Journal of the Physical Society of Japan
Vol. 75, No. 11, November, 2006, 114007
©2006 The Physical Society of Japan

On-line Learning of Perceptron from Noisy Data by One and Two Teachers

Tatsuya UEZU*, Yoshiko MAEDA† and Sachi YAMAGUCHI[1]‡

Graduate School of Sciences and Humanities, Nara Women's University, Nara 630-8506
[1]*Faculty of Sciences, Department of Physics, Nara Women's University, Nara 630-8506*

(Received June 24, 2006; accepted September 5, 2006; published November 10, 2006)

We analyze the on-line learning of a Perceptron from signals produced by a single Perceptron suffering from external noise or by two independent Perceptrons without noise. We adopt typical three learning rules in both single-teacher and two-teacher cases. For the single-teacher case, we treat the input and output noises and for the two-teacher case, we assume that signals are given by two teachers with a definite probability. In the single-teacher case, in order to improve the learning when it does not succeed in the sense that the student vector does not converge to the teacher vector, we use two methods: a method based on the optimal learning rate and an averaging method. Furthermore, we obtain an asymptotic form of the generalization error using an optimal learning rate for the three learning rules, and we estimate noise parameters using the simulation data by the averaging method. In the two-teacher case, for the Hebbian rule, we give analytical solutions of order parameters. Furthermore, we estimate noise parameters using the Perceptron rule by the averaging method. The theoretical results agree quite well with the numerical simulations.

KEYWORDS: perceptron, on-line learning, generalization error, noise
DOI: 10.1143/JPSJ.75.114007

1. Introduction

We study the on-line learning by a single Perceptron[1)] from signals produced by a single teacher or by two teachers.

In the single-teacher case, we assume that the data is contaminated by noise and we adopt the Hebbian,[2)] Perceptron,[1)] and AdaTron[3)] rules as learning rules.[4)] There have been many studies that focus on the case of a single teacher.[5–10)] The main purpose of the present paper is to offer some methods of identifying the teacher vector and estimating noise parameters when the learning is not successful in the sense that the student vector does not converge to the teacher vector.[11)] In the two-teacher case, few previous studies exist.[11)] In this case, we study a situation in which signals are given by two teachers with a definite probability, and by adopting the Hebbian, Perceptron, and AdaTron rules as learning rules, we then estimate the probability. The results are as follows: In the single-teacher case, when the learning fails, the teacher can be identified using the optimal learning rate or by taking the average of the student vector at different times. In particular, noise parameters can be determined using the averaging method. Furthermore, we can obtain an asymptotic form of the generalization error using an optimal learning rate for the three learning rules. In the two-teacher case, the student vector approaches the two-dimensional space Σ spanned by the teacher vectors for the Hebbian rule. On the other hand, for the Perceptron and AdaTron rules, the student vector does not approach Σ, but the time-averaged student vector does. Using this fact, by the averaging method we estimate the probability that the signals are sent by the teachers in the Perceptron rule. Furthermore, both in the single-teacher and two-teacher cases, in the averaging method, we find that the behaviors of the convergence of learning are quite similar when the starting time at which the average is taken is larger than the time at which the student vector starts to rotate around the teacher vector.

The paper is organized as follows: In §2, the formulation in the case of a single teacher is given. In §3 and §4, the cases of output noise and input noise are analyzed, respectively. In §5, the formulation and analysis of the two-teacher case are given. Section 6 is devoted to a summary and discussions.

2. Formulation in Single-Teacher Case

We consider the supervised learning of a Perceptron in the presence of noise. Let $\boldsymbol{J}$ and $\boldsymbol{B}$ be the student and teacher vectors, respectively. We assume that these are N-dimensional vectors. We also assume that $|\boldsymbol{B}| = 1$. Let $\boldsymbol{\xi}$ be an N-dimensional example vector. We assume that its component ξ_i takes ±1 and is drawn independently with the probability $P(\xi = 1) = 1 - P(\xi = -1) = 1/2$. The output S generated by the student $\boldsymbol{J}$ for $\boldsymbol{\xi}$ is given by

$$S = \mathrm{sgn}(\boldsymbol{J} \cdot \boldsymbol{\xi}), \tag{1}$$

where $\boldsymbol{J} \cdot \boldsymbol{\xi}$ denotes the inner product of $\boldsymbol{J}$ and $\boldsymbol{\xi}$, $\mathrm{sgn}(x) = 1$ for $x \geq 0$, and $\mathrm{sgn}(x) = -1$ for $x < 0$. When there is no noise, the output T generated by the teacher for $\boldsymbol{\xi}$ is given by

$$T = \mathrm{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}). \tag{2}$$

In this paper, we treat the cases in which noise exists. We consider the output noise and input noise. Let $\mathcal{P}$ be the probability of $T = 1$. In the output noise model, $\mathcal{P}$ is given by

$$\mathcal{P}(y) = \frac{1}{2}(1 + k\,\mathrm{sgn}(y)), \tag{3}$$

where $y = \boldsymbol{B} \cdot \boldsymbol{\xi}$. That is, for $y > 0$, the probability of $T = 1$ is $(1 + k)/2$. In the input noise model, T is given by

*E-mail: uezu@cc.nara-wu.ac.jp
†Present address: Mitsubishi Electric Company.
‡Present address: Graduate School of Sciences and Humanities, Nara Women's University.

$$T = \mathrm{sgn}(\boldsymbol{B} \cdot (\boldsymbol{\xi} + \boldsymbol{\zeta})), \tag{4}$$

where each component ζ_i of $\boldsymbol{\zeta}$ is assumed to be independently drawn from the Gaussian distribution of the mean 0 and the standard deviation σ. Then, $\mathscr{P}$ is expressed as

$$\mathscr{P}(y) = 1 - H\left(\frac{y}{\sigma}\right), \tag{5}$$

where $H(y) = \int_y^\infty \mathrm{D}u$ and $\mathrm{D}u = (\mathrm{d}u/\sqrt{2\pi})\mathrm{e}^{-u^2/2}$. We adopt the following learning algorithm

$$\boldsymbol{J}\left(t + \frac{1}{N}\right) = \boldsymbol{J}(t) + \frac{1}{N}\eta\boldsymbol{\xi}T\mathcal{F}[|\boldsymbol{J}|; \boldsymbol{J} \cdot \boldsymbol{\xi}, T], \tag{6}$$

where η is the learning rate and $\mathcal{F}$ is the learning rule and is assumed to depend on $|\boldsymbol{J}|$, $\boldsymbol{J} \cdot \boldsymbol{\xi}$ and T. Here, $|\boldsymbol{J}|$ is the norm of $\boldsymbol{J}$. We consider the following three learning rules

$$\text{Hebbian rule: } \mathcal{F} = 1, \tag{7}$$

$$\text{Perceptron rule: } \mathcal{F} = \Theta(-TS), \tag{8}$$

$$\text{AdaTron rule: } \mathcal{F} = |\boldsymbol{\xi} \cdot \boldsymbol{J}|\Theta(-TS), \tag{9}$$

where $\Theta(x) = 1$ for $x \geq 0$ and $\Theta(x) = 0$ for $x < 0$. As for the order parameters, we adopt $Q = \boldsymbol{J}^2$ and $R = \boldsymbol{J} \cdot \boldsymbol{B}$. From eq. (6), we obtain the differential equations for Q and R:[8]

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 2\eta\langle(\boldsymbol{J} \cdot \boldsymbol{\xi})T\mathcal{F}\rangle_\Xi + \eta^2\langle\mathcal{F}^2\rangle_\Xi, \tag{10}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta\langle(\boldsymbol{B} \cdot \boldsymbol{\xi})T\mathcal{F}\rangle_\Xi. \tag{11}$$

Here, we assume self-averaging[12] and $\langle\cdot\rangle_\Xi$ denotes the average over examples and noises. Let us define $J = |\boldsymbol{J}|$, $\widehat{\boldsymbol{J}} \equiv \boldsymbol{J}/J$, and $x \equiv \widehat{\boldsymbol{J}} \cdot \boldsymbol{\xi}$. Since $\mathcal{F}$ is expressed as $\mathcal{F}[J; Jx, T]$, these equations are rewritten as

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 2\eta J\langle xT\mathcal{F}[J; Jx, T]\rangle_\Xi + \eta^2\langle\mathcal{F}^2[J; Jx, T]\rangle_\Xi, \tag{12}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta\langle yT\mathcal{F}[J; Jx, T]\rangle_\Xi. \tag{13}$$

In addition to Q and R, $J = \sqrt{Q}$ and $\omega = R/J$ are also used, and their equations are

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \eta\langle xT\mathcal{F}[J; Jx, T]\rangle_\Xi + \frac{\eta^2}{2J}\langle\mathcal{F}^2[J; Jx, T]\rangle_\Xi \tag{14}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\eta}{J}\langle(y - \omega x)T\mathcal{F}[J; Jx, T]\rangle_\Xi - \frac{\omega\eta^2}{2J^2}\langle\mathcal{F}^2[J; Jx, T]\rangle_\Xi. \tag{15}$$

The generalization error E is given by

$$E = \langle\Theta(-ST)\rangle_\Xi. \tag{16}$$

The probability distribution $P(x, y)$ of x and y is given by the Gaussian distribution with $\langle x\rangle = 0$, $\langle y\rangle = 0$, $\langle x^2\rangle = 1$, $\langle y^2\rangle = 1$ and $\langle xy\rangle = \omega$,

$$P(x, y) = \frac{1}{2\pi\sqrt{1 - \omega^2}}\exp\left[-\frac{1}{2(1 - \omega^2)}(x^2 + y^2 - 2\omega xy)\right]. \tag{17}$$

Thus, the average over examples $\boldsymbol{\xi}$ of A, $\langle A\rangle_\xi$, is replaced by $\langle A\rangle_{x,y} \equiv \int \mathrm{d}x\,\mathrm{d}yP(x, y)A$. The average over noise $\boldsymbol{\zeta}$ of a quantity $A(T)$ is given as follows:

$$\begin{aligned}\langle A\rangle_\zeta &= A(1)\mathscr{P}(y) + A(-1)(1 - \mathscr{P}(y)) \\ &= A(-1) + \mathscr{P}(y)(A(1) - A(-1)).\end{aligned} \tag{18}$$

3. Output Noise Model

In the output noise model, $\mathscr{P}(y) = (1/2)(1 + k\,\mathrm{sgn}(y))$. Then, the average of $A(T)$ over noise is given by

$$\begin{aligned}\langle A\rangle_\zeta &= \frac{1}{2}(A_+ + A_-) + \frac{k}{2}(A_+ - A_-)\,\mathrm{sgn}(y) \\ &= A_\mathrm{s} + kA_\mathrm{as}\,\mathrm{sgn}(y),\end{aligned} \tag{19}$$

where

$$A_+ = A(1), \quad A_- = A(-1),$$

$$A_\mathrm{s} = \frac{1}{2}(A_+ + A_{-1}),$$

and

$$A_\mathrm{as} = \frac{1}{2}(A_+ - A_{-1}).$$

Since

$$\langle T\mathcal{F}\rangle_\zeta = (T\mathcal{F})_\mathrm{s} + k(T\mathcal{F})_\mathrm{as}\,\mathrm{sgn}(y) = \mathcal{F}_\mathrm{as} + k\mathcal{F}_\mathrm{s}\,\mathrm{sgn}(y),$$

we obtain

$$\begin{aligned}\frac{\mathrm{d}Q}{\mathrm{d}t} &= 2\eta J\langle x\{\mathcal{F}_\mathrm{as} + k\mathcal{F}_\mathrm{s}\,\mathrm{sgn}(y)\}\rangle_{x,y} \\ &\quad + \eta^2\langle\mathcal{F}_\mathrm{s}^2 + k\mathcal{F}_\mathrm{as}^2\,\mathrm{sgn}(y)\rangle_{x,y},\end{aligned} \tag{20}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta\langle y\{\mathcal{F}_\mathrm{as} + k\mathcal{F}_\mathrm{s}\,\mathrm{sgn}(y)\}\rangle_{x,y}. \tag{21}$$

By performing the average over x and y, we get equations for Q, R, J, and ω. The generalization error $E = \langle\Theta(-TS)\rangle_\Xi$ is given by

$$\begin{aligned}E &= \frac{1 - k}{2} + \frac{k}{\pi}\cos^{-1}(\omega), \\ &= E_\mathrm{min} + \frac{k}{\pi}\cos^{-1}(\omega),\end{aligned}$$

where $E_\mathrm{min} = (1 - k)/2$ is the minimum value of the generalization error. Then $\tilde{E} \equiv E - E_\mathrm{min}$ is expressed as

$$\tilde{E} = \frac{k}{\pi}\cos^{-1}(\omega). \tag{22}$$

In the next subsection, we study the learning behavior when the learning rate η is constant.

3.1 *Case of constant learning rate*

We summarize the learning behavior in each learning rule.

In the Hebbian rule, the equations for R, J, and ω are

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta k\sqrt{\frac{2}{\pi}}, \tag{23}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \eta k\sqrt{\frac{2}{\pi}}\omega + \frac{\eta^2}{2J}, \tag{24}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\eta k}{J}\sqrt{\frac{2}{\pi}}(1 - \omega^2) - \frac{\eta^2\omega}{2J^2}. \tag{25}$$

This case has been studied previously and these equations have been solved analytically.[13] The solutions for R, J, and ω with initial conditions $R(0) = 0$, $J(0) = 1$, and $\omega(0) = 0$ are

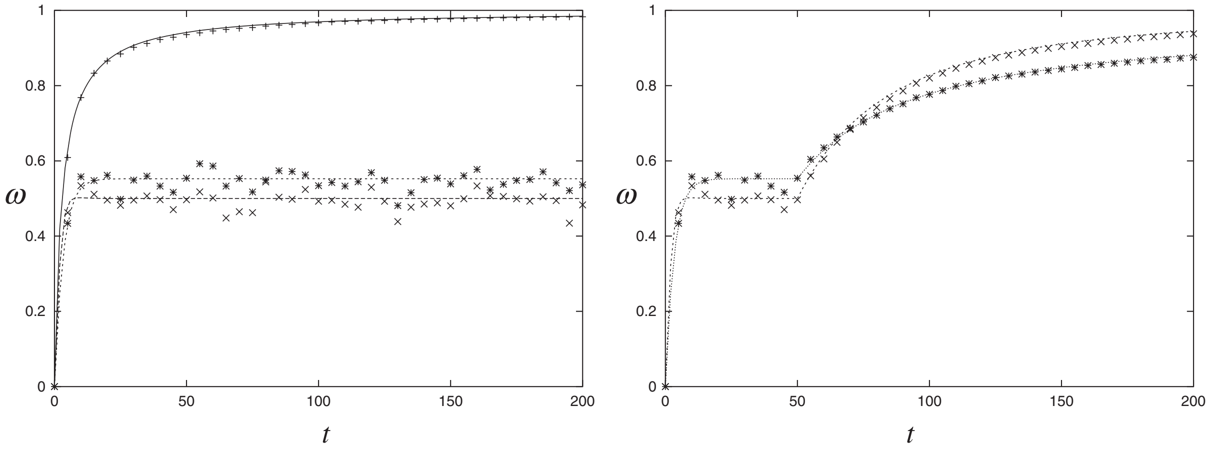

Fig. 1. Time dependence of ω in output noise model. $k = 0.5$. Left panel: Constant learning rate, $\eta = 1$. Theoretical results (RKG): solid curve: Hebbian, dashed curve: Perceptron, dotted curve: AdaTron. Simulations ($N = 1000$); $+$: Hebbian, $\times$: Perceptron, $*$: AdaTron. Right panel: Optimal learning rate. For $t < 50$, $\eta = 1$ and for $t \geq 50$, $\eta = \eta_{\text{opt}}(t)$. Theoretical results (RKG): dashed curve: Perceptron, dotted curve: AdaTron. Simulations ($N = 1000$); $\times$: Perceptron, $*$: AdaTron.

$$R = \eta k\sqrt{\frac{2}{\pi}}t, \tag{26}$$

$$J = \sqrt{1 + \eta^2 t\left(1 + \frac{2}{\pi}k^2 t\right)}, \tag{27}$$

$$\omega = \left[1 + \frac{\pi}{2}\frac{1 + \eta^2 t}{k^2\eta^2 t^2}\right]^{-1/2}. \tag{28}$$

Thus, $J \to \infty$ and $\omega \to 1$ as $t \to \infty$. Therefore, learning succeeds in the sense that the student vector converges to the teacher vector, even if noise exists. In the Perceptron rule, the equations for J and ω are

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}}(k - \omega), \tag{29}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}}(\omega k - 1) + \frac{\eta^2}{2J}\left(\frac{1}{2} - \frac{k}{\pi}\sin^{-1}(\omega)\right), \tag{30}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\eta k}{\sqrt{2\pi}J}(1 - \omega^2) - \frac{\eta^2\omega}{2J^2}\left(\frac{1}{2} - \frac{k}{\pi}\sin^{-1}(\omega)\right). \tag{31}$$

From these equations, we obtain the following stationary state:

$$J_P^* = \eta\sqrt{\frac{\pi}{2}}\frac{\frac{1}{2} - \frac{k}{\pi}\sin^{-1}(k)}{1 - k^2}, \quad \omega_P^* = k.$$

Since $\omega_P^* < 1$, learning fails.

In the AdaTron rule, the equations for J and ω are

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{k}{\pi}\eta J\sqrt{1 - \omega^2} - \eta\omega J\left(\frac{1}{2} - \frac{k}{\pi}\sin^{-1}(\omega)\right), \tag{32}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \eta J\left(\frac{\eta}{2} - 1\right)\left(\frac{1}{2} - \frac{k}{\pi}\left(\omega\sqrt{1 - \omega^2} + \sin^{-1}(\omega)\right)\right), \tag{33}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{k\eta}{\pi}(1 - \omega^2)^{3/2} - \frac{\eta^2}{2}\omega\left(\frac{1}{2} - \frac{k}{\pi}\left(\omega\sqrt{1 - \omega^2} + \sin^{-1}(\omega)\right)\right). \tag{34}$$

The equation for ω does not include J. The factor

$$\left(\frac{1}{2} - \frac{k}{\pi}\left(\omega\sqrt{1 - \omega^2} + \sin^{-1}(\omega)\right)\right)$$

in the equation for J is positive for $0 < k < 1$ and $0 < \omega \leq 1$. Thus, as $t \to \infty$, $J \to 0$ for $\eta < 2$, $J =$ constant for $\eta = 2$ and $J \to \infty$ for $\eta > 2$. $\omega \to \omega_A^*$ as $t \to \infty$. Here, ω_A^* is the solution of $\mathrm{d}\omega/\mathrm{d}t = 0$ and is less than 1. As in the case of the Perceptron rule, learning fails.

As shown in the left panel of Fig. 1, in each learning rule there is agreement between the simulation results and the theoretical ones obtained using the Runge–Kutta–Gill (RKG) method.

As seen above, learning fails for the Perceptron and AdaTron rules. That is, ω does not tend to 1. In the following subsections, we consider two methods to improve the learning for these two cases. First, we introduce the time-dependent learning rate η and second, we take the time average.

3.2 *Optimal learning rate*

Now, let us discuss the optimal learning rate η_{opt}. η_{opt} is defined by the following relation:[8)]

$$^\forall t \geq 0: \quad \frac{\partial}{\partial\tilde{\eta}}\left(\frac{\mathrm{d}}{\mathrm{d}t}E\right) = 0. \tag{35}$$

$\tilde{\eta}$ is η/J for the Hebbian and Perceptron rules, and is η for the AdaTron rule. Since $E = (1 - k)/2 + (k/\pi)\cos^{-1}(\omega)$, the relationship is equivalent to

$$^\forall t \geq 0: \quad \frac{\partial}{\partial\tilde{\eta}}\left(\frac{\mathrm{d}}{\mathrm{d}t}\omega\right) = 0. \tag{36}$$

For each of the three learning rules, it is shown that $\omega \to 1$ when $\tilde{\eta}_{\text{opt}}$ is adopted. See Appendix A. In the right panel of Fig. 1, we display the numerical results for ω in each rule. We found excellent agreement between the theoretical and numerical results. In the theoretical calculation, we used the asymptotic forms of $\tilde{\eta}_{\text{opt}}$. In Table I, the time dependences of the optimal $\tilde{\eta}_{\text{opt}}$ and $\tilde{E}_{\text{opt}}$, where the latter is $\tilde{E}$ obtained using $\tilde{\eta}_{\text{opt}}$ for large t, are given for each learning rule. In Table II, we summarize the asymptotic behavior of ω, $\tilde{E}$, and J for a constant η and for the optimal η, η_{opt}, for each learning rule. Here, $\eta_{\text{opt}} = \tilde{\eta}_{\text{opt}}|J|$ for the Hebbian and Perceptron rules, and $\eta_{\text{opt}} = \tilde{\eta}_{\text{opt}}$ for the AdaTron rule.

From Table I, we note that the asymptotic form of $\tilde{E}_{\text{opt}}$ is

Table I. Asymptotic form of optimal learning rate and $\widetilde{E}_{\rm opt}$ for $t \gg 1$ in output noise model.

Learning rule	Hebbian	Perceptron	AdaTron
$\widetilde{\eta}_{\rm opt}(t)$	$\frac{1}{k}\sqrt{\frac{\pi}{2}}t^{-1}$	$2\sqrt{2\pi}t^{-1}$ $(k=1)$, $\frac{\sqrt{2\pi}}{k}t^{-1}$ $(k<1)$	$\frac{3}{2}$ $(k=1)$, $\left(\frac{\pi^2}{4k^2(1-k)}\right)^{1/4}t^{-3/4}$ $(k<1)$
$\widetilde{E}_{\rm opt}$	$\frac{1}{\sqrt{2\pi}}t^{-1/2}$	$\frac{4}{\pi}t^{-1}$ $(k=1)$, $\sqrt{\frac{1-k}{\pi}}t^{-1/2}$ $(k<1)$	$\frac{4}{3}t^{-1}$ $(k=1)$, $\left(\frac{k^2(1-k)}{4\pi^2}\right)^{1/4}t^{-1/4}$ $(k<1)$

Table II. Asymptotic behavior with constant η and $\eta_{\rm opt}$ in output noise model for $k<1$.

Learning rule	Hebbian	Perceptron	AdaTron
Asymptotic behavior with $\eta=$ constant $(=1)$	$\omega \to 1$ $\widetilde{E} \to 0$ $J \to \infty$	$\omega \to \omega_P^*$ (<1) $\widetilde{E} \to \widetilde{E}_P^*$ $J \to J_P^*$	$\omega \to \omega_A^*$ (<1) $\widetilde{E} \to \widetilde{E}_A^*$ $J \to 0$
Asymptotic behavior with $\eta=\eta_{\rm opt}$	$\omega \to 1$ $\widetilde{E} \to 0$ $J \to \infty$	$\omega \to 1$ $\widetilde{E} \to 0$ $J \to 0$	$\omega \to 1$ $\widetilde{E} \to 0$ $J \to 0$

proportional to $t^{-1/2}$ for the Hebbian and Perceptron rules, whereas it is proportional to $t^{-1/4}$ for the AdaTron rule for $k<1$. That is, the convergence speed of the Hebbian rule and that of the Perceptron rule are comparable, but that of the AdaTron is much lower than those of the Hebbian and Perceptron rules.

Next, we study the averaging method used to improve learning.

3.3 Time averaging method

In the Perceptron rule, $\omega \to \omega_P^*$ and in the AdaTron rule, $\omega \to \omega_A^*$ as $t \to \infty$. Both ω_P^* and ω_A^* are less than 1. Thus, we consider that $\widehat{\boldsymbol{J}} = \boldsymbol{J}/J$ rotates around or is scattered around $\boldsymbol{B}$ as time progresses. Therefore, we expect that by taking the time average of $\widehat{\boldsymbol{J}}$, the direction of the time-averaged vector $\langle\widehat{\boldsymbol{J}}\rangle$ of $\widehat{\boldsymbol{J}}$ tends toward the direction of $\boldsymbol{B}$ as the number of samples in the average increases. Here, $\langle\widehat{\boldsymbol{J}}\rangle$ is defined by

$$\langle\widehat{\boldsymbol{J}}\rangle \equiv \frac{1}{L}\sum_{i=1}^{L}\widehat{\boldsymbol{J}}(t_i), \tag{37}$$

where $0 \leqq t_1 < t_2 < \cdots < t_L$. In Fig. 2, we display the results of this averaging method.

Since $J \to J_P^*$ for the Perceptron rule, we used the time-averaged vector $\langle\boldsymbol{J}\rangle$ of $\boldsymbol{J}$ only. As shown in the left panel of Fig. 2, $(\boldsymbol{B}\cdot\langle\boldsymbol{J}\rangle)/|\langle\boldsymbol{J}\rangle|$ increases and seems to approach 1 as the number of samples in the average, which is denoted by L in eq. (37), increases. For the AdaTron rule, since $J \to 0$, we used both $\langle\boldsymbol{J}\rangle$ and $\langle\widehat{\boldsymbol{J}}\rangle$, and found that we could get $\omega \to 1$ using only $\langle\widehat{\boldsymbol{J}}\rangle$, as L increases.

In Fig. 3, we display the dependence of the convergence of ω on the starting time t_1 when the average is taken. As can be seen in the right panel of Fig. 3, the $t-t_1$ dependences are quite similar for $t_1 = 5$, 10, 25, 50, 100, and 150 except for $t_1 = 0$. This result is attributed to the fact that the student vector already starts to rotate around the teacher vector for these values of t_1 as is seen in the left panel of Fig. 3. On the other hand, for $t_1 = 0$, where all data are used to take the average, the convergence is slower than for other cases because ω is still approaching ω_P^*. We also obtained similar results using the AdaTron rule. Furthermore, we can estimate k from the relation $|\langle\widehat{\boldsymbol{J}}_P\rangle| = k$ or $|\langle\widehat{\boldsymbol{J}}_A\rangle| = \omega_A^*$ in the Perceptron or AdaTron rule, respectively. Indeed, k was estimated as 0.501 and 0.504 when $k=0.5$ using the value $|\langle\widehat{\boldsymbol{J}}_P\rangle|$ in the Perceptron rule and $|\langle\widehat{\boldsymbol{J}}_A\rangle|$ in the AdaTron rule at $t=1000$, respectively.

In the next section, we study the input noise model.

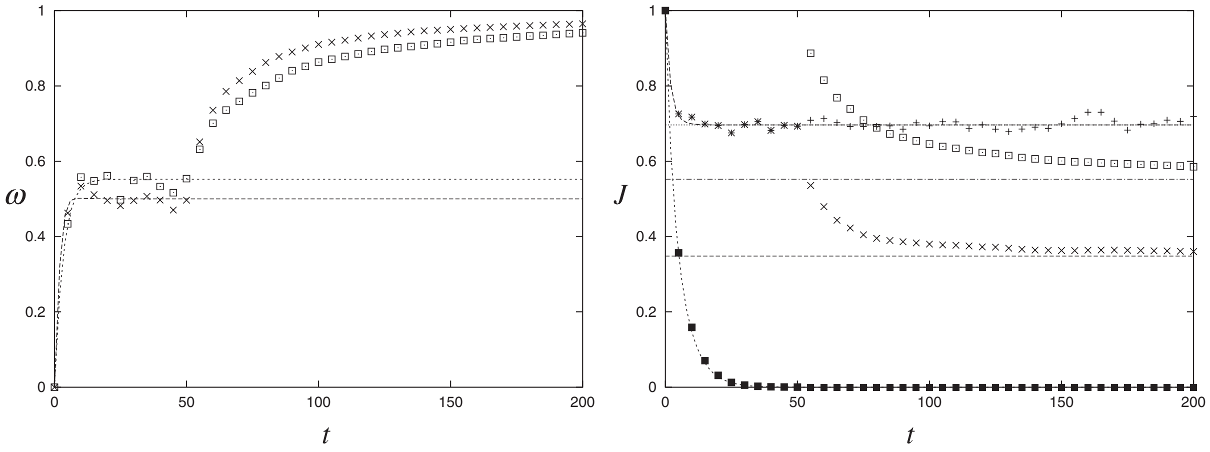

Fig. 2. Averaging method for Perceptron and AdaTron rules in output noise model. $k=0.5$. The average is taken for $t \geq 50$; that is, t_1 in eq. (31) is 50. Symbols denote simulation data for $N=1000$. ×: Perceptron (not normalized), square: AdaTron (normalized). Curves denote the theoretical results for $\eta=1$ without averaging. Dashed: Perceptron, dotted: AdaTron. Left panel: Time dependence of ω. Right panel: Time dependence of J. Data without averaging for $\eta=1$ are also depicted. +: Perceptron; closed square: AdaTron. The theoretical results for J_P^* (dotted line), $J_P^*\omega_P^*$ (dashed line) and ω_A^* (dashed-dotted line) are depicted in the right panel.

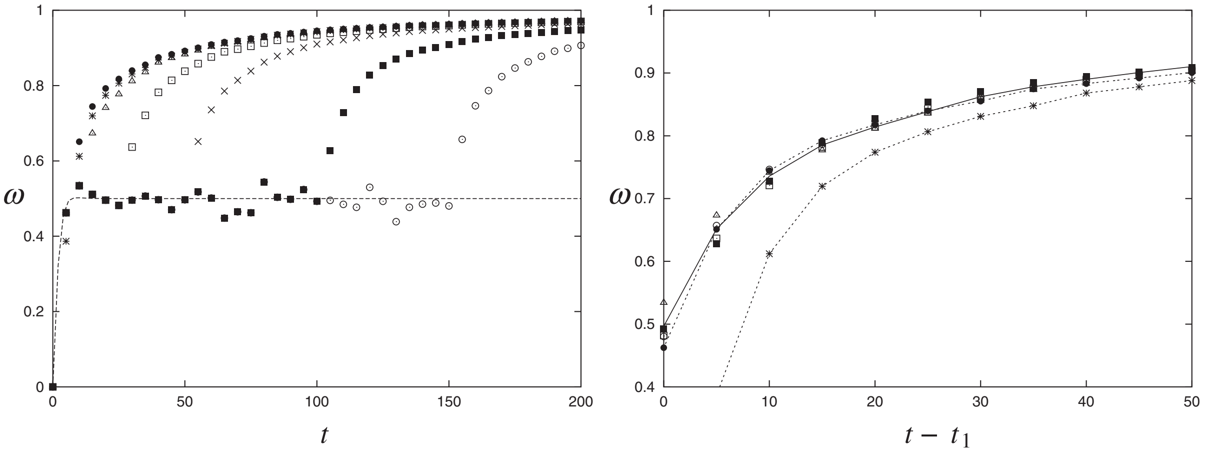

Fig. 3. Averaging method for Perceptron rule in output noise model. $k = 0.5$. t_1 dependence of convergence. Symbols denote simulation data for $N = 1000$. Left panel: t dependence of ω. $*$: $t_1 = 0$, closed circle: $t_1 = 5$, triangle: $t_1 = 10$, square: $t_1 = 25$, ×: $t_1 = 50$, closed square: $t_1 = 100$, circle: $t_1 = 150$. Right panel: $t - t_1$ dependence of ω. Each symbol corresponds to the same value of t_1 as in the left panel except for the case of $t_1 = 50$, in which data are connected by a solid line without symbols. Furthermore, for $t_1 = 0$ and 5, symbols are connected by a dashed line.

4. Input Noise Model

In the input noise model, $\mathcal{P}(y) = 1 - H(y/\sigma)$. Then, the average of $A(T)$ over noise is given by

$$\langle A\rangle_\zeta = A_+ - (A_+ - A_-)H\left(\frac{y}{\sigma}\right) = A_+ - 2A_{\rm as}H\left(\frac{y}{\sigma}\right). \tag{38}$$

Since

$$\begin{aligned}\langle T\mathcal{F}\rangle_\zeta &= (T\mathcal{F})_+ - 2(T\mathcal{F})_{\rm as}H\left(\frac{y}{\sigma}\right)\\ &= \mathcal{F}_+ - 2\mathcal{F}_{\rm s}H\left(\frac{y}{\sigma}\right),\end{aligned} \tag{39}$$

we obtain

$$\begin{aligned}\frac{{\rm d}Q}{{\rm d}t} &= 2\eta J\left\langle x\left\{\mathcal{F}_+ - 2\mathcal{F}_{\rm s}H\left(\frac{y}{\sigma}\right)\right\}\right\rangle_{x,y}\\ &\quad + \eta^2\left\langle \mathcal{F}_+^2 - 2\mathcal{F}_{\rm as}^2H\left(\frac{y}{\sigma}\right)\right\rangle_{x,y}\end{aligned} \tag{40}$$

$$\frac{{\rm d}R}{{\rm d}t} = \eta\left\langle y\left\{\mathcal{F}_+ - 2\mathcal{F}_{\rm s}H\left(\frac{y}{\sigma}\right)\right\}\right\rangle_{x,y}. \tag{41}$$

By taking the average over x and y, we get equations for Q, R, J, and ω. The generalization error is given by

$$E = \frac{1}{\pi}\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right). \tag{42}$$

The minimum value of E is

$$E_{\rm min} = \frac{1}{\pi}\cos^{-1}\left(\frac{1}{\sqrt{1+\sigma^2}}\right). \tag{43}$$

Thus, $\tilde{E} = E - E_{\rm min}$ is

$$\tilde{E} = E - \frac{1}{\pi}\cos^{-1}\left(\frac{1}{\sqrt{1+\sigma^2}}\right). \tag{44}$$

In the next subsection, we study the learning behavior when the learning rate η is constant.

4.1 *Case of constant learning rate*

In the Hebbian rule, we obtain

$$\frac{{\rm d}R}{{\rm d}t} = \sqrt{\frac{2}{\pi(1+\sigma^2)}}\eta, \tag{45}$$

$$\frac{{\rm d}J}{{\rm d}t} = \sqrt{\frac{2}{\pi(1+\sigma^2)}}\eta\omega + \frac{\eta^2}{2J}, \tag{46}$$

$$\frac{{\rm d}\omega}{{\rm d}t} = \frac{\eta}{J}\sqrt{\frac{2}{\pi(1+\sigma^2)}}(1-\omega^2) - \frac{\omega\eta^2}{2J^2}. \tag{47}$$

This case has also been studied previously, and these equations have been solved analytically.[13] These equations and their solutions can be obtained from eqs. (23)–(28), replacing k by $1/\sqrt{1+\sigma^2}$. Thus, $J \to \infty$ and $\omega \to 1$ as $t \to \infty$. Therefore, the learning succeeds even if noise exists. In the Perceptron rule, we obtain

$$\frac{{\rm d}R}{{\rm d}t} = \frac{\eta}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{1+\sigma^2}} - \omega\right), \tag{48}$$

$$\frac{{\rm d}J}{{\rm d}t} = \frac{\eta}{\sqrt{2\pi}}\left(\frac{\omega}{\sqrt{1+\sigma^2}} - 1\right) + \frac{\eta^2}{2\pi J}\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right), \tag{49}$$

$$\frac{{\rm d}\omega}{{\rm d}t} = \frac{\eta}{J}\frac{1-\omega^2}{\sqrt{2\pi(1+\sigma^2)}} - \frac{\omega\eta^2}{2\pi J^2}\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right). \tag{50}$$

From these equations, we get the stationary state as

$$J_P^* = \frac{1+\sigma^2}{\sigma^2}\frac{\eta}{\sqrt{2\pi}}\cos^{-1}\left(\frac{1}{1+\sigma^2}\right), \quad \omega_P^* = \frac{1}{\sqrt{1+\sigma^2}}. \tag{51}$$

Thus, learning fails for $\sigma > 0$. In the AdaTron rule, we obtain

$$\frac{{\rm d}R}{{\rm d}t} = \frac{\eta J}{\pi}\left\{\frac{\sqrt{1+\sigma^2-\omega^2}}{1+\sigma^2} - \omega\right\}, \tag{52}$$

$$\begin{aligned}\frac{{\rm d}J}{{\rm d}t} &= \frac{\eta J}{\pi}\left(\frac{\eta}{2} - 1\right)\\ &\quad \times\left\{\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right) - \frac{\omega\sqrt{1+\sigma^2-\omega^2}}{1+\sigma^2}\right\},\end{aligned} \tag{53}$$

$$\begin{aligned}\frac{{\rm d}\omega}{{\rm d}t} &= \eta\left(1 - \omega^2 + \frac{\eta}{2}\omega^2\right)\frac{\sqrt{1+\sigma^2-\omega^2}}{\pi(1+\sigma^2)}\\ &\quad - \frac{\eta^2}{2\pi}\omega\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right).\end{aligned} \tag{54}$$

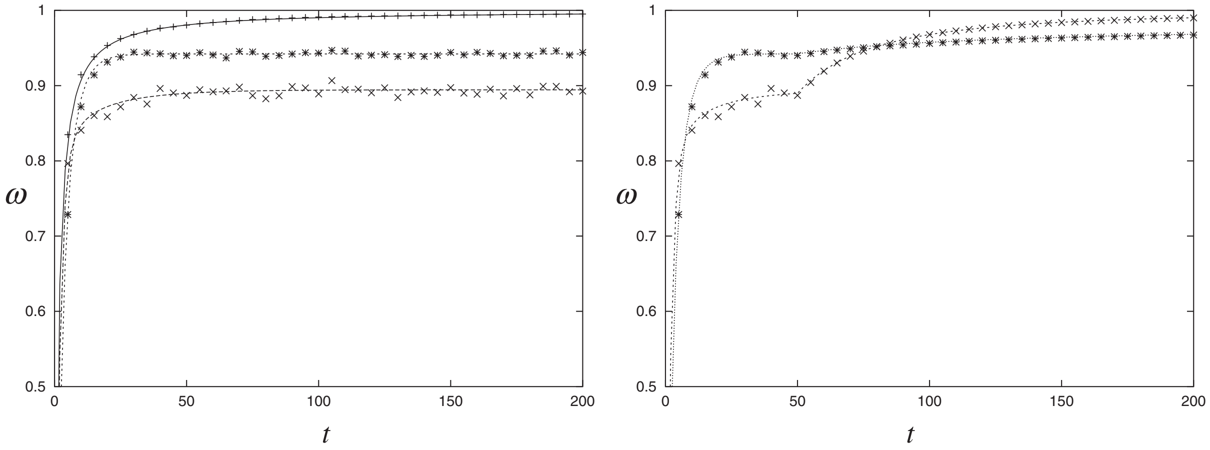

Fig. 4. Time dependence of ω in input noise model. $\sigma = 0.5$. Left panel: Constant learning rate. $\eta = 1$. Theoretical results (RKG); solid curve: Hebbian, dashed curve: Perceptron, dotted curve: AdaTron. Simulations ($N = 1000$); $+$: Hebbian, $\times$: Perceptron, $*$: AdaTron. Right panel: Optimal learning rate. For $t < 50$, $\eta = 1$ and for $t \geq 50$, $\eta = \eta_{\mathrm{opt}}(t)$. Theoretical results (RKG); dashed curve: Perceptron, dotted curve: AdaTron. Simulations ($N = 1000$); $\times$: Perceptron, $*$: AdaTron.

Table III. Asymptotic form of optimal learning rate and $\widetilde{E}_{\mathrm{opt}}$ for $t \gg 1$ in input noise model for $\sigma > 0$.

	Hebbian	Perceptron	AdaTron
$\widetilde{\eta}_{\mathrm{opt}}(t)$	$\sqrt{\frac{\pi(1+\sigma^2)}{2}}t^{-1}$	$\sqrt{2\pi(1+\sigma^2)}t^{-1}$	$\frac{\pi(1+\sigma^2)}{\sigma}t^{-1}$
$\widetilde{E}_{\mathrm{opt}}$	$\frac{1+\sigma^2}{4\sigma}t^{-1}$	$\frac{1+\sigma^2}{2\pi\sigma}\cos^{-1}\left(\frac{1}{\sqrt{1+\sigma^2}}\right)t^{-1}$	$\frac{(1+\sigma^2)^2}{2\sigma^3}\left\{\cos^{-1}\left(\frac{1}{\sqrt{1+\sigma^2}}\right)-\frac{\sigma}{1+\sigma^2}\right\}t^{-1}$

The equation for ω does not include J. The factor

$$\left\{\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right)-\frac{\omega\sqrt{1+\sigma^2-\omega^2}}{1+\sigma^2}\right\}$$

in the equation for J is positive for $\sigma > 0$ and $0 \leq \omega \leq 1$. Thus, as $t \to \infty$, $J \to 0$ for $\eta < 2$, $J =$ constant for $\eta = 2$ and $J \to \infty$ for $\eta > 2$. $\omega \to \omega_A^*$ as $t \to \infty$. Here, ω_A^* is the solution of $\mathrm{d}\omega/\mathrm{d}t = 0$ and is less than 1. Thus, learning fails.

As shown in Fig. 4, in each learning rule the agreement between the simulation results and the theoretical ones is very good.

Since learning fails for the Perceptron and AdaTron rules, in order to improve the learning, we consider the optimal learning rate and the averaging method.

4.2 *Optimal learning rate*

The behaviors of ω and J in the the limit of $t \to \infty$ in the three rules are the same as in the case of the output noise model as shown in Table II. See Fig. 4. In Table III, asymptotic forms of optimal learning rate and $\widetilde{E}_{\mathrm{opt}}$ for $t \gg 1$ in the input noise model are shown.

From Table III, we note that the asymptotic form of $\widetilde{E}_{\mathrm{opt}}$ is proportional to t^{-1} for the three learning rules.

4.3 *Averaging method*

As in the output noise model, ω tends to ω^*, which is less than 1, in the Perceptron and AdaTron rules. Therefore, we take the time averages of $\boldsymbol{J}$ and $\widehat{\boldsymbol{J}}$ for the Perceptron and AdaTron rules, respectively. As shown in Fig. 5, ω for the averaged vector increases and seems to approach 1 as L increases. In Fig. 6, we display the dependence of the convergence of ω on the starting time t_1 when the average is taken. As seen in the right panel of Fig. 6, the $t - t_1$ dependences of ω are quite similar for $t_1 = 50$, 100, and 150, but the behaviors of ω are different for $t_1 = 0$, 5, 10, 15, and 25. The reason is the same as that in the case of the output noise model; that is, for $t_1 = 50$, 100, and 150, the student vector already starts to rotate around the teacher vector, whereas for $t_1 = 0$, 5, 10, 15, and 25, ω is still approaching ω_P^*, as seen in the left panel of Fig. 6.

We obtained similar results using the AdaTron rule. Furthermore, we can estimate σ from the relationship $|\langle\widehat{\boldsymbol{J}}_P\rangle| = \omega_P^*$ and $|\langle\widehat{\boldsymbol{J}}_A\rangle| = \omega_A^*$. When $\sigma = 0.5$, we estimate $\sigma = 0.498$ and $\sigma = 0.496$ in the Perceptron and AdaTron rules at $t = 1000$, respectively.

5. Two-Teacher Model

5.1 *Formulation of two-teacher model*

We consider the case in which signals are given by two teacher Perceptrons. Let $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ be the N-dimensional teacher vectors. For simplicity, we assume $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are orthogonal to each other and are normalized, $\boldsymbol{B}_1 \cdot \boldsymbol{B}_2 = 0$ and $|\boldsymbol{B}_1| = |\boldsymbol{B}_2| = 1$. Let $\boldsymbol{\xi}$ be an N-dimensional example vector. We assume that its component ξ_i takes ± 1 and is drawn independently with the probability $P(\xi = 1) = 1 - P(\xi = -1) = 1/2$. The output T_i of $\boldsymbol{B}_i$ for $\boldsymbol{\xi}$ is given by

$$T_i = \mathrm{sgn}(\boldsymbol{B}_i \cdot \boldsymbol{\xi}), \quad i = 1, 2. \tag{55}$$

Furthermore, we assume that the student receives a signal from $\boldsymbol{B}_1$ or $\boldsymbol{B}_2$ randomly. Let r_i be the probability that a

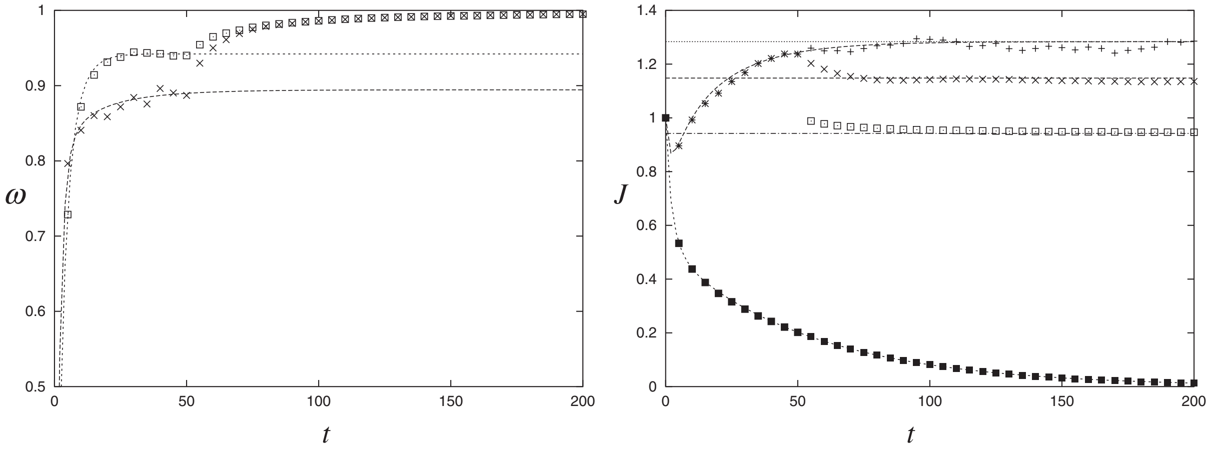

Fig. 5. Averaging method for Perceptron and AdaTron rules in input noise model. $\sigma = 0.5$. The average is taken for $t \geq 50$. Symbols denote the simulation data for $N = 1000$. ×: Perceptron (not normalized), square: AdaTron (normalized). Curves denote the theoretical results for $\eta = 1$ without averaging. Dashed: Perceptron, dotted: AdaTron. Left panel: Time dependence of ω. Right panel: Time dependence of J. Data without averaging for $\eta = 1$ are also depicted. +: Perceptron, closed square: AdaTron. The theoretical results for J_P^* (dotted line) and $J_P^*\omega_P^*$ (dashed line) and ω_A^* (dashed-dotted line) are depicted in the right panel.

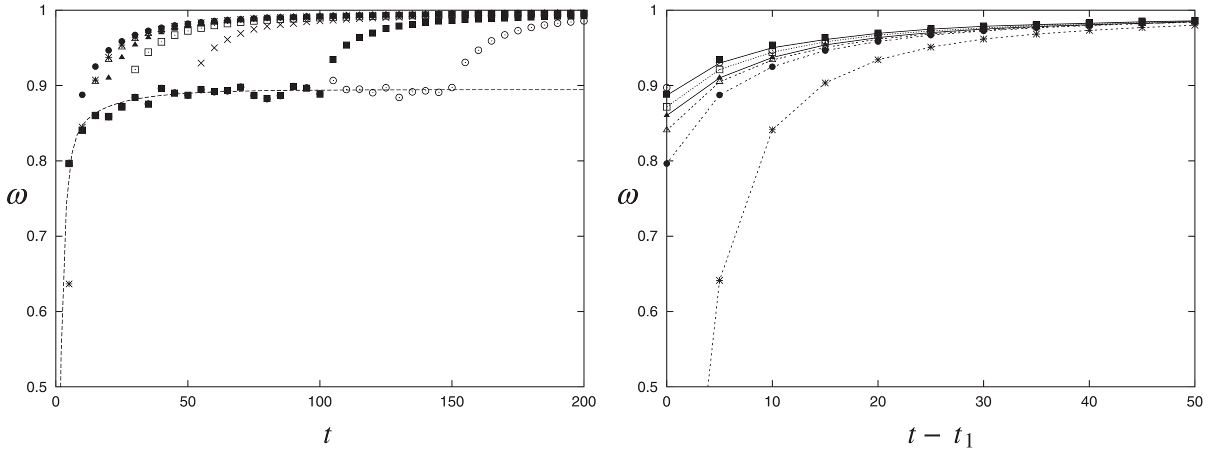

Fig. 6. Averaging method for Perceptron rule in input noise model. $\sigma = 0.5$. t_1 dependence of convergence. Symbols denote simulation data for $N = 1000$. Left panel: t dependence of ω. $*$: $t_1 = 0$, closed circle: $t_1 = 5$, triangle: $t_1 = 10$, closed triangle: $t_1 = 15$, square: $t_1 = 25$, ×: $t_1 = 50$, closed square: $t_1 = 100$, circle: $t_1 = 150$. Right panel: $t - t_1$ dependence of ω. Each symbol corresponds to the same value of t_1 as in the left panel except for the case of $t_1 = 50$, in which data are connected by a solid line without symbols. Furthermore, for $t_1 = 0, 5, 10, 15$, and 25, symbols are connected by a dashed line.

signal is from the teacher $\boldsymbol{B}_i$ for $i = 1$ and 2. Then, $r_1 + r_2 = 1$ holds. Let $\boldsymbol{J}$ be the N-dimensional student vector. The output S of the student $\boldsymbol{J}$ for $\boldsymbol{\xi}$ is given by

$$S = \mathrm{sgn}(\boldsymbol{J} \cdot \boldsymbol{\xi}). \tag{56}$$

The learning algorithm is given by

$$\boldsymbol{J}\left(t + \frac{1}{N}\right) = \boldsymbol{J}(t) + \frac{1}{N}\eta\boldsymbol{\xi}T\mathcal{F}[|\boldsymbol{J}|; \boldsymbol{J}\cdot\boldsymbol{\xi}, T], \tag{57}$$

where η is the learning rate and $\mathcal{F}$ is the learning rule. The order parameters are $Q = \boldsymbol{J}^2$ and $R_i = \boldsymbol{J} \cdot \boldsymbol{B}_i$ $(i = 1, 2)$. The generalization error E is calculated as

$$E = \langle\Theta(-ST)\rangle = \frac{1}{\pi}[r\cos^{-1}\omega_1 + (1-r)\cos^{-1}\omega_2], \tag{58}$$

where $\omega_i = R_i/J$ $(i = 1, 2)$ with $J = |\boldsymbol{J}|$. We also obtain the differential equations for Q, R_1 and R_2 for each learning rule. In the following, we study the learning for each rule.

In the Hebbian rule, we get

$$\frac{\mathrm{d}R_1}{\mathrm{d}t} = \eta\sqrt{\frac{2}{\pi}}r, \tag{59}$$

$$\frac{\mathrm{d}R_2}{\mathrm{d}t} = \eta\sqrt{\frac{2}{\pi}}(1-r), \tag{60}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \eta\sqrt{\frac{2}{\pi}}\{r\omega_1 + (1-r)\omega_2\} + \frac{\eta^2}{2J}, \tag{61}$$

$$\frac{\mathrm{d}\omega_1}{\mathrm{d}t} = \frac{\eta}{J}\sqrt{\frac{2}{\pi}}\left[r(1-\omega_1^2) - (1-r)\omega_1\omega_2\right] - \frac{\omega_1\eta^2}{2J^2}, \tag{62}$$

$$\frac{\mathrm{d}\omega_2}{\mathrm{d}t} = \frac{\eta}{J}\sqrt{\frac{2}{\pi}}\left[(1-r)(1-\omega_2^2) - r\omega_1\omega_2\right] - \frac{\omega_2\eta^2}{2J^2}, \tag{63}$$

where $r = r_1$. Defining $\Omega = r\omega_1 + (1-r)\omega_2$ and $R = J\Omega$, we obtain

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta\sqrt{\frac{2}{\pi}}\Omega_H^*, \tag{64}$$

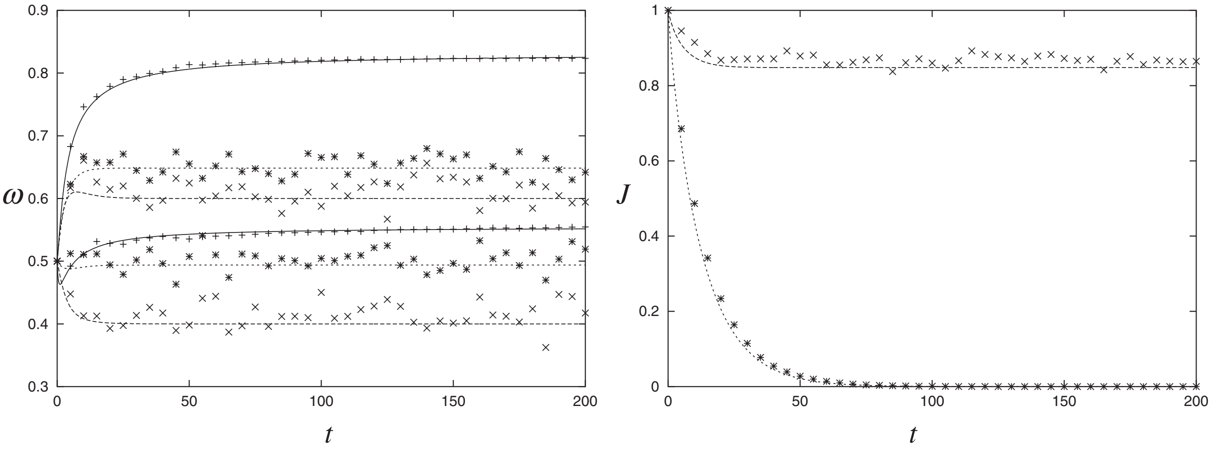

Fig. 7. Time dependence of ω_1, ω_2 (left) and J (right) for three rules. $r = 0.6$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and $+$: Hebbian, dashed curve and $\times$: Perceptron, dotted curve and $*$: AdaTron.

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 2\eta\sqrt{\frac{2}{\pi}}R + \eta^2, \tag{65}$$

where $\Omega_H^* = \sqrt{r^2 + (1-r)^2}$. These equations are solved analytically and solutions with $R_1(0) = R_2(0) = 0$ and $Q(0) = 1$ are

$$R_1 = \eta\sqrt{\frac{2}{\pi}}rt, \tag{66}$$

$$R_2 = \eta\sqrt{\frac{2}{\pi}}(1-r)t, \tag{67}$$

$$R = \eta\sqrt{\frac{2}{\pi}}\Omega_H^* t, \tag{68}$$

$$Q = \eta^2\frac{2}{\pi}\,\Omega_H^{*}t^2 + \eta^2 t + 1, \tag{69}$$

$$J = \sqrt{\eta^2\frac{2}{\pi}\,\Omega_H^{*}t^2 + \eta^2 t + 1}, \tag{70}$$

$$\omega_1 = \frac{\eta\sqrt{\frac{2}{\pi}}rt}{\sqrt{\eta^2\frac{2}{\pi}\Omega_H^* t^2 + \eta^2 t + 1}}, \tag{71}$$

$$\omega_2 = \frac{\eta\sqrt{\frac{2}{\pi}}(1-r)t}{\sqrt{\eta^2\frac{2}{\pi}\Omega_H^* t^2 + \eta^2 t + 1}}. \tag{72}$$

Thus, the generalization error is given by

$$E = \frac{1}{\pi}\left[r\cos^{-1}\left(\frac{\eta\sqrt{\frac{2}{\pi}}rt}{\sqrt{\eta^2\frac{2}{\pi}\Omega_H^* t^2 + \eta^2 t + 1}} \right) + (1-r)\cos^{-1}\left(\frac{\eta\sqrt{\frac{2}{\pi}}(1-r)t}{\sqrt{\eta^2\frac{2}{\pi}\Omega_H^* t^2 + \eta^2 t + 1}} \right) \right]. \tag{73}$$

Furthermore, we obtain $J \to \infty$, $\omega_1 \to \omega_{H,1}^*$, $\omega_2 \to \omega_{H,2}^*$, $\Omega \to \Omega_H^*$ and $E \to E_H^* = (1/\pi)[r\cos^{-1}\omega_{H,1}^* + (1-r)\cos^{-1}\omega_{H,2}^*]$ as $t \to \infty$. Here, $\omega_{H,1}^*$ and $\omega_{H,2}^*$ are defined as

$$\begin{aligned} \omega_{H,1}^* &= \frac{r}{\sqrt{r^2 + (1-r)^2}} = \frac{r}{\Omega_H^*}, \\ \omega_{H,2}^* &= \frac{1-r}{\sqrt{r^2 + (1-r)^2}} = \frac{1-r}{\Omega_H^*}. \end{aligned} \tag{74}$$

Since $(\omega_{H,1}^*)^2 + (\omega_{H,2}^*)^2 = 1$, $\widehat{\boldsymbol{J}}\ (\equiv \boldsymbol{J}/J)$ tends to the plane Σ, which is spanned by $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. In Fig. 7, we display the numerical and theoretical results. From the figure, we note that the numerical results agree with the theoretical ones very well, although there exists a small fluctuation in the simulation because the student cannot learn from both teachers. To eliminate fluctuations, it is useful to take the time average of $\widehat{\boldsymbol{J}}$, $\langle\widehat{\boldsymbol{J}}\rangle$. We confirmed that this procedure really works and that the fluctuations are reduced. Using this method, we can obtain the vector $\langle\widehat{\boldsymbol{J}}\rangle$ that lies on Σ. We denote this using $\langle\widehat{\boldsymbol{J}}_H\rangle$. If we can find another vector on Σ independent of $\langle\widehat{\boldsymbol{J}}_H\rangle$ using the simulation, we can identify $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$.

In the Perceptron rule, we get

$$\frac{\mathrm{d}R_1}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}}(r - \omega_1), \tag{75}$$

$$\frac{\mathrm{d}R_2}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}}(1 - r - \omega_2), \tag{76}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}}[r\omega_1 + (1-r)\omega_2 - 1] + \frac{\eta^2}{2J}E, \tag{77}$$

$$\frac{\mathrm{d}\omega_1}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}J}\left[r(1-\omega_1^2) - (1-r)\omega_1\omega_2\right] - \frac{\eta^2\omega_1}{2J^2}E, \tag{78}$$

$$\frac{\mathrm{d}\omega_2}{\mathrm{d}t} = \frac{\eta}{\sqrt{2\pi}J}\left[(1-r)(1-\omega_2^2) - r\omega_1\omega_2\right] - \frac{\eta^2\omega_2}{2J^2}E. \tag{79}$$

The stationary states for ω_1 and ω_2 are obtained as

$$\omega_{P,1}^* = r, \quad \omega_{P,2}^* = 1 - r. \tag{80}$$

The stationary state for J, J_P^*, is given by

$$J_P^* = \frac{\sqrt{2\pi}\eta}{4r(1-r)}E(r), \tag{81}$$

where $E(r) = (1/\pi)\{r\cos^{-1}(r) + (1-r)\cos^{-1}(1-r)\}$ is the

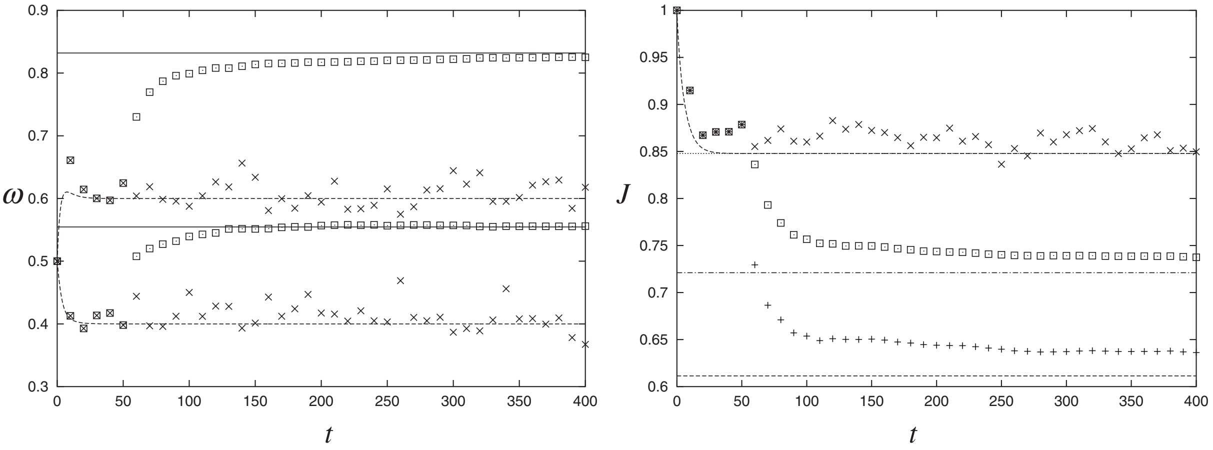

Fig. 8. Averaging method for Perceptron. $r = 0.6$. The average is taken for $t \geq 50$. Symbols denote simulation data for $N = 1000$. $+$: averaging (not normalized), square: averaging (normalized). Dashed curves and × denote the theoretical and numerical results for $\eta = 1$ without averaging. Left panel: Time dependence of ω_1 and ω_2. In this case, since results with and without normalization are the same, only the former is depicted. The theoretical results for $\omega^*_{H,1}$ and $\omega^*_{H,2}$ are also depicted (full line). Right panel: Time dependence of J. Theoretical results for J^*_P (dotted line), Ω^*_H (dashed-dotted line) and $J^*_P\Omega^*_H$ (dashed line) are depicted.

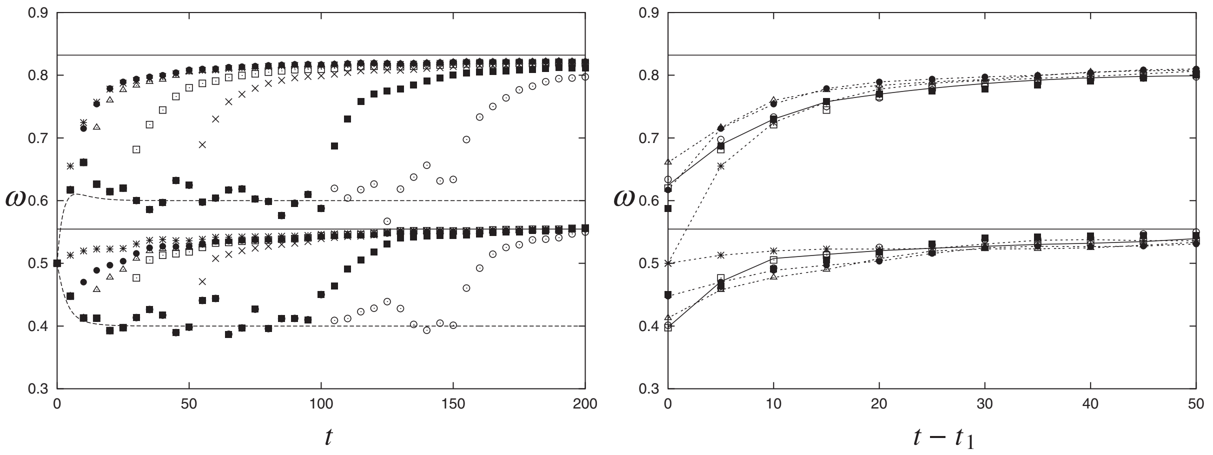

Fig. 9. Averaging method for Perceptron. $r = 0.6$. t_1 dependence of convergence. Symbols denote simulation data (normalized) for $N = 1000$. Left panel: t dependence of ω_1 and ω_2. $*$: $t_1 = 0$, closed circle: $t_1 = 5$, triangle: $t_1 = 10$, square: $t_1 = 25$, ×: $t_1 = 50$, closed square: $t_1 = 100$, circle: $t_1 = 150$. The theoretical results for $\omega^*_{H,1}$ and $\omega^*_{H,2}$ are also depicted (solid line). Right panel: $t - t_1$ dependence of ω_1 and ω_2. Each symbol corresponds to the same value of t_1 as in the left panel except for the case of $t_1 = 50$, in which data are connected by a solid line without symbols. Furthermore, for $t_1 = 0$, 5, and 10, symbols are connected by a dashed line.

generalization error for $t \to \infty$. As shown in Fig. 7, the theoretical and numerical results are in close agreement, although the fluctuation in the simulation is larger in this case than in the case of the Hebbian rule. This is because $\boldsymbol{J}$ does not approach the plain Σ, but rotates around both $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ with angles $\cos^{-1} r$ and $\cos^{-1}(1-r)$, respectively. Therefore, if we take the time average of $\boldsymbol{J}$, $\langle \boldsymbol{J} \rangle$ will converge in Σ and ω_i converges to $\omega^*_{H,i} = r_i/\sqrt{r_1^2 + r_2^2}$ for $i = 1, 2$ as the number of samples in the average increases.

Figure 8 shows this to be the case. In Fig. 9, we display the dependence of the convergence of ω_1 and ω_2 on the starting time t_1 when the average is taken. As seen in the right panel of Fig. 9, the $t - t_1$ dependences are quite similar for $t_1 = 25$, 50, 100, and 150 except for $t_1 = 0$, 5, and 10. This result is attributed to the fact that the student vector already starts to rotate around the teacher vectors for $t_1 = 25$, 50, 100, and 150, whereas for $t_1 = 0$, 5, and 10, ω_i is still approaching $\omega^*_{A,i}$. Furthermore, getting the value of $|\langle \widehat{\boldsymbol{J}} \rangle|$ from the simulation, we can numerically determine r. Let us denote $\langle \boldsymbol{J} \rangle$ and $\langle \widehat{\boldsymbol{J}} \rangle$ using $\langle \boldsymbol{J}_P \rangle$ and $\langle \widehat{\boldsymbol{J}}_P \rangle$, respectively. From the relationship $|\langle \widehat{\boldsymbol{J}}_P \rangle| = \Omega^*_H$, we estimated $r = 0.905$, 0.648, and 0.618, when $r = 0.9$, 0, 6, and 0.52, respectively. As r decreases, the discrepancy between the estimated r and the true value of r becomes larger. Since $\langle \boldsymbol{J}_P \rangle$ is proportional to $\langle \widehat{\boldsymbol{J}}_H \rangle$ obtained using the Hebbian rule, we need another vector independent of $\widehat{\boldsymbol{J}}_H$ on Σ in order to identify $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$.

In the AdaTron rule, we get

$$\frac{\mathrm{d}R_1}{\mathrm{d}t} = \eta J \left[\frac{r}{\pi} \sqrt{1 - \omega_1^2} - \omega_1 E \right], \tag{82}$$

$$\frac{\mathrm{d}R_2}{\mathrm{d}t} = \eta J \left[\frac{1-r}{\pi} \sqrt{1 - \omega_2^2} - \omega_2 E \right], \tag{83}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = -\eta J \left(1 - \frac{\eta}{2} \right) \times \left[E - \frac{r}{\pi} \omega_1 \sqrt{1 - \omega_1^2} - \frac{1-r}{\pi} \omega_2 \sqrt{1 - \omega_2^2} \right], \tag{84}$$

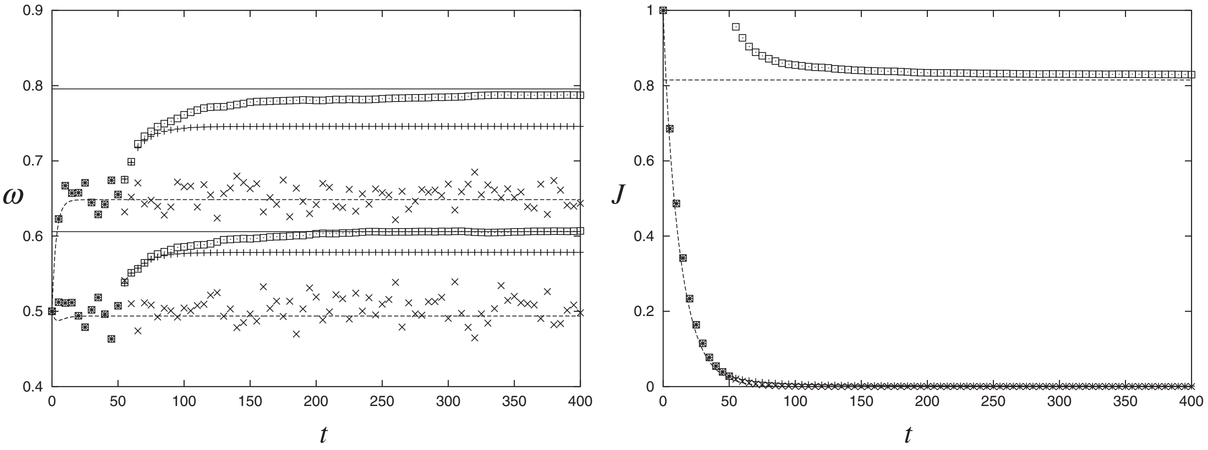

Fig. 10. Averaging method for AdaTron. $r = 0.6$. The average is taken for $t \geq 50$. Symbols denote simulation data for $N = 1000$. +: averaging (not normalized), square: averaging (normalized). Dashed curves and × denote the theoretical and numerical results for $\eta = 1$ without averaging. Left panel: Time dependence of ω_1 and ω_2. The theoretical results for $\tilde{\omega}^*_{A,1}$ and $\tilde{\omega}^*_{A,2}$ are also depicted. (solid line). Right panel: Time dependence of J. The theoretical result for $\sqrt{(\omega^*_{A,1})^2 + (\omega^*_{A,1})^2}$ is also depicted (dashed line).

$$\frac{\mathrm{d}\omega_1}{\mathrm{d}t} = -\frac{\eta^2}{2}\omega_1 E + \frac{r\eta}{\pi}\sqrt{1-\omega_1^2}\left[1 - \left(1-\frac{\eta}{2}\right)\omega_1^2\right] - \eta\left(1-\frac{\eta}{2}\right)\frac{1-r}{\pi}\omega_1\omega_2\sqrt{1-\omega_2^2}, \tag{85}$$

$$\frac{\mathrm{d}\omega_2}{\mathrm{d}t} = -\frac{\eta^2}{2}\omega_2 E + \frac{(1-r)\eta}{\pi}\sqrt{1-\omega_2^2}\left[1 - \left(1-\frac{\eta}{2}\right)\omega_2^2\right] - \eta\left(1-\frac{\eta}{2}\right)\frac{r}{\pi}\omega_1\omega_2\sqrt{1-\omega_1^2}. \tag{86}$$

For the stationary states of ω_1 and ω_2, we obtain the following relationship:

$$\frac{r\sqrt{1-{\omega^*_{A,1}}^2}}{\omega^*_{A,1}} = \frac{(1-r)\sqrt{1-{\omega^*_{A,2}}^2}}{\omega^*_{A,2}} = \rho. \tag{87}$$

From this, we get

$$\omega^*_{A,2} = \frac{(1-r)\omega^*_{A,1}}{\sqrt{r^2 + (1-2r){\omega^*_{A,1}}^2}}. \tag{88}$$

$\omega^*_{A,1}$ and $\omega^*_{A,2}$ are determined by eq. (87) and the following relationship:

$$\frac{\eta\pi}{2}E^*_A = \rho\left\{1 - \left(1-\frac{\eta}{2}\right)({\omega^*_{A,1}}^2 + {\omega^*_{A,2}}^2)\right\}, \tag{89}$$

where

$$E^*_A = \frac{1}{\pi}\{r\cos^{-1}(\omega^*_{A,1}) + (1-r)\cos^{-1}(\omega^*_{A,2})\}.$$

Furthermore, we obtain for $t \gg 1$

$$\frac{\mathrm{d}J}{\mathrm{d}t} \simeq -\left(1-\frac{\eta}{2}\right)J\frac{\rho}{\pi}(1 - {\omega^*_{A,1}}^2 - {\omega^*_{A,2}}^2). \tag{90}$$

It is proved that ${\omega^*_{A,1}}^2 + {\omega^*_{A,2}}^2 < 1$ for $0 < r < 1$. Thus, as $t \to \infty$, $J \to 0$ for $\eta < 2$, $J =$ constant for $\eta = 2$, and $J \to \infty$ for $\eta > 2$.

As shown in Fig. 10, the theoretical and numerical results agree very well, although the fluctuation in the simulation is larger in this case than in the case of the Hebbian rule. This is because $\boldsymbol{J}$ does not approach the plain Σ but rotates around both $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ with angles $\cos^{-1}\omega^*_{A,1}$ and $\cos^{-1}\omega^*_{A,2}$, respectively. We expect that the time average of $\widehat{\boldsymbol{J}}$, $\langle\widehat{\boldsymbol{J}_A}\rangle$, tends to the plain Σ as $t \to \infty$. From Fig. 10, it seems that $\boldsymbol{B}_i \cdot \langle\widehat{\boldsymbol{J}_A}\rangle$ tends to $\tilde{\omega}^*_{A,i}/\sqrt{(\tilde{\omega}^*_{A,1})^2 + (\tilde{\omega}^*_{A,2})^2}$ and $\hat{J}_A$ tends to $\sqrt{(\tilde{\omega}^*_{A,1})^2 + (\tilde{\omega}^*_{A,2})^2}$. However, it turned out that $\langle\widehat{\boldsymbol{J}_A}\rangle$ did not converge in Σ. Therefore, it is difficult to identify $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ numerically.

As for the dependence of the convergence of ω_1 and ω_2 on the time t_1, we obtained the same result as that in the Perceptron rule. That is, when t_1 is larger than the time when the student vector starts to rotate around $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, the $t - t_1$ dependences of ω_i are quite similar. See Fig. 11.

6. Summary and Discussion

First, we summarize the results of the single-teacher case. We studied the output and the input noise models using the Hebbian, Perceptron and AdaTron learning rules. Since we obtained almost the same results in the output and the input noise models, the following summary is for both cases unless otherwise mentioned explicitly. In the Hebbian rule, it has been found in a previous study[13] that learning succeeds in the sense that the student vector converges to the teacher vector even if noise exists. On the other hand, in the Perceptron and AdaTron rules, learning fails, but using the optimal learning rate, we proved that $\omega \to 1$ as $t \to \infty$ in the three learning rules. In the Perceptron and AdaTron rules, we found that ω converges to a value less than 1 as $t \to \infty$. This implies that the student vector rotates around the teacher vector with a constant angle. Thus, by taking the average over time, we expected that the direction of the student vector would converge to that of the teacher vector. The numerical results supported this speculation. Furthermore, using the averaging method, we estimated the parameters that characterize noise: k in the output noise and σ in the input noise. Furthermore, we studied the starting time (t_1) dependence of the convergence of learning. We found that the behaviors of ω are quite similar when t_1 is larger than the time when the student vector starts to rotate around the teacher vector. We found that the longer the learning proceeds and the larger the number of samples in

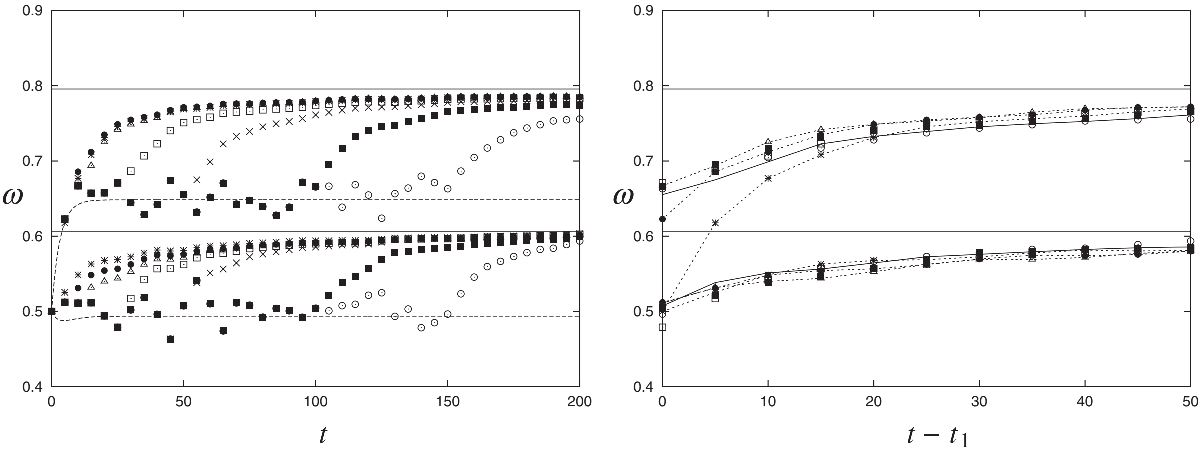

Fig. 11. Averaging method for AdaTron. $r = 0.6$. t_1 dependence of convergence. Symbols denote simulation data (normalized) for $N = 1000$. Left panel: t dependence of ω_1 and ω_2. $*$: $t_1 = 0$, closed circle: $t_1 = 5$, triangle: $t_1 = 10$, square: $t_1 = 25$, ×: $t_1 = 50$, closed square: $t_1 = 100$, circle: $t_1 = 150$. The theoretical results for $\tilde{\omega}^*_{A,1}$ and $\tilde{\omega}^*_{A,2}$ are also depicted (solid line). Right panel: $t - t_1$ dependence of ω_1 and ω_2. Each symbol corresponds to the same value of t_1 as in the left panel except for the case of $t_1 = 50$, in which data are connected by a solid line without symbols. Furthermore, for $t_1 = 0$, 5, and 10, symbols are connected by a dashed line.

the average becomes, the closer ω approaches 1 and the better the estimate of the parameter becomes. In conclusion, the teacher vector $\boldsymbol{B}$ and the noise parameters k and σ can be identified using these methods. As for the asymptotic decay of the generalization error, we found the asymptotic form of $\widetilde{E} = E - E_{\min}$ using the optimal learning rate for the output and input noise models and for the three learning rules. In the output noise model, $\widetilde{E}_{\rm opt} \propto t^{-1/2}$ for the Hebbian and Perceptron rules, whereas $\widetilde{E}_{\rm opt} \propto t^{-1/4}$ for the AdaTron rule. On the other hand, in the input noise model, we obtained $\widetilde{E}_{\rm opt} \propto t^{-1}$ for the three rules.

Next, let us summarize the results of the two-teacher case. We studied a situation where signals are given by two teachers $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ with a definite probability. We adopted the Hebbian, Perceptron and AdaTron learning rules. For the Hebbian rule, we obtained the analytical solutions for order parameters and the generalization error. The student vector converges to the space Σ spanned by $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. On the other hand, for the Perceptron and AdaTron rules, it turned out that the normalized student vector $\widehat{\boldsymbol{J}}$ did not converge to Σ. As in the single-teacher case, we expected that by taking the average of $\widehat{\boldsymbol{J}}$ over time, the averaged vector $\langle\widehat{\boldsymbol{J}}\rangle$ would converge to Σ, and since $\langle\widehat{\boldsymbol{J}}_P\rangle$ and $\langle\widehat{\boldsymbol{J}}_A\rangle$ were theoretically expected to converge to different vectors on Σ, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ could be identified by these vectors. Indeed, we found that $\langle\widehat{\boldsymbol{J}}_P\rangle$ converges to Σ as the number of samples in the average increases. Using this vector, we could identify the probability that signals are sent by two teachers. On the other hand, it turned out that $\langle\widehat{\boldsymbol{J}}_A\rangle$ does not converge to Σ, although $\boldsymbol{B}_i \cdot \langle\widehat{\boldsymbol{J}}_A\rangle$ and $|\langle\widehat{\boldsymbol{J}}_A\rangle|$ seem to converge to the expected values, respectively. The reason that $\langle\widehat{\boldsymbol{J}}_A\rangle$ does not converge to Σ is considered to be due to the fact that the fluctuation of $\langle\widehat{\boldsymbol{J}}_A\rangle$ might be large and not uniform in the orthogonal complement of Σ.

As for the starting time (t_1) dependence of the convergence of learning, as in the single-teacher case, we found that the behavior of ω_i is quite similar when t_1 is larger than the time when the student vector starts to rotate around the teacher vectors.

Next, let us discuss the results in this paper.

We compare the convergence speed of learning in the single-teacher case. If noise does not exist, the asymptotic form of $\widetilde{E}_{\rm opt}$ is expressed as $\widetilde{E}_{\rm opt} \propto t^{-1/2}$ for the Hebbian rule and $\widetilde{E}_{\rm opt} \propto t^{-1}$ for the Perceptron and AdaTron rules, so the convergence speed of learning is faster in the Perceptron and AdaTron rules than in the Hebbian rule. On the other hand, in the output noise case, the convergence speed of learning is faster in the Hebbian and Perceptron rules than in the AdaTron rule, whereas in the input noise case, it is of the same order for all three rules. That is, the convergence speed of learning depends on whether or not noise exists, and also on the type of noise.

In this paper, we studied the single-teacher and two-teacher cases. We can also consider a many-teacher case. Let us assume that there are n teachers and a signal is produced by the i-th teacher with a probability r_i. For simplicity, let us assume that the norm of the teacher vectors is 1 and that any two teacher vectors are orthogonal to each other. Then, we can prove that the student vector tends to the space Σ spanned by n teachers as $t \to \infty$ for the Hebbian rule.

We studied the averaging method numerically. It is desirable to also study this method theoretically. Recently, a theoretical study of the averaging method in the learning of a linear Perceptron in the presence of noise has been performed.[14] Extending theories about nonlinear Perceptrons will be an interesting subject.

Appendix: Proof of $\omega \to 1$ as $t \to \infty$ for $\tilde{\eta} = \tilde{\eta}_{\rm opt}$

The equation for ω in the common form for the three learning rules is as follows:

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = a\tilde{\eta} - \frac{1}{2}b\tilde{\eta}^2. \tag{A·1}$$

See Table A·I. $\tilde{\eta}_{\rm opt}$ is determined by $(\partial/\partial\tilde{\eta})(\mathrm{d}\omega/\mathrm{d}t) = 0$. Thus, we obtain

$$\tilde{\eta} = \frac{a}{b}, \tag{A·2}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{a^2}{2b}. \tag{A·3}$$

Table A·I. a and b for each learning rule in output and input noise models.

Output noise model	Hebbian	Perceptron	AdaTron
a	$k\sqrt{\frac{2}{\pi}}(1-\omega^2)$	$\frac{k}{\sqrt{2\pi}}(1-\omega^2)$	$\frac{k}{\pi}(1-\omega^2)^{3/2}$
b	ω	$\omega\left(\frac{1}{2}-\frac{k}{\pi}\sin^{-1}\omega\right)$	$\omega\left[\frac{1}{2}-\frac{k}{\pi}\left(\omega\sqrt{1-\omega^2}+\sin^{-1}\omega\right)\right]$
Input noise model	Hebbian	Perceptron	AdaTron
a	$\sqrt{\frac{2}{\pi(1+\sigma^2)}}(1-\omega^2)$	$\frac{1}{\sqrt{2\pi(1+\sigma^2)}}(1-\omega^2)$	$\frac{\sqrt{1+\sigma^2-\omega^2}}{\pi(1+\sigma^2)}(1-\omega^2)$
b	ω	$\frac{\omega}{2\pi}\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right)$	$\frac{\omega}{\pi}\left[\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right)-\frac{\sqrt{1+\sigma^2-\omega^2}}{1+\sigma^2}\omega\right]$

Therefore, if $b \neq 0$, $d\omega/dt = 0$ is equivalent to $a = 0$. $b > 0$ is easily proved for $0 < \omega \leq 1$ for input and output noise models. Since $a \propto 1 - \omega^2$, we obtain the unique stationary state $\omega^* = 1$.

1) F. Rosenblatt: *Principles of Neurodynamics* (Spartan, New York, 1962).

2) D. O. Hebb: *The Organization of Behavior* (Wiley, New York, 1949).

3) J. K. Anlauf and M. Biehl: Europhys. Lett. **10** (1989) 687.

4) W. Kinzel and M. Opper: in *Physics of Neural Networks*, ed. J. L. van Hemmen, E. Domany and K. Schulten (Springer-Verlag, New York, 1991) Vol. 1, Chap. 4.

5) T. L. H. Watkin, A. Rau and M. Biehl: Rev. Mod. Phys. **65** (1993) 499.

6) O. Kinouchi and N. Caticha: J. Phys. A **25** (1992) 6243.

7) O. Kinouchi and N. Caticha: J. Phys. A **26** (1993) 6161.

8) C. W. H. Mace and A. C. C. Coolen: Stat. Comput. **8** (1998) 55.

9) *On-line Learning in Neural Networks*, ed. D. Saad (Cambridge University Press, Cambridge, 2001).

10) A. Engel and C. Van den Broeck: *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

11) Y. Maeda: Masters Thesis, Graduate School of Humanities and Sciences, Nara Women's University, Nara (2002) [in Japanese].

12) G. Reents and R. Urbanczik: Phys. Rev. Lett. **80** (1998) 5445.

13) M. Biehl, P. Riegler and M. Stechert: Phys. Rev. E **52** (1995) R4624.

14) S. Miyoshi, T. Uezu and M. Okada: J. Phys. Soc. Jpn. **75** (2006) 084007.