Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program
Wako-shi, Hirosawa 2-1, Saitama 351-01, JAPAN
fax: +81-48-462-9881
amari@zoo.riken.go.jp

Abstract

When a parameter space has a certain underlying structure, the ordinary gradient of a function does not represent its steepest direction but the natural gradient does. Information geometry is used for calculating the natural gradients in the parameter space of perceptrons, the space of matrices (for blind source separation) and the space of linear dynamical systems (for blind source deconvolution). The dynamical behavior of natural gradient on-line learning is analyzed and is proved to be Fisher efficient, implying that it has asymptotically the same performance as the optimal batch estimation of parameters. This suggests that the plateau phenomenon which appears in the backpropagation learning algorithm of multilayer perceptrons might disappear or might be not so serious when the natural gradient is used. An adaptive method of updating the learning rate is proposed and analyzed.

1 Introduction

The stochastic gradient method (Widrow, 1963; Amari, 1967; Tsypkin, 1973; Rumelhart et al, 1986) is a most popular learning method in the general nonlinear optimization framework. However, the parameter space is not Euclidean but has a Riemannian metric structure in many cases as will be shown in the following. In such a case, the ordinary gradient does not give the steepest direction of a target function. The steepest direction is given by the natural (or contravariant) gradient in such a case. The Riemannian metric structures are introduced by means of information geometry (Amari, 1985; Murray and Rice, 1993; Amari, 1997a). The present paper gives the natural gradients explicitly in the case of the space of perceptrons for neural learning, the space of matrices for blind source separation and the space of linear dynamical systems for blind multichannel source deconvolution. This is an extended version of an earlier NIPS paper (Amari, 1996) including new results.

How good is natural gradient learning compared to conventional gradient learning? The asymptotic behavior of on-line natural gradient learning is studied for this purpose. Training examples can be used only once in on-line learning when they appear. Therefore, its asymptotic performance cannot be better than the optimal batch procedure where all the examples can be reused again and again. However, we prove that natural gradient on-line learning gives the Fisher efficient estimator in the sense of asymptotic statistics when the loss function is differentiable, so that it is asymptotically equivalent to the optimal batch procedure (see also Amari, 1995; Opper, 1996). When the loss function is non-differentiable the accuracy of asymptotic on-line learning is worse than batch learning by a factor of 2 (see, for example, Van den Broeck and P. Reimann, 1996).

It is not easy to calculate the natural gradient explicitly in multilayer perceptrons. However, a preliminary analysis (Yang and Amari, 1997a) by using a simple model shows that the performance of natural gradient learning is remarkably good and it is sometimes free from being trapped in plateaus which give rise to slow convergence of the backpropagation learning method (Saad and Solla, 1995). This suggests that the Riemannian structure might eliminate such plateaus or might make it not so serious.

On-line learning is flexible, because it can truck slow fluctuations of the target. Such on-line dynamics was first analyzed in Amari (1967) and then by many researchers recently. Sompolinski et al. (1995), and Barkai et al. (1995) proposed an adaptive method of adjusting the learning rate (see also Amari, 1967). We generalize their idea and evaluate its performance based on the Riemannian metric of errors.

The paper is organized as follows. The natural gradient is defined in section 2. Section 3 formulates the natural gradient in various problems of stochastic descent learning. Section 4 gives the statistical analysis of efficiency of on-line learning, and section 5 is devoted to the problem of adaptive changes in the learning rate. Calculations of the Riemannian metric and explicit forms of the natural gradients are given finally in sections 6, 7 and 8.

2 Natural Gradient

Let $S = \{\boldsymbol{w} \in \boldsymbol{R}^n\}$ be a parameter space on which a function $L(\boldsymbol{w})$ is defined. When S is a Euclidean space with an orthonormal coordinate system $\boldsymbol{w}$, the squared length of a small incremental vector $d\boldsymbol{w}$ connecting $\boldsymbol{w}$ and $\boldsymbol{w} + d\boldsymbol{w}$ is given by

$$\| d\boldsymbol{w} \|^2 = \sum_{i=1}^{n} (dw_i)^2,$$

where dw_i are the components of $d\boldsymbol{w}$. However, when the coordinate system is non-orthonormal, the squared length is given by the quadratic form

$$\| d\boldsymbol{w} \|^2 = \sum_{i,j} g_{ij}(\boldsymbol{w}) dw_i dw_j. \tag{2.1}$$

When S is a curved manifold, there is no orthonormal linear coordinates, and the length of $d\boldsymbol{w}$ is always written as (2.1). Such a space is a Riemannian space. We show in later sections that parameter spaces of neural networks have the Riemannian character. The $n \times n$ matrix $G = (g_{ij})$ is called the Riemannian metric tensor and it depends in general on $\boldsymbol{w}$. It reduces to

$$g_{ij}(\boldsymbol{w}) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases}$$

in the Euclidean orthonormal case, so that G is the unit matrix I in this case.

The steepest descent direction of a function $L(\boldsymbol{w})$ at $\boldsymbol{w}$ is defined by the vector $d\boldsymbol{w}$ that minimizes $L(\boldsymbol{w}+d\boldsymbol{w})$ where $\| d\boldsymbol{w} \|$ has a fixed length, that is, under the constraint

$$\| d\boldsymbol{w} \|^2 = \varepsilon^2 \tag{2.2}$$

for a sufficiently small constant ε.

Theorem 1. The steepest descent direction of $L(\boldsymbol{w})$ in a Riemannian space is given by

$$-\tilde{\nabla} L(\boldsymbol{w}) = -G^{-1}(\boldsymbol{w})\nabla L(\boldsymbol{w}) \tag{2.3}$$

where $G^{-1} = (g^{ij})$ is the inverse of the metric $G = (g_{ij})$ and ∇L is the conventional gradient,

$$\nabla L(\boldsymbol{w}) = \left(\frac{\partial}{\partial w_1} L(\boldsymbol{w}) \cdots, \frac{\partial}{\partial w_n} L(\boldsymbol{w})\right)^T,$$

the superscript T denoting the transposition.

Proof. We put

$$d\boldsymbol{w} = \varepsilon \boldsymbol{a},$$

and search for the $\boldsymbol{a}$ that minimizes

$$L(\boldsymbol{w}+d\boldsymbol{w}) = L(\boldsymbol{w}) + \varepsilon \nabla L(\boldsymbol{w}) \cdot \boldsymbol{a}$$

under the constraint

$$\| \boldsymbol{a} \|^2 = \sum g_{ij} a_i a_j = 1.$$

By the Lagrangean method, we have

$$\frac{\partial}{\partial a_i}\{\nabla L(\boldsymbol{w})^T \boldsymbol{a} - \lambda \boldsymbol{a}^T G \boldsymbol{a}\} = 0.$$

This gives

$$\nabla L(\boldsymbol{w}) = 2\lambda G \boldsymbol{a}$$

or

$$\boldsymbol{a} = \frac{1}{2\lambda} G^{-1} \nabla L(\boldsymbol{w})$$

where λ is determined from the constraint.

We call

$$\tilde{\nabla} L(\boldsymbol{w}) = G^{-1} \nabla L(\boldsymbol{w})$$

the natural gradient of L in the Riemannian space. Thus, $-\tilde{\nabla}L$ represents the steepest descent direction of L. (If we use the tensorial notation, this is nothing but the contravariant form of $-\nabla L$.) When the space is Euclidean and the coordinate system is orthonormal, we have

$$\tilde{\nabla}L = \nabla L. \tag{2.4}$$

This suggests the natural gradient descent algorithm of the form

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \tilde{\nabla}L(\boldsymbol{w}_t), \tag{2.5}$$

where η_t is the learning rate which determines the step size.

3 Natural Gradient Learning

Let us consider an information source which generates a sequence of independent random variables $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_t, \cdots$, subject to the same probability distribution $q(\boldsymbol{z})$. The random signals $\boldsymbol{z}_t$ are processed by a processor (like a neural network) which has a set of adjustable parameters $\boldsymbol{w}$. Let $l(\boldsymbol{z}, \boldsymbol{w})$ be a loss function when signal $\boldsymbol{z}$ is processed by the processor whose parameter is $\boldsymbol{w}$. Then, the risk function or the average loss is

$$L(\boldsymbol{w}) = E[l(\boldsymbol{z}, \boldsymbol{w})] \tag{3.1}$$

where E denotes the expectation with respect to $\boldsymbol{z}$. Learning is a procedure to search for the optimal $\boldsymbol{w}^*$ that minimizes $L(\boldsymbol{w})$.

The stochastic gradient descent learning method can be formulated in general as

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t C(\boldsymbol{w}_t) \nabla l(\boldsymbol{z}_t, \boldsymbol{w}_t), \tag{3.2}$$

where η_t is a learning rate which may depend on t and $C(\boldsymbol{w})$ is a suitably chosen positive definite matrix (see Amari, 1967). In the natural gradient on-line learning method it is proposed to put $C(\boldsymbol{w})$ equal to $G^{-1}(\boldsymbol{w})$ when the Riemannian structure is defined. We give a number of examples to be studied in more detail.

A. Statistical estimation of probability density function

In the case of statistical estimation, we assume a statistical model $\{p(\boldsymbol{z}, \boldsymbol{w})\}$, and the problem is to obtain the probability distribution $p(\boldsymbol{z}, \hat{\boldsymbol{w}})$ which approximates the unknown density function $q(\boldsymbol{z})$ in the best way. That is to estimate the true $\boldsymbol{w}$ or to obtain the optimal approximation $\boldsymbol{w}$ from the observed data. A typical loss function is

$$l(\boldsymbol{z}, \boldsymbol{w}) = -\log p(\boldsymbol{z}, \boldsymbol{w}). \tag{3.3}$$

The expected loss is then given by

$$\begin{aligned} L(\boldsymbol{w}) &= -E[\log p(\boldsymbol{z}, \boldsymbol{w})] \\ &= E_q\left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}, \boldsymbol{w})}\right] + H_Z \end{aligned}$$

where H_Z is the entropy of $q(\boldsymbol{z})$ not depending on $\boldsymbol{w}$. Hence, minimizing L is equivalent to minimizing the Kullback-Leibler divergence

$$D[q(\boldsymbol{z}) \ : \ p(\boldsymbol{z}, \boldsymbol{w})] = \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}, \boldsymbol{w})} d\boldsymbol{z} \tag{3.4}$$

of two probability distributions $q(\boldsymbol{z})$ and $p(\boldsymbol{z}, \boldsymbol{w})$. When the true distribution $q(\boldsymbol{z})$ is written as $q(\boldsymbol{z}) = p(\boldsymbol{z}, \boldsymbol{w}^*)$, this is equivalent to obtain the maximum likelihood estimator $\hat{\boldsymbol{w}}$.

The Riemannian structure of the parameter space of a statistical model is defined by the Fisher information (Rao, 1945; Amari, 1985)

$$g_{ij}(\boldsymbol{w}) = E\left[\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{w})}{\partial w_i} \frac{\partial \log p(\boldsymbol{x}, \boldsymbol{w})}{\partial w_j}\right] \tag{3.5}$$

in the component form. This is the only invariant metric to be given to the statistical model (Chentsov, 1972; Campbell, 1985; Amari, 1985). The learning equation (3.2) gives a sequential estimator $\hat{\boldsymbol{w}}_t$.

B. Multilayer neural network

Let us consider a multilayer feedforward neural network which is specified by a vector parameter $\boldsymbol{w} = (w_1, \cdots w_n)^T \in \boldsymbol{R}^n$. The parameter $\boldsymbol{w}$ is composed of modifiable connection weights and thresholds. When input $\boldsymbol{x}$ is applied, the network processes it and calculates the outputs $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})$. The input $\boldsymbol{x}$ is subject to an unknown probability distribution $q(\boldsymbol{x})$. Let us consider a teacher network which, by receiving $\boldsymbol{x}$, generates the corresponding output $\boldsymbol{y}$

subject to a conditional probability distribution $q(\boldsymbol{y}|\boldsymbol{x})$. The task is to obtain the optimal $\boldsymbol{w}^*$ from examples such that the student network approximates the behavior of the teacher.

Let us denote by $l(\boldsymbol{x}, \boldsymbol{w})$ a loss when input signal $\boldsymbol{x}$ is processed by a network having parameter $\boldsymbol{w}$. A typical loss is given

$$l(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}) = \frac{1}{2}|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})|^2, \tag{3.6}$$

where $\boldsymbol{y}$ is the output given by the teacher.

Let us consider a statistical model of neural networks such that its output $\boldsymbol{y}$ is given by a noisy version of $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})$,

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}) + \boldsymbol{n}, \tag{3.7}$$

where $\boldsymbol{n}$ is a multivariate Gaussian noise with zero mean and unit covariance matrix I. By putting $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ which is an input-output pair, the model specifies the probability density of $\boldsymbol{z}$ as

$$p(\boldsymbol{z}, \boldsymbol{w}) = cq(\boldsymbol{x}) \exp\{-\frac{1}{2}|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})|^2\}, \tag{3.8}$$

where c is a normalizing constant, and the loss function (3.6) is rewritten as

$$l(\boldsymbol{z}, \boldsymbol{w}) = \text{const} + \log q(\boldsymbol{x}) - \log p(\boldsymbol{z}, \boldsymbol{w}). \tag{3.9}$$

Given a sequence of examples $(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_t, \boldsymbol{y}_t), \cdots$, the natural gradient on-line learning algorithm is written as

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \tilde{\nabla} l(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{w}_t). \tag{3.10}$$

Information geometry (Amari, 1985) shows that the Riemannian structure is given to the parameter space of multilayer networks by the Fisher information matrix

$$g_{ij}(\boldsymbol{w}) = E\left[\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})}{\partial w_i} \frac{\partial p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})}{\partial w_j}\right]. \tag{3.11}$$

We will show how to calculate $G = (g_{ij})$ and its inverse in a later section.

C. Blind separation of sources

Let us consider m signal sources which produce m independent signals $s_i(t)$, $i = 1, \cdots, m$, at discrete times $t = 1, 2, \cdots$. We assume that $s_i(t)$ are independent at different times and that the expectations of s_i are 0. Let $r(\boldsymbol{s})$ be the joint probability density function of $\boldsymbol{s}$. Then, it is written in the product form

$$r(\boldsymbol{s}) = \prod_{i=1}^{m} r_1(s_1) \cdots r_m(s_m). \tag{3.12}$$

Consider the case where we cannot have direct access to the source signals $\boldsymbol{s}(t)$ but we can observe their m instantaneous mixtures $\boldsymbol{x}(t)$,

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) \tag{3.13}$$

or

$$x_i(t) = \sum_{j=1}^{m} A_{ij} s_j(t)$$

where $A = (A_{ij})$ is an $m \times m$ nonsingular mixing matrix which does not depend on t, and $\boldsymbol{x} = (x_1, \cdots, x_m)^T$ is the observed mixtures.

Blind source separation is the problem of recovering the original signals $\boldsymbol{s}(t)$, $t = 1, 2, \cdots$ from the observed signals $\boldsymbol{x}(t)$, $t = 1, 2, \cdots$ (Jutten and Heráult, 1991). If we know A, this is trivial, because we have

$$\boldsymbol{s}(t) = A^{-1}\boldsymbol{x}(t).$$

The "blind" implies that we do not know the mixing matrix A nor the probability distribution densities $r_i(s_i)$.

A typical algorithm to solve the problem is to transform $\boldsymbol{x}(t)$ into

$$\boldsymbol{y}(t) = W_t\boldsymbol{x}(t), \tag{3.14}$$

where W_t is an estimate of A^{-1}. It is modified by the following learning equation

$$W_{t+1} = W_t - \eta_t F(\boldsymbol{x}_t, W_t). \tag{3.15}$$

Here, $F(\boldsymbol{x}, W)$ is a special matrix function satisfying

$$E[F(\boldsymbol{x}, W)] = 0 \tag{3.16}$$

for any density functions $r(\boldsymbol{s})$ of the form (3.12) when $W = A^{-1}$. For W_t of (3.15) to converge to A^{-1}, (3.16) is necessary but is not sufficient, because the stability of the equilibrium is not considered here.

Let $K(W)$ be an operator which maps a matrix to a matrix. Then

$$\tilde{F}(\boldsymbol{x}, W) = K(W)F(\boldsymbol{x}, W)$$

satisfies (3.16) when F does. The equilibrium of F and $\tilde{F}$ is the same, but their stability can be different. However, the natural gradient does not alter the stability of an equilibrium, because G^{-1} is positive-definite.

Let $l(\boldsymbol{x}, W)$ be a loss function whose expectation

$$L(W) = E[l(\boldsymbol{x}, W)]$$

is the target function which is minimized at $W = A^{-1}$. A typical function F is obtained by the gradient of l with respect to W,

$$F(\boldsymbol{x}, W) = \nabla l(\boldsymbol{x}, W). \tag{3.17}$$

Such a F is also obtained by heuristic arguments. Amari and Cardoso (1997) gave the complete family of F satisfying (3.16), and elucidated the statistical efficiency of related algorithms.

From the statistical point of view, the problem is to estimate $W = A^{-1}$ from observed data $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(t)$. However, the probability density function of $\boldsymbol{x}$ is written as

$$p_X(\boldsymbol{x}; W, r) = |W| r(W\boldsymbol{x}) \tag{3.18}$$

which is specified not only by W to be estimated but also by an unknown function r of the form (3.12). Such a statistical model is said to be semiparametric, and is a difficult problem to solve (Bickel et al., 1993), because it includes an unknown function of infinite degrees of freedom. However, we can apply the information-geometrical theory of estimating functions (Amari and Kawanabe, 1997) to this problem.

When F is given by the gradient of a loss function (3.17), where ∇ is the gradient $\partial/\partial W$ with respect to a matrix, the natural gradient is given by

$$\tilde{\nabla} l = G^{-1} \circ \nabla l. \tag{3.19}$$

Here, G is an operator transforming a matrix to a matrix so that it is an $m^2 \times m^2$ matrix. G is the metric given to the space $Gl(m)$ of all the nonsingular $m \times m$ matrices. We give its explicit form in a later section based on the Lie group structure. The inverse of G is also given explicitly. Another important problem is the stability of the equilibrium of the learning dynamics. This has recently been solved by using the Riemannian structure (Amari et al. 1997a; see also Cardoso and Laheld, 1996). The superefficiency of some algorithm has been also proved in Amari (1997b) under a certain conditions.

D. Blind source deconvolution

When the original signals $\boldsymbol{s}(t)$ are mixtured not only instantaneously but also with past signals as well, the problem is called blind source deconvolution or equalization. By introducing the time delay operator z^{-1},

$$z^{-1}\boldsymbol{s}(t) = \boldsymbol{s}(t-1), \tag{3.20}$$

we have a mixing matrix filter $\boldsymbol{A}$ denoted by

$$\boldsymbol{A}(z) = \sum_{k=0}^{\infty} A_k z^{-k} \tag{3.21}$$

where A_k are $m \times m$ matrices. The observed mixtures are

$$\boldsymbol{x}(t) = \boldsymbol{A}(z)\boldsymbol{s}(t) = \sum_{k} A_k \boldsymbol{s}(t-k). \tag{3.22}$$

To recover the original independent sources, we use the FIR model

$$\boldsymbol{W}(z) = \sum_{k=0}^{d} W_k z^{-1} \tag{3.23}$$

of degree d. The original signals are recovered by

$$\boldsymbol{y}(t) = \boldsymbol{W}_t(z)\boldsymbol{x}(t), \tag{3.24}$$

where $\boldsymbol{W}_t$ is adaptively modified by

$$\boldsymbol{W}_{t+1}(z) = \boldsymbol{W}_t(z) - \eta_t \nabla l\{\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \cdots, \boldsymbol{W}_t(z)\}. \tag{3.25}$$

Here, $l(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \cdots, \boldsymbol{W})$ is a loss function which includes some past signals. We can summarize the past signals into a current state variable in the on-line learning algorithm. Such a loss function is obtained by the maximum entropy method (Bell and Sejnowski, 1995), independent component analysis (Common, 1994) or by the statistical likelihood method.

In order to obtain the natural gradient learning algorithm

$$\boldsymbol{W}_{t+1}(z) = \boldsymbol{W}_t(z) - \eta_t \tilde{\nabla} l(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \cdots, \boldsymbol{W}_t),$$

we need to define the Riemannian metric in the space of all the matrix filters (multiterminal linear systems). Such a study was initiated by Amari (1987). It is possible to define G and to obtain G^{-1} explicitly (see section 8). A preliminary investigation into the performance of the natural gradient learning algorithm has been undertaken by Douglas et al (1996) and Amari et al (1997b).

4 Natural Gradient Gives Fisher-Efficient On-Line Learning Algorithms

The present section studies the accuracy of natural gradient learning from the statistical point of view. A statistical estimator which gives asymptotically the best result is said to be Fisher-efficient. We prove that natural gradient learning attains the Fisher-efficiency.

Let us consider multilayer perceptrons as an example. We study the case of a realizable teacher, that is, the behavior of the teacher is given by $q(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^*)$. Let $D_T = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_T, \boldsymbol{y}_T)\}$ be T independent input-output examples generated by the teacher network having parameter $\boldsymbol{w}^*$. Then, minimizing the log loss

$$l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = -\log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$$

over the training data D_T is to obtain $\hat{\boldsymbol{w}}_T$ that minimizes the training error

$$L_{\text{train}}(\boldsymbol{w}) = \frac{1}{T} \sum_{t=1}^{T} l(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w}). \tag{4.1}$$

This is equivalent to maximizing the likelihood $\prod_{t=1}^{T} p(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w})$. Hence, $\hat{\boldsymbol{w}}_T$ is the maximum likelihood estimator. The Cramér-Rao theorem states that the expected squared error of an unbiased estimator satisfies

$$E[(\hat{\boldsymbol{w}}_T - \boldsymbol{w}^*)(\hat{\boldsymbol{w}}_T - \boldsymbol{w}^*)^T] \geq \frac{1}{T} G^{-1}, \tag{4.2}$$

where the inequality holds in the sense of positive-definiteness of matrices. An estimator is said to be efficient or Fisher efficient when it satisfies (4.2) with equality for large T. The maximum likelihood estimator is Fisher-efficient, implying that it is the best estimator attaining the Cramér-Rao bound asymptotically,

$$\lim_{T\to\infty} TE[(\hat{\boldsymbol{w}}_T - \boldsymbol{w}^*)(\hat{\boldsymbol{w}}_T - \boldsymbol{w}^*)^T] = G^{-1}, \tag{4.3}$$

where G^{-1} is the inverse of the Fisher information matrix $G = (g_{ij})$ defined by (3.11).

Examples $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \cdots$ are given one at a time in the case of on-line learning. Let $\tilde{\boldsymbol{w}}_t$ be an on-line estimator at time t. At the next time $t+1$, the estimator $\tilde{\boldsymbol{w}}_t$ is modified to give a new estimator $\tilde{\boldsymbol{w}}_{t+1}$ based on the current observation $(\boldsymbol{x}_t, \boldsymbol{y}_t)$. The old observations $(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})$ cannot be reused to obtain $\tilde{\boldsymbol{w}}_{t+1}$, so that the learning rule is written as

$$\tilde{\boldsymbol{w}}_{t+1} = \boldsymbol{m}(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}, \tilde{\boldsymbol{w}}_t).$$

The process $\{\tilde{\boldsymbol{w}}_t\}$ is Markovian. Whatever learning rule $\boldsymbol{m}$ is chosen, the behavior of the estimator $\tilde{\boldsymbol{w}}_t$ is never better than that of the optimal batch estimator $\hat{\boldsymbol{w}}_t$ because of this restriction. The gradient on-line learning rule

$$\tilde{\boldsymbol{w}}_{t+1} = \tilde{\boldsymbol{w}}_t - \eta_t C \frac{\partial l(\boldsymbol{x}_t, \boldsymbol{y}_t; \tilde{\boldsymbol{w}}_t)}{\partial \boldsymbol{w}}$$

was proposed where C is a positive-definite matrix, and its dynamical behavior was studied by Amari [1967] when the learning constant $\eta_t = \eta$ is fixed. Heskes and Kappen (1991) obtained similar results, which ignited research of on-line learning. When η_t satisfies some condition, say $\eta_t = c/t$ for a positive constant c, the stochastic approximation guarantees that $\tilde{\boldsymbol{w}}_t$ is a consistent estimator converging to $\boldsymbol{w}^*$. However, it is not Fisher efficient in general.

There arises a question of whether there exists a learning rule that gives an efficient estimator. If it exists, the asymptotic behavior of on-line learning is equivalent to that of the best batch estimation method. The present paper answers the question affirmatively, by giving an efficient on-line learning rule (see Amari, 1995, see also Opper, 1996).

Let us consider the natural gradient learning rule

$$\tilde{\boldsymbol{w}}_{t+1} = \tilde{\boldsymbol{w}}_t - \frac{1}{t}\tilde{\nabla} l(\boldsymbol{x}_t, \boldsymbol{y}_t, \tilde{\boldsymbol{w}}_t). \tag{4.4}$$

Theorem 2. Under the learning rule (4.4), the natural gradient on-line estimator $\tilde{\boldsymbol{w}}_t$ is Fisher efficient.

Proof. Let us denote the covariance matrix of estimator $\tilde{\boldsymbol{w}}_t$ by

$$\tilde{V}_{t+1} = E[(\tilde{\boldsymbol{w}}_{t+1} - \boldsymbol{w}^*)(\tilde{\boldsymbol{w}}_{t+1} - \boldsymbol{w}^*)^T]. \tag{4.5}$$

This shows the expectation of the squared error. We expand

$$\frac{\partial l(\boldsymbol{x}_t, \boldsymbol{y}_t; \tilde{\boldsymbol{w}}_t)}{\partial \boldsymbol{w}} = \frac{\partial l(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w}^*)}{\partial \boldsymbol{w}} + \frac{\partial^2 l(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w}^*)}{\partial \boldsymbol{w} \partial \boldsymbol{w}}(\tilde{\boldsymbol{w}}_t - \boldsymbol{w}^*) + O(|\tilde{\boldsymbol{w}}_t - \boldsymbol{w}^*|^2).$$

By subtracting $\boldsymbol{w}^*$ from the both sides of (4.4), and by taking the expectation of the square of the both sides, we have

$$\tilde{V}_{t+1} = \tilde{V}_t - \frac{2}{t}\tilde{V}_t + \frac{1}{t^2}G^{-1} + O\left(\frac{1}{t^3}\right), \tag{4.6}$$

where we used

$$E\left[\frac{\partial l(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w}^*)}{\partial \boldsymbol{w}}\right] = 0, \tag{4.7}$$

$$E\left[\frac{\partial^2 l(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w}^*)}{\partial \boldsymbol{w} \partial \boldsymbol{w}}\right] = G(\boldsymbol{w}^*), \tag{4.8}$$

$$G(\tilde{\boldsymbol{w}}_t) = G(\boldsymbol{w}^*) + O\left(\frac{1}{t}\right),$$

because $\tilde{\boldsymbol{w}}_t$ converges to $\boldsymbol{w}^*$ as guaranteed by stochastic approximation under a certain conditions (see Kushner and Clark, 1978). The solution of (4.6) is written asymptotically as

$$\tilde{V}_t = \frac{1}{t}G^{-1} + O\left(\frac{1}{t^2}\right),$$

proving the theorem.

The present theory can be extended to be applicable to the unrealizable teacher case, where

$$K(\boldsymbol{w}) = E\left[\frac{\partial^2}{\partial \boldsymbol{w} \partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})\right] \tag{4.9}$$

should be used instead of $G(\boldsymbol{w})$ in order to obtain the same efficient result as the optimal batch procedure. This is locally equivalent to the Newton-Raphson method. It is immediate to state. The results can be stated in terms of the generalization error instead of the covariance of the estimator, and we can obtain more universal results (see Amari, 1993; Amari and Murata, 1993).

Remark: In the cases of blind source separation and deconvolution, the models are semi-parametric including the unknown function r, cf. (3.18). In such cases, the Cramér-Rao bound does not necessarily holds. Therefore, Theorem 2 does not hold in these cases. It holds when we can estimate the true r of the source probability density functions and use it to define the loss function $l(\boldsymbol{x}, W)$. Otherwise, (4.8) does not hold. The stability of the true solution is not necessarily guaranteed, either. Amari et al (1997a) has analyzed this situation and has proposed a universal method of attaining the stability of the equilibrium solution.

5 Adaptive Learning Constant

The dynamical behavior of the learning rule (3.2) was studied in Amari (1967) when η_t is a small constant η. In this case, $\boldsymbol{w}_t$ fluctuates around the (local) optimal value $\boldsymbol{w}^*$ for large t. The expected value and variance of $\boldsymbol{w}_t$ was studied and the trade-off between the convergence speed and accuracy of convergence was demonstrated.

When the current $\boldsymbol{w}_t$ is far from the optimal $\boldsymbol{w}^*$, it is desirable to use a relatively large η to accelerate the convergence. When it is close to $\boldsymbol{w}^*$, a small η is preferred in order to eliminate fluctuations. An idea of an adaptive change of η was discussed in Amari (1967) and was called learning of learning rules.

Sompolinski et al (1995) (see also Barkai et al, 1995) proposed a rule of adaptive change of η_t, which is applicable to the pattern classification problem where the expected loss $L(\boldsymbol{w})$ is not differentiable at $\boldsymbol{w}^*$. The present paper generalizes their idea to a more general case where $L(\boldsymbol{w})$ is differentiable, and analyzes its behavior by using the Riemannian structure.

We propose the following learning scheme:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \tilde{\nabla} l(\boldsymbol{x}_t, \boldsymbol{y}_t; \hat{\boldsymbol{w}}_t) \tag{5.1}$$

$$\eta_{t+1} = \eta_t \exp\{\alpha[\beta l(\boldsymbol{x}_t, \boldsymbol{y}_t; \hat{\boldsymbol{w}}_t) - \eta_t]\}, \tag{5.2}$$

where α and β are constants. We also assume that the training data are generated by a realizable deterministic teacher, and that $L(\boldsymbol{w}^*) = 0$ holds at the optimal value. See Murata et al. (1996) for a more general case. We try to analyze the dynamical behavior of learning by using the continuous version of the algorithm for the sake of simplicity,

$$\frac{d}{dt}\boldsymbol{w}_t = -\eta_t G^{-1}(\boldsymbol{w}_t)\frac{\partial}{\partial \boldsymbol{w}} l(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{w}_t), \tag{5.3}$$

$$\frac{d}{dt}\eta_t = \alpha\eta_t[\beta l(\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{w}_t) - \eta_t]. \tag{5.4}$$

In order to show the dynamical behavior of $(\hat{\boldsymbol{w}}_t, \eta_t)$, we use the averaged version of the above equation with respect to the current input-output pair $(\boldsymbol{x}_t, \boldsymbol{y}_t)$. The averaged learning equation (Amari, 1967; 1977) is written as

$$\frac{d}{dt}\boldsymbol{w}_t = -\eta_t G^{-1}(\boldsymbol{w}_t)\left\langle \frac{\partial}{\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_t)\right\rangle, \tag{5.5}$$

$$\frac{d}{dt}\eta_t = \alpha\eta_t\{\beta\langle l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_t)\rangle - \eta_t\}, \tag{5.6}$$

where $\langle \quad \rangle$ denotes the average over the current $(\boldsymbol{x}, \boldsymbol{y})$. We also use the asymptotic evaluations

$$\left\langle \frac{\partial}{\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_t)\right\rangle = \left\langle \frac{\partial}{\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}^*)\right\rangle + \left\langle \frac{\partial^2}{\partial \boldsymbol{w}\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}^*)(\boldsymbol{w}_t - \boldsymbol{w}^*)\right\rangle = G^*(\boldsymbol{w}_t - \boldsymbol{w}^*),$$

$$\langle l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_t)\rangle = \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}^*)^T G^*(\boldsymbol{w}_t - \boldsymbol{w}^*),$$

where $G^* = G(\boldsymbol{w}^*)$ and we used $L(\boldsymbol{w}^*) = 0$. We then have

$$\frac{d}{dt}\boldsymbol{w}_t = -\eta_t(\boldsymbol{w}_t - \boldsymbol{w}^*), \tag{5.7}$$

$$\frac{d}{dt}\eta_t = \alpha\eta_t\left\{\frac{\beta}{2}(\boldsymbol{w}_t - \boldsymbol{w}^*)^T G^*(\boldsymbol{w}_t - \boldsymbol{w}^*) - \eta_t\right\}. \tag{5.8}$$

Now we introduce the squared error variable

$$e_t = \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}^*)^T G^*(\boldsymbol{w}_t - \boldsymbol{w}^*), \tag{5.9}$$

where e_t is the Riemannian magnitude of $\boldsymbol{w}_t - \boldsymbol{w}^*$. It is easy to show

$$\frac{d}{dt}e_t = -2\eta_t e_t, \tag{5.10}$$

$$\frac{d}{dt}\eta_t = \alpha\beta\eta_t e_t - \alpha\eta_t^2. \tag{5.11}$$

The behavior of the above equations is interesting : The origin $(0, 0)$ is its attractor. However, the basin of attraction has a boundary of fractal structure. Anyway, starting from an adequate initial value, it has the solution of the form

$$e_t = \frac{a}{t},$$

$$\eta_t = \frac{b}{t}.$$

The coefficients a and b are determined from

$$\begin{aligned} a &= 2ab \\ b &= -\alpha\beta ab + \alpha b^2. \end{aligned}$$

This gives

$$\begin{aligned} b &= \frac{1}{2}, \\ a &= \frac{1}{\beta}\left(\frac{1}{2} - \frac{1}{\alpha}\right), \qquad \alpha > 2. \end{aligned}$$

This proves the $1/t$ convergence rate of the generalization error, that is the optimal order for any estimator $\hat{\boldsymbol{w}}_t$ converging to $\boldsymbol{w}^*$. The adaptive η_t shows a nice characteristic when the target teacher is slowly fluctuating or changes suddenly.

6 Natural Gradient in the Space of Perceptrons

The Riemannian metric and its inverse are calculated in this section to obtain the natural gradient explicitly. We begin with an analog simple perceptron whose input-output behavior is given by

$$y = f(\boldsymbol{w} \cdot \boldsymbol{x}) + n \tag{6.1}$$

where n is a Gaussian noise subject to $N(0, \sigma^2)$ and

$$f(u) = \frac{1 - e^{-u}}{1 + e^{-u}}. \tag{6.2}$$

The conditional probability density of y when $\boldsymbol{x}$ is applied is

$$p(y|\boldsymbol{x}; \boldsymbol{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}[y - f(\boldsymbol{w} \cdot \boldsymbol{x})]^2\right\}. \tag{6.3}$$

The distribution $q(\boldsymbol{x})$ of inputs $\boldsymbol{x}$ is assumed to be the normal distribution $N(0, I)$. The joint distribution of $(\boldsymbol{x}, y)$ is

$$p(y, \boldsymbol{x}; \boldsymbol{w}) = q(\boldsymbol{x})p(y|\boldsymbol{x}; \boldsymbol{w}).$$

In order to calculate the metric G of (3.11) explicitly, let us put

$$w^2 = |\boldsymbol{w}|^2 = \sum w_i^2 \tag{6.4}$$

where $|\boldsymbol{w}|$ is the Euclidean norm. We then have the following theorem.

Theorem 3. The Fisher information metric is

$$G(\boldsymbol{w}) = w^2 c_1(w) I + \{c_2(w) - c_1(w)\}\boldsymbol{w}\boldsymbol{w}^T, \tag{6.5}$$

where $c_1(w)$ and $c_2(w)$ are given by

$$\begin{aligned} c_1(w) &= \frac{1}{4\sqrt{2\pi}\sigma^2 w^2}\int \{f^2(w\varepsilon) - 1\}^2 \exp\left\{-\frac{1}{2}\varepsilon^2\right\} d\varepsilon, \\ c_2(w) &= \frac{1}{4\sqrt{2\pi}\sigma^2 w^2}\int \{f^2(w\varepsilon) - 1\}^2 \varepsilon^2 \exp\left\{-\frac{1}{2}\varepsilon^2\right\} d\varepsilon. \end{aligned}$$

Proof. We have

$$\log p(y, \boldsymbol{x}; \boldsymbol{w}) = \log q(\boldsymbol{x}) - \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}[y - f(\boldsymbol{w}\cdot\boldsymbol{x})]^2.$$

Hence,

$$\begin{aligned} \frac{\partial}{\partial w_i}\log p(y, \boldsymbol{x}; \boldsymbol{w}) &= \frac{1}{\sigma^2}\{y - f(\boldsymbol{w}\cdot\boldsymbol{x})\}f'(\boldsymbol{w}\cdot\boldsymbol{x})x_i \\ &= \frac{1}{\sigma^2} n f'(\boldsymbol{w}\cdot\boldsymbol{x})x_i. \end{aligned}$$

The Fisher information matrix is given by

$$\begin{aligned} g_{ij}(\boldsymbol{w}) &= E\left[\frac{\partial}{\partial w_i}\log p \frac{\partial}{\partial w_j}\log p\right] \\ &= \frac{1}{\sigma^2}E[\{f'(\boldsymbol{w}\cdot\boldsymbol{x})\}^2 x_i x_j], \end{aligned}$$

where $E[n^2] = \sigma^2$ is taken into account. This can be written, in the vector-matrix form, as

$$G(\boldsymbol{w}) = \frac{1}{\sigma^2}E[(f')^2\boldsymbol{x}\boldsymbol{x}^T].$$

In order to show (6.5), we calculate the quadratic form $\boldsymbol{r}^T G(\boldsymbol{w})\boldsymbol{r}$ for arbitrary $\boldsymbol{r}$. When $\boldsymbol{r} = \boldsymbol{w}$,

$$\boldsymbol{w}^T G\boldsymbol{w} = \frac{1}{\sigma^2}E[\{f'(\boldsymbol{w}\cdot\boldsymbol{x})\}^2(\boldsymbol{w}\cdot\boldsymbol{x})^2].$$

Since $u = \boldsymbol{w}\cdot\boldsymbol{x}$ is subject to $N(0, w^2)$, we put $u = w\varepsilon$, where ε is subject to $N(0,1)$. Noting that

$$f'(u) = \frac{1}{2}\{1 - f^2(u)\},$$

we have

$$\boldsymbol{w}^T G(\boldsymbol{w})\boldsymbol{w} = \frac{w^2}{4\sqrt{2\pi}\sigma^2} \int \varepsilon^2 \{f^2(w\varepsilon) - 1\}^2 \exp\{-\frac{\varepsilon^2}{2}\} d\varepsilon,$$

which confirms (6.5) when $\boldsymbol{r} = \boldsymbol{w}$. We next put $\boldsymbol{r} = \boldsymbol{v}$, where $\boldsymbol{v}$ is an arbitrary unit vector orthogonal to $\boldsymbol{w}$ (in the Euclidean sense). We then have

$$\boldsymbol{v}^T G(\boldsymbol{w})\boldsymbol{v} = \frac{1}{4\sigma^2} E[\{f^2(\boldsymbol{w}\cdot\boldsymbol{x}) - 1\}^2 (\boldsymbol{v}\cdot\boldsymbol{x})^2].$$

Since $u = \boldsymbol{w}\cdot\boldsymbol{x}$ and $v = \boldsymbol{v}\cdot\boldsymbol{x}$ are independent, and v is subject to $N(0,1)$, we have

$$\begin{aligned} \boldsymbol{v}^T G(\boldsymbol{w})\boldsymbol{v} &= \frac{1}{4\sigma^2} E[(\boldsymbol{v}\cdot\boldsymbol{x})^2] E[(f^2\{\boldsymbol{w}\cdot\boldsymbol{x}) - 1\}^2] \\ &= \frac{1}{4\sqrt{2\pi}\sigma^2} \int \{f^2(w\varepsilon) - 1\}^2 \exp\left\{-\frac{\varepsilon^2}{2}\right\} d\varepsilon. \end{aligned}$$

Since $G(\boldsymbol{w})$ in (6.5) is determined by then quadratic forms for n independent $\boldsymbol{w}$ and $\boldsymbol{v}$'s, this proves (6.5).

To obtain the natural gradient, it is necessary to have an explicit form of G^{-1}. We can calculate $G^{-1}(\boldsymbol{w})$ explicitly in the perceptron case.

Theorem 4. The inverse of the Fisher information metric is

$$G^{-1}(\boldsymbol{w}) = \frac{1}{w^2 c_1(w)} I + \frac{1}{w^4}\left(\frac{1}{c_2(w)} - \frac{1}{c_1(w)}\right)\boldsymbol{w}\boldsymbol{w}^T. \tag{6.6}$$

This can easily be proved by direct calculation of GG^{-1}. The natural gradient learning equation (3.10) is then given by

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t + \eta_t\{y_t - f(\boldsymbol{w}_t.\boldsymbol{x}_t)\} f'(\boldsymbol{w}_t\cdot\boldsymbol{x}_t) \\ &\quad \left[\frac{1}{w_t^2 c_1(w_t)}\boldsymbol{x}_t + \frac{1}{w_t^4}\left(\frac{1}{c_2(w_t)} - \frac{1}{c_1(w_t)}\right)(\boldsymbol{w}_t\cdot\boldsymbol{x}_t)\boldsymbol{w}_t\right]. \end{aligned} \tag{6.7}$$

We now show some other geometrical characteristics of the parameter space of perceptrons. The volume V_n of the manifold of simple perceptrons is measured by

$$V_n = \int \sqrt{|G(\boldsymbol{w})|} d\boldsymbol{w} \tag{6.8}$$

where $|G(\boldsymbol{w})|$ is the determinant of $G = (g_{ij})$ which represents the volume density by the Riemannian metric. It is interesting to see that the manifold of perceptrons has a finite volume.

Bayesian statistics considers that $\boldsymbol{w}$ is randomly chosen subject to a prior distribution $\pi(\boldsymbol{w})$. A choice of $\pi(\boldsymbol{w})$ is the Jeffrey prior or non-informative prior given by

$$\pi(\boldsymbol{w}) = \frac{1}{V_n}\sqrt{|G(\boldsymbol{w})|}. \tag{6.9}$$

The Jeffrey prior is calculated as follows.

Theorem 5. The Jeffrey prior and the volume of the manifold are given, respectively, by

$$\sqrt{|G(\boldsymbol{w})|} = \frac{w}{V_n}\sqrt{c_2(w)\{c_1(w)\}^{n-1}}, \tag{6.10}$$

$$V_n = a_{n-1}\int\sqrt{c_2(w)\{c_1(w)\}^{n-1}}w^n dw, \tag{6.11}$$

respectively, where a_{n-1} is the area of the unit $(n-1)$-sphere.

The Fisher metric G can also be calculated for multilayer perceptrons. Let us consider a multilayer perceptron having m hidden units with sigmoidal activation functions and a linear output unit. The input-output relation is

$$y = \sum v_i f(\boldsymbol{w}_i \cdot \boldsymbol{x}) + n$$

or the conditional probability is

$$\begin{aligned} &p(y|\boldsymbol{x};\boldsymbol{v},\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m) \\ &\quad = c\exp\left[-\frac{1}{2}\{y-\sum v_i f(\boldsymbol{w}_i\cdot\boldsymbol{x})\}^2\right]. \end{aligned} \tag{6.12}$$

The total parameter $\boldsymbol{w}$ consist of $\{\boldsymbol{v},\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m\}$. Let us calculate the Fisher information matrix G. It consists of $m+1$ blocks corresponding to these $\boldsymbol{w}_i$'s and $\boldsymbol{v}$.

From

$$\frac{\partial}{\partial \boldsymbol{w}_i}\log p(y|\boldsymbol{x};\boldsymbol{w}) = nv_i f'(\boldsymbol{w}_i\cdot\boldsymbol{x})\boldsymbol{x},$$

we easily obtain the block submatrix corresponding to $\boldsymbol{w}_i$ as

$$\begin{aligned} E\left[\frac{\partial}{\partial \boldsymbol{w}_i}\log p\frac{\partial}{\partial \boldsymbol{w}_i}\log p\right] &= E[n^2]v_i^2 E[\{f'(\boldsymbol{w}_i\cdot\boldsymbol{x})\}^2\boldsymbol{x}\boldsymbol{x}^T] \\ &= \sigma^2 v_i^2 E[\{f'(\boldsymbol{w}_i\cdot\boldsymbol{x})\}^2\boldsymbol{x}\boldsymbol{x}^T]. \end{aligned}$$

This is exactly the same as the simple perceptron case except for a factor of $(v_i)^2$. For the off-diagonal block, we have

$$E\left[\frac{\partial}{\partial \boldsymbol{w}_i}\log p\frac{\partial}{\partial \boldsymbol{w}_j}\log p\right]$$

$$= \sigma^2 v_i v_j E[f'(\boldsymbol{w}_i \cdot \boldsymbol{x}) f'(\boldsymbol{w}_j \cdot \boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^T].$$

In this case, we have the following form

$$G_{\boldsymbol{w}_i \boldsymbol{w}_j} = c_{ij} I + d_{ii} \boldsymbol{w}_i \boldsymbol{w}_i^T + d_{ij} \boldsymbol{w}_i \boldsymbol{w}_j^T + d_{ji} \boldsymbol{w}_j \boldsymbol{w}_i^T + d_{jj} \boldsymbol{w}_j \boldsymbol{w}_j^T, \tag{6.13}$$

where the coefficients c_{ij} and d_{ij}'s are calculated explicitly by similar methods.

The $\boldsymbol{v}$ block, and $\boldsymbol{v}$ and $\boldsymbol{w}_i$ block are also calculated similarly. However, the inversion of G is not easy except for simple cases. It required inversion of a $2(m+1)$ dimensional matrix. However, this is much better than the direct inversion of the original $(n+1)m$-dimensional matrix of G. Yang and Amari (1997a) performed a preliminary study on the performance of the natural gradient learning algorithm for a simple multilayer perceptron. The result shows that natural gradient learning might be free from the plateau phenomenon: Once the learning trajectory is trapped in a plateau, it takes a long time to get out of it.

7 Natural Gradient in the Space of Matrices and Blind Source Separation

We now define a Riemannian structure to the space of all the $m \times m$ nonsingular matrices, which forms a Lie group denoted by $Gl(m)$, for the purpose of introducing the natural gradient learning rule to the blind source separation problem. Let dW be a small deviation of a matrix from W to $W + dW$. The tangent space T_W of $Gl(m)$ at W is a linear space spanned by all such small deviations dW_{ij}'s, and is called the Lie algebra.

We need to introduce an inner product at W by defining the squared norm of dW

$$ds^2 = \langle dW, dW \rangle_W = \parallel dW \parallel^2 .$$

By multiplying W^{-1} from the right, W is mapped to $WW^{-1} = I$, the unit matrix, and $W + dW$ is mapped to $(W + dW)W^{-1} = I + dX$, where

$$dX = dWW^{-1}. \tag{7.1}$$

This shows that a deviation dW at W is equivalent to the deviation dX at I by the correspondence given by multiplication of W^{-1}. The Lie group invariance requires that the metric is kept invariant under this correspondence, that is, the inner product of dW at W is equal to the inner product of dWY at WY for any Y,

$$\langle dW, dW\rangle_W = \langle dWY, dWY\rangle_{WY}. \tag{7.2}$$

When $Y = W^{-1}$, $WY = I$. This principle was used to derive the natural gradient in Amari et al (1996), see also Yang and Amari (1997b) for detail. Here we give its analysis by using dX.

We define the inner product at I by

$$\langle dX, dX\rangle_I = \sum_{i,j}(dX_{ij})^2 = \mathrm{tr}(dX^T dX). \tag{7.3}$$

We then have the Riemannian metric structure at W as

$$\langle dW, dW\rangle_W = \mathrm{tr}\{(W^{-1})^T dW^T dW W^{-1}\}. \tag{7.4}$$

We can write down the metric tensor G in the component form. It is a quantity having four indices $G_{ij,kl}(W)$ such that

$$\begin{aligned} ds^2 &= \sum G_{ij,kl}(W) dW_{ij} dW_{kl}, \\ G_{ij,kl}(W) &= \sum_m \delta_{ik} W^{-1}_{jm} W^{-1}_{lm}, \end{aligned} \tag{7.5}$$

where W^{-1}_{jm} are the components of W^{-1}. While it may not appear to be straightforward to obtain the explicit form of G^{-1} and natural gradient $\tilde{\nabla} L$, in fact it can be calculated as shown below.

Theorem 6. The natural gradient in the matrix space is given by

$$\tilde{\nabla} L = \nabla L W^T W. \tag{7.6}$$

Proof. The metric is Euclidean at I, so that both $G(I)$ and its inverse $G^{-1}(I)$ are the identity. Therefore, by mapping dW at W to dX at I, the natural gradient learning rule in terms of dX is written as

$$\frac{dX}{dt} = -\eta_t G^{-1}(I)\frac{\partial L}{\partial X} = -\eta_t \frac{\partial L}{\partial X}, \tag{7.7}$$

where the continuous time version is used. We have from (7.1)

$$\frac{dX}{dt} = \frac{dW}{dt} W^{-1}. \tag{7.8}$$

The gradient $\partial L/\partial X$ is calculated as

$$\frac{\partial L}{\partial X} = \frac{\partial L(W)}{\partial W}\left(\frac{\partial W^T}{\partial X}\right) = \frac{\partial L}{\partial W}W^T.$$

Therefore, the natural gradient learning rule is

$$\frac{dW}{dt} = -\eta_t \frac{\partial L}{\partial W} W^T W,$$

which proves (7.6).

The $dX = dWW^{-1}$ forms a basis of the tangent space at W. But this is not integrable, that is, we cannot find any matrix function $X = X(W)$ that satisfies (7.1). Such a basis is called a nonholonomic basis. This is a locally defined basis but is convenient for our purpose. Let us calculate the natural gradient explicitly. To this end, we put

$$l(\boldsymbol{x}, W) = -\log\det|W| - \sum_{i-1}^{n} \log f_i(y_i), \tag{7.9}$$

where $\boldsymbol{y} = W\boldsymbol{x}$ and $f_i(y_i)$ is an adequate probability distribution. The expected loss is

$$L(W) = E[l(\boldsymbol{x}, W)]$$

which represents the entropy of the output $\boldsymbol{y}$ after a componentwise nonlinear transformation (Nadal and Praga, 1994; Bell and Sejnowski, 1996). The independent component analysis or the mutual information criterion also gives a similar loss function (Comon 1993, Amari et al., 1996). See also Oja and Karhunen, 1995. When f_i is the true probability density function of the ith source, $l(\boldsymbol{x}, W)$ is the negative of the log likelihood.

The natural gradient of l is calculated as follows. We calculate the differential

$$dl = l(\boldsymbol{x}, W + dW) - l(\boldsymbol{x}, W) = -d\log\det|W| - \sum d\log f_i(y_i)$$

due to change dW. Then,

$$\begin{aligned} d\log\det|W| &= \log\det|W + dW| - \log\det|W| \\ &= \log\det|(W + dW)W^{-1}| = \log(\det|I + dX|) \\ &= \mathrm{tr} dX. \end{aligned}$$

Similarly, from $d\boldsymbol{y} = dW\boldsymbol{x}$,

$$\begin{aligned} \sum d\log f_i(y_i) &= -\varphi(\boldsymbol{y})^T dW\boldsymbol{x} \\ &= -\varphi(\boldsymbol{y})^T dX\boldsymbol{y}, \end{aligned}$$

where $\varphi(\boldsymbol{y})$ is the column vector

$$\begin{aligned} &\varphi(\boldsymbol{y}) = [\varphi_1(y_1), \cdots, \varphi_m(y_m)], \\ &\varphi_i(y_i) = -\frac{d}{dy}\log f_i(y_i). \end{aligned} \tag{7.10}$$

This gives $\partial L/\partial X$, and the natural gradient learning equation is

$$\frac{dW}{dt} = \eta_t(I - \varphi(\boldsymbol{y})^T\boldsymbol{y})W. \tag{7.11}$$

The efficiency of this equation is studied from the statistical and information geometrical point of view (Amari and Kawanabe, 1997; Amari and Cardoso, 1997). We further calculate the Hessian by using the natural frame dX,

$$d^2 l = \boldsymbol{y}^T dX^T \dot{\varphi}(\boldsymbol{y}) dX\boldsymbol{y} + \varphi(\boldsymbol{y})^T dX dX\boldsymbol{y}, \tag{7.12}$$

where $\dot{\varphi}(\boldsymbol{y})$ is the diagonal matrix with diagonal entries $d\varphi_i(y_i)/dy_i$. Its expectation can be explicitly calculated (Amari et al. 1997a). The Hessian is decomposed into diagonal elements and two-by-two diagonal blocks (see also Cardoso and Laheld, 1996). Hence, the stability of the above learning rule is easily checked. Thus, in terms of dX, we can solve the two fundamental problems : the efficiency and stability of learning algorithms of blind source separation (Amari and Cardoso, 1997; Amari et al. 1997a).

8 Natural Gradient in Systems Space

The problem is how to define the Riemannian structure in the parameter space $\{W(z)\}$ of systems, where z is the time-shift operator. This was given in Amari (1987) from the point of view of information geometry (Amari, 1985; Murray and Rice, 1993; Amari, 1997a).

We show here only ideas (see Douglas et al. 1996; Amari et al, 1997b for preliminary studies).

In the case of multiterminal deconvolution, a typical loss function l is given by

$$l = -\log\det|W_0| - \sum_i \int p\{y_i; \boldsymbol{W}(z)\} \log f_i(y_i) dy_i, \tag{8.1}$$

where $p\{y_i; \boldsymbol{W}(z)\}$ is the marginal distribution of $\boldsymbol{y}(t)$ which is derived from the past sequence of $\boldsymbol{x}(t)$ by matrix convolution $\boldsymbol{W}(z)$ of (3.22). This type of loss function is obtained from maximization of entropy, independent component analysis or maximum likelihood.

The gradient of l is given by

$$\nabla_m l = -(W_0^{-1})^T \delta_{0m} + \boldsymbol{\varphi}(\boldsymbol{y}_t) \boldsymbol{x}^T(t-m), \tag{8.2}$$

where

$$\nabla_m = \frac{\partial}{\partial W_m},$$

and

$$\nabla l = \sum_{m=0}^{d} (\nabla_m l) z^{-m}. \tag{8.3}$$

In order to calculate the natural gradient, we need to define the Riemannian metric G in the manifold of linear systems. The geometrical theory of the manifold of linear systems by Amari (1986) defines the Riemannian metric and a pair of dual affine connections in the space of linear systems.

Let

$$d\boldsymbol{W}(z) = \sum_m d\boldsymbol{W}_m z^{-m} \tag{8.4}$$

be a small deviation of $\boldsymbol{W}(z)$. We postulate that the inner product $\langle d\boldsymbol{W}(z), d\boldsymbol{W}(z)\rangle$ is invariant under the operation of any matrix filter $\boldsymbol{Y}(z)$,

$$\langle d\boldsymbol{W}(z), d\boldsymbol{W}(z)\rangle_{\boldsymbol{W}(z)} = \langle d\boldsymbol{W}(z)\boldsymbol{Y}(z), d\boldsymbol{W}(z)\boldsymbol{Y}(z)\rangle_{\boldsymbol{WY}}, \tag{8.5}$$

where $\boldsymbol{Y}(z)$ is any system matrix. If we put

$$\boldsymbol{Y}(z) = \{\boldsymbol{W}(z)\}^{-1}$$

which is a general system not necessarily belonging to FIR,

$$\boldsymbol{W}(z)\{\boldsymbol{W}(z)\}^{-1} = \boldsymbol{I}(z),$$

which is the identity system

$$\boldsymbol{I}(z) = I$$

not including any z^{-m} terms. The tangent vector $d\boldsymbol{W}(z)$ is mapped to

$$dX(z) = d\boldsymbol{W}(z)\{\boldsymbol{W}(z)\}^{-1}. \tag{8.6}$$

The inner product at I is defined by

$$\langle dX(z), dX(z)\rangle_I = \sum_{m,ij} (dX_{m,ij})^2, \tag{8.7}$$

where $dX_{m,ij}$ are the elements of matrix dX_m.

The natural gradient

$$\tilde{\nabla} l = G^{-1} \circ \nabla l$$

of the manifold of systems is given as follows.

Theorem 7. The natural gradient of the manifold of systems is given by

$$\tilde{\nabla} l = \nabla l(z) \boldsymbol{W}^T(z^{-1}) \boldsymbol{W}(z), \tag{8.8}$$

where operator z^{-1} should be operated adequately.

The proof is omitted. It should be remarked that $\tilde{\nabla} l$ does not belong to the class of FIR systems. Nor does it satisfy the causality condition either. Hence, in order to obtain an on-line learning algorithm, we need to introduce time delay to map it to the space of causal FIR systems. The present paper only shows the principles involved, and details will published in a separate paper.

Conclusions

The present paper introduces the Riemannian structures to the parameter spaces of multilayer perceptrons, blind source separation and blind source deconvolution by means of information geometry. The natural gradient learning method is then introduced and is shown to be statistically efficient. This implies that optimal on-line learning is as efficient as optimal batch learning when the Fisher information matrix exists. It is also suggested that natural gradient learning might be easier to get out of plateaus than conventional stochastic gradient learning.

Acknowledgments

I would like to than Professor A. Cichocki, Dr. A. Back and Dr. H. Yang at RIKEN Frontier Research Program for their discussions.

References

[1] S.Amari (1967) Theory of adaptive pattern classifiers, *IEEE Trans.*, **EC-16**, No.3, pp.299–307.

[2] S. Amari (1977) Neural theory of association and concept-formation, Biological Cybernetics, **26**, pp.175–185.

[3] S. Amari (1985) *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics **28**, New York: Springer-Verlag.

[4] S. Amari (1987) Differential geometry of a parametric family of invertible linear systems — Riemannian metric, dual affine connections and divergence, *Mathematical Systems Theory*, **20**, pp.53–82.

[5] S. Amari (1993) Universal theorem on learning curves, *Neural Networks*, **6**, pp.161–166.

[6] S.Amari (1995) Learning and statistical inference, In the Handbook of Brain Theory and Neural Networks, ed. M.A. Arbib, MIT Press, pp.522–526.

[7] S. Amari (1996) Neural learning in structured parameter spaces — Natural Riemannian gradient, in *NIPS'96*, vol.**9**, MIT Press.

[8] S. Amari (1997a) Information geometry, *Contemporary Mathematics*, to appear.

[9] S. Amari (1997b) Superefficiency in blind source separation, submitted.

[10] S. Amari and J.F. Cardoso (1997) Blind source separation — Semi-parametric statistical approach, submitted.

[11] S. Amari, T.-P. Chen and A. Cichocki (1997a) Stability analysis of learning algorithms for blind source separation, *Neural Networks*, accepted.

[12] S. Amari, A. Cichocki and H.H. Yang (1996) A new learning algorithm for blind signal separation, in *NIPS'95*, vol.**8**, MIT Press, Cambridge, Mass.

[13] S. Amari, S.C. Douglas, A. Cichocki and H.H. Yang (1997b) Multichannel blind deconvolution and equalization using the natural gradient, *Signal Processing Advance in Wireless Communication Workshop*, Paris.

[14] S. Amari and M. Kawanabe (1997) Information geometry of estimating functions in semiparametric statistical models *Bernoulli*, **3**.

[15] S. Amari and N. Murata (1993) Statistical theory of learning curves under entropic loss criterion, *Neural Computation*, **5**, pp.140–153.

[16] N. Barkai, H.S. Seung and H. Sompolinski (1995) Local and global convergence of on-line learning, *Phys. Rev. Lett.*, **75**, 1415–1418.

[17] A.J. Bell and T.J. Sejnowski (1995) An information-maximization approach to blind separation and blind deconvolution, *Neural Computation*, **7**, pp.1129–1159.

[18] P.J. Bickel, C.A.J. Klassen, Y. Ritov and J.A. Wellner (1993) *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press.

[19] L.L. Campbell (1985) The relation between information theory and the differential-geometric approach to statistics, *Information Sciences*, **35**, pp.199–210.

[20] J.F. Cardoso and B. Laheld (1996) Equivariant adaptive source separation, *IEEE Trans. on Signal Processing*, **44**, 3017–3030.

[21] N.N. Chentsov (1972) *Statistical decision rules and optimal inference* (in Russian), Moscow: Nauka [translated in English (1982), Rhode Island: AMS].

[22] P. Common (1994) Independent component analysis, a new concept?, *Signal Processing*, **36**, pp.287–314.

[23] S.C. Douglas, A. Cichocki and S. Amari (1996) Fast convergence filtered regressor algorithms for blind equalization, *Electronics Letters*, **32**, pp.2114–2115.

[24] T. Heskes and B. Kappen (1991) Learning process in neural networks, *Physical Review*, **A44**, pp.2718–2762.

[25] C. Jutten and J. Hérault (1991) Blind separation of sources, an adaptive algorithm based on neuromimetic architecture, *Signal Processing*, **24**, No.1, pp.1–31.

[26] H.J. Kushner and D.S. Clark (1978) *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag.

[27] N. Murata, K.-R. Müller, A. Ziehe and S. Amari (1996) Adaptive on-line learning in changing environments, in *NIPS'96*, MIT Press, vol. **9**.

[28] M.K. Murray and J.W. Rice (1993) *Differential Geometry and Statistics*, New York: Chapman & Hall.

[29] J.P. Nadall and N. Parga (1994) Nonlinear neurons in the low noise limit — A factorial code maximizes information transfer, *Network*, **5**, pp.561–581.

[30] E. Oja and J. Karhunen (1995) Signal separation by nonlinear Hebbian learning, in M. Palaniswami et al. (Eds.), *Computational Intelligence — A Dynamic Systems Perspective*, IEEE Press, New York, pp.83–97.

[31] M. Opper (1996) Online versus offline learning from random examples: General results, *Phys. Rev. Lett.*, **77**, pp.4671–4674.

[32] C.R. Rao (1945) Information and accuracy attainable in the estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society*, **37**, pp.81–91.

[33] D.E. Rumelhart, G.E. Hinton and R.J. Williams (1986) Learning internal representations by error propagation, *Parallel Distrib. Process.*, Cambridge, MA: MIT Press, **1**, pp.318–362.

[34] D. Saad and S.A. Solla (1995) On-line learning in soft committee machines, *Phys. Rev.* E, **52**, pp.4225–4243.

[35] H. Sompolinsky, N. Barkai and H.S. Seung (1995) On-line learning of dichotomies: algorithms and learning curves, *Neural Networks: The statistical Mechanics Perspective*, Proceedings of the CTP-PBSRI Joint Workshop on Theoretical Physics, J.-H. Oh et al eds, pp.105–130.

[36] Ya. Z. Tsypkin (1973) *Foundation of the Theory of Learning Systems*, Academic Press, New York.

[37] C. Van den Broeck and P. Reimann (1996) Unsupervised learning by examples: On-line versus off-line, *Phys. Rev. Lett.*, **76**, pp.2188–2191.

[38] B. Widrow (1963) *A Statistical Theory of Adaptation*, Pergamon Press, Oxford.

[39] H.H. Yang and S. Amari (1997a) Application of natural gradient in training multilayer perceptrons, in preparation.

[40] H.H. Yang and S. Amari (1997b) Adaptive on-line learning algorithms for blind separation — Maximum entropy and minimal mutual information, *Neural Computation*, accepted.