Finite Size Effects in Bayesian Model Selection and Generalisation

Glenn Marion † § and David Saad ‡ ¶

† Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, UK

‡ Dept. of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, UK

PACS numbers: 02.50.-r,05.20.-y,02.50.Rj,02.60.Ed,42.79.Ta

§ glenn@stams.strath.ac.uk

¶ D.Saad@helios.aston.ac.uk

Abstract.

We show that in supervised learning from a supplied data set Bayesian model selection, based on the evidence, does not optimise generalisation performance even for a learnable linear problem. This is demonstrated by examining the finite size effects in hyperparameter assignment from the evidence procedure and the resultant generalisation 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 hyperparameter 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.

1. Introduction

The problem of supervised learning, or learning from examples, has been much studied using the techniques of statistical physics (see *e.g.* Krogh and Hertz (92), Seung *et al* (92) and Watkin *et al* (93)). A major advantage of such studies over the usual analytical 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. 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 traditional 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 this limits the applicability of these theoretical results to the real world. In this paper we address the problem by calculating first order corrections to the thermodynamic limit, that is we explore finite size effects. Finite size effects in supervised learning have been studied previously by Sollich (94) and Barber *et al* (95). Before discussing the main focus of our study a brief introduction to the supervised learning paradigm is in order.

In this context one is presented with a set of data $\mathcal{D}= \{(y_t(\mathbf{x}_\mu), \mathbf{x}_\mu) : \mu = 1..p\}$ consisting of p example pairs of an otherwise unknown *teacher* mapping denoted by the distribution $P(y_t \mid \mathbf{x})$. This notation accommodates, for example, teachers with deterministic outputs corrupted by noise. Furthermore, we assume that the N_I 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 $\mathcal{D}$ to set the N parameters, $\mathbf{w}$, of some model (or student), with output $y_s(\mathbf{x})$, such that it *learns* to mimic the underlying mapping as closely as possible on all inputs drawn from the distribution $P(\mathbf{x})$ (*i.e.* not simply those in the training set $\mathcal{D}$). A popular measure of this performance is the *generalisation error* which we define formally in section 2.2. We regard minimisation of this error, to which one does not have direct access, as the principal goal of the learning or training process. The question is then how to conduct training so as to obtain the best possible performance. One frequently used approach 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. Provided γ is non-zero this serves to alleviate the problem of *over-fitting* of noisy data which can degrade performance. It is the setting of the, so-called, hyperparameters β and γ which we will examine in this presentation.

If stochastic gradient descent is used to minimise the composite cost function, $\beta E_{\mathbf{w}}(\mathcal{D}) + \gamma C(\mathbf{w})$, one obtains a Gibbs distribution of students, (*i.e.* the post training distribution over the parameters $\mathbf{w}$) (Seung *et al* 92). If we wish to make a prediction on a novel input using the average, or the maximum, of this distribution then this prediction depends solely on the hyperparameters. Thus, the selection of β and γ can be regarded as a model selection. In practice, since a decision must be based only on the training data there are essentially two choices in terms of hyperparameter assignment . Firstly one can attempt to estimate the generalisation error (*e.g.* by cross-validation (Stone 74)) and then optimise this measure with respect to the hyperparameters. However, such an approach can be computationally expensive. Secondly, one can optimise some other measure and hope that the resulting assignments produce low generalisation error. In particular, MacKay (92) advocates a quantity derived from Bayesian statistics, termed the *evidence*, as such a measure. In the main we will explore this latter approach, defining the evidence in section 2.1.

Model selection based on the evidence, in the *learnable* case of a linear student and teacher, has been studied by Bruce and Saad (94) in the thermodynamic limit. Their results show that optimising the average, over all possible data sets, of the log evidence simultaneously with respect to both hyperparameters optimises the average generalisation error. In an *unlearnable* scenario Marion and Saad (95) show that in the thermodynamic limit hyperparameter assignment from the average log evidence does not optimise performance. *Self averaging* is said to hold if the variance of relevant quantities vanishes as the thermodynamic limit is approached. Since both these studies were conducted in the thermodynamic limit and the self averaging property was assumed the analyses were average case. In this paper we show that self averaging does indeed hold in relation to model selection based on the evidence in the learnable linear case. However, we will explore the optimality of the evidence in a system of finite size where the variance over data sets is non-vanishing. Furthermore, rather than conduct an average case analysis we seek to examine hyperparameter assignment *based on individual data sets*.

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 hyperparameter assignments which minimise the generalisation 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 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.

The remainder of the paper is organised as follows. In the next section we review the evidence framework and the performance measures we will deal with. In section 3, we write down the evidence and the performance measures for the learnable linear case. The problem of consistency, that is the behaviour in the limit of large amounts of data, is then explored along with an average case approach. In addition, employing some of the results of Sollich (94), we demonstrate that, for large N, the variances, over data sets, of the evidence and generalisation error are $O(1/N)$, in other words that self averaging holds. In section 4 we avoid the average case approach examining hyperparameter assignment from the evidence in relation to the optimal hyperparameters using finite size corrections to the thermodynamic limit. We corroborate these results with numerical simulations of small systems. The impact of these assignments on performance is studied in section 5. In particular 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. A numerical study of a low dimensional system in section 6 allows a comparison of model selection based on the cross-validation error and on the evidence. Finally we summarise our main results in section 7.

2. Objective functions

2.1. The evidence

Since $E_{\mathbf{w}}(\mathcal{D})$ is the sum squared error then, 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 $P(\mathcal{D} \mid \beta, \mathbf{w}) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D})}$. 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 post training or student distribution, $P(\mathbf{w} \mid \mathcal{D}, \gamma, \beta) \propto e^{-\beta E_{\mathbf{w}}(\mathcal{D}) - \gamma C(\mathbf{w})}$. As noted earlier stochastic minimisation of the composite cost function also gives rise to this distribution. Indeed, Buntine and Weigend (91) refer to this process as *Bayesian Backpropagation.*

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

$$P(\mathcal{D} \mid \gamma, \beta) = \int \prod_j dw_j P(\mathcal{D} \mid \beta, \mathbf{w}) P(\mathbf{w} \mid \gamma) \quad . \tag{2.1}$$

That is, the probability of (or evidence for) the data set ($\mathcal{D}$) given the hyperparameters β and γ. The evidence can thus, be calculated from the data set, $\mathcal{D}$, alone. Throughout this paper we refer to the *evidence procedure* as the process of fixing the hyperparameters to the values that simultaneously maximise the evidence for a given data set. Thus, although the Bayesian framework outlined here envisages the hyperparameters as defining the whole distribution of input-output pairs, the assignments from the evidence procedure will depend on the data set at hand. Indeed, 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 note that this is the way in which the evidence procedure is used in practice (Mackay (92)).

2.2. *The performance measures*

In contrast to the evidence, the performance measures we review here can not be calculated from the data alone. Before proceeding we will introduce the notation $\langle f(z)\rangle_{P(z)}$ to denote the average of the quantity $f(z)$ over the distribution $P(z)$. However, we will use the short hand $\langle . \rangle_{\mathbf{w}}$ to mean the average over the post training distribution $P(\mathbf{w} \mid \mathcal{D}, \gamma, \beta)$. Thus, the average student output at $\mathbf{x}$ conditioned on the training data, $\mathcal{D}$, is $\langle y_s(\mathbf{x})\rangle_{\mathbf{w}}$.

As the principal performance measure we choose the expected squared difference over the input distribution $P(\mathbf{x})$ between the average student and the average teacher. That is, the data dependent generalisation error

$$\epsilon_g(\mathcal{D}) = \langle (\langle y_t(\mathbf{x})\rangle_{P(y_t|\mathbf{x})} - \langle y_s(\mathbf{x})\rangle_{\mathbf{w}})^2 \rangle_{P(\mathbf{x})} \quad . \tag{2.2}$$

If we were to *average over all possible data sets* of fixed size then this would correspond to the generalisation error studied by Bruce and Saad (94) and Krogh and Hertz (92). 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 generalisation error and our focus will be on the performance of the evidence procedure in relation to this. However, in section 3.1 we will consider an average case approach. Further, in section 5 we will also consider the effect of defining the optimal hyperparameter 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.

Another feature we can consider is 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

$$\delta_c(\mathcal{D}) = \langle \{y_s(\mathbf{x}) - \langle y_s(\mathbf{x})\rangle_{\mathbf{w}}\}^2\rangle_{\mathbf{w},P(\mathbf{x})} - \epsilon_g(\mathcal{D}) \quad . \tag{2.3}$$

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 a 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.

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 paper the evidence procedure is unbiased in a particular sense. We then explore the approach to the thermodynamic limit demonstrating that the system is self averaging. However, initialy we must calculate the evidence and the performance measures.

Since the student is linear with output $y(\mathbf{x}) = \mathbf{w}.\mathbf{x}/\sqrt{N}$, the number of parameters equals the dimension of input space, $N_I = N$. 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 regularisation 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 $f(\mathcal{D}) = -1/N \ln P(\mathcal{D} \mid \lambda, \beta)$, where we have introduced the weight decay parameter $\lambda = \gamma/(\beta\sigma_x^2)$. We can write the quantity $f(\mathcal{D})$ which is analogous to a free energy as

$$f(\mathcal{D}) = -\frac{1}{2}\ln\frac{\lambda}{\pi} - \frac{\alpha}{2}\ln\frac{\beta}{\pi} + \frac{1}{2}\ln 2 - \frac{1}{2N}\ln\det\,\mathbf{g} + \mathbf{y}.\mathbf{n} + \beta\left(\lambda\sigma_x^2\sigma_w^2 + \mathbf{n}^T\Gamma\mathbf{n} + a_{ev}\right) \quad (3.1)$$

where

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

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

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

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 from $\mathcal{N}(0,\sigma)$. The term a_{ev} does not fluctuate with the noise but only with the inputs $\mathbf{x}^\mu$.

The generalisation error and the consistency can be calculated from $f(\mathcal{D})$ by averaging appropriate expressions over the input distribution $P(\mathbf{x})$. The generalisation error is given by

$$\epsilon_g(\mathcal{D}) = \mathbf{n}^T\Delta\mathbf{n} + \mathbf{z}.\mathbf{n} + a_{\epsilon_g} \quad (3.2)$$

† Where $\mathcal{N}(\bar{\mathbf{x}},\sigma)$ denotes a normal distribution with mean $\bar{\mathbf{x}}$ and variance σ^2.

where

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

$$\text{and} \qquad a_{\epsilon_g} = -\frac{\lambda^2\sigma_x^2}{N}(\mathbf{w^o})^T \frac{\partial \mathbf{g}}{\partial \lambda}\mathbf{w^o} \quad .$$

Finally, the consistency is

$$\delta_c(\mathcal{D}) = \frac{1}{2\beta N}\operatorname{tr}\mathbf{g} - \epsilon_g(\mathcal{D}) \quad . \tag{3.3}$$

We note here that the generalisation error depends only on the weight decay, λ, thus in the remainder of this paper we refer to the optimal weight decay $\lambda_{opt}(\mathcal{D})$ as that which minimises $\epsilon_g(\mathcal{D})$. 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 generalisation error. We denote the hyperparameters that simultaneously maximise the evidence as $\lambda_{ev}(\mathcal{D})$ and $\beta_{ev}(\mathcal{D})$. Thus, the term *optimal* refers to the optimisation of, or with respect to, the performance measures whilst *evidence optimal* refers to maximisation of the evidence.

3.1. Consistency and unbiasedness

Firstly we consider the question of asymptotic consistency, that is, we examine the free energy, $f(\mathcal{D})$, and the generalisation error in the limit of large amounts of data (*i.e.* as $p \to \infty$ with N fixed). This term is not to be confused with the consistency measure defined above. Using the fact, shown in appendix A, that, for large p, $g_{ij} = \delta_{ij}N/p + \mathrm{O}(1/p^{3/2})$ we can find the asymptotic evidence optimal hyperparameter assignments, namely

$$\lim_{p\to\infty} \lambda_{ev}(\mathcal{D}) = \lambda_0 + \mathrm{O}\left(\frac{1}{\sqrt{p}}\right) \quad \text{and} \quad \lim_{p\to\infty} \beta_{ev}(\mathcal{D}) = \beta_0 + \mathrm{O}\left(\frac{1}{\sqrt{p}}\right) \tag{3.4}$$

where the noise to signal ratio $\lambda_0 = \sigma^2/(\sigma_x^2\sigma_w^2)$ and $\beta_0 = 1/(2\sigma^2)$. In addition it can be shown that, to first order in p^{-1}, the generalisation error is independent of λ. As we shall see later in the context of large N this insensitivity of the generalisation 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 generalisation error is independent of the weight decay for large p implies that *any* scheme for setting λ, and in particular the evidence assignments, will achieve optimal performance asymptotically (*i.e.* , generalisation error tends to zero irrespective of λ). However, as we shall see in section 4 this does not imply that the evidence assignments correspond to the optimal hyperparameters. 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 generalisation error when assessing the performance of the evidence assignments we will focus on the fractional increase in generalisation error from the optimal incurred by their use. That is on

$$\kappa_{\epsilon_g}(\lambda_{ev}, \mathcal{D}) \equiv \frac{\epsilon_g(\lambda_{ev}, \mathcal{D}) - \epsilon_g(\lambda_{opt}, \mathcal{D})}{\epsilon_g(\lambda_{opt}, \mathcal{D})} \quad . \tag{3.5}$$

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

$$\kappa_{\delta_c}(\lambda_{ev}, \beta_{ev}, \mathcal{D}) \equiv \frac{\delta_c(\lambda_{ev}, \beta_{ev}, \mathcal{D})}{\epsilon_g(\lambda_{ev}, \mathcal{D})} \quad . \tag{3.6}$$

In section 5 we examine the behaviour of both $\kappa_{\epsilon_g}(\mathcal{D})$ and $\kappa_{\delta_c}(\mathcal{D})$ in the thermodynamic limit.

However, before considering this regime we examine average case behaviour. Using the result of appendix B it can be shown that

$$\langle \epsilon_g(\mathcal{D}) \rangle_{P(\mathcal{D})} = \sigma^2 G_{\rm av} + \lambda \partial_\lambda G_{\rm av}(\sigma^2 - \lambda \sigma_x^2 \sigma_w^2) \quad , \tag{3.7}$$

where the response function $G_{\rm av} = \langle \mathrm{tr}\, \mathbf{g} \rangle_{P(\mathcal{D})}$ is unknown in general. The average generalisation 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 average case analysis shows that the evidence procedure is unbiased in the sense that maximisation of the average evidence optimises average performance. However, we now show that the fluctuations around this average optimum performance become increasingly important as the system size, N, decreases.

3.2. Self averaging

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

$$\begin{aligned} \mathrm{Var}(f(\mathcal{D})) = {} & 2\sigma^4 \langle \mathrm{tr}(\Gamma\Gamma) \rangle_{P(\{\mathbf{x}^\mu : \mu=1..p\})} + \sigma^2 \langle \mathrm{tr}(\mathbf{y}^T \mathbf{y}) \rangle_{P(\{\mathbf{x}^\mu : \mu=1..p\})} \\ & + \beta^2 \langle a_{ev}^2 \rangle_{P(\{\mathbf{x}^\mu : \mu=1..p\})} - \beta^2 \langle a_{ev} \rangle^2_{P(\{\mathbf{x}^\mu : \mu=1..p\})} \quad . \end{aligned} \tag{3.8}$$

† Alternatively one can show this result using diagrammatic methods.

Here we have explicitly performed the noise average and the remaining average over the input points is with respect to $P(\{\mathbf{x}^\mu : \mu = 1..p\})$. As shown in appendix C, 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 $\mathrm{O}(1/N)$ as we approach the thermodynamic limit. Thus, the variance of the free energy is $\mathrm{O}(1/N)$, *i.e.* it is self averaging. Similarly, it can be shown that the generalisation error and consistency measure are also self averaging. This means that in the thermodynamic limit the behaviour exhibited by the system for any particular 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 for *any particular data set* because their results were obtained in the thermodynamic limit.

4. Data dependent hyperparameter assignment

Having now established, in addition to the self averaging, that the evidence procedure is unbiased and consistent in a crude sense we now wish to examine the finite system behaviour for data sets of finite size. This is clearly the regime of interest to *real world* applications since one is then in the business of optimising performance based on the supplied data set. To obtain the hyperparameter assignments made by the evidence procedure we must simultaneously solve $\partial_\lambda f(\mathcal{D}) = 0$ and $\partial_\beta f(\mathcal{D}) = 0$, where $\partial_\theta f \equiv \partial f/\partial\theta$. We can linearise these equations, close to the thermodynamic limit, by expanding around $\lambda = \lambda_0$ and $\beta = \beta_0$. Doing so we obtain

$$\begin{pmatrix} \Delta\lambda_{ev} \\ \Delta\beta_{ev} \end{pmatrix} = \begin{pmatrix} \partial^2_\lambda f & \partial_\beta\partial_\lambda f \\ \partial_\lambda\partial_\beta f & \partial^2_\beta f \end{pmatrix}^{-1} \begin{pmatrix} \partial_\lambda f \\ \partial_\beta f \end{pmatrix}_{\lambda_0,\beta_0} . \tag{4.1}$$

Where the evidence optimal hyperparameters are $\lambda_{ev}(\mathcal{D}) \approx \lambda_0 + \Delta\lambda_{ev}(\mathcal{D})$ and $\beta_{ev}(\mathcal{D}) \approx \beta_0 + \Delta\beta_{ev}(\mathcal{D})$. In the notation adopted here the data dependence is implicit and the right hand side is evaluated at $\lambda = \lambda_0$ and $\beta = \beta_0$.

Similarly, we can expand the true optimal hyperparameters about the thermodynamic limit, obtaining $\lambda_{opt}(\mathcal{D}) \approx \lambda_0 + \Delta\lambda_{opt}(\mathcal{D})$ from the generalisation error with

$$\Delta\lambda_{opt} = \left(-\frac{\partial_\lambda \epsilon_g}{\partial^2_\lambda \epsilon_g}\right)_{\lambda_0,\beta_0} . \tag{4.2}$$

Since we regard the optimal consistency as zero (see Section 2.2) we obtain $\beta_{opt}(\mathcal{D}) \approx \beta_0 + \Delta\beta_{opt}(\mathcal{D})$ where

$$\Delta\beta_{opt}(\mathcal{D}) = \frac{(\epsilon_g(\mathcal{D}) - (\epsilon_g(\mathcal{D}))_0)\,\mathrm{tr}\,\mathbf{g}}{2N(\epsilon_g(\mathcal{D}))_0^2} , \tag{4.3}$$

and the notation $(h)_0$ denotes the value of the function h in the thermodynamic limit.

The (co)-variances of these quantities are $\mathrm{O}(1/N)$; an example calculation is outlined in appendix D. Figure 1 shows, to first order in N, the scaled variances † in the evidence optimal weight decay, $\widetilde{\mathrm{Var}}(\lambda_{ev})$ and that in the true optimal weight decay, $\widetilde{\mathrm{Var}}(\lambda_{opt})$ for various values of λ_0 . In the limit of large α we find

$$Var(\lambda_{ev}) \approx \frac{2\lambda_0^2}{\alpha N}(1+2\lambda_0) \text{ and } \quad Var(\lambda_{opt}) \approx \frac{\lambda_0 \alpha}{N} \quad . \tag{4.4}$$

The asymptotic $\mathrm{O}(1/\alpha)$ decay of the former reflects the fact that, as discussed in section 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 generalisation error to the weight decay for large α. The divergence of both curves for small α is order $\mathrm{O}(1/(N\alpha))$ and in fact, for $p = 1$ it can be shown analytically that these quantities are $\mathrm{O}(1)$. 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 is zero in the limit of no noise. Thus, if there is no noise the evidence procedure can only set the weight decay with confidence for $\alpha > 1$, whilst the optimal policy is to accept the data completely for all α (*i.e.* $\lambda_0 = 0$).

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

$$\| \lambda_{ev} - \lambda_{opt} \|^2 \equiv < (\lambda_{ev}(\mathcal{D}) - \lambda_{opt}(\mathcal{D}))^2 >_{P(\mathcal{D})} \quad . \tag{4.5}$$

As one would expect this average separation increases with the noise. However, in the limit of zero noise whilst $\| \lambda_{ev} - \lambda_{opt} \|^2$ 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 the average distance between the optimal weight decay and the evidence assignment diverges linearly, indeed for large α we find that

$$\| \lambda_{ev} - \lambda_{opt} \|^2 \approx \mathrm{Var}(\lambda_{opt}) \quad . \tag{4.6}$$

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 regularisation parameter diverges.

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 $\mathrm{O}(1)$ as shown in figure 2. The normalised correlation between two fluctuating quantities $h(\mathcal{D})$ and $k(\mathcal{D})$ is written $C(h(\mathcal{D}), k(\mathcal{D})) = (\langle hk\rangle_{P(\mathcal{D})} - \langle h\rangle_{P(\mathcal{D})}\langle k\rangle_{P(\mathcal{D})})/(\mathrm{Var}(h)\mathrm{Var}(k))^{1/2}$. For

† *i.e.* N times the true variances

small α the non-monotonic behaviour of $C(\lambda_{ev}, \lambda_{opt})$ 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 optimal parameter does not fluctuate in this 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 variance 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.

Finally, in the large α limit we have

$$\lim_{\alpha\to\infty} C(\lambda_{ev}, \lambda_{opt}) = -\frac{\sqrt{2\lambda_0}}{\sqrt{2\lambda_0 + 1}} \quad , \tag{4.7}$$

and

$$\lim_{\alpha\to\infty} C(\beta_{ev}, \beta_{opt}) \approx 4\lambda_0^2 \alpha^{-7/2} \quad . \tag{4.8}$$

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 $O(1/N)$ the evidence assignments correlate rather poorly with the optimal assignments.

When defining the evidence procedure, we could have chosen to optimise the evidence with respect to each of the hyperparameters whilst holding the other fixed rather than 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 hyperparameters (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 hyperparameters.

To summarise, we note that our results in this section are in stark contrast to the average case result of section 3.1 and reveal the inadequacies of the latter approach. In addition, despite mean square consistency the evidence assignments are in fact far from the optimal values both asymptotically and for finite α. Indeed, in section 5 we will see that this has a deleterious effect on performance.

4.1. Simulations

To qualitatively corroborate our results we performed simulations of one dimensional linear perceptron students and teachers. In these simulations we generated random data sets and found the evidence procedure and the true optimal hyperparameter assignments. Then by averaging over many such data sets we calculated the variances and correlations of these parameter assignments. Some results from these simulations are shown in figure 3. 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 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 3 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, over different data sets, of the evidence and the optimal assignments. For a small number of examples, p, the distribution of evidence assignments looks qualitatively the same as that of the optimal assignments. Thus, there are many occasions where λ_{ev} and λ_{opt} are coincident and the correlation between them is positive although as we can see in figure 3 the variances in the assignments are large. As p grows the evidence assignments begin to cluster around λ_0 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. However, as p grows the distribution of the optimal assignment remains similar to its small p form but the variance in λ_{opt} becomes larger also in accord with our theoretical results. Given the differences between these two distributions it is hardly surprising that the correlation between the two corresponding hyperparameter assignments is not positive in this region.

5. Effects on performance

We now examine the effects on performance of these sub-optimal hyperparameter assignments. Firstly, for the generalisation error to order $\mathrm{O}(1/\sqrt{N})$ 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 $\mathrm{O}(1/N)$ they differ, thus we can write the correlation between them, somewhat suggestively, as $1 - \mathrm{O}(1/N)$. Unfortunately, we are unable to calculate this correlation to $\mathrm{O}(1/N)$. Therefore, we examine the increase in error invoked by use of the evidence procedure

$$\Delta\epsilon(\mathcal{D}) \equiv \epsilon_g(\lambda_{ev}, \mathcal{D}) - \epsilon_g(\lambda_{opt}, \mathcal{D})$$

$$= \Delta\lambda_{ev}\partial_\lambda\epsilon_g + \frac{1}{2}\Delta\lambda_{ev}^2\partial_\lambda^2\epsilon_g + \frac{1}{2}\Delta\lambda_{opt}^2\partial_\lambda^2\epsilon_g + \mathrm{O}\left(\frac{1}{N^2}\right) \quad , \tag{5.1}$$

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 4.5. Thus, we find that

$$< \Delta\epsilon(\mathcal{D}) >_{P(\mathcal{D})} = \frac{1}{2}(\partial_\lambda^2\epsilon_g)_0 \parallel \lambda_{ev} - \lambda_{opt} \parallel^2 + \mathrm{O}\left(\frac{1}{N^2}\right) \quad . \tag{5.2}$$

Whilst, the calculation of this average is then straight forward that of the variance is more tricky. The variance is $\mathrm{O}(1/N^2)$ and thus we would have to calculate the variance of the response function $\mathrm{tr}\,\mathbf{g}/N$ 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 $\mathrm{O}(1/N)$, a lower bound on the *typical* error invoked by use of the evidence procedure is the average degradation of equation 5.2 plus the square root of its variance over the noise.

In figure 4, 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 generalisation error. This quantity which is a scaled estimate of the fractional degradation defined in equation 3.5 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 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 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 will range between 0.01 and 0.29 and for a larger sized system the evidence procedure results in closer to optimal behaviour. In the large α limit we find that, for the *average* fractional degradation

$$\lim_{\alpha\to\infty} < \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})} = \frac{1}{N} + \frac{2(\lambda_0 + 1)}{N\alpha} + \mathrm{O}\left(\frac{1}{N\alpha^2}\right) \quad . \tag{5.3}$$

Note that the average relative degradation, $< \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})}$, does not decay with α despite the fact that the average degradation in performance, $< \Delta\epsilon(\mathcal{D}) >_{P(\mathcal{D})}$, is itself $\mathrm{O}(1/\alpha N)$. 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) 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 $O(1/\alpha N)$. Indeed, the fluctuations do not, in general, qualitatively change the behaviour of the average fractional error, $\langle \kappa_{\epsilon_g}(\lambda_{ev}) \rangle_{P(\mathcal{D})}$ and the relative size of the fluctuation term as a fraction of the typical error is most important for a mid range $\alpha \approx 2$.

As the noise level increases so does $< \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})}$ which is a reflection of the increasing uncertainty in λ_{ev} as shown in the right-hand graph of figure 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 $\alpha > 1$ we find

$$\lim_{\lambda_0 \to 0} < \kappa_{\epsilon_g}(\lambda_{ev}) >_{P(\mathcal{D})} = \frac{1}{N}\frac{\alpha+1}{\alpha-1} \quad , \tag{5.4}$$

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 phase transition found in the thermodynamic limit by Bruce and Saad (94). 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 average 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 assignment is optimal. 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 5.4, 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 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* generalisation error. As we saw in section 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' generalisation error, $\langle \epsilon_g(\lambda_0, \mathcal{D}) \rangle_{P(\mathcal{D})}$, is to first order in α^{-1} 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 5.3. In contrast, for this new optimal, at small α we find qualitatively similar behaviour in the fractional degradation to that displayed in figure 4. 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 generalisation 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 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 generalisation error gives an estimate of the fractional error, κ_{δ_c} defined in equation 3.6. To first order this quantity, scaled by $\sqrt{N}$ and denoted by $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$, is plotted in figure 5. In general, $\widetilde{\kappa}_{\delta_c}^{\mathbf{typ}}$ is much larger than $\widetilde{\kappa}_{\epsilon_g}^{\mathbf{typ}}$. 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 generalisation error becomes larger and we are able to estimate it, using the consistency criterion, to a greater degree of accuracy when it is larger.

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 (see *e.g.* 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 generalisation error to the optimal. The results, averaged over 1000 realisations of the data set for each value of p, are plotted in figure 6. These results corroborate the results of the previous section in that they show the evidence procedure to be sub-optimal. Further, they also reveal that cross-validation produces closer to optimal performance. The left-hand graph in figure 6 shows that the resulting error from the cross-validatory estimate correlates more strongly with the optimal generalisation error than does that resulting from the evidence estimate.

In addition, the right-hand graph shows that the average fractional increase in the generalisation error, $\kappa_{\epsilon_g}(\lambda)$, is considerably larger for the evidence procedure than for cross-validation.

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 hyperparameter 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 generalisation performance reflected this sub-optimality. 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. In addition, our numerical studies indicate that for small learnable linear systems leave one out 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 generalisation 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 future work we hope to explore the finite size effects associated with the cross validatory procedure and to compare these analytic results with those obtained here for the evidence procedure. We also note the average case results for discrete mappings obtained by Meir and Merhav (94) on the consistency of hyperparameter assignment via minimisation of the stochastic complexity for a realizable case. Given our results and the analogy between the evidence and the stochastic complexity it would also be interesting to examine finite size effects in model selection based on this quantity.

Acknowledgments

We are very grateful to Alastair Bruce and Peter Sollich for useful discussions. Many thanks are due to the Department of Physics at the University of Edinburgh where much of this work was conducted. GM would also like to acknowledge the support provided by an E.P.S.R.C. studentship.

References

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

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

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

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

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

Graybill F A 1983 *Matrices with Applications in Statistics* Belmont California: Wadsworth Statistics/Probability series)

Gelfand A E and Dey D K 1994 Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser.B* **36**:501-514.

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

Krogh A and Hertz J 1992 Generalisation 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.

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

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

Meir R and Merhav N 1994 On the stochastic complexity of learning realizable and unrealizable rules. *preprint.*

Plutowski M, Sakata S and White H 1994 Cross-validation estimates integrated mean squared error. In *Advances in neural information processing systems* **6**: Cowan *et al* , (Eds.), Morgan Kaufmann Publishers, San Mateo, CA

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.* .**88/422**:486-494.

Sollich P 1994 Finite-size effects in learning and generalisation in linear perceptrons. *J. of Phys. A: Math. Gen.* **27**:7771-7784

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 Akaike's criterion *J. Roy. Statist. Soc. Ser. B* **39**:44-47

Stone M 1977b Asymptotics for and against cross-validation. *Biometrika* **64**:29-35.

Watkin T L H, Rau A, Biehl M 1993 The statistical mechanics of learning a rule. *Reviews of Modern Physics* **65**:499-556.

Appendix A

Here we calculate the matrix $\mathbf{g}$ in the large p limit using the following result for the inverse of the patterned matrix $\mathsf{C} = (a-b)\mathsf{I} + b\mathsf{J}$ (Graybill 83)

$$\mathsf{C}^{-1} = \frac{1}{a-b}\left(\mathsf{I} - \frac{b}{a+(k-1)b}\mathsf{J}\right) \quad , \tag{A.1}$$

where J is the square matrix with all its entries 1. Now in the large p limit, using the central limit theorem we can write

$$\mathbf{g}^{-1} \approx \left(\frac{p}{N} - \sqrt{p}\sigma_x^2\right)\mathsf{I} + \sqrt{p}\sigma_x^2\mathsf{J} \quad , \tag{A.2}$$

where the contribution from the $\lambda\mathsf{I}$ term is negligible. Thus we can write,

$$\mathbf{g}_{ii} \approx \frac{N}{p} + \mathrm{O}\left(\frac{1}{p^{3/2}}\right) \qquad \text{and} \qquad \mathbf{g}_{ij} \approx \mathrm{O}\left(\frac{1}{p^{3/2}}\right) \quad i \neq j \quad . \tag{A.3}$$

This result is in agreement with that for the inverse of the correlation matrix $\mathbf{A}$ which has an inverse Wishart distribution. In the large p limit this also has a fluctuation of $\mathrm{O}(1/p^{3/2})$ around a mean of N/p (for example see Eaton (83)).

Appendix B

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

$$g_{ij} = \lambda^{-1} - \lambda^{-2}A_{ij} + \lambda^{-3}A_{ik}A_{kj}.... \tag{B.1}$$

$$\text{where} \quad A_{ij} = \frac{1}{N\sigma_x^2}\Sigma_{\mu=1}^{p} x_i^{\mu} x_j^{\mu} \quad .$$

A typical term is then

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

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.

Appendix C

In this appendix we show that quantities in equation 3.9 are $\mathrm{O}(1/N)$. Firstly, $\mathrm{tr}\,\Gamma\Gamma$

$$\mathrm{tr}\,\Gamma\Gamma = \left(\frac{(\mathbf{x}_\mu)^T\mathbf{g}\mathbf{x}_\nu}{N^2\sigma_x^2} + \frac{\delta_{\mu\nu}}{N}\right)\left(\frac{(\mathbf{x}_\nu)^T\mathbf{g}\mathbf{x}_\mu}{N^2\sigma_x^2} + \frac{\delta_{\nu\mu}}{N}\right) \quad , \tag{C.1}$$

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

$$\langle \mathrm{tr}\, \Gamma\Gamma \rangle_{P(\{\mathbf{x}^\mu:\mu=1..p\})} = \frac{1}{N}\left(\alpha - 1 + \lambda^2 \partial_\lambda G\right) \quad . \tag{C.2}$$

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

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

$$\mathrm{Var}(a_{ev}) = \sigma_x^4 \lambda^4 Var\left(\frac{1}{N}(\mathbf{w^o})^T \mathbf{g} \mathbf{w^o}\right) \quad . \tag{C.3}$$

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

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

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

Appendix D

Here, as an example we calculate the correlation between λ_{ev} and λ_{opt}. From equation 4.1 we find

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

where we have defined

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

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 $\mathrm{O}(1/\sqrt{N})$. However, the second derivatives do not vanish at this point and so $\partial_\beta^2 f$ and $\partial_\beta \partial_\lambda f$ are $\mathrm{O}(1)$. Thus, expanding up to first order we obtain

$$\Delta\lambda_{ev} = -\frac{1}{(\det \mathsf{M})_0}\{(\partial_\beta^2 f)_0 \; \partial_\lambda f \quad - \quad (\partial_\beta \partial_\lambda f)_0 \; \partial_\beta f\}_{\lambda_0,\beta_0} + \mathrm{O}\left(\frac{1}{N}\right) \quad . \tag{D.3}$$

Similarly, from equation 4.2, we can write

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

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

$$\begin{aligned} \langle \lambda_{opt}\lambda_{ev}\rangle_{P(\mathcal{D})} = & -\frac{1}{(\det \mathsf{M})_0(\partial^2_\lambda \epsilon_g)_0}\{(\partial^2_\beta f)_0 \ \langle \partial_\lambda f \ \partial_\lambda \epsilon_g\rangle_{P(\mathcal{D})} \\ & - (\partial_\beta \partial_\lambda f)_0 \ \langle \partial_\beta f \ \partial_\lambda \epsilon_g\rangle_{P(\mathcal{D})}\}_{\lambda_0,\beta_0} + \mathrm{O}\left(\frac{1}{N^{3/2}}\right) \quad . \end{aligned} \tag{D.5}$$

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

$$\begin{aligned} \langle \partial_\lambda f \ \partial_\lambda \epsilon_g\rangle_{P(\mathcal{D})} = & \ \mathrm{Cov}(\mathbf{n}^T\Gamma^{'}\mathbf{n}, \mathbf{n}^T\boldsymbol{\Delta}^{'}\mathbf{n}) \\ & + \mathrm{Cov}(\mathbf{n}.\mathbf{y}^{'}, \mathbf{n}.\mathbf{z}^{'}) + \beta_0 \mathrm{Cov}(a^{'}_{ev}, a^{'}_{\epsilon_g}) + \mathrm{O}\left(\frac{1}{N}\right) \quad . \end{aligned} \tag{D.6}$$

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 3.1 and 3.2. Equation D.6 can then expressed in terms of the response function as we saw in Appendix C. The second term, $\langle \partial_\beta f \ \partial_\lambda \epsilon_g\rangle_{P(\mathcal{D})}$, is similar.

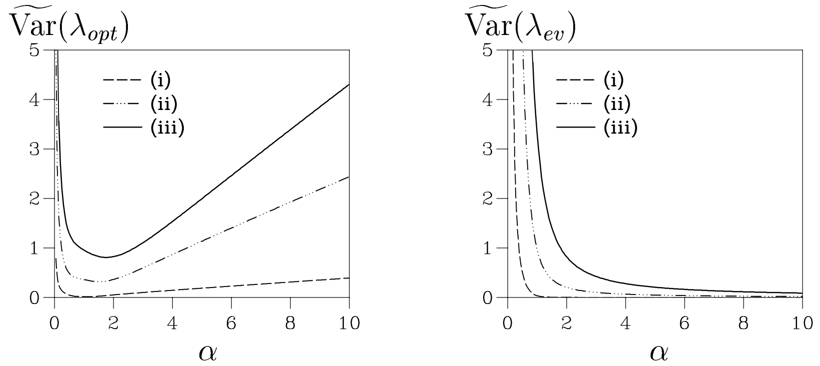

Figure 1. The scaled variance in the optimal weight decay, $\widetilde{\mathrm{Var}}(\lambda_{opt})$, for various noise levels, (i) $\lambda_0 = 0.04$, (ii) $\lambda_0 = 0.25$ and (iii) $\lambda_0 = 0.44$ is shown in the left-hand graph. Notice the linear divergence in α which corresponds to our result in section 3.1 that, for sufficiently large p, the generalisation 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 $\mathrm{O}(1/\alpha)$ decay of this quantity is a reflection of the fact that for large p the evidence optimal weight decay $\lambda_{ev}(\mathcal{D}) = \lambda_0$.

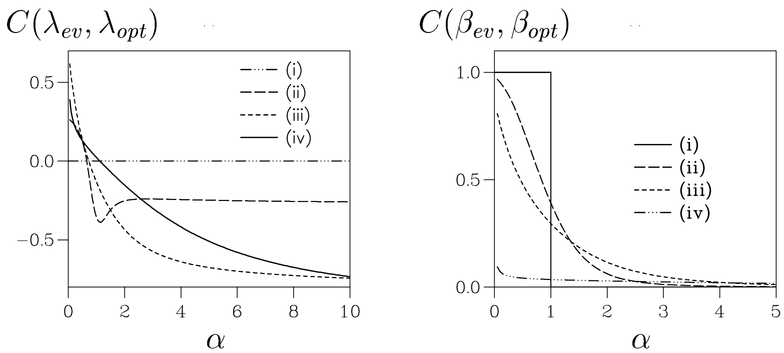

Figure 2. The correlation between the optimal weight decay and the evidence optimal weight decay $C(\lambda_{ev}, \lambda_{opt})$ is shown, in the left-hand graph, for (i) $\lambda_0 \to 0.0$, (ii) $\lambda_0 = 0.01$, (iii) $\lambda_0 = 1$ and (iv) $\lambda_0 = 4$. The right-hand graph shows the correlation 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) $\lambda_0 = 16$.

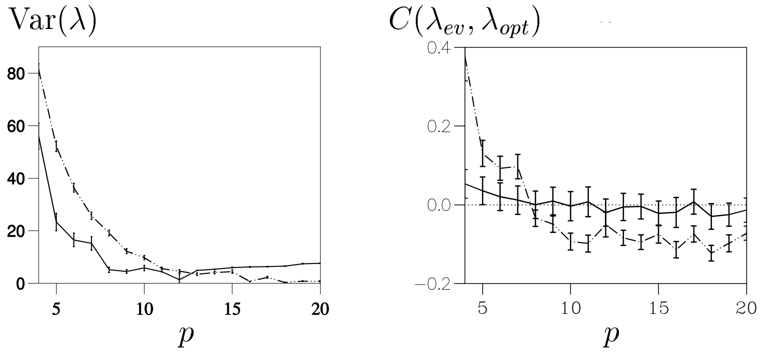

Figure 3. 1-D simulation results: The left-hand graph shows the variance in the optimal weight decay λ_{opt} (solid curve) and that in evidence optimal λ_{ev} (dot-dashed curve) both for $\lambda_0 = 1.0$. The latter curve has been scaled by a factor of 0.01 for ease of presentation and standard error bars are shown. Qualitatively, both curves show similar characteristics to the theoretical curves of figure 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).

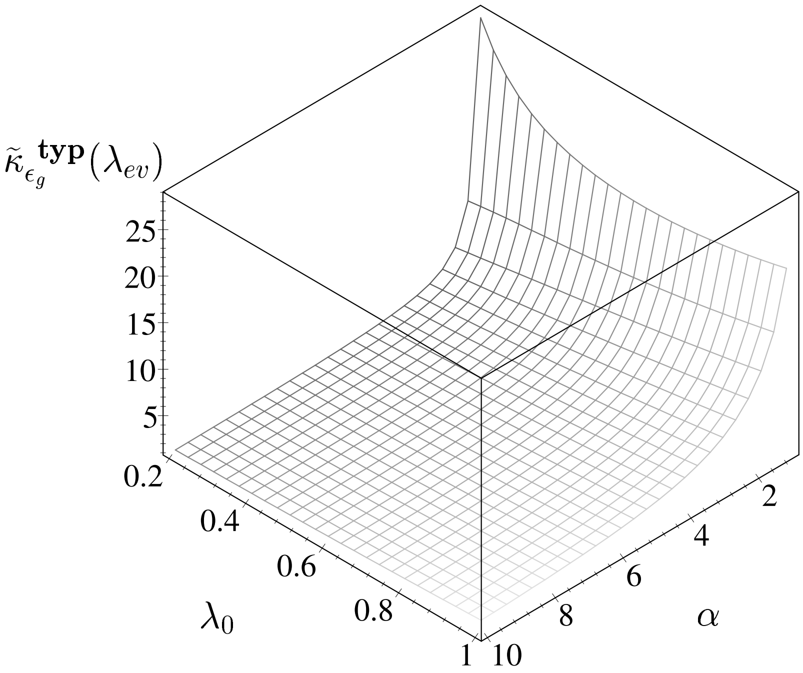

Figure 4. Scaled estimate of the fractional error κ_{ϵ_g}: for a system size of N dividing $\tilde{\kappa}_{\epsilon_g}{}^{\mathbf{typ}}(\lambda_{ev})$ by N gives the an estimate of the true fractional increase in error above the optimal incurred by using the evidence procedure. $\tilde{\kappa}_{\epsilon_g}{}^{\mathbf{typ}}(\lambda_{ev})$ diverges as $\lambda_0 \to \infty$ and as $\alpha \to 0$. For large α $\tilde{\kappa}_{\epsilon_g}{}^{\mathbf{typ}}(\lambda_{ev})$ tends to $1/N$ and for small noise it diverges for $\alpha < 1$ (see text).

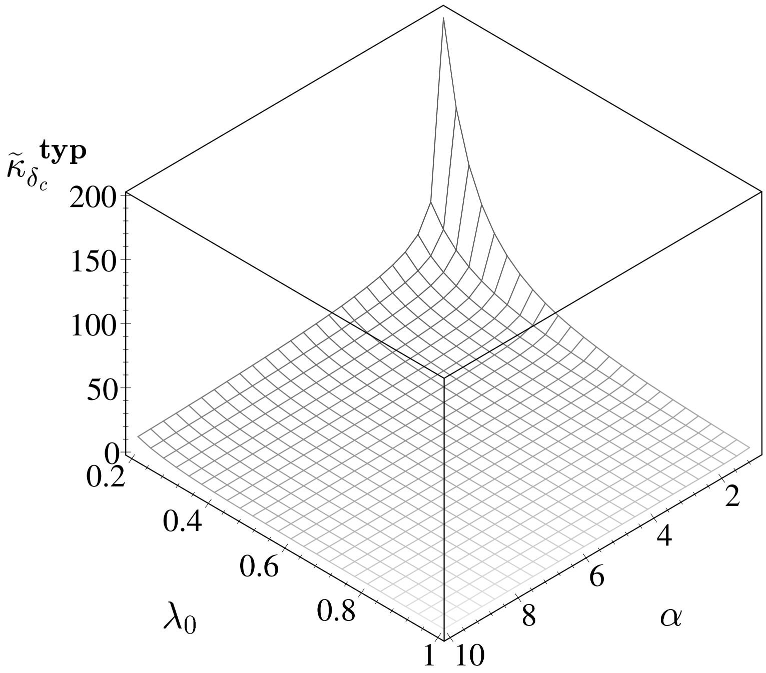

Figure 5. Scaled estimate of the fractional error κ_{δ_c}: for a system size of N dividing $\widetilde{\kappa}_{\delta_c}^{\,\mathbf{typ}}$ by $N^{1/2}$ gives an estimate of the true fractional error in estimating the generalisation error from the variance of the student distribution. $\widetilde{\kappa}_{\delta_c}^{\,\mathbf{typ}}$ diverges as $\alpha \to 0$ and as $\lambda_0 \to 0$.

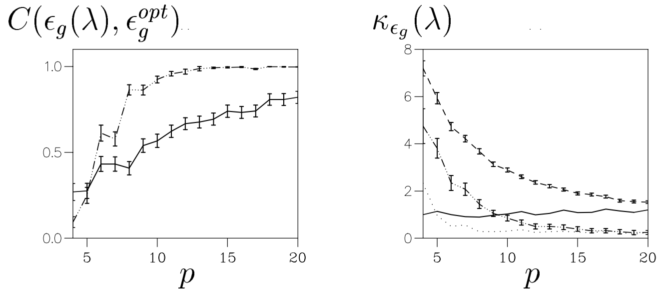

Figure 6. 1-D simulation results: The left-hand graph shows the correlation between the optimal generalisation error and those obtained using the evidence (solid) and cross-validation (chain) with $\lambda_0 = 1.0$. The right-hand graph shows the fractional increase in generalisation error $\kappa_{\epsilon_g}(\lambda) = (\epsilon_g(\lambda) - \epsilon_g(\lambda_{opt}))/\epsilon_g(\lambda_{opt})$. 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.