Replica-symmetry breaking in noise-optimal neural networks



1995 J. Phys. A: Math. Gen. 28 7105

(http://iopscience.iop.org/0305-4470/28/24/011)

J. Phys. A: Math. Gen. 28 (1995) 7105–7111. Printed in the UK

# Replica-symmetry breaking in noise-optimal neural networks

W Whyte†, D Sherrington† and K Y M Wong‡

† Theoretical Physics, Department of Physics, 1 Keble Road, Oxford OX1 3NP, UK

‡ Department of Physics, The Hong Kong University Of Science And Technology, Clear Water Bay, Kowloon, Hong Kong

Received 9 May 1995

**Abstract.** Recent studies of optimization in neural networks trained with noisy data have shown that replica-symmetric solutions are unstable in the low-noise region of parameter space. We calculate the 1-step replica-symmetry broken solution in this region, which joins the replica-symmetric solution continuously at the de Almeida–Thouless line. These solutions yield satisfactory agreement with simulations for the aligning field distribution, better than those given by the replica-symmetric ansatz.

## 1. Introduction

An important problem in the theory of neural networks concerns the determination of optimal behaviour for reproducing correct output for example input data. It can be formulated in terms of the minimization of an appropriate cost function and studied analytically by formulating a statistical mechanical analogue of annealing in which the cost function plays the role of the Hamiltonian, the temperature is a controllable measure of stochasticity, and the free energy is minimized and its zero-temperature limit taken. Replica theory is normally employed to effect such a study averaged over sample data. Wong and Sherrington (WS) [1] have recently employed such a procedure within the replica-symmetric ansatz (RS) to extend the analysis to optimize systems trained with noisy data. The RS ansatz is, however, unstable against small replica-symmetry breaking (RSB) fluctuations as the noise is decreased beneath a critical de Almeida–Thouless (AT) line [2]. Furthermore, within RS theory, in the region of noise beneath a critical line, itself beneath this AT line, one finds a bandgap in the aligning field distribution, which has also been shown recently to imply RSB [3].

This paper extends the analysis of WS to allow for 1-step RSB, analogous to the procedure used for noise-free systems beyond capacity by Erichsen and Theumann [4] and by Majer *et al* [5]. RSB is found everywhere within the region of sufficiently low noise, disappearing continuously at the AT line. 1-step RSB reduces the gap in the aligning field distribution in accord with simulations.

## 2. Model

The system we are interested in is a perceptron with $N$ input nodes (labelled $i = 1, \ldots, N$) and one output node, obeying the update rule

$$S_{\text{output}} = \text{sign}\left(\sum_i J_i S_i\right) \qquad S_i \in \{\pm 1\} \tag{1}$$

0305-4470/95/247105+07$19.50 

with the $\{J_i\}$ constrained by the spherical rule $\sum_i J_i^2 = N$. This perceptron is trained to store correctly as many as possible of an ensemble of $Qp$ noisy examples $\{\eta^{\mu\nu}\}(\mu = 1, \ldots, Q,\ \nu = 1, \ldots, p)$, which are generated from a corresponding set of $p = \alpha N$ clean patterns $\{\xi^\mu\}$ by the rule

$$P(\eta_i^{\mu\nu}) = (1 + \eta_i^{\mu\nu}\xi_i^\mu m_t)/2 \tag{2}$$

where $m_t$ is known as the 'training overlap'; correspondingly the training noise is the complement $d_t \equiv \frac{1}{2}(1 - m_t)$.

The clean patterns $\{\xi^\mu\}$ are drawn at random from $\{\pm 1\}^N$. We define the 'aligning field' $\lambda^{\mu\nu}$ of an example $\eta^{\mu\nu}$ by

$$\lambda^{\mu\nu} = \frac{\xi^\mu_{\text{output}}}{|J|} \sum_i J_i \eta_i^{\mu\nu}. \tag{3}$$

The stability of a clean pattern is defined analogously, $\lambda^\mu = \xi^\mu_{\text{output}} \sum_i J_i \xi_i^\mu / |J|$. Given the aim of storing as many as possible of these examples, our cost function to be minimized is simply

$$E = -\sum_{\mu\nu} \text{sign}(\lambda^{\mu\nu}). \tag{4}$$

In the limit $Q \to \infty$, we can average over the quenched cost function above to obtain an annealed cost function defined only with respect to the stabilities of the clean patterns:

$$E = \sum_\mu g(\lambda^\mu) \qquad g(\lambda) = -\text{erf}\left(\frac{m_t \lambda}{\sqrt{2[1 - m_t^2]}}\right). \tag{5}$$

This is the function which we attempt to minimize in the following calculations.

## 3. Theory

We now attempt to calculate the minimum value of the cost function defined above by considering the free energy of the system,

$$\begin{aligned} f(\xi) &= -\lim_{N\to\infty} \frac{1}{N\beta} \log Z \\ &= -\lim_{N\to\infty} \frac{1}{N\beta} \log \int \prod_{i=1}^N \mathrm{d}J_i\, \delta((J)^2 - N) \mathrm{e}^{-\beta E}. \end{aligned} \tag{6}$$

Taking the limit $\beta \to \infty$ gives the minimum cost. Our interest is in the average of this minimum cost over pattern choices, $\bar{E} \equiv \langle \min E \rangle = \lim_{\beta\to\infty} \langle f \rangle$, where $\langle\ \rangle$ denotes the average over the patterns. To perform the average over the patterns with the distribution given above we use the replica trick, $\langle \log Z \rangle = \lim_{n\to 0}(\langle Z^n \rangle - 1)/n$. Using standard techniques [6, 5] one can express $\langle Z^n \rangle$ in the form

$$\langle Z^n \rangle = \int \prod_{\alpha\beta} \mathrm{d}q_{\alpha\beta}\, \mathrm{e}^{N\Phi(q_{\alpha\beta})} \tag{7}$$

where $q_{\alpha\beta} \equiv \frac{1}{N} \sum_i J_i^\alpha J_i^\beta$ is a measure of the similarity of two replicated networks that both optimize the cost function.

In order to progress from here, we have to make an ansatz about the form of the replica solution. In the replica-symmetric ansatz, we take $q_{\alpha\beta} = q \,\forall\, \alpha \neq \beta$. This ansatz becomes unstable at the AT line given by [1]

$$\alpha \int \mathrm{D}z \left(1 - \frac{\partial \lambda_0}{\partial z}\right)^2 = 1 \tag{8}$$

where $\mathrm{D}z \equiv \mathrm{d}z \exp(-z^2/2)/\sqrt{2\pi}$ is the Gaussian measure, and $\lambda_0$ minimizes the function $g(\lambda_0)+(\lambda_0-z)^2/2x$ for given values of $z$ and $x$, $x$ itself being determined by a saddle-point equation in [1]. It will asymptotically approach $m_t \sim 0.84$ as $\alpha \to \infty$. When $\alpha$ or $m_t$ is larger than their values prescribed by (8), the replica-symmetric solution becomes unstable.

The simplest non-replica-symmetric ansatz for the solution of equations (6), (7) is given by 1-step replica-symmetry breaking (1-step RSB or RSB1) in which we assume that the matrix $q_{\alpha\beta}$ has a block structure, with blocks of size $m \times m$ such that the diagonal blocks have 1 in their diagonal entries and $q_1$ on their off-diagonal, while the off-diagonal blocks have $q_0$ in all their entries. In other words:

$$\begin{aligned} q_{\alpha\beta} &= 1 \qquad && \alpha = \beta \\ q_{\alpha\beta} &= q_1 && \text{if the integer parts of } \alpha/m \text{ and } \beta/m \text{ are the same} \\ q_{\alpha\beta} &= q_0 && \text{otherwise.} \end{aligned} \tag{9}$$

Majer *et al* [5] have shown that in the RSB1 regime, the minimum free energy for any cost function $g(\lambda)$ is

$$\langle f \rangle = \lim_{\beta\to\infty} \min_{x,q_0,w} \left[ \frac{q_0}{2x(1+w\Delta q)} + \frac{\log(1+w\Delta q)}{2wx} + \frac{\alpha}{wx} \int \mathrm{D}z_0 \log \int \mathrm{D}z_1 \exp\left(-wx\left[g(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right)\right] \tag{10}$$

where $w = m/(1-q_1)$ and $\lambda_0$ minimizes the final square bracket for given values of $z_0, z_1, q_0, \Delta q \equiv (1-q_0)$ and $x \equiv \beta(1-q_1)$. Denoting $z = z_0\sqrt{q_0} + z_1\sqrt{\Delta q}$, this minimization requires us to invert the function $z(\lambda) = xg'(\lambda) + \lambda$. For high values of $(x, m_t)$, this inverse function is multiple-valued and we are required to perform a Maxwell construction in order to make it single-valued. This, in turn, gives a discontinuity in $\lambda_0(z)$ which will be reflected in a gap in the distribution of aligning fields, $\rho(\lambda)$. Having found $\lambda_0(z)$ we are left with a formula for $f$ that must be minimized numerically with respect to $x, q_0, w$.

As well as the free energy of the system, we are interested in the distribution of aligning fields, $\rho(\lambda)$. This is the relative volume of solution space with aligning field $\lambda$:

$$\rho(\lambda) = \lim_{\beta\to\infty} \frac{1}{Z} \int \prod_{i=1}^{N} \mathrm{d}J_i \, \delta((\boldsymbol{J})^2 - N) \mathrm{e}^{-\beta \sum_\mu g(\lambda^\mu)} \delta(\lambda - \lambda^\nu). \tag{11}$$

It too can be calculated by the replica method. The RSB1 solution for the aligning field distribution is

$$\rho(\lambda) = \int \mathrm{D}z_0 \frac{\int \mathrm{D}z_1 \exp\left(-wx\left[g(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right) \delta(\lambda - \lambda_0)}{\int \mathrm{D}z_1 \exp\left(-wx\left[g(\lambda_0) + \frac{(\lambda_0 - z_0\sqrt{q_0} - z_1\sqrt{\Delta q})^2}{2x}\right]\right)}. \tag{12}$$

The values of the parameters are those obtained by minimization of (10).

The third quantity of interest is the optimal output overlap $f_{m_t}(m_{in})$ with a stored pattern when the input overlap is $m_{in}$ and the network has been trained with an overlap $m_t$. This has been shown [1] to be

$$f_{m_t}(m_{in}) = -\int d\lambda\, \rho_{m_t}(\lambda) g_{m_{in}}(\lambda) \tag{13}$$

where $g_{m_{in}}$ is given by equation (5) with $m_t$ replaced by $m_{in}$. When $m_{in} = m_t$ the optimal output overlap becomes equal to $\tilde{E}$, the averaged minimum of the cost functions, and we refer to it as the 'performance' of the network.

## 4. Results

We have evaluated the average minimum cost at several values of $\alpha$ and $m_t$, on both sides of the AT line; note that asymptotically, as $\alpha \to \infty$, the critical AT value of $m_t \sim 0.84$, but it goes to zero as $\alpha \to 0$. Results for $q_0(m_t)$ are exhibited in figure 1 for several values of $\alpha$, along with the corresponding AT values. They clearly demonstrate that replica symmetry is broken for noise values less than the AT values and that its onset is continuous as the line is crossed from the noisier side. Figure 2 shows results for $q_0(\alpha)$ as a function of $\alpha$ for three values of $m_t$. As expected, for $m_t = 0.8$ there is no RSB1 until $\alpha$ is reduced beneath the corresponding AT value, $\alpha_{AT} = 2.80$, whereas for $m_t = 0.99$ and 0.85, replica-symmetry is broken for all values of $\alpha$. In all the cases shown $q_0$ increases as $\alpha$ is reduced towards zero, in accord with the expectation that all optimal networks become Hebb-like for a finite number of patterns, and replica symmetry can at most be marginally broken [8]. In the case of $m_t = 0.99$, $q_0$ approaches 1 as $\alpha$ is reduced to order 2, reflecting the fact that the $m_t = 1$ limit corresponds to using the Gardner–Derrida [7] minimal stability cost function which is replica symmetric for $\alpha < 2$. The decrease of $q_0$ as $\alpha$ is reduced from high values to order 1, seen for $m_t = 0.85$ and for $m_t = 0.8$ below the AT value, is a reflection of the migration of these $m_t$ values further from the (decreasing) AT line and deeper into the RSB region.

Next we examine the effects of 1-step RSB on the local aligning field distribution $\rho(\lambda)$. For this we take two points, $\alpha = 1.5$, $m_t = 0.9$ and $\alpha = 4$, $m_t = 0.85$, both of which are in

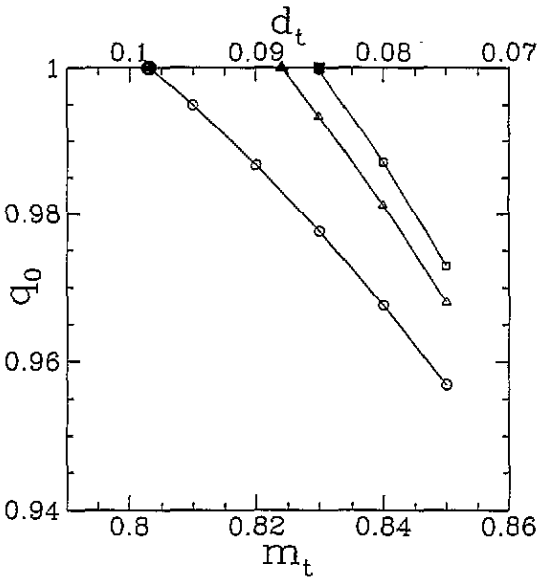


**Figure 1.** $q_0$, the measure of replica-symmetry breaking, as a function of $m_t$ for $\alpha = 3$ (circles), 6.5 (triangles), 10 (squares). The open circles are results of numerical minimization of (9); the full circles are the calculated AT values. The curves are to guide the eye.

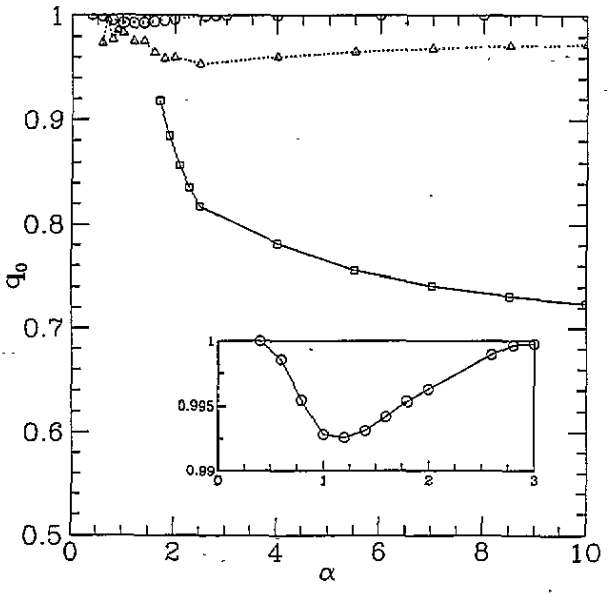


**Figure 2.** $q_0$ as a function of $\alpha$ for $m_t = 0.99$ (squares), 0.85 (triangles), 0.8 (circles). *The curves are to guide the eye.* The inset shows the $m_t = 0.8$ result in more detail.

the RSB1 region but which under RS, respectively, do and do not have a gap in $\rho(\lambda)$. The results for $\alpha = 1.5, m_t = 0.9$ under RS and RSB1 are shown in figure 3. As observed in the corresponding clean-system study of Majer *et al* [5], RSB1 has a marked effect on the aligning field distributions; in particular, the bandgap is narrowed significantly.

These results can be compared with those obtained by direct simulational training on the annealed cost function for a realised set of patterns. The aligning field distributions displayed in figure 4 were obtained on networks of 200, 400 and 800 neurons for $\alpha = 1.5, m_t = 0.9$. A simple gradient descent algorithm was used to minimize the cost function $E = -\sum_\mu \mathrm{erf}(m_t\lambda^\mu/\sqrt{2[1-m_t^2]})$. As can be seen by comparison with figure 3, RSB1 improves the agreement of the theoretical and experimental curves.

In the region where RSB exists but there is no gap in the aligning field distribution,

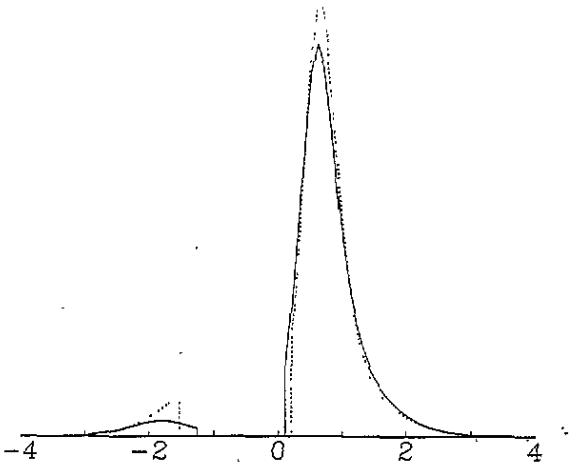


**Figure 3.** The distribution of aligning fields, $\rho(\lambda)$, for $m_t = 0.9$, $\alpha = 1.5$ for the replica-symmetric (dotted curve) and RSB1 (full curve) ansätze.

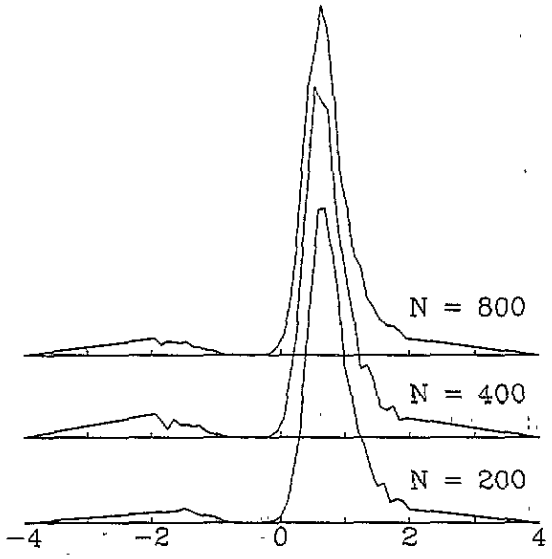


**Figure 4.** The aligning field distribution, $\rho(\lambda)$, obtained for $m_t = 0.9$, $\alpha = 1.5$ on networks of 200, 400 and 800 neurons by gradient descent on the cost function $g(\lambda)$.

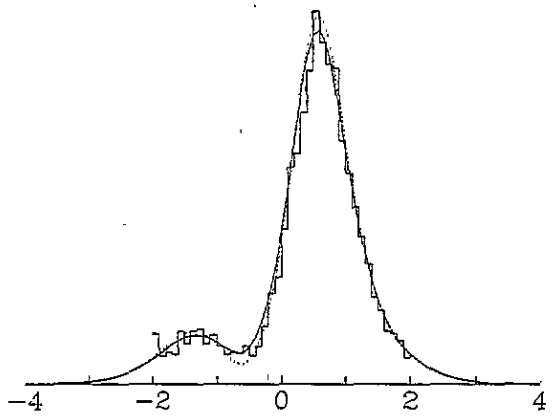


**Figure 5.** The distribution of aligning fields, $\rho(\lambda)$, for $m_t = 0.85$, $\alpha = 4$ for the replica-symmetric (dotted curve) and RSB1 (full curve) ansätze, and as obtained by direct simulational training on a network of 200 neurons (histogram represents average over ten training runs.

the effect of moving from RS to RSB1 on the distribution is to raise the minimum in the pseudogap, which brings the results shown in figure 5 for the case $\alpha = 4, m_t = 0.85$ more into agreement with those obtained from simulations.

Since RSB occurs when the replica-symmetric ansatz no longer gives a global minimum for $\tilde{E}$, its effect will be to decrease the performance of the network. Figure 6 compares the performances under the RS and RSB1 ansätze for the three values of $m_t$ used above. For $m_t = 0.8$, 0.85 the effect of RSB is small, as might have been expected from the fact that $q_0$ is very near 1 in the RSB1 regime for these $m_t$ values. For $m_t = 0.99$, however, the effect of RSB1 is to cause a marked decrease in the performance, particularly at high values of $\alpha$.

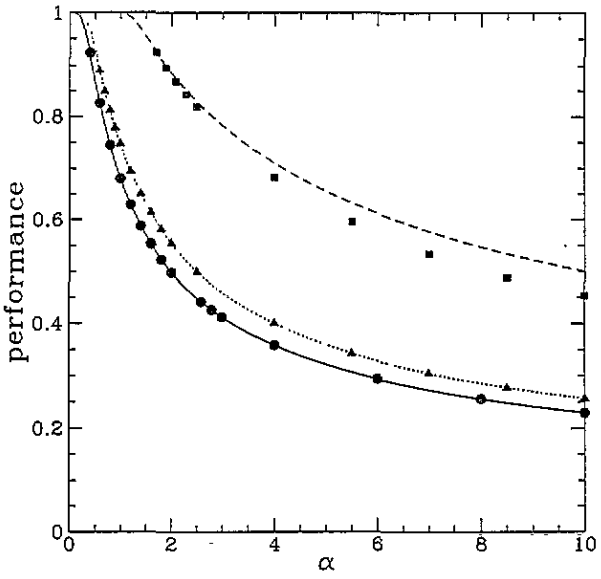


**Figure 6.** The performance overlaps obtained as a function of $\alpha$ using RS (curves) and RSB1 (points) for $m_t = 0.8$ (full curve, circles), $m_t = 0.85$ (dotted curve, triangles) and $m_t = 0.99$ (broken curve, squares).

## Acknowledgments

The authors would like to thank the British Council for providing a scholarship for one of them (WW) and a grant (UK/HK Joint Research Scheme) to enable collaboration between the UK and Hong Kong.

## References

[1] Wong K Y M and Sherrington D 1993 *Phys. Rev.* E **47** 4465; 1994 *Phys. Rev.* E **50** 1727
[2] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
[3] Bouten M 1994 *J. Phys. A: Math. Gen.* **27** 6021
[4] Erichsen R Jr and Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** L61
[5] Majer P, Engel A and Zippelius A 1993 *J. Phys. A: Math. Gen.* **26** 7405
[6] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659
[7] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[8] Wong K Y M, Rau A and Sherrington D 1992 *Europhys. Lett.* **19** 559