Journal of the Physical Society of Japan
Vol. 79, No. 9, September, 2010, 094003

Online Learning of Perceptron from Noisy Data: A Case in which Both Student and Teacher Suffer from External Noise

Tatsuya UEZU*, Sachi YAMAGUCHI†, Mika YOSHIDA‡, and Mami TOMIYASU[1]

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

(Received April 20, 2010; accepted July 1, 2010; published September 10, 2010)

We analyze the online learning of a Perceptron (student) from signals produced by a single Perceptron (teacher) in which both the student and the teacher suffer from external noise. We adopt three typical learning rules and treat the input and output noises. In order to improve learning when it fails in the sense that the student vector does not converge to the teacher vector, we use a method based on the optimal learning rate. Furthermore, in order to control learning, we propose a concrete method for the Perceptron rule in the output noise model. Finally, we analyze time domain ensemble online learning. The theoretical results agree quite well with the numerical simulation results.

KEYWORDS: perceptron, online learning, generalization error, noise, optimal learning rate, control of learning, time domain ensemble learning
DOI: 10.1143/JPSJ.79.094003

1. Introduction

We study the online learning of a single Perceptron[1)] from signals produced by a single teacher. We assume that both the teacher and the student suffer from external noise, and we adopt the Hebbian,[2)] Perceptron[1)] and AdaTron[3)] rules as learning rules.[4)] In our previous paper, we studied a similar system in which only the teacher suffers from external noise.[5,6)] There have been many other studies that focus on the case of a single teacher.[7–12)] The main purpose of the present study is to improve learning when it fails in the sense that the student vector does not converge to the teacher vector.[13)] The results are as follows: When learning fails, the teacher can be identified using the optimal learning rate. We can obtain the asymptotic form of the generalization error using the optimal learning rate for the three learning rules. Furthermore, in order to control learning, we propose a concrete method for the Perceptron rule in the output noise model. Finally, we analyze time domain ensemble learning and derive the formula of the direction cosine between the teacher vector and the averaged student vector.

The paper is organized as follows: In §2, the formulation for online learning is given. In §3 and §4, the learning with a constant learning rate and that with an optimal learning rate are analyzed, respectively. In §5, we study the control of learning, and in §6, we analyze the time domain ensemble learning. Section 7 is devoted to the summary and discussion. In the appendix, we list useful integration formulas in order to derive the differential equations for order parameters.

2. Formulation

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 $P(\xi = 1) = 1 - P(\xi = -1) = 1/2$. When there is no noise, 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 $\boldsymbol{B}$ 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 both the teacher and student noises exist. We assume that the teacher and student noises are independent. We consider the output and input noises. Let $\mathcal{P}_\mathrm{T}$ be the probability of $T = 1$ and $\mathcal{P}_\mathrm{S}$ be the probability of $S = 1$. Since $\mathcal{P}_\mathrm{T}$ and $\mathcal{P}_\mathrm{S}$ depend on $y = \boldsymbol{B} \cdot \boldsymbol{\xi}$ and $x = \hat{\boldsymbol{J}} \cdot \boldsymbol{\xi}$ in the present model, respectively, we denote them as $\mathcal{P}_\mathrm{T}(y)$ and $\mathcal{P}_\mathrm{S}(x)$. Here, $\hat{\boldsymbol{J}} = \boldsymbol{J}/J$, and $J = |\boldsymbol{J}|$ is the norm of $\boldsymbol{J}$. In the output noise model, these are given by

$$\mathcal{P}_\mathrm{T}(y) = \frac{1}{2}[1 + k_\mathrm{T}\,\mathrm{sgn}(y)], \tag{3}$$

$$\mathcal{P}_\mathrm{S}(x) = \frac{1}{2}[1 + k_\mathrm{S}\,\mathrm{sgn}(x)]. \tag{4}$$

In the input noise model, T and S are given by

$$T = \mathrm{sgn}[\boldsymbol{B} \cdot (\boldsymbol{\xi} + \boldsymbol{\zeta}^\mathrm{T})], \tag{5}$$

$$S = \mathrm{sgn}[\hat{\boldsymbol{J}} \cdot (\boldsymbol{\xi} + \boldsymbol{\zeta}^\mathrm{S})], \tag{6}$$

where $\boldsymbol{\zeta}^\mathrm{T}$ and $\boldsymbol{\zeta}^\mathrm{S}$ are the teacher and student noises, respectively. Each component ζ_i^T of $\boldsymbol{\zeta}^\mathrm{T}$ is assumed to be independently drawn from the Gaussian distribution of the mean 0 and the standard deviation σ_T, and each component ζ_i^S of $\boldsymbol{\zeta}^\mathrm{S}$ is assumed to be independently drawn from the Gaussian distribution of the mean 0 and the standard deviation σ_S. Then, $\mathcal{P}_\mathrm{T}$ and $\mathcal{P}_\mathrm{S}$ are expressed as

$$\mathcal{P}_\mathrm{T}(y) = H\left(-\frac{y}{\sigma_\mathrm{T}}\right), \tag{7}$$

$$\mathcal{P}_\mathrm{S}(x) = H\left(-\frac{x}{\sigma_\mathrm{S}}\right), \tag{8}$$

*E-mail: uezu@cc.nara-wu.ac.jp
†Present address: Department of Biology, Kyushu University.
‡Present address: Denso Information Technology Corporation.

where $H(y) = \int_y^\infty Du$ and $Du = \mathrm{d}u/\sqrt{2\pi}\exp(-u^2/2)$. We adopt the following learning algorithm for the output noise model

$$\boldsymbol{J}\left(t+\frac{1}{N}\right) = \boldsymbol{J}(t) + \frac{1}{N}\eta\boldsymbol{\xi}T\mathcal{F}, \tag{9}$$

and for the input noise model

$$\boldsymbol{J}\left(t+\frac{1}{N}\right) = \boldsymbol{J}(t) + \frac{1}{N}\eta(\boldsymbol{\xi}+\boldsymbol{\zeta}^{\mathrm{S}})T\mathcal{F}, \tag{10}$$

where η is the learning rate and $\mathcal{F}$ is the learning rule. In the latter case, the term $\boldsymbol{\xi}$ in the noiseless case is replaced by $\boldsymbol{\xi}+\boldsymbol{\zeta}^{\mathrm{S}}$, because the student receives examples which suffer from external noises.

We consider the following three learning rules:

$$\text{Hebbian rule: } \mathcal{F} = 1, \tag{11}$$

$$\text{Perceptron rule: } \mathcal{F} = \Theta(-TS), \tag{12}$$

$$\text{AdaTron rule:} \tag{13}$$

$$\mathcal{F} = |\boldsymbol{\xi}\cdot\boldsymbol{J}|\Theta(-TS) \text{ for the output noise model,} \tag{14}$$

$$\mathcal{F} = |(\boldsymbol{\xi}+\boldsymbol{\zeta}^{\mathrm{S}})\cdot\boldsymbol{J}|\Theta(-TS) \text{ for the input noise model,} \tag{15}$$

where $\Theta(x) = 1$ for $x \geq 0$ and $\Theta(x) = 0$ for $x < 0$. We define the order parameters $Q = \boldsymbol{J}^2$ and $R = \boldsymbol{J}\cdot\boldsymbol{B}$. In addition to Q and R, $J = \sqrt{Q}$ and $\omega = R/J$ are also used.

The generalization error ϵ_{g} is defined by

$$\epsilon_{\mathrm{g}} = \langle\Theta(-ST)\rangle_{\Xi}, \tag{16}$$

where $\langle\cdot\rangle_{\Xi}$ denotes the average over examples and noises. Now, let us consider a way of taking the average over examples and noises. As an example, let us consider a function of x, y, S, and T, $f = f(x,y,S,T)$. The average over noises $\langle f\rangle_{\mathrm{noise}}$ is taken using $\mathcal{P}_{\mathrm{T}}(y)$ and $\mathcal{P}_{\mathrm{S}}(x)$ as

$$\langle f\rangle_{\mathrm{noise}} = \sum_{S=\pm1}\sum_{T=\pm1}\mathcal{P}_{\mathrm{T}}(Ty)\mathcal{P}_{\mathrm{S}}(Sx)f(x,y,S,T), \tag{17}$$

since the student and teacher noises are assumed to be independent. Here, we use the fact that the probabilities of $S = \pm1$ and $T = \pm1$ are expressed as $\mathcal{P}_{\mathrm{S}}(Sx)$ and $\mathcal{P}_{\mathrm{T}}(Ty)$, respectively. See eqs. (3), (4), (7), and (8). The average over examples $\boldsymbol{\xi}$ is replaced by the average over x and y. Because ξs are independent random variables, $x = \sum_i \hat{J}_i\xi_i$ and $y = \sum_i B_i\xi_i$ are random variables and obey a Gaussian distribution by the central limit theorem. The means, variances, and covariance for x and y are calculated as $\langle x\rangle = 0$, $\langle y\rangle = 0$, $\langle x^2\rangle = 1$, $\langle y^2\rangle = 1$, and $\langle xy\rangle = \omega$. Thus, the Gaussian probability density for x and y, $P(x,y)$, is given as

$$P(x,y) = \frac{1}{2\pi\sqrt{1-\omega^2}} \times \exp\left[-\frac{1}{2(1-\omega^2)}(x^2+y^2-2\omega xy)\right]. \tag{18}$$

Therefore, $\langle f\rangle_{\Xi}$ is calculated using

$$\begin{aligned}\langle f\rangle_{\Xi} = \langle\langle f\rangle_{\mathrm{noise}}\rangle_{\mathrm{sample}} &= \int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}y\,P(x,y)\langle f\rangle_{\mathrm{noise}} \\ &= \int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}y\,P(x,y) \\ &\quad\times\sum_{S=\pm1}\sum_{T=\pm1}\mathcal{P}_{\mathrm{T}}(Ty)\mathcal{P}_{\mathrm{S}}(Sx)f(x,y,S,T).\end{aligned} \tag{19}$$

Then, ϵ_{g} is calculated using

$$\begin{aligned}\epsilon_{\mathrm{g}} &= \langle\Theta(-ST)\rangle_{\Xi} \\ &= \int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}y\,P(x,y)\{\mathcal{P}_{\mathrm{T}}(y)[1-\mathcal{P}_{\mathrm{S}}(x)] \\ &\quad+[1-\mathcal{P}_{\mathrm{T}}(y)]\mathcal{P}_{\mathrm{S}}(x)\}.\end{aligned} \tag{20}$$

In the next section, for a constant learning rate, we derive the differential equations for order parameters for both the output and input noise models, and compare the theoretical and numerical results.

3. Constant Learning Rate

3.1 Output noise model

The generalization error ϵ_{g} is calculated as

$$\epsilon_{\mathrm{g}} = \frac{1-k_{\mathrm{T}}k_{\mathrm{S}}}{2} + \frac{k_{\mathrm{T}}k_{\mathrm{S}}}{\pi}\cos^{-1}(\omega). \tag{21}$$

From eq. (9), we obtain the differential equations for Q and R:[10]

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 2\eta\langle(\boldsymbol{J}\cdot\boldsymbol{\xi})T\mathcal{F}\rangle_{\Xi} + \eta^2\langle\mathcal{F}^2\rangle_{\Xi}, \tag{22}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta\langle(\boldsymbol{B}\cdot\boldsymbol{\xi})T\mathcal{F}\rangle_{\Xi}. \tag{23}$$

Here, we assume self-averaging.[14] Since $\mathcal{F}$ is expressed as $\mathcal{F}[J,x,T,S]$, these equations are rewritten as

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 2\eta J\langle xT\mathcal{F}[J,x,T,S]\rangle_{\Xi} + \eta^2\langle\mathcal{F}^2[J,x,T,S]\rangle_{\Xi}, \tag{24}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta\langle yT\mathcal{F}[J,x,T,S]\rangle_{\Xi}. \tag{25}$$

The equations for J and ω are

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \eta\langle xT\mathcal{F}[J,x,T,S]\rangle_{\Xi} + \frac{\eta^2}{2J}\langle\mathcal{F}^2[J,x,T,S]\rangle_{\Xi}, \tag{26}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\eta}{J}\langle(y-\omega x)T\mathcal{F}[J,x,T,S]\rangle_{\Xi} - \frac{\omega\eta^2}{2J^2}\langle\mathcal{F}^2[J,x,T,S]\rangle_{\Xi}. \tag{27}$$

The average over noises and examples is calculated using $\mathcal{P}_{\mathrm{S}}(x)$, $\mathcal{P}_{\mathrm{T}}(y)$, and $P(x,y)$. By performing the average over x and y, we get equations for Q, R, J, and ω.

In the Hebbian rule, we get the differential equations for order parameters as

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \eta k_{\mathrm{T}}\sqrt{\frac{2}{\pi}}, \tag{28}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = 2J\eta k_{\mathrm{T}}\sqrt{\frac{2}{\pi}}\omega + \eta^2, \tag{29}$$

$$\frac{\mathrm{d}J}{\mathrm{d}t} = \eta k_{\mathrm{T}}\sqrt{\frac{2}{\pi}}\omega + \frac{\eta^2}{2J}, \tag{30}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\eta k_{\mathrm{T}}}{J}\sqrt{\frac{2}{\pi}}(1-\omega^2) - \frac{\omega\eta^2}{2J^2}. \tag{31}$$

These are the same as those in the case in which only the teacher suffers from noise. This case has been studied and the differential equations have been solved analytically.[15]

In the Perceptron rule, we get the differential equations for order parameters as

$$\frac{dR}{dt} = \frac{\eta}{\sqrt{2\pi}}(k_T - \omega k_S), \tag{32}$$

$$\frac{dQ}{dt} = \frac{2\eta J}{\sqrt{2\pi}}(\omega k_T - k_S) + \eta^2 \epsilon_g, \tag{33}$$

$$\frac{dJ}{dt} = \frac{\eta}{\sqrt{2\pi}}(\omega k_T - k_S) + \frac{\eta^2}{2J}\epsilon_g, \tag{34}$$

$$\frac{d\omega}{dt} = \frac{\eta k_T}{\sqrt{2\pi}J}(1-\omega^2) - \frac{\omega\eta^2}{2J^2}\epsilon_g. \tag{35}$$

In the AdaTron rule, the equations for order parameters are given as

$$\frac{dR}{dt} = \frac{\eta J}{2}\left[(k_T - k_S)\omega + \frac{2}{\pi}k_T\sqrt{1-\omega^2} - \frac{2k_T}{\pi}\omega\cos^{-1}(\omega)\right], \tag{36}$$

$$\frac{dQ}{dt} = \eta\left(\frac{\eta}{2} - k_S\right)J^2 + \eta J^2 k_T(2-\eta k_S) \times \left[\frac{1}{2} - \frac{1}{\pi}\cos^{-1}(\omega) + \frac{\omega}{\pi}\sqrt{1-\omega^2}\right], \tag{37}$$

$$\frac{dJ}{dt} = \frac{\eta}{2}\left(\frac{\eta}{2} - k_S\right)J + \frac{\eta J}{2}k_T(2-\eta k_S) \times \left[\frac{1}{2} - \frac{1}{\pi}\cos^{-1}(\omega) + \frac{\omega}{\pi}\sqrt{1-\omega^2}\right], \tag{38}$$

$$\frac{d\omega}{dt} = \frac{\eta k_T}{\pi}(1-\omega^2)^{3/2} - \frac{\eta^2\omega}{2}\left(\epsilon_g - k_T k_S \frac{\omega}{\pi}\sqrt{1-\omega^2}\right). \tag{39}$$

3.2 *Input noise model*

The generalization error ϵ_g is calculated as

$$\epsilon_g = \frac{1}{\pi}\cos^{-1}\left[\frac{\omega}{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)}}\right]. \tag{40}$$

From eq. (10), we obtain the differential equations for Q and R:[10]

$$\frac{dQ}{dt} = 2\eta\langle(\boldsymbol{J}\cdot(\boldsymbol{\xi}+\zeta^S)T\mathcal{F}\rangle_\Xi + \frac{1}{N}\eta^2\langle(\boldsymbol{\xi}+\boldsymbol{\zeta}^S)^2\mathcal{F}^2\rangle_\Xi, \tag{41}$$

$$\frac{dR}{dt} = \eta\langle(\boldsymbol{B}\cdot(\boldsymbol{\xi}+\boldsymbol{\zeta}^S)T\mathcal{F}\rangle_\Xi. \tag{42}$$

Here, we assume self-averaging. Now, $\mathcal{F}$ depends on J, x, T, S, and v, where $v = \hat{\boldsymbol{J}}\cdot\boldsymbol{\zeta}^S$. Since $S = \mathrm{sgn}[J(x+v)] = \mathrm{sgn}(x+v)$, $\mathcal{F}$ is expressed as $\mathcal{F}[J,x,v,T]$. The factor $(\boldsymbol{\xi}+\boldsymbol{\zeta}^S)^2$ in the second term on the right-hand side of eq. (41) can be calculated as

$$(\boldsymbol{\xi}+\boldsymbol{\zeta}^S)^2 = N + \sum_i(2\boldsymbol{\xi}_i\cdot\boldsymbol{\zeta}_i^S + (\boldsymbol{\zeta}_i^S)^2) = N + \mathcal{O}(N^0) + N\sigma_S^2, \tag{43}$$

where we used the fact that N is very large, and $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}^S$ are statistically independent.[16] Then, the above equations are rewritten as

$$\frac{dQ}{dt} = 2\eta J\langle(x+v)T\mathcal{F}[J,x,v,T]\rangle_\Xi + \eta^2(1+\sigma_S^2)\langle\mathcal{F}^2[J,x,v,T]\rangle_\Xi, \tag{44}$$

$$\frac{dR}{dt} = \eta\langle(y+u)T\mathcal{F}[J,x,v,T]\rangle_\Xi, \tag{45}$$

where $u = \boldsymbol{B}\cdot\boldsymbol{\zeta}^S$. The equations for $J = \sqrt{Q}$ and $\omega = R/J$ are

$$\frac{dJ}{dt} = \eta\langle(x+v)T\mathcal{F}[J,x,v,T]\rangle_\Xi + \frac{\eta^2}{2J}(1+\sigma_S^2)\langle\mathcal{F}^2[J,x,v,T]\rangle_\Xi, \tag{46}$$

$$\frac{d\omega}{dt} = \frac{\eta}{J}\langle[y+u-\omega(x+v)]T\mathcal{F}[J,x,v,T]\rangle_\Xi - \frac{\omega\eta^2}{2J^2}(1+\sigma_S^2)\langle\mathcal{F}^2[J,x,v,T]\rangle_\Xi. \tag{47}$$

The average over the teacher noise $\boldsymbol{\zeta}^T$ is taken independently of other averages and is calculated using the probability $\mathcal{P}_T(y)$. On the other hand, the average over the student noise $\boldsymbol{\zeta}^S$ of the quantity A is replaced with $\langle A\rangle_{u,v} \equiv \int_{-\infty}^{\infty}dx\int_{-\infty}^{\infty}dy\,P_2(u,v)A$. Here, the probability distribution $P_2(u,v)$ is given by the Gaussian distribution with $\langle u\rangle = 0$, $\langle v\rangle = 0$, $\langle u^2\rangle = \langle v^2\rangle = \sigma_S^2$ and $\langle uv\rangle = \omega\sigma_S^2$. The average over examples $\boldsymbol{\xi}$ is calculated using $P(x,y)$, as in the output noise case. By performing the average over examples and noises, we get equations for Q, R, J, and ω.

In the Hebbian rule, we get the differential equations for order parameters as

$$\frac{dR}{dt} = \eta\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{1+\sigma_T^2}}, \tag{48}$$

$$\frac{dQ}{dt} = 2\eta J\sqrt{\frac{2}{\pi}}\frac{\omega}{\sqrt{1+\sigma_T^2}} + \eta^2(1+\sigma_S^2), \tag{49}$$

$$\frac{dJ}{dt} = \eta\sqrt{\frac{2}{\pi}}\frac{\omega}{\sqrt{1+\sigma_T^2}} + \frac{\eta^2}{2J}(1+\sigma_S^2), \tag{50}$$

$$\frac{d\omega}{dt} = \frac{\eta}{J}\sqrt{\frac{2}{\pi}}\frac{1-\omega^2}{\sqrt{1+\sigma_T^2}} - \frac{\eta^2\omega}{2J^2}(1+\sigma_S^2). \tag{51}$$

In the Perceptron rule, we get the differential equations for order parameters as

$$\frac{dR}{dt} = \frac{\eta}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{1+\sigma_T^2}} - \omega\sqrt{1+\sigma_S^2}\right), \tag{52}$$

$$\frac{dQ}{dt} = \frac{2\eta J}{\sqrt{2\pi}}\left(\frac{\omega}{\sqrt{1+\sigma_T^2}} - \sqrt{1+\sigma_S^2}\right) + \eta^2(1+\sigma_S^2)\epsilon_g, \tag{53}$$

$$\frac{dJ}{dt} = \frac{\eta}{\sqrt{2\pi}}\left(\frac{\omega}{\sqrt{1+\sigma_T^2}} - \sqrt{1+\sigma_S^2}\right) + \frac{\eta^2}{2J}(1+\sigma_S^2)\epsilon_g, \tag{54}$$

$$\frac{d\omega}{dt} = \frac{\eta}{\sqrt{2\pi}J}\frac{1-\omega^2}{\sqrt{1+\sigma_T^2}} - \frac{\omega\eta^2}{2J^2}(1+\sigma_S^2)\epsilon_g. \tag{55}$$

In the AdaTron rule, the equations for order parameters are given as

$$\frac{dR}{dt} = \frac{\eta J}{\pi}\frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)-\omega^2}}{1+\sigma_T^2} - \eta J\omega(1+\sigma_S^2)\epsilon_g, \tag{56}$$

$$\frac{dQ}{dt} = 2\eta J^2\left[\frac{\eta}{2}(1+\sigma_S^2) - 1\right] \times \left[(1+\sigma_S^2)\epsilon_g - \frac{\omega}{\pi}\frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)-\omega^2}}{1+\sigma_T^2}\right], \tag{57}$$

$$\frac{dJ}{dt} = \eta J\left[\frac{\eta}{2}(1+\sigma_S^2) - 1\right]$$

Table I. Asymptotic values of J and ω.

Model	Learning rule	ω^*	J^*	
Output	Hebbian	1	∞	
	Perceptron	1	∞	$k_T > k_S$
		1	∞ ($J \simeq \sqrt{t}$)	$k_T = k_S$
		$\frac{k_T}{k_S}$	$\eta\sqrt{\frac{\pi}{2}}\frac{k_S}{k_S^2 - k_T^2}\epsilon_g$	$k_T < k_S$
	AdaTron	$0 < \omega^* < 1$	∞ or 0	
Input	Hebbian	1	∞	
	Perceptron	$\frac{1}{\sqrt{(1+\sigma_S^2)(1+\sigma_T^2)}}$	$\frac{\eta}{2}\frac{\sqrt{2\pi(1+\sigma_S^2)}}{1-[(1+\sigma_T^2)(1+\sigma_S^2)]^{-1}}$	
	AdaTron	$0 < \omega^* < 1$	0	$\sigma_S < \sqrt{\frac{2}{\eta} - 1}, (\eta < 2)$
		$0 < \omega^* < 1$	∞	$\sigma_S > \sqrt{\frac{2}{\eta} - 1}, (\eta < 2)$
		$0 < \omega^* < 1$	constant	$\sigma_S = \sqrt{\frac{2}{\eta} - 1}, (\eta < 2)$
		$0 < \omega^* < 1$	∞	$\eta \geq 2$

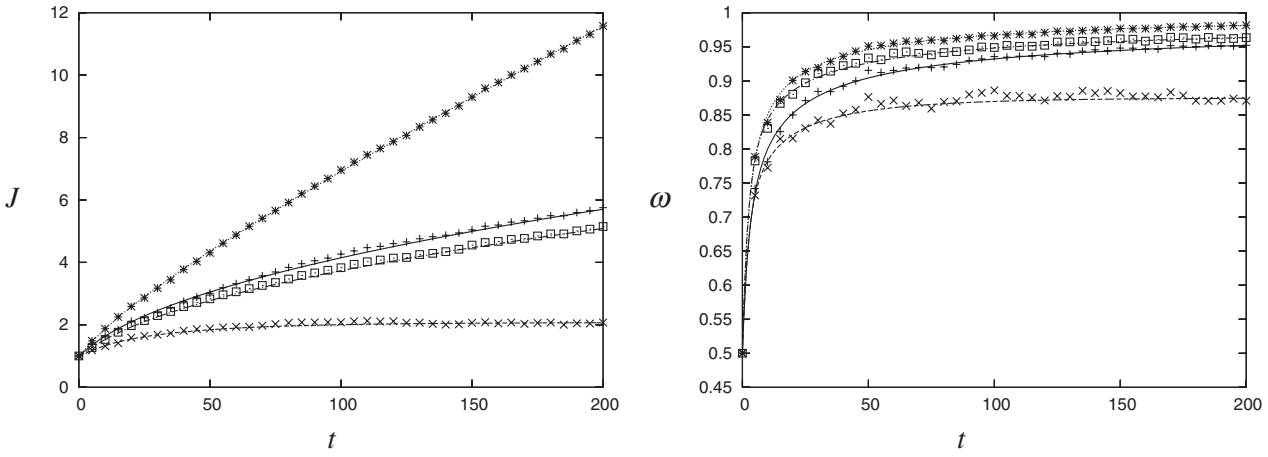

Fig. 1. Time dependences of J and ω for output noise model and Perceptron learning rule. $\eta = 1$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and +: $k_T = 0.7, k_S = 0.7$; dashed curve and ×: $k_T = 0.7, k_S = 0.8$; dotted curve and ∗: $k_T = 0.8, k_S = 0.7$; dotted-dashed curve and square: $k_T = 0.8, k_S = 0.8$. Left panel, J. Right panel, ω.

$$\times \left[(1+\sigma_S^2)\epsilon_g - \frac{\omega}{\pi}\frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2) - \omega^2}}{1+\sigma_T^2} \right], \quad (58)$$

$$\frac{d\omega}{dt} = \eta\left(\left\{ 1 + \omega^2 \left[\frac{\eta}{2}(1+\sigma_S^2) - 1 \right] \right\} \times \frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2) - \omega^2}}{\pi(1+\sigma_T^2)} - \frac{\eta\omega}{2}(1+\sigma_S^2)^2\epsilon_g \right). \quad (59)$$

The asymptotic values of J and ω are given in Table I.

3.3 *Numerical results*

In this subsection, we give the results of numerical integrations of differential equations by the Runge–Kutta–Gill (RKG) method and the results of numerical simulations.

In Figs. 1–4, we show the numerical and theoretical results for Perceptron and AdaTron rules in the output and input noise models. The agreements between the numerical and theoretical results are quite well.

For the Perceptron rule in the output noise model, learning succeeds for $k_T \geq k_S$, but fails for $k_T < k_S$. See Fig. 1. For other cases, learning always fails. See Figs. 2–4. In Fig. 4, we show the results of the three cases of $\sigma_S < \sqrt{(2/\eta) - 1}$, $\sigma_S = \sqrt{(2/\eta) - 1}$, and $\sigma_S > \sqrt{(2/\eta) - 1}$ and confirm that, as $t \to \infty$, J tends to 0, constant and ∞, respectively.

In order to improve learning in the case that learning fails, we consider the optimal learning rate firstly.

4. Optimal Learning Rate

We study the optimal learning rate η_{opt}, which is determined by the following relation:[10]

$$^\forall t \geq 0: \quad \frac{\partial}{\partial\tilde{\eta}}\left(\frac{d}{dt}\epsilon_g\right) = 0. \quad (60)$$

$\tilde{\eta}$ is η/J for the Hebbian and Perceptron rules, and η for the AdaTron rule. Since ϵ_g is a function of ω, the relationship is equivalent to

$$^\forall t \geq 0: \quad \frac{\partial}{\partial\tilde{\eta}}\left(\frac{d}{dt}\omega\right) = 0. \quad (61)$$

The differential equation for ω is expressed as

$$\frac{d}{dt}\omega = a\tilde{\eta} - \frac{b}{2}\tilde{\eta}^2. \quad (62)$$

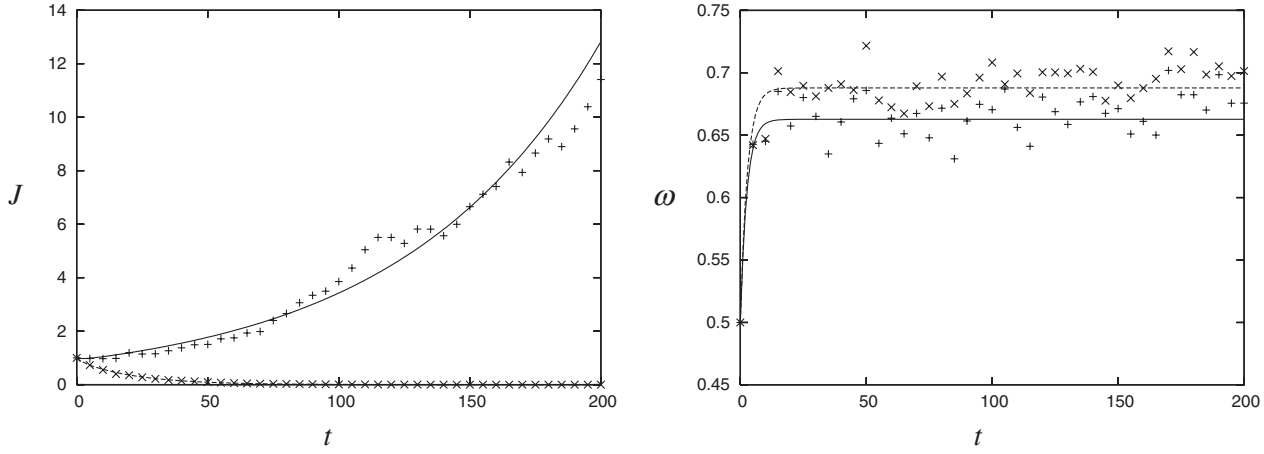

Fig. 2. Time dependences of J and ω for output noise model and AdaTron learning rule. $\eta = 1, k_{\mathrm{T}} = 0.7$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and $+$: $k_{\mathrm{S}} = 0.8$; dashed curve and $\times$: $k_{\mathrm{S}} = 0.9$. Left panel, J. Right panel, ω.

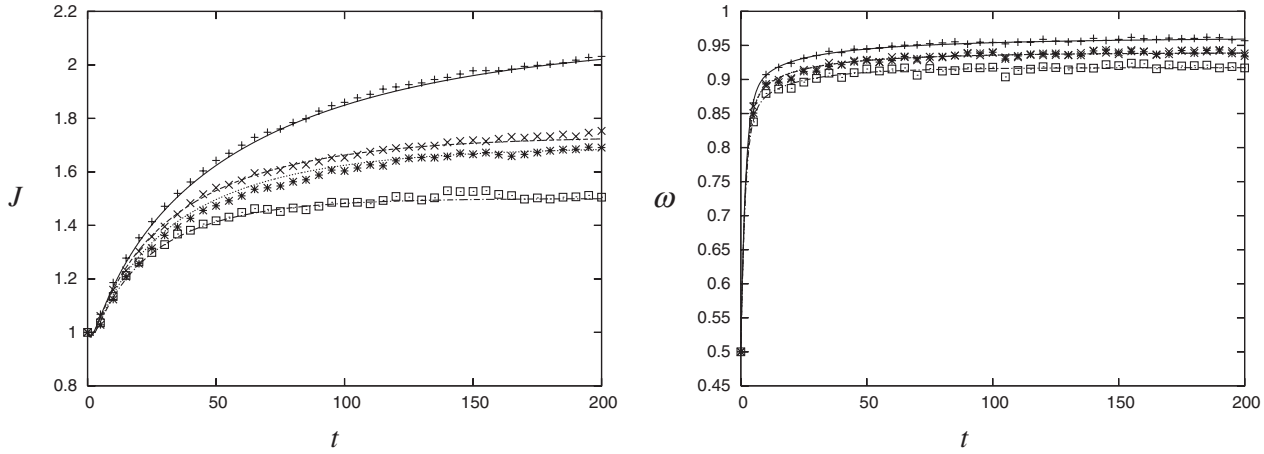

Fig. 3. Time dependences of J and ω for input noise model and Perceptron learning rule. $\eta = 1$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and $+$: $\sigma_{\mathrm{T}} = 0.2, \sigma_{\mathrm{S}} = 0.2$; dashed curve and $\times$: $\sigma_{\mathrm{T}} = 0.2, \sigma_{\mathrm{S}} = 0.3$; dotted curve and $*$: $\sigma_{\mathrm{T}} = 0.3, \sigma_{\mathrm{S}} = 0.2$; dotted-dashed curve and square: $\sigma_{\mathrm{T}} = 0.3, \sigma_{\mathrm{S}} = 0.3$. Left panel, J. Right panel, ω.

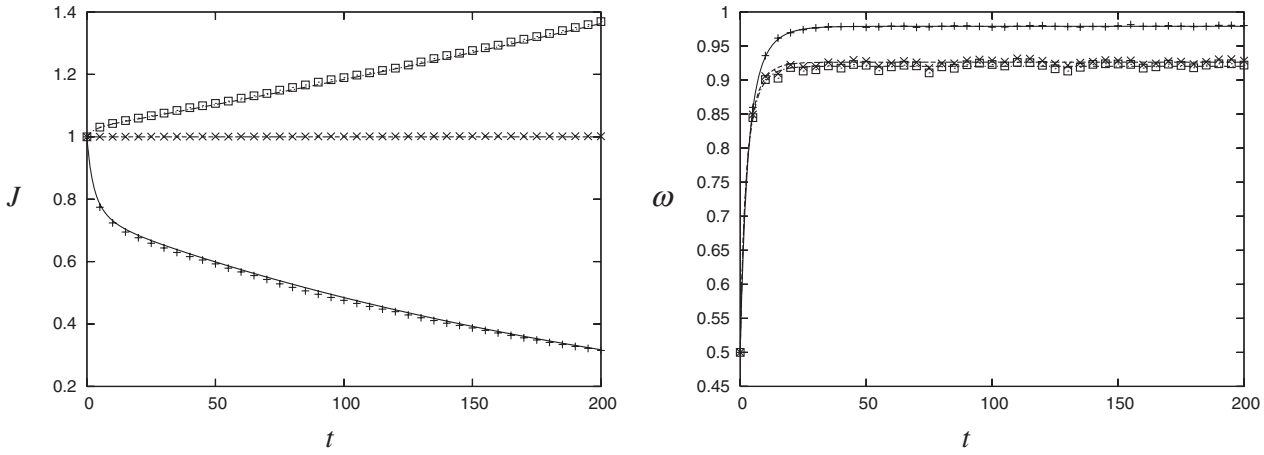

Fig. 4. Time dependences of J and ω for input noise model and AdaTron learning rule. $\sigma_{\mathrm{T}} = \sigma_{\mathrm{S}} = 0.2$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and $+$: $\eta = 1$ ($\sigma_{\mathrm{S}} < \sqrt{(2/\eta) - 1}$); dashed curve and $\times$: $\eta = 2/1.04$ ($\sigma_{\mathrm{S}} = \sqrt{(2/\eta) - 1}$); dotted-dashed curve and square: $\eta = 2/1.01$ ($\sigma_{\mathrm{S}} > \sqrt{(2/\eta) - 1}$). Left panel, J. Right panel, ω.

Table II. a and b for each learning rule in output and input noise models.

Output	Hebbian	Perceptron	AdaTron
a	$k_{\mathrm{T}}\sqrt{\frac{2}{\pi}}(1-\omega^2)$	$\frac{k_{\mathrm{T}}}{\sqrt{2\pi}}(1-\omega^2)$	$\frac{k_{\mathrm{T}}}{\pi}(1-\omega^2)^{3/2}$
b	ω	$\omega\epsilon_{\mathrm{g}}$	$\omega\left(\epsilon_{\mathrm{g}}-\frac{k_{\mathrm{T}}k_{\mathrm{S}}\omega}{\pi}\sqrt{1-\omega^2}\right)$
Input	**Hebbian**	**Perceptron**	**AdaTron**
a	$\sqrt{\frac{2}{\pi(1+\sigma_{\mathrm{T}}^2)}}(1-\omega^2)$	$\frac{1}{\sqrt{2\pi(1+\sigma_{\mathrm{T}}^2)}}(1-\omega^2)$	$\frac{\sqrt{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-\omega^2}}{\pi(1+\sigma_{\mathrm{T}}^2)}(1-\omega^2)$
b	$\omega(1+\sigma_{\mathrm{S}}^2)$	$\omega(1+\sigma_{\mathrm{S}}^2)\epsilon_{\mathrm{g}}$	$\frac{\omega}{\pi}(1+\sigma_{\mathrm{S}}^2)\left[(1+\sigma_{\mathrm{S}}^2)\pi\epsilon_{\mathrm{g}}-\omega\frac{\sqrt{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-\omega^2}}{1+\sigma_{\mathrm{T}}^2}\right]$

Table III. Asymptotic forms of optimal learning rate $\tilde{\eta}_{\mathrm{opt}}$ and $\tilde{\epsilon}_{\mathrm{g,opt}}$ for $t\gg 1$. $\epsilon^*_{\mathrm{g,min}}=\epsilon_{\mathrm{g}}(\omega=1)$.

Output	Hebbian	Perceptron ($k_{\mathrm{T}}<k_{\mathrm{S}}$)	AdaTron
$\tilde{\eta}_{\mathrm{opt}}(t)$	$\frac{1}{k_{\mathrm{T}}}\sqrt{\frac{\pi}{2}}t^{-1}$	$\frac{\sqrt{2\pi}}{k_{\mathrm{T}}}t^{-1}$	$\left(\frac{\pi^2}{4k_{\mathrm{T}}^2(1-k_{\mathrm{T}}k_{\mathrm{S}})}\right)^{1/4}t^{-3/4}$
$\tilde{\epsilon}_{\mathrm{g,opt}}$	$\frac{k_{\mathrm{S}}}{\sqrt{2\pi}}t^{-1/2}$	$k_{\mathrm{S}}\sqrt{\frac{1-k_{\mathrm{T}}k_{\mathrm{S}}}{\pi}}t^{-1/2}$	$\left(\frac{k_{\mathrm{T}}^2(1-k_{\mathrm{S}}k_{\mathrm{T}})}{4\pi^2}\right)^{1/4}k_{\mathrm{S}}t^{-1/4}$
Input	**Hebbian**	**Perceptron ($k_{\mathrm{T}}<k_{\mathrm{S}}$)**	**AdaTron**
$\tilde{\eta}_{\mathrm{opt}}(t)$	$\sqrt{1+\sigma_{\mathrm{T}}^2}\sqrt{\frac{\pi}{2}}t^{-1}$	$\sqrt{2\pi(1+\sigma_{\mathrm{T}}^2)}t^{-1}$	$\frac{\pi(1+\sigma_{\mathrm{T}}^2)}{\sqrt{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-1}}t^{-1}$
$\tilde{\epsilon}_{\mathrm{g,opt}}$	$\frac{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)}{4\sqrt{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-1}}t^{-1}$	$\frac{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)\epsilon^*_{\mathrm{g}}}{\sqrt{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-1}}t^{-1}$	$\frac{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)\{\pi(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)\epsilon^*_{\mathrm{g}}-\sqrt{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-1}\}}{2\{(1+\sigma_{\mathrm{T}}^2)(1+\sigma_{\mathrm{S}}^2)-1\}^{3/2}}t^{-1}$

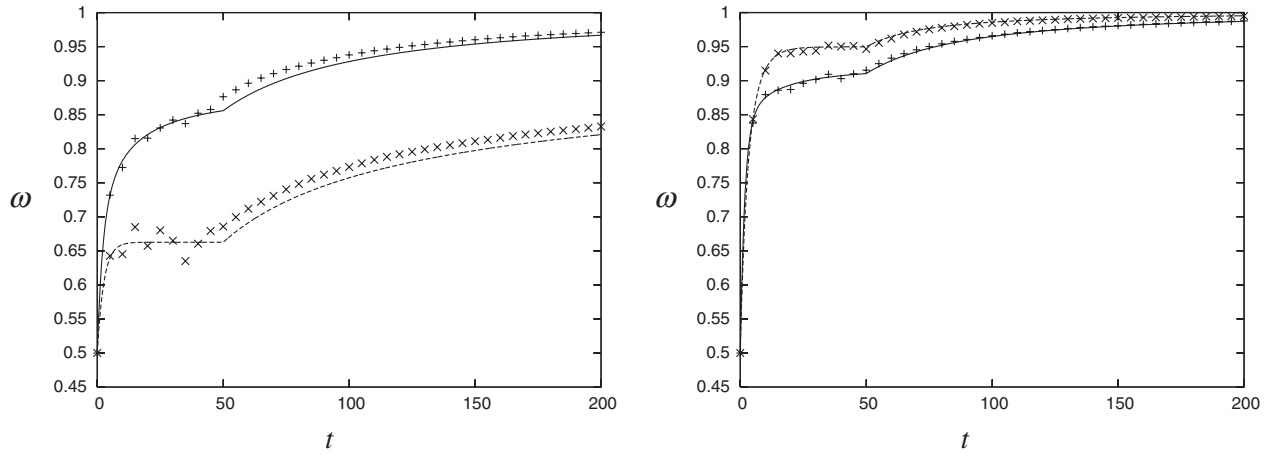

Fig. 5. Time dependence of ω for the optimal learning rate. Curves are theoretical results (RKG) and symbols are numerical results ($N=1000$). Solid curve and +: Perceptron; dashed curve and ×: AdaTron. For $t<50, \eta=1$ and for $t\geq 50, \eta=\tilde{\eta}_{\mathrm{opt}}$. Left panel: output noise model. $k_{\mathrm{T}}=0.7, k_{\mathrm{S}}=0.8$. Right panel: input noise model. $\sigma_{\mathrm{T}}=0.3, \sigma_{\mathrm{S}}=0.3$.

From this, the optimal learning rate $\tilde{\eta}_{\mathrm{opt}}$ and the differential equation obtained using $\tilde{\eta}_{\mathrm{opt}}$ are derived as

$$\tilde{\eta}_{\mathrm{opt}}=\frac{a}{b},\quad \frac{\mathrm{d}}{\mathrm{d}t}\omega=\frac{a^2}{2b}, \tag{63}$$

where a and b are given in Table II. It is easily proved that $\omega\to 1$ as $t\to\infty$, that is, learning succeeds. Let us define $\tilde{\epsilon}_{\mathrm{g}}\equiv\epsilon_{\mathrm{g,min}}-\epsilon_{\mathrm{g}}(\omega)$, where $\epsilon_{\mathrm{g,min}}=\epsilon_{\mathrm{g}}(\omega=1)$. Asymptotic forms of $\tilde{\eta}_{\mathrm{opt}}$ and $\tilde{\epsilon}_{\mathrm{g,opt}}$, which is $\tilde{\epsilon}_{\mathrm{g}}$ evaluated for $\tilde{\eta}_{\mathrm{opt}}$, are shown in Table III. In Fig. 5, we show the numerical and theoretical results for ω in the Perceptron and AdaTron rules. In the theoretical calculation and numerical simulations, we used the asymptotic forms of $\tilde{\eta}_{\mathrm{opt}}$ as η. We find that the agreements between the theoretical and numerical results are fairly well.

5. Control of Learning

In the output noise model for the Perceptron learning rule, learning succeeds for $k_{\mathrm{T}}\geq k_{\mathrm{S}}$, but fails for $k_{\mathrm{T}}<k_{\mathrm{S}}$. This is a rather surprising result. Let us consider the situation of $k_{\mathrm{S}}=a>k_{\mathrm{T}}$, where a is some positive value less than 1. In this case, learning fails. If k_{S} is decreased from a with k_{T} fixed, learning becomes better and, at $k_{\mathrm{S}}=k_{\mathrm{T}}$, it succeeds. That is, when the student noise is increased, learning becomes better and when it is decreased, learning becomes worse. We can use this fact to control learning by intentionally reversing the student's output. Suppose that learning fails. We introduce the control parameter k_{SC} and reverse the student output with the probability $(1-k_{\mathrm{SC}})/2$. Then, the net probability that the student's output is reversed is

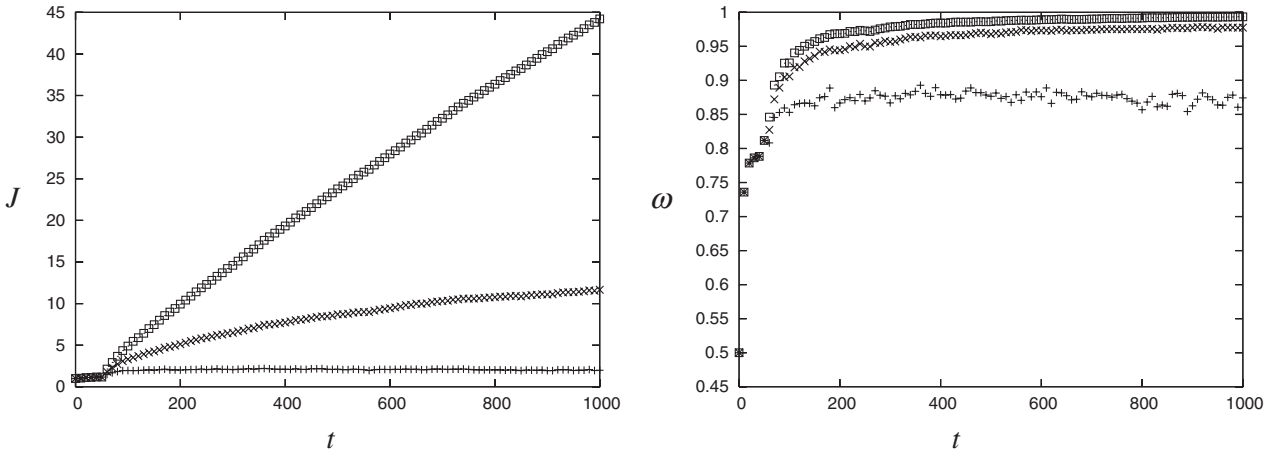

Fig. 6. Time dependence of J and ω for output noise model and Perceptron learning rule. $k_{\mathrm{T}} = 0.7, k_{\mathrm{S}} = 0.9$. In this case, $k^*_{\mathrm{SC}} = 7/9$. We reverse the student output with the probability $(1 - k_{\mathrm{SC}})/2$ for $t > 50$. Numerical simulations ($N = 1000$). +: $k_{\mathrm{SC}} = 8/9$, ×: $k_{\mathrm{SC}} = 7/9$, square $k_{\mathrm{SC}} = 6/9$, Left panel, J. Right panel, ω.

$(1 - k_{\mathrm{S}}k_{\mathrm{SC}})/2$. That is, the substantial parameter for the student is $\hat{k} = k_{\mathrm{S}}k_{\mathrm{SC}}$, instead of k_{S}. If $k_{\mathrm{T}} > k_{\mathrm{S}}k_{\mathrm{SC}}$, learning succeeds. Suppose that we observe the generalization error ϵ_{g} by decreasing k_{SC} from 1. When k_{SC} reaches its critical value k^*_{SC}, ϵ_{g} reaches its minimum value $\epsilon_{\mathrm{g,min}}$. The critical value is given by

$$k^*_{\mathrm{SC}} = \frac{k_{\mathrm{T}}}{k_{\mathrm{S}}}. \tag{64}$$

For $k_{\mathrm{SC}} < k^*_{\mathrm{SC}}$, ϵ_{g} is constant and is $\epsilon_{\mathrm{g,min}}$. Thus, k^*_{SC} is estimated by identifying the value at which ϵ_{g} becomes constant as k_{SC} is decreased. On the other hand, $\epsilon_{\mathrm{g,min}}$ is given as

$$\epsilon_{\mathrm{g,min}} = \frac{1 - k_{\mathrm{T}}k_{\mathrm{S}}k^*_{\mathrm{SC}}}{2} = \frac{1 - k^2_{\mathrm{T}}}{2}. \tag{65}$$

By numerically measuring k^*_{SC} and $\epsilon_{\mathrm{g,min}}$, we can estimate both k_{S} and k_{T} using eqs. (64) and (65). We show results of numerical simulations in Fig. 6 and find that when k_{SC} becomes smaller than the critical value $k^*_{\mathrm{SC}} = 7/9$, learning succeeds.

6. Time Domain Ensemble Learning

Now, we consider another method of making learning successful when it fails. The method is time averaging. Previously, we formulated it when only the teacher suffers from external noise.[6] We extend the theory to the present case in which both the teacher and the student suffer from external noise. Let us briefly explain the formulation of time domain ensemble learning.

We consider a two-time-correlation function $q(t, s) \equiv \boldsymbol{J}(t) \cdot \boldsymbol{J}(s)$.[17] The differential equation for $q(t, s)$ with respect to s for $t \leq s$ is given by

$$\frac{\partial q(t,s)}{\partial s} = \eta J(t) \langle x^t_s T(s) \mathcal{F}(s) \rangle_{\Xi_s} \quad \text{output noise model,} \tag{66}$$

$$\frac{\partial q(t,s)}{\partial s} = \eta J(t) \langle (x^t_s + v^t_s) T(s) \mathcal{F}(s) \rangle_{\Xi_s} \quad \text{input noise model,} \tag{67}$$

where $x^t_s = \widehat{\boldsymbol{J}}(t) \cdot \boldsymbol{\xi}_s$ and $v^t_s = \widehat{\boldsymbol{J}}(t) \cdot \boldsymbol{\zeta}^{\mathrm{S}}_s$. $\boldsymbol{\xi}_s$ and $\boldsymbol{\zeta}^{\mathrm{S}}_s$ are a sample and a student noise given at a time s, respectively. $\langle \rangle_{\Xi_s}$ denotes the average over samples and noises at a time s, and $\mathcal{F}(s)$ denotes $\mathcal{F}$ estimated at a time s.

We define the time-averaged student vectors $\overline{\boldsymbol{J}}(t)$ and $\overline{\widehat{\boldsymbol{J}}}(t)$ as follows:

$$\overline{\boldsymbol{J}}(t) \equiv \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{J}(t + t_i), \tag{68}$$

$$\overline{\widehat{\boldsymbol{J}}}(t) \equiv \frac{1}{K} \sum_{i=1}^{K} \widehat{\boldsymbol{J}}(t + t_i) = \frac{1}{K} \sum_{i=1}^{K} \frac{\boldsymbol{J}(t + t_i)}{J(t + t_i)}, \tag{69}$$

where $t_1 < t_2 < \cdots < t_K$. The order parameters are defined as follows:

$$\overline{R}(t) \equiv \boldsymbol{B} \cdot \overline{\boldsymbol{J}}(t) = \frac{1}{K} \sum_{i=1}^{K} R(t + t_i), \tag{70}$$

$$\overline{Q}(t) \equiv \overline{\boldsymbol{J}}(t)^2 = \frac{2}{K^2} \sum_{i<j} q(t + t_i, t + t_j) + \frac{1}{K^2} \sum_{i=1}^{K} J(t + t_i)^2, \tag{71}$$

$$\overline{\omega}(t) \equiv \frac{\overline{R}(t)}{\sqrt{\overline{Q}(t)}}, \tag{72}$$

$$\overline{\widehat{R}}(t) \equiv \boldsymbol{B} \cdot \overline{\widehat{\boldsymbol{J}}}(t) = \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{B} \cdot \widehat{\boldsymbol{J}}(t + t_i) = \frac{1}{K} \sum_{i=1}^{K} \omega(t + t_i), \tag{73}$$

$$\overline{\widehat{Q}}(t) \equiv \overline{\widehat{\boldsymbol{J}}}(t)^2 = \frac{2}{K^2} \sum_{i<j} \frac{q(t + t_i, t + t_j)}{J(t + t_i) J(t + t_j)} + \frac{1}{K}, \tag{74}$$

$$\overline{\widehat{\omega}}(t) \equiv \frac{\overline{\widehat{R}}(t)}{\sqrt{\overline{\widehat{Q}}(t)}}. \tag{75}$$

We derive the asymptotic expressions for $\overline{\omega}(t)$ and $\overline{\widehat{\omega}}(t)$ as $t \to \infty$ for a finite K, and discuss the efficiency of time domain ensemble learning.

Now, we derive differential equations for $q(t, s)$ with respect to $s(\geq t)$ in both the output and input noise models. Then, in order to obtain the asymptotic forms of

$\overline{\omega}$ and $\overline{\overline{\omega}}$, we study the asymptotic behaviors of $q(t+t_i, t+t_j)$ and $\hat{q}(t+t_i, t+t_j)$, where $\hat{q}(t,s) \equiv q(t,s)/(J(t)J(s))$. In the next subsection, we give a differential equation only for $q(t,s)$.

6.1 Differential equations and asymptotic behaviors

6.1.1 Output noise model

Now, let us study the output noise model. The differential equation for $q(t,s)$ is

$$\frac{\partial q(t,s)}{\partial s} = \eta J(t) \langle x_s^t \{ \mathcal{P}_{\mathrm{T}}(y_s) \mathcal{F}_+(s) - (1 - \mathcal{P}_{\mathrm{T}}(y_s)) \mathcal{F}_-(s) \} \rangle_{x_s^t, x_s^s, y_s}, \tag{76}$$

where $\mathcal{F}_-(s)$ are $\mathcal{F}(s)$ estimated for $T(s) = 1$ and $T(s) = -1$, respectively. That is, $\mathcal{F}_+(s) \equiv \mathcal{F}[J(s), x_s^s, T(s) = +1, S(s)]$ and $\mathcal{F}_-(s) \equiv \mathcal{F}[J(s), x_s^s, T(s) = -1, S(s)]$. Here, $x_s^s = \widehat{\boldsymbol{J}}(s) \cdot \boldsymbol{\xi}_s$ and $y_s = \boldsymbol{B} \cdot \boldsymbol{\xi}_s$. $\langle \cdot \rangle_{x_s^t, x_s^s, y_s}$ denotes the average over the Gaussian distribution $P_3(x_s^t, x_s^s, y_s)$ of x_s^t, x_s^s, and y_s with $\langle x_s^t \rangle = \langle x_s^s \rangle = \langle y_s \rangle = 0$, $\langle (x_s^t)^2 \rangle = \langle (x_s^s)^2 \rangle = \langle y_s^2 \rangle = 1$, $\langle x_s^t x_s^s \rangle = \hat{q}(t,s)$, $\langle x_s^t y_s \rangle = \omega(t)$, and $\langle x_s^s y_s \rangle = \omega(s)$. The initial condition for this equation is $q(t,t) = J(t)^2$.

By performing several integrations in eq. (76), we obtain the differential equation for each learning rule. In the Appendix, we list the integrations used when the differential equations are derived. We omit the details of the calculation and give the resultant differential equations.

In the Hebbian rule, the differential equation for $q(t,s)$ is

$$\frac{\partial q(t,s)}{\partial s} = k_{\mathrm{T}} \eta \sqrt{\frac{2}{\pi}} R(t) \text{ for } s \geq t. \tag{77}$$

The solutions for R, J, and q with the initial conditions $R(0) = 0$, $J(0) = 1$, and $q(t,t) = J(t)^2$ are

$$R(t) = \eta k_{\mathrm{T}} \sqrt{\frac{2}{\pi}} t, \tag{78}$$

$$J(t) = \sqrt{1 + \eta^2 t \left(1 + \frac{2}{\pi} k_{\mathrm{T}}^2 t \right)}, \tag{79}$$

$$q(t,s) = k_{\mathrm{T}} \eta \sqrt{\frac{2}{\pi}} R(t)(s-t) + J(t)^2 \text{ for } s \geq t. \tag{80}$$

From these solutions, it follows that $\lim_{t \to \infty} \overline{\overline{\omega}}(t) = 1$.

In the Perceptron rule, the differential equation for q is

$$\frac{\partial q(t,s)}{\partial s} = \frac{k_{\mathrm{T}} \eta}{\sqrt{2\pi}} R(t) - \frac{\eta k_{\mathrm{S}}}{\sqrt{2\pi}} \frac{q(t,s)}{J(s)} \text{ for } s \geq t. \tag{81}$$

In the AdaTron rule, the differential equation for q is

$$\frac{\partial q(t,s)}{\partial s} = \frac{\eta}{k_{\mathrm{S}}} \left[\frac{1 - k_{\mathrm{S}}^2}{2} - \epsilon_{\mathrm{g}}(s) \right] q(t,s) + \frac{\eta k_{\mathrm{T}}}{\pi} R(t) J(s) \sqrt{1 - \omega(s)^2} \text{ for } s \geq t. \tag{82}$$

6.1.2 Input noise model

Now, we study the input noise model, where $\mathcal{P}_{\mathrm{T}}(y) = H(-y/\sigma_{\mathrm{T}})$. Then, we obtain

$$\frac{\partial q(t,s)}{\partial s} = \eta J(t) \left\langle (x_s^t + v_s^t) \left\{ H\left(-\frac{y_s}{\sigma_{\mathrm{T}}} \right) \mathcal{F}_+(s) - H\left(\frac{y_s}{\sigma_{\mathrm{T}}} \right) \mathcal{F}_-(s) \right\} \right\rangle_{x_s^t, x_s^s, y_s, v_s^t, v_s^s}, \tag{83}$$

where $\mathcal{F}_+(s) \equiv \mathcal{F}[J(s), x_s^s, v_s^s, T = +1]$ and $\mathcal{F}_-(s) \equiv \mathcal{F}[J(s), x_s^s, v_s^s, T = -1]$. Since $\boldsymbol{\xi}_s$ and $\boldsymbol{\zeta}_s^{\mathrm{S}}$ are statistically independent, the average of the quantity A, $\langle A \rangle_{x_s^t, x_s^s, y_s, v_s^t, v_s^s}$, is calculated using the product of $P_3(x_s^t, x_s^s, y_s)$ and $P_v(v_s^t, v_s^s)$. Here, $P_v(v_s^t, v_s^s)$ is the Gaussian distribution with $\langle v_s^t \rangle = \langle v_s^s \rangle = 0$, $\langle (v_s^t)^2 \rangle = \langle (v_s^s)^2 \rangle = \sigma_{\mathrm{S}}^2$, and $\langle v_s^t v_s^s \rangle = \sigma_{\mathrm{S}}^2 \hat{q}(t,s)$.

For the Hebbian rule, the differential equation for $q(t,s)$ is

$$\frac{\partial q(t,s)}{\partial s} = \eta \frac{1}{\sqrt{\pi(1 + \sigma_{\mathrm{T}}^2)}} R(t). \tag{84}$$

This equation and the equations for other order parameters for the input noise model are obtained using those for the output noise model replacing k_{T} by $1/\sqrt{1 + \sigma_{\mathrm{T}}^2}$ and η^2 by $\eta^2(1 + \sigma_{\mathrm{S}}^2)$. Therefore, we obtain $\lim_{t \to \infty} \overline{\overline{\omega}}(t) = 1$.

In the Perceptron rule, the differential equation for q is

$$\frac{\partial q(t,s)}{\partial s} = \frac{\eta}{\sqrt{2\pi}} \left[\frac{R(t)}{\sqrt{1 + \sigma_{\mathrm{T}}^2}} - \sqrt{1 + \sigma_{\mathrm{S}}^2} \frac{q(t,s)}{J(s)} \right] \text{ for } s \geq t. \tag{85}$$

In the AdaTron rule, the equation for q is

$$\frac{\partial q(t,s)}{\partial s} = \frac{\eta}{\pi} R(t) J(s) \frac{\sqrt{(1 + \sigma_{\mathrm{T}}^2)(1 + \sigma_{\mathrm{S}}^2) - \omega(s)^2}}{1 + \sigma_{\mathrm{T}}^2} - \eta (1 + \sigma_{\mathrm{S}}^2) q(t,s) \epsilon_{\mathrm{g}}[\omega(s)] \text{ for } s \geq t. \tag{86}$$

6.2 Asymptotic behavior

Now, let us derive the asymptotic forms of $\overline{\omega}(t)$ and $\overline{\overline{\omega}}(t)$. In order to evaluate them, we have to solve the differential equations for $q(t,s)$. These equations have the following form:

$$\frac{\partial}{\partial s} q(t,s) = f(s) q(t,s) + g(t,s). \tag{87}$$

This is easily solved and we obtain for $t_1 \leq t_2$

$$q(t+t_1, t+t_2) = \left\{ J(t+t_1)^2 + \int_0^{t_2 - t_1} \mathrm{d}\tau \, g(t+t_1, \tau + t + t_1) \times \exp\left[-\int_0^{\tau} \mathrm{d}u \, f(u + t + t_1) \right] \right\} \times \exp\left[\int_0^{t_2 - t_1} \mathrm{d}v \, f(v + t + t_1) \right]. \tag{88}$$

Note that $g(t,s)$ is the product of the function of t and that of s. For the Perceptron rule, $f(s)$ and $g(t,s)$ are as follows:

$$f(s) = -\frac{\eta k_{\mathrm{S}}}{\sqrt{2\pi}} \frac{1}{J(s)} \quad \text{for output noise model,} \tag{89}$$

$$f(s) = -\frac{\eta \sqrt{1 + \sigma_{\mathrm{S}}^2}}{\sqrt{2\pi}} \frac{1}{J(s)} \quad \text{for input noise model,} \tag{90}$$

$$g(t,s) = \frac{k_{\mathrm{T}} \eta}{\sqrt{2\pi}} R(t) \quad \text{for output noise model,} \tag{91}$$

$$g(t,s) = \frac{\eta}{\sqrt{2\pi}} \frac{R(t)}{\sqrt{1 + \sigma_{\mathrm{T}}^2}} \quad \text{for input noise model.} \tag{92}$$

On the other hand, for the AdaTron rule, these functions are as follows:

$$f(s) = -\eta\left\{\frac{k_T - k_S}{2} - \frac{k_T}{\pi}\cos^{-1}[\omega(s)]\right\} \quad \text{for output noise model,} \tag{93}$$

$$f(s) = -\eta(1+\sigma_S^2)\epsilon_g[\omega(s)] \quad \text{for input noise model,} \tag{94}$$

$$g(t,s) = \eta\frac{k_T}{\pi}R(t)J(s)\sqrt{1-\omega(s)^2} \quad \text{for output noise model,} \tag{95}$$

$$g(t,s) = \frac{\eta}{\pi}R(t)J(s)\frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)-[\omega(s)]^2}}{1+\sigma_T^2} \quad \text{for input noise model.} \tag{96}$$

Thus, it follows that, as $t \to \infty$, $J(t) \to J^*$, $\omega(t) \to \omega^*$, $f(t) \to f^*$, and $g(t+\alpha, t+\beta) \to g^*$ for arbitrary constants α and β. When learning fails, in both the output and input noise models, for the Perceptron rule, these limiting values are all finite, but for the AdaTron rule, J^* and g^* are finite or 0 or ∞ depending on the parameters. Then, we obtain for the Perceptron rule

$$\lim_{t\to\infty} q(t+t_1, t+t_2) = \left[(J^*)^2 + \frac{g^*}{f^*}\right]\exp[f^*(t_2-t_1)] - \frac{g^*}{f^*}. \tag{97}$$

For the AdaTron rule, $\lim_{t\to\infty} q(t+t_1, t+t_2)$ takes finite or 0 or ∞. Thus, we consider two cases separately in the following subsections.

6.3 *Perceptron learning rule*

Let us consider the Perceptron rule. In this case, J^* and g^* are nonzero. Then, we consider $\overline{\omega}(t)$. When $t \to \infty$, we obtain

$$\lim_{t\to\infty}\overline{R}(t) = \lim_{t\to\infty}\frac{1}{K}\sum_{i=1}^{K}R(t+t_i) = R^* = J^*\omega^*, \tag{98}$$

$$\lim_{t\to\infty}\overline{Q}(t) = \frac{2}{K^2}\sum_{i<j}\left[(J^*)^2 + \frac{g^*}{f^*}\right]\exp[f^*(t_j-t_i)] - \frac{K-1}{K}\frac{g^*}{f^*} + \frac{(J^*)^2}{K}. \tag{99}$$

$$\lim_{t\to\infty}\overline{\omega}(t) = \frac{J^*\omega^*}{\sqrt{\dfrac{2}{K^2}\displaystyle\sum_{i<j}\left[(J^*)^2 + \frac{g^*}{f^*}\right]\exp[f^*(t_j-t_i)] - \frac{K-1}{K}\frac{g^*}{f^*} + \frac{(J^*)^2}{K}}}. \tag{100}$$

In the output noise model, asymptotic values are calculated as

$$f^* = -\frac{\eta k_S}{\sqrt{2\pi}J^*}, \tag{101}$$

$$g^* = \frac{\eta k_T}{\sqrt{2\pi}}R^*, \tag{102}$$

$$-\frac{g^*}{f^*} = (J^*\omega^*)^2. \tag{103}$$

In the input noise model, asymptotic values are calculated as

$$f^* = -\frac{\eta\sqrt{1+\sigma_S^2}}{\sqrt{2\pi}J^*}, \tag{104}$$

$$g^* = \frac{\eta}{\sqrt{2\pi(1+\sigma_T^2)}}R^*, \tag{105}$$

$$-\frac{g^*}{f^*} = (J^*\omega^*)^2. \tag{106}$$

Thus, we obtain[18)]

$$\lim_{t\to\infty} q(t+t_1, t+t_2) = (J^*)^2\{(\omega^*)^2 + [1-(\omega^*)^2]\exp[f^*(t_2-t_1)]\}, \tag{107}$$

$$\lim_{t\to\infty}\overline{Q}(t) = (J^*)^2\left((\omega^*)^2 + \frac{1}{K}[1-(\omega^*)^2]\left\{1 + \frac{2}{K}\sum_{i<j}\exp[f^*(t_j-t_i)]\right\}\right). \tag{108}$$

Therefore, we obtain the asymptotic form of the overlap between the teacher vector and the averaged student vector in both the output and input noise models as

$$\lim_{t\to\infty}\overline{\omega}(t) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \dfrac{1}{K}[1-(\omega^*)^2]\left\{1 + \dfrac{2}{K}\displaystyle\sum_{i<j}\exp[f^*(t_j-t_i)]\right\}}}. \tag{109}$$

6.4 *AdaTron learning rule*

In this subsection, we consider the AdaTron rule. We consider $\overline{\widehat{\omega}}(t)$. When $t \to \infty$, we obtain

$$\lim_{t\to\infty}\overline{\widehat{R}}(t)=\lim_{t\to\infty}\frac{1}{K}\sum_{i=1}^{K}\omega(t+t_i)=\omega^*, \tag{110}$$

$$\lim_{t\to\infty}\overline{\widehat{Q}}(t)=\frac{2}{K^2}\sum_{i<j}\lim_{t\to\infty}\frac{q(t+t_i,t+t_j)}{J(t+t_i)J(t+t_j)}+\frac{1}{K}. \tag{111}$$

For the AdaTron rule, $g(t,s)$ is expressed as

$$g(t,s)=R(t)J(s)\tilde{g}[\omega(s)]. \tag{112}$$

Thus, we obtain

$$\begin{aligned}\lim_{t\to\infty}\frac{q(t+t_1,t+t_2)}{J(t+t_1)J(t+t_2)}&=\lim_{t\to\infty}\left\{\frac{J(t+t_1)}{J(t+t_2)}+\int_0^{t_2-t_1}\mathrm{d}\tau\,\frac{J(\tau+t+t_1)}{J(t+t_2)}\omega(t+t_1)\tilde{g}[\omega(\tau+t+t_1)]\exp\left[-\int_0^{\tau}\mathrm{d}u\,f(u+t+t_1)\right]\right\}\\&\quad\times\exp\left[\int_0^{t_2-t_1}\mathrm{d}v\,f(v+t+t_1)\right]\\&=\left[B(t_1,t_2)+\int_0^{t_2-t_1}\mathrm{d}\tau\,B(\tau+t_1,t_2)\omega^*\tilde{g}(\omega^*)\exp(-f^*\tau)\right]\exp[f^*(t_2-t_1)].\end{aligned} \tag{113}$$

Here, we define

$$B(t_1,t_2)\equiv\lim_{t\to\infty}\frac{J(t+t_1)}{J(t+t_2)}. \tag{114}$$

This is calculated as follows. The equations for $J(t)$ and $\omega(t)$ have the following forms:

$$\frac{\mathrm{d}J(t)}{\mathrm{d}t}=J(t)\phi(\omega), \tag{115}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t}=\psi(\omega). \tag{116}$$

By solving eq. (116), we obtain $\omega=\omega(t)$. Using this solution, $J(t)$ is given by

$$J(t)=J(0)\exp\left\{\int_0^t\mathrm{d}t'\,\phi[\omega(t')]\right\}. \tag{117}$$

Thus,

$$\begin{aligned}\frac{J(t+t_1)}{J(t+t_2)}&=\exp\left\{-\int_{t+t_1}^{t+t_2}\mathrm{d}t'\,\phi[\omega(t')]\right\}\\&=\exp\left\{-\int_0^{t_2-t_1}\mathrm{d}\tau\,\phi[\omega(\tau+t+t_1)]\right\}.\end{aligned} \tag{118}$$

From this, $B(t_1,t_2)$ is expressed as

$$\begin{aligned}B(t_1,t_2)&\equiv\lim_{t\to\infty}\exp\left\{-\int_0^{t_2-t_1}\mathrm{d}\tau\,\phi[\omega(\tau+t+t_1)]\right\}\\&=\exp[-\phi^*(t_2-t_1)],\end{aligned} \tag{119}$$

where $\phi^*=\phi(\omega^*)$. Therefore, we obtain

$$\begin{aligned}&\lim_{t\to\infty}\frac{q(t+t_1,t+t_2)}{J(t+t_1)J(t+t_2)}\\&\quad=\exp[(f^*-\phi^*)(t_2-t_1)]\left[1-\frac{\omega^*\tilde{g}(\omega^*)}{\phi^*-f^*}\right]+\frac{\omega^*\tilde{g}(\omega^*)}{\phi^*-f^*}.\end{aligned} \tag{120}$$

Then,

$$\begin{aligned}\lim_{t\to\infty}\overline{\widehat{Q}}(t)&=\frac{2}{K^2}\sum_{i<j}\exp[(f^*-\phi^*)(t_j-t_i)]\left[1-\frac{\omega^*\tilde{g}(\omega^*)}{\phi^*-f^*}\right]\\&\quad+\frac{K-1}{K}\frac{\omega^*\tilde{g}(\omega^*)}{\phi^*-f^*}+\frac{1}{K}.\end{aligned} \tag{121}$$

Thus, we obtain

$$\lim_{t\to\infty}\overline{\widehat{\omega}}(t)=\frac{\omega^*}{\sqrt{\dfrac{2}{K^2}\displaystyle\sum_{i<j}\exp[(f^*-\phi^*)(t_j-t_i)]\left[1-\dfrac{\omega^*\tilde{g}(\omega^*)}{\phi^*-f^*}\right]+\dfrac{K-1}{K}\dfrac{\omega^*\tilde{g}(\omega^*)}{\phi^*-f^*}+\dfrac{1}{K}}}. \tag{122}$$

Now, we calculate relevant asymptotic values. First, we treat the output noise model and obtain

$$f^*=-\frac{\eta}{k_{\mathrm{S}}}\left(\epsilon_{\mathrm{g}}^*-\frac{1-k_{\mathrm{S}}^2}{2}\right), \tag{123}$$

$$\tilde{g}^*=\eta\frac{k_{\mathrm{T}}}{\pi}\sqrt{1-(\omega^*)^2}, \tag{124}$$

$$\begin{aligned}\phi^*=\eta\Bigg\{&\frac{1-k_{\mathrm{S}}^2}{2k_{\mathrm{S}}}+\left(\frac{\eta}{2}-\frac{1}{k_{\mathrm{S}}}\right)\\&\times\left[\epsilon_{\mathrm{g}}^*-\frac{k_{\mathrm{T}}k_{\mathrm{S}}}{\pi}\omega^*\sqrt{1-(\omega^*)^2}\right]\Bigg\},\end{aligned} \tag{125}$$

where $\epsilon_{\mathrm{g}}^*=\epsilon_{\mathrm{g}}(\omega^*)$. Now, we estimate $\omega^*\tilde{g}(\omega^*)/(\phi^*-f^*)$. From $\mathrm{d}\omega/\mathrm{d}t=0$, we obtain

$$\frac{k_{\mathrm{T}}\eta}{\pi}[1-(\omega^*)^2]^{3/2}=\frac{\eta^2}{2}\omega^*\left[\epsilon_{\mathrm{g}}^*-\frac{k_{\mathrm{T}}k_{\mathrm{S}}}{\pi}\omega^*\sqrt{1-(\omega^*)^2}\right]. \tag{126}$$

From this, we obtain

$$\epsilon_{\mathrm{g}}^*=\frac{2k_{\mathrm{T}}}{\pi\eta\omega^*}\left[1-\left(1-\frac{\eta k_{\mathrm{S}}}{2}\right)(\omega^*)^2\right]\sqrt{1-(\omega^*)^2}. \tag{127}$$

Thus, we obtain

$$\phi^*-f^*=\frac{\tilde{g}^*}{\omega^*}. \tag{128}$$

Therefore, we obtain

$$\frac{\omega^* \tilde{g}(\omega^*)}{\phi^* - f^*} = (\omega^*)^2. \tag{129}$$

Next, in the input noise model, we obtain

$$f^* = -\eta(1+\sigma_S^2)\epsilon_g^*, \tag{130}$$

$$\tilde{g}^* = \frac{\eta}{\pi}\frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)-(\omega^*)^2}}{1+\sigma_T^2}, \tag{131}$$

$$\phi^* = \eta\left[\frac{\eta(1+\sigma_S^2)}{2} - 1\right] \times \left[(1+\sigma_S^2)\epsilon_g^* - \frac{\omega^*\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)-(\omega^*)^2}}{\pi(1+\sigma_T^2)}\right]. \tag{132}$$

From $d\omega/dt = 0$, we obtain

$$\eta\left[1-(\omega^*)^2 + \frac{\eta}{2}(1+\sigma_S^2)(\omega^*)^2\right] \times \frac{\sqrt{(1+\sigma_T^2)(1+\sigma_S^2)-(\omega^*)^2}}{\pi(1+\sigma_T^2)} = \frac{\eta^2}{2}\omega^*(1+\sigma_S^2)^2\epsilon_g^*. \tag{133}$$

Thus,

$$\phi^* - f^* = \frac{\tilde{g}^*}{\omega^*}. \tag{134}$$

Therefore, we obtain

$$\frac{\omega^* \tilde{g}(\omega^*)}{\phi^* - f^*} = (\omega^*)^2. \tag{135}$$

By substituting the asymptotic values calculated above into eqs. (120)–(122), we obtain in both the output and input noise models

$$\lim_{t\to\infty}\frac{q(t+t_1,t+t_2)}{J(t+t_1)J(t+t_2)} = (\omega^*)^2 + [1-(\omega^*)^2]\exp\left[-\frac{\tilde{g}^*}{\omega^*}(t_2-t_1)\right], \tag{136}$$

$$\lim_{t\to\infty}\overline{\overline{Q}}(t) = (\omega^*)^2 + \frac{1}{K}[1-(\omega^*)^2]\left\{1+\frac{2}{K}\sum_{i<j}\exp\left[-\frac{\tilde{g}^*}{\omega^*}(t_j-t_i)\right]\right\}, \tag{137}$$

$$\lim_{t\to\infty}\overline{\overline{\omega}}(t) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}[1-(\omega^*)^2]\left\{1+\frac{2}{K}\sum_{i<j}\exp\left[-\frac{\tilde{g}^*}{\omega^*}(t_j-t_i)\right]\right\}}}. \tag{138}$$

In the above two subsections, we obtained the asymptotic forms of $\overline{\omega}$ for the Perceptron rule eq. (109) and $\overline{\overline{\omega}}$ for the AdaTron rule eq. (138) as $t \to \infty$ in both the output and input noise models. These two quantities are expressed by one formula as

$$\widetilde{\omega}(K) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}[1-(\omega^*)^2]\left\{1+\frac{2}{K}\sum_{i<j}\exp[-a(t_j-t_i)]\right\}}}, \tag{139}$$

where $a = -f^*$ for the Perceptron rule and $a = \tilde{g}^*/\omega^*$ for the AdaTron rule.

Now, let us consider the behavior of this quantity $\widetilde{\omega}(K)$ as a function of the number of students, K, used for the average, in both the output and input noise models. We assume that $t_i = i \times \Delta t$. Then, the summation in $\widetilde{\omega}(K)$ is calculated as

$$\sum_{i<j}\exp[-a(t_j-t_i)] = \frac{1}{\exp(a\Delta t)-1}\left\{K-1-\frac{1-\exp[-a\Delta t(K-1)]}{\exp(a\Delta t)-1}\right\}. \tag{140}$$

Therefore, we obtain

$$\widetilde{\omega}(K) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}[1-(\omega^*)^2]\left(1+\frac{2}{K}\frac{1}{\exp(a\Delta t)-1}\left\{K-1-\frac{1-\exp[-a\Delta t(K-1)]}{\exp(a\Delta t)-1}\right\}\right)}}. \tag{141}$$

Thus, as $K \to \infty$, we obtain

$$\lim_{K\to\infty}\widetilde{\omega}(K) = 1. \tag{142}$$

That is, the direction of the averaged student vector tends to the direction of the teacher vector as the number of averaged student vectors increases.

6.5 Numerical results

In this subsection, we give the results of numerical integrations of differential equations by the RKG method and the results of numerical simulations.

We show the time dependence of $q(t,s)$ and the K dependence of $\widetilde{\omega}(K)$ in the output and input noise models for the Perceptron and AdaTron rules, in Figs. 7 and 8, respectively. The theoretical results agree with the numerical simulation results quite well.

7. Summary and Discussion

In this paper, we studied the supervised online learning of a Perceptron using the Hebbian, Perceptron and AdaTron learning rules in the case in which both the teacher and the student suffer from output or input noise. We mainly focused

on the case in which learning fails in the sense that the student vector does not converge to the teacher vector. To make learning successful, we investigated the optimal learning rate, the control of learning, and time domain ensemble learning.

First, we summarize the results in the case of a constant learning rate. For the Hebbian rule, learning succeeds in both the output and input noise models. For the Perceptron rule, learning fails when the student noise is smaller than the teacher noise in the output noise model, whereas it always fails in the input noise model. For the AdaTron rule, learning always fails in both the output and input noise models.

Next, in the case of the optimal learning rate, we proved that $\omega \to 1$ as $t \to \infty$ and derived asymptotic forms of the optimal learning rate and the generalization error in the three learning rules and for the output and input noise models. When learning fails for constant learning rates, we compared the numerical and theoretical results obtained using the optimal learning rate and found a fairly good agreement between them.

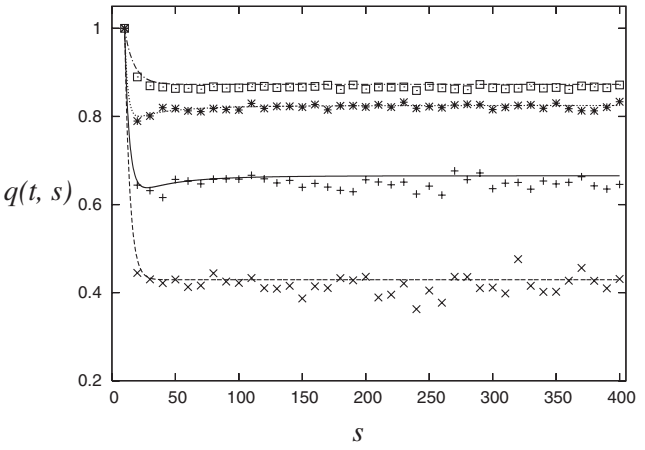

Fig. 7. Time s dependence of $\hat{q}(t,s)$ for $s \geq t$. $t = 10$. $\eta = 1$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and $+$: Perceptron for output noise; dashed curve and $\times$: AdaTron for output noise; dotted curve and $*$: Perceptron for input noise; dotted-dashed curve and square: AdaTron for input noise. For output noise model, $k_T = 0.7$ and $k_S = 0.8$. For input noise model, $\sigma_T = 0.3$ and $\sigma_S = 0.2$.

We also studied the control of learning. For the Perceptron rule in the output noise model, it turned out that learning fails if the student noise is smaller than the teacher noise. Therefore, it is expected that we could make learning successful by reversing the student's output intentionally. By numerical simulations, we confirmed that this method works. Furthermore, we proposed the method to identify the noise parameters k_T and k_S.

Finally, we studied time domain ensemble learning. In the present model, even if learning fails, ω converges to a constant value which is less than 1. This implies that the student vector rotates around the teacher vector with a constant angle. Thus, by taking the average of the student vectors at different times, it is expected that learning succeeds. According to the method developed in our previous study,[6] we analyzed time domain ensemble learning. We found that the formula of the direction cosine between the teacher vector and the averaged student vector can be expressed by the same formula, as in the case in which only the teacher suffers from noise using the terms of the differential equation of $q(t,s)$. We performed numerical simulations for $\hat{q}(t,s)$ and $\tilde{\omega}(K)$ and confirmed that the numerical and theoretical results agree quite well.

Next, let us discuss the results in this paper. Let us compare the convergence speed of learning. If noise does not exist, the asymptotic form of $\tilde{\epsilon}_{g,opt}$ is expressed as $\tilde{\epsilon}_{g,opt} \propto t^{-1/2}$ for the Hebbian rule and $\tilde{\epsilon}_{g,opt} \propto t^{-1}$ for the Perceptron and AdaTron rules, so that the convergence speed of learning is higher in the Perceptron and AdaTron rules than in the Hebbian rule.[5] On the other hand, these behaviors change when both the teacher and the student suffer from noise. In the output noise case, the convergence speed of learning is higher 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.

The present results include the situation in which only the teacher or only the student suffers from external noise. The former case has been studied,[5,6] and the results are obtained by putting $k_S = 1$ or $\sigma_S = 0$. The results of the latter case are obtained by putting $k_T = 1$ or $\sigma_T = 0$. For example, we note that learning succeeds for the Perceptron rule and output noise model, as seen from the Table I by putting $k_T = 1$. As

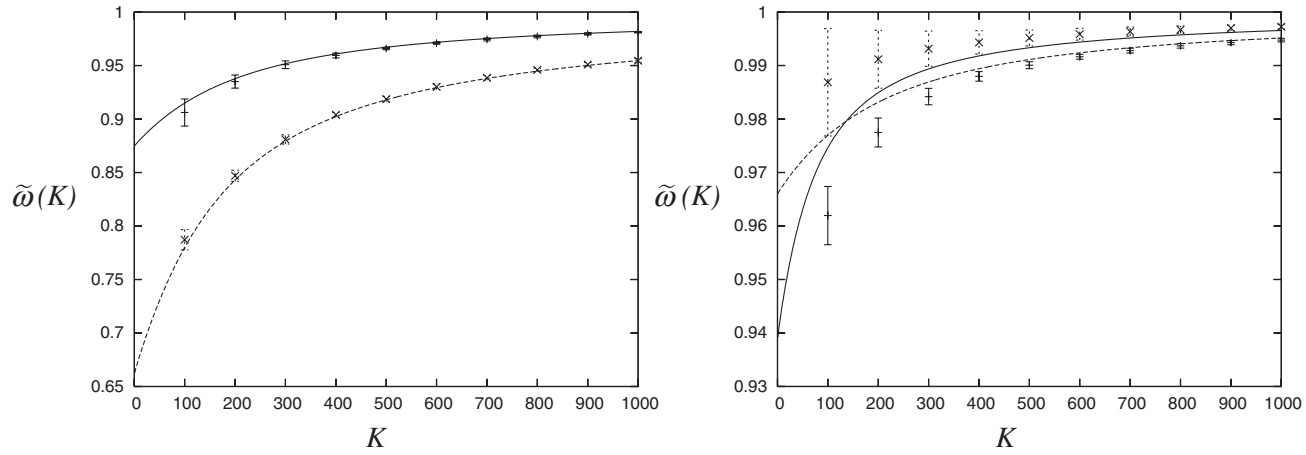

Fig. 8. K dependence of $\tilde{\omega}(K)$. $\eta = 1, \Delta t = 0.1$. Curves are theoretical results and symbols are numerical results which are the averages of 10 samples ($N = 1000$). Solid curve and $+$: Perceptron; dashed curve and $\times$: AdaTron. Left panel: output noise model, $k_T = 0.7, k_S = 0.8$. Right panel: input noise model, $\sigma_T = 0.3, \sigma_S = 0.2$.

a result, we find that the learning speeds and exponents of $\tilde{\eta}_{\rm opt}$ and $\epsilon_{\rm g,opt}$ are the same in the three cases in which only the teacher noise exists, only the student noise exists, and both the teacher and student noise exist.

Appendix: Useful Integration Formulas

We list useful integration formulas in order to derive differential equations for order parameters.

In the following formulas, a and b are real constants unless otherwise noted:

$$\int_{-\infty}^{\infty} Dx\, H(ax+b) = H\left(\frac{b}{\sqrt{1+a^2}}\right), \tag{A·1}$$

$$\int_{-\infty}^{\infty} Dx\, H(ax)H(bx) = \frac{1}{2} - \frac{1}{2\pi}\cos^{-1}\left[\frac{ab}{\sqrt{(1+a^2)(1+b^2)}}\right], \tag{A·2}$$

$$\int_{-\infty}^{\infty} Dx\, x^2 H(ax)H(bx) = \frac{1}{2} - \frac{1}{2\pi}\cos^{-1}\left[\frac{ab}{\sqrt{(1+a^2)(1+b^2)}}\right] + \frac{ab(2+a^2+b^2)}{2\pi(1+a^2)(1+b^2)\sqrt{1+a^2+b^2}}, \tag{A·3}$$

$$\int_{0}^{\infty} Dx\, H(ax) = \frac{1}{2\pi}\cos^{-1}\left(\frac{a}{\sqrt{1+a^2}}\right) = \frac{1}{2} - \frac{1}{2\pi}\cos^{-1}\left(\frac{-a}{\sqrt{1+a^2}}\right), \tag{A·4}$$

$$\int_{0}^{\infty} Dx\, H(ax) = \frac{1}{2\pi}\tan^{-1}\left(\frac{1}{a}\right) \quad a>0, \tag{A·5}$$

where $Dx = {\rm d}x/\sqrt{2\pi}\exp(-x^2/2)$, $H(x) = \int_x^\infty Dt$, and $\cos^{-1}(x)$ and $\tan^{-1}(x)$ are principal values.

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: *Dynamics of Learning*, ed. J. L. van Hemmen, E. Domany, and K. Schulten (Springer, New York, 1991) Physics of Neural Networks.
5) T. Uezu, Y. Maeda, and S. Yamaguchi: J. Phys. Soc. Jpn. **75** (2006) 114007.
6) T. Uezu, S. Miyoshi, M. Izuo, and M. Okada: J. Phys. Soc. Jpn. **76** (2007) 114006.
7) T. L. H. Watkin, A. Rau, and M. Biehl: Rev. Mod. Phys. **65** (1993) 499.
8) O. Kinouchi and N. Caticha: J. Phys. A **25** (1992) 6243.
9) O. Kinouchi and N. Caticha: J. Phys. A **26** (1993) 6161.
10) C. W. H. Mace and A. C. C. Coolen: Stat. Comput. **8** (1998) 55.
11) *On-Line Learning in Neural Networks*, ed. D. Saad (Cambridge University Press, Cambridge, U.K., 2001).
12) A. Engel and C. Van den Broeck: *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, U.K., 2001).
13) Y. Maeda: Masters Thesis, Graduate School of Humanities and Sciences, Nara Women's University, Nara (2002) [in Japanese].
14) G. Reents and R. Urbanczik: Phys. Rev. Lett. **80** (1998) 5445.
15) M. Biehl, P. Riegler, and M. Stechert: Phys. Rev. E **52** (1995) R4624.
16) We can derive the same result using an exact argument.
17) S. Miyoshi, T. Uezu, and M. Okada: J. Phys. Soc. Jpn. **75** (2006) 084007.
18) The equation of $\lim_{t\to\infty} q(t+t_1, t+t_2)$ for the Perceptron learning rule, eq. (56) in ref. 6, is wrong. The correct one is
$$\lim_{t\to\infty} q(t+t_1,t+t_2) = (J^*)^2\left\{(\omega^*)^2 + [1-(\omega^*)^2]\exp\left[-\eta\frac{1}{\sqrt{2\pi}}\frac{1}{J^*}(t_2-t_1)\right]\right\}.$$