Threshold-induced phase transitions in perceptrons

1997 J. Phys. A: Math. Gen. 30 3471

(http://iopscience.iop.org/0305-4470/30/10/023)

J. Phys. A: Math. Gen. **30** (1997) 3471–3496. Printed in the UK PII: S0305-4470(97)75406-4

Threshold-induced phase transitions in perceptrons

Ansgar H L West†‡§ and David Saad‡

† Department of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK
‡ Neural Computing Research Group, Aston University, Aston Triangle, Birmingham B4 7ET, UK

Received 13 June 1996

Abstract. Error rates of a Boolean perceptron with threshold and either spherical or Ising constraint on the weight vector are calculated for storing patterns from biased input and output distributions derived within a one-step replica symmetry breaking (RSB) treatment. For unbiased output distribution and non-zero stability of the patterns, we find a critical load, α_p, above which two solutions to the saddlepoint equations appear; one with higher free energy and zero threshold and a dominant solution with non-zero threshold. We examine this second-order phase transition and the dependence of α_p on the required pattern stability, κ, for both one-step RSB and replica symmetry (RS) in the spherical case and for one-step RSB in the Ising case.

1. Introduction

Since the ground-breaking work of Gardner [1] on the storage capacity of the Boolean perceptron, the replica technique of statistical mechanics has been successfully employed to investigate many aspects of the performance of simple neural network models. While most of the research concentrated on exploring the learning ability and network capacity below saturation (for a review see [2, 3] and references therein), we will concentrate in this paper on the errors of a Boolean perceptron above its saturation limit, or capacity limit α_c, working within a replica framework. Earlier studies [4–6] have particularly examined the cases of zero stability of the stored patterns, the effect of different error functions on the error rates, and the distribution of pattern stabilities. Here, we will extend this work by allowing for a threshold and biased input and output distributions and investigate both real-valued (spherical constraint) and binary weights (Ising constraint).

We find that the Boolean perceptron with threshold has a rich behaviour reflecting the extra degree of freedom introduced by the threshold. In the case of arbitrary input and output distributions we find that the threshold can always compensate for a ferromagnetic bias in the weights but not vice versa, which will allow us to argue that the paradigm of eliminating the threshold in favour of a ferromagnetic bias in the weights, which has been adopted in some papers (e.g. [1, 7]), should be reconsidered. The introduction of a threshold enables the elimination of the input distribution bias, by suitably rescaling the threshold and stability.

Especially intriguing is the role of the threshold for non-zero stability and unbiased output distributions; above some critical pattern load, α_p, we find two solutions to the saddlepoint equations: one has a non-zero threshold and a lower free energy with an

§ E-mail address: A.H.L.West@aston.ac.uk

0305-4470/97/103471+26$19.50

asymptotic error rate of 50%, the other is identical to that of a perceptron without threshold and exhibits a higher free energy with an asymptotic error rate above 50%. The order parameters show a second-order phase transition at the bifurcation point and have different asymptotic values.

This work is further motivated by the fact that the results of this calculation can be applied iteratively to yield the storage capacity of a class of networks with variable architecture produced by constructive algorithms [8, 9] which will be reported elsewhere [10, 11]. This is possible since these algorithms construct the network architecture during training, starting with a simple Boolean perceptron and adding more perceptrons only when needed, i.e. when the existing network is incapable of performing the requested task. The training is performed separately for each perceptron after its creation and the weights are subsequently frozen. Therefore, results for the perceptron are sufficient to calculate the capacity limit of multilayer networks produced by certain constructive algorithms. So far, only an information theoretic upper bound has been derived for two-layer networks with fixed hidden-layer-to-output weights [12]. Statistical-mechanics calculations have been hampered by the inherent difficulties of the replica calculation. Replica symmetric (RS) treatments [13, 14] violate the above-mentioned upper (Mitchison–Durbin) bound. Other efforts [15] break the symmetry of the hidden units explicitly prior to the actual calculation, but the resulting equations are approximations and are difficult to solve for large networks.

The paper is structured as follows. In section 2 we introduce the model, the Boolean perceptron with threshold (and spherical or Ising constraint) and correlated output and input distributions. We briefly explain the replica framework and outline the one-step replica symmetry breaking (RSB) calculations for the two constraints for both the free energy and distribution of pattern stabilities. This is followed in section 3 by a discussion of the error rate and the pattern-stability distribution of the two Boolean perceptron models. We finish with a discussion of the significance of the results and some concluding remarks in section 4.

2. Replica calculation of the Boolean perceptron

In this section we will outline the replica calculation for the Boolean perceptron trying to learn a set of random dichotomies above its saturation limit, α_c. The calculation is similar to [4, 5] for real-valued weights and a spherical constraint and to [6] for binary weights, i.e. an Ising constraint; however, we allow for a threshold and biased output and input distributions. In the following the real-valued weight Boolean perceptron will be referred to as the spherical (Boolean) perceptron, whereas the binary-valued weight Boolean perceptron will be referred to as the Ising (Boolean) perceptron. This section is divided into three parts. In section 2.1, the replica calculation for the free energy of the perceptron above saturation is explained briefly. In section 2.2, the same framework is then extended to calculate the distribution of pattern stabilities for the perceptron. In section 2.3, we will outline the differences for the calculations of the Ising perceptron and present the resulting equations.

2.1. Free energy of the spherical perceptron

In the capacity problem the aim is to adjust the parameters of a spherical perceptron, the synaptic weight vector, $\boldsymbol{W} \in \mathbb{R}^N$, and threshold, $\theta \in \mathbb{R}$, to minimize the error on a set of $p = \alpha N$ input–output mappings, $\boldsymbol{\xi}^\mu \in \{-1, 1\}^N \rightarrow \zeta^\mu \in \{-1, 1\}$ $(\mu = 1, \ldots, p)$, from an N-dimensional binary input space to binary targets. The output of the perceptron is hereby

determined by

$$\sigma^\mu = \mathrm{sgn}\left(\frac{1}{\sqrt{N}}\boldsymbol{W}\cdot\boldsymbol{\xi}^\mu - \theta\right) = \mathrm{sgn}(h^\mu) \tag{1}$$

where $\mathrm{sgn}(x)$ is the sign of x, and h^μ is termed the activation of the perceptron. We define the error function to be

$$E = \sum_\mu \Theta(\kappa - \lambda^\mu) \tag{2}$$

where $\lambda^\mu = \zeta^\mu h^\mu$ and $\Theta(x)$ is the Heaviside step function, which is 1 for $x > 0$ and 0 otherwise, and κ is the stability with which we require the patterns to be stored. This error function, counting the number of misclassifications, is often referred to as the Gardner–Derrida cost function.

The calculation will be performed in the thermodynamic limit, $N \to \infty$, with a finite example load, $\alpha = p/N$. In the following, we will be interested only in the minimum error possible and will therefore consider zero-temperature Gibbs learning, i.e. we consider the free energy, $f = -\beta^{-1}\log Z$ for $\beta \to \infty$, which is assumed to be self-averaging in the thermodynamic limit. Hence

$$\langle\langle f\rangle\rangle = -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{N\beta}\langle\langle\log Z\rangle\rangle = -\lim_{\beta\to\infty}\lim_{N\to\infty}\frac{1}{N\beta}\left\langle\!\!\left\langle \log\int \mathrm{d}\mu\,(\boldsymbol{W})\mathrm{e}^{-\beta E}\right\rangle\!\!\right\rangle \tag{3}$$

where $\langle\langle\cdot\rangle\rangle$ is the quenched average over the distribution of patterns, consisting of integrations over biased input and output distributions. The binary input distribution is independent of the pattern and site indices μ and j:

$$P(\xi_j^\mu) = P(\xi) = \tfrac{1}{2}(1+m_\mathrm{i})\delta(1-\xi) + \tfrac{1}{2}(1-m_\mathrm{i})\delta(1+\xi). \tag{4a}$$

The output distribution is also independent of the pattern index:

$$P(\zeta^\mu) = P(\zeta) = \tfrac{1}{2}(1+m_\mathrm{o})\delta(1-\zeta) + \tfrac{1}{2}(1-m_\mathrm{o})\delta(1+\zeta) \tag{4b}$$

where m_i and m_o represent the input and output bias respectively.

Furthermore, in the case of real-valued weights, we enforce a spherical constraint on the weight vector

$$\mathrm{d}\mu\,(\boldsymbol{W}) = \delta(\boldsymbol{W}\cdot\boldsymbol{W} - N)\prod_{i=1}^{N}\mathrm{d}W_i \tag{5}$$

to avoid the invariance $(\boldsymbol{W},\kappa) \to (\lambda\boldsymbol{W},\lambda\kappa)$. To be able to pick out the two possible error sources (*wrongly on*, where the requested target is $\zeta^\mu = -1$ but the output is $\sigma^\mu = 1$ and *wrongly off*, where $\zeta^\mu = 1$ but $\sigma^\mu = -1$), we introduce auxiliary variables, ϵ^+ and ϵ^-, in the error function (equation (2))

$$E = \sum_\mu \Theta(\kappa - \lambda^\mu)[\epsilon^-\Theta(\zeta^\mu) + \epsilon^+\Theta(-\zeta^\mu)] = \sum_\mu V(\lambda^\mu,\kappa,\zeta^\mu) \tag{6}$$

where V is the error measure for a single example and has been introduced for convenience†. The derivatives of the free energy with respect to ϵ^+ or ϵ^- at $\epsilon^+ = \epsilon^- = 1$ will give us the *wrongly on* and *wrongly off* errors respectively.

† This is also consistent with earlier work [5] and in principle allows a calculation for an arbitrary cost function.

To be able to perform the quenched average we make use of the replica trick $\langle\!\langle \log Z \rangle\!\rangle = \lim_{n\to 0}(\langle\!\langle Z^n \rangle\!\rangle - 1)/n$. After application of standard techniques and introduction of the order parameters†

$$Q_{\sigma\rho} = \frac{1}{N}\boldsymbol{W}^\sigma \cdot \boldsymbol{W}^\rho \qquad \text{for } \sigma < \rho \qquad M_\sigma = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i^\sigma \tag{7}$$

their Lagrange multipliers, $\hat{Q}_{\sigma\rho}$ and $\hat{M}_\sigma$, and the Lagrange multiplier, $\hat{E}_\sigma$, associated with the spherical constraint‡, the replicated partition function is

$$\langle\!\langle Z^n \rangle\!\rangle = \int_{-\infty}^{\infty}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\left(\prod_\sigma \frac{\mathrm{d}M_\sigma \mathrm{d}\hat{E}_\sigma}{2\pi}\right)\left(\prod_{\sigma<\rho}\frac{\mathrm{d}Q_{\sigma\rho}\mathrm{d}\hat{Q}_{\sigma\rho}}{2\pi}\right) \times \exp\left\{N\left[G_0(\hat{Q}_{\sigma\rho},\hat{E}_\sigma) + \alpha G_\mathrm{r}(Q_{\sigma\rho},\theta_\sigma,M_\sigma) + \tfrac{1}{2}\sum_\sigma \hat{E}_\sigma - \sum_{\sigma<\rho} Q_{\sigma\rho}\hat{Q}_{\sigma\rho}\right]\right\} \tag{8}$$

where

$$G_0(\hat{Q}_{\sigma\rho},\hat{E}_\sigma) = \log\left\{\int_{-\infty}^{\infty}\prod_\sigma \mathrm{d}W^\sigma \exp\left[-\tfrac{1}{2}\sum_\sigma \hat{E}_\sigma W^\sigma W^\sigma + \sum_{\sigma<\rho}\hat{Q}_{\sigma\rho}W^\sigma W^\rho\right]\right\} \tag{9}$$

is the prior constraint Hamiltonian and

$$G_\mathrm{r}(Q_{\sigma\rho},\theta_\sigma,M_\sigma) = \log\left\langle\int_{-\infty}^{\infty}\left(\prod_\sigma \frac{\mathrm{d}\lambda_\sigma \mathrm{d}\hat{\lambda}_\sigma}{2\pi}\right)\exp\left\{-\beta V(\lambda_\sigma,\kappa,\zeta) - \mathrm{i}\sum_\sigma \hat{\lambda}_\sigma\lambda_\sigma - \mathrm{i}\zeta\sum_\sigma \hat{\lambda}_\sigma(\theta_\sigma - m_\mathrm{i}M_\sigma) - \tfrac{1}{2}(1-m_\mathrm{i}^2)\left[\sum_\sigma \hat{\lambda}_\sigma^2 + 2\sum_{\sigma<\rho}\hat{\lambda}_\sigma\hat{\lambda}_\rho Q_{\sigma\rho}\right]\right\}\right\rangle_\zeta \tag{10}$$

is the replicated Hamiltonian, and where $\langle\cdot\rangle_\zeta$ denotes an average over the output distribution.

2.1.1. The replica symmetric ansatz. To make further progress one has to make an assumption for the structure of the replica space. The simplest assumption is that replica symmetry holds (which is believed to correspond usually to a connected solution space):

$$\begin{aligned} &Q_{\sigma\rho} = q_1 \qquad \text{and} \qquad \hat{Q}_{\sigma\rho} = \hat{q}_1 \qquad \text{for } \sigma < \rho \\ &M_\sigma = M \qquad \theta_\sigma = \theta \qquad \text{and} \qquad \hat{E}_\sigma = \hat{E} \qquad \text{for all } \sigma. \end{aligned} \tag{11}$$

Inserting the above ansatz into equations (9) and (10) and taking the $n \to \infty$ limit yields

$$\begin{aligned} G_0^\mathrm{RS} &= \frac{1}{2}\frac{\hat{q}_1}{\hat{E}+\hat{q}_1} - \frac{1}{2}\log(\hat{E}+\hat{q}_1) \\ G_\mathrm{r}^\mathrm{RS} &= \left\langle \int \mathrm{D}t \log[\mathcal{F}_\mathrm{RS}(t,\beta,q_1,\kappa,\zeta\theta)]\right\rangle_\zeta \end{aligned} \tag{12}$$

where all integrals without explicit limits are from $-\infty$ to $+\infty$, $\mathrm{D}t = \mathrm{d}t\exp(-t^2/2)/\sqrt{2\pi}$ and the function $\mathcal{F}_\mathrm{RS}$ is given by

$$\mathcal{F}_\mathrm{RS}(t,\beta,q_1,\kappa,\zeta\theta) = \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q_1)}}\exp\left(-\beta\left[V(\lambda,\kappa,\zeta) + \frac{(\psi+\sqrt{q_1}t)^2}{2x}\right]\right) \tag{13}$$

† One could also allow $\rho = \sigma$. In this case $Q_{\sigma\sigma} = 1$ and $\hat{Q}_{\sigma\sigma} = \hat{E}_\sigma$ due to the spherical constraint.

‡ The contribution of $\hat{M}_\sigma$ actually vanishes in the thermodynamic limit.

where $x = \beta(1 - q_1)$ and

$$\psi(\lambda) = \frac{\lambda + \zeta(\theta - m_i M)}{\sqrt{1 - m_i^2}}. \tag{14}$$

When taking the $\beta \to \infty$ in order to access the ground state with least errors only, one has to distinguish two regimes. Below the capacity limit, α_c (above which the training error becomes strictly positive), $q_1 < 1$ even for $\beta \to \infty$. At and above the capacity limit, $q_1 \to 1$ for $\beta \to \infty$, because the volume of the individual solution spaces vanishes. We therefore make the self-consistent ansatz for $\alpha \geqslant \alpha_c$ that $x = \beta(1 - q_1)$ remains finite in the zero-temperature limit. In this case, the integral over λ in (13) can be calculated by the saddlepoint method; the exponential is evaluated at $\lambda = \lambda_0$, where λ_0 minimizes the square bracket for a given t. After calculating $\lambda_0(t)$ for the Gardner–Derrida cost function and eliminating $\hat{q}_1$ and $\hat{E}$, the RS free energy at $\epsilon^+ = \epsilon^- = 1$ simplifies to:

$$\langle\langle f_{RS} \rangle\rangle = \alpha \left\langle \int_{-\tau}^{\sqrt{2x}-\tau} \mathrm{D}t \frac{(t+\tau)^2}{2x} + H\left(\sqrt{2x} - \tau\right) \right\rangle_\zeta - \frac{1}{2x} \tag{15}$$

where

$$\tau = \psi(\kappa) = \frac{\kappa + \zeta(\theta - m_i M)}{\sqrt{1 - m_i^2}} \qquad \text{and} \qquad H(u) = \int_u^\infty \mathrm{D}t. \tag{16}$$

The free energy has to be evaluated at the saddlepoints with respect to the variables x and θ. The capacity limit, α_c, can be calculated from the saddlepoint equations by taking the limit $x \to \infty$. A more detailed examination of the free energy and the saddlepoint equations is deferred to section 2.1.3.

Above the capacity limit α_c it is evident that different solutions can misclassify different patterns and the solution space will in general be disconnected. It has also been previously shown that in this case the replica symmetric saddlepoint is locally unstable [16], and the Parisi scheme of successive steps of RSB [17] must be employed.

2.1.2. The one-step RSB ansatz. Here, we will restrict ourselves to a one-step RSB calculation. We note that it has been shown recently that, for the spherical perceptron with the Gardner–Derrida cost function, infinitely many RSB steps are necessary to derive the correct result [18]. Although one-step RSB is, therefore, incorrect it is a very good approximation, as a two-step RSB calculation carried out for the spherical perceptron without threshold yielded only minor corrections in the free energy [18].

The ansatz for the one-step RSB is that $Q_{\sigma\rho}$ is an $n \times n$ matrix

$$(Q_{\sigma\rho})_{nn} = \begin{pmatrix} Q_1 & Q_0 & \cdots & Q_0 \\ Q_0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & Q_0 \\ Q_0 & \cdots & Q_0 & Q_1 \end{pmatrix}_{nn} \tag{17}$$

where Q_0 is an $m \times m$ matrix with elements q_0 and Q_1 is an $m \times m$ matrix with 0 on the diagonal and q_1 elsewhere. The ansatz for $\hat{Q}_{\sigma\rho}$ has the same block structure as for $Q_{\sigma\rho}$ with matrices $\hat{Q}_0$ and $\hat{Q}_1$. We further assume

$$M_\sigma = M \qquad \theta_\sigma = \theta \qquad \text{and} \qquad \hat{E}_\sigma = \hat{E} \qquad \text{for all } \sigma \tag{18}$$

similar to the RS case (11). The order parameters, q_1 and q_0, can be interpreted as the typical overlap between pairs of weight vectors in the same and different solution spaces

respectively. Clearly, if the solution space is connected $q_0 \equiv q_1$, which is the case for $\alpha \leqslant \alpha_c$, we recover replica symmetry. Again using the above ansatz in equations (9) and (10) and taking the $n \to 0$ limit yields

$$G_0^{\mathrm{RSB}} = \frac{1}{2}\frac{\hat{q}_0}{(\hat{E}+\hat{q}_1)-m(\hat{q}_1-\hat{q}_0)} - \frac{1}{2}\log(\hat{E}+\hat{q}_1) - \frac{1}{2m}\log\left(1-\frac{\hat{q}_1-\hat{q}_0}{\hat{E}+\hat{q}_1}\right)$$
$$G_{\mathrm{r}}^{\mathrm{RSB}} = \left\langle \int \mathrm{D}t \frac{1}{m}\log[\mathcal{F}_{\mathrm{RSB}}(t, m, \beta, q_0, q_1, \kappa, \zeta\theta)] \right\rangle_\zeta \tag{19}$$

where the function $\mathcal{F}_{\mathrm{RSB}}$ is given by

$$\mathcal{F}_{\mathrm{RSB}}(t, m, \beta, q_0, q_1, \kappa, \zeta\theta) = \int \mathrm{D}z \left[\int \frac{\mathrm{d}\lambda}{\sqrt{2\pi(1-q_1)}} \times \exp\left\{ -\beta\left[V(\lambda, \kappa, \zeta) + \frac{(\psi + \sqrt{q_0}\,t + \sqrt{q_1-q_0}\,z)^2}{2\beta\sqrt{1-q_1}} \right] \right\} \right]^m \tag{20}$$

with ψ as in (14).

Similar to the RS case, we are interested in the $\beta \to \infty$ limit where $q_1 \to 1$ with $x = \beta(1-q_1)$ finite. The λ-integral in (20) can again be evaluated at the saddlepoint $\lambda = \lambda_0$, where λ_0 minimizes the square bracket in the exponential for given z and t. Furthermore, the replica space dimension $m \to 0$ ($\beta \to \infty$) as we only access one solution and it becomes exponentially unlikely that any other solution is visited [17]. We therefore make a second self-consistent ansatz that $w = m/(1-q_1)$ remains finite in the zero-temperature limit. After some algebra, including determining $\lambda_0(z, t)$ for the Gardner–Derrida cost function and elimination of $\hat{q}_1$, $\hat{q}_0$ and $\hat{E}$, the one-step RSB free energy for $\epsilon^+ = \epsilon^- = 1$ is given by

$$\langle\langle -f_{\mathrm{RSB}} \rangle\rangle = \frac{\alpha}{wx}\left\langle \int \mathrm{D}t \log[\mathcal{F}_{\mathrm{RSB}}(t, w, x, q_0, \kappa, \zeta\theta)] \right\rangle_\zeta + \frac{q_0}{2x(1+w\Delta q)} + \frac{\log(1+w\Delta q)}{2wx} \tag{21}$$

where τ is as before (16), $\Delta q = 1 - q_0$, and the function $\mathcal{F}_{\mathrm{RSB}}$ has simplified to

$$\mathcal{F}_{\mathrm{RSB}}(t, w, x, q_0, \kappa, \zeta\theta) = \int_{-\frac{\mu}{\sqrt{\Delta q}}}^{\frac{\sqrt{2x}-\mu}{\sqrt{\Delta q}}} \mathrm{D}z \exp\left[-\frac{w}{2}\left(\sqrt{\Delta q}\,z + \mu\right)^2 \right] + H\left(\frac{\mu}{\sqrt{\Delta q}}\right) + \mathrm{e}^{-wx} H\left(\frac{\sqrt{2x}-\mu}{\sqrt{\Delta q}}\right) \tag{22}$$

with $\mu = \tau + \sqrt{q_0}\,t$. The free energy has to be evaluated at the saddlepoints with respect to the variables w, x, q_0 and θ.

2.1.3. Saddlepoint equations and training error. Examining both the RS (15) and the one-step RSB (21) free energies more closely, one sees that the ferromagnetic bias M of the weight vector (7) appears only in the definition of τ (16) and can be set to zero without loss of generality (w.l.o.g.)†. The order parameter, M, is therefore superfluous, i.e. any ferromagnetic bias in the couplings can be compensated by an adjustment of the threshold θ. This is in contrast to the usual paradigm, which eliminates θ in favour of M (e.g. [1, 7]), and therefore reduces the number of actual parameters of the perceptron. However, this is clearly only possible if $m_{\mathrm{i}} \neq 0$ and will lead to large values of M for small m_{i}.

† The fact that M is redundant is a direct consequence of the fact that the integral over $\hat{M}$ does not contribute in the thermodynamic limit.

We further note that the bias of the input distribution, m_i, appears only in the definition of τ (16) also and its sole influence is a rescaling of the threshold and the stability. Therefore, a biased input distribution has the same effect on the performance of the perceptron as the increase of the stability for an unbiased input distribution. This can be understood in geometric terms. If the input distribution is unbiased, input vectors lie randomly distributed on the edges of the unit hypercube and two distinct patterns have a typical overlap of zero. Biased patterns on the other hand are correlated and have a typical overlap of m_i^2 with each other, i.e. they concentrate on a 'conelike' section of the hypercube. The typical distance between patterns is therefore reduced by $\sqrt{1-m_\mathrm{i}^2}$. Any solution of the weight vector corresponds to a hyperplane which separates the two kinds of patterns. The achieved stability is half the distance of the two correctly classified patterns with the shortest separation across this plane and hence the stability decreases by $\sqrt{1-m_\mathrm{i}^2}$ as well. Only at zero stability does the increase of the input bias have no effect on the performance of the perceptron. In the following, we will therefore set $m_\mathrm{i}=0$ w.l.o.g.

The saddlepoint equation of the derivative of the free energy with respect to θ at $\epsilon^+=\epsilon^-=1$ gives

$$0=\left\langle \zeta \int_{-\tau}^{\sqrt{2x}-\tau} \mathrm{D}t\,(t+\tau)\right\rangle_\zeta \qquad \text{and} \qquad 0=\left\langle \zeta \int \mathrm{D}t\,\log[\mathcal{F}_\mathrm{RSB}(t,w,x,q_0,\kappa,\zeta\theta)]\right\rangle_\zeta \tag{23}$$

for RS and one-step RSB respectively. For zero bias, one can readily see that $\theta=0$ is always a solution to this and the other saddlepoint equations; regaining the results of the perceptron without threshold. However, this does not necessarily imply that this is the only solution to the saddlepoint equations, as demonstrated in section 3.

Taking the derivatives of the free energies with respect to ϵ^- and ϵ^+ at $\epsilon^+=\epsilon^-=1$ and dividing by α gives the error rate (i.e. the number of errors divided by the total number of patterns) of *wrongly off* and *wrongly on* patterns respectively

$$\begin{aligned}\epsilon_\mathrm{RS}^{\mathrm{off/on}} &= \tfrac{1}{2}(1\pm m_\mathrm{o})H(\sqrt{2x}-\kappa\mp\theta)\\ \epsilon_\mathrm{RSB}^{\mathrm{off/on}} &= \tfrac{1}{2}(1\pm m_\mathrm{o})\int \mathrm{D}t\,\frac{\mathrm{e}^{-wx}H\left(\sqrt{2x}-\sqrt{q_0}\,t-\kappa\mp\theta\right)}{\mathcal{F}_\mathrm{RSB}(t,w,x,q_0,\kappa,\pm\theta)}\end{aligned} \tag{24}$$

where we have set $m_\mathrm{i}=0$ w.l.o.g.

We note that we find no numerical difference between the total training error† and the free energy in the thermodynamic limit for both RS and one-step RSB and conclude that the normalized entropy, $s=S/N$, must diverge sublinearly or logarithmically for $\beta\to\infty$. One can calculate the first-order finite temperature correction of the free energy for both RS and one-step RSB analytically, and find that it is negative and proportional to $\log(\Delta q)$, and equal to the low-temperature entropy. Unlike in the binary case, where a negative entropy is physically impossible and therefore an indication that the employed replica ansatz breaks down, a negative entropy has no such physical meaning in the real-valued case, due to an arbitrary entropy offset.

2.2. Pattern stability distribution

The pattern stability distribution (PSD) $P(\Lambda)$ is of interest as it provides the distance of stabilized ($\Lambda\geqslant\kappa$) and unstabilized patterns ($\Lambda<\kappa$) to the given threshold stability κ,

† That is the error rates multiplied by α.

i.e. it gives an idea how seriously patterns are misclassified. This extra information will be quite helpful in examining the already mentioned bifurcation point in order-parameter space in section 3. For error functions other than the Gardner–Derrida cost function (e.g. the perceptron or adatron cost function), the integration of the probability density† $p(\Lambda)$ over the unstabilized patterns yields the error rate ϵ [21, 5], which is otherwise inaccessible. The PSD is further of great importance to the dynamics of related attractor neural networks, by determining the basin of attraction of the memory states [19, 20].

The PSD $P(\Lambda|D)$ is in general dependent on the instances of the data set $D = \{(\boldsymbol{\xi}^\mu, \zeta^\mu)|\mu = 1, \ldots, p\}$. As we are interested in its average value $P(\Lambda) = \langle\langle P(\Lambda|D)\rangle\rangle$, we quench over the instances of the examples

$$p(\Lambda) = \langle\langle p(\Lambda|D)\rangle\rangle = \left\langle\!\!\left\langle \frac{1}{Z}\int \mathrm{d}\mu\,(\boldsymbol{W}) \exp\left[-\beta\sum_\mu V(\lambda^\mu, \kappa, \zeta^\mu)\right]\delta(\Lambda - \lambda^1)\right\rangle\!\!\right\rangle \tag{25}$$

where the pattern stability of pattern 1 is calculated w.l.o.g. as the pattern distribution is independent of the pattern index μ. Here, $\mathrm{d}\mu(\boldsymbol{W})$ is the spherical constraint (5), but the above equation holds for any weight prior. In the thermodynamic limit, one can calculate this average using the replica trick. The calculation is similar to that of the free energy except for the average over the first pattern [19]. After some algebra one finds for the RS ansatz

$$p_{\mathrm{RS}}(\Lambda) = \left\langle \int \mathrm{D}t \frac{\mathcal{F}_{\mathrm{RS}}(t, \beta, q_1, \kappa, \zeta\theta)\delta(\Lambda - \lambda)}{\mathcal{F}_{\mathrm{RS}}(t, \beta, q_1, \kappa, \zeta\theta)}\right\rangle_\zeta \tag{26}$$

where $\mathcal{F}_{\mathrm{RS}}(13)$ has to be evaluated at the saddlepoint of the free energy. With the one-step RSB ansatz one finds a similar expression to equation (26) for $p_{\mathrm{RSB}}(\Lambda)$ only with $\mathcal{F}_{\mathrm{RS}}$ replaced by $\mathcal{F}_{\mathrm{RSB}}$ (20).

For $\beta \to \infty$ above the capacity limit α_{c}, both $p_{\mathrm{RS}}(\Lambda)$ and $p_{\mathrm{RSB}}(\Lambda)$ can be simplified along the lines of [21, 5] as the λ-integral in both $\mathcal{F}_{\mathrm{RS}}$ and $\mathcal{F}_{\mathrm{RSB}}$ can be evaluated at their respective saddlepoint λ_0.

Calculating λ_0 for the Gardner–Derrida cost function equates the RS probability density

$$\begin{aligned} p_{\mathrm{RS}}(\Lambda) = \Bigg\langle \delta(\Lambda - \kappa)\int_{-\tau}^{\sqrt{2x}-\tau} \mathrm{D}t + \frac{\boldsymbol{\Theta}(\Lambda - \kappa)}{\sqrt{2\pi(1 - m_{\mathrm{i}}^2)}}\exp\left[-\tfrac{1}{2}\phi^2\right] \\ + \frac{\boldsymbol{\Theta}\left(\kappa - \Lambda - \sqrt{1 - m_{\mathrm{i}}^2}\sqrt{2x}\right)}{\sqrt{2\pi(1 - m_{\mathrm{i}}^2)}}\exp\left[-\tfrac{1}{2}\phi^2\right]\Bigg\rangle_\zeta \end{aligned} \tag{27}$$

where x and θ have to be evaluated at the saddlepoint of the free energy (15) as mentioned above and $\phi = \psi(\Lambda)$ with ψ as in (14). The PSD has three terms, a δ-function contribution for $\Lambda = \kappa$, i.e. at the error boundary, and two Gaussian contributions, which leave a gap of width $\sqrt{1 - m_{\mathrm{i}}^2}\sqrt{2x}$.

In the one-step RSB ansatz, the probability density can be calculated similarly leading to

$$p_{\mathrm{RSB}}(\Lambda) = \left\langle \int \mathrm{D}t \frac{\mathcal{N}_{\mathrm{RSB}}(t, w, x, q_0, \kappa, \zeta\theta)}{\mathcal{F}_{\mathrm{RSB}}(t, w, x, q_0, \kappa, \zeta\theta)}\right\rangle_\zeta \tag{28}$$

† We use uppercase P notations for probability distributions and lowercase p for probability densities.

where the denominator $\mathcal{F}_{\rm RSB}$ is identical to (22) and the numerator is given by

$$\mathcal{N}_{\rm RSB}(t, w, x, q_0, \kappa, \zeta\theta) = \delta(\Lambda - \kappa) \int_{-\frac{\mu}{\sqrt{\Delta q}}}^{\frac{\sqrt{2x}-\mu}{\sqrt{\Delta q}}} \mathrm{D}z \exp\left[-\frac{w}{2}\left(\sqrt{\Delta q}\, z + \mu\right)^2\right]$$
$$+ \frac{\boldsymbol{\Theta}(\Lambda - \kappa)}{\sqrt{2\pi\,\Delta q(1 - m_{\rm i}^2)}} \exp\left[-\frac{\varrho^2}{2\Delta q}\right]$$
$$+ \frac{\boldsymbol{\Theta}\left(\kappa - \Lambda - \sqrt{1 - m_{\rm i}^2}\sqrt{2x}\right)}{\sqrt{2\pi\,\Delta q(1 - m_{\rm i}^2)}} \exp\left[-\frac{\varrho^2}{2\Delta q} - wx\right] \tag{29}$$

where $\varrho = \phi + \sqrt{q_0}\, t$ and the values of the order parameters x, w, q_0 and θ are again determined by the saddlepoint of the free energy (21).

Comparing the one-step RSB with the RS PSDs we find three similar contributions, a δ-peak at the stability κ, and two exponential terms, separated by a gapwidth of $\sqrt{1 - m_{\rm i}^2}\sqrt{2x}$. In general, one finds [5] that one-step RSB has a smaller gap, which is formally due to a reduced saddlepoint value of x, and a reduced weight of the δ-function contribution. The one-step RSB distribution has also lost the Gaussian form of the RS distribution, due to the presence of the denominator and the integration over t. We further find a correction to the third contribution, which represents unstabilized, i.e. erroneous, patterns, which has acquired an extra supressive exponential term, e^{-wx}. As we have already pointed out in section 2.1.3, the role of a non-zero input bias is the rescaling of the threshold θ, the stability κ and the pattern stability Λ with a factor of $\sqrt{1 - m_{\rm i}^2}$, and can therefore be set to zero w.l.o.g.

It is worth mentioning that the gap and the δ-peak are a feature of training algorithms above saturation employing the Gardner–Derrida cost function [22]. This is due to the fact that an algorithm achieving least errors attempts to stabilize the least unstabilized pattern, until any movement of the hyperplane will destabilize a pattern lying on the threshold decision boundary, leading to a fraction of patterns exactly on the decision boundary and leaving a gap between stabilized and unstabilized patterns. The above work has been complemented by a numerical study [23], where the numerical PSD exhibits a gap and a δ-peak which are both finite but smaller than the theoretical one-step RSB predictions within the accuracy of the simulations. This is consistent with a recent proof [18] which showed that any model exhibiting a gap in the PSD necessitates infinitely many RSB steps.

2.3. Ising perceptron

In the case of the Ising perceptron the calculation is very similar. In fact, the calculation of the replicated Hamiltonian $G_{\rm r}$ (10) is exactly the same as it only depends on the quenched average over the training examples. The difference is therefore mainly in the prior constraint Hamiltonian G_0 (9), where the integration over weight space is performed. Since the weight vector of the Ising perceptron is binary, i.e. $\boldsymbol{W} \in \{-1, 1\}^N$, the measure in weight space (see equation (5)) becomes a sum $\int \mathrm{d}\mu\,(\boldsymbol{W}) = \prod_{i=1}^{N} \sum_{W_i = \pm 1}$, and all terms with the Lagrange multiplier, $\hat{E}_\sigma$, associated with the spherical constraint vanish in equation (8). The prior constraint Hamiltonian equates to

$$G_0^{\rm I}(\hat{Q}_{\sigma\rho}) = \log\left\{\prod_{\sigma} \exp\left[-\sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} W^\sigma W^\rho\right]\right\}. \tag{30}$$

Again, using two ansatzes for the structure in replica space, RS and one-step RSB, identical to those made in section 2.1, one finds

$$G_0^{\mathrm{IRS}}(\hat{q}_1) = -\frac{\hat{q}_1}{2} + \int \mathrm{D}t \log\left[2\cosh\left(t\sqrt{\hat{q}_1}\right)\right]$$

$$G_0^{\mathrm{IRSB}}(\hat{q}_1, \hat{q}_0) = -\frac{\hat{q}_1}{2} + \frac{m}{2}(\hat{q}_1 - \hat{q}_0) + \frac{1}{m}\int \mathrm{D}t \log\left[\int \mathrm{D}z\ 2\cosh\left(t\sqrt{\hat{q}_0} + z\sqrt{\hat{q}_1 - \hat{q}_0}\right)\right]^m \tag{31}$$

where IRS(B) stands for the RS or one-step RSB ansatz for the Ising perceptron.

Great care has to be taken in the $\beta \to \infty$ limit, which is discussed in detail in [6], here we will only outline the main results. One finds that the entropy of the RS solution is negative for $\alpha > \alpha_{\mathrm{S}}^{\mathrm{I}}$ with $q_1 < 1$ in the zero-temperature limit, and is therefore incorrect above $\alpha_{\mathrm{S}}^{\mathrm{I}}$. Studying the one-step RSB solutions identifies $\alpha_{\mathrm{S}}^{\mathrm{I}}$ as the capacity limit $\alpha_{\mathrm{c}}^{\mathrm{I}}$. The RS error only becomes strictly positive for $\alpha > \alpha_{\mathrm{E}}^{\mathrm{I}}$ where $q_1 \to 1$ with $x = \beta(1-q_1)$ finite and the RS free energy of the Ising perceptron can be simplified [16], resulting in

$$\langle\!\langle f_{\mathrm{IRS}} \rangle\!\rangle = \alpha\left\langle \int_{-\tau}^{\sqrt{2x}-\tau} \mathrm{D}t \frac{(t+\tau)^2}{2x} + H\left(\sqrt{2x} - \tau\right)\right\rangle_\zeta - \frac{1}{\pi x} \tag{32}$$

which is identical to the RS free energy of the spherical perceptron (15) but for a constant $2/\pi$ in the last α-independent term. The RS solution of the Ising perceptron at α is therefore the same as the RS solution of the spherical perceptron at $\tilde{\alpha} \equiv \pi\alpha/2$, which holds also for error rates and the distribution of pattern stabilities. The RS solution of the Ising perceptron will therefore not be discussed further.

However, as already mentioned above, the RS solution is incorrect for $\alpha > \alpha_{\mathrm{S}}^{\mathrm{I}}$ and $\beta > \beta_{\mathrm{c}}$, where one finds one-step RSB solutions, which are characterized by $q_1 = 1$ and $\hat{q}_1 = \infty$ for finite β. One further finds $m = \beta_{\mathrm{c}}/\beta$, $\hat{q}_0 \to 0$ and makes the self-consistent ansatz that $v = m\beta$ and $y = m\sqrt{\hat{q}_0}$ are finite in the zero-temperature limit. Inserting this ansatz back into (32), one finds $G_0^{\mathrm{IRSB}}(\infty, \hat{q}_0) = G_0^{\mathrm{IRS}}(y^2)/m$. The replicated Hamiltonian $G_{\mathrm{r}}^{\mathrm{IRSB}}$ (19) is calculated similarly to the spherical perceptron, with the above ansatz becoming equivalent to $x \to 0$ and $w \to \infty$ with wx finite. The one-step RSB free energy of the Ising perceptron is therefore given by

$$\langle\!\langle -f_{\mathrm{IRSB}} \rangle\!\rangle = \frac{\alpha}{v}\left\langle \int \mathrm{D}t \log[\mathcal{F}_{\mathrm{IRSB}}(t, v, y, q_0, \kappa, \zeta\theta)]\right\rangle_\zeta + \frac{1}{v}\int \mathrm{D}t \log[2\cosh(yt)] - \frac{\Delta q y^2}{2v} \tag{33}$$

for $\epsilon^+ = \epsilon^- = 1$ and the function $\mathcal{F}_{\mathrm{IRSB}}$ is

$$\mathcal{F}_{\mathrm{IRSB}}(t, v, y, q_0, \kappa, \zeta\theta) = \mathrm{e}^{-v} + (1 - \mathrm{e}^{-v})H\left(\frac{\mu}{\sqrt{\Delta q}}\right) \tag{34}$$

with μ as before. The free energy has to be evaluated at its saddlepoint with respect to the variables v, y, q_0 and θ. The normalized entropy of the Ising perceptron can be shown to be identical to zero [6].

Identical to the spherical perceptron the ferromagnetic bias on the weights M and the bias of the input distribution m_{i} can be set to zero w.l.o.g. We also find as before that $\theta = 0$ is always a solution to the saddlepoint equation for zero output bias and the error rates of *wrongly off* and *wrongly on* patterns are given respectively by

$$\epsilon_{\mathrm{IRSB}}^{\mathrm{off/on}} = \frac{1}{2}(1 \pm m_{\mathrm{o}}) \int \mathrm{D}t \frac{\mathrm{e}^{-v} H\left(\sqrt{2x} - \sqrt{q_0}\, t - \kappa \mp \theta\right)}{\mathcal{F}_{\mathrm{IRSB}}(t, v, y, q_0, \kappa, \pm\theta)}. \tag{35}$$

The PSD density $p_{\mathrm{IRSB}}(\Lambda)$ of the Ising perceptron within a one-step RSB ansatz can be calculated similarly to the spherical perceptron in section 2.2. In the zero-temperature limit, we use $x \to 0$ and $w \to \infty$ with wx finite to find

$$p_{\mathrm{IRSB}}(\Lambda) = \left\langle \int \mathrm{D}t \frac{\mathcal{N}_{\mathrm{IRSB}}(t, v, y, q_0, \kappa, \zeta\theta)}{\mathcal{F}_{\mathrm{IRSB}}(t, v, y, q_0, \kappa, \zeta\theta)} \right\rangle_{\zeta} \tag{36}$$

where the denominator $\mathcal{F}_{\mathrm{IRSB}}$ is identical to (34) and the numerator is given by

$$\mathcal{N}_{\mathrm{IRSB}}(t, v, y, q_0, \kappa, \zeta\theta) = \frac{[\Theta(\Lambda - \kappa) + \mathrm{e}^{-v}\Theta(\kappa - \Lambda)]}{\sqrt{2\pi\,\Delta q(1 - m_{\mathrm{i}}^2)}} \exp\left[-\frac{\varrho^2}{2\Delta q}\right] \tag{37}$$

with ϱ as in (29) and the values of the order parameters y, v, q_0 and θ are evaluated at the saddlepoint of the free energy (33). The PSD of the Ising perceptron has a common Gaussian numerator centred around ϱ, but for an extra exponential suppression of the unstabilized patterns $\Lambda < \kappa$ proportional to e^{-v}.

Comparing the PSDs of the spherical and the Ising perceptron, shows no difference within the RS ansatz besides the already mentioned rescaling of α. However, one finds striking differences within the one-step RSB treatment: the gap in the distribution as well as the δ-peak contribution at the threshold boundary, κ, have vanished in the PSD of the Ising perceptron in contradiction to [22] (see section 2.2). However, this could be explained by the fact that the Ising perceptron cannot adjust its decision boundary continuously, due to the discreteness of the weights. Therefore, one may expect that unstabilized patterns lie arbitrarily close to the decision boundary and the patterns do not accumulate at the threshold stability.

Whereas it has been shown previously that the one-step RSB ansatz for the spherical perceptron is not exact [18], which is formally due to the gap in the PSD, the Ising perceptron does not exhibit this gap and there has been some argument whether one-step RSB is exact for this model†. Krauth and Mézard [6] have carried out a second RSB step and have found no solution different to the one-step RSB result, although one should mention that most of their numerical work was carried out around the capacity limit. Fontanari and Meir [24] have calculated the entropy of the Ising perceptron in a microcanonical approach and found that their RS solution is identical to the one-step RSB solution in the canonical approach. They calculated that the microcanonical RS saddlepoint is locally stable for all α, which also suggests that the ansatz is correct, as a breakdown would require that the RS saddlepoint is locally stable but globally unstable even for $\alpha \to \infty$. A third approach by Horner [25] investigating the learning dynamics using dynamic mean field theory which does not rely on the replica trick, indicates a slightly different picture. He finds that the fluctuation dissipation theorem (FDT) holds for high temperatures and the dynamics are ergodic validating the use of RS. For lower temperatures ergodicity is broken but one finds that a quasi FDT (QFDT) holds, parametrized by a variable m, which has a similar role as the one-step RSB parameter m but has to be chosen inconsistently to the choice of m in replica theory. These dynamics were found to be strictly stable for infinite times indicating that no further RSB steps are necessary in this regime. Furthermore, there exists a third regime with additional diverging time scales which corresponds to further breaking of replica symmetry‡. However, the relevance of dynamic mean-field theory for validating replica ansatz is debatable.

† One-step RSB has been proved to be exact for several models, e.g. for the generalized Sherrington–Kirkpatrick (SK) spin glass with $p = \infty$ spin interactions, which is equivalent to the random energy model and can be solved exactly [26].

‡ An explicit phase diagram is given only for the perceptron and adatron cost functions.

3. Discussion

Calculating the saddlepoint solutions for the order parameters and the error rates as a function of the normalized example number, α, for a range of stabilities, κ, and output biases, m_o, we find striking differences in the solution space to the case of a perceptron without threshold even for zero (output) bias [5, 6].

Since we found the zero bias results the most intriguing, we will limit most of our discussion to this special case, as we find that the introduction of a single free parameter to the perceptron, a threshold, can change the space of solutions accessible to the perceptron radically even for unbiased input and output distributions. We will first examine the order-parameter solution space and the total error rates of the spherical perceptron and the Ising perceptron in sections 3.1 and 3.2. This is followed by a discussion of the PSD in section 3.3 and a discussion of the phase transition found in parameter solution space as a function of the stability κ in section 3.4. We will further assess the influence of a biased output distribution in section 3.5. As we have discussed in section 2.1.3, a biased input distribution can be absorbed through rescaling of the stability and therefore need not be discussed in more detail.

3.1. Error rates and order-parameter solution space of the spherical perceptron

In figure 1 we show the total error rates, ϵ, and the percentrage of *wrongly on* errors, $\epsilon^{\rm on}$ for the spherical perceptron in both the RS and the one-step RSB ansatz, for $m_o = 0$ and $\kappa = 0.1$ as a function α. Below the capacity limit, α_c, the error rate, $\epsilon(\alpha)$, is identically zero. For $\alpha > \alpha_c$, we find that the RS estimate of the error rate is always below the one-step RSB estimate for all $\alpha > \alpha_c$ and replica symmetry is broken as expected [16]. In figure 2, the one-step RSB overlap, q_0, is plotted as a function of α in the same scenario, indicating the degree of replica symmetry breaking.

In figure 1 one can also see that for $\alpha > \alpha_c$, the proportion of *wrongly on* errors, $\epsilon^{\rm on}$, is initially $\frac{1}{2}$. This corresponds to the threshold θ being identical to zero as one

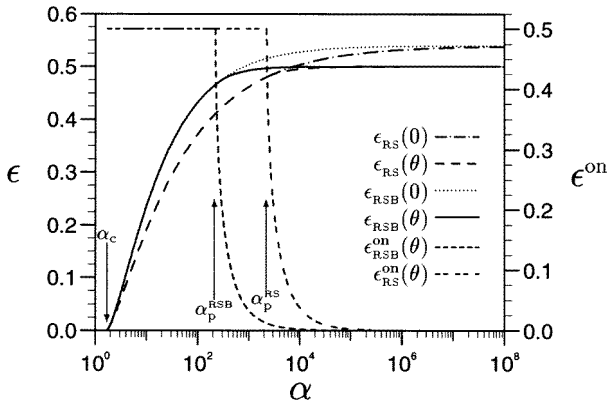

Figure 1. The total error rate, ϵ, of the spherical perceptron as a function of α for $\kappa = 0.1$ is predicted by one-step RSB to be larger than the estimate of RS. For $\alpha > \alpha_c$ both theories initially predict a portion of $\frac{1}{2}$ for *wrongly on* errors, $\epsilon^{\rm on}$, indicating zero threshold (see also figure 2). Above a critical α_p, $\epsilon^{\rm on}$ decreases abruptly and quickly approaches zero signalling a solution with non-zero threshold. This solution exhibits a lower asymptotic error rate than a perceptron without threshold. The predicted value of α_p is smaller for one-step RSB than for RS.

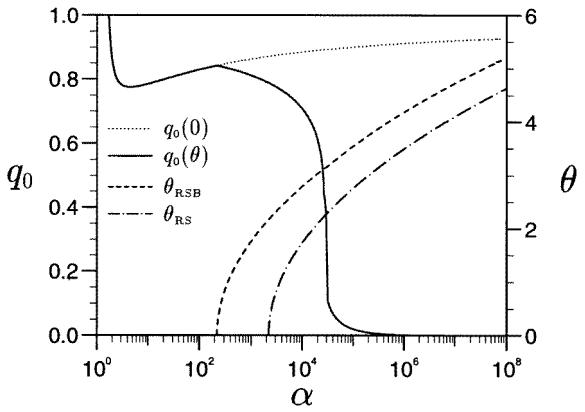

Figure 2. The prediction of the one-step RSB overlap, q_0, for the solution goes to zero as $\alpha \to \infty$ for the perceptron with threshold, whereas it approaches one with zero threshold. The threshold as a function of α in the one-step RSB and the RS ansatz is also included.

can see in figure 2. This solution, for both RS and one-step RSB, could have been expected from examining equations (23). However, above a critical value of the normalized example number, $\alpha > \alpha_p$, we find a second solution to the saddlepoint equations, which is characterized by a non-zero threshold and a fraction of *wrongly on* errors smaller than $\frac{1}{2}$ (see figures 1 and 2). The value of α_p can be seen to be significantly smaller for one-step RSB than for RS. This is found to be true for all finite stabilities, which will be examined in more detail in section 3.4 where we examine the phase transition as a function of the threshold stability κ.

One should note that, although a zero threshold solution (to which we will refer to as θ_0) still exists and is identical to the solution of a perceptron without threshold, it is, however, not a physically viable solution for the perceptron with threshold as it exhibits a higher free energy (i.e. larger error rate, as shown in figure 1) than the non-zero threshold solution (to which we will refer to as θ) and is therefore to be neglected in the thermodynamic limit. This illustrates that a solution to the saddlepoint equations found for any given replica ansatz is not necessarily unique.

Going back to figure 1, one finds for further increasing $\alpha \to \infty$ the error rate of the θ_0 solution approaches an asymptotic error rate which is higher than $\frac{1}{2}$, the asymptotic error rate of the θ solution. The qualitative difference between the error rates can be better understood by examining the PSD and we will therefore defer the discussion of the error limit to section 3.3.

The bifurcation point in solution space is a second-order phase transition as all order parameters (see e.g. $\theta(\alpha)$ and $q_0(\alpha)$ in figure 2) are continuous but non-differentiable for $\alpha = \alpha_p$. In particular, for the threshold the numerical data strongly indicates the functional relationship

$$\theta \propto [\log(\alpha) - \log(\alpha_p)]^\gamma \tag{38}$$

for both RS and one-step RSB theory with an exponent γ which is in very good agreement with the mean-field theory exponent of $\frac{1}{2}$, and a prefactor which is κ dependent and consistently larger for one-step RSB. We further have spontaneous symmetry breaking in the space of thresholds θ as the solution is invariant under sign change of θ. The external field in this case is the output bias, m_o, as it breaks the symmetry in θ space and 'smears' out the phase transition, as will be studied more closely in section 3.5.

The phase transition at α_p stems from the competition between optimizing the weights (or hyperplane angle) and a deterministic bias in the output of the perceptron which is controlled by the threshold. Whereas it is self-evident that for a biased output distribution it is also sensible to bias the output of the student with a non-zero threshold, this is only the case for an unbiased output distribution when the error rate becomes large enough for a given stability κ. To understand this more clearly, the distribution of pattern stabilities together with the total error rate is studied around the phase transition in section 3.3.

In order-parameter space we find qualitatively very different solutions, as can be seen in figure 2 for the order parameters q_0 and θ. For the θ solution, we find the threshold increases towards infinity following the above functional relationship of equation (38) and q_0 decaying to zero, where we find numerically $q_0 \propto 1/\alpha$, with the possibility of minor logarithmic corrections. For the θ_0 solution on the other hand q_0 approaches 1. To investigate the functional behaviour of the θ_0 solution in more detail, one can expand the free energy using the numerically justified ansatz $x \propto 1/\alpha$ and $w \propto \sqrt{\alpha}$ for $\alpha \to \infty$. Although we find the same scaling behaviour as [5], we have found that their prefactors are inconsistent with our analytical solutions and the numerical data. In particular, the solutions of the order parameters are to leading order

$$\Delta q = \frac{2}{\log(\alpha)} \qquad x = \frac{9}{4}\frac{e^{\kappa^2/2}}{\alpha[\log\alpha]^{3/2}} \qquad \text{and} \qquad w = \frac{4}{9}\sqrt{\pi}\,e^{-\kappa^2/4}\sqrt{\alpha}[\log\alpha]^{9/4}. \quad (39)$$

These solutions are, however, only good approximations provided Δq is small and $\log\alpha \gg \log(\log\alpha)$, i.e. in general $\alpha \gg 10^{10}$ and is therefore not very accurate in the region where numerical solutions were obtained. The solutions suggest that for increasing α the degree of RSB becomes more severe as $m[m = wx/\beta]$ and $(1 - q_1)[1 - q_1 = x/\beta]$ decay to zero faster than the temperature.

For the solution with $\theta \neq 0$, we have not been able to find closed-form asymptotic solutions to the saddlepoint equations. In fact, closed-form asymptotic solutions are not even feasible for the much simpler RS theory. The numerical analysis is quite difficult for both x and w; w and wx may at most diverge algebraically in $\log\alpha$ with powers smaller than one, whereas x seems to have a similar $\log\alpha$ behaviour, but the power is even smaller in magnitude and its sign seems to be κ dependent. As the error in the numerical calculation of the order-parameter solutions increase with α and the prefactor in the power laws in $\log\alpha$ are very small, we were not able to determine the value of the powers accurately. A divergent behaviour of wx indicates that the degree of RSB becomes less severe for increasing α, which should be contrasted with the θ_0 solution where the degree of RSB becomes worse.

We find the different asymptotic behaviours for the two sets of order-parameter solutions puzzling; especially, the asymptotics of the order parameter q_0—the typical overlap between two replicas in different solution spaces. Whereas q_0 decays algebraically in α to zero for the θ solution, i.e. weight-vector solutions become totally uncorrelated, it approaches 1 logarithmically for the θ_0 solution, i.e. the weight-vector solutions become absolutely correlated. It has been argued before [5] that this asymptotic behaviour for the spherical perceptron without threshold is incorrect (and one-step RSB must therefore be inexact at least for a high storage level), since one should expect q_0 to approach 0 for $\alpha \to \infty$ as in this limit any weight vector should perform equally well on the training data. More precisely, for loads α greater than the capacity limit α_c, the perceptron classifies only a subset of the examples correctly and misclassifies the rest. For moderate loads and small error rates, there must be a significant overlap between the sets of examples two weight-vector solutions classify correctly. Therefore, the average overlap between weight-vector

solutions should be non-zero and hence, $q_0 > 0$. For very large α and large error rates ϵ, the smallest possible overlap between two sets of correctly classified examples should decrease† and since the patterns are uncorrelated, the correlations between their respective weight-vector solutions should decrease similarly. Hence, the smallest average overlap scale in the replica ansatz should approach 0 for $\alpha \to \infty$.

We will later return to this argument and the issue of the breakdown of one-step RSB in the light of the asymptotics of the order parameter q_0, especially in comparison with the asymptotic solutions of the Ising perceptron, which we will present below.

3.2. Order-parameter solution space of the Ising perceptron

As mentioned in section 2.3, whereas it has been established that one-step RSB is not exact for the spherical perceptron there has been some argument whether one-step RSB is exact for the Ising perceptron, and it is therefore useful to compare the solution in order-parameter space and their asymptotics for the two weight priors.

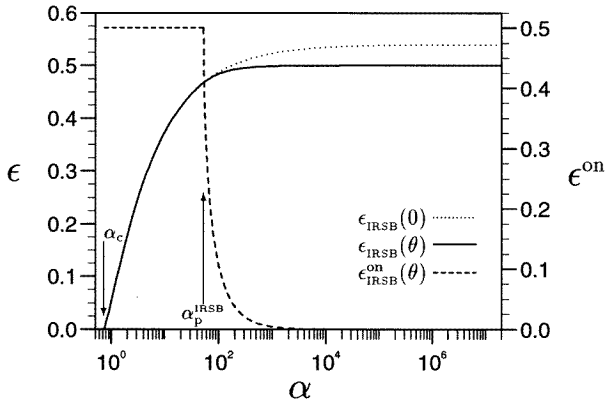

Figure 3. The error rates, ϵ, are shown as a function of α with $\kappa = 0.1$ for the Ising perceptron within the one-step RSB ansatz. Similar to the spherical perceptron there is initially only one solution with a fraction of $\frac{1}{2}$ for *wrongly on* errors, $\epsilon^{\rm on}$, and zero threshold (see figure 4). Again we find a bifurcation point in solution space at a critical $\alpha_{\rm p}$, which is smaller than for the spherical perceptron and similar behaviour of the fraction of $\epsilon^{\rm on}$ errors.

In figure 3 we show the evolution of the error rates and the fractions of *wrongly on* errors and in figure 4 the corresponding values of the order parameters q_0 and θ for the Ising perceptron in the one-step RSB ansatz in the same scenario, i.e. for $m_{\rm o} = 0$ and $\kappa = 0.1$. We find certain similarities but also striking differences to the results for the spherical perceptron. At the capacity limit $\alpha_{\rm c}^{\rm I}$, q_0 does not approach 1 as in the spherical perceptron, indicating a single solution in weight space, but a finite value $q_0 < 1$, i.e. several correlated solutions exist at $\alpha_{\rm c}^{\rm I}$. As for the spherical perceptron, the solution to the saddlepoint equations is initially unique and exhibits a zero threshold. As the error increases for growing α, we find a similar second-order phase transition in order-parameter space, with the emergence of a second solution to the saddlepoint equations characterized by a non-zero threshold at $\alpha = \alpha_{\rm p}$. For the threshold, the numerical data supports the same mean-field power-law behaviour of equation (38).

† In fact, for the perceptron with zero threshold and $\kappa > 0$, one finds $\epsilon > \frac{1}{2}$ for α large enough and the sets of correctly classified patterns for two solutions could be disjoint.

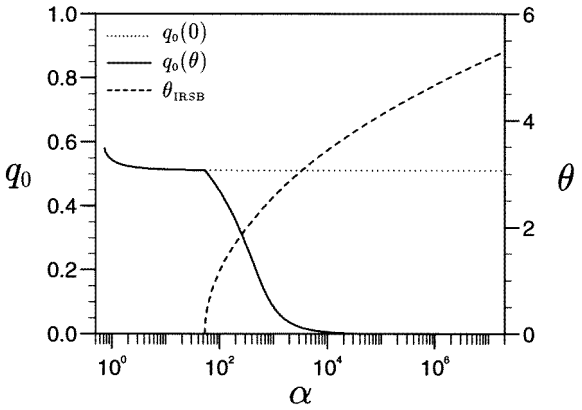

Figure 4. The one-step RSB overlap, q_0, of the Ising perceptron for the θ solution goes to zero as $\alpha \to \infty$, whereas it approaches a finite value ($q_0 = 0.511\,14$) for the θ_0 solution. The threshold as a function of α grows logarithmically to infinity.

In the asymptotic limit of infinite example load, we again find that the RSB overlap, q_0, approaches a finite limit for the θ_0 solution, which is κ dependent but always strictly less than 1, whereas it converges against zero for the θ solution following a power-law decay $q_0 \propto \alpha^{-1}$.

We further find for the Ising perceptron without threshold that the order parameter y approaches a finite value as q_0, whereas v, which is the equivalent of wx in the spherical case, decays as $v \propto 1/\sqrt{\alpha}$, similar to the spherical perceptron, indicating that the degree of RSB becomes more severe for increasing α.

We would like to point out that the asymptotic result of q_0 violates the qualitative argument in [5], which demands $q_0 \to 0$ for $\alpha \to \infty$, although it has been argued that one-step RSB may be exact for the Ising perceptron. In order to exclude with certainty that no solution to the saddlepoint exists which is characterized by $q_0 \to 0$, we have carried out substantial numerical and analytical work for the special case $\kappa = 0$ even for $\alpha > 10^{10}$, where the numerical solutions to the saddlepoint equations (33) become unreliable due to the inherent inaccuracy of the numerical integrations. The saddlepoint equations were expanded in a Taylor series in v, for which the dominant terms of all integrals can be solved analytically for $\kappa = 0$. This expansion was in excellent agreement with previous results and also provided accurate results for α values, where the solutions to the full equations were inaccurate. However, an extensive numeric search for solutions with q_0 and y small was unsuccessful even for $\alpha > 10^{200}$. This could be confirmed by the fact that algebraic saddlepoint equations, obtained by expanding the equations further for small q_0 and y, have only unphysical complex roots.

In the numerical analysis for the θ solution, it is again difficult to find the exact power-law exponents and possible logarithmic corrections. However, we find exact relationships between order parameters. The conjugate order parameter, y, decays as $1/\sqrt{\alpha}$. This suggests a relationship with q_0 as $y^2 \propto \hat{q}_0$, and indeed we find $q_0/y^2 \sim 1$ for large α. The order parameter, v, diverges logarithmically in α and we find $v/\theta \sim 2\kappa$ as the asymptotic behaviour, again indicating that the degree of RSB of the θ solution decreases for large α.

These functional relationships can be confirmed by a series expansion of the free energy around $q_0 = 0$ and $y = 0$, followed by an asymptotic expansion in θ and v, where we

assume† w.l.o.g. $\theta > 0$. The later expansion is, however, only valid in the region where $\theta - \kappa \gg 1$. The saddlepoint equations of $\partial f/\partial y$ and $\partial f/\partial\theta$ give to leading order $q_0 = y^2$ and $v = 2\kappa\theta$, in agreement with the numerical data. Inserting $\partial f/\partial v$ in $\partial f/\partial q_0$ gives

$$\sqrt{q_0} = y = \frac{\log(2)}{\kappa\sqrt{\alpha}}.$$

The remaining saddlepoint equation, $\partial f/\partial v$, determining θ,

$$\exp\left[-\frac{1}{2}(\theta-\kappa)^2\right] - \exp\left[-\frac{1}{2}(\theta+\kappa)^2\right] = \frac{\sqrt{2\pi}\log(2)}{\kappa\alpha} \tag{40}$$

does not have a closed-form solution. However, for $\theta\kappa \gg 0$ an approximate solution can be found

$$\theta \approx \kappa + \sqrt{2}\left[\log\left(\frac{\kappa\alpha}{\sqrt{2\pi}\log(2)}\right)\right]^{1/2}. \tag{41}$$

Whereas the analytical equations for y and q_0 and the solution of θ, obtained by solving equation (40) numerically, fit the numerical solutions of the full saddlepoint equations very well even for moderate values of $2 \leqslant \theta \leqslant 6$, the closed-form solution for θ (41) is only a good approximation for $\kappa \geqslant 1$ in this region.

3.3. Pattern stability distribution

The phase transition in order parameter space is driven by the increase of the error rate ϵ for increasing example load α. If is therefore natural to examine the change in the PSD of the perceptron around the critical load α_p. We first examine the PSD of the Ising perceptron as it has a simpler structure (it lacks the gap and the δ-contribution of the spherical case).

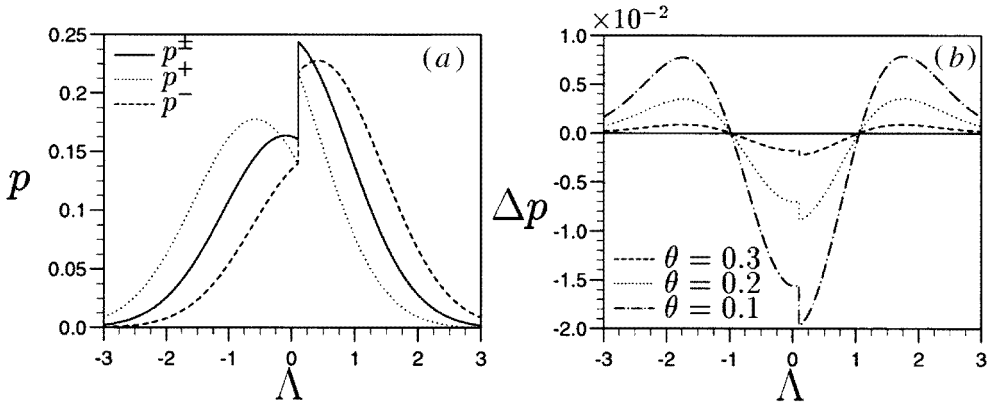

Figure 5. (*a*) The PSDs, $p(\Lambda)$, of the Ising perceptron is shown as a function of the pattern stability, Λ, for $\kappa = 0.1$ for an example load $\alpha(\theta = 0.5) = 59.492$ close to the phase transition point [$\alpha_p(\kappa = 0.1) = 53.021$]. The θ_0 solution predicts the same PSD $p^{\pm}$ for both $\sigma = +1$ and $\sigma = -1$ patterns. For the θ solution this symmetry is broken. (*b*) The difference in the total PSD ($\Delta p \equiv p^{+} + p^{-} - 2p^{\pm}$) as a function of Λ for various values of α : $\alpha(0.1) = 53.266$, $\alpha(0.2) = 54.008$ and $\alpha(0.3) = 55.266$. The asymmetry of $\Delta p(\Lambda)$ caused by the discontinuity at the decision boundary leads to the reduction in the error rate of the θ solution.

† For $\theta < 0$, one has to replace θ by $|\theta|$ in all the equations.

In figure 5(*a*), PSDs of the Ising perceptron for patterns with targets $\sigma = +1$ and $\sigma = -1$ are plotted for both the θ_0 and θ solutions for stability $\kappa = 0.1$. The example load α was chosen slightly larger than α_p and determined as a function of the value of threshold θ, e.g. in figure 5 $\alpha(\theta = 0.5) = 59.492$ (for comparison $\alpha_p(\kappa = 0.1) = 53.021$). The $\sigma = \pm 1$ PSD $p^{\pm}$ of the θ_0 solution are identical. For the θ solution this symmetry is broken and the PSDs p^+ and p^- are distorted around the former. For $\theta > 0$ the probability in the unstablized region, $\Lambda < \kappa$, has increased for $\sigma = +1$ patterns whereas it has reduced for $\sigma = -1$ patterns, and vice versa for the stabilized region $\Lambda \geqslant \kappa$.

All three distributions exhibit a discontinuity at the threshold stability κ which is formally due to the exponentional factor e^{-v} in equation (37). Although the functional form of the PSDs (36) is quite complicated, the PSDs have almost conserved the Gaussian form of the numerator. The means are shifted and dependent on $\sigma\theta$.

To assess the change in total error, it is more accurate to study the difference of the total PSD ($\Delta p \equiv (p^+ + p^-) - 2p^{\pm}$). In figure 5(*b*), Δp is shown for three values of α even closer to the critical point. One can see that the shift of the means of the θ PSDs removes probability mass from the region close to the decision threshold κ. Furthermore, $\Delta p(\Lambda)$ is almost symmetric around κ. If this symmetry were perfect, the total error rate could not be different for the θ_0 and θ solutions. However, we find a distortion in the region $\Lambda \approx \kappa$, which can be most easily depicted by the discontinuity at κ, which grows for increasing $\alpha(\theta)$.

We find quite similar results in the case of the spherical perceptron although due to the gap and the δ-contribution in the PSD lead to a more complex behaviour. To make the effect of these extra features more obivous, we have chosen a larger threshold stability, $\kappa = 1$, for the spherical case. In figure 6(*a*), the one-step RSB PSDs of patterns with targets $\sigma = +1$ and $\sigma = -1$ is shown for both solutions and as an example load of $\alpha(\theta = 0.5) = 2.0901$

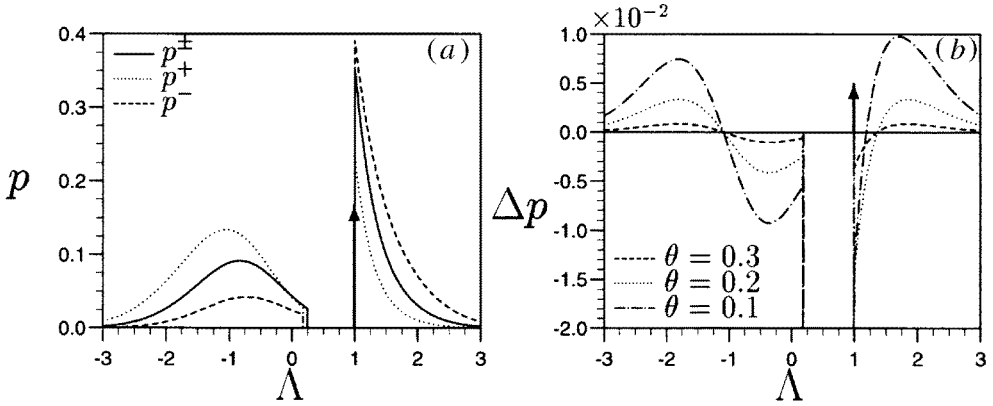

Figure 6. (*a*) The PSDs, $p(\Lambda)$, of the spherical perceptron as a function of the pattern stability, Λ, for $\kappa = 1$ for an example load $\alpha(\theta = 0.5) = 2.0901$ close to the phase transition point [$\alpha_p(\kappa = 1) = 1.8706$]. Again the θ_0 solution predicts the same PSD $p^{\pm}$ for both $\sigma = \pm 1$ patterns, whereas this symmetry is broken for the θ solution. The δ-peak is indicated by the arrows and its probability mass is given by $P_\delta^{\pm} = 9.5251 \times 10^{-2}$, $P_\delta^{+} = 9.1652 \times 10^{-2}$ and $P_\delta^{-} = 1.0563 \times 10^{-1}$. (*b*) The difference in the total PSD ($\Delta p \equiv p^+ + p^- - 2p^{\pm}$) as a function of Λ for various values of α : $\alpha(0.1) = 1.8790$ [$\Delta P_\delta = 2.3490 \times 10^{-4}$], $\alpha(0.2) = 1.9076$ [$\Delta P_\delta = 1.1915 \times 10^{-3}$], and $\alpha(0.3) = 1.9469$ [$\Delta P_\delta = 2.6226 \times 10^{-3}$]. The reduction in the error rate of the θ solution seems to be mainly caused by the increase of the gap.

(for comparison $\alpha_p(\kappa = 1) = 1.8706$). Again we find that the $\sigma = \pm 1$ PSDs of the θ solution are distorted around the PSD of the θ_0 solution.

The distributions have three components. For $\Lambda < \kappa$, the distribution looks similar to a Gaussian hump with means which vary with the value of $-\theta$. This regime is separated by a visible gap to the stabilized patterns, with a gap width which is widened for the θ solution. One further finds that the contribution of the δ-functions at $\Lambda = \kappa$ has increased for the θ solution. The main probability mass of the stabilized patterns is found in the Gaussian-like tail for $\Lambda > \kappa$.

To study the differences of the PSDs, we further show $\Delta p(\Lambda)$ for three values of α closer to α_p in figure 6(*b*). We find less symmetry in Δp than for the Ising perceptron, but again total probability mass has been removed from the vicinity of $\Lambda = \kappa$. The main reduction in the error rate in this case seems to come from the widening of the gap. This difference in probability mass has been partly shifted to the δ-contributions. The increase of probability mass at the δ-peaks and the decrease of probability mass at the widened gap is, however, between a factor of 10–100 times larger (and increasing for $\alpha \to \alpha_p$) than the reduction in the error rate for the α values studied in figure 6(*b*).

It is further interesting to study the limit $\alpha \to \infty$ as the error rate of the θ_0 solution approaches its asymptotic value, which is larger than the asymptotic error rate of the θ solution of $\frac{1}{2}$ as was shown in both figures 1 and 3. The θ solutions in the limit of infinite example load has been shown to be characterized by a threshold increasing to infinity and the portion of *wrongly on* errors decreasing rapidly to zero (see e.g. figures 1 and 2).

To study this limit more closely, we show the PSDs of the spherical perceptron in the one-step RSB ansatz for $\kappa = 1$ and increasing α separately from the θ_0 and θ solutions in figure 7. For the θ_0 solution (which is equivalent to the perceptron without threshold), both

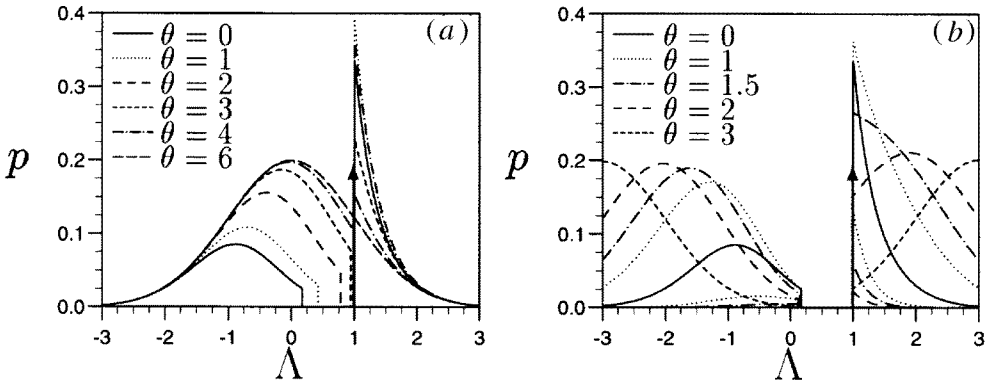

Figure 7. The PSDs, $p(\Lambda)$, of the spherical perceptron as a function of the pattern stability, Λ, for $\kappa = 1$ and increasing example load α. (*a*) The PSD of the θ_0 solution and $\alpha(\theta = 0) = \alpha_p = 1.8706$ [$P_\delta^\pm = 1.1029 \times 10^{-1}$], $\alpha(1) = 2.8878$ [$P_\delta^\pm = 6.1194 \times 10^{-2}$], $\alpha(2) = 10.800$ [$P_\delta^\pm = 1.1017 \times 10^{-2}$], $\alpha(3) = 85.059$ [$P_\delta^\pm = 9.2650 \times 10^{-4}$], $\alpha(4) = 1385.9$ [$P_\delta^\pm = 5.1225 \times 10^{-5}$] and $\alpha(\theta = 6) = 6.2488 \times 10^5$ [$P_\delta^\pm = 3.4349 \times 10^{-8}$]. The total PSD of both $\sigma = \pm 1$ patterns approaches the zero mean unit variance Gaussian distribution. (*b*) Both PSDs of the θ solution for a range of α values (see above and $\alpha(\theta = 1.5) = 4.0890$). The δ-contributions to the $\sigma = \pm 1$-PSDs for $\theta > 0$ are given by (in order of increasing threshold): [$P_\delta^+ = 6.0682 \times 10^{-2}$; $P_\delta^- = 8.0595 \times 10^{-2}$], [$P_\delta^+ = 3.2087 \times 10^{-2}$; $P_\delta^- = 4.9147 \times 10^{-2}$], [$P_\delta^+ = 1.3589 \times 10^{-2}$; $P_\delta^- = 2.4065 \times 10^{-2}$], [$P_\delta^+ = 1.1555 \times 10^{-3}$; $P_\delta^- = 2.7075 \times 10^{-3}$]. Both PSDs approach half of the probability mass of a unit variance Gaussian distribution centred at $\sigma\theta$.

PSDs approach half the probability mass of a Gaussian distribution with zero mean and unit variance. This is as expected, since the examples are uniformly distributed spatially and a random-weight vector on the hypersphere has an average overlap (activation) with the examples which is Gaussian distributed. As all examples with absolute activation smaller than κ are always counted as erroneous, the error rate approaches $\epsilon = 1 - H(\kappa) \geqslant \frac{1}{2}$ in the $\alpha \to \infty$ limit†.

For the θ solution on the other hand, both PSDs also approach (half the probability masses of) unit variance Gaussian distributions but with means centred around $\sigma\theta$. Although any weight vector will have a Gaussian distributed overlap, the activation is shifted due to large threshold. This means that for infinite α, the θ solution classifies the examples deterministically as either all $+1$ or -1 depending on the sign of the (infinite) threshold, resulting in an total error rate of $\frac{1}{2}$ irrespective of the stability κ.

One can assess the convergence rate of the error rate of the perceptron against the asymptotic error rate ϵ^∞ from the numerical solutions of the saddlepoint equations. For the θ_0 solution, we find within the RS ansatz (independent of the weight prior), and within the one-step RSB ansatz for spherical and Ising perceptron respectively

$$\epsilon^\infty - \epsilon_{\mathrm{RS}} \propto \alpha^{-0.3333\pm1} \qquad \epsilon^\infty - \epsilon_{\mathrm{RSB}} \propto \alpha^{-0.490\pm5} \qquad \text{and} \qquad \epsilon^\infty - \epsilon_{\mathrm{IRSB}} \propto \alpha^{-0.500\pm1}$$

where the error indicates the uncertainty in the last significant digit only. The different exponent in the power law for Ising and spherical perceptron in the one-step RSB ansatz is due to a logarithmic correction in the spherical case, as can be confirmed by using the results for the expansions of the saddlepoint equations (39) to calculate the asymptotic error of the spherical perceptron in the RS and similarly the one-step RSB ansatz

$$\epsilon^\infty - \epsilon_{\mathrm{RS}} = \frac{1}{2}\left[\frac{12\mathrm{e}^{-\kappa^2}}{\pi\alpha}\right]^{1/3} \qquad \text{and} \qquad \epsilon^\infty - \epsilon_{\mathrm{RSB}} = \frac{\mathrm{e}^{-\kappa^2/4}}{\sqrt{\pi}}\frac{[\log\alpha]^{1/4}}{\sqrt{\alpha}}. \tag{42}$$

For the θ solution we find to similarly for the total error rate

$$\tfrac{1}{2} - \epsilon_{\mathrm{RS}} \propto \alpha^{-1.0002\pm2} \qquad \tfrac{1}{2} - \epsilon_{\mathrm{RSB}} \propto \alpha^{-1.04\pm4} \qquad \text{and} \qquad \tfrac{1}{2} - \epsilon_{\mathrm{IRSB}} \propto \alpha^{-1.04\pm4}$$

for the three cases respectively. For the θ solution it was again difficult to measure the powers in the one-step RSB cases very accurately due to possible logarithmic corrections. This is supported by comparing the numerical predictions to our analytical results for the Ising perceptron, where we find to leading order

$$\frac{1}{2} - \epsilon_{\mathrm{IRSB}} = \frac{\log(2)}{2\kappa\theta_{\mathrm{s}}}\frac{1}{\alpha} \tag{43}$$

where θ_{s} is the solution for the threshold from equation (40) or its approximation (41), which gives a logarithmic correction to the power law with exponent 1.

Comparing the predictions of the power-law decay of the error rate between the θ_0 and θ solutions, one notes two important differences. First, the exponent of the decay is twice as large for the θ solution, where the error decays linearly with α, and a slower convergence for the θ_0 solution with $\sqrt{\alpha}$. Second, the correction of the θ solution going from RS to one-step RSB is only minor, a logarithmic term, whereas it is substantial for the θ_0 solution, a change in the exponent from $\frac{1}{3}$ to $\frac{1}{2}$. This suggests that the effect of RSB for large α is more severe for the perceptron without threshold than with threshold. It also may indicate that the effect of further RSB breaking should be less pronounced for the θ solution than for the θ_0 solution.

† This means that any random-weight vector on the hypersphere has the same error for $\alpha = \infty$. In the case of $\kappa = 0$ this corresponds to random guessing of the output with 50% chance of success.

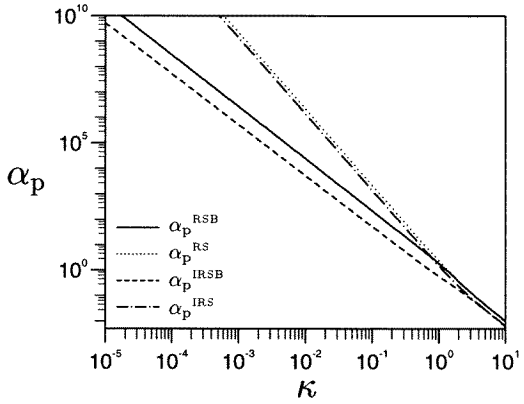

Figure 8. The critical normalized example number, α_p, as a function of the stability, κ, on a log–log scale shows a power-law behaviour for small stability. The predicted power-law behaviour using one-step RSB is significantly different to the one predicted from RS.

3.4. The stability dependence of the phase transition

In this section we will examine the dependence of the phase-transition point in order-parameter solution space on the threshold stability κ. In figure 8, α_p is plotted versus κ on a log–log scale for both spherical and Ising perceptron in the RS and one-step RSB ansatz. The critical point, α_p, in solution space increases for decreasing stability but exists for all non-zero stabilities, and exhibits a power-law dependence on κ for small stabilities with $\alpha_p \to \infty$ as $\kappa \to 0$. The numerical data predicts the exponents of the power laws as

$$\alpha_p^{RS} \propto \kappa^{-3.000\pm1} \qquad \alpha_p^{RSB} \propto \kappa^{-2.04\pm2} \qquad \text{and} \qquad \alpha_p^{IRSB} \propto \kappa^{-2.0000\pm1}$$

where the RS theory of the Ising perceptron only rescales the prefactor with the constant $2/\pi$.

From figure 8, we can further conclude that the phase transition exists for all finite stabilities $\kappa > 0$. The limits $\kappa \to 0$ and $\alpha \to \infty$ are therefore *not* interchangeable, i.e. the 'point' $\{\kappa = 0, \alpha = \infty\}$ is an unstable fixed point. Although, $\kappa = 0$ would have an error rate of $\frac{1}{2}$ at $\alpha = \infty$ irrespective of the threshold, only the θ_0 solution is accessible to the perceptron for any finite α and it has no access to the θ solution for $\alpha \to \infty$.

As the phase transition seems to be triggered by the increase of the error rate above a critical value, we also show the error rate $\epsilon_p = \epsilon(\alpha_p)$ at the critical load, together with its deviation from the asymptotic error rate $\frac{1}{2}$ in figure 9. One can see that the stability has a dominant influence on the occurrence of the phase transition through the error rate. For large stabilities the θ_0 solution becomes already unstable for small error rates, with the limit $\epsilon_p \to 0$ for $\kappa \to \infty$. The difference in the critical error rate between the Ising and spherical perceptron is greatest for moderate stabilities $\kappa \approx 1$, which may be attributed to the gap and the δ-contribution in the PSD of the spherical perceptron.

The RS theory not only underestimates the error for a given load α, and, therefore, gives the incorrect power law for α_p, but also fails to predict the correct critical error rate. RS fails especially for smaller stabilities, i.e. large α as expected. This is especially obvious by looking at the remnant error rate in figure 9, which decays with a power law. The exponents can be also evaluated from the numerical data:

$$\tfrac{1}{2} - \epsilon_p^{RS} \propto \kappa^{1.000\pm1} \qquad \tfrac{1}{2} - \epsilon_p^{RSB} \propto \kappa^{0.993\pm2} \qquad \text{and} \qquad \tfrac{1}{2} - \epsilon_p^{IRSB} \propto \kappa^{1.000\,00\pm1}.$$

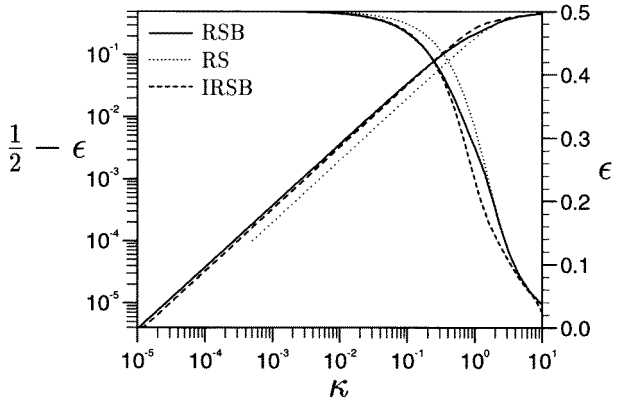

Figure 9. The error rate, $\epsilon(\alpha_p)$, and its deviation from the asymptotic error rate $\frac{1}{2}$ is shown as a function of the stability, κ, on log–lin and log–log scales respectively. The remnant error rate $\frac{1}{2} - \epsilon$ shows a power-law decay for small κ. For larger stabilities, the phase transition occurs for increasingly small error rates.

Although RS seems to give a reasonable power-law decay of the error, the prefactor is blatantly incorrect. An asymptotic expansion for small thresholds and stabilities for the RS theory gives

$$\alpha_p^{RS} = \frac{8}{9}\frac{\sqrt{2\pi}}{\kappa^3} \qquad \text{and} \qquad \tfrac{1}{2} - \epsilon_p^{RS} = \frac{\kappa}{2\sqrt{2\pi}}. \tag{44}$$

Of more interest is the functional behaviour of the one-step RSB solution for small stabilities, as the numerical solutions indicate a deviation from the pure power-law behaviour in both the point of the phase transition as well as the asymptotic error. A similar analytic expansion gives

$$\frac{\alpha_p^{RSB}}{\sqrt{\log \alpha_p^{RSB}}} = \frac{1}{2\kappa^2} \tag{45}$$

for which a closed-form solution does not exist. However, one can see that the deviation from the pure κ^{-2} power-law behaviour of α_p is due to the additional logarithmic term in α_p.

For the Ising perceptron it is not possible to expand all of the equations as the order parameters y and q_0 have finite limits. However, the numerical solutions themselves give us some insight. For the Ising perceptron there is no numerical indication that the critical load, α_p, or its error, ϵ_p, deviate from pure power-law behaviours in contrast with the spherical case, which exhibit logarithmic corrections. Furthermore, for large stabilities the phase transition occurs at a smaller error rate for the Ising than for the spherical perceptron, whereas this characteristic is reversed for small stabilities, where the phase transition occurs at a larger error rate. These differences between the two weight priors could either be attributed to their respective weight-space structures, or it may indicate that one-step RSB is correct in the Ising and incorrect in the spherical case.

3.5. Non-zero output bias, m_o

For non-zero output bias, m_o, the symmetry in the space of thresholds θ is broken and we find only solutions with $\theta \neq 0$ for all α, with $\theta > 0$ for $m_o < 0$ and vice verse. Due to the

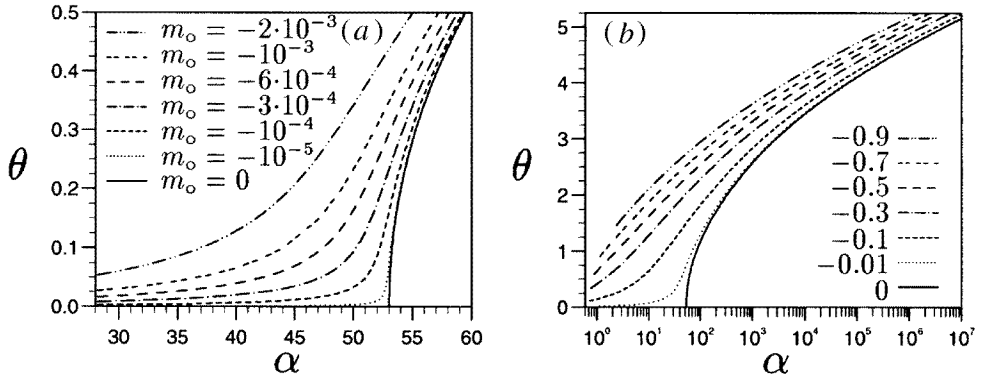

Figure 10. (*a*) The evolution of the threshold, θ, with the example load, α, is shown for several small values of the bias (see legend) around the critical load, α_p, with constant stability $\kappa = 0.1$. The phase transition is increasingly smeared out for growing magnitude of the bias. (*b*) The evolution of $\theta(\alpha)$ over a wide range of α for larger magnitudes of the bias, m_o, shows the same effect. The left-hand starting point of each curve depicts the capacity limit, α_c, increasing with growing magnitude of the bias.

symmetry of the solutions for $m_o \to -m_o \Rightarrow \theta \to -\theta$, one can assume $m_o < 0$ and $\theta > 0$ without loss of generality. Below, we will discuss only the Ising perceptron as we found the behaviour for both binary and real weights to be quite generic.

In figure 10 the threshold of the Ising perceptron is shown as a function of α for various values of the output bias, m_o, at fixed stability $\kappa = 0.1$. In figure 10(*a*), one sees that, for a very small magnitude of the bias, the evolution of the threshold closely approaches the curve for zero bias. Similar behaviour can also be found for the other order parameters. The largest deviations between the zero-bias solution and the finite-bias solution can always be found around the point of the phase transition at α_p. In this sense, the output bias, m_o, can be seen as an external field which 'smears' out the phase transition.

In figure 10(*b*), we show the evolution of the threshold θ for larger magnitudes of the bias over a wide range of loads α. For large α the threshold tends to infinity, whereas the left-hand starting point of each curve depicts the capacity limit, α_c, increasing with increasing magnitude of the bias.

For large α, one can expand the free energy of the Ising perceptron, similarly to the zero-bias case. One finds that the leading order of $\partial f/\partial y$ gives $q_0 = y^2$ as for zero bias case. The leading order of $\partial f/\partial\theta$ implies

$$v = 2|\theta_s|\left[\kappa + \frac{1}{2|\theta_s|}\log\left(\frac{1+|m_o|}{1-|m_o|}\right)\right] = 2|\theta_s|\kappa^* \tag{46}$$

where θ_s is the solution of the threshold for a given load α and κ^* is a modified effective stability, which depends on the bias and on the solution of the threshold (i.e. ultimately on α). Further inserting $\partial f/\partial v$ in $\partial f/\partial q_0$ yields

$$\sqrt{q_0} = y = \frac{\log(2)}{\sqrt{1-m_o^2\kappa^*}}\frac{1}{\sqrt{\alpha}}. \tag{47}$$

The remaining saddlepoint equation $\partial f/\partial v$ to determine θ_s is given by

$$(1+|m_o|)\exp\left[-\frac{1}{2}(|\theta|-\kappa)^2\right] - (1-|m_o|)\exp\left[-\frac{1}{2}(|\theta|+\kappa)^2\right] = \frac{\sqrt{2\pi}\log(2)}{\kappa^*\alpha} \tag{48}$$

and cannot be solved for θ in closed form. The approximation used in the zero-bias case in equations (40) and (41) (see section 3.2), which neglects the less dominant term on the left-hand side of equation (48), still does not make a closed-form solution feasible, due to the θ dependence of κ^*.

For the asymptotic error rate one finds $\epsilon^\infty = \frac{1}{2}(1 - |m_o|)$ irrespective of the stability κ—the intuitive result if one classifies the larger call of example correctly and misclassifies the smaller example class by using a threshold of infinite absolute value. The asymptotic error rate is approached via

$$\epsilon^\infty - \epsilon_{\mathrm{IRSB}} = \frac{\log(2)}{2|\theta_s|\kappa^*}\frac{1}{\alpha}. \tag{49}$$

As both θ_s and κ^* are dependent on α, the asymptotic behaviour deviates from a pure power-law behaviour.

4. Summary and conclusions

In this paper we have investigated the threshold Boolean perceptron above saturation for both spherical and binary weight priors. Even for unbiased input and output distributions, we find that the introduction of a threshold triggers interesting phenomena for finite stabilities $\kappa > 0$ which are not otherwise present. Namely, we find a second-order phase transition in order-parameter space at a stability-dependent critical load, $\alpha_p(\kappa)$, with spontaneous symmetry breaking in the space of thresholds θ. This phase transition is driven by the error rate as we find that the perceptron without threshold exhibits a higher asymptotic error ($\epsilon^\infty = 1 - H(\kappa)$) than the perceptron with threshold ($\epsilon^\infty = \frac{1}{2}$).

Zero stability $\kappa = 0$ constitutes a special case, as one does not find a phase transition for finite α. This means that the limits $\kappa \to 0$ and $\alpha \to \infty$ are *not* interchangeable and the 'point' $\{\kappa = 0, \alpha = \infty\}$ is an unstable fixed point. One could argue that this point is in fact a first-order parameter transition, leading to a discontinuous jump in order-parameter space.

Further we have identified the bias of the output distribution, m_o, with the external magnetic field in spin systems that breaks the symmetry in θ space and 'smears' out the phase transition. Whereas a non-zero output bias has, therefore, a profound effect on the performance of perceptrons, we find that a non-zero input bias can always be absorbed by a rescaling of the target stability κ. These results also suggest that one should not remove the threshold in favour of a ferromagnetic bias in the couplings as we have found that a threshold can always compensate for this bias but not vice versa.

In the asymptotic limit $\alpha \to \infty$ and finite stability $\kappa > 0$, we not only find unequal values for the asymptotic error rate but strikingly different solutions in order-parameter space for the perceptron with and without threshold, especially, for the asymptotics of the one-step RSB overlap q_0. In the case of the spherical weight constraint, we find that q_0 approaches 1 for the perceptron without threshold, whereas q_0 decays to 0 for the perceptron with threshold. For the Ising perceptron we find a similar behaviour: the solution with non-zero threshold is characterized by a vanishing overlap q_0 for increasing α and the solution with zero threshold exhibits a finite limit of q_0 for infinite load which is stability dependent and strictly smaller than 1.

It has been argued previously [5] that the above asymptotic behaviour for the spherical perceptron without threshold indicates that one-step RSB cannot be exact at high load. For a correct solution one would expect the smallest overlap scale q_0 to approach 0 for $\alpha \to \infty$ as in this limit any weight vector should perform equally well.

Recently, it has been shown by performing a two-step RSB calculation [18] that one-step RSB is indeed inexact for the spherical perceptron without threshold. Furthermore, it has been proved [18] that any model with a gap in the PSD (such as the spherical perceptron with or without threshold and Gardner–Derrida cost function) necessitates infinitely many RSB steps to yield the exact result.

These findings give some support to the validity of the qualitative argument made above. A strict application of this argument would imply that one-step RSB is also inexact for the Ising constraint, which has been the source of some debate [6, 24, 25]. As the PSD of the Ising perceptron with the Gardner–Derrida cost function does not exhibit a gap, the proof in [18] is not able to resolve this issue.

We have some doubts if one can have enough confidence in the qualitative argument of [5] to argue that one-step RSB is incorrect in the Ising model. First, we believe that one should be very careful to apply such an intuitive argument to models with discrete weights. For example, whereas all overlaps in the spherical model converge to 1 at the capacity limit, leaving just a single solution, the smallest overlap scale q_0 remains finite but strictly smaller than 1 for the Ising model, which is initially not really intuitive (see [6] for a plausible explanation), as it suggests several solutions at the capacity limit. A similar effect may be present in the limit $\alpha \to \infty$. Second, one may argue, that the argument of [5] can demand $q_0 = 0$ strictly only at $\alpha = \infty$, whereas it implicitly assumes a smooth transition of $q_0 \to 0$ for $\alpha \to \infty$, which does not take into account the possibility of a discontinuous transition. We have arguably found a possibility for such a discontinuous transition for the case $\kappa = 0$ at $\alpha = \infty$, from the θ_0 solution with $q_0 = 1$ to the θ solution with $q_0 = 0$.

To resolve the issue of the exactness of one-step RSB in the Ising perceptron with Gardner–Derrida cost function, it may be worthwhile to re-examine the two-step RSB solution in [6] numerically for large α and/or to calculate the stability of the one-step RSB solution.

Nevertheless, results concerning the asymptotic behaviour of the error rate and the order parameters in this paper suggest that the effect of further RSB breaking may be even smaller for both the Ising and the spherical perceptron with threshold in the regime of the θ solution than has been found for the θ_0 solution of the spherical perceptron in [18]. The one-step RSB solution may therefore remain sufficiently accurate for many practical purposes like calculating the capacity of multilayer networks produced by constructive algorithms [10, 11], where a treatment with a two-step RSB solution is computationally not feasible.

Acknowledgments

AHLW would like to gratefully acknowledge financial support by the EPSRC, a research scholarship of the Department of Physics at the University of Edinburgh, and the financial support and hospitality of the Neural Computing Research Group at Aston University, where part of this research was carried out. This research was further supported financially by EU grant ERB CHRX-CT92-0063. The authors would like to thank Peter Sollich, and especially Andreas Engel for interesting discussions. We would also like to thank David Barber for helpful comments and careful reading of the manuscript.

References

[1] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257–70
[2] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499–556
[3] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056–91

[4] Erichsen Jr R and Thuemann W K 1993 *J. Phys. A: Math. Gen.* **26** L61–8
[5] Majer P, Engel A and Zippelius A 1993 *J. Phys. A: Math. Gen.* **26** 7405–16
[6] Krauth W and Mézard M 1989 *J. Physique* **20** 3057–66
[7] Wendemuth A, Opper M and Kinzel W 1993 *J. Phys. A: Math. Gen.* **26** 3165–85
[8] Mézard M and Nadal J-P 1989 *J. Phys. A: Math. Gen.* **22** 2191–203
Nadal J-P 1989 *Int. J. Neural Syst.* **1** 55–9
[9] Frean M 1990 *Neural Comput.* **2** 198–209
[10] West A H L and Saad D 1997 *Mathematics of Neural Networks: Models, Algorithms and Applications* ed S W Ellacott *et al* (Dordrecht: Kluwer)
[11] West A H L and Saad D 1997 in preparation
[12] Mitchison G J and Durbin R M 1989 *Biol. Cybernet.* **60** 345–56
[13] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146–61
[14] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 *Phys. Rev.* A **45** 7590–609
[15] Saad D 1994 *J. Phys. A: Math. Gen.* **27** 2719–34
[16] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271–84
Bouten M 1994 *J. Phys. A: Math. Gen.* **27** 6021–3
Derrida B 1994 *J. Phys. A: Math. Gen.* **27** 6025
[17] Mézard M, Parisi G and Virasoro M G 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[18] Whyte W and Sherrington D 1996 *J. Phys. A: Math. Gen.* **29** 3063–73
[19] Kepler T B and Abbott L F 1988 *J. Physique* **50** 3057–66
[20] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969–74
[21] Griniasty M and Gutfreund H 1990 *J. Phys. A: Math. Gen.* **24** 715–34
[22] Wendemuth A 1995 *J. Phys. A: Math. Gen.* **28** 5423–36
[23] Wendemuth A 1995 *J. Phys. A: Math. Gen.* **28** 5485–93
[24] Fontanari J F and Meir R 1993 *J. Phys. A: Math. Gen.* **26** 1077–89
[25] Horner H 1992 *Z. Phys.* B **86** 291–308
[26] Gross D J and Mézard M 1984 *Nucl. Phys.* B **240** 431–52