One-step replica-symmetry-breaking solution for a perceptron learning with weight mismatch

One-step replica-symmetry-breaking solution for a perceptron learning with weight mismatch

Kibeom Park[†] C Kwon[‡] and Youngah Park[‡]

[†] Department of Physics, Seoul National University, Seoul 151-742, Korea
[‡] Department of Physics, Myoung Ji University, Yongin, Kyonggi-do, Seoul 449-728, Korea

Received 31 May 1995, in final form 1 December 1995

Abstract. We investigate the properties of the one-step replica-symmetry-breaking (1RSB) solution for a perceptron learning from examples with weight mismatch where the entropy zero line crosses the Almeida–Thouless (AT) line of the RS solution. For a small number of examples we find the optimal 1RSB solution which has the maximum free energy, non-negative entropy and satisfies the stability condition, the AT criterion for the 1RSB solution. The transition from RS to 1RSB is continuous or discontinuous depending on whether the RS AT line is above or below the zero entropy line. However, for a relatively large number of examples, the 1RSB solution which maximizes the free energy becomes unstable, and should be replaced by higher-step RSB solutions. We also obtain the AT line for the 1RSB solution.

1. Introduction

Feedforward layered networks, so-called perceptrons, are considered as having an appropriate neural architecture where various learning mechanisms can be studied [1–4]. Gardner [5] showed that a statistical–mechanical approach can be useful for studying the properties of feedforward networks, whence there have been many valuable results in studies of storage capacity [5–7] and learning [8–10]. Most of the studies have been done for a single-layer perceptron, the simplest feedforward network. Recently, some progress in studies on a double-layer perceptron, the committee machine, has been reported [11–14]. Further studies on a more realistic network, with more than two layers or multiple outputs, are in progress [15].

In this paper, we revisit the problem of learning in a single-layer perceptron, where there was an unresolved question about the nature of the low-temperature solution. The network is composed of N input nodes, N synapses with weights W_i $(i = 1, \ldots, N)$, and a single output node. A teacher network provides P sets of examples in the form of input–output pairs, $(\boldsymbol{S}^l, \sigma_0(\boldsymbol{S}^l))$, with $l = 1, \ldots, P$. Input variables S_i^l, interpreted as a problem (pattern), are ith components of vectors $\boldsymbol{S}^l$. The resultant output σ_0^l, as the answer (pattern code), is generated via synaptic weights and the transfer function g_0 as $\sigma_0^l = g_0(N^{-1/2}\boldsymbol{S}^l \cdot \boldsymbol{W}^0)$. A student network with a given transfer function g is trained by tuning the synaptic weights to minimize the error of the trial answer $\sigma^l = g(N^{-1/2}\boldsymbol{S}^l \cdot \boldsymbol{W})$ from the correct one given

0305-4470/96/071397+13$19.50

by the teacher network. The error function E is defined as

$$E\left(\boldsymbol{W};\{\boldsymbol{S}^l\}\right) = \sum_{l=1}^{P} \epsilon(\boldsymbol{W};\boldsymbol{S}^l) = \sum_{l=1}^{P} \tfrac{1}{2}\left[g\ (N^{-1/2}\boldsymbol{S}^l\cdot\boldsymbol{W}) - g_0\ (N^{-1/2}\boldsymbol{S}^l\cdot\boldsymbol{W}^0)\right]^2 . \qquad (1.1)$$

The number P of examples is known to scale as αN so as to reach a learning stage. Regarding the error function in equation (1.1) as the Hamiltonian of a thermodynamic system, this learning problem becomes a statistical mechanics problem of a disordered system.

When the architectures of the teacher and the student network are different from each other, the student network might not reproduce the target output of the teacher network exactly. Among these unrealizable cases, the case of weight mismatch is interesting, where the student network has discrete weights, $W_i = \pm 1$, while the teacher has continuous ones. Output made by the student cannot be correct, instead there are many optimal answers with the same level of error, which correspond to many degenerate ground states in the sense of statistical mechanics. In fact, Seung *et al* found that a spin-glass phase exists in a low-temperature region. As an approximate estimate for the phase boundary of the spin-glass phase, they suggested the line of zero entropy below which the replica-symmetric (RS) solution has a negative entropy [9]. Particularly when the transfer function is Boolean, they found that the one-step replica-symmetry-breaking (1RSB) solution can be found very easily and proposed it as an exact solution below the zero entropy line as in the study of storage capacity [7]. Later, Kwon *et al* obtained an instability line corresponding to the Almeida–Thouless (AT) line of the RS solution and proposed it as the exact phase boundary of the spin-glass phase. They found the AT line when the transfer function is linear in figure 1 of [10], showing that there is a crossing between the AT line and the zero entropy line. We reproduce this in figure 1. We also find the AT line when the transfer function is Boolean in figure 2, which shows the AT line lies below the zero

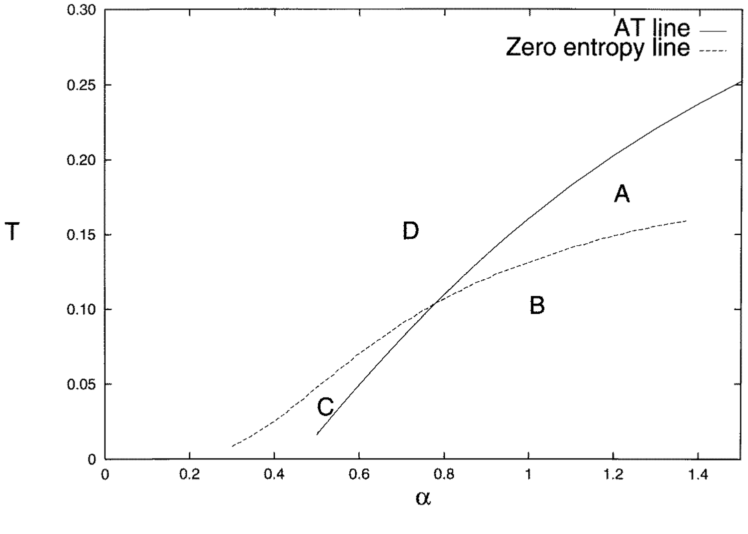

Figure 1. Zero entropy line and the AT line of the RS and 1RSB solution for the linear transfer function.

entropy line for all regions. It means that a discontinuous transition should occur from the RS solution across the zero entropy line, confirming the earlier proposition by Seung *et al*.

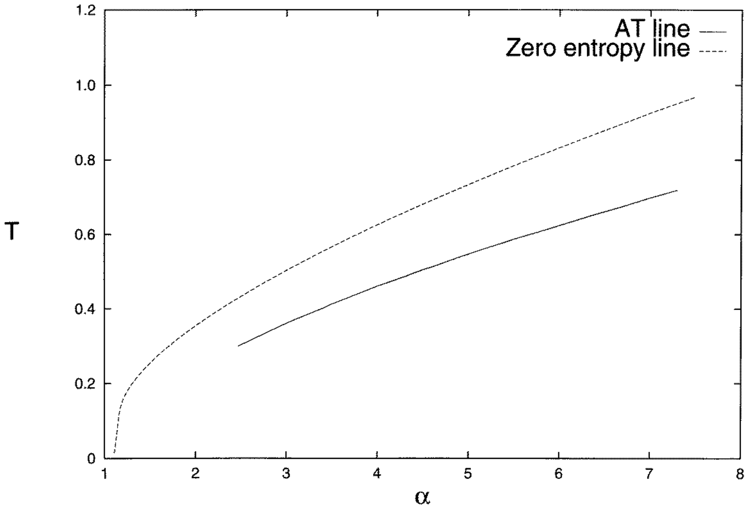

Figure 2. Zero entropy line and the AT line of the RS solution for the Boolean transfer function. The AT line drops to the $T = 0$ axis at $\alpha = 2.47$.

The case of the linear transfer function, however, left a problem unresolved. As seen in figure 1, the AT line is below the zero entropy line for small numbers of examples, so should not be regarded as the phase boundary. Note that one of the eigenvalues, determining the curvature of the free energy, vanishes at the AT line, signalling a continuous phase transition. In the region where the AT line is below the zero entropy line, there might be a first-order phase transition above the AT line. The first-order phase transition is determined by comparing the RS solution, correct at high temperatures, with the low-temperature solution having a certain broken replica symmetry. We consider the 1RSB solution as a candidate, at least as a more physical one than the RS one. In section 2, we derive the self-consistent equations for the 1RSB solution which by no means show a simple form in contrast to the case of the Boolean transfer function. In section 3, we derive the stability condition for the 1RSB solution in order to check whether it may be stable at low temperature. In section 4, we present our results which show quite complicated behaviour, not observed in the case of the Boolean transfer function. We find that there exist 1RSB solutions at a low temperature with non-negative entropy. Via the stability analysis of the 1RSB solution we find, however, the 1RSB solution which satisfies the stationarity with respect to the RSB parameter m in the Parisi replica-symmetry-breaking scheme [16] becomes unstable for a relatively large number of examples. We obtain the stability line for the 1RSB solution and show how transitions from the RS to the 1RSB solution occur either discontinuously or continuously in $q_1 - q_0$, according to the number of examples. We summarize our study in section 5.

2. One-step RSB solution

The free-energy functional f per neuron was found using the replica trick and the saddle-point method in the limit of large number N of input nodes, expressed as

$$
\begin{aligned}
&n\beta f[R_\sigma, \hat{R}_\sigma, Q_{\sigma\rho}, \hat{Q}_{\sigma\rho}] \\
&\quad = \sum_\sigma \hat{R}_\sigma R_\sigma + \sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} Q_{\sigma\rho} - \ln \mathrm{Tr}_{\{W^\sigma\}} \exp\left[\sum_\sigma \hat{R}_\sigma W^\sigma W^0 + \sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} W^\sigma W^0\right] \\
&\qquad -\alpha \ln \int \prod_\sigma \frac{\mathrm{d}x^\sigma\, \mathrm{d}\hat{x}^\sigma}{2\pi} \int \frac{\mathrm{d}y\, \mathrm{d}\hat{y}}{2\pi} \exp\left[-\tfrac{1}{2}\beta \sum_\sigma (g(x^\sigma) - g_0(y))^2\right. \\
&\qquad \left. + \sum_\sigma \mathrm{i}\hat{x}^\sigma x^\sigma - \mathrm{i}\hat{y} y + \sum_{\sigma<\rho} Q_{\sigma\rho} \hat{x}^\sigma \hat{x}^\rho - \tfrac{1}{2} \sum_\sigma \hat{x}_\sigma^2 - \hat{y} \sum_\sigma R_\sigma \hat{x}^\sigma - \tfrac{1}{2}\hat{y}^2 \right]. \qquad (2.1)
\end{aligned}
$$

Here, σ and ρ are replica indices and n is the number of replicas. The $n \to 0$ limit is taken afterwards. The weight W^0 of the teacher network is quenched, having a Gaussian distribution with a variance of unity. The weights W^σ of the replicated student networks are either $+1$ or -1. We restrict ourselves to the case of unrealizable learning solely due to weight mismatch. So, $g(x) = g_0(x) = x$. The free-energy functional is made stationary by the saddle-point solution given by the order parameters, $Q_{\sigma\rho}$ and R_σ, and their conjugate parameters, $\hat{Q}_{\sigma\rho}$ and $\hat{R}_\sigma$.

Now we assume the saddle-point solution has one-step replica-symmetry breaking. Following the Parisi replica-symmetry-breaking scheme [16], we divide $(n \times n)$ matrices, $Q_{\sigma\rho}$ and $\hat{Q}_{\sigma\rho}$, into $(n/m)^2$ blocks of size m. Then, the order parameter $Q_{\sigma\rho}$ can be written as

$$
Q_{\sigma\rho} = (1 - q_1)\delta_{\sigma\rho} + (q_1 - q_0)\epsilon_{\sigma\rho} + q_0 \qquad (2.2)
$$

where $\delta_{\sigma\rho}$ is the Kronecker delta function and the matrix $\epsilon_{\sigma\rho}$ is defined as

$$
\epsilon_{\sigma\rho} = \begin{cases} 1 & \text{if } \sigma \text{ and } \rho \text{ are in a diagonal block} \\ 0 & \text{otherwise}. \end{cases} \qquad (2.3)
$$

$\hat{Q}_{\sigma\rho}$ can be written in a similar way. The order parameters R_σ and $\hat{R}_\sigma$ are independent of replica indices, equal to R and $\hat{R}$, respectively. Inserting the 1RSB order parameters into (2.1), and taking the limit $n \to 0$, we can write the 1RSB free-energy functional f_{RSB} as

$$
\begin{aligned}
-\beta f_{\mathrm{RSB}} &= \ln 2 - \tfrac{1}{2}[1 + (m-1)q_1](\hat{q}_1 - \hat{q}_0) - \tfrac{1}{2}[1 - q_1 + m(q_1 - q_0)]\hat{q}_0 \\
&\quad - \frac{1}{2} \frac{R^2}{1 - q_1 + m(q_1 - q_0)} + \frac{1}{m} \int \mathrm{D}z \ln \int \mathrm{D}z_1 \cosh^m\left(\sqrt{\hat{q}_1 - \hat{q}_0}\, z_1 + \sqrt{\hat{q}_0}\, z\right) \\
&\quad + \alpha \left[-\frac{1}{2} \ln(1 + \beta(1 - q_1)) - \frac{1}{2m} \ln\left(1 + \frac{m\beta(q_1 - q_0)}{1 + \beta(1 - q_1)}\right) \right. \\
&\quad \left. - \frac{\beta(1 - 2R + q_0)}{2\,(1 + \beta(1 - q_1) + m\beta(q_1 - q_0))} \right] \qquad (2.4)
\end{aligned}
$$

where $\int \mathrm{D}x = \int \frac{\mathrm{d}x}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ is used and the change of variable $\hat{q}_0 + \hat{R}^2 \to \hat{q}_0$ is carried out. f_{RSB} is stationary at the saddle-point characterized by the 1RSB order parameters, $q_0, q_1, \hat{q}_0, \hat{q}_1, R$ and $\hat{R}$.

Then, the saddle-point equations are obtained from the stationary condition, written as

$$R = \frac{\alpha\beta\,(1-q_1+m(q_1-q_0))}{1+\beta(1-q_1)+m\beta(q_1-q_0)} \tag{2.5a}$$

$$q_0 = \int \mathrm{D}z \left(\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh Z}{\int \mathrm{D}z_1 \cosh^m Z}\right)^2 \tag{2.5b}$$

$$q_1 = \int \mathrm{D}z \frac{\int \mathrm{D}z_1 \cosh^m Z \tanh^2 Z}{\int \mathrm{D}z_1 \cosh^m Z} \tag{2.5c}$$

$$\hat{q}_0 = \frac{\alpha\beta^2(1-2R+q_0)}{[1+\beta(1-q_1)+m\beta(q_1-q_0)]^2} + \frac{R^2}{[1-q_1+m(q_1-q_0)]^2} \tag{2.5d}$$

$$\hat{q}_1 = \frac{\alpha\beta^2(q_1-q_0)}{\left[1+\beta(1-q_1)\right]\left[1+\beta(1-q_1)+m\beta(q_1-q_0)\right]} + \hat{q}_0 \tag{2.5e}$$

where $Z = \sqrt{\hat{q}_1-\hat{q}_0}z_1 + \sqrt{\hat{q}_0}z$ is used. The allowed range of m is $0 \leqslant m \leqslant 1$ as n goes to zero. It can be easily shown that after substituting $m = 0$ or $m = 1$, the free energy and entropy reduce to the RS results, i.e.

$$f_{\mathrm{RSB}}(m=0) = f_{\mathrm{RSB}}(m=1) = f_{\mathrm{RS}} \tag{2.6a}$$

$$S_{\mathrm{RSB}}(m=0) = S_{\mathrm{RSB}}(m=1) = S_{\mathrm{RS}}\,. \tag{2.6b}$$

When $m = 0$ ($m = 1$), the free energy and entropy do not depend on q_0 (q_1), and q_1 (q_0) plays the role of q_{RS}. Since the free energy is convex with respect to m, the 1RSB free energy is always higher than the RS free energy for $0 < m < 1$.

If an m state that has a fixed value of m is locally stable, it can be an appropriate solution for the system. In the thermodynamic limit, however, the optimum m which is relevant in the thermodynamic limit is known to maximize the free energy [17]. In order to obtain the 1RSB solution with the maximal allowed free energy, we need to find the range of m that gives the stable solution. Then the optimum 1RSB solution can be found for $0 < m < 1$. To study the stability of the 1RSB solution, we will derive the stability condition of the 1RSB solution in the next section.

3. Stability analysis of the one-step RSB solution

The stability of the solution can be examined by expanding the free-energy functional given in (2.1) with respect to variations $\delta R_\sigma, \delta\hat{R}_\sigma, \delta Q_{\sigma\rho}, \delta\hat{Q}_{\sigma\rho}$ of $R_\sigma, \hat{R}_\sigma, Q_{\sigma\rho}, \hat{Q}_{\sigma\rho}$ from the values of the one-step RSB solution. The fluctuation determining the stability of the solution is assumed to come only from $\delta Q_{\sigma\rho}, \delta\hat{Q}_{\sigma\rho}$ [10]. Up to second order in these, the variation of the free-energy functional is expressed as

$$\delta^2 n\beta f = \sum_{\sigma,\rho} \delta Q_{\sigma\rho}\delta\hat{Q}_{\sigma\rho} - \tfrac{1}{2}\sum_{\sigma,\rho}\sum_{\gamma,\delta}\hat{\Gamma}_{\sigma\rho,\gamma\delta}\delta\hat{Q}_{\sigma\rho}\delta\hat{Q}_{\gamma\delta} - \tfrac{1}{2}\alpha\sum_{\sigma,\rho}\sum_{\gamma,\delta}\Gamma_{\sigma\rho,\gamma\delta}\delta Q_{\sigma\rho}\delta Q_{\gamma\delta} \tag{3.1}$$

where the detailed expressions for $\hat{\Gamma}_{\sigma\rho,\gamma\delta}$, $\Gamma_{\sigma\rho,\gamma\delta}$ are given in appendix A.

Following the formalisms of [17], the eigenvalues responsible for the stability of the 1RSB saddle-point solution are associated with block fluctuations, $\sum_{\sigma,\rho}\delta Q_{\sigma\rho}$ and $\sum_{\sigma,\rho}\delta\hat{Q}_{\sigma\rho}$ where each of σ and ρ belongs to a block of size m. For these eigenvalues, we have a reduced stability matrix M,

$$M = \begin{bmatrix} \alpha\Lambda_0 & -1 \\ -1 & \hat{\Lambda}_0 \end{bmatrix} \tag{3.2}$$

where Λ_0, $\hat{\Lambda}_0$ are also given in detail in appendix A. Two eigenvalues are found as

$$\lambda_\pm = \tfrac{1}{2}[(\alpha\Lambda_0 + \hat{\Lambda}_0) \pm [(\alpha\Lambda_0 - \hat{\Lambda}_0)^2 + 4]^{1/2}]\,. \tag{3.3}$$

λ_- is always negative, while the sign of λ_+ may change. Then, the stability condition is given by

$$\Lambda = 1 - \alpha\Lambda_0\hat{\Lambda}_0 \geqslant 0\,. \tag{3.4}$$

The equality holds when λ_+ is equal to zero, giving the instability plane in the α–T–m space. For given α and T, the stability condition gives the allowed range of m. Hence we can find an optimal solution which has lowest free energy and non-negative entropy as far as satisfying the stability condition.

4. Results

We solve the self-consistent equations (2.5) numerically and plot the free energy and entropy for several values of α and temperatures in figures 3 and 4. The free energy is convex with respect to m. The entropy is also convex, and positive for any m above the zero entropy line in the α–T plane. Below it the entropy is only positive for $m_1 < m < m_2$, vanishing at $m = m_1, m_2$. We also calculate Λ given in (3.4) numerically, which is monotonically decreasing with m and intersects the m axis at $m = m_c$. As a result, a stable solution should be found for $m \leqslant m_c$.

In region A of figure 1 where the RS zero entropy line is above the RS AT line, we find the 1RSB solution whose free energy is maximum at $m = m_{\max}$ with $m_1 < m_{\max} < m_2$. Also we found that $m_{\max}$ is smaller than m_c so that the 1RSB solution which is stationary with respect to the RSB order parameter m satisfies the stability condition. As we approach the zero entropy line by increasing the temperature, $m_{\max}$ also increases and finally at the zero entropy line, $m_{\max}$ becomes one. Thus we conclude that there occurs a phase transition from the RS to the 1RSB along the RS zero entropy line. As $m_{\max}$ goes to one, $q_1 - q_0$ remains non-zero, which indicates that the transition occurs discontinuously. In regions B and C, the RS AT line is above the zero entropy line. In this region, q_0 goes to q_1 as we

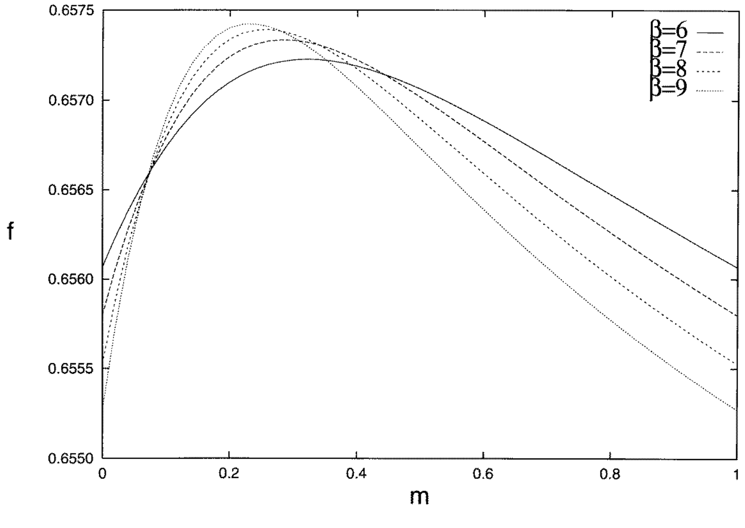

Figure 3. Free energy of the 1RSB solution as a function of m at $\alpha = 4.0$.

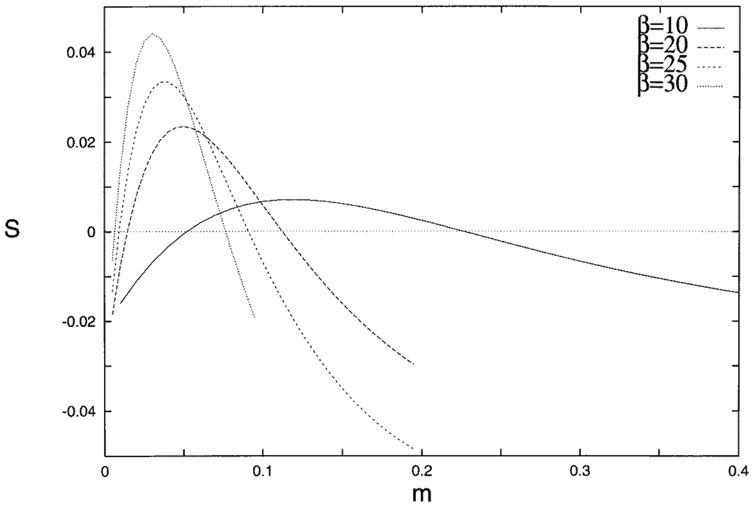

Figure 4. Entropy of the 1RSB solution as a function of m at $\alpha = 4.0$.

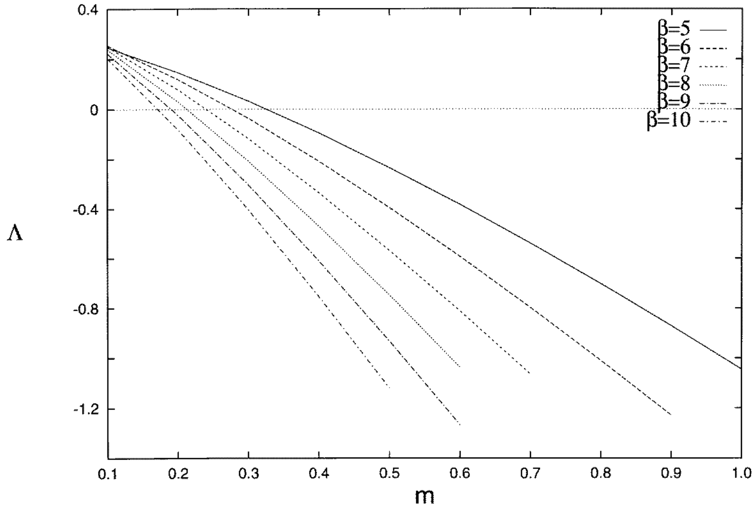

Figure 5. Right-hand side Λ of the stability condition in equation (3.4) as a function of m.

approach the RS AT line, which signals the continuous phase transition from the RS to the 1RSB phase along the RS AT line. The continuous transition line meets the discontinuous transition line at a tricritical point $\alpha_c = 0.79$.

However, when α is relatively large, quite a different behaviour occurs. We plot the free energy, entropy and the stability eigenvalue Λ for $\alpha = 4.0$ at several temperatures in figures 3–5. In this region, $m_{\max}$ is slightly larger than m_c so that the stationary 1RSB solution becomes unstable when the number of examples becomes large. We obtained the RSB AT line satisfying $m_{\max} = m_c$ numerically, and showed it in figure 1. The line crosses the RS AT line at $\alpha_0 = 1.18$. In the replica theory, it has been argued that the most relevant 1RSB solution among the m states is the one satisfying the stationarity with respect to m. If one wanted to find the stable RSB solution which is also stationary with respect to m, one

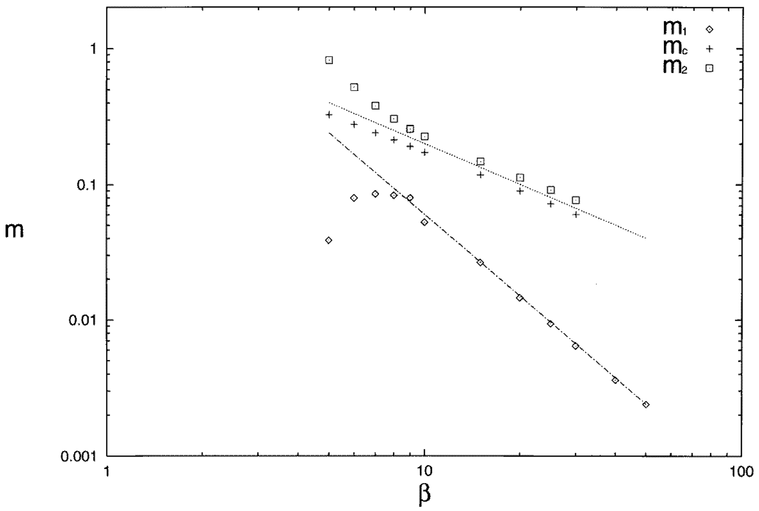

Figure 6. Temperature dependence of m_1, m_c and m_2. The slope of the upper line is 1, and that of the lower line is 2.

should go to the higher-step RSB scheme in regions D and E. Nevertheless, we expect that the high-step RSB solution will not be very much different from the 1RSB solution we found here, because $m_{\max}$ is very close to m_c.

Figure 6 shows the temperature dependence of m_1, m_2 and m_c as T approaches zero. As far as shown by numerical calculations, the m_c and m_2 are linear with T while m_1 goes as T^2. $m_{\max}$ lies between m_1 and m_2, so the entropy of the optimum 1RSB solution will go to zero as T approaches zero. $m_{\max}$ is exactly proportional to T in the Boolean case, which is the case only near $T = 0$ in the linear case. In the former, q_1 is exactly equal to unity, which is crucial in finding the RSB solution below the RS zero entropy line. On the other hand, q_1 goes to unity only as T goes to zero in the latter. Numerical analysis on the self-consistent equations obtained in section 2 becomes difficult near $T = 0$.

We can derive the free-energy functional and the self-consistent equations at zero temperature. We expect that $q_1 \to 1$ and $m \to 0$ as $T \to 0$. We include this result in appendix B. Again, the 1RSB solution at $T = 0$ can be obtained numerically. We do not include detailed numerical results in this paper. It seems enough for our purpose in this paper to note that the self-consistent equations depend only on $m\beta$. This agrees with the numerical observation that $m_{\max}$ is proportional to T as T goes to zero.

5. Conclusion

We revisit a perceptron learning problem where there is weight mismatch between the teacher and student networks. We study the property of the spin-glass phase with the 1RSB solution resulting from multi-degeneracy due to the weight mismatch. We derive the self-consistent equations for the 1RSB solution of block size m from the saddle-point condition of the free-energy functional. In order to examine the stability with respect to variation from the saddle point, we find the stability condition from one of the eigenvalues of the stability matrix associated with the variation of the free-energy functional. The 1RSB solution exists in the region where the RS solution becomes unphysical, and we obtain the 1RSB solution numerically. The 1RSB solution satisfying the stationarity with respect to the 1RSB parameter

m is stable only when the number of examples is small. When α is less than α_c the transition occurs discontinuously at the zero entropy line. When $\alpha_c < \alpha < \alpha_0$, the transition from the RS to the RSB solution occurs continuously along the RS AT line.

In summary, the one-step RSB solution can be adopted as a solution in the region where the RS solution becomes unphysical. When the number of examples becomes large, however, the 1RSB solution also becomes unstable, implying that one should go to the higher step RSB scheme for a more general solution.

Acknowledgments

We would like to thank Dr J-H Oh and Dr M Y Choi for useful discussions. One of authors (CK) would especially like to thank Dr H-J Sommers for informing us of his studies with Dr A Crisanti which stimulated this work. This work was supported in part by the KOSEF through the CTP of the Seoul National University, in part by the Non-Directed Fund,the Korea Research Foundation, 1993, and in part by the Ministry of Education through the Basic Science Research Institute of Seoul National University.

Appendix A

The stability analysis of the 1RSB solution for a spin-glass problem with p-spin interaction is well established by Crisanti and Sommers [17]. Their analysis on possible eigenmodes can be applied to our problem. We apply their formalism to our problem. We start with the expression of $\Gamma_{\sigma\rho,\gamma\delta}$ and $\hat{\Gamma}_{\sigma\rho,\gamma\delta}$ in (3.1), given by

$$\Gamma_{\sigma\rho,\gamma\delta} = \langle \hat{x}^\sigma \hat{x}^\rho \hat{x}^\gamma \hat{x}^\delta \rangle - \langle \hat{x}^\sigma \hat{x}^\rho \rangle \langle \hat{x}^\gamma \hat{x}^\delta \rangle \tag{A.1}$$

$$\hat{\Gamma}_{\sigma\rho,\gamma\delta} = \langle W^\sigma W^\rho W^\gamma W^\delta \rangle - \langle W^\sigma W^\rho \rangle \langle W^\gamma W^\delta \rangle \tag{A.2}$$

where

$$\langle \cdots \hat{x}^\sigma \cdots \rangle = \int \prod_\sigma \frac{\mathrm{d}x^\sigma \, \mathrm{d}\hat{x}^\sigma}{2\pi} \int \mathrm{D}y (\cdots \hat{x}^\sigma \cdots) \mathrm{e}^L \left(\int \prod_\sigma \frac{\mathrm{d}x^\sigma \, \mathrm{d}\hat{x}^\sigma}{2\pi} \int \mathrm{D}y \, \mathrm{e}^L \right)^{-1} \tag{A.3}$$

$$\langle \cdots W^\sigma \cdots \rangle = \frac{\mathrm{Tr}_{\{W^\sigma\}} (\cdots W^\sigma \cdots) \exp\left[\sum_\sigma \hat{R}_\sigma W^\sigma W^0 + \sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} W^\sigma W^0\right]}{\mathrm{Tr}_{\{W^\sigma\}} \exp\left[\sum_\sigma \hat{R}_\sigma W^\sigma W^0 + \sum_{\sigma<\rho} \hat{Q}_{\sigma\rho} W^\sigma W^0\right]} \tag{A.4}$$

with

$$L = -\frac{\beta}{2} \sum_\sigma \{g(x^\sigma) - g(y)\}^2 + \sum_\sigma \mathrm{i}\hat{x}^\sigma (x^\sigma - Ry) + \frac{R^2}{2} \left(\sum_\sigma \hat{x}^\sigma \right)^2 - \frac{1}{2} \sum_{\sigma,\rho} Q_{\sigma\rho} \hat{x}^\sigma \hat{x}^\rho \, . \tag{A.5}$$

There are two sources in variation, $Q_{\sigma\rho}$ and $\hat{Q}_{\sigma\rho}$, two non-Gaussian averages over x^σ and W^σ. This makes our analysis quite complicated, compared to the one in [17].

In order to calculate (A.1) and (A.2), we should count all the possible choices of $\delta Q_{\sigma\rho}$ and $\delta Q_{\gamma\delta}$. For example, (i) $(\sigma, \rho) = (\gamma, \delta)$ and $\epsilon_{\sigma\rho} = 1$, (ii) one of $(\sigma, \rho) =$ one of (γ, δ), $\epsilon_{\sigma\rho} = 1$ and $\epsilon_{\gamma\delta} = 0$, etc. There are eleven such ways of pairing (σ, ρ) and (γ, δ), shown in figure A1. Each circle indicates one of n/m diagonal blocks, i.e. A_1 represents the case that $(\sigma, \rho) = (\gamma, \delta)$ and two replicas, σ and ρ, are in the same block, etc. A_i, B_i, and C_i

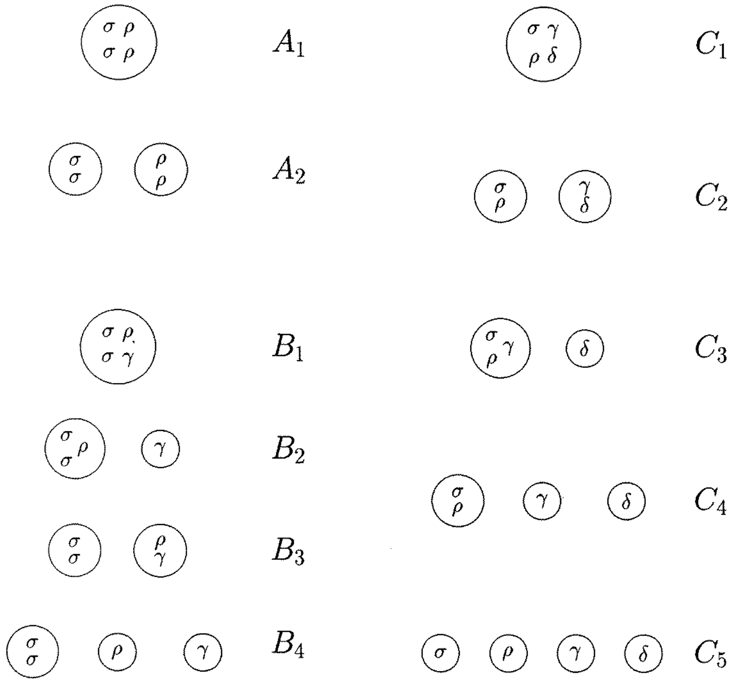

Figure A1. Eleven possible ways of pairing (σ, ρ) and (γ, δ).

can be written in integral forms and expressed as functions of q_1, q_0, m, and β, which do not depend on particular replica indices. Arranging all the coefficients properly,

$$
\begin{aligned}
\sum_{\sigma,\rho,\gamma,\delta} \Gamma_{\sigma\rho,\gamma\delta}\delta Q_{\sigma\rho}\delta Q_{\gamma\delta} = {}& 2\{(A_1 - A_2) - 2(B_1 - B_3) + (C_1 - C_2)\}\sum_{\sigma,\rho} \epsilon_{\sigma\rho}(\delta Q_{\sigma\rho})^2 \\
& +4\{(B_1 - 2B_2 - B_3 + 2B_4) - (C_1 - C_2 - 2C_3 + 2C_4)\}\sum_{\sigma}(\boldsymbol{\epsilon}\cdot\delta\boldsymbol{Q})^2_{\sigma\rho} \\
& +8(B_2 - B_4 - C_3 + C_4)\sum_{\sigma,\rho,\delta} \epsilon_{\sigma\rho}\delta Q_{\sigma\rho}\delta Q_{\sigma\gamma} \\
& +4(B_3 - B_4 - C_2 + C_4)\,\mathrm{Tr}\,\delta\boldsymbol{Q}\cdot\boldsymbol{\epsilon}\cdot\delta\boldsymbol{Q} + 4(B_4 - C_4)\sum_{\sigma,\rho}(\delta\boldsymbol{Q}\cdot\delta\boldsymbol{Q})_{\sigma\rho} \\
& +2(A_2 - 2B_3 + C_2)\sum_{\sigma,\rho}(\delta Q_{\sigma\rho})^2 \\
& +(C_1 - 3C_2 - 4C_3 + 12C_4 - 6C_5)\sum_{\sigma,\rho}\epsilon_{\sigma\rho}(\boldsymbol{\epsilon}\cdot\delta\boldsymbol{Q})_{\sigma\sigma}(\boldsymbol{\epsilon}\cdot\delta\boldsymbol{Q})_{\rho\rho} \\
& +(C_2 - 2C_4 + C_5)\Big(\sum_{\sigma,\rho}\epsilon_{\sigma\rho}\delta Q_{\sigma\rho}\Big)^2 \\
& +4(C_3 - 3C_4 + 2C_5)\sum_{\sigma,\rho\,\gamma,\delta}\epsilon_{\sigma\rho}\epsilon_{\sigma\gamma}\delta Q_{\sigma\rho}\delta Q_{\gamma\delta} \\
& +2(C_4 - C_5)\sum_{\sigma,\rho,\gamma,\delta}\epsilon_{\sigma\rho}\delta Q_{\sigma\rho}\delta Q_{\gamma\delta} + C_5\Big(\sum_{\sigma,\rho}\delta Q_{\sigma\rho}\Big)^2
\end{aligned}
$$

$$+4(C_4 - C_5)\sum_{\sigma,\rho}(\delta\boldsymbol{Q}\cdot\boldsymbol{\epsilon}\cdot\delta\boldsymbol{Q})_{\sigma\rho} + 2(C_2 - 2C_4 + C_5)\,\mathrm{Tr}(\boldsymbol{\epsilon}\cdot\delta\boldsymbol{Q})^2\,. \tag{A.6}$$

In this equation the bold italic letters stand for matrices. Replacing all A_i, B_i, and C_i by $\hat{A}_i$, $\hat{B}_i$ and $\hat{C}_i$ we can get a similar expression for $\sum_{\sigma,\rho,\gamma,\delta}\hat{\Gamma}_{\sigma\rho,\gamma\delta}\delta\hat{Q}_{\sigma\rho}\delta\hat{Q}_{\gamma\delta}$.

The eigenmode responsible for the stability of the solution comes from block fluctuations [17]. Then the contribution Λ_0 to $\Gamma_{\sigma\rho,\gamma\delta}$ in equation (A.6) from this eigenmode is given by

$$\Lambda_0 = (A_2 - 2B_3 + C_2) + 2m(B_3 - B_4 - C_2 + C_4) + m^2(C_2 - 2C_4 + C_5)\,. \tag{A.7}$$

After some manipulation, we have

$$A_2 - 2B_3 + C_2 = \frac{\beta^2}{[1+\beta(1-q_1)]^2} \tag{A.8a}$$

$$B_3 - B_4 + C_4 - C_2 = \frac{-\beta^3(q_1-q_0)}{[1+\beta(1-q_1)]^2[1+\beta(1-q_1)+m\beta(q_1-q_0)]} \tag{A.8b}$$

$$C_2 - 2C_4 + C_5 = \frac{\beta^4(q_1-q_0)^2}{[1+\beta(1-q_1)]^2[1+\beta(1-q_1)+m\beta(q_1-q_0)]^2}\,. \tag{A.8c}$$

So Λ_0 is simplified as

$$\Lambda_0 = \frac{\beta^2}{[1+\beta(1-q_1)+m\beta(q_1-q_0)]^2}\,. \tag{A.9}$$

The corresponding contribution $\hat{\Lambda}_0$ to $\hat{\Gamma}_{\sigma\rho,\gamma\delta}$ can also be found in a similar way,

$$\hat{\Lambda}_0 = (\hat{A}_2 - 2\hat{B}_3 + \hat{C}_2) + 2m(\hat{B}_3 - \hat{B}_4 - \hat{C}_2 + \hat{C}_4) + m^2(\hat{C}_2 - 2\hat{C}_4 + \hat{C}_5)\,. \tag{A.10}$$

The coefficients are given by

$$\hat{A}_2 = 1 \tag{A.11a}$$

$$\hat{B}_3 = \int \mathrm{D}z\,\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh^2 Z}{\int \mathrm{D}z_1 \cosh^m Z} \tag{A.11b}$$

$$\hat{B}_4 = \int \mathrm{D}z\left(\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh Z}{\int \mathrm{D}z_1 \cosh^m Z}\right)^2 \tag{A.11c}$$

$$\hat{C}_2 = \int \mathrm{D}z\left(\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh^2 Z}{\int \mathrm{D}z_1 \cosh^m Z}\right)^2 \tag{A.11d}$$

$$\hat{C}_4 = \int \mathrm{D}z\,\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh^2 Z}{\int \mathrm{D}z_1 \cosh^m Z}\left(\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh Z}{\int \mathrm{D}z_1 \cosh^m Z}\right)^2 \tag{A.11e}$$

$$\hat{C}_5 = \int \mathrm{D}z\left(\frac{\int \mathrm{D}z_1 \cosh^m Z \tanh Z}{\int \mathrm{D}z_1 \cosh^m Z}\right)^4 \tag{A.11f}$$

where $Z = \sqrt{\hat{q}_1 - \hat{q}_0}z_1 + \sqrt{\hat{q}_0}z$ is used. Finally, we get the stability condition in equation (3.4) from a reduced matrix M given in terms of Λ_0 and $\hat{\Lambda}_0$, shown in (3.2).

Appendix B

We checked our results also at the zero-temperature limit. In this limit, we can rewrite the free energy taking $\beta \to \infty$ with $\beta(1-q_1)$ and $m\beta$ kept constant:

$$f = \frac{1}{2}x_1\hat{x}_1 + \frac{1}{2}(x_1 + m\beta x_0)\hat{x}_0 + \frac{1}{2}\frac{R^2}{x_1 + m\beta x_0} + \frac{\alpha}{2}\frac{2-2R-x_0}{1+x_1+m\beta x_0}$$
$$-\frac{1}{m\beta}\int \mathrm{d}z_1\sqrt{\frac{\hat{x}_1}{2\pi}}\mathrm{e}^{-\frac{1}{2}\hat{x}_1 z_1^2}\ln A_+ - \frac{\alpha}{2m\beta}\ln\left\{\frac{1+x_1}{1+x_1+m\beta x_0}\right\} \tag{B.1}$$

where

$$A_\pm \equiv \mathrm{e}^{-m\beta z_1\sqrt{\hat{x}_1\hat{x}_0}}\int_{z_1\sqrt{\hat{x}_0}-m\beta\sqrt{\hat{x}_1}}^{\infty}\mathrm{D}z \pm \mathrm{e}^{m\beta z_1\sqrt{\hat{x}_1\hat{x}_0}}\int_{-z_1\sqrt{\hat{x}_0}-m\beta\sqrt{\hat{x}_1}}^{\infty}\mathrm{D}z$$

and order parameters are redefined as

$$x_1 \equiv \beta(1-q_1) \tag{B.2a}$$
$$x_0 \equiv 1 - q_0 \tag{B.2b}$$
$$\hat{x}_1 \equiv m^4\beta^2(\hat{q}_1 - \hat{q}_0) \tag{B.2c}$$
$$\hat{x}_0 \equiv m^4\beta^2\hat{q}_0\,. \tag{B.2d}$$

The saddle-point equations are

$$x_1 = \sqrt{\frac{2}{\pi}}\mathrm{e}^{-\frac{1}{2}(m\beta)^2\hat{x}_1}\int\frac{\mathrm{d}z_1}{\sqrt{2\pi}}\frac{\exp\left[-\frac{1}{2}(\hat{x}_1+\hat{x}_0)z_1^2\right]}{A_+} \tag{B.3a}$$
$$x_0 = 1 - \sqrt{\hat{x}_1}\int\frac{\mathrm{d}z_1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\hat{x}_1 z_1^2}\left(\frac{A_-}{A_+}\right)^2 \tag{B.3b}$$
$$R = \frac{\alpha(x_1+m\beta x_0)}{1+x_1+m\beta x_0} \tag{B.3c}$$
$$\hat{x}_1 = \frac{\alpha x_0}{(1+x_1)(1+x_1+m\beta x_0)} \tag{B.3d}$$
$$\hat{x}_0 = \frac{\alpha(\alpha+2-2R-x_0)}{(1+x_1+m\beta x_0)^2}\,. \tag{B.3e}$$

References

[1] Rumelhart D E and McClelland J L 1986 *Parallel Distributed Processing* (Cambridge, MA: MIT Press)

[2] Denker J, Schwartz D, Wittner B, Solla S, Howard R, Jackel L and Hopfield J 1987 *Complex Syst.* **1** 877

[3] Domany E, Meir R and Kinzel W 1986 *Europhys. Lett.* **2** 275
Meir R and Domany E 1987 *Phys. Rev. Lett.* **59** 359; 1988 *Phys. Rev.* A **37** 608

[4] Tishby N, Levin E and Solla S 1989 *Proc. Int. Joint Conf. on Neural Networks (Washington, DC)* vol 2 (New York: IEEE) p 2043

[5] Gardner E 1987 *Europhys. Lett.* **4** 1205; 1988 *J. Phys. A: Math. Gen.* **21** 257

[6] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983

[7] Gross D J and Mézard M 1984 *Nucl. Phys.* B **240** 431

[8] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683

[9] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056

[10] Kwon C, Park P and Oh J-H 1993 *Phys. Rev.* E **47** 3707

[11] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146

[12] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 *Phys. Rev.* A **45** 7590

[13] Schwarze H and Hertz 1993 *J. Europhys. Lett.* **21** 785

[14] Kang K, Oh J-H, Kwon C and Park Y 1993 *Phys. Rev.* E **48** 4805

[15] Kang K, Oh J-H, Kwon C and Park Y 1994 *Phys. Rev. Lett.* submitted

[16] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[17] Crisanti A and Sommers H-J 1992 *Z. Phys.* B **87** 341