Neural Networks 102 (2018) 21–26

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Non-monotonic convergence of online learning algorithms for perceptrons with noisy teacher

Kazushi Ikeda [a,*], Arata Honda [a,1], Hiroaki Hanzawa [a,2], Seiji Miyoshi [b]

[a] *Nara Institute of Science and Technology, Ikoma, Nara, Japan*
[b] *Kansai University, Suita, Osaka, Japan*

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form 29 November 2017
Accepted 9 February 2018
Available online 21 February 2018

Keywords:
Learning curve
Perceptron
Online learning
Statistical mechanics
Asymptotic analysis

ABSTRACT

Learning curves of simple perceptron were derived here. The learning curve of the perceptron learning with noisy teacher was shown to be non-monotonic, which has never appeared even though the learning curves have been analyzed for half a century. In this paper, we showed how this phenomenon occurs by analyzing the asymptotic property of the perceptron learning using a method in systems science, that is, calculating the eigenvalues of the system matrix and the corresponding eigenvectors. We also analyzed the AdaTron learning and the Hebbian learning in the same way and found that the learning curve of the AdaTron learning is non-monotonic whereas that of the Hebbian learning is monotonic.

1. Introduction

Statistical mechanical methods can apply to problems in information science such as neural networks (Nishimori, 2001), communication theory (Tanaka, 2002), and adaptive filters (Miyoshi & Kajikawa, 2013). One successful application is the analyses of the perceptron learning algorithm (Biehl & Schwarze, 1992; Rosenblatt, 1961) and its variations (Hara & Okada, 2004; Inoue & Nishimori, 1997; Miyoshi, Hara, & Okada, 2005; Miyoshi & Okada, 2006a, b; Uezu, Miyoshi, Izuo, & Okada, 2007).

The perceptron learning is an online learning algorithm where the student updates its weight vector of a linear dichotomy according to the teacher's signal (Rosenblatt, 1961). Biehl and Schwarze (1992) introduced the statistical mechanics to the analysis of the perceptron learning and Inoue and Nishimori (1997) applied the method to the AdaTron learning in unlearnable cases. Hara and Okada (2004) discussed the perceptron learning with a margin and Miyoshi and his colleagues extended the analysis to the ensemble learning and/or noisy cases (Miyoshi et al., 2005; Miyoshi & Okada, 2006a, b; Uezu et al., 2007).

In this paper, we consider the case where the teacher has noise in its output while the student does not. In this case, the learning curve, which is defined as the average prediction error, is not monotonically decreasing but has an overshoot, differently from other cases analyzed so far (Ikeda, Hanzawa, & Miyoshi, 2013). Although an analysis for this problem was partially given by some of the authors (Ikeda et al., 2013), some part was given not theoretically but numerically.

This paper gives a theoretically rigorous and complete analysis above. In addition, we extend the analysis to other online algorithms for perceptrons, that is, the AdaTron learning and the Hebbian learning. Our analysis consisted of three steps. In the first step, we applied the statistical mechanical method to our problem, i.e., we introduced three order parameters assuming the thermodynamic limit, and derived a system of differential equations for the three algorithms. In the second step, we calculated the ensemble averages that appeared in the differential equations for each algorithm using Gaussian approximations. Note this had not been derived analytically yet in Ikeda et al. (2013). In the last step, we applied an asymptotic analysis to our dynamical system, i.e., we linearized the equations around their convergence point and analyzed their behaviors by the eigenvalues and eigenvectors of the state-transition matrix a.k.a. the system matrix. The three steps elucidated how and why the overshoot phenomenon occurs.

The remainder of this paper is organized as follows. Section 2 formulates the problem we treated. Sections 3–5 are devoted to the three steps, that is, statistical mechanical analysis, the calculation of the ensemble averages and the asymptotic analysis of the system, respectively. We conclude the paper in Section 6.

[*] Corresponding author.
E-mail addresses: kazushi@is.naist.jp (K. Ikeda), arata.honda@excite.jp (A. Honda), h.hanzawax68@gmail.com (H. Hanzawa), miyoshi@kansai-u.ac.jp (S. Miyoshi).

[1] A. Honda is currently with Excite.

[2] H. Hanzawa is currently with Yahoo! Japan.

https://doi.org/10.1016/j.neunet.2018.02.009
0893-6080/© 2018 Elsevier Ltd. All rights reserved.

2. Problem statement

Two linear perceptrons are treated: a teacher and a student, whose connection weights are $B = (B_1, \ldots, B_N) \in R^N$ and $J = (J_1, \ldots, J_N) \in R^N$, respectively. The initial value of each of the components is independently drawn from the normal distribution $N(0, 1)$, that is,

$$\langle B_i \rangle = 0, \qquad \left\langle (B_i)^2 \right\rangle = 1, \tag{1}$$

$$\langle J_i \rangle = 0, \qquad \left\langle (J_i)^2 \right\rangle = 1, \tag{2}$$

where $\langle \cdot \rangle$ denotes the mean of $\cdot$, as was in Nishimori (2001).

The mth input vector $x^m = (x_1^m, \ldots, x_N^m) \in R^N$ is independently drawn from the N-dimensional normal distribution $N(0, I/N)$ and the corresponding output y^m of the teacher is produced as

$$y^m = \mathrm{sgn}(v_m), \qquad v_m = B \cdot x^m + n_B^m, \tag{3}$$

where n_B^m is an observation noise obeying $N(0, \sigma_B^2)$.

The learning rule is either the standard perceptron learning (Biehl & Schwarze, 1992; Nishimori, 2001; Rosenblatt, 1961), the AdaTron learning (Nishimori, 2001), or the Hebbian learning (Nishimori, 2001). In the perceptron learning, given the mth input vector x^m, the student updates its weight vector J^m as

$$J^{m+1} = J^m + f^m x^m, \tag{4}$$

$$f^m = \eta y^m \Theta(-y^m J^m \cdot x^m), \tag{5}$$

where η is a learning coefficient and $\Theta(\cdot)$ is the Heaviside function,

$$\Theta(t) = \begin{cases} 1 & t \geq 0, \\ 0 & t < 0. \end{cases} \tag{6}$$

This means that it updates its weight vector when its output does not coincide with the teacher's one.

In the AdaTron learning and the Hebbian learning, the update functions f^m are changed to

$$f^m = \eta y^m J^m x^m \Theta(-y^m J^m \cdot x^m), \tag{7}$$

$$f^m = \eta y^m, \tag{8}$$

respectively.

As the learning proceeds and m increases, the weight vector, J^m, of the student approaches the teacher's one, B. The problem of learning curves is to evaluate how fast the covariance coefficient between J^m and B,

$$R^m = \frac{B \cdot J^m}{\|B\| \, \|J^m\|}, \tag{9}$$

approaches unity in noiseless cases and another value in noisy cases (0.70 in Fig. 1, for example).

3. Statistical mechanical analysis

3.1. Theory

The method to derive the learning curve of the student is essentially the same as Nishimori (2001). We introduce auxiliary order parameters, R^m in (9) and

$$l^m = \|J^m\| / \sqrt{N}, \tag{10}$$

and consider the thermodynamic limit, $N, m \to \infty$ and $m/N = t$. Then,

$$\|B\| = \sqrt{N}, \qquad \|J^0\| = \sqrt{N}, \qquad \|x^m\| = 1, \tag{11}$$

hold and the random vector of the inner products, (u, v), where

$$v^m = B \cdot x^m, \qquad u^m l^m = J^m \cdot x^m \tag{12}$$

obeys the two-dimensional normal distribution $N(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 1 & R^m \\ R^m & 1 \end{pmatrix}. \tag{13}$$

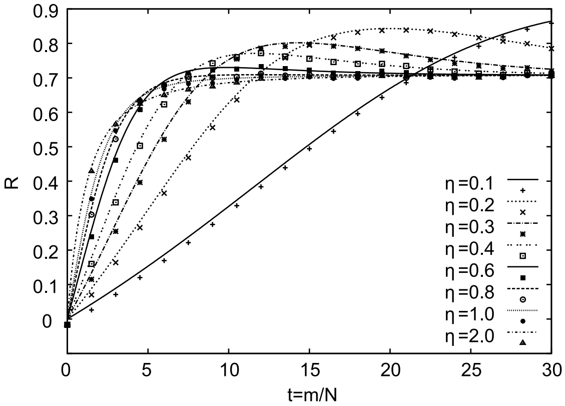

Fig. 1. Dynamics of R. $\sigma_B^2 = 1.0$, $N = 10^4$, $\eta = 0.1, \ldots, 2.0$, plots: experiments, lines: theory (modified from Ikeda et al., 2013).

By self-averaging and omitting the step index m in (4) hereafter, we get the simultaneous differential equations of the order parameters,

$$\dot{l} = \langle fu \rangle + \frac{\langle f^2 \rangle}{2l}, \tag{14}$$

$$\dot{R} = \frac{\langle fv \rangle - \langle fu \rangle R}{l} - \frac{R}{2l^2} \langle f^2 \rangle, \tag{15}$$

where $\langle \cdot \rangle$ expresses the average over (u, v) and $n_B \sim N(0, \sigma_B^2)$ (Nishimori, 2001).

3.2. Experiments

To confirm the validity of the theory above, we conducted some computer simulations of the perceptron learning under the condition in Section 2. The experimental values of R coincided well with the theoretical values for any learning coefficient, η (Fig. 1).

As a result of the experiments, the value of R converged to 0.70 for any η due to the noise on the teacher's output. One notable property was that R was not monotonically increasing but had an overshoot. This overshoot phenomenon does not occur in other cases analyzed so far (Hara & Okada, 2004; Inoue & Nishimori, 1997; Miyoshi et al., 2005; Miyoshi & Okada, 2006a; Nishimori, 2001).

A quantitative analysis of this phenomenon is given using an asymptotic dynamical system theory in Section 5.

4. Calculation of ensemble averages

The ensemble averages $\langle fv \rangle$, $\langle fu \rangle$ and $\langle f^2 \rangle$ in (14) and (15) are difficult to calculate analytically, in general. In fact, we calculated those for the perceptron learning numerically (Ikeda et al., 2013). However, we theoretically derived the ensemble averages for the perceptron learning. In addition, we also calculated those for the AdaTron learning and the Hebbian learning, which will be given below.

The ensemble averages $\langle fv \rangle$, $\langle fu \rangle$ and $\langle f^2 \rangle$ for the perceptron learning are expressed as

$$\begin{aligned} \langle fv \rangle &= \langle \eta \Theta(-u(v + n_B)) \mathrm{sgn}(v + n_B) v \rangle \\ &= \eta \int_{-\infty}^{\infty} dn_B \int_{-n_B}^{\infty} dv \int_{-\infty}^{0} du P(u, v) P(n_B) v \end{aligned}$$

$$
-\ \eta \int_{-\infty}^{\infty} dn_B \int_{-\infty}^{-n_B} dv \int_{0}^{\infty} du P(u,v)P(n_B)v, \tag{16}
$$

$$
\begin{aligned}
\langle fu\rangle &= \langle \eta\Theta(-u(v+n_B))\mathrm{sgn}(v+n_B)u\rangle \\
&= \eta \int_{-\infty}^{\infty} dn_B \int_{-n_B}^{\infty} dv \int_{-\infty}^{0} du P(u,v)P(n_B)u \\
&\quad - \eta \int_{-\infty}^{\infty} dn_B \int_{-\infty}^{-n_B} dv \int_{0}^{\infty} du P(u,v)P(n_B)u,
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
\left\langle f^2\right\rangle &= \left\langle \eta^2\Theta(-u(v+n_B))\right\rangle \\
&= \eta^2 \int_{-\infty}^{\infty} dn_B \int_{-n_B}^{\infty} dv \int_{-\infty}^{0} du P(u,v)P(n_B) \\
&\quad + \eta^2 \int_{-\infty}^{\infty} dn_B \int_{-\infty}^{-n_B} dv \int_{0}^{\infty} du P(u,v)P(n_B),
\end{aligned} \tag{18}
$$

respectively, where $P(u,v)$ and $P(n_B)$ are Gaussian distributions and $\mathrm{erf}\,(x)$ and $\mathrm{erfc}\,(x)$ are defined as

$$
\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) dt, \tag{19}
$$

$$
\mathrm{erfc}\,(x) = 1 - \mathrm{erf}\,(x)\,, \tag{20}
$$

respectively (Ikeda et al., 2013). Then, (16) is analytically calculated as

$$
\begin{aligned}
\langle fv\rangle &= \frac{\eta}{2\pi\sqrt{1-R^2}} \frac{1}{\sqrt{2\pi\sigma_B^2}} \int_{-\infty}^{\infty} dn_B \int_{-n_B}^{\infty} dv \int_{-\infty}^{0} du \\
&\quad v \exp\left(-\frac{u^2+v^2-2Ruv}{2\left(1-R^2\right)}\right) \exp\left(-\frac{n_B^2}{2\sigma_B^2}\right) \\
&\quad - \frac{\eta}{2\pi\sqrt{1-R^2}} \frac{1}{\sqrt{2\pi\sigma_B^2}} \int_{-\infty}^{\infty} dn_B \int_{-\infty}^{-n_B} dv \int_{0}^{\infty} du \\
&\quad v \exp\left(-\frac{u^2+v^2-2Ruv}{2\left(1-R^2\right)}\right) \exp\left(-\frac{n_B^2}{2\sigma_B^2}\right)
\end{aligned} \tag{21}
$$

$$
\begin{aligned}
&= \frac{\eta}{2\pi\sigma_B} \int_{-\infty}^{\infty} dn_B \\
&\quad \exp\left(-\frac{n_B^2\left(1+\sigma_B^2\right)}{2\sigma_B^2}\right) \mathrm{erfc}\left(-\frac{Rn_B}{\sqrt{2\left(1-R^2\right)}}\right) \\
&\quad - \frac{\eta R}{2\pi\sigma_B} \int_{-\infty}^{\infty} dn_B \\
&\quad \exp\left(-\frac{n_B^2}{2\sigma_B^2}\right) \mathrm{erfc}\left(-\frac{n_B}{\sqrt{2\left(1-R^2\right)}}\right)
\end{aligned} \tag{22}
$$

$$
= -\frac{\eta}{\sqrt{2\pi}}R + \frac{\eta}{\sqrt{2\pi\left(1+\sigma_B^2\right)}} \tag{23}
$$

using $\mathrm{erf}\,(-\infty) = -1$, $\mathrm{erf}\,(\infty) = 1$, $\mathrm{erfc}\,(-\infty) = 2$ and $\mathrm{erfc}\,(\infty) = 0$. In the same way, (17) is calculated as

$$
\langle fu\rangle = \frac{\eta}{\sqrt{2\pi\left(1+\sigma_B^2\right)}}R - \frac{\eta}{\sqrt{2\pi}}, \tag{24}
$$

which is also a linear function of R. However, (18) cannot analytically be calculated but can be rewritten as

$$
\begin{aligned}
\left\langle f^2\right\rangle &= \frac{\eta^2}{2} + \frac{\eta^2}{2\pi\sigma_B} \int_{-\infty}^{0} du \int_{-\infty}^{\infty} dn_B \\
&\quad \exp\left(-\frac{u^2}{2}\right) \mathrm{erf}\left(\frac{n_B+Ru}{\sqrt{2\left(1-R^2\right)}}\right) \exp\left(-\frac{n_B^2}{2\sigma_B^2}\right),
\end{aligned} \tag{25}
$$

which is approximately calculated using the Taylor expansion of $\mathrm{erfc}\,(\cdot)$ with respect to u around zero up to the ninth order, that is,

$$
\begin{aligned}
\left\langle f^2\right\rangle &\approx \frac{\eta^2}{2} - \frac{\eta^2 R}{\pi\left(1-R^2+\sigma_B^2\right)^{1/2}} \\
&\quad + \frac{\eta^2 R^3}{3\pi\left(1-R^2+\sigma_B^2\right)^{3/2}} - \frac{\eta^2 R^5}{5\pi\left(1-R^2+\sigma_B^2\right)^{5/2}} \\
&\quad + \frac{\eta^2 R^7}{7\pi\left(1-R^2+\sigma_B^2\right)^{7/2}} - \frac{\eta^2 R^9}{9\pi\left(1-R^2+\sigma_B^2\right)^{9/2}} \\
&= \frac{\eta^2}{2} + \sum_{n=1}^{5} \frac{(-1)^n\eta^2R^{2n-1}}{(2n-1)\pi\left(1-R^2+\sigma_B^2\right)^{(2n-1)/2}}.
\end{aligned} \tag{26}
$$

In the same way, the ensemble averages for the AdaTron learning and those for the Hebbian learning are calculated as

$$
\langle fv\rangle = \langle -\eta u\Theta(-u(v+n_B))v\rangle \tag{27}
$$

$$
\begin{aligned}
&= \frac{\eta}{\pi} \frac{(1-R^2)^2}{\sqrt{1-R^2+\sigma_B^2}} \\
&\quad + \frac{\eta R^2}{\pi}\left(\frac{\sigma_B^2}{1+\sigma_B^2}\right) \frac{\left(1-R^2\right)}{\sqrt{1-R^2+\sigma_B^2}} + R\,\langle fu\rangle\,,
\end{aligned} \tag{28}
$$

$$
\begin{aligned}
\langle fu\rangle &\approx -\frac{\eta}{2} - \sum_{n=1}^{5} \frac{(-1)^n\eta R^{2n-1}}{(2n-1)\pi(1-R^2+\sigma_B^2)^{\frac{2n-1}{2}}} \\
&\quad + \frac{\eta}{\pi} \frac{R(1-R^2)}{\sqrt{1-R^2+\sigma_B^2}} \\
&\quad + \frac{\eta R^3}{\pi}\left(\frac{\sigma_B^2}{1+\sigma_B^2}\right) \frac{\left(1-R^2\right)}{\sqrt{1-R^2+\sigma_B^2}},
\end{aligned} \tag{29}
$$

$$
\left\langle f^2\right\rangle = -\eta\,\langle fu\rangle\,, \tag{30}
$$

and

$$
\langle fv\rangle = \langle \eta\,\mathrm{sgn}(v+n_B)v\rangle
$$

$$
= \frac{2\eta}{\sqrt{2\pi(1+\sigma_B^2)}}, \tag{31}
$$

$$
\langle fu\rangle = \frac{2\eta}{\sqrt{2\pi(1+\sigma_B^2)}}R, \tag{32}
$$

$$
\left\langle f^2\right\rangle = \eta^2, \tag{33}
$$

respectively.

In order to confirm the validity of the derived equations for the perceptron learning, we compared the experimental values of $\langle fv\rangle$ and $\langle fu\rangle$ with the theoretical values when $\sigma_B^2 = 1$ for the perceptron learning (Fig. 2). Note the experimental values were used in Ikeda et al. (2013). In addition, we confirmed the validity of the other equations for AdaTron learning and the Hebbian learning in the same way.

5. Asymptotic analysis of dynamical system

To see how the nonmonotonicity of the learning curves in Fig. 1 appears, we analyzed the behaviors of the nonlinear differential equations (14) and (15) for the perceptron learning when $t \to \infty$. Later, the same analysis method was applied to the AdaTron learning and the Hebbian learning.

In the following, we set $\sigma_B^2 = 1$ and calculate the values numerically because the ensemble average of f^2 for the perceptron

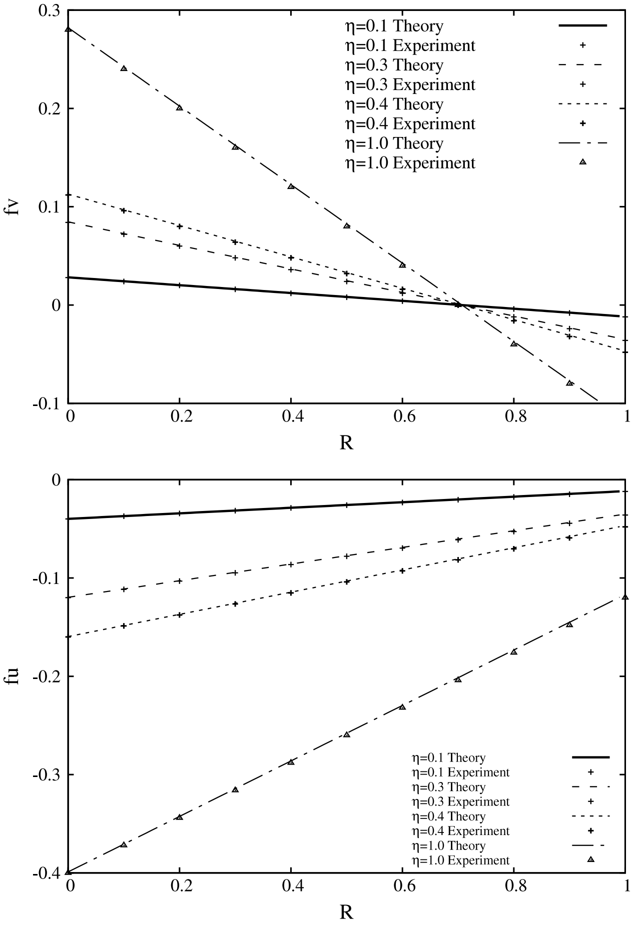

Fig. 2. The values of $\langle fv \rangle$ and $\langle fu \rangle$ by our theory and experiments in the perceptron learning.

learning is not given analytically. For $\sigma_B^2 = 1$, (14) and (15) for the perceptron learning are reduced to

$$\dot{l} = 0.28\eta R - 0.4\eta + \frac{-0.24\eta^2 R + 0.5\eta^2}{2l}, \tag{34}$$

$$\dot{R} = \frac{0.28\eta(1-R^2)}{l} - \frac{R}{2l^2}(-0.24\eta^2 R + 0.5\eta^2), \tag{35}$$

respectively, using the results in the previous section. Although the dynamical system is nonlinear, its asymptotic behavior can be analyzed by linearization around the equilibrium of the system. Hence, we first transformed their variables R and l to

$$e = (1-R) - 0.30, \qquad d = 1/l - 1.23/\eta, \tag{36}$$

so that the equilibrium becomes the origin, where R in the equilibrium does not depend on η. Then,

$$\begin{aligned}\dot{e} &= 0.1162\eta^2 d^2 + 0.1428\eta d - 0.593542\eta e d - 0.605574e \\ &\quad - 0.082\eta^2 e d^2 - 0.01494\eta e^2 d + 0.163172e^2 - 0.12\eta^2 e^2 d^2,\end{aligned} \tag{37}$$

$$\begin{aligned}\dot{d} &= -0.408\eta d^2 - 0.250699d - 0.16241\eta e d^2 + 0.144509ed \\ &\quad - 0.166\eta^2 d^3 + 0.200152\frac{e}{\eta} - 0.12\eta^2 e d^3,\end{aligned} \tag{38}$$

which is linearized around the origin to

$$\begin{pmatrix}\dot{e} \\ \dot{d}\end{pmatrix} = \begin{pmatrix}-0.606 & 0.143\eta \\ 0.200/\eta & -0.251\end{pmatrix}\begin{pmatrix}e \\ d\end{pmatrix} \tag{39}$$

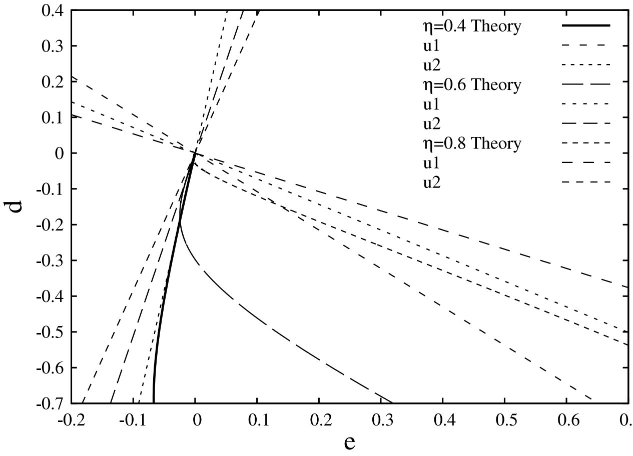

Fig. 3. Eigenvectors and traces of (e, d). $\sigma_B{}^2 = 1.0$ of the perceptron learning. The difference of the eigenvalues and the direction of the corresponding eigenvectors induce the curves.

in matrix form, as is done in systems science. This matrix is called the system matrix and its eigenvalues and the corresponding eigenvectors determine the behaviors of the dynamics, which were explicitly calculated as

$$\lambda_1 = -0.67, \qquad u_1 = \begin{pmatrix}\eta \\ -0.43\end{pmatrix}, \tag{40}$$

$$\lambda_2 = -0.18, \qquad u_2 = \begin{pmatrix}\eta \\ 3.07\end{pmatrix}. \tag{41}$$

The state vector, $(e, d)^T$, converges to the origin because $\lambda_1 < \lambda_2 < 0$ and the component along u_2 decreases more slowly than that along u_1. In addition, u_2 is in the first/third quadrant of the (e, d)-plane while u_1 is in the second/fourth quadrant. This means that the points in the fourth (lower-right) quadrant move to the third quadrant once and then go to the origin along u_2 (Fig. 3). Since $R = 0.70 - e$, the above explains how the overshoot of R appears in the perceptron learning.

In the same way, we can reduce (14) and (15) for the AdaTron learning with $\sigma_B^2 = 1$ to

$$\begin{aligned}\dot{e} &= 0.0701122\eta^2 d^2 + 0.138224\eta d - 0.3208\eta e d - 0.6e \\ &\quad - 0.0104\eta^2 e d^2 - 0.41\eta e^2 d - 0.205\eta^2 e^2 d^2,\end{aligned} \tag{42}$$

$$\begin{aligned}\dot{d} &= -0.5008\eta d^2 - 0.5008d - 0.82\eta e d^2 - 0.82ed \\ &\quad - 0.1252\eta^2 d^3 - 0.205\eta^2 e d^3,\end{aligned} \tag{43}$$

by transforming the variables R and l to

$$e = (1-R) - 0.44, \qquad d = 1/l - 2.0/\eta, \tag{44}$$

so that $e, d \to 0$ as $t \to \infty$. Since the eigenvalues and the corresponding eigenvalues of the system matrix are

$$\lambda_1 = -0.6, \qquad u_1 = \begin{pmatrix}1 \\ 0\end{pmatrix}, \tag{45}$$

$$\lambda_2 = -0.5008, \qquad u_2 = \begin{pmatrix}1.39332\eta \\ 1\end{pmatrix}. \tag{46}$$

These values and the orientations mean the AdaTron learning has the same property as the perceptron learning (Fig. 4). In fact, the AdaTron learning has a learning curve with an overshoot (Fig. 5).

In the case of the Hebbian learning, the nonmonotonicity does not appear (Fig. 6). The reason is elucidated by the theoretical analysis below. We can explicitly rewrite (14) and (15) for the

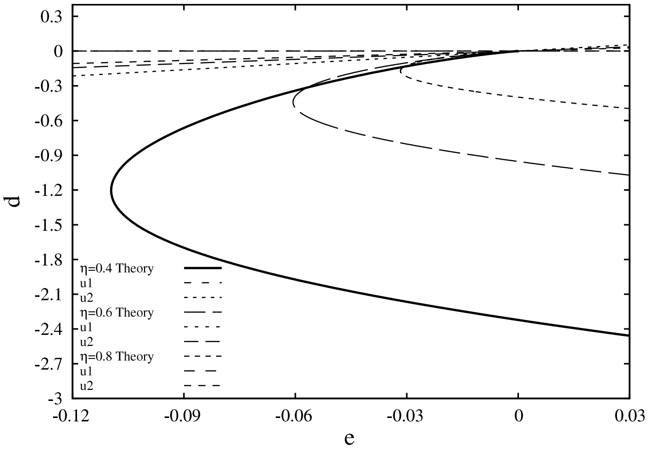

Fig. 4. Eigenvectors and traces of (e, d). $\sigma_B{}^2 = 1.0$ of the AdaTron learning. The difference of the eigenvalues and the direction of the corresponding eigenvectors induce the curves.

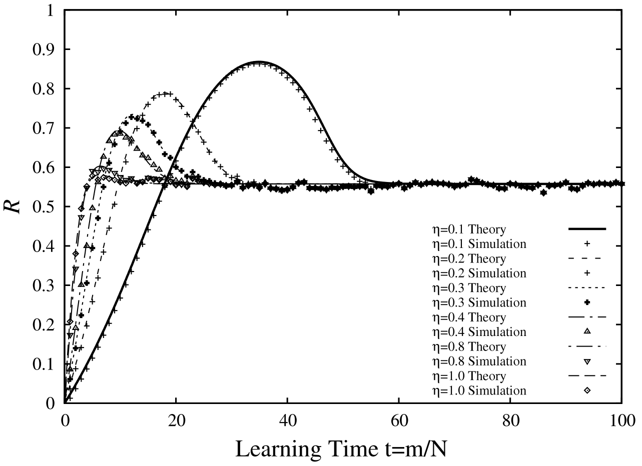

Fig. 5. Dynamics of R of the AdaTron learning. $\sigma_B^2 = 1.0, N = 10^4, \eta = 0.1, \dots, 1.0$, plots: experiments, lines: theory.

Hebbian learning to

$$\dot{e} = -\frac{2\eta}{\sqrt{\pi}}de + \frac{\eta^2}{2}d^2 + \frac{\eta}{\sqrt{\pi}}de^2 - \frac{\eta^2}{2}d^2e, \tag{47}$$

$$\dot{d} = -\frac{\eta}{\sqrt{\pi}}d^2 + \frac{\eta}{\sqrt{\pi}}d^2e - \frac{\eta^2}{2}d^3, \tag{48}$$

by transforming the variables R and l to

$$e = 1 - R, \qquad d = 1/l, \tag{49}$$

so that $e, d \to 0$ as $t \to \infty$.

The system matrix becomes null since the right-hand sides of (47) and (48) have no first order terms of e and d. This means that our systems scientific method is not applicable. Instead, these differential equations are explicitly solvable in the asymptotic of $e, d \to 0$ in the same way as Nishimori (2001). The lowest order of (14) and (15) are the second ones. Hence, they are simplified in the asymptotic as

$$\dot{e} = -\frac{2\eta}{\sqrt{\pi}}de + \frac{\eta^2}{2}d^2, \tag{50}$$

$$\dot{d} = -\frac{\eta}{\sqrt{\pi}}d^2, \tag{51}$$

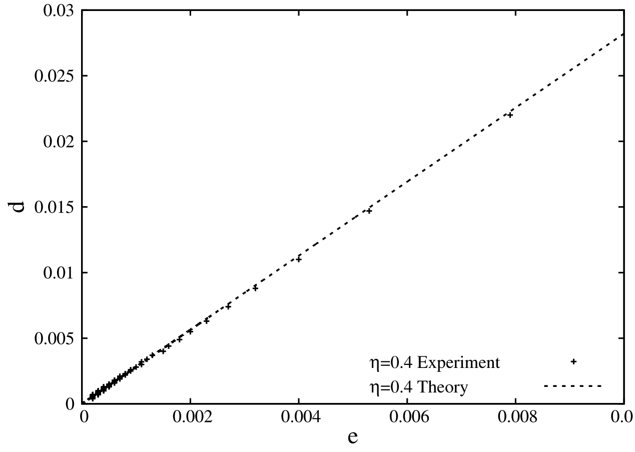

Fig. 6. Eigenvectors and traces of (e, d). $\sigma_B{}^2 = 1.0$ of the Hebbian learning. In this case, (e, d) converges to the origin linearly and hence the error does not have nonmonotonicity.

respectively. (51) is easily solved and d is expressed as

$$d = \frac{\sqrt{\pi}}{\eta}t^{-1}. \tag{52}$$

By substituting d in (50) with (52), we get

$$\dot{e} = -2t^{-1}e + \frac{\pi}{2}t^{-2}, \tag{53}$$

which leads to

$$e = \frac{\pi}{2}t^{-1}. \tag{54}$$

These theoretical values matched the simulation results (Fig. 6).

6. Conclusions

In this paper, we analyzed convergence properties of the perceptron learning, the AdaTron learning and the Hebbian learning, when the teacher was noisy. The learning curves in these cases were analytically derived using a statistical mechanical method and were consistent with the experimental results in our simulation. Our analyses showed that the learning curves of the perceptron learning and the AdaTron learning have an overshoot, that is, the covariance coefficient R of the teacher and the student exceeds the convergence value once. However, the Hebbian learning does not have this property. We showed that these phenomena result from the difference of the eigenvalues and eigenvectors of the system matrix using the asymptotic analysis of dynamical systems. This result may give a method for controlling the learning coefficient η to achieve a faster convergence speed and a lower residual error in the future.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Numbers 25280083, 25118009, 15H01620 from MEXT, Japan.

References

Biehl, M., & Schwarze, H. (1992). Online learning of a time-dependent rule. *Europhysics Letters*, *2*, 733–738.

Hara, K., & Okada, M. (2004). On-line learning through simple perceptron learning with a margin. *Neural Networks*, *17*, 215–223.

Ikeda, K., Hanzawa, H., & Miyoshi, S. (2013). Convergence properties of perceptron learning with noisy teacher. In *LNCS*: *Vol. 7751*. *Intelligent science and intelligent data engineering* (pp. 417–424).

Inoue, J., & Nishimori, H. (1997). On-line AdaTron learning of unlearnable rules. *Physical Review E*, *55*(4), 4544–4551.

Miyoshi, S., Hara, K., & Okada, M. (2005). Analysis of ensemble learning using simple perceptrons based on online learning theory. *Physical Review E*, *71*, 036116.

Miyoshi, S., & Kajikawa, Y. (2013). Statistical-mechanical analysis of the FXLMS algorithm with nonwhite reference signals. In *Proc. ICASSP 2013* (pp. 5652–5656).

Miyoshi, S., & Okada, M. (2006a). Analysis of on-line learning when a moving teacher goes around a true teacher. *Journal of the Physical Society of Japan*, *75*(2), 024003.

Miyoshi, S., & Okada, M. (2006b). Statistical mechanics of online learning for ensemble teachers. *Journal of the Physical Society of Japan*, *75*(4), 044002.

Nishimori, H. (2001). *Statistical physics of spin glasses and information processing: An introduction*. Oxford, UK: Oxford Univ. Press.

Rosenblatt, F. (1961). Principle of neurodynamics. Spartan, Washington D. C.

Tanaka, T. (2002). A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Transaction on Information Theory*, *48*(11), 2888–2910.

Uezu, T., Miyoshi, S., Izuo, M., & Okada, M. (2007). Theory of time domain ensemble on-line learning of perceptron under the existence of external noise. *Journal of the Physical Society of Japan*, *76*(11), 114006.