Learning from correlated patterns by simple perceptrons

2009 J. Phys. A: Math. Theor. 42 015005

(http://iopscience.iop.org/1751-8121/42/1/015005)

IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **42** (2009) 015005 (12pp) doi:10.1088/1751-8113/42/1/015005

Learning from correlated patterns by simple perceptrons

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 8 September 2008, in final form 22 October 2008
Published 26 November 2008
Online at stacks.iop.org/JPhysA/42/015005

Abstract

Learning behavior of simple perceptrons is analyzed for a teacher–student scenario in which output labels are provided by a teacher network for a set of possibly correlated input patterns, and such that the teacher and student networks are of the same type. Our main concern is the effect of statistical correlations among the input patterns on learning performance. For this purpose, we extend to the teacher–student scenario a methodology for analyzing randomly labeled patterns recently developed in Shinzato and Kabashima 2008 *J. Phys. A: Math. Theor.* **41** 324013. This methodology is used for analyzing situations in which orthogonality of the input patterns is enhanced in order to optimize the learning performance.

PACS numbers: 02.50.−r, 84.35.+i

1. Introduction

Learning from examples is a fundamental technique for analyzing real-world data, and simple perceptrons are included in widely used devices for this purpose. In the past two decades, the structural similarity between statistical learning and statistical mechanics of disordered systems has been commonly recognized [1]. This similarity has promoted statistical mechanical analysis of perceptron learning [2–4]. Such research has led to the discovery of various types of learning behavior of perceptrons and the development of computationally feasible approximate algorithms that had not previously been known in conventional learning research [5–13].

Numerous studies have been published on perceptron learning. However, there still remain several research directions to explore. Learning from correlated patterns is a typical example. As a first step in this direction, the authors recently developed methodologies to analyze learning from randomly labeled patterns that are correlated in a certain manner on the basis of a formula involving rectangular random matrices [14, 15]. This paper is concerned with a second step; more precisely, we extend the methodologies developed for randomly labeled patterns to the cases of a teacher–student scenario in which output labels are provided by a teacher network, and both the teacher and student networks are of the same type.

In earlier studies, asymptotic behavior of learning curves and a critical pattern ratio of perfect learning in which the teacher network is completely identified from a reference data set of the same order as the network size have been assessed for continuous and discrete weights, respectively, for the case of independently and identically distributed (i.i.d.) patterns [6, 7]. Therefore, our main concern here is how these assessments are influenced by correlations among input patterns. For dealing with real-world data, various preprocessing methods are proposed in order to improve the performance of extracting relevant information from a given set of patterns. Orthogonalizing the patterns is a typical example of such methods. Hence, we particularly examine how the learning performance can be improved by enhancing orthogonality of the input patterns. Recent deeper understanding of the relations among learning, communication and information theories has suggested that a perceptron can be a useful building block for various coding schemes [16–18]. The analysis here may also be a useful guideline for developing efficient schemes to be used in information and communication engineering.

This paper is organized as follows. In the next section, we define the model that we shall investigate. In section 3, the main section of this manuscript, we will extend a scheme to handle correlated patterns, utilizing a formula involving rectangular random matrices, which was developed in [14, 15], to perceptron learning of a teacher–student scenario. In section 4, the extended scheme will be applied to several examples. The final section is devoted to a summary and future work.

2. Model definition

For an N-dimensional input pattern vector $\vec{x}$, a single layer perceptron of weight $\vec{w}$ of dimension N returns a binary output $y \in \{+1, -1\}$ given by

$$y = \mathrm{sgn}\left(\frac{1}{\sqrt{N}}\vec{x}^{\mathrm{T}}\vec{w}\right) = \begin{cases} 1, & N^{-1/2}\vec{x}^{\mathrm{T}}\vec{w} > 0, \\ -1, & \text{otherwise}, \end{cases} \tag{1}$$

where T denotes the matrix transpose and the prefactor $1/\sqrt{N}$ is introduced to keep relevant variables $O(1)$ as $N \to \infty$. Let us suppose a situation in which a student perceptron infers the weight vector of a teacher perceptron, $\vec{w}_0$, based on a given reference data set $\xi^p = \{(\vec{x}_1^{\mathrm{T}}, y_1), (\vec{x}_2^{\mathrm{T}}, y_2), \ldots, (\vec{x}_p^{\mathrm{T}}, y_p)\}$, where output labels are provided by the teacher as $y_\mu = \mathrm{sgn}(N^{-1/2}\vec{x}_\mu^{\mathrm{T}}\vec{w}_0)$ $(\mu = 1, 2, \ldots, p)$. The problem we consider here is how the inference accuracy depends on the pattern ratio $\alpha = p/N$ and correlations in the pattern matrix $X = N^{-1/2}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p)^{\mathrm{T}}$ as N and p tend to infinity, while keeping $\alpha = p/N$ finite.

As the basis for our analysis, we introduce a representation of the singular value decomposition

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

where $D = \mathrm{diag}(d_k)$ is a $p \times N$ diagonal matrix consisting of singular values $d_k (k = 1, 2, \ldots, \min(p, N))$, and U and V denote $p \times p$ and $N \times N$ orthogonal matrices, respectively. Linear algebra guarantees that any $p \times N$ matrices can be decomposed according to equation (2). The singular values are linked to the eigenvalues λ_k $(k = 1, 2, \ldots, N)$ of the correlation matrix $X^{\mathrm{T}}X$ via $\lambda_k = d_k^2$ $(k = 1, 2, \ldots, \min(p, N))$ and 0 otherwise, where

$\min(p, N)$ denotes the lesser value of p and N. Orthogonal matrices U and V constitute the right and left eigenbases of X, respectively; i.e., they are the eigenbases of XX^{T} and $X^{\mathrm{T}}X$.

To handle the correlations in X somewhat analytically, we assume hereafter that the following two properties hold for the pattern matrix X:

(i) The eigenvalue spectrum of the correlation matrix $X^{\mathrm{T}}X$, $\rho_{X^{\mathrm{T}}X}(\lambda) = N^{-1}\sum_{k=1}^{N}\delta(\lambda - \lambda_k)$, tends to a certain specific distribution $\rho(\lambda)$ in the limit as $N \to \infty$ for typical samples of X. Controlling $\rho(\lambda)$ allows us to characterize various second-order correlations in X.

(ii) U and V are independently generated from the uniform distributions of $p \times p$ and $N \times N$ orthogonal matrices (the Haar measures), respectively. This assumption makes it possible to characterize the correlations in X using only the eigenvalue spectrum $\rho(\lambda)$.

XX^{T} and $X^{\mathrm{T}}X$ generally represent correlation matrices among row and column vectors of X, respectively. In general, even if either of the eigenbases of XX^{T} or $X^{\mathrm{T}}X$ is given, it is difficult to infer that of the other. Therefore, we consider the assumption of independence of U and V as natural. The most crucial issue in the above setting may be the assumption of generation of U and V from the Haar measures. This might not be directly acceptable when U and/or V include components of large values, which indicates that U and/or V are an atypical sample from the Haar measures. However, such eigenbases usually provide us with information about certain regularities of the patterns (means, periodicities, etc), which can be utilized for decomposing the patterns into regular and (seemingly) irregular parts before the learning. In practice, techniques of machine learning are usually used for analyzing such irregular parts, since it is relatively easy to extract relevant information from the regular parts by human eyes in many cases. Therefore, the above assumptions may be reasonable when dealing with such irregular parts, for which characteristic features of U and V are reduced.

3. Analytical scheme

3.1. Expression for the average free energy

Given ξ^p, the volume of weight vectors that are compatible with ξ^p, which serves as the partition function of the current system, can be expressed as

$$Z(\xi^p) = \sum_{\vec{w}} P(\vec{w}) \prod_{\mu=1}^{p} \Theta\left(\frac{y_\mu}{\sqrt{N}}\sum_{k=1}^{N} x_{\mu k} w_k\right), \qquad (3)$$

where $P(\vec{w})$ represents the prior distribution of $\vec{w}$ and $\Theta(x) = 1$ for $x \geqslant 0$ and 0, otherwise. The conventional scheme of statistical mechanics of disordered systems indicates that typical properties of learning can be examined by assessing the average free energy

$$\Phi = -\frac{1}{N}[\log Z(\xi^p)]_{\xi^p}. \qquad (4)$$

Here, $[\cdots]_{\xi^p}$ represents an average taken over the reference data set $\xi^p = \{X, \vec{y}\}$ with respect to a distribution

$$P(\xi^p) = P(X)\sum_{\vec{w}_0} P(\vec{w}_0) \prod_{\mu=1}^{p} \Theta\left(\frac{y_\mu}{\sqrt{N}}\sum_{k=1}^{N} x_{\mu k} w_{0k}\right) = P(X)Z(\xi^p), \qquad (5)$$

where $P(X)$ denotes the distribution of the pattern matrix $X = UDV^{\mathrm{T}}$. This implies that equation (4) can be evaluated as

$$\Phi = -\lim_{n\to 1}\frac{\partial}{\partial n}\frac{1}{N}\log\left(\sum_{\xi^p} P(X)Z^n(\xi^p)\right). \qquad (6)$$

3.2. Replica analysis

Equation (6) can be evaluated by the replica method. For this, we first evaluate $\overline{Z^n(\xi^p)} = \sum_{\xi^p} P(X) Z^n(\xi^p)$ for $n = 1, 2, \ldots$ utilizing the expression

$$Z(\xi^p) = \sum_{\vec{w}} \prod_{k=1}^{N} P(w_k) \int \prod_{\mu=1}^{p} \mathrm{d}\Delta_\mu \Theta(y_\mu \Delta_\mu) \delta\left(\Delta_\mu - \frac{1}{\sqrt{N}} \sum_{k=1}^{N} x_{\mu k} w_k\right)$$
$$= \sum_{\vec{w}} \int \mathrm{d}\vec{u} \prod_{k=1}^{N} P(w_k) \times \prod_{\mu=1}^{p} \widetilde{\Theta}(y_\mu u_\mu) \times \mathrm{e}^{-\mathrm{i}\vec{u}^T X \vec{w}}, \tag{7}$$

where $\mathrm{i} = \sqrt{-1}$, $\widetilde{\Theta}(x) = (2\pi)^{-1} \int \mathrm{d}t\, \Theta(t)\, \mathrm{e}^{\mathrm{i}tx}$ and $\vec{u} = (u_1, u_2, \ldots, u_p)^{\mathrm{T}}$. We have assumed a factorizable prior $P(\vec{w}) = \prod_{k=1}^{N} P(w_k)$ for analytical tractability. Taking nth powers, for $n(= 1, 2, \ldots)$, equation (7) yields an expression

$$\exp\left[-\mathrm{i} \sum_{a=1}^{n} \vec{u}_a^{\mathrm{T}} X \vec{w}_a\right] = \exp\left[-\mathrm{i} \sum_{a=1}^{n} (U^{\mathrm{T}} \vec{u}_a)^{\mathrm{T}} D (V^{\mathrm{T}} \vec{w}_a)\right]. \tag{8}$$

For evaluating the average of this equation with respect to X, it is useful to note that for fixed sets of dynamical variables $\{\vec{u}_a\} = \{\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n\}$ and $\{\vec{w}_a\} = \{\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_n\}$, $\widetilde{\vec{u}}_a = U^{\mathrm{T}} \vec{u}_a$ and $\widetilde{\vec{w}}_a = V^{\mathrm{T}} \vec{w}_a$ behave as continuous random variables which satisfy the strict constraints

$$\frac{1}{N} \widetilde{\vec{w}}_a^{\mathrm{T}} \widetilde{\vec{w}}_b = \frac{1}{N} \vec{w}_a^{\mathrm{T}} \vec{w}_b = q_{ab}^{w}, \tag{9}$$

$$\frac{1}{p} \widetilde{\vec{u}}_a^{\mathrm{T}} \widetilde{\vec{u}}_b = \frac{1}{p} \vec{u}_a^{\mathrm{T}} \vec{u}_b = q_{ab}^{u}, \tag{10}$$

$(a, b = 1, 2, \ldots, n)$ when U and V are independently sampled from the Haar measures. This indicates that $\overline{Z^n(\xi^p)}$ can be evaluated by the saddle point method with respect to the macroscopic order parameters $\mathcal{Q}^w = \left(q_{ab}^w\right)$ and $\mathcal{Q}^u = \left(q_{ab}^u\right)$ in the limit $N, p \to \infty$, keeping $\alpha = p/N \sim O(1)$. Furthermore, due to the intrinsic permutation symmetry with respect to the replica indices $a = 1, 2, \ldots, n$, it is natural to assume that the relevant saddle point is replica symmetric (RS). This assumption can be expressed as

$$q_{ab}^{w} = \begin{cases} \chi_w + q_w, & (a = b), \\ q_w, & (a \neq b), \end{cases} \qquad q_{ab}^{u} = \begin{cases} \chi_u - q_u, & (a = b), \\ -q_u, & (a \neq b), \end{cases} \tag{11}$$

which yields

$$\frac{1}{N} \log\left(\int \mathcal{D}X P(X)\, \mathrm{e}^{-\mathrm{i}\sum_{a=1}^{n} \vec{u}_a^{\mathrm{T}} X \vec{w}_a}\right) = (n-1) F(\chi_w, \chi_u) + F(\chi_w + n q_w, \chi_u - n q_u), \tag{12}$$

for fixed sets of $\{\vec{w}_a\}$ and $\{\vec{u}_a\}$, where

$$F(x, y) = \mathop{\mathrm{Extr}}_{\Lambda_x, \Lambda_y} \left\{ -\frac{\alpha - 1}{2} \log \Lambda_y - \frac{1}{2} \langle \log(\Lambda_x \Lambda_y + \lambda) \rangle + \frac{\Lambda_x x}{2} + \frac{\alpha \Lambda_y y}{2} \right\}$$
$$- \frac{1}{2} \log x - \frac{\alpha}{2} \log y - \frac{\alpha + 1}{2}, \tag{13}$$

and $\int \mathcal{D}X$ denotes integration with respect to the pattern matrix X [14, 15, 19]. $\langle (\cdots) \rangle$ means an average of $(\cdots)$ with respect to $\rho(\lambda)$. $\mathrm{Extr}_x\{\cdots\}$ denotes the operation of extremization with respect to x, which corresponds to the saddle point evaluation of a certain complex integral and does not refer to maximization or minimization.

Equation (11) and assessment of the volumes of the dynamical variables yield a saddle point evaluation of $\overline{Z^n(\xi^p)}$ for $n = 1, 2, \ldots$. However, the functional form of $\overline{Z^n(\xi^p)}$ that we obtain can be defined for real values of n as well. Therefore, we analytically continue the expression from $n = 1, 2, \ldots$ to $n \in \mathbb{R}$ to evaluate equation (6). For $n \to 1$, the normalization constraint $\overline{Z(\xi^p)} = \sum_{\xi^p} P(X)Z(\xi^p) = \sum_{\xi^p} P(\xi^p) = 1$ implies that the relations

$$\chi_w + q_w = T_w = \sum_w P(w)w^2, \qquad \chi_u - q_u = 0, \tag{14}$$

$$\frac{\partial F(\chi_w + q_w, \chi_u - q_u)}{\partial \chi_w} = 0, \qquad \frac{\partial F(\chi_w + q_w, \chi_u - q_u)}{\partial \chi_u} = \frac{1}{2}\langle\lambda\rangle T_w, \tag{15}$$

must hold. These yield a formula for calculating the average free energy as

$$\Phi = -\mathop{\rm Extr}_{q_w, q_u} \{\mathcal{A}_{wu}(q_w, q_u) + \mathcal{A}_w(q_w) + \alpha\mathcal{A}_u(q_u)\}, \tag{16}$$

where

$$\mathcal{A}_{wu}(q_w, q_u) = F(T_w - q_w, q_u) + \frac{1}{2}\langle\lambda\rangle T_w q_u, \tag{17}$$

$$\mathcal{A}_w(q_w) = \mathop{\rm Extr}_{\widehat{q}_w} \left\{ -\frac{\widehat{q}_w q_w}{2} + \int Dz\mathcal{P}(z; \widehat{q}_w) \log \mathcal{P}(z; \widehat{q}_w) \right\}, \tag{18}$$

and

$$\mathcal{A}_u(q_u) = \mathop{\rm Extr}_{\widehat{q}_u} \left\{ -\frac{\widehat{q}_u q_u}{2} + 2\int DzH(\gamma z) \log H(\gamma z) \right\}, \tag{19}$$

given a particular eigenvalue spectrum $\rho(\lambda)$. Here, $Dz = \mathrm{d}z\, \mathrm{e}^{-z^2/2}/\sqrt{2\pi}$ represents the Gaussian measure, $H(u) = \int_u^{+\infty} Dz$, $\mathcal{P}(z; \widehat{q}_w) = \sum_w P(w) \exp[-\widehat{q}_w w^2/2 + \sqrt{\widehat{q}_w} zw]$ and $\gamma = \sqrt{\widehat{q}_u/(\langle\lambda\rangle T_w/\alpha - \widehat{q}_u)}$. Equations (16)–(19) contain the main results of this paper.

Three points are noteworthy here. First, q_w being determined by the saddle point condition of equation (16) physically means that there is a typical overlap $N^{-1}\vec{w}_0^{\rm T}\vec{w}$ between the teacher $\vec{w}_0$ and student $\vec{w}$ perceptrons after learning. Therefore the learning performance can be assessed by solving the saddle point problem of equation (16). In earlier studies for i.i.d. patterns, a generalization error, which represents the probability of wrongly predicting the output label for a novel example that is independently generated from an identical input distribution, was widely used for evaluating the performance of perceptron learning [2, 3]. However, the way of defining the generalization error is nontrivial in the current framework because given patterns are not independent but assumed to be correlated with one another. Therefore we do not employ such measures here. The second issue is about the formalism of the present analysis. In many earlier studies, the replica analysis is based on not equation (3) but on the Gardner volume $W(\xi^p) = \sum_{\vec{w}} \prod_{\mu=1}^p \Theta\big(y_\mu N^{-1/2} \sum_{k=1}^N x_{\mu k} w_k\big)$, for which the counterpart of equation (4) represents a typical entropy (per component) S of the posterior distribution $P(\vec{w}|\xi^p) = Z^{-1}(\xi^p)P(\vec{w}) \prod_{\mu=1}^p \Theta\big(y_\mu N^{-1/2} \sum_{k=1}^N x_{\mu k} w_k\big)$. Actually, the two formalisms are equivalent, since S and Φ are generally linked as

$$S = S_0 - \Phi, \tag{20}$$

where $S_0 = -N^{-1}\sum_{\vec{w}} P(\vec{w}) \log P(\vec{w})$ is the entropy (per component) of the prior distribution $P(\vec{w})$. Nevertheless, we dare to adopt the present formalism, which is useful for relating the learning problem to that of information and communication engineering. More precisely, equation (16) itself represents the mutual information (per component) between $\vec{w}$ and $\vec{y}$

for a given typical pattern matrix X, which measures the quantity of information about $\vec{w}$ that can typically be gained from output labels $\vec{y}$ when X is fixed. This correspondence can be directly generalized to the case of finite temperature where $\Theta(u)$ is replaced with $\widetilde{\Theta}(u) = [\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})\Theta(u)]/(1+\mathrm{e}^{-\beta})$ utilizing an inverse temperature $\beta > 0$, which corresponds to a binary symmetric communication channel [20]. Therefore, the current scheme is useful for characterizing potential capabilities of simple perceptrons when they are used for communication purposes. Finally, the assumption that the teacher and student networks are of the same type corresponds to the Nishimori condition known in spin glass research, which implies that the RS solution constructed above is expected to be correct [21–23]. Therefore, we do not proceed to the replica symmetry breaking (RSB) analysis here. Treatment in a more general setting, including the local stability analysis of the RS solution and the expression of the 1RSB free energy, can be found in [14, 15].

4. Examples

4.1. Independently and identically distributed patterns

In order to show consistency with the existing results, we first apply the scheme that we have developed here to the case of i.i.d. patterns, in which entries of the pattern matrix X are independently generated from an identical distribution with mean zero and variance $1/N$. In the current framework, this case is characterized by the Marčenko–Pastur distribution

$$\rho(\lambda) = [1-\alpha]^+\delta(\lambda) + \frac{\sqrt{[\lambda-\lambda_-]^+[\lambda_+-\lambda]^+}}{2\pi\lambda}, \tag{21}$$

where $\lambda_\pm = (1 \pm \sqrt{\alpha})^2$ and

$$[x]^+ = \begin{cases} x, & (x \geqslant 0), \\ 0, & (x < 0). \end{cases} \tag{22}$$

Plugging equation (21) into equation (13) yields

$$F(x, y) = -\frac{\alpha}{2}xy. \tag{23}$$

Applying this to equation (17) and utilizing the relation $\langle\lambda\rangle = \alpha$ which holds for equation (21) yields the result that $\mathcal{A}_{wu}(q_w, q_u) = \alpha q_w q_u/2$. This implies that $\widehat{q}_u = q_w$ at the extremum, where $\widehat{q}_u$ is the auxiliary variable in equation (19). These conditions mean that the average free energy can be calculated as

$$\Phi = -\mathop{\mathrm{Extr}}_{q_w, \tilde{q}_w} \left\{ \int Dz \log\left(\sum_w P(w)\,\mathrm{e}^{-\frac{1}{2}\widehat{q}_w w^2 + \sqrt{\widehat{q}_w} z w}\right) - \frac{1}{2}\widehat{q}_w q_w \right.$$
$$\left. + 2\alpha \int_{-\infty}^{\infty} Dz H\left(\sqrt{\frac{q_w}{T_w - q_w}} z\right) \log H\left(\sqrt{\frac{q_w}{T_w - q_w}} z\right) \right\}, \tag{24}$$

which is equivalent to the known expression for the free energy of the teacher–student scenario with i.i.d. patterns [10].

4.2. Asymptotic learning curve for spherical weights

The relation between a measure of learning performance and the amount of reference data p or the pattern ratio α is sometimes termed as a *learning curve*. In statistical learning theory, the asymptotic behavior of learning curves is frequently examined, which is, however, limited mostly to the cases of i.i.d. patterns [24–27]. We employ here the methodology that has been developed for the analysis of the asymptotic learning curve in order to investigate the effect of correlations in the pattern matrix. Investigations of this kind may be useful for *active learning* or *experimental design* contexts in which the pattern matrix can be designed to optimize learning performance [28–30].

As a representative example, let us consider the case of spherical weights $P(\vec{w}) \propto \delta(|\vec{w}|^2 - N)$, which implies that $T_w = 1$. For generality, we investigate a model for which the second-order correlations of the pattern matrix X are characterized by an eigenvalue spectrum

$$\rho(\lambda) = (1-\kappa)\delta(\lambda) + \kappa\widetilde{\rho}(\lambda), \tag{25}$$

where $\widetilde{\rho}(\lambda)$ is a distribution, the support of which is defined over a certain region of $\lambda > 0$. $0 \leqslant \kappa \leqslant 1$ is introduced to include the possibility of rank deficiency of the pattern matrix.

For this situation, we evaluate q_w, which serves as a performance measure representing the overlap between teacher and student perceptrons, for $\alpha \gg 1$ solving the saddle point problem of equation (16). For spherical weights, Λ_w, which is the counterpart of Λ_x in equation (13) for $x = T_w - q_w = 1 - q_w$, is always fixed to unity at the saddle point. This yields four coupled equations relevant to the calculation of q_w thus:

$$\widehat{q}_u = \frac{\langle\lambda\rangle}{\alpha} + \frac{2}{\alpha}\frac{\partial F(1-q_w, q_u)}{\partial q_u} = \frac{\langle\lambda\rangle}{\alpha} + \Lambda_u - \frac{1}{q_u}, \tag{26}$$

$$q_u = \left(1 - \frac{\kappa}{\alpha}\right)\frac{1}{\Lambda_u} + \frac{\kappa}{\alpha}\left\langle\frac{1}{\Lambda_u + \lambda}\right\rangle_{\widetilde{\rho}}, \tag{27}$$

$$q_u = \frac{\alpha}{\pi\langle\lambda\rangle\sqrt{1 - \alpha\widehat{q}_u/\langle\lambda\rangle}}\int Dz\frac{\mathrm{e}^{-\alpha\widehat{q}z^2/(2\langle\lambda\rangle)}}{H\left(\sqrt{\alpha\widehat{q}_u/\langle\lambda\rangle}z\right)}, \tag{28}$$

$$1 - q_w = (1-\kappa) + \kappa\left\langle\frac{\Lambda_u}{\Lambda_u + \lambda}\right\rangle_{\widetilde{\rho}}, \tag{29}$$

where $\langle\cdots\rangle_{\widetilde{\rho}}$ represents an average with respect to $\widetilde{\rho}(\lambda)$. For $\alpha \gg 1$, equations (27) and (28) yield asymptotic relations $q_u \simeq (1 - \kappa/\alpha)/\Lambda_u$ and $1 - \alpha\widehat{q}_u/\langle\lambda\rangle \simeq (c\alpha/\pi\langle\lambda\rangle)^2\big/q_u^2 \simeq (c\alpha/\pi\langle\lambda\rangle)^2\Lambda_u^2$, where $c = \int Dz\,\mathrm{e}^{-z^2/2}/H(z) \simeq 2.263$. Inserting these relations into equation (26) yields $\Lambda_u \simeq \kappa\langle\lambda\rangle(\pi/c)^2/\alpha^2 \simeq 1.926\kappa\langle\lambda\rangle/\alpha^2$. From this result and equation (29), we obtain the asymptotic learning curve

$$q_w \simeq \kappa - \frac{1.926\kappa^2\langle\lambda\rangle\langle\lambda^{-1}\rangle_{\widetilde{\rho}}}{\alpha^2} + O(\alpha^{-3}). \tag{30}$$

Two issues are noteworthy here. First, in the current model, κ denotes a fraction of the relevant dimensions that the pattern matrix X spans. Convergence $q_w \to \kappa$ as $\alpha \to \infty$ in equation (30) indicates that weights concerning the relevant dimensions are correctly identified, while no information is obtained from the irrelevant dimensions for perceptron learning. The rate of convergence scales as $O(\kappa^2\alpha^{-2}) = O((\kappa N)^2/p^2)$, which indicates that the irrelevant dimensions do not affect the learning performance of the relevant weights. This is in accordance with the existing results for singular statistical models in which some of the eigenvalues of the Fisher information matrix vanish, similar to the cases of equation (25) such that $0 \leqslant \kappa < 1$ [27]. Second, the inequality $\langle\lambda^{-1}\rangle_{\widetilde{\rho}} \geqslant \langle\lambda\rangle_{\widetilde{\rho}}^{-1} = \kappa\langle\lambda\rangle^{-1}$, which holds because λ is positive and $\langle\lambda\rangle = \kappa\langle\lambda\rangle_{\widetilde{\rho}}$ is satisfied by equation (25), implies that q_w is asymptotically bounded above:

$$q_w \lesssim \kappa - \frac{1.926\kappa^3}{\alpha^2} + O(\alpha^{-3}), \tag{31}$$

where equality holds when $\langle \lambda^{-1}\rangle_{\widetilde{\rho}} = \langle \lambda \rangle_{\widetilde{\rho}}^{-1}$ is satisfied. This property is asymptotically satisfied for the i.i.d. patterns since equation (21) yields $\langle \lambda^{-1}\rangle = (\alpha-1)^{-1} \simeq \alpha^{-1} = \langle \lambda \rangle^{-1}$ for $\alpha \gg 1$. In addition, $\kappa = 1$ holds for $\alpha \gg 1$ of the i.i.d. patterns, which maximizes the value of convergence κ to unity, reproducing the known asymptotic learning curve $\cos^{-1}(q_w)/\pi \simeq 0.625/\alpha$ [7]. Therefore, the i.i.d. patterns are asymptotically optimal for the leading order of the learning curve, although certain improvements can be gained for the next order by optimally designing the pattern matrix.

4.3. Presumably optimal performance in the non-asymptotic region

The above argument characterizes the optimal learning performance of simple perceptrons in the asymptotic region $\alpha \gg 1$. On the other hand, in information theory, it is known that when $\vec{w}$ is transformed into $\vec{y}$ via $\vec{y} = X\vec{w} + \vec{n}$, where $\vec{n}$ is a noise vector whose components are i.i.d. Gaussian random numbers, X which is characterized by the eigenvalue spectrum

$$\rho_{\rm POPT}(\lambda) = \begin{cases} (1-\alpha)\delta(\lambda) + \alpha\delta(\lambda-1), & (0 \leqslant \alpha \leqslant 1), \\ \delta(\lambda-\alpha), & (\alpha > 1), \end{cases} \tag{32}$$

maximizes the mutual information between $\vec{w}$ and $\vec{y}$, $\forall\alpha \geqslant 0$ under the condition that the power of each column in X is equally constrained to α [31]. For $0 \leqslant \alpha \leqslant 1$, this spectrum can be realized by composing X of $p = N\alpha$ randomly chosen orthonormal row vectors. On the other hand, for $\alpha > 1$, patterns of randomly constructed N orthogonal column vectors of dimension $p = N\alpha$ and length $\sqrt{\alpha}$ satisfy equation (32). A set of row vectors of such pattern matrices are sometimes referred to as Welch bound equality (WBE) sequences [32].

Intuitively, WBE sequences constitute a set of $p\,(> N)$ row vectors which are 'as orthogonal as possible' with one another under a constraint that lengths are typically fixed to unity in the N-dimensional space. It may be worthy of emphasizing that such row vectors are highly interdependent. In terms of the learning problem, this means that one has to design the optimal pattern set *before* starting the learning if the number of available patterns is fixed, implying that the optimal learning performance cannot be achieved in the manners of *online* learning [3].

One scheme to generate WBE sequences for large p and N is as follows: first, we prepare a $p \times N$ matrix G by independently generating each element from the standard normal distribution $\mathcal{N}(0, 1)$. Applying the singular value decomposition to G yields an expression $G = UDV^{\rm T}$, computational cost for which is $O(N^3)$ when $p/N = \alpha \sim O(1)$. In this expression, we replace D with a $p \times N$ matrix $\widetilde{D} = (\sqrt{\alpha}\delta_{ij})$. A set of row vectors of the resulting matrix $X = U\widetilde{D}V^{\rm T}$, which can be evaluated in an $O(N^3)$ computational time, serves as a sample of WBE sequences.

Note that these sequences achieve the upper bound of equation (31) in the asymptotic region, since $\langle \lambda^{-1}\rangle = \langle \lambda \rangle^{-1}$ holds. This and the formal similarity between the channel problem and perceptron learning imply that pattern matrices characterized by equation (32) may maximize the learning performance of perceptrons, in terms of overlap q_w or mutual information Φ, for $\forall\alpha \geqslant 0$ as well, although it is apparent that for $\forall s > 0$ an arbitrary rescaled distribution $\widetilde{\rho}(\lambda) = s^{-1}\rho_{\rm POPT}(\lambda/s)$ yields the same performance as that for $\rho_{\rm POPT}(\lambda)$ in the case of perceptron learning, since only the signs of inner products between the patterns and the weight vector are relevant. Unfortunately, q_w and Φ depend on the eigenvalue spectrum in a complicated manner, and finding the optimal spectrum as a solution to a certain variational problem as shown in [31] is nontrivial for the current system. Therefore, as the final example, we analyze the case of equation (32), utilizing the methodology of equation (16) in the

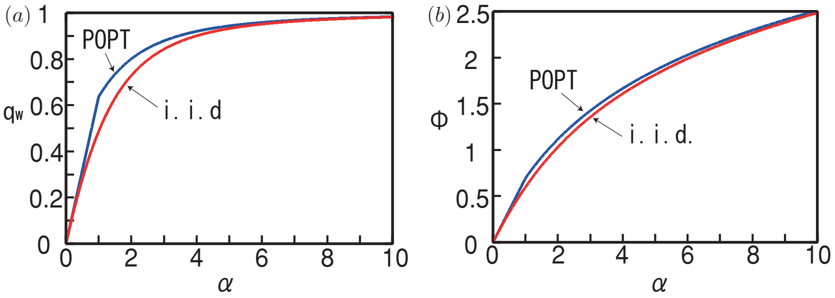

Figure 1. Performance curves for spherical weights. (*a*) q_w versus α. (*b*) Φ versus α. 'POPT' and 'i.i.d.' denote data for the presumed optimal and i.i.d. patterns, respectively. Similarly for other figures.

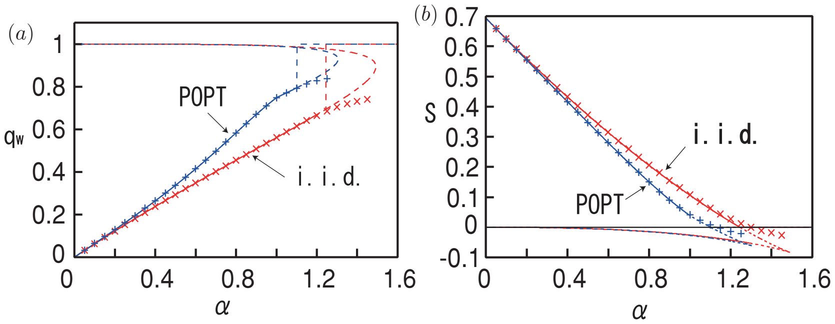

Figure 2. (*a*) q_w versus α. (*b*) S versus α. The curves represent the theoretical prediction evaluated by the replica method. The markers are obtained from 10^4 experiments for $N = 100$ systems utilizing a Thouless–Anderson–Palmer type mean field method [15].

non-asymptotic region of $\alpha \sim O(1)$ and compare its learning performance to that of the i.i.d. patterns to probe optimality.

For spherical weights, the saddle point problem of equation (16) can be analytically solved for equation (32) for $0 \leqslant \alpha \leqslant 1$, yielding the solution $q_w = 2\alpha/\pi$, $q_u = 2/\pi$, which in turn implies that

$$\Phi = \alpha \log 2. \tag{33}$$

For $\alpha > 1$, analytical construction of the solution is difficult, and we resorted to a numerical method. Figures 1(*a*) and (*b*) show a comparison of the learning performance between the (presumably optimal) case of equation (32) and the i.i.d. patterns of equation (21). For both the teacher–student overlap q_w (figure 1(*a*)) and mutual information (free energy) Φ (figure 1(*b*)), equation (32) results in a better learning performance than that of the i.i.d. patterns over the entire region of $\alpha \geqslant 0$.

As another representative learning model, we examined the case of binary (Ising) weights, the results of which are shown in figures 2(*a*) and (*b*). For the i.i.d. patterns, it is known that simple perceptrons with binary weights exhibit perfect learning at $\alpha = \alpha_c^{\text{i.i.d.}} \simeq 1.245$, completely identifying the teacher network [7]. Such behavior is also observed for the case of equation (32) at a certain critical ratio α_c^{POPT}, which is characterized by the vanishing entropy condition $S = S_0 - \Phi = \log 2 - \Phi = 0$. Figure 2(*a*) yields $\alpha_c^{\text{POPT}} \simeq 1.101$ for equation (32), implying a better learning performance than that of the i.i.d. patterns. In terms

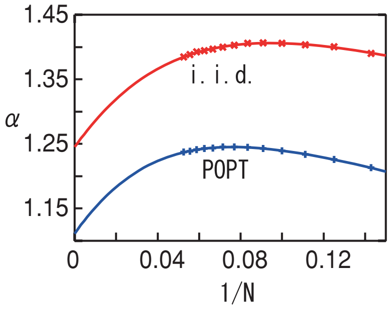

Figure 3. Critical ratio of perfect learning estimated by exhaustive search experiments. Data (markers) are obtained by 10^6 experiments for systems of $N = 6, \ldots, 19$. Error bars are omitted because they are smaller than the markers. We fitted a fourth-degree polynomial with respect to $1/N$, which is supported by a model selection scheme based on the leave-one-out cross-validation, for assessing the values as $N \to \infty$. Assessed values are $\lim_{N\to\infty} \alpha_c^{\mathrm{POPT}}(N) \simeq 1.111$ and $\lim_{N\to\infty} \alpha_c^{\mathrm{i.i.d.}}(N) \simeq 1.245$ for the presumably optimal and i.i.d. patterns, respectively. These are in very close agreement with the theoretical predictions $\alpha_c^{\mathrm{POPT}} \simeq 1.101$ and $\alpha_c^{\mathrm{i.i.d.}} \simeq 1.245$.

of q_w, patterns of equation (32) are also superior to the i.i.d. patterns (figure 2(*b*)). In both figures 2(*a*) and (*b*), numerical data for $N = 100$ systems obtained from 10^4 experiments based on a Thouless–Anderson–Palmer (TAP) type mean field method [15, 33], which offers not a sample of actual compatible weight vectors but approximate marginal distributions for their entries, is in very close agreement with the saddle point solutions of equation (16) (curves), which justifies the methodology based on equation (16). Note that the correct solutions for $\alpha > \alpha_c^{\mathrm{i.i.d.}}$ and $\alpha > \alpha_c^{\mathrm{POPT}}$ are the branches of perfect learning which are characterized by $q_w = 1$ and $S = 0$. Figures 2(*a*) and (*b*) indicate that the TAP method cannot correctly follow the transitions to the perfect learning branches due to the limitation of the approximation accuracy although a certain unphysical solution is found for each sample for $\alpha > \alpha_c^{\mathrm{i.i.d.}}$ and $\alpha > \alpha_c^{\mathrm{POPT}}$ as well.

The superiority of equation (32) is also confirmed by experimental assessment of α_c^{POPT}. Figure 3 shows the result of exhaustive search experiments for small systems ($6 \leqslant N \leqslant 19$). In order to characterize the critical ratio for finite systems, for each pair of N and p, we estimated the probability $r(N, p)$ that at least one weight vector $\vec{w}$ that differs from the teacher vector $\vec{w}_0$ is completely compatible with a given reference data set ξ^p, utilizing 10^6 experiments. For each N, the critical ratio is defined as $\alpha_c^{\mathrm{POPT}}(N) = N^{-1} \sum_{p=1}^{p_{\max}} r(N, p)$, where $p_{\max}$ is a sufficiently large threshold value to truncate the summation. We set $p_{\max} = 4N$. $\alpha_c^{\mathrm{POPT}}(N)$ is expected to converge to α_c^{POPT} as N tends to infinity. In figure 3, the data plotted versus $1/N$ are asymmetric on either side of a peak, which implies that it is necessary to use a higher order polynomial for estimating $\lim_{N\to\infty} \alpha_c^{\mathrm{POPT}}(N)$ by extrapolation. Therefore, we fitted a fourth-degree polynomial, which is supported by minimization of the leave-one-out cross-validation error (table 1). The value $\lim_{N\to\infty} \alpha_c^{\mathrm{POPT}}(N) \simeq 1.111$ assessed by extrapolation agrees closely with the theoretical estimate $\alpha_c^{\mathrm{POPT}} \simeq 1.101$. This is considerably smaller than the counterpart of the i.i.d. patterns, $\lim_{N\to\infty} \alpha_c^{\mathrm{i.i.d.}}(N) \simeq 1.245$, the theoretical value of which is $\alpha_c^{\mathrm{i.i.d.}} \simeq 1.245$, indicating the superiority of equation (32).

In conclusion, the above analyses for spherical and binary weights indicate that the eigenvalue spectrum of equation (32) always yields a better learning performance than that of the i.i.d. patterns. This lends some support to our conjecture that equation (32) achieves optimal learning performance for perceptrons operating under a fixed power constraint on the pattern matrix.

Table 1. Comparison among polynomial fittings. The first column represents the degree of polynomials. The second and third columns show leave-one-out cross validation error per data point and the value of $\alpha_c^{\mathrm{i.i.d.}}$ estimated by extrapolation for i.i.d data shown in figure 3, respectively. The fourth and fifth columns from the left show those for the POPT data. For both patterns, these data support the fitting by the fourth-degree polynomial.

Degree	LOOE(i.i.d.)	$\alpha_c^{\mathrm{i.i.d.}}$	LOOE(POPT)	α_c^{POPT}
1	0.004 076	1.414 710	0.001 321	1.268 553
2	0.000 256	1.335 408	0.000 137	1.205 526
3	0.000 262	1.295 237	0.000 048	1.158 679
4	0.000 103	1.245 047	0.000 001	1.111 172
5	0.000 145	1.282 783	0.000 158	1.118 968
6	0.007 534	1.251 661	0.000 290	1.275 169
7	4.164 808	1.300 419	0.000 053	1.000 935
8	2.414 898	−6.608 969	11.556 052	1.329 810
9	69.426 278	−2.924 355	3.854 332	0.552 925

5. Summary

In summary, we have investigated the learning performance of simple perceptrons extending a methodology for handling correlated patterns developed in [14, 15] to a teacher–student scenario. The scheme allows us to characterize various second-order correlations among the input patterns by an eigenvalue spectrum of the cross-correlation matrix under an assumption that the right and left eigenbases of the pattern matrix are independently generated from the Haar measure. Using this characterization, we have offered a general formula that relates the eigenvalue spectrum to the average free energy, which, in the current context, is a measure of the mutual information between the weight vector and output labels, given a typical pattern matrix. The formula is used to examine cases for which column or row vectors in the pattern matrix are orthogonalized under a fixed power constraint, the learning performance of which is optimal for the asymptotic region and presumed to be optimal in general. Results from numerical experiments based on a Thouless–Anderson–Palmer type mean field method and exhaustive search examinations for small systems are in agreement with theoretical predictions obtained from the formula.

A mathematical proof of the optimality of the eigenvalue spectrum (32) and applications of the current scheme to various problems in learning and communication are promising future research directions.

Acknowledgments

The authors thank A Pagnani, M Weigt and S Watanabe for several exhaustive discussions and fruitful comments. This work was partly supported by the JSPS Global COE program ‘Computationism as a Foundation for the Sciences’ (TS and YK) and Grants-in-Aid MEXT, Japan, No. 18079006 (YK).

References

[1] Levin E, Tishby N and Solla S A 1990 *Proc. IEEE* **78** 1568
[2] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[3] Engel A and van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)

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

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

[6] Györgyi G 1990 *Phy. Rev.* A **41** 7097

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

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

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

[10] 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

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

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

[13] Braunstein A and Zecchina R 2006 *Phys. Rev. Lett.* **96** 030201

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

[15] Shinzato T and Kabashima Y 2008 *J. Phys. A. Math. Theor.* **41** 324013

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

[17] Kinzel W and Kanter I 2002 *Proc. ICONIP'02* vol 3 p 1351

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

[19] Kabashima Y 2008 An integral formula for large random rectangular matrices and its application to analysis of linear vector channels *Proc. 1st Workshop on Physics-Inspired Paradigms in Wireless Communications and Networks* (Berlin, Germany) arXiv:0802.1372

[20] Cover T M and Thomas J A 2001 *Elements of Information Theory* (New York: Wiley)

[21] Nishimori H 1981 *Prog. Theor. Phys.* **66** 1169

[22] Nishimori H 1993 *J. Phys. Soc. Jpn.* **62** 2973

[23] Nishimori H and Sherrington D 2001 *Disordered and Complex Systems* ed P Sollich, A C C Coolen, L P Hughston and R F Streater (New York: AIP) p 67

[24] Baum E B and Haussler D 1990 *Neural Comput.* **1** 151

[25] Amari S, Fujita N and S Shinomoto 1992 *Neural Comput.* **4** 605

[26] Murata N, Yoshizawa S and Amari S 1994 *IEEE Trans. Neural Net.* **5** 865

[27] Watanabe S 2001 *Neural Comput.* **13** 899

[28] Fukumizu K 1996 *Adv. Neural Inf. Process. Syst.* **8** 312

[29] Sollich P 1994 *Phys. Rev.* E **49** 4637

[30] Seeger M W 2008 *J. Mach. Learn. Res.* **9** 759

[31] Kitagawa K and Tanaka T 2008 *Proc. 2008 IEEE Int. Symp. Information Theory* (Tronto, Canada) p 1373

[32] Welch L R 1974 *IEEE Trans. Inf. Theor.* **IT-20** 397

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