Perceptron capacity revisited: classification ability for correlated patterns

2008 J. Phys. A: Math. Theor. 41 324013

(http://iopscience.iop.org/1751-8121/41/32/324013)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **41** (2008) 324013 (18pp) doi:10.1088/1751-8113/41/32/324013

Perceptron capacity revisited: classification ability for correlated patterns

Takashi Shinzato and Yoshiyuki Kabashima

Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan

E-mail: shinzato@sp.dis.titech.ac.jp and kaba@dis.titech.ac.jp

Received 25 December 2007, in final form 27 February 2008
Published 30 July 2008
Online at stacks.iop.org/JPhysA/41/324013

Abstract

In this paper, we address the problem of how many randomly labeled patterns can be correctly classified by a single-layer perceptron when the patterns are correlated with each other. In order to solve this problem, two analytical schemes are developed based on the replica method and the Thouless–Anderson–Palmer (TAP) approach by utilizing an integral formula concerning random rectangular matrices. The validity and relevance of the developed methodologies are shown for one known result and two example problems. A message-passing algorithm to perform the TAP scheme is also presented.

PACS numbers: 02.50.−r, 84.35.+i

1. Introduction

Learning from examples is one of the most significant problems in information science, and (single-layer) perceptrons are often included in widely used devices for solving this problem. In the last two decades, the structural similarity between the learning problem and the statistical mechanics of disordered systems has been observed, thus promoting cross-disciplinary research on perceptron learning with the use of methods from statistical mechanics [1, 2]. This research activity has successfully contributed to the discovery of various behaviors in the learning process of perceptrons [3–5] and to the development of computationally feasible approximate learning algorithms [6, 7] that had never been discovered by conventional approaches in information science, particularly for the non-asymptotic regimes in which the ratio between the numbers of examples p and weight parameters N, $\alpha = p/N$, is $O(1)$.

Although such statistical mechanical methodologies have been successfully applied to learning problems, there still remain several research directions to explore. Learning from correlated patterns is a typical example of such a problem. In most of the earlier studies, it was assumed, for simplicity, that the input patterns used for learning were independently and identically distributed (IID) [3–5]. However, this assumption is obviously not practical since real-world data are usually somewhat biased and correlated across components, which makes

1751-8113/08/324013+18$30.00 © 2008 IOP Publishing Ltd Printed in the UK

it difficult to utilize the developed schemes directly for learning beyond a conceptual level. In order to increase the practical relevance of the statistical mechanical approach, it is necessary to generalize the approach to handle correlated patterns.

As a first step for such a research direction, we address the problem of correctly classifying many randomly labeled patterns by a single-layer perceptron when the patterns are correlated with each other. Finding a regularity in a given set of patterns is highly demanded in many real-world problems of data analysis. The addressed problem is of practical importance as an assessment of null hypotheses that state no regularity of classification represented by the perceptron underlies the labeled correlated patterns. In addition, recent deepening of the relations across learning, information and communication theories shows that the perceptron can be utilized as a useful building block for various coding schemes [8–11]. Therefore, exploration to handle learning from correlated patterns may lead to the development of better schemes used for information and communication engineering.

This paper is organized as follows. In the following section, we introduce the problem we are studying. In section 3, which is the main part of this article, we develop two schemes for analyzing the problem on the basis of the replica method and Thouless–Anderson–Palmer (TAP) approach. Statistical mechanical techniques that can handle correlated patterns have already been developed by Opper and Winther [12–14]. However, their schemes, which apply to densely connected networks of two-body interactions, are highly general, and therefore properties that hold specifically for perceptrons are not fully utilized. Hence, in this paper, we offer specific methodologies that can be utilized for perceptron-type networks. We show that an integral formula provided for ensembles of rectangular random matrices plays important roles for the provided methods. A message-passing algorithm to solve the developed TAP scheme is also presented. In section 4, the validity and utility of the methods are shown by applications to one known result and two example problems. The final section is a summary.

2. Problem definition

In a general scenario, for an N-dimensional input pattern vector $\boldsymbol{x}$, a perceptron which is parametrized by an N-dimensional weight vector $\boldsymbol{w}$ can be identified with an indicator function of class label $y = \pm 1$,

$$\mathcal{I}(y|\Delta), \tag{1}$$

where $\mathcal{I}(y|\Delta) = 1 - \mathcal{I}(-y|\Delta)$ takes 1 or 0 depending on the value of internal potential $\Delta = N^{-1/2}\boldsymbol{w}\cdot\boldsymbol{x}$. Prefactor $N^{-1/2}$ is introduced to keep relevant variables $O(1)$ as $N \to \infty$. Equation (1) indicates that a perceptron specified by $\boldsymbol{w}$ correctly classifies a given labeled pattern $(\boldsymbol{x}, y)$ if $\mathcal{I}(y|\Delta) = 1$; otherwise, it does not make the correct classification. Let us suppose that a set of patterns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ is given. The problem we consider here is whether the perceptron can typically classify the patterns correctly by only adjusting $\boldsymbol{w}$ when the class label of each pattern $\boldsymbol{x}_\mu$, $y_\mu \in \{+1, -1\}$, is independently and randomly assigned with a probability of $1/2$ for $\mu = 1, 2, \ldots, p$ as N and p tend to infinity, keeping the pattern ratio $\alpha = p/N$ of the order of unity.

In general, entries of pattern matrix $X = N^{-1/2}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)^{\mathrm{T}}$ are correlated with each other, where T denotes the matrix transpose. As a basis for dealing with such correlations, we introduce an expression of the singular value decomposition

$$X = UDV^{\mathrm{T}}, \tag{2}$$

of the pattern matrix X, where $D = \mathrm{diag}(d_k)$ is a $p \times N$ diagonal matrix composed of singular values d_k $(k = 1, 2, \ldots, \min(p, N))$, and U and V are $p \times p$ and $N \times N$ orthogonal matrices, respectively. $\min(p, N)$ denotes the lesser value of p and N. Linear algebra guarantees that an arbitrary $p \times N$ matrix can be decomposed according to equation (2). The singular values d_k are linked to eigenvalues of the correlation matrix $X^{\mathrm{T}}X$, $\lambda_k(k = 1, 2, \ldots, N)$, as $\lambda_k = d_k^2(k = 1, 2, \ldots, \min(p, N))$ and 0 otherwise. The orthogonal matrices U and V constitute the eigen bases of correlation matrices XX^{T} and $X^{\mathrm{T}}X$, respectively. In order to handle correlations in X analytically, we assume that U and V are uniformly and independently generated from the Haar measures of $p \times p$ and $N \times N$ orthogonal matrices, respectively, and that the empirical eigenvalue spectrum of $X^{\mathrm{T}}X$, $N^{-1}\sum_{k=1}^{N}\delta(\lambda-\lambda_k) = (1-\min(p, N)/N)\delta(\lambda)+N^{-1}\sum_{k=1}^{\min(p,N)}\delta\big(\lambda-d_k^2\big)$, converges to a certain specific distribution $\rho(\lambda)$ in the large system limit of N, $p \to \infty$, $\alpha = p/N \sim O(1)$. Controlling $\rho(\lambda)$ allows us to characterize various second-order correlations in pattern matrix X.

For generality and analytical tractability, let us assume that $\boldsymbol{w}$ obeys a factorizable distribution $P(\boldsymbol{w}) = \prod_{i=1}^{N} P(w_i)$ *a priori*. Given a labeled pattern set $\xi^p = (X, \boldsymbol{y})$, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^{\mathrm{T}}$, it is possible to assess the volumes of $\boldsymbol{w}$ that are compatible with ξ^p as

$$V(\xi^p) = \operatorname*{Tr}_{\boldsymbol{w}} \prod_{i=1}^{N} P(w_i) \prod_{\mu=1}^{p} \mathcal{I}(y_\mu|\Delta_\mu), \tag{3}$$

where $\Delta_\mu = N^{-1/2}\boldsymbol{w}\cdot\boldsymbol{x}_\mu(\mu = 1, 2, \ldots, p)$ and $\mathrm{Tr}_{\boldsymbol{w}}$ denotes the summation (or integral) over all possible states of $\boldsymbol{w}$. Equation (3), which is sometimes referred to as the *Gardner volume*, is used for assessing whether ξ^p can be classified by a given type of perceptron because it is possible to choose an appropriate $\boldsymbol{w}$ that is fully consistent with ξ^p if and only if $V(\xi^p)$ does not vanish [15].

In the large system limit, $V(\xi^p)$ typically vanishes and, therefore, ξ^p cannot be correctly classified by perceptrons of the given type when α becomes larger than a certain critical value α_{c}, which is often termed *perceptron capacity* [15, 16]. Since the mid 1980s, much effort has been made in the cross-disciplinary field of statistical mechanics and information science to assess α_{c} in various systems [4]: in particular, for pattern matrices entries of which are independently drawn from an identical distribution of zero mean and variance N^{-1}. Such situations are characterized by the Marčenko–Pastur law $\rho(\lambda) = [1-\alpha]^+\delta(\lambda) + (2\pi)^{-1}\lambda^{-1}\sqrt{[\lambda-\lambda_-]^+[\lambda_+-\lambda]^+}$ in the current framework, where $[x]^+ = x$ for $x > 0$ and 0, otherwise, and $\lambda_\pm = (\sqrt{\alpha}\pm 1)^2$ [17]. However, it seems that little is known about how the correlations in pattern matrices, which are characterized by $\rho(\lambda)$ here, influence the perceptron capacity α_{c}. Therefore, the main objective of the present paper is to answer this question.

3. Analysis

3.1. A generalization of the Itzykson–Zuber integral

The expression

$$\begin{aligned} V(\xi^p) &= \operatorname*{Tr}_{\boldsymbol{w}} \prod_{i=1}^{N} P(w_i) \prod_{\mu=1}^{p} \left(\int \mathrm{d}\Delta_\mu \mathcal{I}(y_\mu|\Delta_\mu)\delta(\Delta_\mu - N^{-1/2}\boldsymbol{w}\cdot\boldsymbol{x}_\mu) \right) \\ &= \int \prod_{\mu=1}^{p} \left(\frac{\mathrm{d}u_\mu \mathrm{d}\Delta_\mu}{2\pi} \exp[-\mathrm{i}u_\mu\Delta_\mu]\mathcal{I}(y_\mu|\Delta_\mu) \right) \operatorname*{Tr}_{\boldsymbol{w}} P(w_i)\exp[\mathrm{i}\boldsymbol{u}^{\mathrm{T}}X\boldsymbol{w}] \\ &= \operatorname*{Tr}_{\boldsymbol{u},\boldsymbol{w}} \prod_{\mu=1}^{p} \widehat{\mathcal{I}}_{y_\mu}(u_\mu) \prod_{i=1}^{N} P(w_i)\exp[\mathrm{i}\boldsymbol{u}^{\mathrm{T}}X\boldsymbol{w}] \end{aligned} \tag{4}$$

constitutes the basis for analyzing the behavior of equation (3), where $\mathrm{i} = \sqrt{-1}$, $\boldsymbol{u} = (u_1, u_2, \ldots, u_p)^{\mathrm{T}}$ and $\widehat{\mathcal{I}}_{y_\mu}(u_\mu) = \int \mathrm{d}\Delta_\mu \exp[-\mathrm{i}u_\mu \Delta_\mu]\mathcal{I}(y_\mu|\Delta_\mu)/(2\pi)$. In order to evaluate the average of $V(\xi^p)$, we substitute equation (2) into (4) and take the average with respect to the orthogonal matrices U and V. For this assessment, it is worthwhile noting that for the fixed sets of dynamical variables $\boldsymbol{w}$ and $\boldsymbol{u}$, $\widetilde{\boldsymbol{w}} = V^{\mathrm{T}}\boldsymbol{w}$ and $\widetilde{\boldsymbol{u}} = U^{\mathrm{T}}\boldsymbol{u}$ behave as continuous random variables that are uniformly generated under the strict constraints

$$\frac{1}{N}|\widetilde{\boldsymbol{w}}|^2 = \frac{1}{N}|\boldsymbol{w}|^2 = Q_w, \tag{5}$$

$$\frac{1}{p}|\widetilde{\boldsymbol{u}}|^2 = \frac{1}{p}|\boldsymbol{u}|^2 = Q_u, \tag{6}$$

when U and V are independently and uniformly generated from the Haar measures. In the limit as $N, p \to \infty$, keeping $\alpha = p/N \sim O(1)$, this yields the expression

$$\begin{aligned}\frac{1}{N}\ln[\overline{\exp[\mathrm{i}\boldsymbol{u}^{\mathrm{T}}X\boldsymbol{w}]}] &= \frac{1}{N}\ln\left[\frac{\int \mathrm{d}\widetilde{\boldsymbol{w}}\,\mathrm{d}\widetilde{\boldsymbol{u}}\,\delta(|\widetilde{\boldsymbol{w}}|^2 - NQ_w)\delta(|\widetilde{\boldsymbol{u}}|^2 - pQ_u)\exp[\mathrm{i}\widetilde{\boldsymbol{u}}^{\mathrm{T}}D\widetilde{\boldsymbol{w}}]}{\int \mathrm{d}\widetilde{\boldsymbol{w}}\,\mathrm{d}\widetilde{\boldsymbol{u}}\,\delta(|\widetilde{\boldsymbol{w}}|^2 - NQ_w)\delta(|\widetilde{\boldsymbol{u}}|^2 - pQ_u)}\right]\\ &= F(Q_w, Q_u),\end{aligned} \tag{7}$$

where $\overline{\cdots}$ denotes averaging with respect to the Haar measures, the function $F(x, y)$ is assessed as

$$\begin{aligned}F(x, y) = \mathop{\mathrm{Extr}}_{\Lambda_x, \Lambda_y}&\left\{-\frac{1}{2}\langle\ln(\Lambda_x\Lambda_y + \lambda)\rangle_\rho - \frac{\alpha - 1}{2}\ln\Lambda_y + \frac{\Lambda_x x}{2} + \frac{\alpha\Lambda_y y}{2}\right\}\\ &-\frac{1}{2}\ln x - \frac{\alpha}{2}\ln y - \frac{1+\alpha}{2},\end{aligned} \tag{8}$$

and $\langle\cdots\rangle_\rho$ indicates averaging with respect to the asymptotic eigenvalue spectrum of $X^{\mathrm{T}}X$, $\rho(\lambda)$ [18]. The derivation of equations (7) and (8) is shown in appendix A. $\mathrm{Extr}_\theta\{\cdots\}$ represents extremization with respect to θ. This corresponds to the saddle-point assessment of a complex integral and does not necessarily mean the operation of a minimum or maximum. Expressions analogous to these equations are known as the Itzykson–Zuber integral or *G*-function for ensembles of square (symmetric) matrices [19–27]. Equation (7) implies that the annealed average of equation (3) is evaluated as

$$\frac{1}{N}\ln[V(\xi^p)]_{\xi^p} = \mathop{\mathrm{Extr}}_{Q_w, Q_u}\{F(Q_w, Q_u) + A_w(Q_w) + \alpha A_u(Q_u)\}, \tag{9}$$

where $[\cdots]_{\xi^p} = 2^{-p}\,\mathrm{Tr}_{\boldsymbol{y}}\,\overline{(\cdots)}$ represents the average with respect to a set of randomly labeled patterns ξ^p and

$$A_w(Q_w) = \mathop{\mathrm{Extr}}_{\widehat{Q}_w}\left\{\frac{\widehat{Q}_w Q_w}{2} + \ln\left[\mathop{\mathrm{Tr}}_{w} P(w)\exp\left[-\frac{\widehat{Q}_w}{2}w^2\right]\right]\right\}, \tag{10}$$

$$A_u(Q_u) = \mathop{\mathrm{Extr}}_{\widehat{Q}_u}\left\{\frac{\widehat{Q}_u Q_u}{2} + \ln\left[\frac{1}{2}\mathop{\mathrm{Tr}}_{u,y}\widehat{\mathcal{I}}_y(u)\exp\left[-\frac{\widehat{Q}_u}{2}u^2\right]\right]\right\}. \tag{11}$$

Normalization constraints $\mathrm{Tr}_y\,\mathcal{I}(y|\Delta) = 1$ guarantee that $[V(\xi^p)]_{\xi^p} = 2^{-p}$, which implies that for any $\boldsymbol{w}$ the probability that each randomly labeled pattern $(\boldsymbol{x}_\mu, y_\mu)$ $(\mu = 1, 2, \ldots, p)$ is correctly classified is equal to $1/2$ and, therefore, the size of feasible volume $V(\xi^p)$ decreases as 2^{-p} on average, regardless of correlations in X. In addition, in conjunction with equations (9)–(11), this implies that $Q_w = \mathrm{Tr}_w w^2 P(w)$, $Q_u = 0$, $\widehat{Q}_w = 0$ and $\widehat{Q}_u = \alpha^{-1}Q_w\langle\lambda\rangle_\rho$. The physical implication is that, due to the central limit theorem, $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_p)^{\mathrm{T}}$ follows an isotropic Gaussian distribution

$$P(\boldsymbol{\Delta}) = \frac{1}{(2\pi \widehat{Q}_u)^{p/2}} \exp\left[-\frac{|\boldsymbol{\Delta}|^2}{2\widehat{Q}_u}\right] = \frac{\alpha^{p/2}}{(2\pi Q_w \langle\lambda\rangle_\rho)^{p/2}} \exp\left[-\frac{\alpha|\boldsymbol{\Delta}|^2}{2Q_w \langle\lambda\rangle_\rho}\right], \tag{12}$$

in the limit as $N, p \to \infty, \alpha = p/N \sim O(1)$ when $\boldsymbol{w}$ is generated from $P(\boldsymbol{w}) = \prod_{i=1}^N P(w_i)$, and U and V are independently and uniformly generated from the Haar measures.

3.2. Replica analysis

Now we are ready to analyze the typical behavior of equation (3). Because ξ^p is a set of quenched random variables, we resort to the replica method [28–30]. This indicates that we evaluate the nth moment of $V(\xi^p)$ for natural numbers $n \in \mathbb{N}$ as

$$[V^n(\xi^p)]_{\xi^p} = \mathop{\mathrm{Tr}}_{\{\boldsymbol{u}^a\},\{\boldsymbol{w}^a\}} \prod_{\mu=1}^{p} \left(\frac{1}{2} \mathop{\mathrm{Tr}}_{y_\mu} \prod_{a=1}^{n} \widehat{\mathcal{I}}_{y_\mu}\left(u_\mu^a\right)\right) \times \prod_{i=1}^{N} \left(\prod_{a=1}^{n} P\left(w_i^a\right)\right) \overline{\exp\left[\mathrm{i} \sum_{a=1}^{n} (\boldsymbol{u}^a)^{\mathrm{T}} X \boldsymbol{w}^a\right]}, \tag{13}$$

and assess the quenched average of the free energy with respect to the labeled pattern set ξ^p as $N^{-1}[\ln V(\xi^p)]_{\xi^p} = \lim_{n\to 0} \frac{\partial}{\partial n} N^{-1} \ln[V^n(\xi^p)]_{\xi^p}$ by analytically continuing expressions obtained for equation (13) from $n \in \mathbb{N}$ to real numbers $n \in \mathbb{R}$. Here, $\{\boldsymbol{w}^a\}$ and $\{\boldsymbol{u}^a\}$ represent sets of dynamical variables $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^n$ and $\boldsymbol{u}^1, \ldots, \boldsymbol{u}^n$, respectively, where $1, 2, \ldots, n$ denote the n replicas of perceptrons.

For this procedure, an explanation similar to that for the evaluation of equation (7) is useful. Namely, for fixed sets of dynamical variables $\{\boldsymbol{u}^a\}$ and $\{\boldsymbol{w}^a\}$, $\widetilde{\boldsymbol{u}}^a = U^{\mathrm{T}}\boldsymbol{u}^a$ and $\widetilde{\boldsymbol{w}}^a = V^{\mathrm{T}}\boldsymbol{w}^a$ behave as continuous random variables which satisfy strict constraints

$$\frac{1}{N}\widetilde{\boldsymbol{w}}^a \cdot \widetilde{\boldsymbol{w}}^b = \frac{1}{N}\boldsymbol{w}^a \cdot \boldsymbol{w}^b = q_w^{ab}, \tag{14}$$

$$\frac{1}{p}\widetilde{\boldsymbol{u}}^a \cdot \widetilde{\boldsymbol{u}}^b = \frac{1}{p}\boldsymbol{u}^a \cdot \boldsymbol{u}^b = q_u^{ab}, \tag{15}$$

$(a, b = 1, \ldots, n)$ when U and V are independently and uniformly generated from the Haar measures. This indicates that equation (13) can be evaluated by the saddle-point method with respect to sets of macroscopic parameters $\mathcal{Q}_w = \left(q_w^{ab}\right)$ and $\mathcal{Q}_u = \left(q_u^{ab}\right)$ in the limit as $N, p \to \infty, \alpha = p/N \sim O(1)$. In addition, intrinsic permutation symmetry among replicas indicates that it is natural to assume that the $n \times n$ matrices $\mathcal{Q}_w$ and $\mathcal{Q}_u$ are of the replica symmetric (RS) form

$$\begin{aligned}
\mathcal{Q}_w &= \begin{pmatrix} \chi_w + q_w & q_w & \cdots & q_w \\ q_w & \chi_w + q_w & \cdots & q_w \\ \vdots & \vdots & \ddots & \vdots \\ q_w & q_w & \cdots & \chi_w + q_w \end{pmatrix} \\
&= E \times \left(\begin{array}{c|cccc} \chi_w + nq_w & 0 & 0 & \cdots & 0 \\ \hline 0 & \chi_w & 0 & \cdots & 0 \\ 0 & 0 & \chi_w & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \chi_w \end{array}\right) \times E^{\mathrm{T}},
\end{aligned} \tag{16}$$

and

$$
\begin{aligned}
\mathcal{Q}_u &= \begin{pmatrix} \chi_u - q_u & -q_u & \ldots & -q_u \\ -q_u & \chi_u - q_u & \ldots & -q_u \\ \vdots & \vdots & \ddots & \vdots \\ -q_u & -q_u & \ldots & \chi_u - q_u \end{pmatrix} \\
&= E \times \left(\begin{array}{c|cccc} \chi_u - nq_u & 0 & 0 & \ldots & 0 \\ \hline 0 & \chi_u & 0 & \ldots & 0 \\ 0 & 0 & \chi_u & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \chi_u \end{array}\right) \times E^{\mathrm{T}},
\end{aligned} \tag{17}
$$

at the saddle point. Here, $E = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ denotes an n-dimensional orthonormal basis composed of $\boldsymbol{e}_1 = (n^{-1/2}, n^{-1/2}, \ldots, n^{-1/2})^{\mathrm{T}}$ and $n-1$ orthonormal vectors $\boldsymbol{e}_2, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_n$, which are orthogonal to $\boldsymbol{e}_1$. Equations (16) and (17) indicate that under the RS ansatz, the n replicas that are coupled with each other in equations (14) and (15) can be decoupled by rotating $\{\widetilde{\boldsymbol{w}}^a\}$ and $\{\widetilde{\boldsymbol{u}}^a\}$ with respect to the replica coordinates simultaneously with the use of the identical orthogonal matrix E. The already decoupled expression $\sum_{a=1}^n (\boldsymbol{u}^a)^{\mathrm{T}} X \boldsymbol{w}^a = \sum_{a=1}^n (\widetilde{\boldsymbol{u}}^a)^{\mathrm{T}} D \widetilde{\boldsymbol{w}}^a$ is kept invariant under this rotation. These operations imply that, in the new coordinates, the average with respect to U and V over uniform distributions of the Haar measures can be evaluated individually for each of the n decoupled modes, which yields

$$
\begin{aligned}
&\frac{1}{N} \ln \left[\overline{\exp\left[\mathrm{i} \sum_{a=1}^n (\boldsymbol{u}^a)^{\mathrm{T}} X \boldsymbol{w}^a \right]} \right] \\
&\quad = \frac{1}{N} \ln \left[\frac{\int \prod_{a=1}^n \mathrm{d}\widetilde{\boldsymbol{w}}^a \, \mathrm{d}\widetilde{\boldsymbol{u}}^a \, \mathcal{C}_{\text{coupled}} \exp\left[\mathrm{i} \sum_{a=1}^n (\widetilde{\boldsymbol{u}}^a)^{\mathrm{T}} D \widetilde{\boldsymbol{w}}^a \right]}{\int \prod_{a=1}^n \mathrm{d}\widetilde{\boldsymbol{w}}^a \, \mathrm{d}\widetilde{\boldsymbol{u}}^a \, \mathcal{C}_{\text{coupled}}} \right] \\
&\quad = \frac{1}{N} \ln \left[\frac{\int \prod_{a=1}^n \mathrm{d}\widetilde{\boldsymbol{w}}^a \mathrm{d}\widetilde{\boldsymbol{u}}^a \mathcal{C}_{\text{decoupled}} \exp\left[\mathrm{i} \sum_{a=1}^n (\widetilde{\boldsymbol{u}}^a)^{\mathrm{T}} D \widetilde{\boldsymbol{w}}^a \right]}{\int \prod_{a=1}^n \mathrm{d}\widetilde{\boldsymbol{w}}^a \, \mathrm{d}\widetilde{\boldsymbol{u}}^a \, \mathcal{C}_{\text{decoupled}}} \right] \\
&\quad = F(\chi_w + nq_w, \chi_u - nq_u) + (n-1) F(\chi_w, \chi_u),
\end{aligned} \tag{18}
$$

where

$$
\begin{aligned}
\mathcal{C}_{\text{coupled}} = {} & \prod_{a=1}^n \delta(|\widetilde{\boldsymbol{w}}^a|^2 - N(\chi_w + q_w)) \prod_{a>b} \delta(\widetilde{\boldsymbol{w}}^a \cdot \widetilde{\boldsymbol{w}}^b - Nq_w) \\
& \times \prod_{a=1}^n \delta(|\widetilde{\boldsymbol{u}}^a|^2 - p(\chi_u - q_u)) \prod_{a>b} \delta(\widetilde{\boldsymbol{u}}^a \cdot \widetilde{\boldsymbol{u}}^b + pq_u),
\end{aligned} \tag{19}
$$

and

$$
\begin{aligned}
\mathcal{C}_{\text{decoupled}} = {} & \delta(|\widetilde{\boldsymbol{w}}^1|^2 - N(\chi_w + nq_w)) \prod_{a=2}^n \delta(|\widetilde{\boldsymbol{w}}^a|^2 - Nq_w) \\
& \times \delta(|\widetilde{\boldsymbol{u}}^1|^2 - p(\chi_u - nq_u)) \prod_{a=2}^n \delta(|\widetilde{\boldsymbol{u}}^a|^2 + pq_u).
\end{aligned} \tag{20}
$$

Equation (18) and evaluation of the volumes of the dynamical variables $\{\boldsymbol{w}^a\}$ and $\{\boldsymbol{u}^a\}$ under constraints (14) and (15) of the RS ansatz (16) and (17) provide an expression for the average free energy

$$\frac{1}{N}[\ln V(\xi^p)]_{\xi^p} = \lim_{n\to 0}\frac{\partial}{\partial n}\frac{1}{N}\ln[V^n(\xi^p)]_{\xi^p}$$
$$= \mathop{\mathrm{Extr}}_{\boldsymbol{\Theta}}\left\{\mathcal{A}_0(\chi_w,\chi_u,q_w,q_u)+\mathcal{A}_w(\chi_w,q_w)+\alpha\mathcal{A}_u(\chi_u,q_u)\right\}, \tag{21}$$

where $\boldsymbol{\Theta} = (\chi_w,\chi_u,q_w,q_u)$,

$$\mathcal{A}_0(\chi_w,\chi_u,q_w,q_u) = F(\chi_w,\chi_u)+q_w\frac{\partial F(\chi_w,\chi_u)}{\partial\chi_w}-q_u\frac{\partial F(\chi_w,\chi_u)}{\partial\chi_u}, \tag{22}$$

$$\mathcal{A}_w(\chi_w,q_w) = \mathop{\mathrm{Extr}}_{\widehat{\chi}_w,\widehat{q}_w}\left\{\frac{\widehat{\chi}_w}{2}(\chi_w+q_w)-\frac{\widehat{q}_w}{2}\chi_w + \int \mathrm{D}z\ln\left[\mathop{\mathrm{Tr}}_{w}P(w)\exp\left[-\frac{\widehat{\chi}_w}{2}w^2+\sqrt{\widehat{q}_w}zw\right]\right]\right\}, \tag{23}$$

and

$$\mathcal{A}_u(\chi_u,q_u) = \mathop{\mathrm{Extr}}_{\widehat{\chi}_u,\widehat{q}_u}\left\{\frac{\widehat{\chi}_u}{2}(\chi_u-q_u)+\frac{\widehat{q}_u}{2}\chi_u + \frac{1}{2}\mathop{\mathrm{Tr}}_{y}\int \mathrm{D}z\ln\left[\int \mathrm{D}x\mathcal{I}(y|\sqrt{\widehat{\chi}_u}x+\sqrt{\widehat{q}_u}z)\right]\right\}. \tag{24}$$

Here, $\mathrm{D}s = \mathrm{d}s\exp[-s^2/2]/\sqrt{2\pi}$ represents the Gaussian measure.

Two points should be noted here. The first is that the current formalism can be applied not only to the RS analysis presented above but also to that of replica symmetry breaking (RSB) [29, 30]. An expression of the average free energy under the one-step RSB (1RSB) ansatz is shown in appendix B. In addition, analysis of the local instability condition of the RS solutions (16) and (17) subject to infinitesimal perturbation of the form of 1RSB yields

$$\left(1-2\frac{\partial^2 F}{\partial\chi_w^2}\chi_w^{(2)}\right)\left(1-\frac{2}{\alpha}\frac{\partial^2 F}{\partial\chi_u^2}\chi_u^{(2)}\right)-\frac{4}{\alpha}\left(\frac{\partial^2 F}{\partial\chi_w\partial\chi_u}\right)^2\chi_w^{(2)}\chi_u^{(2)}<0, \tag{25}$$

where

$$\chi_w^{(2)} = \int \mathrm{D}z\left(\frac{\partial^2}{\partial(\sqrt{\widehat{q}_w}z)^2}\ln\left[\mathop{\mathrm{Tr}}_{w}P(w)\exp\left[-\frac{\widehat{\chi}_w}{2}w^2+\sqrt{\widehat{q}_w}zw\right]\right]\right)^2, \tag{26}$$

and

$$\chi_u^{(2)} = \frac{1}{2}\mathop{\mathrm{Tr}}_{y}\int \mathrm{D}z\left(\frac{\partial^2}{\partial(\sqrt{\widehat{q}_u}z)^2}\ln\left[\int \mathrm{D}x\mathcal{I}(y|\sqrt{\widehat{\chi}_u}x+\sqrt{\widehat{q}_u}z)\right]\right)^2. \tag{27}$$

Equation (25) corresponds to the de Almeida–Thouless (AT) condition for the current system [31]. The second point is that although randomly labeled patterns are assumed here, one can develop a similar framework for analyzing the teacher–student scenario, which assigns pattern labels by a *teacher* perceptron, and which has a deep link to a certain class of modern wireless communication systems [8, 24, 25, 32–38]. One can find details of the framework in [18, 39].

3.3. The Thouless–Anderson–Palmer approach and message-passing algorithm

The scheme developed so far is used for investigating typical macroscopic properties of perceptrons which are averaged over a pattern set ξ^p. However, another method is necessary to evaluate microscopic properties of a perceptron for an individual sample of ξ^p. The Thouless–Anderson–Palmer (TAP) approach [40], originating in spin glass research, offers a useful guideline for this purpose. Although several formalisms are known for this approximation scheme [6], we follow the one based on the Gibbs free energy because of its generality and wide applicability [14, 22].

Let us suppose a situation for which the microscopic averages of the dynamical variables,

$$\begin{aligned}\boldsymbol{m}_w &= \mathop{\mathrm{Tr}}_{\boldsymbol{w}} \boldsymbol{w} P(\boldsymbol{w}|\xi^p) \\ &= \frac{1}{V(\xi^p)} \mathop{\mathrm{Tr}}_{\boldsymbol{u},\boldsymbol{w}} \boldsymbol{w} \prod_{\mu=1}^{p} \widehat{\mathcal{I}}_{y_\mu}(u_\mu) \prod_{i=1}^{N} P(w_i) \exp[\mathrm{i}\boldsymbol{u}^{\mathrm{T}} X \boldsymbol{w}],\end{aligned} \tag{28}$$

and

$$\boldsymbol{m}_u = \frac{1}{V(\xi^p)} \mathop{\mathrm{Tr}}_{\boldsymbol{u},\boldsymbol{w}} (\mathrm{i}\boldsymbol{u}) \prod_{\mu=1}^{p} \widehat{\mathcal{I}}_{y_\mu}(u_\mu) \prod_{i=1}^{N} P(w_i) \exp[\mathrm{i}\boldsymbol{u}^{\mathrm{T}} X \boldsymbol{w}], \tag{29}$$

are required, where $P(\boldsymbol{w}|\xi^p) = \prod_{\mu=1}^{p} \mathcal{I}(y_\mu|\Delta_\mu) \prod_{i=1}^{N} P(w_i)/V(\xi^p)$ denotes the posterior distribution of $\boldsymbol{w}$ given ξ^p. The Gibbs free energy

$$\Phi(\boldsymbol{m}_w, \boldsymbol{m}_u) = \mathop{\mathrm{Extr}}_{\boldsymbol{h}_w,\boldsymbol{h}_u} \{\boldsymbol{h}_w \cdot \boldsymbol{m}_w + \boldsymbol{h}_u \cdot \boldsymbol{m}_u - \ln[V(\boldsymbol{h}_w, \boldsymbol{h}_u)]\}, \tag{30}$$

where

$$V(\boldsymbol{h}_w, \boldsymbol{h}_u) = \mathop{\mathrm{Tr}}_{\boldsymbol{u},\boldsymbol{w}} \prod_{\mu=1}^{p} \widehat{\mathcal{I}}_{y_\mu}(u_\mu) \prod_{i=1}^{N} P(w_i) \exp[\boldsymbol{h}_w \cdot \boldsymbol{w} + \boldsymbol{h}_u \cdot (\mathrm{i}\boldsymbol{u}) + (\mathrm{i}\boldsymbol{u})^{\mathrm{T}} X \boldsymbol{w}], \tag{31}$$

offers a useful basis because the extremization conditions of equation (30) generally agree with equations (28) and (29). This indicates that one can evaluate the microscopic averages in equations (28) and (29) by extremization, which leads to assessment of the correct free energy, since $\ln V(\xi^p) = -\mathrm{Extr}_{\{\boldsymbol{m}_w,\boldsymbol{m}_u\}}\{\Phi(\boldsymbol{m}_w, \boldsymbol{m}_u)\}$ hold, once the function of Gibbs free energy (30) is provided.

Unfortunately, an exact evaluation of equation (30) is computationally difficult and therefore we resort to approximation. For this purpose, we put parameter l in front of X in equation (31), which yields the generalized Gibbs free energy as

$$\widetilde{\Phi}(\boldsymbol{m}_w, \boldsymbol{m}_u; l) = \mathop{\mathrm{Extr}}_{\boldsymbol{h}_w,\boldsymbol{h}_u} \{\boldsymbol{h}_w \cdot \boldsymbol{m}_w + \boldsymbol{h}_u \cdot \boldsymbol{m}_u - \ln[V(\boldsymbol{h}_w, \boldsymbol{h}_u; l)]\}, \tag{32}$$

where $V(\boldsymbol{h}_w, \boldsymbol{h}_u; l)$ is defined by replacing X with lX in equation (31). This implies that the correct Gibbs free energy in equation (30) can be obtained as $\Phi(\boldsymbol{m}_w, \boldsymbol{m}_u) = \widetilde{\Phi}(\boldsymbol{m}_w, \boldsymbol{m}_u; l = 1)$ by setting $l = 1$ in the generalized expression (32). One scheme for utilizing this relation is to perform the Taylor expansion around $l = 0$, for which $\widetilde{\Phi}(\boldsymbol{m}_w, \boldsymbol{m}_u; l)$ can be analytically calculated as an exceptional case, and substitute $l = 1$ into the expression obtained, which is sometimes referred to as the Plefka expansion [41]. However, evaluation of higher-order terms, which are non-negligible for correlated patterns in general, requires a complicated calculation in this expansion, which sometimes prevents the scheme from being practical. In order to avoid this difficulty, we take an alternative approach here, which is inspired by a derivative of equation (32),

$$\frac{\partial \widetilde{\Phi}(\boldsymbol{m}_w, \boldsymbol{m}_u; l)}{\partial l} = -\langle (\mathrm{i}\boldsymbol{u})^{\mathrm{T}} X \boldsymbol{w} \rangle_l, \tag{33}$$

where $\langle \cdots \rangle_l$ represents the average with respect to the generalized weight $\prod_{\mu=1}^{p} \widehat{\mathcal{I}}_{y_\mu}(u_\mu) \times \prod_{i=1}^{N} P(w_i) \times \exp[\boldsymbol{h}_w \cdot \boldsymbol{w} + \boldsymbol{h}_u \cdot (\mathrm{i}\boldsymbol{u}) + (\mathrm{i}\boldsymbol{u})^{\mathrm{T}}(lX)\boldsymbol{w}]$, and $\boldsymbol{h}_w$ and $\boldsymbol{h}_u$ are determined to satisfy $\langle \boldsymbol{w} \rangle_l = \boldsymbol{m}_w$ and $\langle (\mathrm{i}\boldsymbol{u}) \rangle_l = \boldsymbol{m}_u$, respectively [14]. The right-hand side of this equation is the average of a quadratic form containing many random variables. The central limit theorem implies that such an average does not depend on details of the objective distribution but is determined only by the values of the first and second moments. In order to construct a simple approximation scheme, let us assume that the second moments are characterized macroscopically by $\langle|\boldsymbol{w}-\langle\boldsymbol{w}\rangle_l|^2\rangle_l = N\chi_w$ and $\langle|\boldsymbol{u}-\langle\boldsymbol{u}\rangle_l|^2\rangle_l = p\chi_u$. Evaluating the right-hand side of equation (33) using a Gaussian distribution for which the first and second moments are constrained as $\langle\boldsymbol{w}\rangle_l = \boldsymbol{m}_w$, $\langle(\mathrm{i}\boldsymbol{u})\rangle_l = \boldsymbol{m}_u$, $\langle|\boldsymbol{w}-\langle\boldsymbol{w}\rangle_l|^2\rangle_l = N\chi_w$ and $\langle|\boldsymbol{u}-\langle\boldsymbol{u}\rangle_l|^2\rangle_l = p\chi_u$, and integrating from $l = 0$ to $l = 1$ yields

$$\widetilde{\Phi}(\chi_w,\chi_u,\boldsymbol{m}_w,\boldsymbol{m}_u;1) - \widetilde{\Phi}(\chi_w,\chi_u,\boldsymbol{m}_w,\boldsymbol{m}_u;0) \simeq -\boldsymbol{m}_u^{\mathrm{T}}X\boldsymbol{m}_w - NF(\chi_w,\chi_u), \tag{34}$$

where the function $F(x, y)$ is provided as in equation (8) by the empirical eigenvalue spectrum of $X^{\mathrm{T}}X$, $\rho(\lambda) = N^{-1}\sum_{k=1}^{N}\delta(\lambda-\lambda_k)$ and the macroscopic second moments χ_w and χ_u are included in arguments of the Gibbs free energy because the right-hand side of equation (33) depends on them. Utilizing this and evaluating $\widetilde{\Phi}(\chi_w,\chi_u,\boldsymbol{m}_w,\boldsymbol{m}_u;0)$, which is not computationally difficult since interaction terms are not included, yield an approximation of the Gibbs free energy as

$$\begin{aligned}\Phi(\chi_w,\chi_u,\boldsymbol{m}_w,\boldsymbol{m}_u) &\simeq -\boldsymbol{m}_u^{\mathrm{T}}X\boldsymbol{m}_w - NF(\chi_w,\chi_u)\\ &+ \mathop{\mathrm{Extr}}_{\widehat{\chi}_w,\boldsymbol{h}_w}\left\{\boldsymbol{h}_w\cdot\boldsymbol{m}_w - \frac{1}{2}\widehat{\chi}_w(N\chi_w+|\boldsymbol{m}_w|^2) - \sum_{i=1}^{N}\ln\left[\mathop{\mathrm{Tr}}_{w}P(w)\,\mathrm{e}^{-\frac{1}{2}\widehat{\chi}_w w^2 + h_{wi}w}\right]\right\}\\ &+ \mathop{\mathrm{Extr}}_{\widehat{\chi}_u,\boldsymbol{h}_u}\left\{\boldsymbol{h}_u\cdot\boldsymbol{m}_u - \frac{1}{2}\widehat{\chi}_u(p\chi_u-|\boldsymbol{m}_u|^2) - \sum_{\mu=1}^{p}\ln\left[\int Dx\mathcal{I}(y_\mu|\sqrt{\widehat{\chi}_u}x+h_{u\mu})\right]\right\},\end{aligned} \tag{35}$$

which is a general expression of the TAP free energy of the current system. Extremization of this equation provides a set of TAP equations

$$m_{wi} = \frac{\partial}{\partial h_{wi}}\ln\left[\mathop{\mathrm{Tr}}_{w}P(w)\,\mathrm{e}^{-\frac{1}{2}\widehat{\chi}_w w^2 + h_{wi}w}\right], \tag{36}$$

$$\chi_w = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial h_{wi}^2}\ln\left[\mathop{\mathrm{Tr}}_{w}P(w)\,\mathrm{e}^{-\frac{1}{2}\widehat{\chi}_w w^2 + h_{wi}w}\right], \tag{37}$$

$$m_{u\mu} = \frac{\partial}{\partial h_{u\mu}}\ln\left[\int Dx\mathcal{I}(y_\mu|\sqrt{\widehat{\chi}_u}x+h_{u\mu})\right], \tag{38}$$

$$\chi_u = -\frac{1}{p}\sum_{\mu=1}^{p}\frac{\partial^2}{\partial h_{u\mu}^2}\ln\left[\int Dx\mathcal{I}(y_\mu|\sqrt{\widehat{\chi}_u}x+h_{u\mu})\right], \tag{39}$$

where

$$\boldsymbol{h}_w = X^{\mathrm{T}}\boldsymbol{m}_u - 2\frac{\partial}{\partial\chi_w}F(\chi_w,\chi_u)\boldsymbol{m}_w, \tag{40}$$

$$\widehat{\chi}_w = -2\frac{\partial}{\partial\chi_w}F(\chi_w,\chi_u), \tag{41}$$

$$\boldsymbol{h}_u = X\boldsymbol{m}_w + \frac{2}{\alpha}\frac{\partial}{\partial\chi_u}F(\chi_w,\chi_u)\boldsymbol{m}_u, \tag{42}$$

$$\widehat{\chi}_u = -\frac{2}{\alpha}\frac{\partial}{\partial\chi_u}F(\chi_w,\chi_u), \tag{43}$$

solutions of which represent approximate values of the first and second moments of the posterior distribution $P(\boldsymbol{w}, \boldsymbol{u}|\xi^p)$ for a fixed sample of ξ^p. In equations (40) and (42), $-2(\partial/\partial\chi_w)F(\chi_w, \chi_u)\boldsymbol{m}_w$ and $(2/\alpha)(\partial/\partial\chi_u)F(\chi_w, \chi_u)\boldsymbol{m}_u$ are generally referred to as the Onsager reaction terms. The counterparts of these equations for systems of two-body interactions have been presented in an earlier paper [22].

Solving TAP equations (36)–(43) is not a trivial task. Empirically, naive iterative substitution of these equations does not converge in most cases. Conversely, it is reported that message-passing (MP) algorithms of a certain type, which are developed on the basis of the belief propagation [42], exhibit excellent solution search performance for pattern sets entries of which are IID with low-computational cost [32, 43]. Therefore, we developed an MP algorithm as a promising heuristic that reproduces known, efficient algorithms for IID pattern matrices. A pseudocode of the proposed algorithm is shown in figure 1. One can generalize this algorithm to the case of probabilistic perceptrons by replacing the indicator function $\mathcal{I}(y|\Delta)$ with a certain conditional probability $P(y|\Delta)$. It should be noted that Λ_w and Λ_u in the algorithm denote the counterparts of Λ_x and Λ_y in equation (8) for $x = \chi_w$ and $y = \chi_u$, respectively. Solving (χ_u, Λ_u) and (χ_w, Λ_w) in **H-Step** and **V-Step**, respectively, can be performed efficiently by the use of the bisection method. Solving the TAP equations employing this algorithm yields approximate estimates of the free energy $\ln V(\xi^p)$ and its derivatives as well as $\boldsymbol{m}_w$ and $\boldsymbol{m}_u$, which can be utilized for assessing whether the given specific sample ξ^p can be correctly classified by the perceptron.

Although we have assumed single macroscopic constraints as characterizing the second moments, the current formalism can be generalized to include componentwise multiple constraints for constructing more accurate approximations. By doing this, the current formalism leads to the adaptive TAP approach or, more generally, to the expectation consistent approximate schemes developed by Opper and Winther [12–14].

4. Examples

4.1. Independently and identically distributed patterns

In order to investigate the relationship with existing results, let us first apply the developed methodologies to the case in which the entries of X are IID of zero mean and variance N^{-1}. This case can be characterized by the eigenvalue spectrum of the Marčenko–Pastur-type, which was already mentioned in section 2 and yields

$$F(x, y) = -\frac{\alpha}{2}xy. \tag{44}$$

This implies that equation (22) can be expressed as

$$\mathcal{A}_0(\chi_w, \chi_u, q_w, q_u) = -\frac{\alpha}{2}(\chi_w\chi_u + q_w\chi_u - q_u\chi_w). \tag{45}$$

Inserting this into equation (21) and then performing an extremization with respect to χ_u and q_u yields

$$\widehat{\chi}_u = \chi_w, \qquad \widehat{q}_u = q_w, \tag{46}$$

where $\widehat{\chi}_u$ and $\widehat{q}_u$ are the variational variables used in equation (24). This implies that the replica symmetric free energy (21) can be expressed as

$$\frac{1}{N}[\ln V(\xi^p)]_{\xi^p} = \mathop{\mathrm{Extr}}_{\chi_w,q_w}\left\{\mathcal{A}_w(\chi_w, q_w) + \frac{\alpha}{2}\mathop{\mathrm{Tr}}_{y}\int Dz\ln\left[\int Dx\mathcal{I}(y|\sqrt{\chi_w}x + \sqrt{q_w}z)\right]\right\}. \tag{47}$$

```
**MPforPerceptron**{
        Perform **Initialization**;
        Iterate **H-Step** and **V-Step** alternately sufficient times;
}
**Initialization**{
        $\chi_w \leftarrow \frac{1}{N}\sum_{i=1}^{N} w_i^2 P(w_i);\quad \widehat{\chi}_w \leftarrow 0;\quad \Lambda_w \leftarrow \frac{1}{\chi_w} - \widehat{\chi}_w;$
        $m_{wi} \leftarrow \mathop{\mathrm{Tr}}_{w_i} w_i P(w_i) \qquad (i = 1, 2, \ldots, N);$
        $\boldsymbol{h}_u \leftarrow X\boldsymbol{m}_w;\quad \boldsymbol{m}_u \leftarrow \boldsymbol{0};$
}
**H-Step**{
        Search $(\chi_u, \Lambda_u)$ for given $(\chi_w, \Lambda_w)$ to satisfy conditions
            $\chi_w = \left\langle \frac{\Lambda_u}{\Lambda_w\Lambda_u + \lambda} \right\rangle_\rho$ and $\chi_u = (1-\alpha^{-1})\frac{1}{\Lambda_u} + \alpha^{-1}\left\langle \frac{\Lambda_w}{\Lambda_w\Lambda_u + \lambda} \right\rangle_\rho;$
        $\widehat{\chi}_u \leftarrow \frac{1}{\chi_u} - \Lambda_u;$
        $\boldsymbol{h}_u \leftarrow \boldsymbol{h}_u - \widehat{\chi}_u \boldsymbol{m}_u;$
        $m_{u\mu} \leftarrow \frac{\partial}{\partial h_{u\mu}} \ln\left[\int Dx\mathcal{I}(y_\mu|\sqrt{\widehat{\chi}_u}x + h_{u\mu})\right] \qquad (\mu = 1, 2, \ldots, p);$
        $\boldsymbol{h}_w \leftarrow X^{\mathrm{T}}\boldsymbol{m}_u;$
        $\chi_u \leftarrow -\frac{1}{p}\sum_{\mu=1}^{p} \frac{\partial^2}{\partial h_{u\mu}^2} \ln\left[\int Dx\mathcal{I}(y_\mu|\sqrt{\widehat{\chi}_u}x + h_{u\mu})\right];$
        $\Lambda_u \leftarrow \frac{1}{\chi_u} - \widehat{\chi}_u;$
}
**V-Step**{
        Search $(\chi_w, \Lambda_w)$ for given $(\chi_u, \Lambda_u)$ to satisfy conditions
            $\chi_w = \left\langle \frac{\Lambda_u}{\Lambda_w\Lambda_u + \lambda} \right\rangle_\rho$ and $\chi_u = (1-\alpha^{-1})\frac{1}{\Lambda_u} + \alpha^{-1}\left\langle \frac{\Lambda_w}{\Lambda_w\Lambda_u + \lambda} \right\rangle_\rho;$
        $\widehat{\chi}_w \leftarrow \frac{1}{\chi_w} - \Lambda_w;$
        $\boldsymbol{h}_w \leftarrow \boldsymbol{h}_w + \widehat{\chi}_w \boldsymbol{m}_w;$
        $m_{wi} \leftarrow \frac{\partial}{\partial h_{wi}} \ln\left[\mathop{\mathrm{Tr}}_{w} P(w)e^{-\frac{1}{2}\widehat{\chi}_w w^2 + h_{wi}w}\right] \qquad (i = 1, 2, \ldots, \mathrm{N});$
        $\boldsymbol{h}_u \leftarrow X\boldsymbol{m}_w;$
        $\chi_w \leftarrow \frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2}{\partial h_{wi}^2} \ln\left[\mathop{\mathrm{Tr}}_{w} P(w)e^{-\frac{1}{2}\widehat{\chi}_w w^2 + h_{wi}w}\right];$
        $\Lambda_w \leftarrow \frac{1}{\chi_w} - \widehat{\chi}_w;$
}
```

Figure 1. Pseudocode of the proposed message-passing algorithm **MPforPerceptron**. ';' and '←' represent the end of a command line and the operation of substitution, respectively.

This is equivalent to the general expression of the replica symmetric free energy of a single-layer perceptron for the IID pattern matrices and randomly assigned labels [4, 44].

4.2. Rank deficient patterns versus spherical weights

In data analysis, the property of pattern components strongly correlated with each other is referred to as *multicollinearity*, which sometimes requires special treatment. As a second example, we utilize the developed framework to examine how this property influences α_c.

Strong correlations among components can be modeled by rank deficiency of the cross-correlation matrix $X^{\mathrm{T}}X$. In the current framework, this is characterized by an eigenvalue spectrum of the form

$$\rho(\lambda) = (1-c)\delta(\lambda) + c\widetilde{\rho}(\lambda), \tag{48}$$

where $0 < c \leqslant 1$ denotes the ratio between the rank of $X^{\mathrm{T}}X$ and N, and $\tilde{\rho}(\lambda)$ is a certain distribution the support of which is defined over a region of $\lambda > 0$. For simplicity, let us limit ourselves to the case of simple perceptron and spherical weights, for which $\mathcal{I}(y|\Delta) = 1$ for $y|\Delta > 0$ and 0, otherwise, and $P(\boldsymbol{w}) \propto \delta(|\boldsymbol{w}|^2 - N)$. Inserting these into equation (21) offers a set of saddle-point equations. Among them, those relevant for capacity analysis are

$$\chi_w = (1-c)\frac{1}{\Lambda_w} + c\left\langle \frac{\Lambda_u}{\Lambda_w\Lambda_u + \lambda}\right\rangle_{\widetilde{\rho}}, \tag{49}$$

$$\chi_u = \left(1 - \frac{c}{\alpha}\right)\frac{1}{\Lambda_u} + \frac{c}{\alpha}\left\langle \frac{\Lambda_w}{\Lambda_w\Lambda_u + \lambda}\right\rangle_{\widetilde{\rho}}, \tag{50}$$

$$\widehat{\chi}_u = -\frac{2}{\alpha}\frac{\partial F(\chi_w, \chi_u)}{\partial \chi_u} = \frac{1}{\chi_u} - \Lambda_u, \tag{51}$$

$$\chi_u = -\int \mathrm{D}z \frac{\partial^2}{(\partial\sqrt{\widehat{q}_u}z)^2} \ln H\left(\sqrt{\frac{\widehat{q}_u}{\widehat{\chi}_u}}z\right), \tag{52}$$

where $H(x) = \int_x^{+\infty} \mathrm{D}z$.

Let us assume that no RSB occurs for $\alpha < \alpha_c$, as is the case for IID patterns. Under this assumption, a critical condition is offered by taking a limit $\widehat{\chi}_u \to 0$, which implies that the variance of $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_p)^{\mathrm{T}}$ of the posterior distribution for a given sample ξ^p typically vanishes. Applying an asymptotic form, $\ln H(x) \simeq -x^2/2$ for $x \gg 1$, to equation (52) in conjunction with equation (51) yields

$$\Lambda_u \simeq \frac{1}{2\chi_u}. \tag{53}$$

Inserting this into equation (50) gives

$$\frac{2c}{\alpha} - 1 \simeq \frac{c}{\alpha}\left\langle \frac{2\Lambda_w}{\Lambda_w + 2\lambda\chi_u}\right\rangle_{\widetilde{\rho}} \geqslant 0. \tag{54}$$

This means that no RS solution can exist for $\alpha > 2c$, indicating that the perceptron capacity is given as

$$\alpha_c = 2c, \tag{55}$$

regardless of $\widetilde{\rho}(\lambda)$. Equation (55) is consistent with the known result $\alpha_c = 2$ for IID patterns [15, 16], for which $c = 1$ as $X^{\mathrm{T}}X$ is typically of full rank for $\alpha > 1$. Numerical experiments for rank deficient pattern matrices support the present analysis, which is shown in figure 2.

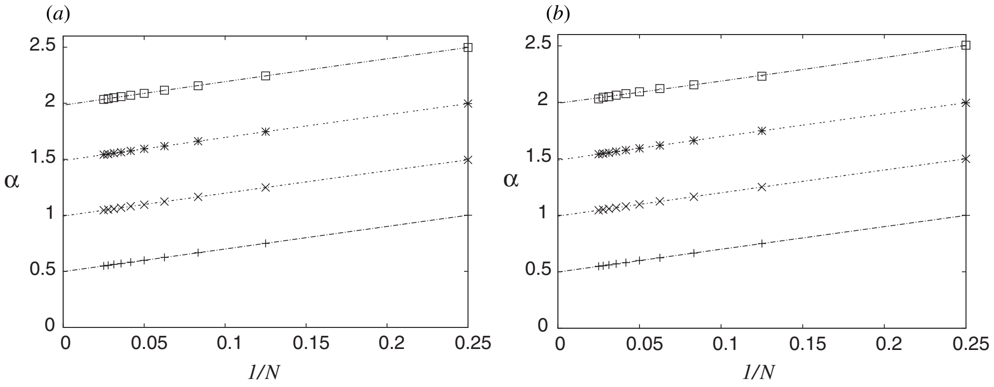

Figure 2. Assessment of $\alpha_{\rm c}$ of spherical weights for rank deficient pattern matrices. For $N = 4, 8, 12, \ldots, 40$, the critical pattern ratio $\alpha_{\rm c}(N)$, which is defined as the average of the maximum pattern ratio above which no weight can correctly classify a given sample of ξ^p, was assessed from 10^4 experiments. Each estimate of $\alpha_{\rm c}(N)$ was obtained by extrapolating $\alpha_{\rm c}(N, T_{\max})$, which is an average value of α above which the perception learning algorithm [48] does not converge after the number of updates reaches $T_{\max}$ for a given sample of ξ^p with respect to $T_{\max} = 10^3 \sim 2 \times 10^4$. The capacity is estimated by a quadratic fitting under the assumption of $\alpha_{\rm c}(N) \simeq \alpha_{\rm c} + aN^{-1} + bN^{-2}$ where a and b are adjustable parameters. (*a*) and (*b*) represent results for $\widetilde{\rho}(\lambda) = (2\pi\lambda)^{-1}\sqrt{[\lambda - (\sqrt{\alpha/c} - 1)^2]^+[(\sqrt{\alpha/c} + 1)^2 - \lambda]^+}$ and $\delta(\lambda - 1)$, respectively. For both cases, each data corresponds to $c = 1/4, 1/2, 3/4$ and 1 from the bottom. The estimates of $\alpha_{\rm c}$ show excellent consistency with the theoretical prediction $\alpha_{\rm c} = 2c$ regardless of $\widetilde{\rho}(\lambda)$.

4.3. Random orthogonal patterns versus binary weights

Equation (55) means that the capacity depends only on the rank of the cross-correlation matrix $X^{\rm T}X$ in the case of spherical weights; however, this is not always the case. To show this, we present a capacity problem of binary weights $\boldsymbol{w} = \{+1, -1\}^N$ as the final example.

It is known that in typical cases, simple perceptrons of binary weights can correctly classify randomly labeled IID patterns for $\alpha < \alpha_{\rm c} \simeq 0.833$ [45–47]. Our question here is how $\alpha_{\rm c}$ is modified when the pattern matrix X is generated randomly in such a way that patterns $\boldsymbol{x}_\mu$ are orthogonal to each other.

To answer this question, we employ the replica and TAP methods developed in the preceding sections for $\rho(\lambda) = (1 - \alpha)\delta(\lambda) + \alpha\delta(\lambda - 1)$, which represents the eigenvalue spectrum of the random orthogonal patterns assuming $0 < \alpha < 1$ and yields

$$F(x, y) = -1 + \left(\mathcal{L} - \frac{1}{2}\ln\mathcal{L}\right), \qquad (56)$$

where $\mathcal{L} = 2^{-1}(1 \pm \sqrt{1 - 4\alpha xy})$. Here, ± 1 is chosen so that the operation of $\mathrm{Extr}_{\Lambda_x, \Lambda_y}\{\cdots\}$ in equation (8) corresponds to the correct saddle-point evaluation of equation (A.3). Figure 3(*a*) shows how the entropy of $\boldsymbol{w}$ depends on the pattern ratio α. The curve denotes the theoretical prediction of the replica analysis and the markers denote the averages of the entropy obtained by the TAP method over 100 samples for $N = 500$ systems. The error bars are smaller than the markers. Solutions of the TAP method are obtained by **MPforPerceptron**, shown in figure 1. Although the curve and the markers exhibit excellent agreement for data points $\alpha = 0.1, 0.2, \ldots, 0.8$, we were not able to obtain a reliable result for $\alpha = 0.9$, at which point this algorithm does not converge in most cases, even after 1000 iterations. This may be a consequence of RSB since the replica analysis indicates that the AT stability of the RS

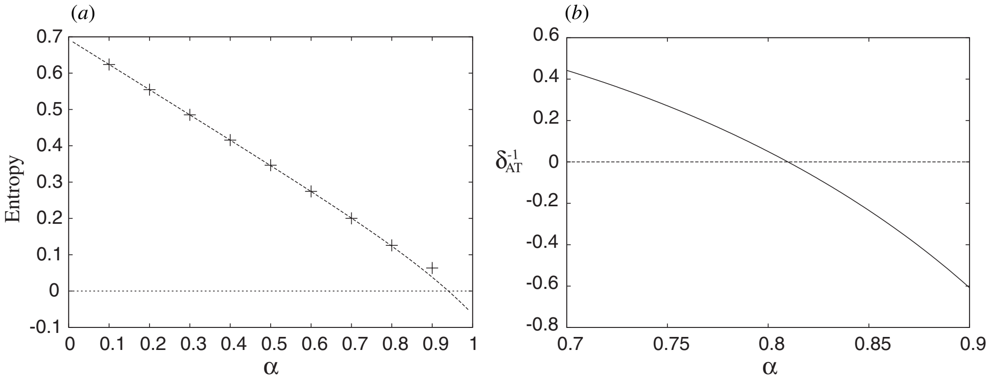

Figure 3. (*a*) Entropy of $\boldsymbol{w}$ (per element) versus the pattern ratio α. The curve represents the theoretical prediction assessed by the replica method and the markers denote experimental data obtained by **MPforPerceptron** for 100 samples of ξ^p of $N = 500$ systems. (*b*) Diagnosis of the AT stability. $\delta_{\rm AT}^{-1}$, which is the inverse of the left-hand side of equation (25), is plotted versus α for the assessed RS solution. $\delta_{\rm AT}^{-1}$ becomes negative for $\alpha > \alpha_{\rm AT} \simeq 0.810$, indicating the occurrence of RSB.

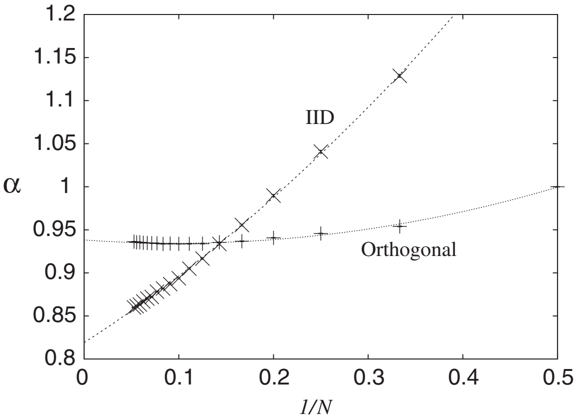

Figure 4. Results of exhaustive search experiments. For $N = 2, 3, \ldots, 20, \alpha_{\rm c}(N)$, which is defined in figure 2, were estimated from 10^6 experiments performed by an exhaustive search of binary weights. The values of capacity $\alpha_{\rm c}$ are estimated by employing a quadratic fitting similar to that explained in figure 2. For IID patterns, this yields an estimate of $\alpha_{\rm c} \simeq 0.819$, whereas the theoretical prediction is 0.833 and is considered as exact. The estimate $\alpha_{\rm c} \simeq 0.938$ for the random orthogonal patterns is reasonably close to the theoretical prediction 0.940, which is obtained from the unstable RS solution.

solution shown in figure 3(*a*) is broken for α beyond $\alpha_{\rm AT} \simeq 0.810$ (see figure 3(b)). Therefore $\alpha_{\rm c} \simeq 0.940$, indicated by the condition of vanishing entropy is not regarded as the exact, but as an approximate value provided by the unstable RS solution. However, extrapolation of the results of direct numerical experiments for finite-size systems indicates that $\alpha_{\rm c} \simeq 0.938$, as shown in figure 4, which implies that the effect of RSB is not significant for the evaluation of $\alpha_{\rm c}$ in this particular case.

5. Summary

We developed a framework for analyzing the classification problems of perceptrons for randomly labeled patterns. The development is intended to handle correlated patterns. For this purpose, we developed two methodologies based on the replica method and the Thouless–Anderson–Palmer (TAP) approach, which are standard techniques from the statistical mechanics of disordered systems, and introduced a certain specific random assumption about the singular value decomposition of the pattern matrix. In both schemes, an integral formula, which can be regarded as a generalization of the Itzykson–Zuber integral known for square (symmetric) matrices, plays an important role. As a promising heuristic for solving TAP equations, we provided a message-passing algorithm **MPforPerceptron**. The validity and utility of the developed schemes are shown for one known result and two novel problems.

Investigation of the properties of **MPforPerceptron**, as well as application of the developed framework to real-world data analysis [43, 49] and various models of information and communication engineering [17, 50], are promising topics for future research.

Acknowledgments

This research was supported in part by the JSPS Global COE program ‘Computationism as a Foundation for the Sciences’ (TS and YK) and Grants-in-Aid MEXT/JSPS, Japan, Nos. 17340116 and 18079006 (YK).

Appendix A. Derivation of equations (7) and (8)

The expressions

$$\delta(|\tilde{\boldsymbol{w}}|^2 - Nx) = \frac{1}{2\pi \mathrm{i}}\int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \frac{\mathrm{d}\Lambda_x}{2} \exp\left[-\frac{\Lambda_x}{2}(|\tilde{\boldsymbol{w}}|^2 - Nx)\right], \tag{A.1}$$

$$\delta(|\tilde{\boldsymbol{u}}|^2 - py) = \frac{1}{2\pi \mathrm{i}}\int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \frac{\mathrm{d}\Lambda_y}{2} \exp\left[-\frac{\Lambda_y}{2}(|\tilde{\boldsymbol{u}}|^2 - py)\right], \tag{A.2}$$

yield an integral

$$\begin{aligned}
&\int \mathrm{d}\tilde{\boldsymbol{w}}\,\mathrm{d}\tilde{\boldsymbol{u}}\,\delta(|\tilde{\boldsymbol{w}}|^2 - Nx)\delta(|\tilde{\boldsymbol{u}}|^2 - py)\exp[\mathrm{i}\tilde{\boldsymbol{u}}^{\mathrm{T}} D\tilde{\boldsymbol{w}}] \\
&\quad = \frac{1}{(4\pi \mathrm{i})^2}\int \mathrm{d}\Lambda_x\,\mathrm{d}\Lambda_y \left(\int \mathrm{d}\tilde{\boldsymbol{w}}\,\mathrm{d}\tilde{\boldsymbol{u}} \exp\left[-\frac{\Lambda_x|\tilde{\boldsymbol{w}}|^2}{2} - \frac{\Lambda_y|\tilde{\boldsymbol{u}}|^2}{2} + \mathrm{i}\tilde{\boldsymbol{u}}^{\mathrm{T}} D\tilde{\boldsymbol{w}}\right]\right) \\
&\qquad \times \exp\left[\frac{N\Lambda_x x}{2} + \frac{p\Lambda_y y}{2}\right]. \\
&\quad = \frac{(2\pi)^{(N+p)/2}}{(4\pi \mathrm{i})^2}\int \mathrm{d}\Lambda_x\,\mathrm{d}\Lambda_y \left(\det\left[\begin{array}{c|c}\Lambda_x I_N & -\mathrm{i}D^{\mathrm{T}} \\ -\mathrm{i}D & \Lambda_y I_p\end{array}\right]\right)^{-1/2} \\
&\qquad \times \exp\left[\frac{N\Lambda_x x}{2} + \frac{p\Lambda_y y}{2}\right],
\end{aligned} \tag{A.3}$$

where I_N and I_p are $N \times N$ and $p \times p$ identity matrices, respectively. Linear algebra can be used to generate the expression

$$\ln\det\left[\begin{array}{c|c}\Lambda_x I_N & -\mathrm{i}D^{\mathrm{T}} \\ \hline -\mathrm{i}D & \Lambda_y I_p\end{array}\right] = \sum_{k=1}^{\min(p,N)} \ln(\Lambda_x\Lambda_y+\lambda_k) + (N-\min(p,N))\ln\Lambda_y$$
$$\simeq N(\langle\ln(\Lambda_x\Lambda_y+\lambda)\rangle_\rho + (\alpha-1)\ln\Lambda_y), \qquad \text{(A.4)}$$

in the large system limit $N, p \to \infty$, keeping $\alpha = p/N \sim O(1)$. This implies that equation (A.3) can be evaluated by the saddle-point method as

$$\frac{1}{N}\ln\left[\int \mathrm{d}\tilde{\boldsymbol{w}}\,\mathrm{d}\tilde{\boldsymbol{u}}\,\delta(|\tilde{\boldsymbol{w}}|^2-Nx)\delta(|\tilde{\boldsymbol{u}}|^2-py)\exp[\mathrm{i}\tilde{\boldsymbol{u}}^{\mathrm{T}}D\tilde{\boldsymbol{w}}]\right]$$
$$= \mathop{\mathrm{Extr}}_{\Lambda_x,\Lambda_y}\left\{-\frac{1}{2}\langle\ln(\Lambda_x\Lambda_y+\lambda)\rangle_\rho - \frac{\alpha-1}{2}\ln\Lambda_y + \frac{\Lambda_x x}{2} + \frac{\alpha\Lambda_y y}{2}\right\} + \mathrm{const}, \qquad \text{(A.5)}$$

where const represents constant terms that do not depend on either x or y. In particular, inserting $p \times N$ zero matrix $D = 0_{p,N}$ into this expression leads to

$$\frac{1}{N}\ln\left[\int \mathrm{d}\tilde{\boldsymbol{w}}\,\mathrm{d}\tilde{\boldsymbol{u}}\,\delta(|\tilde{\boldsymbol{w}}|^2-Nx)\delta(|\tilde{\boldsymbol{u}}|^2-py)\right]$$
$$= \mathop{\mathrm{Extr}}_{\Lambda_x,\Lambda_y}\left\{-\frac{1}{2}\ln\Lambda_x - \frac{\alpha}{2}\ln\Lambda_y + \frac{\Lambda_x x}{2} + \frac{\alpha\Lambda_y y}{2}\right\}$$
$$= \frac{1}{2}\ln x + \frac{\alpha}{2}\ln y + \frac{1+\alpha}{2} + \mathrm{const}. \qquad \text{(A.6)}$$

Equations (A.5) and (A.6) are used in equations (7) and (8).

Appendix B. Assessment of free energy under the 1RSB ansatz

The argument in section 3 implies that when $n \times n$ matrices $\mathcal{Q}_w = \left(q_w^{ab}\right)$ and $\mathcal{Q}_u = \left(q_u^{ab}\right)$ are simultaneously diagonalized by an identical orthogonal matrix, the average of the replicated coupling term with respect to U and V is evaluated as

$$\frac{1}{N}\ln\left[\overline{\exp\left[\mathrm{i}\sum_{a=1}^{n}(\boldsymbol{u}^a)^{\mathrm{T}}X\boldsymbol{w}^a\right]}\right] = \sum_{a=1}^{n}F\left(t_w^a, t_u^a\right), \qquad \text{(B.1)}$$

where t_w^a and t_u^a $(a = 1, 2, \ldots, n)$ denote a pair of eigenvalues of $\mathcal{Q}_w$ and $\mathcal{Q}_u$ and correspond to an identical eigen vector. Under the 1RSB ansatz, n replica indices are divided into n/m groups of identical size m, and the relevant saddle point is characterized as

$$\left(q_w^{ab}, q_u^{ab}\right) = \begin{cases}(\chi_w+v_w+q_w,\ \chi_u-v_u-q_u), & a=b,\\ (v_w+q_w,\ -v_u-q_u), & a \text{ and } b \text{ belong to an identical group},\\ (q_w,\ -q_u), & \text{otherwise},\end{cases} \qquad \text{(B.2)}$$

where m serves as Parisi's RSB parameter after analytical continuation. $\mathcal{Q}_w$ and $\mathcal{Q}_u$ of the form of equation (B.2) can be simultaneously diagonalized, which yields pairs of eigenvalues as

$$\left(t_w^a, t_u^a\right) = \begin{cases}(\chi_w+mv_w+nq_w,\ \chi_u-mv_u-nq_u), & 1,\\ (\chi_w+mv_w,\ \chi_u-mv_u), & n/m-1,\\ (\chi_w,\ \chi_u), & n-n/m,\end{cases} \qquad \text{(B.3)}$$

where the numbers in the right-most column represent the degeneracies of the pair of eigenvalues denoted in the middle column. This gives

$$\frac{1}{N}\ln\left[\overline{\exp\left[\mathrm{i}\sum_{a=1}^{n}(\boldsymbol{u}^a)^{\mathrm{T}}X\boldsymbol{w}^a\right]}\right]$$
$$= F(\chi_w+mv_w+nq_w,\,\chi_u-mv_u-nq_u)+\left(\frac{n}{m}-1\right)F(\chi_w+mv_w,\,\chi_u-mv_u)$$
$$+\left(n-\frac{n}{m}\right)F(\chi_w,\,\chi_u). \qquad \text{(B.4)}$$

Equation (B.4) and assessment of the volumes of dynamical variables $\{\boldsymbol{w}^a\}$ and $\{\boldsymbol{u}^a\}$ under the 1RSB ansatz (B.2), in conjunction with analytical continuation from $n \in \mathbb{N}$ to $n \in \mathbb{R}$, lead to the expression of the 1RSB free energy as

$$\frac{1}{N}[\ln V(\xi^p)]_{\xi^p} = \lim_{n\to 0}\frac{\partial}{\partial n}\frac{1}{N}\ln[V^n(\xi^p)]_{\xi^p} = \mathop{\mathrm{Extr}}_{\boldsymbol{\Theta},m}\left\{\mathcal{A}_0^{\mathrm{1RSB}}(\chi_w,\chi_u,v_w,v_u,q_w,q_u;m)\right.$$
$$\left.+\mathcal{A}_w^{\mathrm{1RSB}}(\chi_w,v_w,q_w;m)+\alpha\mathcal{A}_u^{\mathrm{1RSB}}(\chi_u,v_u,q_u;m)\right\}, \qquad \text{(B.5)}$$

where $\boldsymbol{\Theta} = (\chi_w,\chi_u,v_w,v_u,q_w,q_u)$,

$$\mathcal{A}_0^{\mathrm{1RSB}}(\chi_w,\chi_u,v_w,v_u,q_w,q_u;m) = F(\chi_w,\chi_u)+\frac{1}{m}(F(\chi_w+mv_w,\chi_u-mv_u)-F(\chi_w,\chi_u))$$
$$+q_w\frac{\partial F(\chi_w+mv_w,\chi_u-mv_u)}{\partial\chi_w}-q_u\frac{\partial F(\chi_w+mv_w,\chi_u-mv_u)}{\partial\chi_u}, \qquad \text{(B.6)}$$

$$\mathcal{A}_w^{\mathrm{1RSB}}(\chi_w,v_w,q_w;m)$$
$$=\mathop{\mathrm{Extr}}_{\widehat{\chi}_w,\widehat{v}_w,\widehat{q}_w}\left\{\frac{\widehat{\chi}_w(\chi_w+v_w+q_w)}{2}-\frac{\widehat{v}_w(\chi_w+m(v_w+q_w))}{2}-\frac{\widehat{q}_w(\chi_w+mv_w)}{2}\right.$$
$$\left.+\frac{1}{m}\int Dz\ln\left[\int Dy\left(\mathop{\mathrm{Tr}}_{w}P(w)\,\mathrm{e}^{-\frac{\widehat{\chi}_w}{2}w^2+(\sqrt{\widehat{v}_w}y+\sqrt{\widehat{q}_w}z)w}\right)^m\right]\right\}, \qquad \text{(B.7)}$$

and

$$\mathcal{A}_u^{\mathrm{1RSB}}(\chi_u,v_u,q_u;m)=\mathop{\mathrm{Extr}}_{\widehat{\chi}_u,\widehat{v}_u,\widehat{q}_u}\left\{\frac{\widehat{\chi}_u(\chi_u-v_u-q_w)}{2}+\frac{\widehat{v}_u(\chi_u-m(v_u+q_u))}{2}+\frac{\widehat{q}_u(\chi_u-mv_u)}{2}\right.$$
$$\left.+\frac{1}{2m}\mathop{\mathrm{Tr}}_{y}\int Dz\ln\left[\int Ds\left(\int Dx\mathcal{I}(y|\sqrt{\widehat{\chi}_u}x+\sqrt{\widehat{v}_u}s+\sqrt{\widehat{q}_u}z)\right)^m\right]\right\}. \qquad \text{(B.8)}$$

References

[1] Levin E, Tishby N and Solla S A 1990 *Proc. IEEE* **78** 1568

[2] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific) p 3

[3] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499

[4] Engel A and van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)

[5] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing—An Introduction* (Oxford: Oxford University Press)

[6] Opper M and Saad D ed 2001 *Advanced Mean Field Methods: Theory and Practice* (Cambridge, MA: MIT Press)

[7] Mézard M, Parisi G and Zecchina R 2002 *Science* **297** 812

[8] Tanaka T 2002 *IEEE Trans. Inform. Theory* **48** 2888

[9] Hosaka T, Kabashima Y and Nishimori H 2002 *Phys. Rev.* E **66** 066126

[10] Kinzel W and Kanter I 2002 *Proc. ICONIP02́* vol 3 p 1351

[11] Mimura K and Okada M 2006 *Phys. Rev.* E **74** 026108

[12] Opper M and Winther O 2001 *Phys. Rev. Lett.* **86** 3695

[13] Opper M and Winther O 2001 *Phys. Rev.* E **64** 056131

[14] Opper M and Winther O 2005 *J. Mach. Lear. Res.* **6** 2177

[15] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[16] Cover T M 1965 *IEEE Trans. Electron. Comput.* **14** 326

[17] Tulino A M and Verdú S 2004 *Random Matrix Theory and Wireless Communications* (Hanover, MA: Now Publishers)

[18] Kabashima Y 2008 *J. Phys. Conf. Ser.* **95** 012001

[19] Itzykson C and Zuber J B 1980 *J. Math. Phys.* **21** 411

[20] Voiculescu D V, Dykema K J and Nica A 1992 *Free Random Variables* (Providence, RI: American Mathematical Society)

[21] Marinari E, Parisi G and Ritort F 1994 *J. Phys. A: Math. Gen.* **27** 7647

[22] Parisi G and Potters M 1995 *J. Phys. A: Math. Gen.* **28** 5267

[23] Cherrier R, Dean D S and Lefèvre A 2003 *Phys. Rev.* E **67** 046112

[24] Takeda K, Uda S and Kabashima Y 2006 *Europhys. Lett.* **76** 1193

[25] Takeda K, Hatabu A and Kabashima Y 2007 *J. Phys. A: Math. Theor.* **40** 14085

[26] Müller R R, Guo D and Moustakas A L 2007 Vector precoding for wireless MIMO systems: a replica analysis *Preprint* 0706.1169

[27] Tanaka T 2008 *J. Phys. Conf. Ser.* **95** 012002

[28] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792

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

[30] Dotsenko V S 2001 *Introduction to the Replica Theory of Disordered Statistical Systems* (Cambridge: Cambridge University Press)

[31] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983

[32] Kabashima Y 2003 *J. Phys. A: Math. Gen.* **36** 11111

[33] Müller R R 2003 *IEEE Trans. Signal Process.* **51** 2821

[34] Moustakas A L, Simon S H and Sengupta A M 2003 *IEEE Trans. Inform. Theory* **49** 2545

[35] Guo D and Verdú S 2005 *IEEE Trans. Inform. Theory* **51** 1983

[36] Neirotti J P and Saad D 2005 *Europhys. Lett.* **71** 866

[37] Montanari A and Tse D 2006 Analysis of belief propagation for non-linear problems: the example of CDMA (or: How to prove Tanaka's formula) *Proc. IEEE Inform. Theory Workshop* (Punta del Este: Uruguay) (*Preprint* cs/0602028)

[38] Montanari A, Prabhakar B and Tse D 2005 Belief propagation based multi-user detection *Preprint* cs/0510044

[39] Kabashima Y 2008 An integral formula for large random rectangular matrices and its application to analysis of linear vector channels *Preprint* 0802.1372

[40] Thouless D J, Anderson P W and Palmer R G 1977 *Phil. Mag.* **35** 593

[41] Plefka T 1982 *J. Phys. A: Math. Gen.* **15** 1971

[42] Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems* (San Mateo, CA: Morgan Kaufmann Publishers)

[43] Uda S and Kabashima Y 2005 *J. Phys. Soc. Jpn.* **74** 2233

[44] Opper M and Kinzel W 1996 *Models of Neural Networks III* ed E Domany, J L van Hemmen and K Schulten (New York: Springer) p 151

[45] Krauth W and Mézard M 1989 *J. Phys.* **50** 3056

[46] Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** L519

[47] Derrida B, Griffith R B and Prügel-Benett A 1991 *J. Phys. A: Math. Gen.* **24** 4907

[48] Minsky M and Papert S 1969 *Perceptrons* (Campridge, MA: MIT Press)

[49] Braunstein A, Pagnani A, Weigt M and Zecchina R 2008 *J. Phys. Conf. Ser.* **95** 012016

[50] Verdú S 1998 *Multiuser Detection* (Cambridge: Cambridge University Press)