A statistical mechanical analysis of a Bayesian inference scheme for an unrealizable rule

1995 J. Phys. A: Math. Gen. 28 2159

(http://iopscience.iop.org/0305-4470/28/8/010)

J. Phys. A: Math. Gen. 28 (1995) 2159–2171. Printed in the UK

A statistical mechanical analysis of a Bayesian inference scheme for an unrealizable rule

Glenn Marion† and David Saad‡
Department of Physics, University of Edinburgh, Edinburgh, EH9 3JZ, UK

Received 22 August 1994, in final form 13 February 1995

Abstract. Within a Bayesian framework we consider a system that learns from examples. In particular, using a statistical mechanical formalism, 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 nonlinear teacher. The teacher is said to be unrealizable because the student can never model it without error. In fact, our model allows us to interpolate between the known linear case and an unrealizable, nonlinear, case. A comparison of the hyperparameters 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.

1. Introduction

The analysis of supervised learning, or learning from examples, is a major field of research in which techniques from statistical physics have been successfully employed. In general, one has a model mapping (*a student*) parametrized by some N_s-dimensional vector w and some, possibly noisy, examples $\mathcal{D}$ generated by the true mapping (*the teacher*). 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 $E_w(\mathcal{D})$ is some measure of the difference between the student and the teacher outputs over the set $\mathcal{D}$. Clearly, $E_w(\mathcal{D})$ 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 (see, for example, Krogh and Hertz 1992).

Minimization of the training energy, with respect to the weights w, leads to the problem of *over-fitting* and 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 1992). Occams 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 $\beta E_w(\mathcal{D}) + \gamma C(w)$, combining a measure of the performance on the data set and some *complexity cost* $C(w)$ 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. If $C(w) = w \cdot w$ then γ is termed the *weight decay*. The setting of the *hyperparameters*, β and γ, controls the

† E-mail address: glenny@uk.ac.ed
‡ E-mail address: D.Saad@uk.ac.ed

0305-4470/95/082159+13$19.50

learning algorithm. In this paper we will concern ourselves with the question of how to set the hyperparameters.

One can also consider the supervised learning paradigm within the context of Bayesian inference. In particular, MacKay (1992a) advocates the *evidence procedure* as a 'principled' method of setting hyperparameters. Moreover, it 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, for example, Wolpert 1993, MacKay 1993, Wolpert and Strauss 1994). 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. In fact, in some situations the evidence procedure does seem to improve performance (Thodberg 1993) whilst in others, as MacKay points out, it can be misleading (MacKay 1992b). 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. Theoretical results have been obtained for a linear perceptron trained and tested on data produced by a linear perceptron (Bruce and Saad 1994). They suggest that the evidence procedure is a useful guide to optimizing the learning algorithm's performance.

In the remainder of this paper we examine the evidence procedure, in relation to performance, for the case of a linear perceptron learning a nonlinear teacher. In the next section we review the Bayesian scheme, introducing the evidence and the relevant performance measures. In sections 3 and 4 we calculate these quantities in the case where the data is generated by a nonlinear mapping and the student is linear. Finally, in section 5 we examine the effects of the resultant unrealizability on the efficacy of the evidence procedure.

2. Bayesian formalism

2.1. The evidence

We take $E_w(\mathcal{D})$ to be the usual sum-squared error and assume that our 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 model w and β is

$$P(\mathcal{D}|\beta, w) \propto e^{-\beta E_w(\mathcal{D})}. \tag{2.1}$$

In order to incorporate Occams razor we also assume a prior distribution on our models. That is, we believe *a priori* in some rules more strongly than others. Specifically we believe that

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

Multiplying these together we obtain the post-training or student distribution

$$P(w|\mathcal{D}, \gamma, \beta) \propto e^{-\beta E_w(\mathcal{D}) - \gamma C(w)}. \tag{2.3}$$

It is clear that the most probable model w^* is given by minimizing the composite cost function $\beta E_w(\mathcal{D}) + \gamma C(w)$ with respect to 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 (minimization) algorithm can also give rise to this post-training distribution, equation (2.3) (Seung *et al* 1992).

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

$$P(\mathcal{D}|\gamma,\beta)=\int\prod_j \mathrm{d}w_j\, P(\mathcal{D}|\beta,w)P(w|\gamma)\,. \tag{2.4}$$

That is, the probability of (or evidence for) the data set ($\mathcal{D}$) given the hyperparameters β and γ. The evidence procedure fixes the hyperparameters to the values that maximize this probability for a given data set (MacKay 1992a).

2.2. *The performance measures*

Before defining our performance measures we must first introduce some notation. In general, we consider a teacher with output, $y_t(x)$, described by the conditional density $P(y_t|x)$. This accommodates, for example, deterministic teachers whose output is corrupted by noise. Furthermore, the inputs x are N-dimensional vectors sampled with probability $P(x)$. Thus a data set $\mathcal{D}=\{(y_t(x^\mu),x^\mu):\mu=1\ldots p\}$ is generated with probability $P(\mathcal{D})=\prod_{\mu=1}^{p}P(y_t|x^\mu)P(x^\mu)$. In general, we will use the notation $\langle f(z)\rangle_P$ to denote the average of the quantity $f(z)$ over the distribution $P(z)$. However, we will use the short-hand $\langle\cdot\rangle_w$ to mean the average over the post-training distribution $P(w|\mathcal{D},\gamma,\beta)$.

Many performance measures have been introduced in the literature (see, for example, Hansen 1993, Levin *et al* 1989, Krogh and Hertz 1992). Here, we consider the difference between the average student output $\langle y_s(x)\rangle_w$ and that of the teacher $y_t(x)$, squared and averaged over all possible inputs x, and teacher outputs (i.e. over $P(y_t|x)$) and finally, over all possible sets of data. Then, the generalization error

$$\epsilon_g=\langle(y_t(x)-\langle y_s(x)\rangle_w)^2\rangle_{P(y_t|x)P(x)P(\mathcal{D})}\,. \tag{2.5}$$

This is equivalent to the generalization error given by Krogh and Hertz (1992) up to a teacher-dependent constant, namely the variance in the teacher output over $P(y_t|x)$.

Another feature we can consider is the variance of the student output, $y_s(x)$, over the student distribution $\langle\{y_s(x)-\langle y_s(x)\rangle_w\}^2\rangle_{w,P(x)}$. 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 $P(x)$. Bruce and Saad define the consistency measure as the difference between this variance and the generalization error (Bruce and Saad 1994). 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 $\alpha=p/N$.). The consistency measure δ_c is now defined by

$$\delta_c=\langle\{y_s(x)-\langle y_s(x)\rangle_w\}^2\rangle_{w,P(x),P(\mathcal{D})}-\left(\epsilon_g-\epsilon_g^\infty\right). \tag{2.6}$$

In the limit $\alpha\to\infty$ δ_c tends to zero, even though the generalization error may not be zero. We regard $\delta_c=0$ as optimal since then we can estimate our expected error, ϵ_g, from the variance of our student output.

The fact that the quantities (2.5) and (2.6) 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 $\delta_c(\mathcal{D})$. 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 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.

3. The model

In our model the student is simply a linear perceptron and the input dimension N equals the model dimension N_s. The output for an input vector x^μ is given by

$$y_s^\mu = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j x_j^\mu \,. \tag{3.1}$$

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

$$P(y_t|x) \propto \sum_{\Omega=1}^{n} P(y_t|x,\Omega) P(x|\Omega) P_\Omega^t \,. \tag{3.2}$$

where $P(y_t|x,\Omega) \propto \exp([y_t - w^\Omega \cdot x]^2/2\sigma_\Omega^2)$, $P(x|\Omega)$ is $N(\bar{a}_\Omega, \sigma_{x_\Omega}^2)$† and P_Ω^t is chosen such that $\sum_{\Omega=1}^n P_\Omega^t = 1$. The input distribution is $P(x) = \sum_{\Omega=1}^n P(\Omega) P(x|\Omega)$.

One way of visualising the n-teacher mapping is as the average over the conditional distribution $P(y_t|x)$. Figure 1 displays some examples of a 2-teacher with one-dimensional input vector. Figure 1 curve (i) shows the linear case whilst (ii) shows the average of two linear teachers when the distributions $P(x|\Omega)$ are the same, again the average output is a linear function of the input. Finally in figure 1 curve (iii) the distributions $P(x|\Omega = 1)$ and $P(x|\Omega = 2)$ are both centred on the origin but have different variances; the average output is a nonlinear function of the input. In fact, for the general case, where the distributions, $P(x|\Omega)$, 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

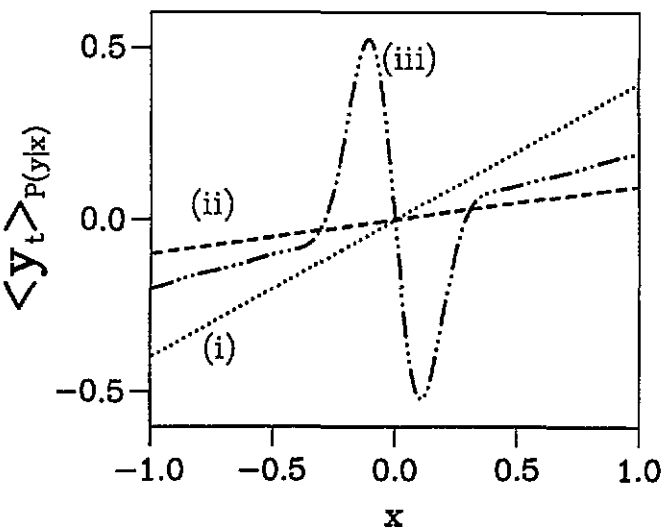

Figure 1. A 2-teacher in 1D: The average output $\langle y_t \rangle_{P(y|x)}$ (i) when the component teacher vectors are aligned, (ii) when they are misaligned but $\sigma_{x_1} = \sigma_{x_2}$ and (iii) with $\sigma_{x_1} \neq \sigma_{x_2}$ and with the teachers misaligned.

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

section of the mapping. In this way a nonlinear teacher is constructed, in a piecewise linear fashion, with n segments. As n grows we can steadily improve our approximation of arbitrary piecewise linear functions.

Given this model a data set $\mathcal{D} = \prod_{\Omega=1}^{n}\{(w^{\Omega}\cdot x_{\Omega}^{\mu_\Omega} + \eta_{\Omega}^{\mu_\Omega}, x_{\Omega}^{\mu_\Omega}) : \mu_\Omega = 1 \ldots p_\Omega\}$. The variables $\eta_{\Omega}^{\mu_\Omega}$ are drawn independently from a Gaussian distribution with zero mean and variance σ_Ω^2 whilst the $x_{\Omega}^{\mu_\Omega}$ are drawn independently from $P(x|\Omega)$. The range of the index μ_Ω is from 1 to p_Ω where on average $p_\Omega = pP_\Omega^t$.

Adopting $C(w) = w\cdot w$ 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 $\ln P(\mathcal{D}|\lambda,\beta) = -Nf(\mathcal{D})$, where we have introduced $\lambda = \gamma/\beta$. The quantity $f(\mathcal{D})$ is analogous to a free energy in statistical physics. This analogy has been noted by others, for example (Neal 1992). The expression for the free energy is of the form,

$$-f(\mathcal{D}) = \frac{1}{2}\ln\frac{\lambda}{\pi} + \frac{\alpha}{2}\ln\frac{\beta}{\pi} + \frac{1}{2}\ln 2\pi + \frac{1}{2N}\ln\det g + \frac{1}{N}\rho_j g_{jk}\rho_k - \kappa \tag{3.3}$$

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 the evidence:

$$\epsilon_g = \left\langle \frac{\sigma_{x_\Omega}^2}{N}P_\Omega^t\left\{w_j^\Omega w_j^\Omega - 2w_j^\Omega\langle w_j\rangle_w + \langle w_j\rangle_w^2\right\}\right\rangle_{P(\mathcal{D})} + P_\Omega^t\sigma_\Omega^2 \tag{3.4}$$

$$\delta_c = \frac{\sigma_{x_{\rm eff}}^2}{2N\beta}\langle \mathrm{tr}\,\mathbf{g}\rangle_{P(\mathcal{D})} - \left(\epsilon_g - \epsilon_g^\infty\right) \tag{3.5}$$

where, $\langle w_j\rangle_w = \rho_k g_{kj}$ and $\sigma_{x_{\rm eff}}^2 = P_\Omega^t\sigma_{x_\Omega}^2$

Due to the sampling assumptions of our model 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 $P(\mathcal{D})$).

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* 1989). 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 of the thermodynamic approximation in more detail elsewhere. 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 $O(1/N)$. Thus, in the thermodynamic limit, the fluctuations, from one data set to the next, vanish.

We now calculate these thermodynamic averages. After the average over the noise variables we are left with the average over the input distribution. In particular, we need to calculate $\langle\langle \mathbf{g} \rangle\rangle$, $\langle\langle \mathbf{A}_\Omega \mathbf{g} \rangle\rangle$ and $\langle\langle \mathbf{A}_2 \mathbf{g} \mathbf{A}_1 \rangle\rangle$, where the double brackets refer to averages, in the thermodynamic limit, over the input distribution. The details are relegated to the appendix, where equation (A.7) defines $NG = \mathrm{tr}\langle\langle \mathbf{g} \rangle\rangle$ and $\langle\langle \mathbf{A}_\Omega \mathbf{g} \rangle\rangle$ and $\langle\langle \mathbf{A}_2 \mathbf{g} \mathbf{A}_1 \rangle\rangle$ are defined by (A.4) and (A.5), respectively.

The averaged free energy f can now be written

$$f = -\frac{1}{2}\ln\frac{\lambda}{\pi} - \frac{\alpha}{2}\ln\frac{\beta}{\pi} - \frac{1}{2}\ln 2\pi - \frac{1}{2N}\langle\langle \ln \det \mathbf{g} \rangle\rangle + \beta\sigma_1^2(P_1^t\alpha - \Psi_1 G) + \beta\sigma_2^2(P_2^t\alpha - \Psi_2 G) + \beta G(\sigma_{w_1}^2\lambda\Psi_1 + \sigma_{w_2}^2\lambda\Psi_2 + \Psi_1\Psi_2 D_w) \tag{4.1}$$

the generalization error as

$$\epsilon_g = P_1^t\sigma_{w_1}^2\sigma_{x_1}^2 + P_2^t\sigma_{w_2}^2\sigma_{x_2}^2 + P_1^t\sigma_1^2 + P_2^t\sigma_2^2 - 2P_1^t\sigma_{x_1}^2(\Psi_1\sigma_{w_1}^2 + \Psi_2\theta_w)G - 2P_2^t\sigma_{x_2}^2(\Psi_2\sigma_{w_2}^2 + \Psi_1\theta_w)G + \sigma_{x_{\mathrm{eff}}}^2\frac{\partial}{\partial\lambda}\left[G\left\{\Psi_1\Psi_2 D_w + \Psi_1(\lambda\sigma_{w_1}^2 - \sigma_1^2) + \Psi_2(\lambda\sigma_{w_2}^2 - \sigma_2^2)\right\}\right] \tag{4.2}$$

whilst the consistency measure becomes

$$\delta_c = \frac{\sigma_{x_{\mathrm{eff}}}^2}{2\beta}G - \left(\epsilon_g - \epsilon_g^\infty\right) \tag{4.3}$$

and we have defined

$$D_w = \frac{1}{N}|w^1 - w^2|^2 \qquad \sigma_{w_\Omega}^2 = \frac{1}{N}w^\Omega \cdot w^\Omega \qquad \text{and} \qquad \theta_w = \frac{1}{N}w^1 \cdot w^2 .$$

The variable D_w is a measure of the Euclidean distance in weight space between the two teachers whilst, $\sigma_{w_\Omega}^2$ measures the magnitude of teacher Ω and θ_w is the overlap between the two teachers. We note that in two limits we recover the learnable, linear teacher, case. Specifically, if the probability of picking one of the component teachers is zero or if both component teacher vectors are aligned. We can now examine the evidence and the performance measures for our unlearnable problem.

5. Results and discussion

5.1. The performance measures

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

$$\epsilon_g \sim \frac{P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w}{\sigma_{x_{\mathrm{eff}}}^2} + P_\Omega^t \sigma_\Omega^2 + \frac{P_1^t P_2^t \sigma_{x_1}^4 \sigma_{x_2}^4 D_w}{\sigma_{x_{\mathrm{eff}}}^6}\frac{1}{\alpha} + O\left(\frac{1}{\alpha^2}\right). \tag{5.1}$$

Similarly, also for large α,

$$|\delta_c| \sim \frac{1}{\alpha}\left(\frac{P_1^t P_2^t \sigma_{x_1}^4 \sigma_{x_2}^4 D_w}{\sigma_{x_{\mathrm{eff}}}^6} + \frac{1}{\beta\sigma_{x_{\mathrm{eff}}}^4}\right) + O\left(\frac{1}{\alpha^2}\right). \tag{5.2}$$

In the limit of infinite α, $|\delta_c|$ tends to zero and $\epsilon_g^\infty = P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w/\sigma_{x_{\mathrm{eff}}}^2 + P_\Omega^t \sigma_\Omega^2$. 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.

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

$$\epsilon_g = \tau\left(\alpha, P_\Omega^t, \sigma_{x_\Omega}^2, \sigma_\Omega^2, \sigma_{w_\Omega}^2\right) + \frac{\alpha \sigma_{x_{\text{eff}}}^2}{1-\alpha}\left(\frac{P_1^t \sigma_1^2}{\sigma_{x_1}^2} + \frac{P_2^t \sigma_2^2}{\sigma_{x_2}^2} + \alpha P_1^t P_2^t D_w\right) \tag{5.3}$$

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

$$\epsilon_g \sim \left(\frac{\sigma_1^2 \sigma_{x_2}^2 P_1^t + \sigma_2^2 \sigma_{x_1}^2 P_2^t + P_1^t P_2^t \sigma_{x_2}^2 \sigma_{x_1}^2 D_w}{\sigma_{x_{\text{eff}}}^2 \sigma_{x_1}^2 \sigma_{x_2}^2}\right)(\alpha - 1)^{-1}. \tag{5.4}$$

In accord with standard results (for example, Krogh and Hertz 1992, Dunmur and Wallace 1993), if there is no noise and $D_w = 0$, the generalization error is proportional to $1 - \alpha$ for $\alpha < 1$ and zero for $\alpha > 1$. Figure 2(a) shows the generalization error in the zero-γ limit. The case of a 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(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 maxima can easily be understood; if there is no noise or γ is large enough then there is a steady reduction in ϵ_g (figure 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 effect of the noise begins 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 1994b).

Figure 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 5.2.1, in this limit, the consistency is optimized by the evidence procedure for the linear case only. A nonlinear case is shown in figure 2(c) curve (iv) where the temperature is set by the evidence procedure but the true optimal consistency measure is actually zero.

Finally, figure 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

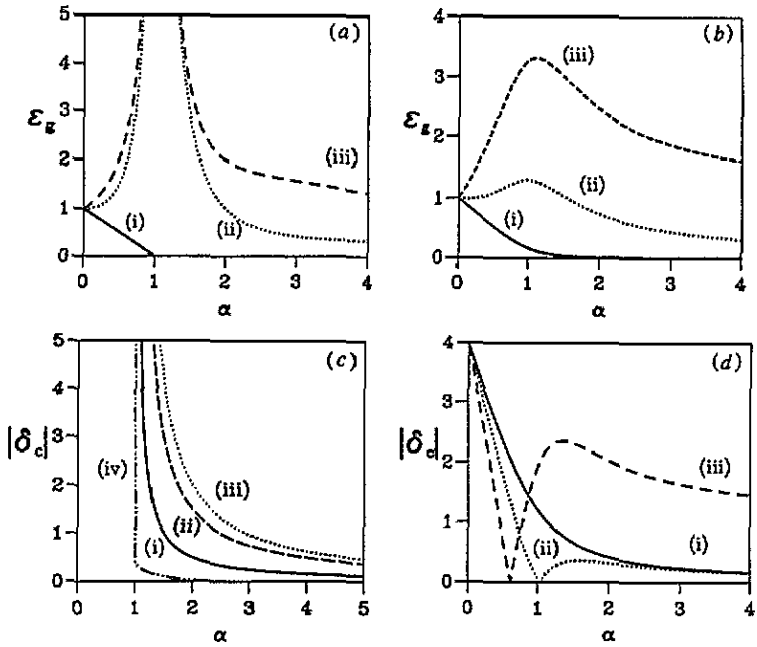

Figure 2. The performance measures: graph (*a*) shows the generalization error versus α for zero γ. (*a*) Curves (i) and (ii) are the realizable case without noise and with noise, respectively; (iii) is an unlearnable case where we can see that the unlearnability acts qualitatively in the same manner as noise. (*b*) Shows ϵ_g for finite γ. Curves (i) and (ii) are learnable scenarios in the latter case with noise; (iii) shows that the effect of *adding* unlearnability is qualitatively the same as adding noise. (*c*) Shows $|\delta_c|$ for $\gamma \to 0$, note that for $\alpha < 1$ the consistency measure diverges. Graph (i) shows the learnable linear case; (ii) shows the unlearnable but linear case and (iii) is the nonlinear case; (iv) shows the effect of setting the learning temperature to T_{ev}; the evidence optimal temperature. In this latter case the optimal value of the consistency measure is $|\delta_c| = 0$. (*d*) Shows the modulus of the consistency error versus α for finite γ. Curves (i) and (ii) are the learnable case without and with noise respectively; (iii) is an unlearnable case with the same noise level. (*a*) ϵ_g for zero γ, (*b*) ϵ_g for finite γ, (*c*) δ_c for zero γ and (*d*) δ_c for finite γ.

of the zero of the consistency measure is a reflection of the trade-off between the effective noise and the weight decay described above (figure 2(*d*) curves (ii) and (iii) 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 figure 2(*d*) curve (i)).

5.2. *The evidence procedure*

We now turn to the evidence and, in particular, to the assignments of the hyperparameters we can make from it. We define $\beta_{ev}(\gamma)$ and $\gamma_{ev}(\beta)$ to be the hyperparameters which maximise the evidence with respect to fixed γ and β respectively. The evidence procedure picks the point in hyperparameter space where these curves coincide. Furthermore, we define β_{ev}^{∞} and γ_{ev}^{∞} to be the solutions to $\lim_{\alpha\to\infty} \frac{\partial f}{\partial \beta}|_{\gamma=\text{constant}} = 0$ and $\lim_{\alpha\to\infty} \frac{\partial f}{\partial \gamma}|_{\beta=\text{constant}} = 0$, 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

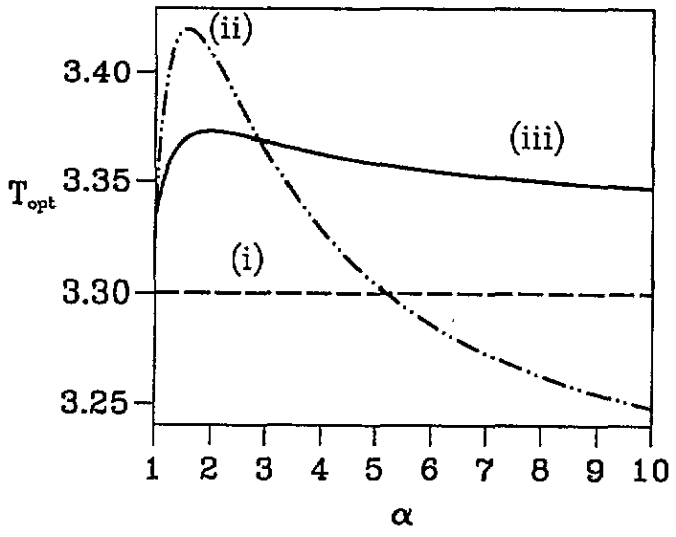

Figure 3. Optimal temperatures in the $\gamma \to 0$ model. (*i*) The evidence procedure estimate T_{ev} and that which optimizes the consistency measure T_{δ_c} coincides in the linear regime. In the nonlinear regime (ii) shows the dependence of T_{ev} on α and (iii) shows that of T_{δ_c}.

analogue of curve (iii) in figure 1, we shall speak of the nonlinear 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.

5.2.1. The $\gamma \to 0$ limit. The simplest case is the unregularized limit where we have only one hyperparameter (β) to optimize. In this limit, for $\alpha < 1$, the evidence is maximal for $T \equiv 1/\beta = 0$ whereas, for $\alpha > 1$, the evidence optimal temperature (T_{ev}) is finite as shown in figure 3 curves (i) and (ii). In fact this transition in behaviour is analogous to the phase transition found by Bruce and Saad (1994). In the regime $\alpha < 1$ there is not even enough data to specify the perceptron weights and in consequence, there is no option but to believe the data completely. Thus, the evidence optimal learning temperature is zero. However, for $\alpha > 1$ and increasing we can make steadily better estimates of the noise in the examples. 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 independent of the learning temperature for $\alpha < 1$. We also comment that the generalization error is a function of λ 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

$$T_{ev}^{\infty} = 2(P_1^t\sigma_1^2 + P_2^t\sigma_2^2) + \frac{2P_1^tP_2^t\sigma_{x_1}^2\sigma_{x_2}^2 D_w}{\sigma_{x_{eff}}^2}. \tag{5.5}$$

In the linear regime T_{ev} is constant ($\forall\alpha > 1$) as shown in figure 3 curve (i) whereas in the nonlinear regime figure 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 nonlinear case is shown in figure 2(*c*) curve (iv) where the optimal $|\delta_c|$ is actually zero. The learning temper-

ature which minimizes the consistency (T_{δ_c}) is shown for this case in figure 3 curve (iii) (this is the same case as figure 3 curve (ii) which shows T_{ev}). In the limit $\alpha \to \infty$, T_{δ_c} becomes

$$T_{\delta_c}^{\infty} = \frac{2(\sigma_1^2\sigma_{x_1}^2 P_1^t + \sigma_2^2\sigma_{x_2}^2 P_2^t)}{\sigma_{x_{eff}}^2} + \frac{2P_1^t P_2^t \sigma_{x_1}^4 \sigma_{x_2}^4 D_w}{\sigma_{x_{eff}}^6}. \tag{5.6}$$

Contrasting this with (5.5) 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 ($T_{ev} = T_{\delta_c}$ = constant $\forall\alpha$ s.t. $\alpha > 1$). However, for a nonlinear teacher or noise that varies across the input space, even in the large-α limit, it does not.

5.2.2. 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 (5.5), whilst

$$\frac{1}{\gamma_{ev}^{\infty}} = \frac{2(\sigma_{w_1}^2\sigma_{x_1}^2 P_1^t + \sigma_{w_2}^2\sigma_{x_2}^2 P_2^t)}{\sigma_{x_{eff}}^2} - \frac{2P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w}{\sigma_{x_{eff}}^4}. \tag{5.7}$$

These asymptotic assignments can be understood intuitively. The setting of T_{ev}^{∞} reflects the average noise in the examples ($P_1^t\sigma_1^2 + P_2^t\sigma_2^2$) and the noise due to the unlearnability, $P_1^t P_2^t \sigma_{x_1}^2 \sigma_{x_2}^2 D_w/\sigma_{x_{eff}}^2$, discussed earlier. The weight decay term is not as easy to interpret. However, in the linear regime we have $N/2\gamma_{ev}^{\infty} = |w^1 P_1^t + w^2 P_2^t|^2$; 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.

In order to assess 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 hyperparameters (for finite α) coincide with β_{ev}^{∞} and γ_{ev}^{∞} and also optimize ϵ_g and δ_c in agreement with Bruce and Saad (1994). This is shown in figure 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

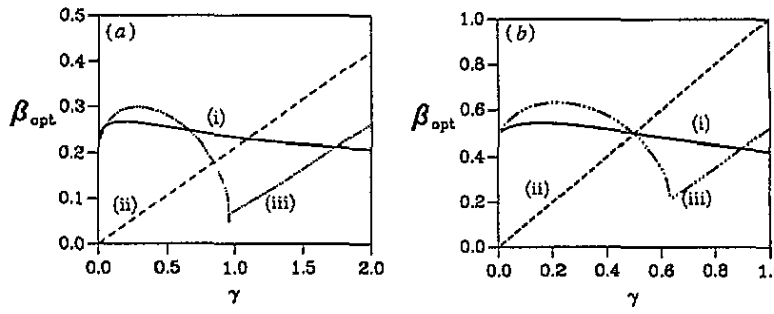

Figure 4. The evidence procedure: optimal β versus γ. (a) Linear case. (b) Nonlinear case. In both graphs the β which optimizes evidence $\beta_{ev}(\gamma)$ is curve (i), that which optimizes the *generalization error $\beta_{\epsilon_g}(\gamma)$ curves (ii) and that which optimizes the consistency measure $\beta_{\delta_c}(\gamma)$* curve (iii). In (a) the evidence procedure picks the point, in the γ–β plane, where all three curves coincide. In (b) the evidence procedure point coincides only with curve (i). (a) Linear case, (b) nonlinear case.

the β–γ plane) which the evidence procedure picks. However, we note here that if one of the hyperparameters is poorly determined then the evidence procedure is a poor guide to optimizing performance even in the linear case.

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 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 nonlinear regime is shown in figure 4(b). In this instance the parameters picked by *the evidence procedure neither minimize* ϵ_g *nor* 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.

Any Bayesian scheme must make assumptions concerning the process generating the data (i.e. assumptions concerning the teacher) and, in general, such assumptions will not be valid. In this paper, in the nonlinear 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 nonlinear regime, if we have different real noise levels associated with 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 violation of our Bayesian assumptions. One further question we might ask concerns robustness; given that the evidence procedure does not optimise performance in the nonlinear regime how far from optimality is it? We simply note here that we have explored this issue elsewhere (Marion and Saad 1995).

6. Conclusion

In this work 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 nonlinear 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. However, for a nonlinear teacher (and a linear student) the evidence procedure breaks down in that it fails to optimize the performance measures. Whether or not such a breakdown of the evidence procedure is a generic feature of a mismatch between the hypothesis (student) space and the teacher space is a matter for further study.

Acknowledgments

We are very grateful to Alastair Bruce and Peter Sollich for useful discussions and to David Barber for a thorough reading of the manuscript. GM is supported by a SERC studentship.

Appendix

In this appendix we calculate the averages over the input distribution, $P(x)$, required in section 4.

We note that $P(x) = P(x|\Omega = 1)P_1^t + P(x|\Omega = 2)P_2^t$ and in what follows $\langle\langle \mathbf{g} \rangle_1\rangle_2 = \langle\langle \mathbf{g} \rangle_2\rangle_1 = \langle\langle \mathbf{g} \rangle\rangle$, where $\langle \cdots \rangle_1$ and $\langle \cdots \rangle_2$ refer to averages over the distributions $P(x|\Omega = 1)$ and $P(x|\Omega = 2)$.

Firstly let us 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 $P(x|\Omega = 1)$. This step is similar to the calculation, in Hertz *et al* (1989), 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,

$$\langle \mathbf{g} \rangle_1 = (\mathbf{\Gamma} + \Sigma_1 \mathbf{I})^{-1} \qquad \text{where} \qquad \Sigma_\Omega = \frac{\alpha P_\Omega^t \sigma_{x_\Omega}^2}{1 + \sigma_{x_\Omega}^2 \frac{1}{N} \langle \mathrm{tr}\, \mathbf{g} \rangle_\Omega} . \tag{A.1}$$

We can then rewrite (A.1) as

$$\langle \mathbf{g} \rangle_1 \mathbf{A}_2 = \mathbf{I} - \lambda \langle \mathbf{g} \rangle_1 - \langle \mathbf{g} \rangle_1 \Sigma_1 . \tag{A.2}$$

Now we wish to perform the average over the second distribution $P(x|\Omega = 2)$ but the last term in (A.2) is potentially problematic. However, if following the diagrammatic method of (Hertz *et al* 1989) we examine the diagrams for this term we see that the 'crossings', or *interactions*, between $\langle \mathbf{g} \rangle_1$ and Σ_1 are $O(1/N^2)$ and can be ignored in the thermodynamic limit. Thus, we can average the two factors independently, ignoring any interaction between them. This leads to

$$\langle\langle \mathbf{g} \mathbf{A}_2 \rangle\rangle = \mathbf{I} - \lambda \langle\langle \mathbf{g} \rangle\rangle - \langle\langle \mathbf{g} \rangle\rangle \langle\langle \Sigma_1 \rangle\rangle . \tag{A.3}$$

Using the matrix identity $\mathbf{g}\mathbf{A}_2 = \mathbf{I} - \lambda \mathbf{g} - \mathbf{A}_1 \mathbf{g}$ and defining $\Psi_1 = \langle \Sigma_1 \rangle_2$ we obtain $\langle\langle \mathbf{g} \mathbf{A}_1 \rangle\rangle = \Psi_1 \langle\langle \mathbf{g} \rangle\rangle$. If we perform these averages the other way around and define $\Psi_2 = \langle \Sigma_2 \rangle_1$ we find the analogous expression. Thus, in general we have

$$\langle\langle \mathbf{g} \mathbf{A}_\Omega \rangle\rangle = \Psi_\Omega \langle\langle \mathbf{g} \rangle\rangle . \tag{A.4}$$

Now by multiplying (A.2) by $\mathbf{A}_2$, averaging over the distribution $P(x|\Omega = 2)$ 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

$$\langle\langle \mathbf{A}_2\, \mathbf{g} \mathbf{A}_1 \rangle\rangle = \Psi_1 \Psi_2 \langle\langle \mathbf{g} \rangle\rangle . \tag{A.5}$$

We now have all the averages we require in terms of the average $\langle\langle \mathbf{g} \rangle\rangle$. 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

$$(\Psi_1 + \Psi_2 + \lambda) \langle\langle \mathbf{g} \rangle\rangle = \mathbf{I} . \tag{A.6}$$

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

$$G^{-1} = \lambda + \frac{\alpha P_1^t \sigma_{x_1}^2}{1 + \sigma_{x_1}^2 G} + \frac{\alpha P_2^t \sigma_{x_2}^2}{1 + \sigma_{x_2}^2 G} \tag{A.7}$$

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 this response function could be calculated using the more general method of Sollich (1994a).

References

Bruce A D and Saad D 1994 Statistical mechanics of hypothesis evaluation *J. Phys. A: Math. Gen.* **27** 3355–63

Dunmur A P and Wallace D J 1993 Learning and generalization in a linear perceptron stochastically trained with noisy data *J. Phys. A: Math. Gen.* **26** 5767–79

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

Hertz J, Krogh A and Thorbergsson G 1989 *J. Phys. A: Math. Gen.* **22** 2133–50

Krogh A and Hertz J 1992 Generalization in a linear perceptron in the presence of noise *J. Phys. A: Math. Gen.* **25** 1135–47

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* (San Mateo, CA: Kauffmann) pp 245–60

MacKay D J C 1992a Bayesian interpolation *Neural Comp.* **4** 415–47

——1992b A practical Bayesian framework for backprop networks *Neural Comput.* **4** 448–72

——1994 Hyperparameters: optimise or integrate out? *Maximum entropy and Bayesian methods (Santa Barbara, 1993)* ed G Heidbreder (Dordrecht: Kluwer)

Marion G and Saad D 1995 Hyperparameters, evidence and generalization for an unrealisable rule. *Advances in Neural Information Processing Systems* vol 7, ed Tesauro *et al* (Cambridge MA: MIT Press), in press

Neal R M 1992 Bayesian training of backpropagation networks by the hybrid Monte Carlo method *Technical Report* CRG-TR-92-1 University of Toronto, Dept. of Computer Science.

Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev.* A **45** 6056–91

Thodberg H H 1994 Bayesian backprop in action: pruning, ensembles, error bars and application to spectroscopy *Advances in Neural Information Processing Systems* vol 6 ed Cowan *et al* (San Mateo, CA: Kauffmann) pp 208–15

Sollich P 1994a Finite-size effects in learning and generalization in linear perceptrons *J. Phys. A: Math. Gen.* **27** 7771–84

——1994b Minimum entropy queries for linear students learning nonlinear rules, in preparation

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* vol 5 ed Hanson *et al* (San Mateo, CA: Kauffmann) pp 539–46

Wolpert D H and Strauss C E M 1994 What Bayes has to say about the evidence procedure *Maximum entropy and Bayesian methods* ed G Heidbreder (Dordrecht: Kluwer) to appear