Statistical mechanics of hypothesis evaluation

Statistical mechanics of hypothesis evaluation

A D Bruce and D Saad

Department of Physics, The University of Edinburgh, Edinburgh EH9 3JZ, UK

Received 23 August 1993

Abstract. Following ideas of Gull, Skilling and MacKay, we develop and explore a statistical-mechanics framework through which one may assign values to the parameters of a model for a 'rule' (instanced, here, by the noisy linear perceptron), on the basis of data instancing the rule. The 'evidence' which the data offers in support of a given assignment, is likened to the free energy of a system with quenched variables (the data): the most probable (MAP) assignments of parameters are those which minimize this free-energy; tracking the free-energy minimum may lead to 'phase transitions' in the preferred assignments. We explore the extent to which the MAP assignments lead to optimal performance.

1. Introduction

The task of parameterizing a model of a set of data is one of the recurring problems of science. In general terms the problem is as follows. We are given a set of data $\mathcal{D}$ comprising p members; each member instances an unknown rule (or mapping) $\mathcal{R}^0$ connecting the components of the member (the 'input' and the 'output', in the language of neural networks). We attempt to model $\mathcal{R}^0$ by some rule $\mathcal{R}(\{w\})$, specified by a parameter set $\{w\}$ having N members. Prior convictions (hypotheses) about the rule $\mathcal{R}^0$ are expressed, *de facto*, in the form chosen for $\mathcal{R}$, and in the values assigned to a further set of parameters $\{\beta\}$, characterizing (for example) constraints on the set $\{w\}$. The modelling process has many facets, both practical and conceptual; it may be addressed from a wide range of perspectives and with a range of techniques. Here, we focus on one particular issue—the strategy underlying the assignment of what we shall call the *hypothesis parameters* $\{\beta\}$; we do so in a context—that of sufficiently large N and p—where the methods of statistical mechanics are appropriate. Although the large-N assumption distances the analysis from the more usual modelling tasks, it is relevant to neural-network modelling (see e.g. [1]) and, in concept at least, to the general task of image restoration (see e.g. [2]).

Two recent bodies of work provide the context and motivation for this study.

First, building on work of Gull [3] and Skilling [4], MacKay [5, 6] has developed and explored a Bayesian framework in which hypothesis parameters $\{\beta\}$ are assigned values such as to maximize the conditional probability $P(\mathcal{D}|\{\beta\})$, which, following these authors, we shall refer to as the 'evidence' for the parameters. The rationale for the name lies in the fact that, in the absence of prior information on the hypothesis parameters, the evidence $P(\mathcal{D}|\{\beta\})$ provides a direct measure of the conditional probability $P(\{\beta\}|\mathcal{D})$. The parameters which maximize the evidence are thus 'optimal' in the sense that they represent the single most likely (maximum *a posteriori*, MAP) values given the data.

Second, in a contribution to a substantial programme devoted to the statistical mechanics of learning, Krogh and Hertz [7] have explored the particular case in which the rule

0305-4470/94/103355+09$19.50 © 1994 IOP Publishing Ltd

connecting input and output is *linear*, the data is corrupted with *Gaussian noise*, and the parameters $\{w\}$ of the model rule are drawn from a *Gaussian prior*. For this case (the 'noisy linear perceptron with weight decay', to be referred to as NLP) they identify a *line* through the two-dimensional space of hypothesis parameters, characterizing assignments that are 'optimal' in the sense† that they minimize the typical value of the error with which the model will predict the the output associated with a new input (the *generalization error*).

In this paper we develop each of these strands, and explore their interconnections. We calculate the evidence for the NLP—or, more precisely its logarithm, averaged over an ensemble of data sets $\mathcal{D}$; it plays the role of the free energy of a many-body system with quenched random variables (the data). While this function has been calculated by others [10] within the replica formulation of learning theory [9], the result given here, based on the methods of Krogh and Hertz [7], is explicit and transparent, allowing us to explore the role of the evidence as a predictor of 'optimal' parameters. The notion of the 'evidence' as a free energy has also been noted elsewhere [11]; here we point out that, in consequence, one may naturally expect to find different 'phases' of hypothesis space, and we show that the NLP (in the limit of zero weight decay) provides a simple and explicit example of a phase transition separating a 'phase' in which a hypothesis parameter is sharply determined by the data, from a phase in which it is undetermined. We also explore the relationship between the evidence-based assignments of 'optimal' parameters and assignments based on other performance measures such as the generalization error. The results suggest that the evidence provides an effective guide to performance only if the model is at least potentially well matched to the underlying reality.

2. Calculation: evidence and performance measures

We suppose that our data $\mathcal{D}$ comprises p input–output pairs $(\boldsymbol{x}^\mu, y^\mu)$ where the vector $\boldsymbol{x}^\mu$ has elements x_j^μ $(j = 1 \ldots N)$. We suppose that the input–output relation $\mathcal{R}^0$ is linear, and subject to corruption by a Gaussian noise:

$$y^\mu = \mathcal{R}^0(\boldsymbol{x}^\mu) + \eta = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j^0 x_j^\mu + \eta \tag{2.1}$$

where w_j^0 $(j = 1 \ldots N)$ are the elements of a rule vector $\boldsymbol{w}^0$, and η is a random Gaussian-noise variable of variance σ. We suppose that the elements x_j^μ of the inputs are also Gaussian random variables of variance σ_x.

We proceed to examine the effectiveness of 'models' $\mathcal{R}$ of the data-generating process which take the rule to be linear, parameterized by a rule vector $\boldsymbol{w}$,

$$\mathcal{R}(\boldsymbol{x}^\mu) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j x_j^\mu \tag{2.2}$$

with elements $\{w\}$ that are Gaussian distributed with variance $1/2\gamma$, and which take the corruption process to be Gaussian with variance $1/2\beta$. The quantities β and γ constitute the 'hypothesis parameters'.

† The word 'optimal' is used with a wide range of meanings in the relevant literature. Thus, for example, Watkin [8] considers 'optimal learning', meaning the behaviour of a network *all* of whose parameters are chosen so as to minimize the generalization error.

The hypothesis about the rule itself is expressed through the conditional probability (prior):

$$P(\boldsymbol{w}|\gamma) = \left(\frac{\gamma}{\pi}\right)^{N/2} \mathrm{e}^{-\gamma\sum_{j=1}^{N} w_j^2} . \tag{2.3}$$

The hypothesis about the corruption process is expressed through the further conditional probability (likelihood):

$$P(\mathcal{D}|\boldsymbol{w},\beta) = \left(\frac{\beta}{\pi}\right)^{p/2} \prod_{\mu=1}^{p} \mathrm{e}^{-\beta(y^\mu - \frac{1}{\sqrt{N}}\sum_{j=1}^{N} w_j x_j^\mu)^2} . \tag{2.4}$$

Appealing to Bayes theorem [12], these two conditional probabilities may be combined to give a conditional probability (the posterior) for the model rule, parameterized by the rule vector $\boldsymbol{w}$:

$$P(\boldsymbol{w}|\mathcal{D},\gamma,\beta) = \frac{P(\mathcal{D}|\boldsymbol{w},\beta)P(\boldsymbol{w}|\gamma)}{P(\mathcal{D}|\gamma,\beta)} . \tag{2.5}$$

The normalizing factor appearing in this relation,

$$P(\mathcal{D}|\gamma,\beta) = \sum_{\boldsymbol{w}} P(\mathcal{D}|\boldsymbol{w},\beta)P(\boldsymbol{w}|\gamma) \tag{2.6}$$

constitutes the 'evidence' which the data provides for the assignment of the model parameters [5, 6]. In the present context the 'sum' over rule space which it entails is implemented by an integral over the elements w_j of the model rule:

$$\sum_{\boldsymbol{w}} \longrightarrow \int \prod_j \mathrm{d}w_j . \tag{2.7}$$

Writing the expression for the evidence explicitly one obtains

$$P(\mathcal{D}|\gamma,\beta) = \left(\frac{\gamma}{\pi}\right)^{N/2} \left(\frac{\beta}{\pi}\right)^{p/2} \int \prod_j \mathrm{d}w_j \mathrm{e}^{-\mathcal{H}} \tag{2.8}$$

where

$$\mathcal{H} = \tfrac{1}{2}\sum_{jk} r_j \Lambda_{jk}^{-1} r_k - \sum_j \rho_j r_j + \beta \sum_\mu \eta_\mu^2 + \gamma \sum_j (w_j^0)^2 \tag{2.9}$$

while $r_j \equiv w_j^0 - w_j$ and

$$\begin{aligned} \Lambda_{jk}^{-1} &\equiv \frac{2\beta}{N} \sum_\mu x_j^\mu x_k^\mu + 2\gamma\delta_{jk} \\ \rho_j &\equiv -\frac{2\beta}{\sqrt{N}} \sum_\mu x_j^\mu \eta_\mu + 2\gamma w_j^0 . \end{aligned}$$

Performing the Gaussian integrals over rule space we obtain

$$\begin{aligned} \ln P(\mathcal{D}|\gamma,\beta) = {} & \tfrac{1}{2}N \ln\left(\frac{\gamma}{\pi}\right) + \tfrac{1}{2}p \ln\left(\frac{\beta}{\pi}\right) + \tfrac{1}{2}\ln \det\Lambda + \tfrac{1}{2}N\ln(2\pi) \\ & -\beta \sum_\mu \eta_\mu^2 - \gamma \sum_j (w_j^0)^2 + \tfrac{1}{2}\sum_{jk} \rho_j \Lambda_{jk} \rho_k . \end{aligned} \tag{2.10}$$

At this point there is a choice to be made. One may proceed to consider the 'optimal' parameters which follow from maximizing the evidence for a *specific* realization of the data $\mathcal{D}$, deferring averages over $\mathcal{D}$. Alternatively (the route we shall actually take here) one may proceed directly to consider the ensemble average of the log-evidence, which may be thought of as defining a free-energy density of a system with 'quenched' data variables:

$$f = -\left\langle\left\langle \frac{1}{N} \ln P(\mathcal{D}|\gamma, \beta) \right\rangle\right\rangle. \tag{2.11}$$

The quenched average $\langle\langle\cdot\rangle\rangle$ extends over the ensemble of input data, and noise. The averaging may be done by adapting the analysis of Hertz *et al* [13] to show that

$$\langle\langle \Lambda_{jk} \rangle\rangle = (2\beta\sigma_x^2)^{-1} G(\lambda, \alpha)\delta_{jk} \tag{2.12}$$

where

$$G(\lambda, \alpha) = \frac{1 - \alpha - \lambda + \sqrt{(\lambda + \alpha - 1)^2 + 4\lambda}}{2\lambda} \tag{2.13}$$

with $\alpha \equiv p/N$ and

$$\lambda \equiv \frac{\gamma}{\beta\sigma_x^2}. \tag{2.14}$$

Using this result in conjunction with (2.11) and (2.10) we then obtain

$$f = \frac{1}{2}\left(\overline{\gamma} - \overline{\beta}\right)(1 - \lambda G(\lambda, \alpha)) + \frac{\alpha}{2}\left(\overline{\beta} - \ln\overline{\beta} - \ln(2\pi\sigma^2)\right) - \frac{1}{2}\int_\lambda^\infty \mathrm{d}\lambda' \left(G(\lambda', \alpha) - \frac{1}{\lambda'}\right) \tag{2.15}$$

where the last term comes from solving a differential equation for $\langle\langle \ln \det \Lambda \rangle\rangle$. The new parameters $\overline{\gamma}$ and $\overline{\beta}$ are scaled versions of the original hypothesis parameters:

$$\overline{\gamma} \equiv 2\sigma_w^2\gamma \qquad \overline{\beta} \equiv 2\sigma^2\beta$$

while

$$\sigma_w^2 \equiv \frac{1}{N}\sum_j (w_j^0)^2 \tag{2.16}$$

gives the variance of the elements of the true rule vector $\boldsymbol{w}^0$.

Equation (2.15) completes the calculation of the evidence measure; we shall explore its structure in the following section, with particular emphasis on the MAP parameter estimates defined by its extrema. We will also be concerned with the implications of these MAP assignments for performance measures, which we proceed to identify.

The most widely used performance measure is the *generalization error* ϵ_g defined by

$$\sigma^2\epsilon_g \equiv \left\langle\left\langle \left[\langle \mathcal{R}^0(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{x})\rangle\right]^2 \right\rangle\right\rangle. \tag{2.17}$$

The inner average $\langle\cdot\rangle$ extends over the ensemble of model rules for given data (characterized by the distribution (2.5)); the outer average $\langle\langle\cdot\rangle\rangle$ extends over the quenched variables (the data and the noise). Recognising that (cf equations (2.8)–(2.10))

$$\langle r_j \rangle = \frac{\partial P(\mathcal{D}|\gamma, \beta)}{\partial \rho_j} = \sum_k \Lambda_{jk}\rho_k \tag{2.18}$$

we may write

$$\begin{aligned}\sigma^2\epsilon_g &\equiv \frac{1}{N}\left\langle\left\langle \left[\left\langle \sum_{j=1}^N r_j x_j \right\rangle\right]^2 \right\rangle\right\rangle \\ &= \frac{\sigma_x^2}{N}\sum_{j=1}^N \left\langle\left\langle \left[\sum_k \Lambda_{jk}\rho_k\right]^2 \right\rangle\right\rangle\end{aligned}$$

which may be evaluated using (2.12) to give

$$\epsilon_{\rm g} = G(\lambda,\alpha) + \lambda G'(\lambda,\alpha)\left(1 - \frac{\lambda}{\nu}\right) \tag{2.19}$$

where $G'(\lambda,\alpha)$ is the derivative of the function G with respect to λ, and

$$\nu \equiv \frac{\sigma^2}{\sigma_x^2 \sigma_w^2} \tag{2.20}$$

provides a measure of the ratio of noise to signal associated with a typical element of the input vector. Equation (2.19) recovers a result of Krogh and Hertz [7].

The generalization error measures only one aspect of the effectiveness of the modelling procedure. It gives a measure of the typical (mean-square) difference between the output of the *mean* model rule for given fixed data and the output of the true rule, for a typical novel input. A small value of $\epsilon_{\rm g}$ implies that the model makes accurate predictions. The *variance* of the model output for given fixed data and test input provides the ingredients for a second performance measure. It is desirable that the confidence limits on the model output, implied by this variance, should correspond as closely as possible to the true error, whose average is $\epsilon_{\rm g}$. With this motivation we define a measure of the generalization *consistency*

$$\sigma^2 \delta_g \equiv \langle\langle\ \langle [\mathcal{R}(\boldsymbol{x}) - \langle \mathcal{R}(\boldsymbol{x})\rangle]^2 \rangle\ \rangle\rangle - \sigma^2 \epsilon_{\rm g}\,. \tag{2.21}$$

A small value of $\delta_{\rm g}$ implies that the model predicts its own errors accurately. In the present context we find that

$$\begin{aligned}
\langle\langle\ \langle [\mathcal{R}(\boldsymbol{x}) - \langle \mathcal{R}(\boldsymbol{x})\rangle]^2 \rangle\ \rangle\rangle &\equiv \frac{1}{N}\left\langle\!\!\left\langle\ \left\langle \left[\sum_{j=1}^{N} (r_j - \langle r_j \rangle) x_j\right]^2 \right\rangle\ \right\rangle\!\!\right\rangle \\
&= \frac{\sigma_x^2}{N} \sum_{j=1}^{N} \langle\langle \Lambda_{jj} \rangle\rangle \\
&= \frac{\sigma^2}{\overline{\beta}} G(\lambda,\alpha)
\end{aligned}$$

so that

$$\delta_{\rm g} = \frac{G(\lambda,\alpha)}{\overline{\beta}} - \epsilon_{\rm g}\,. \tag{2.22}$$

While appropriate for our own purposes, (2.17) and (2.21) do not exhaust the possible forms of performance measure. We note two others in particular. First, Hansen [14] defines a 'generalization error' as a typical (mean-square) difference between mean model rule output and *corrupted* rule output. This quantity is an amalgam of $\epsilon_{\rm g}$ and $\delta_{\rm g}$ defined here. Second, Levin *et al* [15] define a 'prediction error' as a measure of the *likelihood* one would assign to an observed novel corrupted output, given the rule distribution implied by the data. It is straightforward to show that this quantity is simply related to the derivative of the log-evidence with respect to the number p of members of the defining data set.

3. Analysis and discussion

In this section we explore the implications of using the evidence extrema (the minima of the free energy f, equation (2.15)) to set the values of the hypotheses parameters β and γ—in particular, the consequences for the performance measures $\epsilon_{\rm g}$ and $\delta_{\rm g}$ (equations (2.19), (2.22)). We divide our discussion into two parts. First we examine the $\gamma \to 0$ limit, which provides a simple illustration of a phase transition in the space of hypothesis parameters. The $\gamma \neq 0$ case is more complex, but provides insights into the phenomena which may occur when the model is not well matched to the underlying rule.

3.1. A special case: the $\gamma \to 0$ limit

The $\gamma \to 0$ limit of (2.15) has to be taken with some care, since the 'weight-decay' term in the cost function (2.9) is needed to regularize the integral in (2.8), when $p < N$ (i.e. when $\alpha < 1$). We find

$$\lim_{\gamma\to 0} f = \begin{cases} \dfrac{(\alpha-1)}{2}[\overline{\beta} - \ln\overline{\beta}] - \dfrac{1}{2}\ln\overline{\gamma} + f_0(\alpha) + \mathrm{O}(\gamma) & \alpha > 1 \\ -\dfrac{\alpha}{2}\ln\overline{\gamma} + f_0(\alpha) + \mathrm{O}(\gamma) & \alpha < 1 \end{cases} \tag{3.1}$$

where the unspecified function $f_0(\alpha)$ is independent of $\overline{\beta}$. The result for the regime $\alpha > 1$ (where the regularizer is unnecessary) can be found in the work of Levin *et al* [15]. In the regime $\alpha < 1$ the leading contributions to the free energy are independent of $\overline{\beta}$: in this regime the data provides no 'evidence' to guide the choice of this hypothesis parameter. In the regime $\alpha > 1$ the free energy is minimized (the evidence maximized) by the choice $\overline{\beta} = 1$. The point $\alpha = 1$ locates a phase transition between these two regions of hypothesis space. The approach to the phase transition as $\alpha \to 1^+$ is signalled by a divergence of the 'susceptibility'

$$\chi_\beta \equiv \left(\frac{\partial^2 f}{\partial\beta^2}\right)^{-1} = \frac{2\beta^2}{\alpha - 1} \tag{3.2}$$

and by fluctuations in the optimal value of the hypothesis parameter located by the extrema of the $\mathcal{D}$-dependent evidence (2.10), which we shall report on elsewhere. The singular behaviour at $\alpha = 1$ also manifests itself in the performance measures. From (2.19) we obtain, in accord with Krogh and Hertz,

$$\lim_{\gamma\to 0} \epsilon_{\mathrm{g}} = \begin{cases} \dfrac{1}{\alpha - 1} + \mathrm{O}(\gamma) & \alpha > 1 \\ \dfrac{\alpha}{1-\alpha} + \dfrac{1-\alpha}{\nu} + \mathrm{O}(\gamma) & \alpha < 1 \end{cases} \tag{3.3}$$

while from (2.22) we find

$$\lim_{\gamma\to 0} \delta_{\mathrm{g}} = \begin{cases} \dfrac{1-\overline{\beta}}{\overline{\beta}(\alpha - 1)} + \mathrm{O}(\gamma) & \alpha > 1 \\ \dfrac{1-\alpha}{\overline{\gamma}\nu}(1 + \mathrm{O}(\gamma)) & \alpha < 1 \end{cases} . \tag{3.4}$$

These results show that, without equivocation in this case, the evidence provides a good guide to parameter choice. In the regime $\alpha > 1$ the data (through the evidence) identifies the value $\overline{\beta} = 1$ which optimizes (sets to zero) the consistency measure δ_{g}. In the regime $\alpha < 1$ both performance measures ϵ_{g} and δ_{g} are indifferent to the choice of $\overline{\beta}$ mirroring the behaviour of the evidence. This behaviour reflects the fact that, in this regime, the data is insufficient to anchor the model rule in the vicinity of the true rule: the consequent large fluctuations in rule space are controlled predominantly by the prior on the weights.

3.2. The general case: $\gamma \neq 0$

It is straightforward to show analytically that the free energy (2.15) has a turning point in the $\overline{\gamma}$, $\overline{\beta}$ plane at $\overline{\gamma} = \overline{\beta} = 1$. Numerical studies indicate (we have not proved it generally) that this turning point is invariably a global minimum. This is to be expected. The condition $\overline{\beta} = 1$ corresponds to a prior assignment ($\beta = 1/2\sigma^2$) which accords to the observed data the appropriate degree of confidence, given the magnitude of the noise; the condition $\overline{\gamma} = 1$ ($\gamma = 2\sigma_w^2$) tunes the typical member of the rule parameter set $\{w\}$ to the typical size of the elements of the true rule vector w^{o}.

The generalization error ϵ_{g} (2.19) is actually sensitive only to the *ratio* of the two hypothesis parameters, through its dependence on $\lambda = \nu\overline{\gamma}/\overline{\beta}$. It is easily shown that ϵ_{g} is minimal along the *line* $\overline{\gamma} = \overline{\beta}$, where $\lambda = \nu$. By inspection of (2.19) and (2.22) one sees that, on this line, δ_{g} vanishes (is optimized) with the choice $\overline{\beta} = 1$. Thus the evidence-based assignments optimize the performance measures. This is also eminently reasonable. Indeed, quite generally we must expect that, *if* the hypothesis is well-matched to the world (the hypothesis space contains the 'truth' about the data), the evidence framework will locate the 'truth' and will presumably optimize all reasonable performance measures.

It is, however, of rather more interest to know how the evidence framework functions when the hypothesis is *not* well-matched to the world. This situation is realized in a primitive way in the present framework when one of the hypothesis parameters is clamped at a sub-optimal value. We identify two potentially generic features of this situation.

First, *optimizing the evidence ovet an inadequate hypothesis space does not guarantee optimal performance measures.* Consider, for example, the evidence-optimized parameter $\overline{\beta}_f(\overline{\gamma})$, defined by the solution to the equation $\partial f/\partial\overline{\beta}|_{\overline{\gamma}} = 0$. It is easy to show from (2.15) that

$$\overline{\beta}_f = \overline{\gamma} + (1-\overline{\gamma})\left[1 - \frac{\lambda}{\alpha}(1+G)(G+\lambda G')\right]^{-1} . \qquad (3.5)$$

This form captures the 'truth' ($\overline{\beta} = 1$) if $\overline{\gamma} = 0$, when there is no bias at all from the weight decay term; if $\overline{\gamma} = 1$, when there is no false bias from the weight decay term; and if $\alpha \to \infty$ or $\mu \to 0$, when the data contains enough information about the corrupting noise to overwhelm any bias from the weight decay term. In general, however, $\overline{\beta}_f$ depends upon the value assigned to $\overline{\gamma}$. Significantly (as one finds, already, if one examines the $\mathrm{O}(\gamma)$ corrections to the leading behaviour recorded in (3.1), (3.3) and (3.4)) the assignment $\overline{\beta} = \overline{\beta}_f(\overline{\gamma})$ does *not* optimize the performance measures ϵ_{g} and δ_{g}. Equation (2.19) shows that ϵ_{g} is optimized for a given $\overline{\gamma}$ by the assignment $\overline{\beta} = \overline{\beta}_\epsilon = \overline{\gamma}$. The optimization of δ_{g} (2.22) provides a criterion for identifying a further 'optimal' parameter, $\overline{\beta}_\delta$. The $\overline{\gamma}$-dependence of the three 'optimal' parameters is displayed in figure 1, for the case $\alpha = 1.5$. It is clear that the parameter choice producing the most 'effective' rule performance depends upon the performance measure; and the *likeliest* parameter values identified by the evidence framework provide little guidance as to which parameter assignments will best compensate for the deficiencies of the hypothesis space.

The second key feature to be noted is that *regions of hypothesis space remote from the truth may have sufficient structure to support phase transitions.* Figure 2 provides a simple example: it shows the *difference* $\Delta f = f_1 - f_2$ between the free energies associated with two sub-optimal hypotheses, plotted as a function of the effective number of examples $\alpha = p/N$. The two hypotheses are characterized by values $\overline{\gamma} = 0.1$, $\overline{\beta} = 1$ and $\overline{\gamma} = 0.1$, $\overline{\beta} = 1.5$, respectively. Evidently the first hypothesis is to be 'preferred' (in the sense of being more likely) for large enough α, where the data is sufficient to overwhelm the false bias resulting

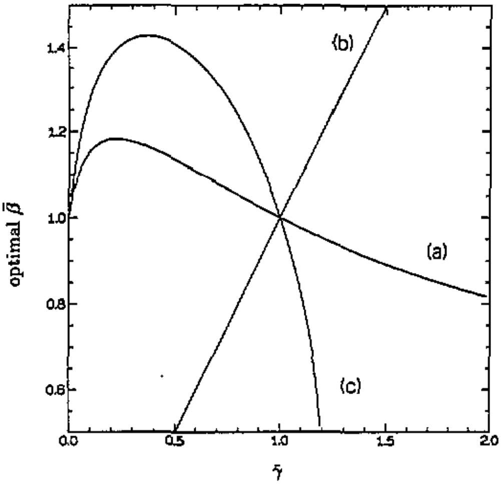

Figure 1. The 'optimal' values of the hypothesis parameter $\overline{\beta}$ (for $\alpha = 1.5$), as a function of the hypothesis parameter $\overline{\gamma}$, identified on the basis of three different criteria:(*a*) minimization of the free energy f (giving $\overline{\beta}_f$); (*b*) mininization of the generalization error ϵ_g (giving $\overline{\beta}_\epsilon$) and (*c*) minimization of the magnitude of the consistency measure δ_g (giving $\overline{\beta}_\delta$). The three results coincide only at $\overline{\gamma} = 1$.

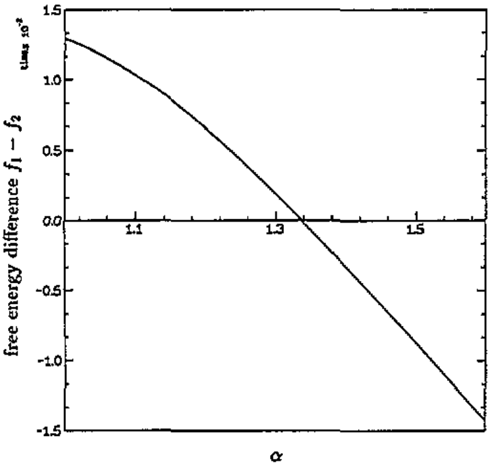

Figure 2. The difference between the free energies f_1 and f_2 associated with two 'hypotheses', characterized by assignments $\overline{\gamma} = 0.1, \overline{\beta} = 1$ and $\overline{\gamma} = 0.1, \overline{\beta} = 1.5$ respectively. The point $\alpha_0 \simeq 1.34$ locates a boundary between a 'phase' in which the first hypothesis is favoured ($\alpha > \alpha_0$) and one in which the second hypothesis is favoured.

from the assignment of the weight decay parameter; but the second hypothesis is favoured at lower values of α. The crossing of the evidence surfaces occurring at intermediate α values locates a transition between sectors, or 'phases', of parameter space favouring the two hypotheses.

4. Summary

In this paper we have explored the general issue of hypothesis-parameter assignment, and its consequences for performance measures, in the context of the noisy linear perceptron. We have seen that evidence-based MAP assignments of hypothesis parameters do not, in general, optimize performance measures: the *most likely* parameter values, given a less-than-perfect prior will rarely be the *most effective*. We have also seen two instances of phase transitions in the space of hypothesis parameters, driven by a variation of the number of members of the data set. We anticipate that such occurrences may prove to be a general feature of hypothesis-parameter assignment, reflecting the shift in the relative influence of model rule error ('energy') and model rule flexibility ('entropy'), with the growth of the information content of the data set.

Acknowledgment

The work of one author (DS) is supported by SERC grant No GRF 79719.

References

[1] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)

[2] Geman S and Geman D 1984 *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-6** 721–41

[3] Gull S F 1988 *Maximum Entropy and Bayesian Methods* ed J Skilling (Cambridge: Cambridge University Press) 53–71; *Maximum Entropy and Bayesian Methods in Science and Engineering—vol 1: Foundations* ed G J Erickson and C R Smith (Dordrecht: Kluwer) 53–74

[4] Skilling J 1993 Physics and Probability ed W T Grandy Jr and P Milonni (Cambridge: Cambridge University Press)

[5] MacKay D J C 1992 *Neural Computation* **4** 415–47

[6] MacKay D J C 1992 *Neural Computation* **5** 698–714

[7] Krogh A and Hertz J A 1992 *J. Phys. A: Math. Gen.* **25** 1135–47

[8] Watkin T L H 1993 *Europhys. Lett.* **21** 871–6

[9] Seung H S, Sompolinsky H and Tishby N 1993 *Phys. Rev.* A **45** 6056–91

[10] Dunmur A P and Wallace D J 1993 *J. Phys. A: Math. Gen.* **26** 5767-79

[11] Neal R M 1992 *Technical Report* CRG-TR-92-1 University of Toronto, Dept of Computer Science

[12] Papoulis A 1986 *Probability, Random Variables and Stochastic Processes* 2nd edn (Singapore: McGraw Hill)

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

[14] Hansen L K 1993 *Neural Networks* **6** 393–97

[15] Levin E, Tishby N and Solla S 1989 *Proc 2nd Workshop on Computational Learning Theory* (San Mateo: Morgan Kaufmann)