A Statistical Study on On-line Learning

Noboru Murata
Lab. for Information Synthesis
RIKEN Brain Science Institute
mura@brain.riken.go.jp

Abstract

In this paper we examine on-line learning with statistical framework. Firstly we study the cases with fixed and annealed learning rate. It can be shown that on-line learning with $1/t$ annealed learning rate minimizes the generalization error with the same rate as batch learning in the asymptotic regime, that is, on-line learning can be as effective as batch learning asymptotically. Using these analyses, we study an adaptive learning rate algorithm which is based on Sompolinsky-Barkai-Seung algorithm and which achieves $1/t$-annealing automatically.

1 Batch Learning and On-line Learning

Let us consider a learning system which is specified by a parameter vector $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^m)^T \in \boldsymbol{R}^m$. Let $(\boldsymbol{x}, \boldsymbol{y})$ be a input-output pair, which the system learns, where $\boldsymbol{x} = (x_1, \ldots, x_r)^T \in \boldsymbol{R}^r$ and $\boldsymbol{y} = (y_1, \ldots, y_s)^T \in \boldsymbol{R}^s$. For each input-output pair, we define a loss function

$$d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) \tag{1}$$

which measures the performance of learning system $\boldsymbol{\theta}$ for given input $\boldsymbol{x}$ and desired output $\boldsymbol{y}$. In general, loss function $d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ can be divided into two part, pointwise loss $l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ and regularization $r(\boldsymbol{\theta})$

$$d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) + \lambda r(\boldsymbol{\theta}), \tag{2}$$

where λ determines the strength of regularization. We assume that loss $d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ is differentiable up to the appropriate order and we use ∇ and $\nabla\nabla$ to indicate

vector and matrix differential operators

$$\nabla = \begin{pmatrix} \dfrac{\partial}{\partial \theta^1} \\ \vdots \\ \dfrac{\partial}{\partial \theta^m} \end{pmatrix}, \tag{3}$$

$$\nabla\nabla = \begin{pmatrix} \dfrac{\partial^2}{\partial \theta^1 \partial \theta^1} & \cdots & \dfrac{\partial^2}{\partial \theta^1 \partial \theta^m} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2}{\partial \theta^m \partial \theta^1} & \cdots & \dfrac{\partial^2}{\partial \theta^m \partial \theta^m} \end{pmatrix}. \tag{4}$$

A typical example of the loss function is the squared error for the three-layered perceptron with l_2 regularization, such as

$$d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{s} |y_i - f_i(\boldsymbol{x}; \boldsymbol{\theta})|^2 + \lambda \sum_{j=1}^{m} |\theta^j|^2,$$

where

$$f_i(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{j=1}^{h} v_{ij} \phi \left(\sum_{k=1}^{r} w_{jk} x_k - b_j \right), \quad i = 1, \ldots, s,$$

h is the number of hidden units, $\phi(x)$ is a sigmoidal function, and $\boldsymbol{\theta}$ denotes a set of all the modifiable parameters v_{ij}, w_{jk} and b_j.

We also define a total loss function averaged over all the possible inputs and desired outputs

$$\begin{aligned} D(\boldsymbol{\theta}) &= E_p \left(d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}) \right) \\ &= \int_{\mathbf{R}^{r+s}} d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}, \end{aligned} \tag{5}$$

where E_p denotes expectation taken under distribution $p(\boldsymbol{x}, \boldsymbol{y})$. The optimal parameter is defined as a parameter which minimizes total loss function $D(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_* = \operatorname*{argmin}_{\boldsymbol{\theta}} D(\boldsymbol{\theta}). \tag{6}$$

The goal of learning is to obtain optimal parameter $\boldsymbol{\theta}_*$, however, in usual settings, the true input-output distribution $p(\boldsymbol{x}, \boldsymbol{y})$ is unknown and only a finite number of input-output examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ are available, which are i.i.d. random variables sampled from the target system $p(\boldsymbol{x}, \boldsymbol{y})$. A plausible approach is defining the empirical total loss function by

$$\begin{aligned} \hat{D}(\boldsymbol{\theta}) &= E_{\hat{p}} \left(d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}) \right) \\ &= \int d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) \hat{p}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \end{aligned} \tag{7}$$

with the empirical distribution

$$\hat{p}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{y}_i), \tag{8}$$

and defining an estimator by

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \hat{D}(\boldsymbol{\theta}). \tag{9}$$

Note that estimator $\hat{\boldsymbol{\theta}}$ is a function of given examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_n), \tag{10}$$

therefore $\hat{\boldsymbol{\theta}}$ is a random variable which depends on examples.

This procedure is called batch learning or non-sequential learning. From the statistical point of view, batch learning is a sort of statistical inference, and it has close relation with estimating function method (see for examples, [7, 8]). As well known in statistics, if the optimal parameter gives the strict minimum of the total loss, i.e.

$$D(\boldsymbol{\theta}_*) < D(\boldsymbol{\theta}), \quad |\boldsymbol{\theta} - \boldsymbol{\theta}_*| > 0, \tag{11}$$

the ensemble mean and the variance of estimator $\hat{\boldsymbol{\theta}}$ are asymptotically given by the following lemma.

Lemma 1. *Let $\hat{\boldsymbol{\theta}}$ be an estimator which minimizes empirical total loss function* (7)*. The mean and the variance of $\hat{\boldsymbol{\theta}}$ are asymptotically given by*

$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_*, \tag{12}$$

$$V(\hat{\boldsymbol{\theta}}) = \frac{1}{n} Q_*^{-1} G_* Q_*^{-1}, \tag{13}$$

where E and V denote the mean and the variance over all the possible sets of n i.i.d. examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ from $p(\boldsymbol{x}, \boldsymbol{y})$, and G_* and Q_* are symmetric matrices defined by

$$G_* = E_p\left(\nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}_*) \nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}_*)^T\right), \tag{14}$$

$$Q_* = E_p\left(\nabla \nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}_*)\right), \tag{15}$$

To find estimator $\hat{\boldsymbol{\theta}}$, non-linear optimization methods are applied, such as gradient descent, Newton and quasi-Newton method. For a huge number of parameters, Newton-like methods don't have any specific advantage because of large computational complexity, hence many acceleration methods for gradient descent and approximated or abbreviated methods based on Newton method are proposed.

We have to note that in artificial neural network learning, a problem of local minima often becomes a subject of discussion. There are two main problems:

- How many local minima exist in parameter space?
- How local minima distribute around the optimal parameter?

Unfortunately nobody succeeded to give a rigorous answer yet. In this paper, we don't go into the matter of local minima deeply, but we just note that by substituting the locally optimal parameter for the optimal parameter, most of lemmas and theorems hold in each basin of attraction of local minima.

Also when optimal parameter $\boldsymbol{\theta}_*$ is not a point but a connected region in parameter space the above lemma doesn't hold. For example, in over-realizable case of teacher-student scenario, that is the case in which the student networks have bigger structure than the teacher network, experimentally it is known that distribution of estimator has a longer tail than regular cases. This is a problem of degenerated Fisher information matrices in statistics, and asymptotic distributions of estimators are unknown except for special cases [6].

Let us consider the following updating rule.

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - \eta_t C_t \nabla d(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}; \hat{\boldsymbol{\theta}}_t), \tag{16}$$

where $(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})$ is a pair of input and desired output, η_t is a positive value and C_t is a positive definite matrix which can depend on $\hat{\boldsymbol{\theta}}_t$. This procedure is called on-line learning, sequential learning or stochastic descent method. Main difference of on-line learning from batch learning is that observing examples and modifying parameters are alternatively taken place. Namely, on-line update rule uses only information available at time t. Estimator $\hat{\boldsymbol{\theta}}_t$ is a function of previous estimator $\hat{\boldsymbol{\theta}}_{t-1}$ and a given example $(\boldsymbol{x}_t, \boldsymbol{y}_t)$, hence it can be thought as a function of initial value $\boldsymbol{\theta}_0 \equiv \hat{\boldsymbol{\theta}}_0$ and all the given examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^t$

$$\begin{aligned}\hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_t \left(\hat{\boldsymbol{\theta}}_{t-1}, \boldsymbol{x}_t, \boldsymbol{y}_t\right) \\ &= \hat{\boldsymbol{\theta}}_t \left(\boldsymbol{\theta}_0, \boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_t, \boldsymbol{y}_t\right),\end{aligned}$$

and it is a random variable which depends on a sequence of examples. Unless there is any confusion, we use abbreviated form $\hat{\boldsymbol{\theta}}_t$.

Note that other update rules can be applicable to on-line learning. An example is to use a small set of examples $\{(\boldsymbol{x}_{t+1}^j, \boldsymbol{y}_{t+1}^j)\}_{j=1}^k$ at time t

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - \eta_t C_t \sum_{j=1}^{k} \nabla d(\boldsymbol{x}_{t+1}^j, \boldsymbol{y}_{t+1}^j; \hat{\boldsymbol{\theta}}_t).$$

Although the detailed behavior of this sort of gradient based on-line learning differs from the simple ones, still update rule (16) is essential. Also we can use more general update rule such as

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - \eta_t F(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}; \hat{\boldsymbol{\theta}}_t),$$

however, in usual cases the above update rule can be well approximated by Equation (16) by utilizing the dependency of C_t on $\hat{\boldsymbol{\theta}}_t$, and in order to investigate

asymptotic property, this approximation is sufficient. Therefore, in this paper we deal with the simplest update rule.

The procedure (16) is validated as a method that reduces total loss at each time in an average sense, i.e.

$$\begin{aligned}
&E_p(D(\hat{\boldsymbol{\theta}}_{t+1})) - D(\hat{\boldsymbol{\theta}}_t) \\
&= E_p\left(D\left(\hat{\boldsymbol{\theta}}_t - \eta_t C_t \nabla d(\boldsymbol{X}, \boldsymbol{Y}; \hat{\boldsymbol{\theta}}_t)\right)\right) - D(\hat{\boldsymbol{\theta}}_t) \\
&= -\nabla D(\hat{\boldsymbol{\theta}}_t)^T E_p\left(\eta_t C_t \nabla d(\boldsymbol{X}, \boldsymbol{Y}; \hat{\boldsymbol{\theta}}_t)\right) + O\left(\eta_t^2\right) \\
&= -\eta_t \nabla D(\hat{\boldsymbol{\theta}}_t)^T C_t \nabla D(\hat{\boldsymbol{\theta}}_t) + O\left(\eta_t^2\right) \\
&< 0,
\end{aligned}$$

if η_t and C_t are appropriately chosen. When C_t is fixed, from the theory of a stochastic approximation [17], a sufficient condition for convergence is given by

$$\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty. \tag{17}$$

Hence learning rate η_t plays an important role in on-line learning.

Compared with batch learning, on-line learning has several advantages. One is low computational cost because less memory is needed to store examples and experimentally on-line learning is said to be faster than batch learning in terms of computational time. Another advantage is adaptation ability to changing environment. Since in on-line learning, used examples are abandoned and never referred, newly given examples have more influence to the estimator. This forgetting effect helps to follow the gradual change of the environment. On the other hand, estimation of the parameter is thought to be less accurate because of its stochastic property.

2 Asymptotic Behavior of On-line Learning

To have a good intuition of on-line learning, we calculate the mean and the variance of estimator $\hat{\boldsymbol{\theta}}_t$

$$\boldsymbol{\theta}_t = E(\hat{\boldsymbol{\theta}}_t), \tag{18}$$

$$V_t = V(\hat{\boldsymbol{\theta}}_t). \tag{19}$$

In the following we discuss convergence property around local minima which are stable fixed points of dynamics. Except for special models such as simple perceptrons, it is difficult to discuss the whole process of on-line learning. If the estimator is in the neighborhood of the optimal parameter or the locally optimal parameter, the evolution of the mean and the variance are approximated in a simple form that is given by Amari [2]. In the case that there exist local minima, learning dynamics is conditioned within the basin of attraction of a certain local minimum and the similar statement holds in each basin.

Lemma 2 (Amari[2]). *If η_t is sufficiently small, the mean value of a smooth function $f(\boldsymbol{\theta})$ is approximated by recursive equation*

$$\begin{aligned} E^{\hat{\boldsymbol{\theta}}_{t+1}}\left(f(\hat{\boldsymbol{\theta}}_{t+1})\right) &= E^{\hat{\boldsymbol{\theta}}_t}\left(f(\hat{\boldsymbol{\theta}}_t)\right) - \eta_t E^{\hat{\boldsymbol{\theta}}_t}\left(\nabla f(\hat{\boldsymbol{\theta}}_t)^T C_t \nabla D(\hat{\boldsymbol{\theta}}_t)\right) \\ &\quad + \frac{\eta_t^2}{2}\operatorname{tr}\left(E^{\hat{\boldsymbol{\theta}}_t}\left(C_t G(\hat{\boldsymbol{\theta}}_t) C_t^T \nabla\nabla f(\hat{\boldsymbol{\theta}}_t)\right)\right) + O(\eta_t^3), \end{aligned} \tag{20}$$

where $E^{\hat{\boldsymbol{\theta}}}$ denotes expectation with respect to $\hat{\boldsymbol{\theta}}$.

From the above lemma, the evolution of the mean and the variance of the estimator are given as follows in the neighborhood of the optimal parameter, where total loss function $D(\boldsymbol{\theta})$ is well approximated by quadratic form

$$D(\boldsymbol{\theta}) = D(\boldsymbol{\theta}_*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*). \tag{21}$$

Lemma 3. *In the neighborhood of the optimal parameter, evolution of mean and the variance are approximated by recursive equations*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t C_t Q_* (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*), \tag{22}$$

$$\begin{aligned} V_{t+1} &= V_t - \eta_t (C_t Q_* V_t + V_t Q_* C_t^T) + \eta_t^2 C_t G_* C_t^T \\ &\quad - \eta_t^2 C_t Q_* (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)^T Q_* C_t^T. \end{aligned} \tag{23}$$

In this section we investigate two specific cases:

- C_t is a constant matrix C and η_t is a constant value η,
- C_t is a constant matrix C and η_t is controlled as $O(1/t)$.

Introducing notations

$$K_* = CQ_*, \tag{24}$$

$$V_* = Q_*^{-1} G_* Q_*^{-1}, \tag{25}$$

we hereafter use simplified recursive equations

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t K_* (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*), \tag{26}$$

$$\begin{aligned} V_{t+1} &= V_t - \eta_t (K_* V_t + V_t K_*^T) + \eta_t^2 K_* V_* K_*^T \\ &\quad - \eta_t^2 K_* (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)^T K_*^T. \end{aligned} \tag{27}$$

Since Q_* and C are positive definite, all the eigenvalues of matrix K_* are positive and we refer the eigenvalues of K_* by

$$\lambda_1, \ldots, \lambda_m, \quad \lambda_1 \geq \cdots \geq \lambda_m > 0. \tag{28}$$

Also we use two linear operators Ξ and Ω which are defined by

$$\Xi_A B = AB + (AB)^T, \tag{29}$$

$$\Omega_A B = ABA^T, \tag{30}$$

where A and B are square matrices. The eigenvalues of Ξ_A and Ω_A are given by

$$\nu_i + \nu_j, \tag{31}$$
$$\nu_i \nu_j, \quad i, j = 1, \ldots, m \tag{32}$$

respectively, where ν_i's are eigenvalues of matrix A.

2.1 Fixed Learning Rate

In the case that learning rate η_t does not vary during the training process, i.e. $\eta_t = \eta$, Equations (26) and (27) can be solved directly.

Theorem 1. *For fixed learning rate η, the mean and the variance of the estimator are given by*

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + (I - \eta K_*)^t (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*), \tag{33}$$
$$V_t = \left\{ I - (I - \Xi_{\eta K_*})^t \right\} \eta V_\infty - \left\{ (\Omega_{I - \eta K_*})^t - (I - \Xi_{\eta K_*})^t \right\} V_0, \tag{34}$$

where $\boldsymbol{\theta}_0$ is the initial value and

$$V_0 = (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T, \tag{35}$$
$$V_\infty = \Xi_{K_*}^{-1} \Omega_{K_*} V_*. \tag{36}$$

Proof is given in Appendix A.3. Also we can do various analysis of fixed learning rate case, such as step response and frequency response. They are given in Appendix C.

From the above theorem, we know that with an appropriate η, the mean and the variance of the estimator converge as $t \to \infty$

$$\lim_{t \to \infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_* \tag{37}$$
$$\lim_{t \to \infty} V_t = \eta V_\infty. \tag{38}$$

The mean converges to the optimal parameter, however there remains a fluctuation of order $O(\eta)$. To avoid the divergence of the variance, η must be smaller than the inverse of the maximal eigenvalue $1/\lambda_1$. On the other hand, the convergence speed of the mean and the variance is dominated by the minimal eigenvalue λ_m. As the first term of the variance represents the remaining fluctuation of the estimator around the optimal parameter, small η is preferable. On the other hand, to eliminate the dependency on the initial value in the second term as fast as possible, large η is better. Roughly speaking, if the learning rate η is large, the speed of approaching to optimal parameter $\boldsymbol{\theta}_*$ is fast, however, large fluctuation remains even if time t becomes infinitely large. And if η is small, fluctuation is small but it takes a long time for the convergence.

This effect can be seen from the viewpoint of generalization errors [13, 14]. The ensemble average of total loss function is called generalization error or predictive error, and its behavior through time is called learning curve. By using the above result and expanding the total loss around the optimal parameter, i.e.

$$
\begin{aligned}
&E\left(D(\hat{\boldsymbol{\theta}}_t)\right) \\
&= D(\boldsymbol{\theta}_*) + \frac{1}{2}\operatorname{tr}\left(Q_* E\left((\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)^T + (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)^T\right)\right),
\end{aligned}
$$

we know the asymptotic behavior of the learning curve.

Corollary 2. *In the case of fixed rate on-line learning, the learning curve is asymptotically given by*

$$
E\left(D(\hat{\boldsymbol{\theta}}_t)\right) = D(\boldsymbol{\theta}_*) + \frac{1}{2}\eta \operatorname{tr}\left(Q_* V_\infty\right) + \frac{1}{2}\operatorname{tr}\left(Q_*(I - \Xi_{\eta K_*})^t (V_0 - \eta V_\infty)\right). \quad (39)
$$

The first term is possible minimal loss, and the second term is additional loss caused by the fluctuation of estimation. Only the third term depends on time and it describes decreasing speed of errors. Here we emphasize that it is not possible to determine an universal optimal rate because there is a trade-off between convergence speed and accuracy of learning. However, for example, if the training time is limited up to T_{end}, we can define an optimal rate in terms of minimal expected generalization errors from Equation (39) by

$$
\eta_{\mathrm{opt}} = \operatorname*{argmin}_{\eta} E\left(D(\hat{\boldsymbol{\theta}}_{T_{\mathrm{end}}})\right). \quad (40)
$$

It is interesting to compare the stochastic descent and the true gradient descent. Suppose that we know true input-output distribution $p(\boldsymbol{x}, \boldsymbol{y})$, then we can use the following update rule to obtain the optimal parameter:

$$
\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta C E_p\left(\nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}_t)\right). \quad (41)
$$

In this case, estimator $\boldsymbol{\theta}_t$ is not a random variable and it obeys the same recursive equation of the ensemble mean of on-line estimator $\hat{\boldsymbol{\theta}}_t$. Hence we can calculate the learning curve of this case as follows.

Corollary 3. *In the case of batch learning, the learning curve is approximated by*

$$
D(\boldsymbol{\theta}_t) = D(\boldsymbol{\theta}_*) + \frac{1}{2}\operatorname{tr}\left(Q_*(\Omega_{I-\eta K_*})^t V_0\right). \quad (42)
$$

The second term describes decreasing speed of errors, and it is slightly slower than the corresponding term of the on-line leaning. However, there is no fluctuation in this case, the total performance is better than the on-line learning.

2.2 Annealed Learning Rate

An effective adaptation of the learning rate is $\eta_t = O(1/t)$, because the convergence is guaranteed by conditions

$$\sum_{t=1}^{\infty} \frac{1}{t} = \infty, \quad \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty \tag{43}$$

from the theory of stochastic approximation [17]. Here we investigate how this annealing works asymptotically. Since we can not solve the equation of the variance directly, we give only the leading order of the solution.

Theorem 4. *Suppose the learning rate is $\eta_t = 1/(t+1)$, then mean and the variance of estimator $\hat{\boldsymbol{\theta}}_t$ are asymptotically given by*

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + S_t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*), \quad t \geq 2, \quad \boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_1 \tag{44}$$

$$V_t = \begin{cases} R_t V' & \lambda_m < \frac{1}{2} \\ \dfrac{1}{t+1} \left(\Xi_{K_*} - I\right)^{-1} K_* V_* K_*^T & \textit{otherwise}, \end{cases} \tag{45}$$

where

$$S_t = \prod_{k=2}^{t} \left(I - \frac{1}{k} K_* \right), \tag{46}$$

$$R_t = \prod_{k=2}^{t} \Xi_{\frac{1}{2} I - \frac{1}{k} K_*}, \tag{47}$$

and V' is a positive definite matrix, which depends on the initial value and the learning process and is not able to be written in general form.

Proof is found in Appendix A.4. Note that for the simplicity of notations, in this case the learning starts at time $t = 2$ with initial value $\boldsymbol{\theta}_1$. Since the order of S_t is bounded by

$$\begin{aligned} \prod_{k=2}^{t} \left(1 - \frac{\lambda_m}{k}\right) &\sim \prod_{k=2}^{t} e^{-\frac{\lambda_m}{k}} \\ &= e^{-\sum_{k=2}^{t} \frac{\lambda_m}{k}} \\ &\sim e^{-\log \frac{\lambda_m}{t} + \text{const.}} \\ &= O(1/t^{\lambda_m}), \end{aligned}$$

large λ_m is preferable for the fast convergence of the mean. And since the eigenvalues of the operator $\left(\Xi_{K_*} - I\right)^{-1} \Omega_{K_*}$ are represented by

$$\frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j - 1}, \quad i, j = 1, \ldots, m,$$

and the minimum value is achieved when $\lambda_i = 1$, $i = 1, \ldots, m$, the smallest variance can be realized if all the eigenvalues of K_* are 1, which is realized by $C = Q_*^{-1}$ as the simplest case. If the minimum eigenvalue λ_m is larger than $1/2$, the variance has the bigger contribution to the fluctuation of the on-line learning. Therefore we know the following situation is the optimal case of the annealed rate learning.

Corollary 5. *The optimal case of $1/t$-annealed rate is achieved when all the eigenvalues of K_* are 1. Then the mean and the variance of the estimator are asymptotically given by*

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + \frac{1}{t}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*), \tag{48}$$

$$V_t = \frac{1}{t} V_* + O\left(\frac{1}{t^2}\right). \tag{49}$$

When we use the batch learning with t examples, the mean and the variance of the estimator is given by Lemma 1. Expanding the total loss function around the optimal parameter, we obtain the learning curve of batch learning as

$$\begin{aligned}
&E\left(D(\hat{\boldsymbol{\theta}}_t^B)\right) \\
&= D(\boldsymbol{\theta}_*) + \frac{1}{2}\,\mathrm{tr}\left(Q_* E\left((\hat{\boldsymbol{\theta}}_t^B - \boldsymbol{\theta}_*)(\hat{\boldsymbol{\theta}}_t^B - \boldsymbol{\theta}_*)^T\right)\right) + o\left(\frac{1}{t}\right) \\
&= D(\boldsymbol{\theta}_*) + \frac{1}{2t}\,\mathrm{tr}\left(Q_* V_*\right) + o\left(\frac{1}{t}\right),
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}_t^B$ denotes the estimator obtained by t-example batch learning. Similarly, we can calculate the learning curve of optimally annealed on-line learning as

$$\begin{aligned}
&E\left(D(\hat{\boldsymbol{\theta}}_t)\right) \\
&= D(\boldsymbol{\theta}_*) + \frac{1}{2}\,\mathrm{tr}\left(Q_* E\left((\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)^T + (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)^T\right)\right) + o\left(\frac{1}{t}\right) \\
&= D(\boldsymbol{\theta}_*) + \frac{1}{2}\,\mathrm{tr}\left(Q_*\left(\frac{1}{t}V_* + \frac{1}{t^2}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T\right)\right) + o\left(\frac{1}{t}\right),
\end{aligned}$$

The above relationship supports efficiency of optimally annealed on-line learning, and it is summarized as follows.

Corollary 6. *In the case of optimally annealed on-line learning, the learning curve is asymptotically given by*

$$E\left(D(\hat{\boldsymbol{\theta}}_t)\right) = D(\boldsymbol{\theta}_*) + \frac{1}{2t}\,\mathrm{tr}\left(Q_* V_*\right) + o\left(\frac{1}{t}\right). \tag{50}$$

This is coincide with the case of batch learning in the order of $O(1/t)$.

Therefore, optimally annealed on-line learning is as asymptotically effective as batch learning in the sense of the generalization error.

Here we give a quite simple explanation why $1/t$-annealing is feasible in a class of $1/t^\alpha$-annealed rate learning. For $\eta_t = 1/(1+t)^\alpha$ annealing, from lemma 3 we have the solution

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + S_t^\alpha(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*), \tag{51}$$

where

$$S_t^\alpha = \prod_{k=2}^{t}(I - \frac{1}{k^\alpha}K_*). \tag{52}$$

Obviously the convergence speed of the variance is slower than order $O(1/t^\alpha)$. From this fact, we know that the larger α is preferable. On the other hand, if we assume $\alpha > 1$, the slowest term of Equation (52) can be bounded by

$$\begin{aligned}
\prod_{k=2}^{t}(1 - \frac{1}{k^\alpha}\lambda_n) &= \exp\left\{\sum_{k=2}^{t}\log(1 - \frac{1}{k^\alpha}\lambda_n)\right\} \\
&\geq \exp\left\{\sum_{k=2}^{t}\frac{1}{k^\alpha}\lambda_n\right\} \\
&\geq \exp\left\{-\frac{1}{\alpha-1}\left(\frac{1}{2^{\alpha-1}} - \frac{1}{t^{\alpha-1}}\right)\lambda_n\right\} \\
&> 0,
\end{aligned}$$

where we assumed that $\lambda_n/2^\alpha$ is less than 1. This means that if α is larger than 1, the mean of the estimator does not converge to the optimal parameter. In this sense, $1/t$-annealing is feasible.

3 Adaptive Learning Rate

From the results of the previous section, we have an intuitive idea about the learning rate for on-line learning. In practical applications, the learning rate might be scheduled as follows:

- When the estimator $\hat{\boldsymbol{\theta}}_t$ is far from the optimal parameter $\boldsymbol{\theta}_*$, use an appropriately large η.
- When $\hat{\boldsymbol{\theta}}_t$ is close to $\boldsymbol{\theta}_*$, use $1/t$-annealing with an appropriate C.

A simple implementation is

$$\eta_t = \begin{cases} \eta_0 & t < T, \\ \dfrac{\eta_{t-1}}{\eta_{t-1} + \eta_0}\eta_0 & t \geq T, \end{cases}$$

where T is a switching time and η_0 is a initial learning rate. However, it is difficult to give a proper switching time a priori and when the rule changes over time, an annealed learning rate cannot follow the changes fast enough since η_t is too small. Hence we need some sophisticated method to perform the above simple strategy automatically.

3.1 Sompolinsky-Barkai-Seung Algorithm

The idea of an adaptively changing η_t was called learning of learning rule [2, 22]. The Sompolinsky-Barkai-Seung algorithm is extended to the differentiable loss function version as follows.

Definition 1.

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{t+1} &= \hat{\boldsymbol{\theta}}_t - \eta_t Q^{-1}(\hat{\boldsymbol{\theta}}_t) \nabla d(\boldsymbol{x}_t, \boldsymbol{y}_t; \hat{\boldsymbol{\theta}}_t), \\ \eta_{t+1} &= \eta_t + \alpha \eta_t \left\{ \beta \left(d(\boldsymbol{x}_t, \boldsymbol{y}_t; \hat{\boldsymbol{\theta}}_t) - D_* \right) - \eta_t \right\}, \end{aligned} \tag{53}$$

where α and β are positive constants, D_* is the minimal loss function $D(\boldsymbol{\theta}_*)$ and

$$Q(\hat{\boldsymbol{\theta}}_t) = E_p \left(\nabla \nabla d(\boldsymbol{X}, \boldsymbol{Y}, \hat{\boldsymbol{\theta}}_t) \right). \tag{54}$$

Intuitively speaking, in equation system (53) the coefficient η is controlled by the remaining error. When the error is large, the learning rate takes a relatively large value

$$\eta_t \sim \beta \left(d(\boldsymbol{x}_t, \boldsymbol{y}_t; \hat{\boldsymbol{\theta}}_t) - D_* \right).$$

When the error is small, that is, the estimator is close to the optimal parameter, the learning rate approaches to 0 automatically as

$$\eta_{t+1} = \eta_t - \alpha \eta_t^2.$$

In order to know how the algorithm works, let us consider the averaged dynamical behavior of $(\hat{\boldsymbol{\theta}}_t, \eta_t)$ asymptotically. For simplicity of treatment, we consider a continuous version of the algorithm

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\eta(t) Q(\boldsymbol{\theta}(t))^{-1} E_p \left(\nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}(t)) \right), \tag{55}$$

$$\frac{d}{dt}\eta(t) = \alpha \eta(t) \left\{ \beta E_p \left(d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}(t)) - D_* \right) - \eta(t) \right\}, \tag{56}$$

where $\boldsymbol{\theta}(t)$ and $\eta(t)$ denote the mean of the estimator and the learning rate at continuous time t respectively. When the estimator is in the neighborhood of the optimal parameter, we can use the following relation and approximations

$$\begin{aligned} & E_p \left(\nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}_*) \right) = 0, \\ & E_p \left(d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}(t)) \right) \simeq D_* + \frac{1}{2} (\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*)^T Q_* (\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*), \\ & E_p \left(\nabla d(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}(t)) \right) \simeq Q_* \left(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_* \right). \end{aligned}$$

Then the equations are rewritten as

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\eta(t)\left(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*\right), \tag{57}$$

$$\frac{d}{dt}\eta(t) = \alpha\eta(t)\left\{\frac{\beta}{2}(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*)^T Q_*(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*) - \eta(t)\right\}. \tag{58}$$

Finally introducing a squared error variable

$$e(t) = \frac{1}{2}(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*)^T Q_*(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*) \tag{59}$$

we obtain an equation system

$$\begin{cases} \dfrac{d}{dt}e(t) = -2\eta(t)e(t), \\ \dfrac{d}{dt}\eta(t) = \alpha\beta\eta(t)e(t) - \alpha\eta(t)^2. \end{cases} \tag{60}$$

It is easy to check that the equation system has a solution

$$\begin{cases} e(t) = \dfrac{1}{\beta}\cdot\left(\dfrac{1}{2} - \dfrac{1}{\alpha}\right)\cdot\dfrac{1}{t}, \quad \alpha > 2, \\ \eta(t) = \dfrac{1}{2}\cdot\dfrac{1}{t}. \end{cases} \tag{61}$$

Therefore the learning rate is automatically annealed as $1/t$, if the estimator approaches to the optimal parameter. Thus Equations (57) and (58) give us an adaptive on-line learning algorithm in which the learning rate is annealed as $O(1/t)$. This algorithm is also expected to follow slow fluctuation or sudden change of target rule. Note that in this explanation, we omitted the effect of randomness which is a natural characteristics of on-line learning. We will discuss it in the next section.

3.2 Modified Algorithm

From the viewpoint of practical implementation, the algorithm dealt in the previous section has some problems such as

- the Hessian Q_* of the total loss must be calculated at each iteration,
- the minimal value of the loss function must be known.

Here we consider a slightly alleviated learning rule.

Let us consider the averaged dynamics of the estimator in the neighborhood of the optimal parameter:

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\eta(t)K_*(\boldsymbol{\theta}(t) - \boldsymbol{\theta}_*), \tag{62}$$

where the approximation

$$E_p\left(C\nabla d(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\theta}(t))\right) \simeq K_*(\boldsymbol{\theta}(t)-\boldsymbol{\theta}_*) \tag{63}$$

is used. Suppose that we have a vector $\boldsymbol{v}$ which satisfies

$$\boldsymbol{v}^T K_* = \lambda \boldsymbol{v}^T, \tag{64}$$

where λ is an eigenvalue of matrix K_*. Using a new value

$$\xi(t) = E_p\left(\boldsymbol{v}^T C\nabla d(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\theta}(t))\right) \simeq \boldsymbol{v}^T K_*(\boldsymbol{\theta}(t)-\boldsymbol{\theta}_*), \tag{65}$$

we define an update rule for learning rate as

$$\frac{d}{dt}\eta(t) = \alpha\eta(t)\left(\beta|\xi(t)| - \eta(t)\right), \tag{66}$$

where $|\cdot|$ denotes the absolute value. Taking account of relationship between ξ and η, we obtain an equation system

$$\begin{cases} \dfrac{d}{dt}\xi(t) = -\lambda\eta(t)\xi(t), \\ \dfrac{d}{dt}\eta(t) = \alpha\eta(t)\left(\beta|\xi(t)| - \eta(t)\right), \end{cases} \tag{67}$$

and a solution is described as

$$\begin{cases} |\xi(t)| = \dfrac{1}{\beta}\cdot\left(\dfrac{1}{\lambda}-\dfrac{1}{\alpha}\right)\cdot\dfrac{1}{t}, \quad \alpha > \lambda \\ \eta(t) = \dfrac{1}{\lambda}\cdot\dfrac{1}{t}. \end{cases} \tag{68}$$

Intuitively $|\xi|$ plays a role of a distance, where the average gradient is projected to a certain direction. If we choose a clever projection, the learning rate is automatically well annealed, as the estimator approaches the optimal parameter.

It is an important problem to find a good projection direction $\boldsymbol{v}$ and here we use a knowledge of learning process shown in section 2. Usually the learning speed is dominated by the minimum eigenvalue, hence the average trajectory of the estimator is almost parallel to the eigendirection of the minimum eigenvalue after reasonably many iterations. This means that in an average sense the learning process can be seen as a one-dimensional problem along the eigendirection of minimum eigenvalue of K_*. Therefore we can use the averaged gradient of the loss function as $\boldsymbol{v}$

$$\boldsymbol{v} = \frac{E_p\left(C\nabla d(\boldsymbol{X},\boldsymbol{Y};\theta(t))\right)}{\left|E_p\left(C\nabla d(\boldsymbol{X},\boldsymbol{Y};\theta(t))\right)\right|} \tag{69}$$

to approximate the eigenvector of the minimum eigenvalue, where $|\cdot|$ denotes l_2 norm. In this case, ξ is expressed as

$$\xi(t) = \left|E_p\left(C\nabla d(\boldsymbol{X},\boldsymbol{Y};\theta(t))\right)\right|. \tag{70}$$

Based on the above consideration, we propose the following practical implementation by substituting the running average (leaky average) for the ensemble average.

Definition 2.

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{t+1} &= \hat{\boldsymbol{\theta}}_t - \eta_t^{\dagger} C \nabla d(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}; \hat{\boldsymbol{\theta}}_t), \\
\eta_t^{\dagger} &= \min(\eta_0, \eta_t) \\
\boldsymbol{r}_{t+1} &= (1-\delta)\boldsymbol{r}_t + \delta C \nabla d(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}; \hat{\boldsymbol{\theta}}_t), \\
\eta_{t+1} &= \eta_t + \alpha \eta_t \left(\beta |\boldsymbol{r}_{t+1}| - \eta_t\right),
\end{aligned} \tag{71}$$

where δ is a constant between 0 and 1, which controls the leakiness of the average and $\boldsymbol{r}$ is used as an auxiliary variable to calculate the running average of the gradient.

Note that due to the use of the running average, η_t doesn't converge to 0 even though $t \to \infty$. When the size of $\boldsymbol{r}_t$ is large compared with its fluctuation, the above algorithm shows automatic $1/t$-annealing as we expect. However, η_t fluctuates around the mean of $|\boldsymbol{r}_t|$ because of the fluctuation of $\boldsymbol{r}_t$, and this can be expressed as

$$\eta_t \sim \beta E(|\boldsymbol{r}_t|) + \epsilon(\alpha, \beta),$$

where ϵ is a random variable and its amplitude is controlled by α. Roughly speaking, if α is large, modification of η_t is fast and as a result η_t is highly affected by the fluctuation of $\boldsymbol{r}_t$. If α is small, the fluctuation is smoothed and ϵ is small. Also the fluctuation of $\boldsymbol{r}_t$ originates from the running average and the estimation error of $\hat{\boldsymbol{\theta}}_t$, hence $\boldsymbol{r}_t$ is approximated with a zero-mean random variable by

$$\boldsymbol{r}_t \sim \epsilon'(\delta, \eta_t).$$

The amplitude of ϵ' is reduced when δ and η_t become small. According to these mutual interaction, the size of η_t is determined.

For practical applications of this algorithm, refer to Murata *et al.* (1997)[12].

4 Conclusion

In this paper we studied on-line learning with fixed and annealed learning rate with the framework of statistics, and asymptotic evolutions of the mean and the variance of the estimators are investigated,

In the case of fixed learning rate we have found an exponential convergence of the mean of estimators and the variance, i.e. the estimation error is proportional to the learning rate. Also we have shown that with the annealing rule $\eta = O(1/t)$, the same convergence speed as in batch learning can be achieved asymptotically. This means that on-line learning is as efficient as batch learning with optimal $1/t$-annealing.

Also we gave a theoretically motivated adaptive on-line algorithm extending the work of Sompolinsky *et al.* On-line learning is especially important under non-stationary environments, and strategies for the learning of learning rate might be applied in the case of such changing environments.

References

[1] Masafumi Akahira and Kei Takeuchi. *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*, volume 7 of *Lecture Notes in Statistics*. Springer-Verlag, 1981.

[2] Shun-ichi Amari. Theory of adaptive pattern classifiers. *IEEE Trans. EC*, 16(3):299–307, 1967.

[3] Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985.

[4] Shun-ichi Amari, Naotake Fujita, and Sigeru Sinomoto. Four types of learning curves. *Neural Computation*, 4(4):605–618, 1992.

[5] Shun-ichi Amari and Noboru Murata. Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, 5(1):140–153, 1993.

[6] Kenji Fukumizu. Special statistical properties of neural network learning. In *Proceedings of 1997 International Symposium on Nonlinear Theory and Its Applications*, volume 2, pages 747–750. Research Society of Nonlinear Theory and its Applications, IEICE, 1997.

[7] V. P. Godambe, editor. *Estimating Functions*. Oxford University Press, New York, 1991.

[8] Motoaki Kawanabe and Shun-ichi Amari. Estimation of network parameters in semiparametric stochastic perceptron. *Neural Computation*, 6(6):1244–1261, nov 1994.

[9] J. W. Kim and H. Sompolinsky. On-line Gibbs learning. submitted to Physical Review Letters, 1995.

[10] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. In *Proceedings of the IEEE 78*, number 10, pages 1568–1574, 1990.

[11] M. Minsky and S. Papert. *Perceptrons – An Introduction to Computational Geometry (Expanded Edition)*. The MIT Press, 1988.

[12] Noboru Murata, Klaus-Robert Müller, Andreas Ziehe, and Shun-ichi Amari. Adaptive on-line learning in changing environments. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advences in Neural Information Processing Systems*, volume 9, pages 599–605, Cambridge, MA, 1997. The MIT Press.

[13] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen et al., editor, *Artificial Neural Networks*, pages 9–14, Holland, 1991. ICANN, Elsevier Science Publishers.

[14] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. Learning curves, model selection and complexity of neural networks. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 607–614, San Mateo, CA, 1993. Morgan Kaufmann Publishers.

[15] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. Network information criterion — determining the number of hidden units for an artificial neural network model. *IEEE Trans. NN*, 5(6):865–872, 1994.

[16] Manfred Opper and David Haussler. Calculation of the learning curve of Bayes optimal classification algorithm for learning a Perceptron with noise. In *Proceedings of COLT*, pages 75–87, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[17] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[18] F. Rosenblatt. *Principle of Neurodynamics.* Spartan, 1961.

[19] D. Rumelhart, J. L McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* The MIT Press, 1986.

[20] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45:6056–6091, 1992.

[21] H. Sompolinski and N Barkai. Theory of learning from examples. In *IJCNN'93-NAGOYA Tutorial Texts*, pages 221–240. International Joint Conference on Neural Networks, 1993.

[22] H. Sompolinsky, N. Barkai, and H. S. Seung. On-line learning of dichotomies: algorithms and learning curves. In J-H. Oh, C. Kwon, and S. Cho, editors, *Neural Networks: The Statistical Mechanics Perspective*, pages 105–130, Singapore, 1995. World Scientific.

[23] Halbert White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, 1989.

[24] B. Widraw. *A Statistical Theory of Adaptation.* Pergamon Press., 1963.

A Proofs of Lemmas and Theorems

In the following, we adopt Einstein summation convention, that is summation is automatically taken without the summation symbol $\sum$ for those indices which appear as a subscript and a superscript simultaneously

$$a^i b_i \equiv \sum_i a^i b_i,$$

and we use the following abbreviation

$$\partial_i \equiv \frac{\partial}{\partial \theta^i}.$$

For any vector $\boldsymbol{a}$, $(\boldsymbol{a})^i$ denotes the i-th element of $\boldsymbol{a}$, and for any matrix A, $(A)^{ij}$ denotes the ij element of A.

A.1 Proof of Lemma 2

Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be $\hat{\boldsymbol{\theta}}_t$ and $\hat{\boldsymbol{\theta}}_{t+1}$ which are estimators at time t and $t+1$ respectively, and let $\boldsymbol{z}$ be a pair of input-output example $(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})$ which obeys probability density $p(\boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{y})$ and which is used to modify the estimator from $\hat{\boldsymbol{\theta}}_t$ to $\hat{\boldsymbol{\theta}}_{t+1}$. We write the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ by

$$\boldsymbol{\theta}'(\boldsymbol{z}, \boldsymbol{\theta}) = \boldsymbol{\theta} + \delta\boldsymbol{\theta}(\boldsymbol{z}, \boldsymbol{\theta}), \tag{72}$$

where

$$\delta\boldsymbol{\theta}(\boldsymbol{z}, \boldsymbol{\theta}) = -\eta C \nabla d(\boldsymbol{z}; \boldsymbol{\theta}), \tag{73}$$

and η_t and C_t are simply denoted by η and C respectively. First we derive the probability density of $\boldsymbol{\theta}'$ for fixed $\boldsymbol{\theta}$, that is conditioned probability density $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. Let $q(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta}'$ be the probability that estimator θ' is in a rectangle

$$[\boldsymbol{\theta}', \boldsymbol{\theta}' + d\boldsymbol{\theta}'] = [\theta^{'1}, \theta^{'1} + d\theta^{'1}] \times \cdots \times [\theta^{'m}, \theta^{'m} + d\theta^{'m}] \in \mathbf{R}^m, \tag{74}$$

then the following relation holds,

$$q(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta}' = p(\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}'))d\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}'), \tag{75}$$

where $\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $d\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ define a set of examples and a volume elements respectively, such that examples $\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ modify the estimator from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ and volume elements $d\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ specifies region of examples which modify the estimator from $\boldsymbol{\theta}$ to a point in rectangle $[\boldsymbol{\theta}', \boldsymbol{\theta}' + d\boldsymbol{\theta}']$. Let $q(\boldsymbol{\theta})$ be the probability density of estimator $\boldsymbol{\theta}$. By averaging both sides of Equation (75) with respect to $q(\boldsymbol{\theta})$, we obtain the probability density of estimator $\boldsymbol{\theta}'$

$$\begin{aligned} q(\boldsymbol{\theta}')d\boldsymbol{\theta}' &= \left(\int_{\boldsymbol{\theta} \in \mathbf{R}^m} q(\boldsymbol{\theta}'|\boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) d\boldsymbol{\theta}' \\ &= \int_{\boldsymbol{\theta} \in \mathbf{R}^m} p\left(\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')\right) d\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}') q(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \tag{76}$$

Since the domain of integration for $\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is whole $\mathbf{R}^{r+s}$ when the domain of integration for $\boldsymbol{\theta}'$ is $\mathbf{R}^m$ for fixed $\boldsymbol{\theta}$, then

$$\begin{aligned} E^{\boldsymbol{\theta}'}\left(f(\boldsymbol{\theta}')\right) &= \int_{\boldsymbol{\theta}' \in \mathbf{R}^m} f(\boldsymbol{\theta}') q(\boldsymbol{\theta}') d\boldsymbol{\theta}' \\ &= \int_{\boldsymbol{z} \in \mathbf{R}^{r+s}} \int_{\boldsymbol{\theta} \in \mathbf{R}^m} f(\boldsymbol{\theta}'(\boldsymbol{z}, \boldsymbol{\theta})) p(\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}')) d\boldsymbol{z}(\boldsymbol{\theta}, \boldsymbol{\theta}') q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta} \in \mathbf{R}^m} \left(\int_{\boldsymbol{z} \in \mathbf{R}^{r+s}} f(\boldsymbol{\theta}'(\boldsymbol{z}, \boldsymbol{\theta})) p(\boldsymbol{z}) d\boldsymbol{z} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= E^{\boldsymbol{\theta}}\left(E^{\boldsymbol{z}}\left(f(\boldsymbol{\theta}'(\boldsymbol{z}, \boldsymbol{\theta}))\right)\right) \end{aligned} \tag{77}$$

holds, where $E^{\boldsymbol{\theta}}$, $E^{\boldsymbol{\theta}'}$ and $E^{\boldsymbol{z}}$ denote average over $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ and $\boldsymbol{z}$ respectively. Expanding $\boldsymbol{\theta}'$ around $\boldsymbol{\theta}$ and knowing that the order of $\delta\boldsymbol{\theta}$ is $O(\eta)$, we obtain

$$\begin{aligned} E^{\boldsymbol{z}}\left(f(\boldsymbol{\theta}')\right) &= E^{\boldsymbol{z}}\left(f(\boldsymbol{\theta} + \delta\boldsymbol{\theta})\right) \\ &= E^{\boldsymbol{z}}\left(f(\boldsymbol{\theta}) - \partial_i f(\boldsymbol{\theta}) \delta\boldsymbol{\theta}^i + \frac{1}{2} \partial_i \partial_j f(\boldsymbol{\theta}) \delta\boldsymbol{\theta}^i \delta\boldsymbol{\theta}^j + O(\eta^3) \right) \\ &= f(\boldsymbol{\theta}) - \eta \partial_i f(\boldsymbol{\theta}) c^{ij} E^{\boldsymbol{z}}\left(\partial_j d(\boldsymbol{z}; \boldsymbol{\theta})\right) \\ &\quad + \frac{\eta^2}{2} \partial_i \partial_j f(\boldsymbol{\theta}) c^{ik} c^{jh} E_{\boldsymbol{z}}\left(\partial_k d(\boldsymbol{z}; \boldsymbol{\theta}) \partial_h d(\boldsymbol{z}; \boldsymbol{\theta})\right) + O(\eta^3) \\ &= f(\theta) - \eta \partial_i f(\boldsymbol{\theta}) c^{ij} \partial_j D(\boldsymbol{\theta}) + \frac{\eta^2}{2} \partial_i \partial_j f(\boldsymbol{\theta}) c^{ik} c^{jh} g_{kh}(\boldsymbol{\theta}) + O(\eta^3) \\ &= f(\boldsymbol{\theta}) - \eta \nabla f(\boldsymbol{\theta})^T C \nabla D(\boldsymbol{\theta}) + \frac{\eta^2}{2} \operatorname{tr}\left(C G(\boldsymbol{\theta}) C^T \nabla\nabla f(\boldsymbol{\theta})\right) + O(\eta^3), \end{aligned} \tag{78}$$

where c^{ij} and $g_{ij}(\boldsymbol{\theta})$ are the ij elements of C and $G(\boldsymbol{\theta})$ respectively. By taking average with respect to $q(\boldsymbol{\theta})$, the proof is completed. □

A.2 Proof of Lemma 3

Recalling that $D(\boldsymbol{\theta})$ takes the minimum value at optimal parameter $\boldsymbol{\theta}_*$,

$$\nabla D(\boldsymbol{\theta}_*) = 0 \tag{79}$$

holds. $D(\boldsymbol{\theta})$ is expanded at $\boldsymbol{\theta}_*$ as

$$D(\boldsymbol{\theta}) = D(\boldsymbol{\theta}_*) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*) + O(|\boldsymbol{\theta} - \boldsymbol{\theta}_*|^3), \tag{80}$$

where the relation

$$\nabla\nabla D(\boldsymbol{\theta}_*) \equiv Q_*, \tag{81}$$

is used. Hence, gradient of $D(\boldsymbol{\theta})$ can be written

$$\nabla D(\boldsymbol{\theta}) = Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*) + O(|\boldsymbol{\theta} - \boldsymbol{\theta}_*|^2). \tag{82}$$

First, suppose that

$$f(\boldsymbol{\theta}) = \theta^i \tag{83}$$

in Lemma 2, then

$$E^{\hat{\boldsymbol{\theta}}_t}\left(f(\hat{\boldsymbol{\theta}}_t)\right) = \theta_t^i, \tag{84}$$

and

$$\nabla f(\boldsymbol{\theta}) = \begin{pmatrix} \delta_1^i \\ \vdots \\ \delta_m^i \end{pmatrix}, \tag{85}$$

$$\nabla\nabla f(\boldsymbol{\theta}) = 0, \tag{86}$$

where δ_j^i denotes Kronecker's delta, i.e.

$$\delta_j^i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{87}$$

Assuming that $\epsilon_t^2 = E(|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_*|^2)$ is small, we obtain

$$\theta_{t+1}^i = \theta_t^i - \eta_t \left(C_t Q_* (\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)\right)^i + O(\eta_t^3) + O(\epsilon_t^2 \eta_t). \tag{88}$$

By collecting the above relation for all i, we obtain Equation (22).

Next suppose that

$$f(\boldsymbol{\theta}) = \theta^i \theta^j, \tag{89}$$

then

$$E^{\hat{\boldsymbol{\theta}}_t}\left(f(\hat{\boldsymbol{\theta}}_t)\right) = V_t^{ij}, \tag{90}$$

where V_t^{ij} denote the ij element of V_t and

$$\nabla f(\boldsymbol{\theta}) = \begin{pmatrix} \delta_1^i \theta^j + \theta^i \delta_1^j \\ \vdots \\ \delta_m^i \theta^j + \theta^i \delta_m^j \end{pmatrix}, \tag{91}$$

$$\nabla\nabla f(\boldsymbol{\theta}) = \begin{pmatrix} \delta_1^i \delta_1^j + \delta_1^i \delta_1^j & \cdots & \delta_1^i \delta_m^j + \delta_m^i \delta_1^j \\ \vdots & \ddots & \vdots \\ \delta_m^i \delta_1^j + \delta_1^i \delta_m^j & \cdots & \delta_m^i \delta_m^j + \delta_m^i \delta_m^j \end{pmatrix}. \tag{92}$$

Using the relations

$$\begin{aligned} &\nabla f(\boldsymbol{\theta})^T C \nabla D(\boldsymbol{\theta}) \\ &= \left(\delta_1^i \theta^j + \theta^i \delta_1^j, \ldots, \delta_m^i \theta^j + \theta^i \delta_m^j\right) C Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*) + O(|\boldsymbol{\theta} - \boldsymbol{\theta}_*|^2) \\ &= \left(C Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*)\right)^i \theta^j + \theta^i \left(C Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*)\right)^j + O(|\boldsymbol{\theta} - \boldsymbol{\theta}_*|^2) \\ &= \left(C Q_* (\boldsymbol{\theta} - \boldsymbol{\theta}_*) \boldsymbol{\theta}^T\right)^{ij} + \left(\boldsymbol{\theta} (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T Q_* C^T\right)^{ij} + O(|\boldsymbol{\theta} - \boldsymbol{\theta}_*|^2) \end{aligned} \tag{93}$$

and

$$
\begin{aligned}
&\mathrm{tr}\left(CG(\boldsymbol{\theta})C^T\nabla\nabla f(\boldsymbol{\theta})\right)\\
&=\mathrm{tr}\left(CG(\boldsymbol{\theta})C^T\begin{pmatrix}\delta_1^i\delta_1^j+\delta_1^i\delta_1^j & \dots & \delta_1^i\delta_m^j+\delta_m^i\delta_1^j\\ \vdots & \ddots & \vdots\\ \delta_m^i\delta_1^j+\delta_1^i\delta_m^j & \dots & \delta_m^i\delta_m^j+\delta_m^i\delta_m^j\end{pmatrix}\right)\\
&=\left(CG(\boldsymbol{\theta})C^T\right)^{ij}+\left(CG(\boldsymbol{\theta})C^T\right)^{ji}\\
&=2\left(CG(\boldsymbol{\theta})C^T\right)^{ij}\\
&=2\left(CG_*C^T\right)^{ij}+O(|\boldsymbol{\theta}-\boldsymbol{\theta}_*|),
\end{aligned}
\tag{94}
$$

and assuming that $\epsilon_t = E(|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_*|)$ is small we obtain

$$
\begin{aligned}
&E^{\hat{\boldsymbol{\theta}}_{t+1}}\left(\hat{\boldsymbol{\theta}}_{t+1}\hat{\boldsymbol{\theta}}_{t+1}^T\right)\\
&=E^{\hat{\boldsymbol{\theta}}_t}\left(\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^T\right)-\eta_tC_tQ_*E^{\hat{\boldsymbol{\theta}}_t}\left((\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_*)\hat{\boldsymbol{\theta}}_t^T\right)\\
&\quad-\eta_tE^{\hat{\boldsymbol{\theta}}_t}\left(\hat{\boldsymbol{\theta}}_t(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_*)^T\right)Q_*C_t^T+\eta_t^2C_tG_*C_t^T+O(\eta_t^3)+O(\epsilon\eta_t^2)+O(\epsilon^2\eta_t)\\
&=E^{\hat{\boldsymbol{\theta}}_t}\left(\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^T\right)-\eta_tC_tQ_*E^{\hat{\boldsymbol{\theta}}_t}\left(\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^T\right)-\eta_tE^{\hat{\boldsymbol{\theta}}_t}\left(\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^T\right)Q_*C_t^T\\
&\quad+\eta_tC_tQ_*\boldsymbol{\theta}_*\hat{\boldsymbol{\theta}}_t^T+\eta_t\boldsymbol{\theta}_t\hat{\boldsymbol{\theta}}_*^TQ_*C_t^T+\eta_t^2C_tG_*C_t^T+O(\eta_t^3)+O(\epsilon\eta_t^2)+O(\epsilon^2\eta_t).
\end{aligned}
\tag{95}
$$

Noting the definition

$$
V_t = E^{\hat{\boldsymbol{\theta}}_t}(\hat{\boldsymbol{\theta}}_t\hat{\boldsymbol{\theta}}_t^T)-\boldsymbol{\theta}_t\boldsymbol{\theta}_t^T,
\tag{96}
$$

and using the relation

$$
\begin{aligned}
&\boldsymbol{\theta}_{t+1}\boldsymbol{\theta}_{t+1}^T\\
&=\{\boldsymbol{\theta}_t-\eta_tC_tQ_*(\boldsymbol{\theta}_t-\boldsymbol{\theta}_*)\}\{\boldsymbol{\theta}_t-\eta_tC_tQ_*(\boldsymbol{\theta}_t-\boldsymbol{\theta}_*)\}^T+O(\eta_t^3)+O(\epsilon^2\eta_t)\\
&=\boldsymbol{\theta}_t\boldsymbol{\theta}_t^T-\eta_tC_tQ_*\boldsymbol{\theta}_t\boldsymbol{\theta}_t^T-\eta_t\boldsymbol{\theta}_t\boldsymbol{\theta}_t^TQ_*C_t^T\\
&\quad+\eta_tC_tQ_*\boldsymbol{\theta}_*\boldsymbol{\theta}_t^T+\eta_t\boldsymbol{\theta}_t\boldsymbol{\theta}_*^TQ_*C_t^T+\eta_t^2C_tQ_*(\boldsymbol{\theta}_t-\boldsymbol{\theta}_*)(\boldsymbol{\theta}_t-\boldsymbol{\theta}_*)^TQ_*C_t^T\\
&\quad+O(\eta_t^3)+O(\epsilon^2\eta_t),
\end{aligned}
\tag{97}
$$

we obtain

$$
\begin{aligned}
V_{t+1}&=E^{\boldsymbol{\theta}_{t+1}}\left(\hat{\boldsymbol{\theta}}_{t+1}\hat{\boldsymbol{\theta}}_{t+1}^T\right)-\boldsymbol{\theta}_{t+1}\boldsymbol{\theta}_{t+1}^T\\
&=V_t-\eta_tC_tQ_*V_t-\eta_tV_tQ_*C_t^T\\
&\quad+\eta_t^2C_tG_*C_t^T-\eta_t^2C_tQ_*(\boldsymbol{\theta}_t-\boldsymbol{\theta}_*)(\boldsymbol{\theta}_t-\boldsymbol{\theta}_*)^TQ_*C_t^T\\
&\quad+O(\eta_t^3)+O(\epsilon\eta_t^2)+O(\epsilon^2\eta_t).
\end{aligned}
\tag{98}
$$

□

A.3 Proof of Theorem 1

From Equation (26), the solution of the mean is trivial.

From the properties of operators Ξ and Ω,

$$\begin{aligned}\eta^2 K_*(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_*)^T K_*^T &= \Omega_{\eta K_*}\Omega_{(I-\eta K_*)^t}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T \\ &= \Omega_{\eta K_*}(\Omega_{I-\eta K_*})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T \\ &= (\Omega_{I-\eta K_*})^t\Omega_{\eta K_*}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T \end{aligned} \tag{99}$$

holds. Let us define

$$v_t = V_t - \eta V_\infty, \tag{100}$$

$$u_t = (\Omega_{I-\eta K_*})^t\Omega_{\eta K_*}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T, \tag{101}$$

then the recursive equation for the variance can be written by two equations

$$v_{t+1} = (I - \Xi_{\eta K_*})v_t - u_t, \tag{102}$$

$$u_{t+1} = \Omega_{I-\eta K_*}u_t, \tag{103}$$

and they are rewritten in matrix form as

$$\begin{pmatrix} v_{t+1} \\ u_{t+1} \end{pmatrix} = \begin{pmatrix} I - \Xi_{\eta K_*} & -I \\ 0 & \Omega_{I-\eta K_*} \end{pmatrix} \begin{pmatrix} v_t \\ u_t \end{pmatrix}. \tag{104}$$

Noting the commutativity of operators and the relation

$$\Omega_{I-\eta K_*} = I - \Xi_{\eta K_*} + \Omega_{\eta K_*},$$

the solution of the recursive equation system is

$$\begin{pmatrix} v_t \\ u_t \end{pmatrix} = \begin{pmatrix} (I - \Xi_{\eta K_*})^t & -(\Omega_{\eta K_*})^{-1}\{(\Omega_{I-\eta K_*})^t - (I - \Xi_{\eta K_*})^t\} \\ 0 & (\Omega_{I-\eta K_*})^t \end{pmatrix} \begin{pmatrix} v_0 \\ u_0 \end{pmatrix}, \tag{105}$$

where the initial values are given by

$$v_0 = -(\Xi_{\eta K_*})^{-1}\Omega_{\eta K_*}V_* = -\eta(\Xi_{K_*})^{-1}\Omega_{K_*}V_* = -V_\infty \tag{106}$$

$$u_0 = \Omega_{\eta K_*}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)^T = \Omega_{\eta K_*}V_0 \tag{107}$$

By using the definition of v_t, V_t is obtained. □

A.4 Proof of Theorem 4

From Equation (26) and the definition of S_t, the solution of the mean is trivial.

The solution of the variance is given by considering the following two cases.

Let us assume the order of V_t is slower than $O(1/t)$. Considering the contribution to the leading term, Equation (27) is written as

$$\begin{aligned} V_{t+1} &= V_t - \frac{1}{t+1}(K_* V_t + V_t K_*^T) + \frac{1}{(t+1)^2} K_* V_* K_*^T \\ &= \left(\frac{1}{2} I - \frac{1}{t+1} K_*\right) V_t + V_t \left(\frac{1}{2} I - \frac{1}{t+1} K_*\right)^T \end{aligned}$$

by omitting the higher order terms. This recursive equation gives the solution, and knowing that the eigenvalue of $\Xi_{\frac{1}{2}I - \frac{1}{t+1}K_*}$ is represented by

$$1 - \frac{\lambda_i + \lambda_j}{t+1}, \quad i, j = 1, \ldots, m,$$

the order of R_t is bounded by $O(1/t^{2\lambda_m})$. Therefore, when $\lambda_m < 1/2$ holds, $R_t V'$ is slower than $O(1/t)$.

Let us assume the order of V_t is $O(1/t)$ and write V_t with $1/tV$. By omitting the higher order terms, we have the equation for the leading order as

$$\begin{aligned} V_{t+1} &= V_t - \frac{1}{t+1}(K_* V_t + V_t K_*^T) + \frac{1}{(t+1)^2} K_* V_* K_*^T \\ \frac{1}{t+1} V &= \frac{1}{t} V - \frac{1}{t(t+1)}(K_* V + V K_*^T) + \frac{1}{(t+1)^2} K_* V_* K_*^T \end{aligned}$$

$$\frac{1}{t(t+1)} \left(K_* V + V K_*^T - V\right) = \frac{1}{(t+1)^2} K_* V_* K_*^T$$

Then we have the solution

$$V = \frac{t}{t+1} \left(\Xi_{K_*} - I\right)^{-1} K_* V_* K_*^T. \tag{108}$$

□

B Asymptotic Distribution of Fixed Rate Learning

Here we calculate the evolution of the characteristic function of estimator $\hat{\boldsymbol{\theta}}_t$. Let f be

$$f(\boldsymbol{\theta}) = e^{\sqrt{-1} z_i (\theta^i - \theta^i_*)}, \tag{109}$$

where $\boldsymbol{z} = (z_i) \in \mathbf{R}^m$ and m is the dimension of the parameter $\boldsymbol{\theta}$. Without loss of generality, we can assume that the optimal parameter θ^* is at origin, that is, we use such a coordinate for simplicity. Then f can be written as

$$f(\boldsymbol{\theta}) = e^{\sqrt{-1} z_i \theta^i}, \tag{110}$$

The characteristic function of the distribution of $\hat{\boldsymbol{\theta}}_t$ is

$$E^{\hat{\boldsymbol{\theta}}_t}(f(\hat{\boldsymbol{\theta}}_t)) = \varphi_t(\boldsymbol{z}). \tag{111}$$

Knowing that $\hat{\boldsymbol{\theta}}_t \sim O(\sqrt{\eta})$ after sufficient learning steps, ∂D can be expanded around the origin up to order $O(\eta)$ as

$$\partial_i D(\hat{\boldsymbol{\theta}}_t) = q_{ij}\hat{\theta}_t^j + \frac{t_{ijk}}{2}\hat{\theta}_t^j\hat{\theta}_t^k + O(\eta^{3/2}),$$

where

$$q_{ij} = \partial_i\partial_j D(\boldsymbol{\theta}_*), \quad t_{ijk} = \partial_i\partial_j\partial_k D(\boldsymbol{\theta}_*).$$

From Lemma 2, we obtain the recursive equation

$$\begin{aligned}
\varphi_{t+1}(\boldsymbol{z}) &= \varphi_t(\boldsymbol{z}) - \eta E^{\hat{\boldsymbol{\theta}}_t}\left(c^{ij}\sqrt{-1}z_i f(\hat{\boldsymbol{\theta}}_t)\left(q_{jk}\hat{\theta}_t^k + \frac{t_{jkl}}{2}\hat{\theta}_t^k\hat{\theta}_t^l\right)\right) \\
&\quad - \frac{\eta^2}{2}E^{\hat{\boldsymbol{\theta}}_t}\left(c^{ik}c^{jl}g_{kl}z_i z_j f(\hat{\boldsymbol{\theta}}_t)\right) + O(\eta^{5/2}) \\
&= \varphi_t(\boldsymbol{z}) - \eta c^{ij}z_i E^{\hat{\boldsymbol{\theta}}_t}\left(q_{jk}\frac{\partial}{\partial z_k}f(\hat{\boldsymbol{\theta}}_t) - \sqrt{-1}\frac{t_{jkl}}{2}\frac{\partial^2}{\partial z_k\partial z_l}f(\hat{\boldsymbol{\theta}}_t)\right) \\
&\quad - \frac{\eta^2}{2}c^{ik}c^{jl}g_{kl}z_i z_j E^{\hat{\boldsymbol{\theta}}_t}(f(\hat{\boldsymbol{\theta}}_t)) + O(\eta^{5/2}) \\
&= \varphi_t(\boldsymbol{z}) - \eta c^{ij}z_i\left(q_{jk}\frac{\partial}{\partial z_k}\varphi_t(\boldsymbol{z}) - \sqrt{-1}\frac{t_{jkl}}{2}\frac{\partial^2}{\partial z_k\partial z_l}\varphi_t(\boldsymbol{z})\right) \\
&\quad - \frac{\eta^2}{2}c^{ik}c^{jl}g_{kl}z_i z_j\varphi_t(\boldsymbol{z}) + O(\eta^{5/2}),
\end{aligned} \tag{112}$$

where c^{ij} is the ij element of matrix C and relations

$$\partial_k f(\boldsymbol{\theta}) = \sqrt{-1}z_k f(\boldsymbol{\theta}), \ldots \quad \frac{\partial}{\partial z_k}f(\boldsymbol{\theta}) = \sqrt{-1}\theta_k f(\boldsymbol{\theta}), \ldots$$

are used.

Assuming that as $t \to \infty$, a sequence of characteristic functions $\{\varphi_t\}$ converge to a function

$$\varphi_t(\boldsymbol{z}) \to \varphi(\boldsymbol{z}),$$

we have the equation from Equation (112)

$$c^{ij}z_i\left(q_{jk}\frac{\partial}{\partial z_k}\varphi(\boldsymbol{z}) - \sqrt{-1}\frac{t_{jkl}}{2}\frac{\partial^2}{\partial z_k\partial z_l}\varphi(\boldsymbol{z})\right) = -\frac{\eta}{2}c^{ik}c^{jl}g_{kl}z_i z_j\varphi(\boldsymbol{z}) + O(\eta^{3/2}) \tag{113}$$

for any $\boldsymbol{z}$. Using an expansion

$$\varphi(\boldsymbol{z}) = e^{h_0(\boldsymbol{z}) + \eta h_1(\boldsymbol{z}) + O(\eta^{3/2})}, \tag{114}$$

we obtain equations

$$c^{ij}z_i\left(q_{jk}\frac{\partial}{\partial z_k}h_0(\boldsymbol{z})-\sqrt{-1}\frac{t_{jkl}}{2}\frac{\partial^2}{\partial z_k\partial z_l}h_0(\boldsymbol{z})\right)=0 \tag{115}$$

$$c^{ij}z_i\left(q_{jk}\frac{\partial}{\partial z_k}h_1(\boldsymbol{z})-\sqrt{-1}\frac{t_{jkl}}{2}\frac{\partial^2}{\partial z_k\partial z_l}h_1(\boldsymbol{z})\right)=-\frac{1}{2}c^{ik}c^{jl}g_{kl}z_iz_j. \tag{116}$$

Knowing that $\varphi(0)=1$ from the property of the characteristic function, we obtain solutions

$$h_0(\boldsymbol{z})=0 \tag{117}$$

$$h_1(\boldsymbol{z})=-\frac{1}{2}z_iz_jv^{ij}-\sqrt{-1}z_iu^i, \tag{118}$$

where v^{ij} is a symmetric matrix which satisfies

$$c^{ik}q_{kl}v^{lj}+c^{jk}q_{kl}v^{li}=c^{ik}c^{jl}g_{kl}$$

and $\boldsymbol{u}$ is given by

$$u^i=-q^{ij}t_{jkl}v^{kl}.$$

Therefore we have the following theorem and corollary.

Theorem 7. *The characteristic function of the estimator obtained by on-line learning with Equation* (16) *converges to*

$$\varphi(\boldsymbol{z})=\exp\left\{\sqrt{-1}z_i\left(\theta^i_*-\eta u^i\right)-\frac{\eta}{2}z_iz_jv^{ij}+O(\eta^{2/3})\right\}. \tag{119}$$

Corollary 8. *The mean of the estimator with Equation* (16) *converges to*

$$\lim_{t\to\infty}E(\boldsymbol{\theta}_t)=\boldsymbol{\theta}_*-\eta\boldsymbol{u}+O(\eta^{2/3}). \tag{120}$$

There exists bias of order $O(\eta)$ in the mean of the estimator, but the order of standard deviation is $O(\sqrt{\eta})$, which comes from the second derivative of the characteristic function. As η goes to 0, the bias can be neglected asymptotically.

C Dynamical Property of Fixed Rate Learning

C.1 Step Response

Let us consider the situation in which the distribution of examples changes at $t=0$

$$p(\boldsymbol{x},\boldsymbol{y})\to p(\boldsymbol{x},\boldsymbol{y})+\delta p(\boldsymbol{x},\boldsymbol{y}), \tag{121}$$

and as a result, the optimal parameter changes as

$$\boldsymbol{\theta}_*+\delta\boldsymbol{\theta}=\operatorname*{argmin}_{\boldsymbol{\theta}}E_{p+\delta p}(d(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\theta})). \tag{122}$$

For simplicity, we assume that $\delta\boldsymbol{\theta}$ is sufficiently small and the approximations

$$G(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}) \simeq G(\boldsymbol{\theta}_*) = G_*, \tag{123}$$
$$Q(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}) \simeq Q(\boldsymbol{\theta}_*) = Q_* \tag{124}$$

hold. Suppose before $t = 0$ the estimator has converged in an average sense, i.e.

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_*,$$
$$V_t = \eta V_\infty,$$

then from Equation (26), the mean of the estimator at t is given by

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + (I - (I - \eta K_*)^t)\delta\boldsymbol{\theta}. \tag{125}$$

Therefore, the step response of the mean is represented by operator

$$I - (I - \eta K_*)^t. \tag{126}$$

From Equation (27), the variance of the estimator is also calculated as

$$V_t = \eta V_\infty - \left\{ (\Omega_{I-\eta K_*})^t - (I - \Xi_{\eta K_*})^t \right\} \delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^T. \tag{127}$$

The second term corresponds to the transitional response. It is interesting that the variance becomes once slightly smaller than ηV_∞ while the mean is apart from the optimal parameter, and as the mean converges to the optimal parameter the variance goes back to ηV_∞.

The learning curve can be calculated by using above result. Noting

$$\begin{aligned}
&E\left((\hat{\boldsymbol{\theta}}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}))(\hat{\boldsymbol{\theta}}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}))^T\right) \\
&= E\left((\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)^T\right) + E\left((\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}))(\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}))^T\right) \\
&= V_t + (\Omega_{I-\eta K_*})^t \delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^T \\
&= \eta V_\infty + (I - \Xi_{\eta K_*})^t \delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^T,
\end{aligned} \tag{128}$$

and expanding new total loss $D'(\boldsymbol{\theta}_t)$ around the new optimal parameter $\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}$, we obtain

$$\begin{aligned}
&E\left(D'(\boldsymbol{\theta}_t)\right) \\
&= D'(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}) + \frac{1}{2}\operatorname{tr}\left(Q_* E\left((\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}))(\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}))^T\right)\right) \\
&= D'(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}) + \frac{1}{2}\eta \operatorname{tr}\left(Q_* V_\infty\right) + \frac{1}{2}\operatorname{tr}\left((I - \Xi_{\eta K_*})^t Q_* \delta\boldsymbol{\theta}\delta\boldsymbol{\theta}^T\right).
\end{aligned} \tag{129}$$

The third term corresponds to the step response of the learning curve.

C.2 Frequency Response

Let us consider the situation in which the optimal parameter fluctuates around parameter $\boldsymbol{\theta}_*$ as

$$\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + \boldsymbol{\kappa}_i \sin \omega_i t, \tag{130}$$

where $\boldsymbol{\kappa}_i$ is an eigenvector of matrix K_* with the i-th eigenvalue λ_i

$$K_* \boldsymbol{\kappa}_i = \lambda_i \boldsymbol{\kappa}_i. \tag{131}$$

Here we assume that $\delta\boldsymbol{\theta}_t$ is sufficiently small and the approximations

$$G(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t) \simeq G(\boldsymbol{\theta}_*) = G_*, \tag{132}$$
$$Q(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t) \simeq Q(\boldsymbol{\theta}_*) = Q_*. \tag{133}$$

hold. From Equation (26), the evolution of the mean is given by

$$\boldsymbol{\theta}_{t+1} = (I - \eta K_*)\boldsymbol{\theta}_t + \eta K_* \boldsymbol{\theta}_* + \eta \lambda_i \boldsymbol{\kappa}_i \sin \omega_i t. \tag{134}$$

Suppose the stationary solution is written as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + a_i \boldsymbol{\kappa}_i \sin(\omega_i t - \alpha_i). \tag{135}$$

From the addition theorem of the trigonometric functions, the relations

$$\eta \lambda_i \cos \alpha_i = a_i \{\cos \omega_i - (1 - \eta \lambda_i)\} \tag{136}$$
$$\eta \lambda_i \sin \alpha_i = a_i \sin \omega_i \tag{137}$$

are obtained, then a_i and α_i are given by

$$a_i = \frac{\eta \lambda_i}{\sqrt{(\cos \omega_i - (1 - \eta \lambda_i))^2 + (\sin \omega_i)^2}}, \tag{138}$$
$$\tan \alpha_i = \frac{\sin \omega_i}{\cos \omega_i - (1 - \eta \lambda_i)}. \tag{139}$$

When ω_i is small, that is, the fluctuation is slow, the approximated solution is given by

$$a_i = \frac{1}{\sqrt{1 + \alpha_i^2}}, \tag{140}$$
$$\alpha_i = \frac{\omega_i}{\eta \lambda_i}. \tag{141}$$

In this case, the stationary solution is written as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_* + \frac{1}{\sqrt{1 + \alpha_i^2}} \boldsymbol{\kappa}_i \sin(\omega_i t - \alpha_i), \tag{142}$$

and the solution means that when the optimal parameter oscillates with period $2\pi/\omega_i$ along the eigendirection of eigenvalue λ_i, the estimator follows the optimal parameter with delay α_i and the amplitude is reduced as $1/\sqrt{1+\alpha_i^2}$ times. If the direction of oscillation $\delta\boldsymbol{\theta}_t$ is general, we can consider the linear combination of eigenvectors. In any cases, if α_i is sufficiently small

$$\omega_i \ll \eta\lambda_i, \quad i = 1, \ldots, m, \tag{143}$$

the on-line learning can follow the fluctuation of the target.

Similarly the frequency response of the variance can be discussed as follows. Here we assume that α_i is small and the difference between the estimator and the optimal parameter is approximated by

$$\begin{aligned}\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t) &= \boldsymbol{\kappa}_i \left(\frac{1}{\sqrt{1+\alpha^2}} \sin(\omega_i t - \alpha_i) - \sin(\omega_i t) \right) \\ &= \boldsymbol{\kappa}_i \left(\alpha_i \cos(\omega_i t) + O(\alpha_i^2) \right). \end{aligned} \tag{144}$$

By neglecting the higher order term of α_i, the recursive equation for the variance is given by

$$V_{t+1} = (I - \eta\Xi_{K_*})V_t + \eta^2 K_* V_* K_*^T - \eta^2 \alpha_i^2 \cos^2(\omega_i t)\Omega_{K_*}\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T. \tag{145}$$

Let us assume the stationary solution of the variance is

$$V_t = \eta V_\infty + \{b_i \cos(2\omega_i t + \phi_i) + c_i\}\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T. \tag{146}$$

Using the relations

$$\Xi_{K_*}\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T = 2\lambda_i\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T, \tag{147}$$

$$\Omega_{K_*}\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T = \lambda_i^2\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T, \tag{148}$$

the magnification of the amplitude b_i and bias c_i and phase shift ϕ_i satisfy the equations

$$\frac{\eta^2\lambda_i^2\alpha_i^2}{2}\cos\phi_i = b_i(1 - 2\eta\lambda_i - \cos 2\omega_i), \tag{149}$$

$$\frac{\eta^2\lambda_i^2\alpha_i^2}{2}\sin\phi_i = b_i \sin 2\omega_i, \tag{150}$$

$$\frac{\eta^2\lambda_i^2\alpha_i^2}{2} = -2\eta\lambda_i c_i. \tag{151}$$

The solutions are

$$b_i = \frac{\eta^2\lambda_i^2\alpha_i^2}{2\sqrt{(1 - 2\eta\lambda_i - \cos 2\omega_i)^2 + (\sin 2\omega_i)^2}}, \tag{152}$$

$$c_i = -\frac{1}{4}\eta\lambda_i\alpha_i^2, \tag{153}$$

$$\tan\phi_i = \frac{\sin 2\omega_i}{1 - 2\eta\lambda_i - \cos 2\omega_i}. \tag{154}$$

By neglecting the higher order terms in the same way as the mean, we obtain the approximations

$$b_i = \frac{\eta\lambda_i\alpha_i^2}{4\sqrt{1+\alpha_i^2}}, \tag{155}$$

$$c_i = -\frac{\eta\lambda_i\alpha_i^2}{4}, \tag{156}$$

$$\phi_i = \frac{\omega_i}{\eta\lambda_i} = \alpha_i. \tag{157}$$

Therefore the stationary solution is

$$V_t = \eta V_\infty - \frac{\eta\lambda_i\alpha_i^2}{4}\left(1 - \frac{1}{\sqrt{1+\alpha_i^2}}\cos(2\omega_i t + \alpha_i)\right)\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T. \tag{158}$$

Similar to the step response, V_t becomes slightly smaller than the case where the target doesn't move.

Using the relation

$$\begin{aligned}
&E^{\hat{\boldsymbol{\theta}}_t}\left((\hat{\boldsymbol{\theta}}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t))(\hat{\boldsymbol{\theta}}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t))^T\right) \\
&= E^{\hat{\boldsymbol{\theta}}_t}\left((\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)^T\right) + \left((\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t))(\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t))^T\right) \\
&= V_t + (\alpha_i\cos\omega_i t)^2\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T \\
&= \eta V_\infty - \frac{\eta\lambda_i\alpha_i^2}{4}\left(1 - \frac{1}{\sqrt{1+\alpha_i^2}}\cos(2\omega_i t + \alpha_i)\right)\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T \\
&\quad + \frac{\alpha_i^2}{2}(\cos 2\omega_i t + 1)\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T \\
&\simeq \eta V_\infty + \left\{\left(\frac{1}{2} - \frac{\eta\lambda_i}{4}\right) + \left(\frac{1}{2} + \frac{\eta\lambda_i}{4\sqrt{1+\alpha_i^2}}\right)\cos 2\omega_i t\right\}\alpha_i^2\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T,
\end{aligned} \tag{159}$$

the frequency response of the learning curve can be also calculated in the same way as the step response,

$$\begin{aligned}
&E^{\hat{\boldsymbol{\theta}}_t}\left(D_t(\hat{\boldsymbol{\theta}}_t)\right) \\
&= D_t(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t) + \frac{1}{2}\operatorname{tr}\left(Q_* E^{\hat{\boldsymbol{\theta}}_t}\left((\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t))(\boldsymbol{\theta}_t - (\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t))^T\right)\right) \\
&= D_t(\boldsymbol{\theta}_* + \delta\boldsymbol{\theta}_t) + \frac{1}{2}\eta\operatorname{tr}(Q_* V_\infty) \\
&\quad + \frac{1}{2}\left\{\left(\frac{1}{2} - \frac{\eta\lambda_i}{4}\right) + \left(\frac{1}{2} + \frac{\eta\lambda_i}{4\sqrt{1+\alpha_i^2}}\right)\cos 2\omega_i t\right\}\alpha_i^2\operatorname{tr}\left(Q_*\boldsymbol{\kappa}_i\boldsymbol{\kappa}_i^T\right),
\end{aligned} \tag{160}$$

where D_t denotes the total loss defined by the input-output distribution at time t. Note that the third term represents the fluctuation of the learning curve and

its order is

$$O(\alpha_i^2) = O\left(\frac{\omega_i^2}{(\eta\lambda_i)^2}\right). \tag{161}$$

Hence, if

$$\frac{\omega_i^2}{(\eta\lambda_i)^2} \ll \eta$$

holds, the on-line learning can follow the target with negligible loss of the performance.