selection. In particular, this issue was explored in the scenario in which the number of segments had to be chosen for a piece-wise linear student learning from examples generated by a linear teacher. We discussed a method of model selection based on the noise sensitivity signature, **NSS**, and found it to provide a novel method for architecture selection. Furthermore, we considered the utility of **CV(1)** and the evidence in this situation. We found that all three methods could choose the over-parameterised model in cases where the models were inappropriately regularized. Moreover, in such cases use of the evidence or of the **NSS** could result in the selection of the model with the worse generalization ability. These results suggest that models should be optimally regularized to allow fair comparison of different architectures.

Our aim in this thesis was to investigate some of the strengths and weaknesses associated with various model selection criteria. This we have done but many open questions remain. The most obvious of these is to what extent our results, derived in relatively simple learning scenarios, will carry over to more general cases. For example, from general arguments, it seems that the evidence procedure will be optimal on average in the realisable case. However, whilst we have considered the important question as to the effects of over and under realizability in some simple cases, the degree of robustness in evidence based model selection in general is not clear. Furthermore, our finite size corrections revealed subtleties in behaviour which are not apparent in standard

analytical approaches. However, it is not clear to what extent these results will be applicable in general and should be treated with care when applied to real training scenarios. Indeed, it seems probable that such questions will depend crucially on the particular circumstances involved, and perhaps no general statements can be made, at least not on the basis of analysis of specific learning scenarios. It would nonetheless be interesting to extend the work of this thesis to examine more complex learning scenarios in order to assess the generality of our results. Amongst possible avenues are the extension of penalty based approaches to layered neural networks such as the committee machine. Moreover, methods of model selection such as **CV(1)** or those based on the **NSS** could be applied in the on-line training scenario taking advantage of recent developments there.

Bibliography

Anderson J A and Rosenfield E (Eds) 1988 *Neurocomputing: Foundations of research* **Cambridge, Massachusetts: The MIT Press.**

Anthony M 1995 Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. In *The computational and Learning Complexity of Neural Networks.* **Parberry (Ed.)** *in press* **Cambridge, Massachusetts: The MIT Press.**

Barber D, Sollich P and Saad D 1995 Finite size effects in learning in linear perceptrons *J. of Phys. A: Math. Gen.* **28:1325-1334.**

Barber D 1995 *Finite size effects in neural networks.* **University of Edinburgh Thesis.**

Barron A R and Cover T M 1991 Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37:1034-1054.**

Baum, E and Haussler, D 1989 What size net gives valid generalisation? *Neural Computation* **1:151-160.**

Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis.* **Second Edition. New York: Spinger-Verlag.**

Berger J O and Berliner L M 1983 Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *Tech. Report* **83-35. Purdue University, West Lafayette: Department of Statistics.**

Bishop C 1995 *Neural Networks for Pattern Recognition.* **Oxford: Oxford University Press.**

Bishop C, Haynes P S, Smith M E U, Todd T N, Trotman D L and Windsor C G 1995 Real-time control of a Tokamak plasma using neural networks. In *Advances in Neural Information Processing Systems* **7:1007-1014 Tesauro G, Touretzky D S and Leen T K (Eds.). Cambridge, Massachusetts: The MIT Press.**

Bös S, Kinzel W and Opper M 1993 Generalisation ability of perceptrons with continuous outputs *Physical Review E* **47: 1384-1391.**

Brieman L 1994 Understanding non-linear neural networks from their linear relatives. *Tutorial at the Neural Information Processing Systems conference*, **Denver, Colorado.**

Bruce A D and Saad D 1994 Statistical mechanics of hypothesis evaluation. *J. of Phys. A: Math. Gen.* **27:3355-3363.**

Buntine W L and Weigend A S 1991 Bayesian back-propagation. *Complex Systems* **5:603-643.**

Burman P 1989 A comparative study of ordinary cross validation and ν-fold cross validation, and the repeated learning testing methods. *Biometrika* **76:503-514.**

Craven P and Wahba G 1979 Smoothing noisy data with spline functions -Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31:377-403.**

Cybenko G 1989 Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2: 303-314.**

Dunmur A P and Wallace D J 1993 Learning and generalization in a linear perceptron stochastically trained with noisy data. *J. of Phys. A: Math. Gen.* **26:5767-5779.**

Eaton M 1983 *Multivariate Statistics - A Vector Space Approach.* **New York: Wiley.**

Engel A 1994 Uniform convergence bounds for learning from examples.

Hansen L K 1993 Stochastic linear learning: exact test and training error averages. *Neural Networks* **6:393-396.**

Haussler D, Kearns M, Seung H S and Tishby N 1994 Rigorous learning curve bounds from statistical mechanics. In *Proceedings of the Seventh Annual ACM Workshop on Computational Learning Theory (COLT '94)*:**76-87.**

Hertz J, Krogh A and Palmer R G 1991 *Introduction to the theory of neural computation. Santa Fe Institute lectures in the sciences of complexity* **Redwood City, California: Addisson-Wesley/Advanced book program.**

Hertz J, Krogh A and Thorbergsson G 1989 Phase transitions in simple learning. *J. of Phys. A: Math. Gen.* **22:2133-2150.**

Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **USA 79: 2554-2558. Reprinted in Anderson and Rosenfield (88).**

Hornik K, Stinchcombe M and White H 1989 Multilayer feedforward neural networks are universal approximators *Neural Networks* **2: 359-366.**

Jacobs R A 1995 Methods for combining expert's probability assessments. *Neural Computation* **7:867-888.**

John F 1978 *Partial Differential Equations.* **3rd edition, New York: Springer.**

Jordan M I and Jacobs R A 1994 Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6:181-214.**

Kearns M 1996 A bound on the error of cross-validation using the approximation and estimation rates, with consequences for the training-test split. *Advances in Neural Information Processing Systems* **8:183-189. Touretzky D S, Mozer M C and Hasselmo M E (Eds.). Cambridge, Mass: The MIT press.**

Kearns M, Mansour Y, Ng A Y, and Ron D 1995 An experimental and theoretical comparison of model selection methods. *preprint* **:**

Krogh A (1992) Learning with noise in a linear perceptron *J. Phys. A Math.*

Gen. **25:1119 - 1133.**

Krogh A and Hertz J 1992 Generalization in a linear perceptron in the presence of noise. *J. of Phys. A: Math. Gen.* **25:1135-1147.**

Krogh A and Vedelsby J 1995 Neural network ensembles, cross validation and active learning. In *Advances in neural information processing systems* **7:231-238 Tesauro G, Touretzky D S and Leen T K (Eds.). Cambridge, Massachusetts: The MIT Press.**

Levin E, Tishby N and Solla S A 1989 A statistical approach to learning and generalization in a layered neural network. *Proc. 2nd Workshop on Computational Learning Theory*: **245-260. San Mateo: Morgan Kauffmann.**

Lippmann R P 1987 Review of neural networks for speech recognition *Neural Computation* **1:1-38.**

MacKay D J C 1992a Bayesian interpolation. *Neural Comp.* **4:415-447.**

MacKay D J C 1992b A practical Bayesian framework for backprop networks. *Neural Comp.* **4:448-472.**

MacKay D J C 1993 Hyperparameters: optimise or integrate out? *Maximum entropy and Bayesian methods, Santa Barbara*, **G. Heidbreder, (Ed.), Dordrecht, 1994 Kluwer.**

Marion G and Saad D 1995a A statistical mechanical analysis of a Bayesian inference scheme for an unrealisable rule. *J. of Phys A: Math. Gen.* **28:2159-2171.**

Marion G and Saad D 1995b Finite size effects in Bayesian model selection and generalization. Submitted to *J. of Phys A: Math. Gen.*

Marion G and Saad D 1995c Hyperparameters, evidence and generalization for an unrealisable rule. In *Advances in Neural Information Processing Systems* **7:232-239. Tesauro G, Touretzky D S and Leen T K (Eds.). Cambridge, Massachusetts: The MIT Press.**

Marion G and Saad D 1995d Data dependent hyperparameter assignment. *Annals of Mathematics and Artificial Intelligence.* **In press.**

McCulloch W S and Pitts 1943 A logical calculus of ideas immanent in nervous activity*Bulletin of Mathematical Biophysics* **5:115-133. Reprinted in Anderson and Rosenfield (88).**

Meir R and Fontanari J F 1993 Data compression and prediction in neural networks *Physica A* **200:644-654.**

Meir R and Merhav N 1994 On the stochastic complexity of learning realizable and unrealizable rules. *preprint.*

Neal R M 1992 Bayesian training of backpropagation networks by the hybrid monte carlo method. *Technical Report* **CRG-TR-92-1, University of Toronto.**

Neal R M 1993 Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report* **CRG-TR-93-1, University of Toronto.**

Neal R M 1994 Priors for infinite networks. *Technical Report* **University of Toronto. Available by anonymous ftp from:** `ftp.cs.toronto.edu`, **file** `/pub/radford/pin.ps.Z`.

Pearlmutter B A 1989 Learning state space trajectories in recurrent neural networks. *Neural Computation* **1:263-269.**

Plutowski M, Sakata S and White H 1994 Cross - Validation estimates integrated mean squared error. *NIPS* **6 Cowan J D, Tesauro G and Alspector J (Eds.). San Mateo, CA: Morgan Kaufmann Publishers.**

Plutowski M 1994 *Selecting Training Exemplars for Neural Network Learning* **PhD thesis University of California, San Diego.**

Pomerleau D 1989 ALVINN: An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems* **1 Touretzky D S, (Ed.), San Mateo, CA: Morgan Kaufmann Publishers.**

Pryce J M and Bruce A D 1995 Statistical mechanics of image restoration. *J. of Phys. A: Math. Gen.* **28:511-532.**

Ripley B D 1992 Neural networks and related methods for classification. *J. Roy. Statist. Soc. Ser.B* **56:409-456.**

Rissanen J 1978 Modelling by shortest data description. *Automatica* **14:465-471.**

Rissanen J 1986 Stochastic Complexity and modelling. *The Annals of Statistics* **14:1080-1100.**

Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors. *Nature* **323:533-536.**

Saad D and Solla S A 1995a Exact solution for online learning in multilayer neural networks. *Physical Review Letters* **74:4337-4340.**

Saad D and Solla S A 1995b Online learning in soft committee machines. *Physical Review E* **52:4225-4243.**

Schwarze H 1993 Learning a rule in a multilayer neural network *J. Phys. A: Math. Gen.* **26: 5781-5794.**

Sejnowski T J and Rosenberg 1987 Parallel networks that learn to pronounce english text. *Complex Systems*, **1:145-168.**

Seung H S, Sompolinsky H, Tishby N 1992 Statistical mechanics of learning from examples. *Phys. Rev. A*, **45:6056-6091.**

Shao J 1993 Linear model selection by cross-validation. *J Amer. Statist. Assoc.* , **Vol. 88, No. 422:486-494.**

Skilling J 1993 *Physics and Probability.* **Grandy W T and Milloni P (Eds.) Cambridge: Cambridge University Press.**

Sollich P 1994a Finite-size effects in learning and generalization in linear perceptrons. *J. of Phys. A: Math. Gen.* **27:7771-7784.**

Sollich P 1994b Query construction, entropy, and generalization in neural network models. *Phys. Rev. E* **49:4637-4651.**

Sollich P 1995a Learning unrealizable tasks from minimum entropy queries. *J. of Phys. A: Math. Gen.* **28:6125-6142.**

Sollich P 1995b *Asking Intelligent Questions - The Statistical Mechanics of Query Learning.* **PhD thesis, University of Edinburgh, Scotland.**

Sollich P 1995c *Personal communication.*

Stone M 1974 Cross-Validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser.B* **36:111-147.**

Stone M 1977a An asymptotic equivalence of choice of model by cross-validation and Akaikes criterion. *J. Roy. Statist. Soc. Ser. B* **39:44-47.**

Stone M 1977b Asymptotics for and against cross-validation. *Biometrika* **64:29-35.**

Thodberg H H 1994 Bayesian backprop in action:pruning, ensembles, error bars and application to spectroscopy. *Advances in Neural Information Processing Systems* **6:208-215. Cowan J D, Tesauro G and Alspector J (Eds.). San Mateo, CA: Morgan Kauffmann Publishers.**

Valiant L G 1984 A theory of the learnable. *Communications of the ACM* **27:1134-1142.**

Vapnik V 1982 *Estimation of dependencies based on empirical data.* **Springer-Verlag: New York.**

Vapnik V and Chervonenkis A 1971 On the uniform convergence of relative frequencies of events to there probabilities. *Theory Probabil. Appl.* **16:264-280.**

Watkin T L H, Rau A, Biehl M 1993 The statistical mechanics of learning a rule. *Reviews of Modern Physics* **65:499-556.**

Wahba G 1985 A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statitics* **13:1378 - 1402.**

Wang C and Venkatesh S S 1995 Temporal dynamics of generalization in neural networks. In *Advances in Neural Information Processing Systems* **7:263-270 Tesauro G, Touretzky D S and Leen T K (Eds.). Cambridge, Massachusetts: The MIT Press.**

Williams P 1994 Bayesian regularisation and a Laplacian prior *University of Sussex, preprint.*

Williams R J and Zipser D 1989 A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* **1:270-280.**

Wolpert D H 1990 The relationship between Occam's razor and convergent guessing. *Complex Systems* **4:319-368.**

Wolpert D H 1992 On the connection between in sample testing and generalization error. *Complex Systems* **6:47-94.**

Wolpert D H 1993 On the Use of Evidence in Neural Networks. *Advances in Neural Information Processing Systems* **5:539-546. Hanson S J, Cowan J D and Giles C L (Eds.). San Mateo, CA: Morgan Kauffmann Publishers.**

Wolpert D H and Strauss C E M 1994 What Bayes has to say about the evidence procedure. To appear in *Maximum entropy and Bayesian methods*, **G. Heidbreder, (Ed.), Dordrecht, Kluwer.**