# Stability of the replica-symmetric solution for a perceptron learning from examples

C. Kwon and Youngah Park
*Department of Physics, Myong Ji University, Yongin, Kyonggi-do, Seoul 449-728, Korea*

J.-H. Oh
*Department of Physics, Pohang Institute of Science and Technology, P.O. Box 125, Pohang, Kyongbuk 790-600, Korea*

We study the stability of the replica-symmetric solution for a perceptron learning from examples. By examining the replicon mode, we find the Almeida-Thouless (AT) line signaling a spin-glass phase transition. We find an interesting phase diagram where the AT line crosses the line of zero entropy. Entropy is still negative in a low-temperature region above the AT line. The possibility of a discontinuous phase transition to the replica-symmetry-breaking phase is also discussed.

PACS number(s): 87.10.+e, 02.50.−r, 05.20.−y

The feedforward layered network [1–4] is considered as an appropriate neural architecture where various learning mechanisms can be studied. Gardner [5] showed that a statistical-mechanical approach can be useful for studying the properties of feedforward networks. Learning [6–8] in a network with a single layer called a perceptron [9] is the simplest case and is expected to provide useful knowledge about learning in multilayered networks.

Sompolinsky, Tishby, and Seung [8, 10] studied learning from examples in a perceptron and showed that spin-glass phases exist in some cases. As an approximate estimate for the phase boundary of a spin-glass phase they suggested the line of zero entropy below which the replica-symmetric (RS) solution has negative entropy. In this paper we investigate the exact phase boundary of a spin-glass phase for this perceptron learning in a particular case. We expect an instability line corresponding to the Almeida-Thouless (AT) line [11].

Following Refs. [8] and [10], we consider a network with $N_i$ input nodes $S_i (i = 1, \ldots, N)$, $N$ synaptic weights $W_i (i = 1, \ldots, N)$, and a single output node $\sigma = \sigma(\mathbf{W}; \mathbf{S})$. A teacher network provides a set of examples consisting of $P$ input-output pairs $(\mathbf{S}^l, \sigma_0(\mathbf{S}^l))$, with $l = 1, \ldots, P$. An output is generated with a transfer function $g$ as $\sigma = g(N^{-1/2}\mathbf{S}^l \cdot \mathbf{W})$. A student network is trained by tuning the synaptic weights to minimize the error function from the correct answer of the teacher network. The error function $E$ is defined as

$$E(\mathbf{W}; \{\mathbf{S}^l\}) = \sum_{l=1}^{P} \frac{1}{2} \Big[ g(N^{-1/2}\mathbf{S}^l \cdot \mathbf{W}) - g_0(N^{-1/2}\mathbf{S}^l \cdot \mathbf{W}^0) \Big]^2, \quad (1)$$

which is the sum of squared errors over all examples. The number of examples should scale as $P = \alpha N$.

Regarding this error function as the Hamiltonian of a thermodynamic system, this learning problem turns into a statistical-mechanical problem of a disordered system. Synaptic weights $W_i$ become dynamic variables, like spins in magnetic systems, and input variables $\{\mathbf{S}^l\}$ are quenched disorder parameters. Temperature $T$ can be considered as a noise parameter inherent in human brains or computer devices. The zero-temperature limit leads to the original perceptron problem of minimizing the error function. When the teacher and the student networks have the identical weight space and the transfer function, learning is said to be realizable. In this case, perfect learning without error is possible for large $\alpha$ and low $T$. When the architectures of the teacher and the student networks are different, perfect learning is not possible. This case is called unrealizable. The probability distribution of $S_i^l$ is Gaussian with variance unity.

Studies of disordered systems such as spin-glasses and the Hopfield model have been successfully performed using the replica trick [12, 13] and the relaxation dynamics [14, 15]. Sompolinsky and his collaborators studied the perceptron learning using the replica trick and obtained a RS solution [8, 10]. We used the relaxation dynamics based on the Fokker-Planck equation and reproduced the identical result, the detail of which is to be presented elsewhere [16]. In this paper we will investigate the stability of the RS solution, which was not rigorously examined by the previous authors. We find a very interesting phase transition which was not observed in spin-glass or the Hopfield model.

The free-energy functional $f$ per neuron was found using the saddle-point method in large-$N$ limit, expressed as [10]

 

$$n\beta f\left[R_\sigma,\hat{R}_\sigma,Q_{\sigma\rho},\hat{Q}_{\sigma\rho}\right]=\sum_\sigma \hat{R}_\sigma R_\sigma+\sum_{\substack{\sigma,\rho\\(\sigma<\rho)}}\hat{Q}_{\sigma\rho}Q_{\sigma\rho}-\ln \mathrm{Tr}_{\{W^\sigma\}}\exp\left[\sum_\sigma \hat{R}_\sigma W^\sigma W^0+\sum_{\substack{\sigma,\rho\\(\sigma<\rho)}}\hat{Q}_{\sigma\rho}W^\sigma W^\rho\right]$$
$$-\alpha\ln\int\prod_\sigma\frac{dx_\sigma d\hat{x}_\sigma}{2\pi}\int\frac{dy\,d\hat{y}}{2\pi}\exp\Big[-\frac{1}{2}\beta\sum_\sigma\left[g(x_\sigma)-g_0(y)\right]^2+\sum_\sigma i\hat{x}_\sigma x_\sigma-i\hat{y}y$$
$$-\sum_{\substack{\sigma,\rho\\(\sigma<\rho)}}Q_{\sigma\rho}\hat{x}_\sigma\hat{x}_\rho-\frac{1}{2}\sum_\sigma\hat{x}_\sigma^2-\hat{y}\sum_\sigma R_\sigma\hat{x}_\sigma-\frac{1}{2}\hat{y}^2\Big]\,. \tag{2}$$

Here, $\sigma,\rho$ are replica indices and $n$ is the number of replicas of the student network. The $n\to 0$ limit is taken afterwards. The trace is carried out over the weights $W^\sigma$ of the student network. The weight $W^0$ of the teacher network is quenched. We consider two different cases, one with a discrete weight space $W=\pm1$ and one with a continuous weight space whose distribution is Gaussian. The weight space of the teacher network may or may not be the same as that of the student network. We consider the case where the transfer functions of the teacher and the student networks are identical and linear, i.e., $g(x)=g_0(x)=x$. We will report briefly on the network with the Boolean transfer function, $g(x)=\mathrm{sgn}(x)$ at the end of the paper. The weight space of the student is chosen to be discrete. If the weight space of the teacher is also discrete, learning is realizable. On the other hand, if it is Gaussian, learning is unrealizable. In this case there is a possibility of the appearance of a spin-glass phase due to the mismatch of weight space. In the following, we focus on this unrealizable learning with the weight mismatch.

A saddle-point solution for $R_\sigma$, $\hat{R}_\sigma$, $Q_{\sigma\rho}$, and $\hat{Q}_{\sigma\rho}$ can be obtained from the stationary condition of the free-energy functional with respect to variations of those variables. Then the free energy per neuron can be obtained by substituting the saddle-point solution into the free-energy functional. The replica-symmetric assumption is that a saddle-point solution is independent of replica indices; $R_\sigma=R$, $\hat{R}_\sigma=\hat{R}$ for all $\sigma$ and $Q_{\sigma\rho}=q$, $\hat{Q}_{\sigma\rho}=\hat{q}$ for all pairs of $\sigma$ and $\rho$. This RS solution is expected to be correct at high $T$ and for small $\alpha$. It was found that the RS solution has negative entropy at low $T$ [10]. The line of zero entropy is only a lower bound of the instability line of the RS solution.

The stability of the RS solution can be examined by expanding the free-energy functional given in Eq. (2) 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 RS solution. Expanding $f$ up to second order in the variations produces an $n(n+1)\times n(n+1)$ matrix $M$. The stability matrix $M$ is written as

$$M=\begin{bmatrix} A_{\sigma\rho} & \delta_{\sigma\rho} & C_{\sigma(\gamma\delta)} & 0\\ \delta_{\sigma\rho} & \hat{A}_{\sigma\rho} & 0 & \hat{C}_{\sigma(\gamma\delta)}\\ C_{(\alpha\beta)\rho} & 0 & P_{(\alpha\beta)(\gamma\delta)} & \delta_{(\alpha\beta)(\gamma\delta)}\\ 0 & \hat{C}_{(\alpha\beta)\rho} & \delta_{(\alpha\beta)(\gamma\delta)} & \hat{P}_{(\alpha\beta)(\gamma\delta)}\end{bmatrix}, \tag{3}$$

where $\delta_{\sigma\rho},\delta_{(\alpha\beta)(\gamma\delta)}$ are the Kronecker delta functions. The matrix elements are given from second derivatives of the free-energy functional,

$$A_{\sigma\rho}=\frac{\partial^2 n\beta f}{\partial R_\sigma\partial R_\rho}\,,\quad \hat{A}_{\sigma\rho}=\frac{\partial^2 n\beta f}{\partial\hat{R}_\sigma\partial\hat{R}_\rho}\,,$$
$$C_{\sigma(\gamma\delta)}=\frac{\partial^2 n\beta f}{\partial R_\sigma\partial Q_{\gamma\delta}}\,,\quad \hat{C}_{\sigma(\gamma\delta)}=\frac{\partial^2 n\beta f}{\partial\hat{R}_\sigma\partial\hat{Q}_{\gamma\delta}}\,, \tag{4}$$
$$P_{(\alpha\beta)(\gamma\delta)}=\frac{\partial^2 n\beta f}{\partial Q_{\alpha\beta}\partial Q_{\gamma\delta}}\,,\quad \hat{P}_{(\alpha\beta)(\gamma\delta)}=\frac{\partial^2 n\beta f}{\partial\hat{Q}_{\alpha\beta}\partial\hat{Q}_{\gamma\delta}}\,.$$

Here, derivatives are evaluated for the RS solution. There are only a few distinct terms given in the following:

$$A_{\sigma\sigma}=A_1,\ A_{\sigma\rho}=A_2,\ \hat{A}_{\sigma\sigma}=\hat{A}_1,\ \hat{A}_{\sigma\rho}=\hat{A}_2,$$
$$C_{\sigma(\sigma\beta)}=C_1,\ C_{\sigma(\alpha\beta)}=C_2,\ \hat{C}_{\sigma(\sigma\beta)}=\hat{C}_1,\ \hat{C}_{\sigma(\alpha\beta)}=\hat{C}_2, \tag{5}$$
$$P_{(\alpha\beta)(\alpha\beta)}=P_1,\ P_{(\alpha\beta)(\alpha\delta)}=P_2,\ P_{(\alpha\beta)(\gamma\delta)}=P_3,$$
$$\hat{P}_{(\alpha\beta)(\alpha\beta)}=\hat{P}_1,\ \hat{P}_{(\alpha\beta)(\alpha\delta)}=\hat{P}_2,\ \hat{P}_{(\alpha\beta)(\gamma\delta)}=\hat{P}_3.$$

The detailed expressions for the above terms are given in the Appendix.

Most of the eigenvalues of the stability matrix are degenerate. There are three cases for the expression of eigenvectors $\mathbf{u}$. The transposed one $\mathbf{u}^T$ is written by

$$\mathbf{u}^T=\left(\{\delta R_\sigma\},\ \{\delta\hat{R}_\sigma\},\ \{\delta Q_{\sigma\rho}\},\ \{\delta\hat{Q}_{\sigma\rho}\}\right)\,. \tag{6}$$

(i) $\delta R_\sigma=a$, $\delta\hat{R}_\sigma=b$ for all $\sigma$. $\delta Q_{\sigma\rho}=c$, $\delta\hat{Q}_{\sigma\rho}=d$ for all $(\sigma,\rho)$. There are four distinct eigenvalues.

(ii) Let $\nu$ be a given value of replica indices. $\delta R_\sigma=a$, $\delta\hat{R}_\sigma=b$ for $\sigma=\nu$; $\delta R_\sigma=a'$, $\delta\hat{R}_\sigma=b'$ for $\sigma\neq\nu$. $\delta Q_{\sigma\rho}=c$, $\delta\hat{Q}_{\sigma\rho}=d$ for $\sigma=\nu$ or $\rho=\nu$; $\delta Q_{\sigma\rho}=c'$, $\delta\hat{Q}_{\sigma\rho}=d'$ for $\sigma,\ \rho\neq\nu$. Orthogonality to the eigenvectors in case (i) gives

$$a=(1-n)a',\quad b=(1-n)b',$$
$$c=\frac{2-n}{2}c',\quad d=\frac{2-n}{2}d'. \tag{7}$$

There are four distinct eigenvalues, each with a degeneracy of $n-1$.

(iii) Let $\nu,\mu$ be a given pair of replica indices. $\delta R_\sigma=a$, $\delta\hat{R}_\sigma=b$ for $\sigma=\nu$ or $\mu$; $\delta R_\sigma=a'$, $\delta\hat{R}_\sigma=b'$ for $\sigma\neq\nu,\mu$. $\delta Q_{\sigma\rho}=c$, $\delta\hat{Q}_{\sigma\rho}=d$ for $(\sigma,\rho)=(\nu,\mu)$; $\delta Q_{\sigma\rho}=c'$, $\delta\hat{Q}_{\sigma\rho}=d'$ for $\sigma=\nu$ or $\mu$; $\rho\neq\nu,\mu$, $\delta Q_{\sigma\rho}=c''$, $\delta\hat{Q}_{\sigma\rho}=d''$ for $\sigma,\rho\neq\nu,\ \mu$. Orthogonality to the eigenvectors in cases (i) and (ii) gives

$$a = a' = b = b' = 0,$$
$$c = \frac{(n-1)(n-3)}{2}c'', \quad c' = \frac{3-n}{2}c'', \tag{8}$$
$$d = \frac{(n-1)(n-3)}{2}d'', \quad d' = \frac{3-n}{2}d''.$$

There are two distinct eigenvalues, each with a degeneracy of $n(n-3)/2$. This case corresponds to the replicon mode and gives the AT line across which the sign of an eigenvalue changes.

In limit $n \to 0$ the cases (i) and (ii) commonly yield four eigenvalues. They are eigenvalues of a reduced matrix $\tilde{M}$,

$$\tilde{M} = \begin{bmatrix} A & 1 & -C & 0 \\ 1 & \hat{A} & 0 & -\hat{C} \\ 2C & 0 & P & 1 \\ 0 & 2\hat{C} & 1 & \hat{P} \end{bmatrix}, \tag{9}$$

where

$$A = A_1 - A_2, \quad \hat{A} = \hat{A}_1 - \hat{A}_2,$$
$$C = C_1 - C_2, \quad \hat{C} = \hat{C}_1 - \hat{C}_2, \tag{10}$$
$$P = P_1 - 4P_2 + 3P_3, \quad \hat{P} = \hat{P}_1 - 4\hat{P}_2 + 3\hat{P}_3.$$

Two eigenvalues in the case (iii) are found in limit $n \to 0$ as

$$\lambda_\pm = \frac{1}{2}\Big( P_1 + \hat{P}_1 - 2(P_2 + \hat{P}_2) + (P_3 + \hat{P}_3) \pm \{[P_1 - \hat{P}_1 - 2(P_2 - \hat{P}_2) + (P_3 - \hat{P}_3)]^2 + 4\}^{\frac{1}{2}} \Big). \tag{11}$$

$\lambda_-$ is always negative, while the sign of $\lambda_+$ may change when

$$1 - (P_1 - 2P_2 + P_3)(\hat{P}_1 - 2\hat{P}_2 + \hat{P}_3) = 0. \tag{12}$$

This condition gives the AT line for the perceptron learning.

We find that the AT line exists for the unrealizable learning with the weight mismatch. Equation (12) leads to

$$[1 + \beta(1-q)]^2 - \alpha\beta^2 \overline{\mathrm{sech}^4(\sqrt{\hat{q}}z)} = 0. \tag{13}$$

The change of variable $\hat{q} + \hat{R}^2 \to \hat{q}$ is used as in the paper by Seung, Sompolinsky, and Tishby [10] $z$ is a Gaussian variable with variance unity and the overbar denotes the average over $z$. We obtained the identical result for the AT line via the relaxation dynamics [16]. A spin-glass phase appears below the AT line. It is expected because there is no unique configuration $\{W_i\}$ of the student network due to the weight-mismatch. There are many possible configurations minimizing the error function. If the AT line is the only instability line for the RS solution, the line of zero entropy should be below the AT line. However, in a certain region with low $T$ and small $\alpha$, the AT line is below the line of zero entropy and terminates at nonzero $\alpha$, as shown in Fig. 1.

We expect a different kind of phase transition should occur above the line of zero entropy which is not covered by the AT line. We should examine the other eigenvalues.

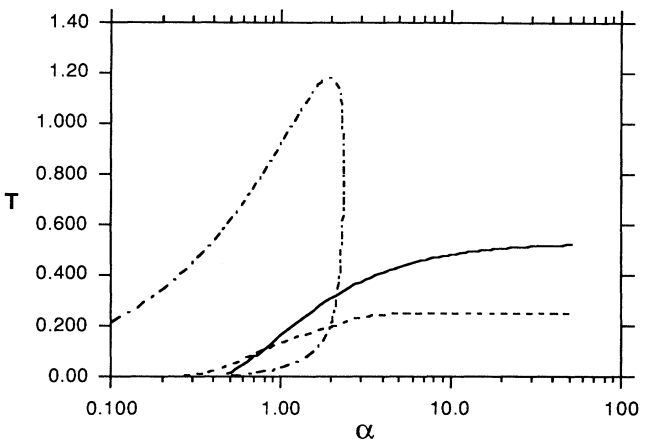


FIG. 1. The phase diagram in $\alpha$-$T$ plane. The solid line is the AT line and the dashed line is the line of zero entropy. The dot-dashed line is the new stability line.

The eigenvalues obtained from the reduced matrix $\tilde{M}$ in Eq. (9) can be complex. There are two pairs of complex-conjugate eigenvalues. The sign of all the eigenvalues does not change in the whole region. Instead, the sign of the real part of one pair of the two changes at a line, plotted in Fig 1. It is not clear whether this line signals a new phase transition. However, it seems that there should be a solution other than the RS solution above the AT line.

We are examining the possibility of the one-step replica-symmetry-breaking (RSB) phase in this problem. In the studies of spin glasses with $p$-spin interaction [17, 18], it was found that the one-step RSB solution is exact. Also, Crisanti and Sommers [19] found that the transition from the RS phase to the one-step RSB phase can be discontinuous in a related problem. In this case, the RS solution is still stable in the sense that the eigenvalues of the stability matrix do not show the change of sign. Krauth and Mézard [20] studied the one-step RSB solution of the learning of random input and ouput mapping originally addressed by Gardner [5]. They also found that the one-step RSB solution is exact. It is interesting to see whether the one-step RSB solution exists and whether the transition is discontinuous in our problem. If the one-step RSB solution exists, it is worthwhile to check whether it is exact out of many multistep RSB solutions as it was in the above problems.

A similar problem occurs even for the realizable learning with discrete weights and a linear transfer function. It was found that a metastable solution presents negative entropy at low temperatures [8, 10]. However, we find that none of the eigenvalues of the stability matrix shows the change of sign. The study of the one-step RSB solution may also resolve this problem.

We briefly report on another unrealizable learning with the weight mismatch where the transfer function is Boolean, $g(x) = \mathrm{sgn}(x)$. In this case the AT line is given by

$$1 - K(\alpha, \beta)\overline{\mathrm{sech}^4\sqrt{\hat{q}}z} = 0 \tag{14}$$

with

$$K(\alpha,\beta)=\frac{\alpha}{\pi(1-q)^2}\int_0^\infty dy\, e^{-(1/2)y^2}\int_{-\infty}^\infty dt\, e^{-(1/2)t^2}\left[\frac{ue^{-(1/2)u^2}}{H(u)+(e^\beta-1)^{-1}}-\frac{\frac{1}{\sqrt{2\pi}}e^{-u^2}}{[H(u)+(e^\beta-1)^{-1}]^2}\right]^2, \tag{15}$$

where

$$u\equiv\frac{t\sqrt{q-R^2}-yR}{\sqrt{1-q}} \tag{16}$$

and

$$H(u)\equiv\int_u^\infty dx\, e^{-(1/2)x^2}\ . \tag{17}$$

It is an interesting problem whether the AT line and the line of zero entropy may intersect.

In summary we studied the stability of the RS solution in the perceptron learning. For the unrealizable case with weight mismatch we obtained the AT line, below which the spin-glass phase appears. This is a plausible result because there are many possible configurations of the student network minimizing the error with respect to the teacher. However, the AT line covers only partially the region of negative entropy. We also discussed the stablity of other eigenvalues other than replicon mode. We suggested a possibility of discontinuous transition to the RSB solution.

We appreciate Dr. H. Seung for sending his results. C. K. would like to thank Dr. David Thouless for discussions and warm hospitality during a visit to the University of Washington and would also like to thank Dr. Hans-Jürgen Sommers for his comments. We would like to thank Dr. Doochul Kim and Dr. Mooyoung Choi for useful discussions. This work was supported in part by the KOSEF through the Center for the Theoretical Physics of the Seoul National University and also in part by the Ministry of Education through the Basic Science Research Institute of POSTECH. J. H. O. appreciates financial support from RIST.

## APPENDIX: MATRIX ELEMENTS OF THE STABILITY MATRIX

The matrix elements in Eq. (5) are given in this section. First, those associated with derivatives with respect to $R_\sigma$ and $Q_{\sigma\rho}$ are given:

$$A_1=-\alpha\left(\langle\hat{x}_\sigma^2\hat{y}^2\rangle-\langle\hat{x}_\sigma\hat{y}\rangle^2\right), \tag{A1}$$

$$A_2=-\alpha\left(\langle\hat{x}_\sigma\hat{x}_\rho\hat{y}^2\rangle-\langle\hat{x}_\sigma\hat{y}\rangle^2\right), \tag{A2}$$

$$C_1=-\alpha\left(\langle\hat{x}_\sigma^2\hat{x}_\beta\hat{y}\rangle-\langle\hat{x}_\sigma\hat{y}\rangle\langle\hat{x}_\alpha\hat{x}_\beta\rangle\right), \tag{A3}$$

$$C_2=-\alpha\left(\langle\hat{x}_\sigma\hat{x}_\alpha\hat{x}_\beta\hat{y}\rangle-\langle\hat{x}_\sigma\hat{y}\rangle\langle\hat{x}_\alpha\hat{x}_\beta\rangle\right), \tag{A4}$$

$$P_1=-\alpha\left(\langle\hat{x}_\alpha^2\hat{x}_\beta^2\rangle-\langle\hat{x}_\alpha\hat{x}_\beta\rangle^2\right), \tag{A5}$$

$$P_2=-\alpha\left(\langle\hat{x}_\alpha^2\hat{x}_\beta\hat{x}_\delta\rangle-\langle\hat{x}_\alpha\hat{x}_\beta\rangle^2\right), \tag{A6}$$

$$P_3=-\alpha\left(\langle\hat{x}_\alpha\hat{x}_\beta\hat{x}_\gamma\hat{x}_\delta\rangle-\langle\hat{x}_\alpha\hat{x}_\beta\rangle^2\right). \tag{A7}$$

In these equations,

$$\langle\cdots\hat{x}_\alpha\cdots\hat{y}\cdots\rangle=\frac{\int\prod_\sigma\frac{dx_\sigma d\hat{x}_\sigma}{2\pi}\int\frac{dy\,d\hat{y}}{2\pi}(\cdots\hat{x}_\alpha\cdots\hat{y}\cdots)e^L}{\int\prod_\sigma\frac{dx_\sigma d\hat{x}_\sigma}{2\pi}\int\frac{dy\,d\hat{y}}{2\pi}e^L}, \tag{A8}$$

where

$$L=-\frac{1}{2}\beta\sum_\sigma[g(x_\sigma)-g(y)]^2+\sum_\sigma i\hat{x}_\sigma x_\sigma-i\hat{y}y$$
$$-q\sum_{\substack{\sigma,\rho\\(\sigma<\rho)}}\hat{x}_\sigma\hat{x}_\rho-\frac{1}{2}\sum_\sigma\hat{x}_\sigma^2-R\hat{y}\sum_\sigma\hat{x}_\sigma-\frac{1}{2}\hat{y}^2. \tag{A9}$$

In $n\to 0$ limit and for the linear transfer function $g(x)=x$, some of the matrix elements of the reduced matrix $\tilde{M}$ are given as

$$A=0, \tag{A10}$$

$$C=\frac{\alpha\beta^2}{[1+\beta(1-q)]^2}, \tag{A11}$$

$$P=-\frac{\alpha\beta^2}{[1+\beta(1-q)]^2}\left(1+\frac{2\beta}{1+\beta(1-q)}(1-2R+q)\right). \tag{A12}$$

The matrix elements associated with derivatives with respect to $\hat{R}_\sigma$ and $\hat{Q}_{\sigma\beta}$ are given as

$$\hat{A}_1=-\left(1-\langle W^\sigma W^0\rangle^2\right), \tag{A13}$$

$$\hat{A}_2=-\left(\langle W^\sigma W^0 W^\rho W^0\rangle-\langle W^\sigma W^0\rangle^2\right), \tag{A14}$$

$$\hat{C}_1=-\left(\langle W^\beta W^0\rangle-\langle W^\sigma W^0\rangle\langle W^\alpha W^\beta\rangle\right), \tag{A15}$$

$$\hat{C}_2=-\left(\langle W^\sigma W^0 W^\alpha W^\beta\rangle-\langle W^\sigma W^0\rangle\langle W^\alpha W^\beta\rangle\right), \tag{A16}$$

$$\hat{P}_1=-\left(1-\langle W^\alpha W^\beta\rangle^2\right), \tag{A17}$$

$$\hat{P}_2=-\left(\langle W^\beta W^\delta\rangle-\langle W^\alpha W^\beta\rangle^2\right), \tag{A18}$$

$$\hat{P}_3=-\left(\langle W^\alpha W^\beta W^\gamma W^\delta\rangle-\langle W^\alpha W^\beta\rangle^2\right). \tag{A19}$$

In these equations,

$$\langle \cdots W^{\sigma} \cdots \rangle = \frac{\mathrm{Tr}\left[ (\cdots W^{\sigma} \cdots) \exp\left( q \sum_{\substack{\sigma,\rho \\ (\sigma<\rho)}} W^{\sigma} W^{\rho} + \hat{R} \sum_{\sigma} W^{\sigma} W^{0} \right) \right]}{\mathrm{Tr}\left[ \exp\left( q \sum_{\substack{\sigma,\rho \\ (\sigma<\rho)}} W^{\sigma} W^{\rho} + \hat{R} \sum_{\sigma} W^{\sigma} W^{0} \right) \right]} . \tag{A20}$$

In $n \to 0$ limit the rest of the matrix elements of $\tilde{M}$ are given as

$$\hat{A} = -\left[ 1 - q - \frac{2\alpha\beta R}{1+\beta(1-q)} + \frac{2}{\sqrt{\hat{q}}} \left( \frac{\alpha\beta}{1+\beta(1-q)} \right)^{2} \overline{z \tanh^{3}(\sqrt{\hat{q}} z)} \right] , \tag{A21}$$

$$\hat{C} = -\left( R - \frac{1}{\sqrt{\hat{q}}} \frac{\alpha\beta}{1+\beta(1-q)} \overline{z \tanh^{3}(\sqrt{\hat{q}} z)} \right) , \tag{A22}$$

$$\hat{P} = -\overline{\left[ 1 - 4\tanh^{2}(\sqrt{\hat{q}} z) + 3\tanh^{4}(\sqrt{\hat{q}} z) \right]} . \tag{A23}$$

The overbar denotes the average over the Gaussian variable $z$ with variance unity. Note that the Gaussian average over the weight $W_0$ of the teacher is carried out after the matrix elements are found for a fixed $W_0$ because it is also a quenched parameter.


[1] D.E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing* (MIT Press, Cambridge, MA, 1986).
[2] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield, Complex Syst. **1**, 877 (1987).
[3] E. Domany, R. Meir, and W. Kinzel, Europhys. Lett. **2**, 275 (1986); R. Meir and E. Domany, Phys. Rev. Lett. **59**, 359 (1987); Phys. Rev. A **37**, 608 (1988).
[4] N. Tishby, E. Levin, and S. Solla, in *Proceedings of the International Joint Conference on Neural Networks, Washington, D.C.* (IEEE, New York, 1989), Vol. 2, p. 2043.
[5] E. Gardner, Europhys. Lett. **4**, 1205 (1987); J. Phys. A **21**, 257 (1988).
[6] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).
[7] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).
[8] H. Sompolinsky, N. Tishby, and H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).
[9] M. L. Minsky and S. Papert, *Perceptron* (MIT Press, Cambridge, MA, 1969).
[10] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[11] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).
[12] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1972 (1975).
[13] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (NY) **173**, 30 (1987).
[14] H. Sompolinsky and A. Zippelius, Phys. Rev. B **25**, 6860 (1988).
[15] H. Rieger, M. Schreckenberg, and J. Zittartz, Z. Phys. B **72**, 523 (1988).
[16] C. Kwon, Youngah Park, and J.-H. Oh (unpublished).
[17] D. J. Gross and M. Mézard, Nucl. Phys. B **240**, 431 (1984).
[18] E. Gardner, Nucl. Phys. B **257**, 747 (1985).
[19] A Crisanti and H-J Sommers (unpublished).
[20] W. Krauth and M. Mézard, J. Phys. (Paris) **50**, 3057 (1989).