# Theory of Time Domain Ensemble On-line Learning of Perceptron under the Existence of External Noise

Tatsuya UEZU*, Seiji MIYOSHI[1], Mika IZUO[2] and Masato OKADA[3]

*Graduate School of Humanities and Sciences, Nara Women's University, Kitauoyanichi-machi, Nara 630-8506*

[1] *Department of Electronic Engineering, Kobe City College of Technology, 8-3 Gakuenhigashimachi, Nishi-ku, Kobe 651-2194*

[2] *Department of Physics, Faculty of Science, Nara Women's University, Kitauoyanichi-machi, Nara 630-8506*

[3] *Division of Transdisciplinary of Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561*

(Received September 12, 2007)

We analyze the time domain ensemble on-line learning of a Perceptron under the existence of external noise. We adopt three typical learning rules, Hebbian, Perceptron and AdaTron rules. We treat the input and output noises. 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 an averaging method and give theoretical analysis of the method. We obtain the precise formula for the overlap between the teacher vector and the time averaged student vector for $t \to \infty$ limit as a function of the number of student vectors to be averaged. We compare the theoretical results with numerical simulations and find that the theoretical results agree quite well with the numerical simulations.

KEYWORDS: perceptron, on-line learning, noise, time domain ensemble

## 1. Introduction

We study the on-line learning by a single Perceptron[1)] from signals produced by a single teacher. 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)] In this paper, we treat the on-line learning of a Perceptron. In the on-line learning, an example vector is chosen randomly and used in order to update a student vector by a learning algorithm. When the student vector is updated next, another example vector is chosen randomly. In contrast to this, for batch learning, many example vectors are stored and all of them are used simultaneously for learning. Although on-line learning seems less efficient than batch learning, in some situations the performance of on-line learning is comparable to that of batch learning.

*E-mail address: uezu@ki-rin.phys.nara-wu.ac.jp

In the framework of on-line learning, we have interest in ensemble learning in time domain.[12–14] Here, we briefly explain ensemble learning by using an example of the on-line learning of Perceptrons. Suppose that there are several student Perceptrons which learn from a teacher Perceptron. When initial students' vectors are randomly distributed, it is shown that the generalization error of the averaged student vector becomes smaller than that of one student vector.[11] In contrast with ensemble learning explained above, ensemble learning in time domain uses only one student who learns from a teacher. Ensemble consists of vectors of the student at different times. A typical situation in which ensemble learning in time domain is efficient is when the student vector rotates around the teacher vector. In this situation, by taking the average of normalized student vectors, the direction of the averaged vector is closer to that of the teacher vector than that of the student vector in any instance.

In the previous study,[13] we investigated time domain ensemble learning for Perceptrons numerically and found that learning is improved by taking average of student vectors over different times. Further, in,[14] we analytically studied time domain ensemble learning for linear Perceptrons.

The main purpose of the present paper is to give a theory for time domain ensemble learning for Perceptrons. We adopt the Hebbian, Perceptron and AdaTron rules as learning rules, and for both the input and output noise cases, we obtain the differential equations for order parameters. We obtain the precise formula for the overlap between the teacher vector and the time averaged student vector for $t \to \infty$ limit as a function of the number of students to be averaged. We compare the theoretical results with numerical simulations and find that the theoretical results agree quite well with the numerical simulations.

The paper is organized as follows: In §2, the formulation of the time domain ensemble learning is given. In §3, we derive differential equations for relevant quantities and obtain asymptotic forms of overlap between the teacher vector and the time averaged student vector. In §4, we give numerical results. Section 5 is devoted to a summary and discussions.

## 2. Formulation in time domain ensemble learning

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 $\xi_i$ takes $\pm 1$ and is drawn independently with the probability $P(\xi = 1) = 1 - P(\xi = -1) = \dfrac{1}{2}$. 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 $\frac{1+k}{2}$. In the input noise model, $T$ is given by

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

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

$$\mathcal{P}(y) = H(-\frac{y}{\sigma}), \tag{5}$$

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

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

where $\eta$ 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)]

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

$$\frac{dR}{dt} = \eta\langle(\boldsymbol{B}\cdot\boldsymbol{\xi})T\mathcal{F}\rangle_\Xi. \tag{11}$$

Here, we assume self-averaging[15)] and $\langle\cdot\rangle_\Xi$ denotes the average over examples and noises. Let us define $J=|\boldsymbol{J}|$, $\widehat{\boldsymbol{J}}\equiv\frac{\boldsymbol{J}}{J}$ 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

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

$$\frac{dR}{dt} = \eta\langle yT\mathcal{F}[J;Jx,T]\rangle_\Xi. \tag{13}$$

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

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

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

The generalization error $E$ is given by

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

Further, we consider a two time correlation function $q(t,s) \equiv \boldsymbol{J}(t) \cdot \boldsymbol{J}(s)$.[14] The differential equation for $q(t,s)$ with respect to $s$ for $t \leq s$ is given by

$$\frac{\partial q(t,s)}{\partial s} = \eta J(t) \langle x_s^t T(s) \mathcal{F}(s) \rangle_{\Xi_s}, \tag{17}$$

where $x_s^s = \widehat{\boldsymbol{J}}(s) \cdot \boldsymbol{\xi}_s$ and $x_s^t = \widehat{\boldsymbol{J}}(t) \cdot \boldsymbol{\xi}_s$. $\boldsymbol{\xi}_s$ is a sample given at time $s$ and $\langle \rangle_{\Xi_s}$ denotes the average over samples at time $s$. Here we abbreviate as $\mathcal{F}(s) = \mathcal{F}[J(s); J(s)x_s^s, T(s)]$.

Now, let us formulate the time domain ensemble. We define the time averaged student vectors $\bar{\boldsymbol{J}}(t)$ and $\bar{\widehat{\boldsymbol{J}}}(t)$ as follows.

$$\overline{\boldsymbol{J}}(t) \equiv \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{J}(t+t_i), \tag{18}$$

$$\overline{\widehat{\boldsymbol{J}}}(t) \equiv \frac{1}{K} \sum_{i=1}^{K} \widehat{\boldsymbol{J}}(t+t_i) = \frac{1}{K} \sum_{i=1}^{K} \frac{\boldsymbol{J}(t+t_i)}{J(t+t_i)}, \tag{19}$$

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

$$\overline{R}(t) \equiv \boldsymbol{B} \cdot \overline{\boldsymbol{J}}(t) = \frac{1}{K} \sum_{i=1}^{K} R(t+t_i), \tag{20}$$

$$\overline{Q}(t) \equiv \overline{\boldsymbol{J}}(t)^2 = \frac{2}{K^2} \sum_{i<j} q(t+t_i, t+t_j) + \frac{1}{K^2} \sum_{i=1}^{K} J(t+t_i)^2, \tag{21}$$

$$\overline{\omega}(t) \equiv \frac{\overline{R}(t)}{\sqrt{\overline{Q}(t)}}, \tag{22}$$

$$\overline{\widehat{R}}(t) \equiv \boldsymbol{B} \cdot \overline{\widehat{\boldsymbol{J}}}(t) = \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{B} \cdot \widehat{\boldsymbol{J}}(t+t_i) = \frac{1}{K} \sum_{i=1}^{K} \omega(t+t_i), \tag{23}$$

$$\overline{\widehat{Q}}(t) \equiv \overline{\widehat{\boldsymbol{J}}}(t)^2 = \frac{2}{K^2} \sum_{i<j} \frac{q(t+t_i, t+t_j)}{J(t+t_i) J(t+t_j)} + \frac{1}{K}, \tag{24}$$

$$\overline{\widehat{\omega}}(t) \equiv \frac{\overline{\widehat{R}}(t)}{\sqrt{\overline{\widehat{Q}}(t)}} \tag{25}$$

In our previous study, we found that it seems that $\overline{\omega}(0)$ tends to 1 for the Perceptron rule and $\overline{\widehat{\omega}}(0)$ tends to 1 for the AdaTron rule as $K \to \infty$[12,13] . In this paper, we derive the asymptotic expressions for $\overline{\omega}(t)$ and $\overline{\widehat{\omega}}(t)$ as $t \to \infty$ for finite $K$, and discuss the efficiency of the time domain ensemble learning.

## 3. Differential equations and asymptotic behaviors

In this section, first we derive differential equations for $q(t,s)$ with respect to $s(\geq t)$ both in the output and input noise models. And then, in order to obtain asymptotic forms of $\overline{\omega}$

and $\overline{\widehat{\omega}}$, we study the asymptotic behaviour of $q(t+t_i, t+t_j)$ and $\frac{q(t+t_i,t+t_j)}{J(t+t_i)J(t+t_j)}$. In the below, we give the differential equation only for $q(t,s)$. See ref.[13] on differential equations for $R, J$ and $\omega$.

Now, let us study the output noise model, where $\mathcal{P}(y) = \frac{1}{2}(1+k\ \mathrm{sgn}(y))$. Then, we obtain the differential equation for $q(t,s)$ as

$$\frac{\partial q(t,s)}{\partial s} = \frac{\eta}{2} J(t) \langle x_s^t \{(1+k\ \mathrm{sgn}(y_s))\mathcal{F}_+(s) - (1-k\ \mathrm{sgn}(y_s))\mathcal{F}_-(s)\} \rangle_{x_s^t, x_s^s, y_s}, \quad (26)$$

where $\mathcal{F}_+(s) = \mathcal{F}[J(s); J(s)x_s^s, +1]$ and $\mathcal{F}_- = \mathcal{F}[J(s); J(s)x_s^s, -1]$, $\langle\cdot\rangle_{x_s^t, x_s^s, y_s}$ denotes the average over the Gaussian distribution of $x_s^t, x_s^s$ and $y_s$ with $\langle (x_s^t)^2 \rangle = 1, \langle (x_s^s)^2 \rangle = 1, \langle y_s^2 \rangle = 1, \langle x_s^t x_s^s \rangle = \frac{q(t,s)}{J(t)J(s)}, \langle x_s^t y_s \rangle = \omega(t)$ and $\langle x_s^s y_s \rangle = \omega(s)$. The initial condition for this equation is $q(t,t) = J(t)^2$.

The generalization error $E = \langle \Theta(-TS) \rangle_\Xi$ is given by

$$E(\omega) = \frac{1-k}{2} + \frac{k}{\pi} \cos^{-1}(\omega). \quad (27)$$

By performing several integrations in eq.(26), we obtain the differential equation for each learning rule. We omit the details of the calculation and summarize the learning behavior in each learning rule.

In the Hebbian rule, the differential equation for $q$ is

$$\frac{\partial q(t,s)}{\partial s} = k\eta \sqrt{\frac{2}{\pi}} R(t) \text{ for } s \geq t. \quad (28)$$

This case has been studied previously and the differential equations for $R$ and $J$ have been solved analytically.[16] The solutions for $R, J$ and $q$ with initial conditions $R(0) = 0, J(0) = 1$ and $q(t,t) = J(t)^2$ are

$$R(t) = \eta k \sqrt{\frac{2}{\pi}} t, \quad (29)$$

$$J(t) = \sqrt{1 + \eta^2 t (1 + \frac{2}{\pi} k^2 t)}, \quad (30)$$

$$q(t,s) = k\eta \sqrt{\frac{2}{\pi}} R(t)(s-t) + J(t)^2 \text{ for } s \geq t. \quad (31)$$

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

In the Perceptron rule, the differential equation for $q$ is

$$\frac{\partial q(t,s)}{\partial s} = \frac{k\eta}{\sqrt{2\pi}} R(t) - \frac{\eta}{\sqrt{2\pi}} \frac{q(t,s)}{J(s)} \text{ for } s \geq t. \quad (32)$$

The stationary state is given by

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

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

In the AdaTron rule, the differential equation for $q$ is

$$\frac{\partial q(t,s)}{\partial s} = \eta[-E(\omega(s))q(t,s) + \frac{k}{\pi}R(t)J(s)\sqrt{1-\omega(s)^2}] \text{ for } s \geq t. \tag{33}$$

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, $\omega_A^*$ is the solution of $\frac{d\omega}{dt} = 0$ and is less than 1. As in the case of the Perceptron rule, learning fails.

Now, we study the input noise model, where $\mathcal{P}(y) = H(-\frac{y}{\sigma})$. Then, we obtain

$$\frac{\partial q(t,s)}{\partial s} = \eta J(t)\langle x_s^t\{H(-\frac{y_s}{\sigma})\mathcal{F}_+(s) - H(\frac{y_s}{\sigma})\mathcal{F}_-(s)\}\rangle_{x_s^t, x_s^s, y_s}. \tag{34}$$

The generalization error is given by

$$E(\omega) = \frac{1}{\pi}\cos^{-1}\left(\frac{\omega}{\sqrt{1+\sigma^2}}\right). \tag{35}$$

For the Hebbian rule, the differential equations of order parameters for the input noise model are obtained by those for the output noise model replacing $k$ by $\frac{1}{\sqrt{1+\sigma^2}}$. Therefore, we obtain $\lim_{t\to\infty}\overline{\widehat{\omega}}(t) = 1$.

In the Perceptron rule, the differential equation for $q$ is

$$\frac{\partial q(t,s)}{\partial s} = \frac{\eta}{\sqrt{2\pi}}\left(\frac{R(t)}{\sqrt{1+\sigma^2}} - \frac{q(t,s)}{J(s)}\right) \text{ for } s \geq t. \tag{36}$$

The stationary state is given by

$$J_P^* = \frac{1+\sigma^2}{\sigma^2}\eta\sqrt{\frac{\pi}{2}}E(\omega_P^*), \quad \omega_P^* = \frac{1}{\sqrt{1+\sigma^2}}. \tag{37}$$

Thus, learning fails for $\sigma > 0$. In the AdaTron rule,[17] the equation for $q$ is

$$\frac{\partial q(t,s)}{\partial s} = \frac{\eta}{\pi}R(t)J(s)\frac{\sqrt{1+\sigma^2-\omega(s)^2}}{1+\sigma^2} - \eta q(t,s)E(\omega(s)) \text{ for } s \geq t. \tag{38}$$

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, $\omega_A^*$ is the solution of $\frac{d\omega}{dt} = 0$ and is less than 1. Thus, learning fails.

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

$$\frac{\partial}{\partial s}q(t,s) = f(s)q(t,s) + g(t,s). \tag{39}$$

Its solution for $s \geq t$ with initial condition $q(t,t) = J(t)^2$ at $s = t$ is given by

$$q(t,s) = \{J(t)^2 + \int_t^s ds'' g(t,s'')e^{-\int_t^{s''} ds' f(s')}\}e^{\int_t^s ds''' f(s''')}. \tag{40}$$

Thus, we obtain for $t_1 \leq t_2$

$$\begin{aligned} q(t+t_1, t+t_2) = &\{J(t+t_1)^2 + \int_0^{t_2-t_1} d\tau g(t+t_1, \tau+t+t_1)e^{-\int_0^\tau du f(u+t+t_1)}\} \\ &\times e^{\int_0^{t_2-t_1} dv f(v+t+t_1)}. \end{aligned} \tag{41}$$

We note that $g(t,s)$ is the product of a function of $t$ and that of $s$. For the Perceptron rule,

$f(s)$ and $g(t,s)$ are as follows.

$$f(s) = -\frac{\eta}{\sqrt{2\pi}}\frac{1}{J(s)} \text{ for output and input noise models,} \tag{42}$$

$$g(t,s) = \frac{k\eta}{\sqrt{2\pi}}R(t) \text{ for output noise model,} \tag{43}$$

$$.g(t,s) = \frac{\eta}{\sqrt{2\pi}}\frac{R(t)}{\sqrt{1+\sigma^2}} \text{ for input noise model.} \tag{44}$$

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

$$f(s) = -\eta E(\omega(s)) \text{ for output and input noise models,} \tag{45}$$

$$g(t,s) = \eta\frac{k}{\pi}R(t)J(s)\sqrt{1-\omega(s)^2} \text{ for output noise model,} \tag{46}$$

$$g(t,s) = \frac{\eta}{\pi}R(t)J(s)\frac{\sqrt{1+\sigma^2-\omega(s)^2}}{1+\sigma^2} \text{ for input noise model.} \tag{47}$$

Thus, it follows that as $t\to\infty$, $J(t)\to J^*, f(t)\to f^*$ and $g(t+\alpha,t+\beta)\to g^*$ for any constants $\alpha$ and $\beta$. In both the output and input models, for the Perceptron rule, these limiting values are all finite, but for the AdaTron rule, $J^*$ and $g^*$ are 0. Then, we obtain for the Perceptron rule

$$\lim_{t\to\infty} q(t+t_1,t+t_2) = ((J^*)^2+\frac{g^*}{f^*})e^{f^*(t_2-t_1)} - \frac{g^*}{f^*}, \tag{48}$$

and for the AdaTron rule

$$\lim_{t\to\infty} q(t+t_1,t+t_2) = 0. \tag{49}$$

Thus, we consider two cases separately in the following subsections.

### *3.1 Case of Perceptron learning rule*

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

$$\lim_{t\to\infty}\overline{R}(t) = \lim_{t\to\infty}\frac{1}{K}\sum_{i=1}^{K}R(t+t_i) = R^* = J^*\omega^*, \tag{50}$$

$$\begin{aligned}\lim_{t\to\infty}\overline{Q}(t) &= \lim_{t\to\infty}\{\frac{2}{K^2}\sum_{i<j}q(t+t_i,t+t_j)+\frac{1}{K^2}\sum_{i=1}^{K}J(t+t_i)^2\}\\ &= \frac{2}{K^2}\sum_{i<j}((J^*)^2+\frac{g^*}{f^*})e^{f^*(t_j-t_i)} - \frac{K-1}{K}\frac{g^*}{f^*}+\frac{(J^*)^2}{K}.\end{aligned} \tag{51}$$

$$\lim_{t\to\infty}\overline{\omega}(t) = \frac{J^*\omega^*}{\sqrt{\frac{2}{K^2}\sum_{i<j}((J^*)^2+\frac{g^*}{f^*})e^{f^*(t_j-t_i)} - \frac{K-1}{K}\frac{g^*}{f^*}+\frac{(J^*)^2}{K}}}. \tag{52}$$

In both output and input noise models, asymptotic values are calculated as

$$f^* = -\frac{\eta}{\sqrt{2\pi}J^*}, \tag{53}$$

$$g^* = \frac{\eta}{\sqrt{2\pi}} J^* (\omega^*)^2, \tag{54}$$

$$-\frac{g^*}{f^*} = (J^* \omega^*)^2. \tag{55}$$

Thus, we get

$$\lim_{t\to\infty} q(t+t_1, t+t_2) = (J^*)^2 [(\omega^*)^2 - \left(1-(\omega^*)^2\right) e^{-\eta \frac{1}{\sqrt{2\pi}} \frac{1}{J^*}(t_2-t_1)}], \tag{56}$$

$$\lim_{t\to\infty} \overline{Q}(t) = (J^*)^2 [(\omega^*)^2 + \frac{1}{K}\left(1-(\omega^*)^2\right) \left(1 + \frac{2}{K} \sum_{i<j} e^{-\eta \frac{1}{\sqrt{2\pi}} \frac{1}{J^*}(t_j - t_i)}\right)]. \tag{57}$$

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

$$\lim_{t\to\infty} \overline{\omega}(t) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}\left(1-(\omega^*)^2\right)\left(1 + \frac{2}{K}\sum_{i<j} e^{-\eta \frac{1}{\sqrt{2\pi}} \frac{1}{J^*}(t_j - t_i)}\right)}}. \tag{58}$$

### *3.2 Case of AdaTron learning rule*

In this subsection, we consider the case of the AdaTron rule in which $J(t) \to 0$. Since $J^*$ and $g^*$ are zero, we consider $\overline{\widetilde{\omega}}(t)$. When $t \to \infty$, we obtain,

$$\lim_{t\to\infty} \overline{\widetilde{R}}(t) = \lim_{t\to\infty} \frac{1}{K} \sum_{i=1}^{K} \omega(t+t_i) = \omega^*, \tag{59}$$

$$\lim_{t\to\infty} \overline{\widetilde{Q}}(t) = \frac{2}{K^2} \sum_{i<j} \lim_{t\to\infty} \frac{q(t+t_i, t+t_j)}{J(t+t_i)J(t+t_j)} + \frac{1}{K}. \tag{60}$$

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

$$g(t,s) = R(t) J(s) \tilde{g}(\omega(s)). \tag{61}$$

Thus, we obtain

$$\begin{aligned} \lim_{t\to\infty} \frac{q(t+t_1, t+t_2)}{J(t+t_1)J(t+t_2)} &= \lim_{t\to\infty} \{ \frac{J(t+t_1)}{J(t+t_2)} \\ &\quad + \int_0^{t_2-t_1} d\tau \frac{J(\tau+t+t_1)}{J(t+t_2)} \omega(t+t_1) \tilde{g}(\omega(\tau+t+t_1)) e^{-\int_0^\tau du f(u+t+t_1)} \} \\ &\quad \times e^{\int_0^{t_2-t_1} dv f(v+t+t_1)} \qquad (62) \\ &= \{ B(t_1,t_2) + \int_0^{t_2-t_1} d\tau B(\tau+t_1, t_2) \omega^* \tilde{g}(\omega^*) e^{-f^*\tau} \} e^{f^*(t_2-t_1)}. \qquad (63) \end{aligned}$$

Here, we define

$$B(t_1,t_2) \equiv \lim_{t\to\infty} \frac{J(t+t_1)}{J(t+t_2)}. \tag{64}$$

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

$$\frac{dJ(t)}{dt} = J(t)\phi(\omega), \tag{65}$$

$$\frac{d\omega}{dt} = \psi(\omega). \tag{66}$$

By solving eq.(66) we obtain $\omega = \omega(t)$. Using this solution, $J(t)$ is given by

$$J(t) = J(0)e^{\int_0^t dt' \phi(\omega(t'))}. \tag{67}$$

Thus,

$$\frac{J(t+t_1)}{J(t+t_2)} = e^{-\int_{t+t_1}^{t+t_2} dt' \phi(\omega(t'))} = e^{-\int_0^{t_2-t_1} d\tau \phi(\omega(\tau+t+t_1))}. \tag{68}$$

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

$$B(t_1, t_2) \equiv \lim_{t\to\infty} e^{-\int_0^{t_2-t_1} d\tau \phi(\omega(\tau+t+t_1))} = e^{-\phi^*(t_2-t_1)}, \tag{69}$$

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

$$\lim_{t\to\infty} \frac{q(t+t_1, t+t_2)}{J(t+t_1)J(t+t_2)} = e^{(f^*-\phi^*)(t_2-t_1)}\{1 - \frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*}\} + \frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*}. \tag{70}$$

Then,

$$\lim_{t\to\infty} \overline{\widehat{Q}}(t) = \frac{2}{K^2}\sum_{i<j} e^{(f^*-\phi^*)(t_j-t_i)}\{1 - \frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*}\} + \frac{K-1}{K}\frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*} + \frac{1}{K}. \tag{71}$$

Thus, we obtain

$$\lim_{t\to\infty} \overline{\widehat{\omega}}(t) = \frac{\omega^*}{\sqrt{\frac{2}{K^2}\sum_{i<j} e^{(f^*-\phi^*)(t_j-t_i)}\{1 - \frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*}\} + \frac{K-1}{K}\frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*} + \frac{1}{K}\}}}. \tag{72}$$

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

$$f^* = -\eta E^*, \tag{73}$$

$$\tilde{g}^* = \eta\frac{k}{\pi}\sqrt{1-(\omega^*)^2}, \tag{74}$$

$$\phi^* = \eta\left(\frac{\eta}{2} - 1\right)\left(E^* - \frac{k}{\pi}\omega^*\sqrt{1-(\omega^*)^2}\right). \tag{75}$$

Now, we estimate $\frac{\omega^*\tilde{g}(\omega^*)}{\phi^* - f^*}$. From $\frac{d\omega}{dt} = 0$, we obtain

$$\frac{k\eta}{\pi}(1-(\omega^*)^2)^{3/2} = \frac{\eta^2}{2}\omega^*\left(E^* - \frac{k}{\pi}\omega^*\sqrt{1-(\omega^*)^2}\right). \tag{76}$$

From this, we get

$$E^* = \frac{2k}{\pi\eta\omega^*}\left(1 - (1-\frac{\eta}{2})(\omega^*)^2\right)\sqrt{1-(\omega^*)^2}. \tag{77}$$

Thus,

$$\phi^* - f^* = \eta\left(\frac{\eta}{2} - 1\right)\left(E^* - \frac{k}{\pi}\omega^*\sqrt{1-(\omega^*)^2}\right) + \eta E^* \tag{78}$$

$$= \frac{k\eta}{\pi\omega^*}\sqrt{1-(\omega^*)^2} = \frac{1}{\omega^*}\tilde{g}^*. \tag{79}$$

Therefore, we obtain

$$\frac{\omega^* \tilde{g}(\omega^*)}{\phi^* - f^*} = (\omega^*)^2. \tag{80}$$

Next, in the input noise model, we obtain

$$f^* = -\eta E^*, \tag{81}$$

$$\tilde{g}^* = \frac{\eta}{\pi} \frac{\sqrt{1+\sigma^2-(\omega^*)^2}}{1+\sigma^2}, \tag{82}$$

$$\phi^* = \eta \left(\frac{\eta}{2} - 1\right) \left\{ E^* - \frac{\omega^* \sqrt{1+\sigma^2-(\omega^*)^2}}{\pi(1+\sigma^2)} \right\}, \tag{83}$$

From $\frac{d\omega}{dt} = 0$, we get

$$\eta(1-(\omega^*)^2 + \frac{\eta}{2}(\omega^*)^2) \frac{\sqrt{1+\sigma^2-(\omega^*)^2}}{\pi(1+\sigma^2)} = \frac{\eta^2}{2} \omega^* E^*. \tag{84}$$

Thus,

$$\begin{aligned} \phi^* - f^* &= \eta \left(\frac{\eta}{2} - 1\right) \left\{ E^* - \frac{\omega^* \sqrt{1+\sigma^2-(\omega^*)^2}}{\pi(1+\sigma^2)} \right\} + \eta E^* \\ &= \frac{1}{\omega^*} \eta \frac{\sqrt{1+\sigma^2-(\omega^*)^2}}{\pi(1+\sigma^2)} = \frac{1}{\omega^*} \tilde{g}^*. \end{aligned} \tag{85}$$

Therefore, we obtain

$$\frac{\omega^* \tilde{g}(\omega^*)}{\phi^* - f^*} = (\omega^*)^2. \tag{86}$$

By substituting asymptotic values calculated above into eqs.(**??**), (71) and (72), we obtain in both the output and input noise models,

$$\lim_{t\to\infty} \frac{q(t+t_1, t+t_2)}{J(t+t_1)J(t+t_2)} = (\omega^*)^2 + \left(1-(\omega^*)^2\right) e^{-\frac{\tilde{g}^*}{\omega^*}(t_2-t_1)}, \tag{87}$$

$$\lim_{t\to\infty} \overline{\widehat{Q}}(t) = (\omega^*)^2 + \frac{1}{K}\left(1-(\omega^*)^2\right) \left(1 + \frac{2}{K} \sum_{i<j} e^{-\frac{\tilde{g}^*}{\omega^*}(t_j-t_i)}\right), \tag{88}$$

$$\lim_{t\to\infty} \overline{\widehat{\omega}}(t) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}\left(1-(\omega^*)^2\right)\left(1 + \frac{2}{K}\sum_{i<j} e^{-\frac{\tilde{g}^*}{\omega^*}(t_j-t_i)}\right)}}. \tag{89}$$

In the above two subsections, we obtained the asymptotic forms of $\overline{\omega}$ for the Perceptron rule (58) and $\overline{\widehat{\omega}}$ for the AdaTron rule (89) as $t \to \infty$ in both the output and input noise models. These two quantities are expressed by one formula as

$$\widetilde{\omega}(K) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}\left(1-(\omega^*)^2\right)\left(1 + \frac{2}{K}\sum_{i<j} e^{-a(t_j-t_i)}\right)}}, \tag{90}$$

where $a = \frac{\eta}{\sqrt{2\pi}} \frac{1}{J^*}$ for the Perceptron rule and $a = \frac{\tilde{g}^*}{\omega^*}$ for the AdaTron rule.

Now, let us consider the behavior of this quantity $\widetilde{\omega}(K)$ as a function of the number of averaged students $K$, 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

$$\sum_{i<j} e^{-a(t_j - t_i)} = \frac{1}{e^{a\Delta t - 1}} \{K - 1 - \frac{1 - e^{-a\Delta t(K-1)}}{e^{a\Delta t - 1}}\}. \tag{91}$$

Therefore, we obtain

$$\widetilde{\omega}(K) = \frac{\omega^*}{\sqrt{(\omega^*)^2 + \frac{1}{K}(1 - (\omega^*)^2)\left(1 + \frac{2}{K}\frac{1}{e^{a\Delta t - 1}}\{K - 1 - \frac{1 - e^{-a\Delta t(K-1)}}{e^{a\Delta t - 1}}\}\right)}}. \tag{92}$$

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

$$\lim_{K \to \infty} \widetilde{\omega}(K) = 1. \tag{93}$$

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.

## 4. Numerical results

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

In figs.1 and 2, we display the time dependence of $q(t, s)$ in the output and input noise models, respectively. In fig.3, we display the $K$ dependence of $\widetilde{\omega}(K)$ in the output and input noise models. Theoretical results agree with numerical simulations quite well.

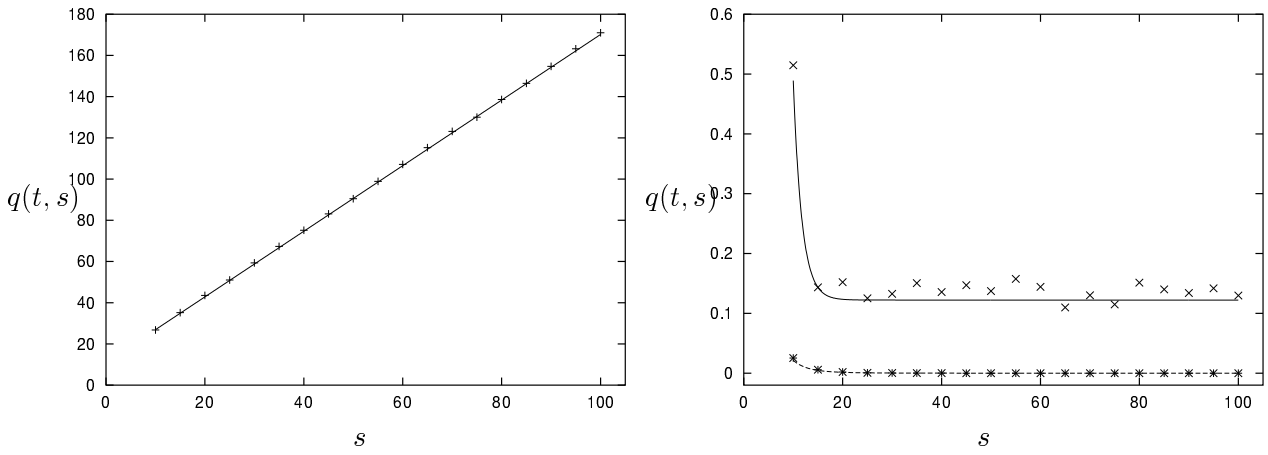


Fig. 1. Output noise model. $\eta = 1, k = 0.5$. Time $s$ dependence of $q(t, s)$ for $s \geq t$. $t = 10$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Left panel. Solid curve and +: Hebbian. Right panel. Solid curve and ×: Perceptron, dashed curve and ∗: AdaTron.

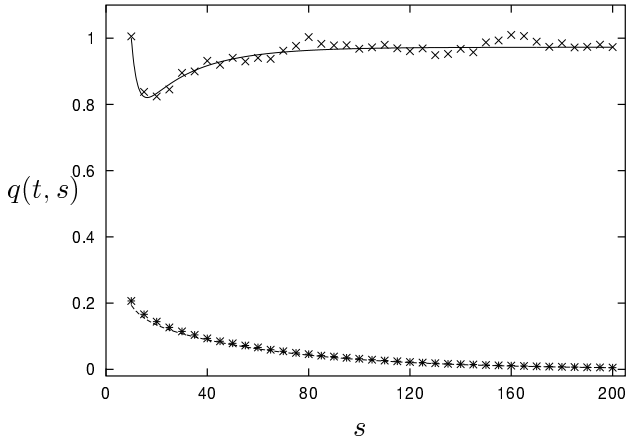


Fig. 2. Input noise model. $\eta = 1, \sigma = 0.5$. Time $s$ dependence of $q(t,s)$ for $s \geq t$. $t = 10$. Curves are theoretical results (RKG) and symbols are numerical results ($N = 1000$). Solid curve and $\times$: Perceptron, dashed curve and $*$: AdaTron.

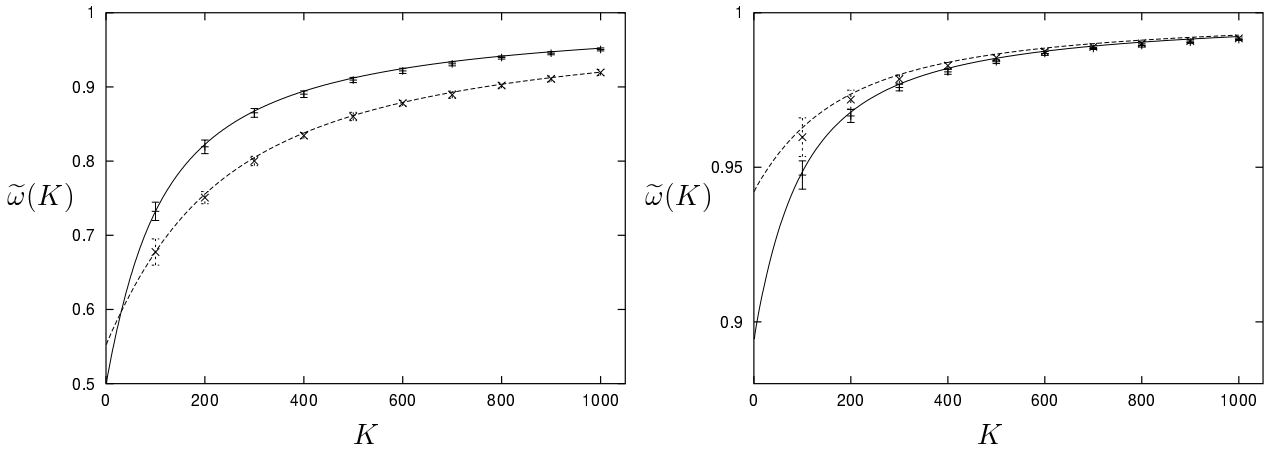


Fig. 3. $K$ dependence of $\widetilde{\omega}(K)$. $\eta = 1, \Delta t = 0.1$. Curves are theoretical results and symbols are numerical results which are the average of 10 samples ($N = 1000$). Solid curve and $+$: Perceptron, dotted curve and $\times$ : AdaTron. Left panel. Output noise model. $k = 0.5$ Right panel. Input noise model. $\sigma = 0.5$

## 5. Summary and discussion

In this paper, we studied the time domain ensemble on-line learning of a perceptron under the existence of input and output noises. We considered three learning rules, the Hebbian, Perceptron and AdaTron rules. In the Hebbian rule, learning succeeds in the sense that the student vector tends to the teacher vector. So, we focused on the study of the Perceptron and AdaTron rules. We obtained the asymptotic forms of the two time correlation functions $q(t+t_1, t+t_2)$ for the Perceptron rule and $\frac{q(t+t_1,t+t_2)}{J(t+t_1)J(t+t_2)}$ for the AdaTron rule in both the input and output noises as $t \to \infty$, eqs.(48) and (70), respectively. Using these forms, we gave the precise formula for the overlap between the teacher vector and the time averaged student

vector for the Perceptron and AdaTron rules in both the input and output noises, eq.(90). By this formula, we conclude that the direction of the averaged student vector tends to the direction of the teacher vector as the number of student vectors increases.

We performed numerical simulations and estimated $q(t,s)$ and $\widetilde{\omega}(K)$ for the Perceptron and AdaTron rules in both the input and output noise models. We compared these results with theoretical ones, and confirmed that the theoretical and numerical results agree quite well.

In this paper, we considered the situation that the teacher vector suffers from external noise but the student vector does not. Our present analysis can be applied to the situation that both the teacher and student vectors suffer from external noise. This study will be reported elsewhere.

## References


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, in *Physics of Neural Networks*, eds. J. L. van Hemmen, E. Domany and K. Schulten (Springer-Verlag, New York, 1991).
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: Statistics and Computing **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) S. Miyoshi, K. Hara and M. Okada: Phys. Rev. E **71** (2005) 036116
12) Y. Maeda: Masters Thesis (in Japanese), Graduate School of Humanities and Sciences, Nara Women's University, Nara, 2002.
13) T. Uezu, Y. Maeda and S. Yamaguchi, J. Phys. Soc. Jpn. **75** (2006) 114007
14) S. Miyoshi, T. Uezu and M. Okada: Phys. Soc. Jpn. **75** (2006) 084007.
15) G. Reents and R. Urbanczik: Phys. Rev. Lett. **80** (1998) 5445.
16) M. Biehl, P. Riegler and M. Stechert: Phys. Rev. E **52** (1995) R4624.
17) The equation for $R$, eq.(52) in ref.,[13] is wrong. The correct one is

$$\frac{dR}{dt} = \frac{\eta J}{\pi}\left(\frac{\sqrt{1+\sigma^2-\omega^2}}{1+\sigma^2} - \pi\omega E(\omega)\right).$$