ARTICLE Communicated by Steven Nowlan and Erkki Oja

Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program, Saitama 351-01, Japan

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 online 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 not be 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, Hinton, & Williams, 1986) is a popular learning method in the general nonlinear optimization framework. The parameter space is not Euclidean but has a Riemannian metric structure in many cases. In these cases, the ordinary gradient does not give the steepest direction of a target function; rather, the steepest direction is given by the natural (or contravariant) gradient. The Riemannian metric structures are introduced by means of information geometry (Amari, 1985; Murray and Rice, 1993; Amari, 1997a; Amari, Kurata, & Nagoska, 1992). This article 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 article (Amari, 1996), including new results.

How good is natural gradient learning compared to conventional gradient learning? The asymptotic behavior of online natural gradient learning is studied for this purpose. Training examples can be used only once in online learning when they appear. Therefore, the asymptotic performance of online learning cannot be better than the optimal batch procedure where all the examples can be reused again and again. However, we prove that natural gradient online 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 nondifferentiable, the accuracy of asymptotic online learning is worse than batch learning by a factor of 2 (see, for example, Van den Broeck & Reimann, 1996). It was shown in Amari et al. (1992) that the dynamic behavior of natural gradient in the Boltzmann machine is excellent.

It is not easy to calculate the natural gradient explicitly in multilayer perceptrons. However, a preliminary analysis (Yang & Amari, 1997), 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 & Solla, 1995). This suggests that the Riemannian structure might eliminate such plateaus or might make them not so serious.

Online learning is flexible, because it can track slow fluctuations of the target. Such online dynamics were first analyzed in Amari (1967) and then by many researchers recently. Sompolinsky, Barkai, and Seung (1995), and Barkai, Seung, and Sompolinsky (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 article 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 online 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 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 nonorthonormal, 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 in equation 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,

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

the superscript T denoting the transposition.

Proof. We put

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

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

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

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 that determines the step size.

3 Natural Gradient Learning

Let us consider an information source that generates a sequence of independent random variables $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t, \ldots$, subject to the same probability distribution $q(\boldsymbol{z})$. The random signals $\boldsymbol{z}_t$ are processed by a processor (like a neural network) that 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 that may depend on t and $C(\boldsymbol{w})$ is a suitably chosen positive definite matrix (see Amari, 1967). In the natural gradient online 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.

3.1 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 $p(\boldsymbol{z}, \hat{\boldsymbol{w}})$ that 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 (see equation 3.2) gives a sequential estimator $\hat{\boldsymbol{w}}_t$.

3.2 Multilayer Neural Network. Let us consider a multilayer feedforward neural network specified by a vector parameter $\boldsymbol{w} = (w_1, \ldots, w_n)^T \in \boldsymbol{R}^n$. 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 that, by receiving $\boldsymbol{x}$, generates the corresponding output $\boldsymbol{y}$ subject to a conditional probability distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$. 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

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

where c is a normalizing constant and the loss function (see equation 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 $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_t, \boldsymbol{y}_t), \ldots$, the natural gradient online 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.

3.3 Blind Separation of Sources. Let us consider m signal sources that produce m independent signals $s_i(t)$, $i = 1, \ldots, m$, at discrete times $t = 1, 2, \ldots$. 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

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

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 that does not depend on t, and $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ is the observed mixtures.

Blind source separation is the problem of recovering the original signals $\boldsymbol{s}(t)$, $t = 1, 2, \ldots$ from the observed signals $\boldsymbol{x}(t)$, $t = 1, 2, \ldots$ (Jutten & Hérault, 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 and 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 $r(\boldsymbol{s})$ in equation 3.12 when $W = A^{-1}$. For W_t of equation 3.15 to converge to A^{-1}, equation 3.16 is necessary but not sufficient, because the stability of the equilibrium is not considered here.

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

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

satisfies equation 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 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 an F is also obtained by heuristic arguments. Amari and Cardoso (in press) gave the complete family of F satisfying equation 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 $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(t)$. 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, Klassen, Ritov, & Wellner, 1993), because it includes an unknown function of infinite degrees of freedom. However, we can apply the information-geometrical theory of estimating functions (Amari & Kawanabe, 1997) to this problem.

When F is given by the gradient of a loss function (see equation 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, Chen, & Chichocki, in press; see also Cardoso & Laheld, 1996). The superefficiency of some algorithms has been also proved in Amari (1997b) under certain conditions.

3.4 Blind Source Deconvolution. When the original signals $\boldsymbol{s}(t)$ are mixed 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 finite impulse response 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

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

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

In order to obtain the natural gradient learning algorithm

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

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, Chichocki, and Amari (1996) and Amari et al. (1997).

4 Natural Gradient Gives Fisher-Efficient Online Learning Algorithms

This section studies the accuracy of natural gradient learning from the statistical point of view. A statistical estimator that gives asymptotically the best result is said to be Fisher efficient. We prove that natural gradient learning attains 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 $q(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}^*)$. Let $D_T = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_T, \boldsymbol{y}_T)\}$ 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 equation 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 equation 3.11.

Examples $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \ldots$ are given one at a time in the case of online learning. Let $\tilde{\boldsymbol{w}}_t$ be an online 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 $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})$ cannot be reused to obtain $\tilde{\boldsymbol{w}}_{t+1}$, so the learning rule is written as

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

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 online 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 into online 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 online learning is equivalent to that of the best batch estimation method. This article answers the question affirmatively, by giving an efficient online 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 (see equation 4.4), the natural gradient online 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 equation 4.4 and 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 certain conditions (see Kushner & Clark, 1978). The solution of equation 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 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. 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 & Murata, 1993).

Remark. In the cases of blind source separation and deconvolution, the models are semiparametric, including the unknown function r (see equation 3.18). In such cases, the Cramér-Rao bound does not necessarily hold. 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 equation 4.8 does not hold. The stability of the true solution is not necessarily guaranteed either. Amari, Chen, & Cichocki (in press) have analyzed this situation and proposed a universal method of attaining the stability of the equilibrium solution.

5 Adaptive Learning Constant

The dynamical behavior of the learning rule (see equation 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."

Sompolinsky 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 $\boldsymbol{w}^*$. This article 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, Müller, Ziehe, and Amari (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 $(\boldsymbol{w}_t, \eta_t)$, we use the averaged version of equations 5.3 and 5.4 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

$$\begin{aligned}\left\langle \frac{\partial}{\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_t)\right\rangle &= \left\langle \frac{\partial}{\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}^*)\right\rangle + \left\langle \frac{\partial^2}{\partial \boldsymbol{w}\partial \boldsymbol{w}} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}^*)(\boldsymbol{w}_t - \boldsymbol{w}^*)\right\rangle \\ &= G^*(\boldsymbol{w}_t - \boldsymbol{w}^*), \\ \langle l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_t)\rangle &= \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}^*)^T G^*(\boldsymbol{w}_t - \boldsymbol{w}^*),\end{aligned}$$

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 equations 5.10 and 5.11 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

$$a = 2ab$$

$$b = -\alpha\beta ab + \alpha b^2.$$

This gives

$$b = \frac{1}{2},$$

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

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

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

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

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

In order to calculate the metric G of equation 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

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

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

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

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 equation 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,

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

which confirms equation 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

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

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

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

The total parameter $\boldsymbol{w}$ consist of $\{\boldsymbol{v}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$. 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

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

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

$$\begin{aligned} E\left[\frac{\partial}{\partial \boldsymbol{w}_i} \log p \frac{\partial}{\partial \boldsymbol{w}_i} \log p\right] &= \frac{1}{\sigma^4} E[n^2] v_i^2 E[\{f'(\boldsymbol{w}_i \cdot \boldsymbol{x})\}^2 \boldsymbol{x}\boldsymbol{x}^T] \\ &= \frac{1}{\sigma^2} v_i^2 E[\{f'(\boldsymbol{w}_i \cdot \boldsymbol{x})\}^2 \boldsymbol{x}\boldsymbol{x}^T]. \end{aligned}$$

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

$$E\left[\frac{\partial}{\partial \boldsymbol{w}_i} \log p \frac{\partial}{\partial \boldsymbol{w}_j} \log p\right] = \frac{1}{\sigma^2} v_i v_j E[f'(\boldsymbol{w}_i \cdot \boldsymbol{x}) f'(\boldsymbol{w}_j \cdot \boldsymbol{x}) \boldsymbol{x}\boldsymbol{x}^T].$$

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 requires 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 (1997) 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

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

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, Cichocki, and Yang (1996); see also Yang and Amari (1997) 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

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

We can write 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_{jm}^{-1} 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 equation 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 equation 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 equation 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 & Parga, 1994; Bell & Sejnowski, 1995). The independent component analysis or the mutual information criterion also gives a similar loss function (Comon, 1994; Amari et al., 1996; see also Oja & 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

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

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 & Kawanabe, 1997; Amari & Cardoso, in press). 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., in press). The Hessian is decomposed into diagonal elements and two-by-two diagonal blocks (see also Cardoso & 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 the stability of learning algorithms of blind source separation (Amari & Cardoso, in press; Amari et al., in press).

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, 1997a; Murray & Rice, 1993). We show here only ideas (see Douglas et al., 1996; Amari, Douglas, Cichocki, & Yang, 1997, 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 equation 3.24. 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 (1987) 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 $\boldsymbol{Y}(z)$,

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

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 online learning algorithm, we need to introduce time delay to map it to the space of causal FIR systems. This article shows only the principles involved; details will published in a separate article by Amari, Douglas, and Cichocki.

9 Conclusions

This article 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 online 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 thank A. Cichocki, A. Back, and H. Yang at RIKEN Frontier Research Program for their discussions.

References

Amari, S. (1967). Theory of adaptive pattern classifiers. *IEEE Trans., EC-16*(3), 299–307.

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

Amari, S. (1985). *Differential-geometrical methods in statistics*. Lecture Notes in Statistics 28. New York: Springer-Verlag.

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

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

Amari, S. (1995). Learning and statistical inference. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 522–526). Cambridge, MA: MIT Press.

Amari, S. (1996). Neural learning in structured parameter spaces—Natural Riemannian gradient. In M. C. Mozer, M. I. Jordan, & Th. Petsche (Eds.), *Advances in neural processing systems, 9*. Cambridge, MA: MIT Press.

Amari, S. (1997a). Information geometry. *Contemporary Mathematics, 203*, 81–95.

Amari, S. (1997b). *Superefficiency in blind source separation*. Unpublished manuscript.

Amari, S., & Cardoso, J. F. (In press). Blind source separation—Semi-parametric statistical approach. *IEEE Trans. on Signal Processing*.

Amari, S., Chen, T.-P., & Cichocki, A. (In press). Stability analysis of learning algorithms for blind source separation. *Neural Networks*.

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation, in *NIPS'95*, vol. 8, Cambridge, MA: MIT Press.

Amari, S., Douglas, S. C., Cichocki, A., & Yang, H. H. (1997). Multichannel blind deconvolution and equalization using the natural gradient. *Signal Processing*

Advance in Wireless Communication Workshop, Paris.

Amari, S., & Kawanabe, M. (1997). Information geometry of estimating functions in semiparametric statistical models, *Bernoulli, 3*, 29–54.

Amari, S., Kurata, K., & Nagaoka, H. (1992). Information geometry of Boltzmann machines. *IEEE Trans. on Neural Networks, 3*, 260–271.

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

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

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

Bickel, P. J., Klassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Baltimore: Johns Hopkins University Press.

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

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

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

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*, 287–314.

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

Heskes, T., & Kappen, B. (1991). Learning process in neural networks. *Physical Review, A44*, 2718–2762.

Jutten, C., & Hérault, J. (1991). Blind separation of sources, an adaptive algorithm based on neuromimetic architecture. *Signal Processing, 24*(1), 1–31.

Kushner, H. J., & Clark, D. S. (1978). *Stochastic approximation methods for constrained and unconstrained systems*. Berlin: Springer-Verlag.

Murata, N., & Müller, K. R., Ziehe, A., & Amari, S. (1996). Adaptive on-line learning in changing environments. In M. C. Mozer, M. I. Jordan, & Th. Petsche (Eds.), *Advaces in neural processing systems, 9*. Cambridge, MA: MIT Press.

Murray, M. K., & Rice, J. W. (1993). *Differential geometry and statistics*. New York: Chapman & Hall.

Nadal, J. P. & Parga, N. (1994). Nonlinear neurons in the low noise limit—A factorial code maximizes information transfer. *Network, 5*, 561–581.

Oja, E., & Karhunen, J. (1995). Signal separation by nonlinear Hebbian learning. In M. Palaniswami et al. (Eds.), *Computational intelligence—A dynamic systems perspective* (pp. 83–97). New York: IEEE Press.

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

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

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

Sompolinsky, H., Barkai, N., & Seung, H. S. (1995). On-line learning of dichotomies: Algorithms and learning curves. In J.-H. Oh et al. (Eds.), *Neural networks: The statistical mechanics perspective* (pp. 105–130). Proceedings of the CTP-PBSRI Joint Workshop on Theoretical Physics. Singapore: World Scientific.

Tsypkin, Ya. Z. (1973). *Foundation of the theory of learning systems*. New York: Academic Press.

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

Widrow, B. (1963). *A statistical theory of adaptation*. Oxford: Pergamon Press.

Yang, H. H., & Amari, S. (1997). *Application of natural gradient in training multilayer perceptrons*. Unpublished manuscript.

Yang, H. H., & Amari, S. (In press). Adaptive on-line learning algorithms for blind separation—Maximum entropy and minimal mutual information. *Neural Computation*.

Received January 24, 1997; accepted May 20, 1997.

This article has been cited by:

1. Manjunath Ramachandra, Pandit PattabhiramaAnalysis of the High-Speed Network Performance through a Prediction Feedback Based Model 162-178. [CrossRef]
2. Wentao Fan, Nizar Bouguila. 2012. Online variational learning of finite Dirichlet mixture models. *Evolving Systems* . [CrossRef]
3. Dimitrije Markovi#, Claudius Gros. 2012. Intrinsic Adaptation in Autonomous Recurrent Neural Networks. *Neural Computation* **24**:2, 523-540. [Abstract] [Full Text] [PDF] [PDF Plus]
4. Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno. 2012. Efficient Blind Dereverberation and Echo Cancellation Based on Independent Component Analysis for Actual Acoustic Signals. *Neural Computation* **24**:1, 234-272. [Abstract] [Full Text] [PDF] [PDF Plus]
5. Zhu Shou Zhong, Liu Zheng, Jiang Wen Li, Guo Kun. 2012. The Key technology of Blind Source Separation of Satellite-Based AIS. *Procedia Engineering* **29**, 3737-3741. [CrossRef]
6. Nihat Ay, Holger Bernigau, Ralf Der, Mikhail Prokopenko. 2011. Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theory in Biosciences* . [CrossRef]
7. Georg Martius, J. Michael Herrmann. 2011. Variants of guided self-organization for robot control. *Theory in Biosciences* . [CrossRef]
8. Filip Jur#í#ek, Blaise Thomson, Steve Young. 2011. Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech & Language* . [CrossRef]
9. Youhei Akimoto, Yuichi Nagata, Isao Ono, Shigenobu Kobayashi. 2011. Theoretical Foundation for CMA-ES from Information Geometry Perspective. *Algorithmica* . [CrossRef]
10. Anthony Lombard, Yuanhang Zheng, Herbert Buchner, Walter Kellermann. 2011. TDOA Estimation for Multiple Sound Sources in Noisy and Reverberant Environments Using Broadband Independent Component Analysis. *IEEE Transactions on Audio, Speech, and Language Processing* **19**:6, 1490-1503. [CrossRef]
11. Inkyung Ahn, Jooyoung Park. 2011. Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Biosystems* . [CrossRef]
12. Bibliography **20110657**, . [CrossRef]
13. Pontus Johannisson, Henk Wymeersch, Martin Sjödin, A. Serdar Tan, Erik Agrell, Peter A. Andrekson, Magnus Karlsson. 2011. Convergence Comparison of the CMA and ICA for Blind Polarization Demultiplexing. *Journal of Optical Communications and Networking* **3**:6, 493. [CrossRef]
14. Rupert C.J. Minnett, Andrew T. Smith, William C. Lennon, Robert Hecht-Nielsen. 2011. Neural network tomography: Network replication from output surface geometry. *Neural Networks* **24**:5, 484-492. [CrossRef]
15. Jianwei Wu. 2011. Estimating source kurtosis directly from observation data for ICA. *Signal Processing* **91**:5, 1150-1156. [CrossRef]
16. Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ond#ej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow. 2011. The subspace Gaussian mixture model—A structured model for speech recognition. *Computer Speech & Language* **25**:2, 404-439. [CrossRef]
17. Haihong Zhang, Cuntai Guan, Yuanqing Li. 2011. A linear discriminant analysis method based on mutual information maximization. *Pattern Recognition* **44**:4, 877-885. [CrossRef]
18. Francesco Nesta, Piergiorgio Svaizer, Maurizio Omologo. 2011. Convolutive BSS of Short Mixtures by ICA Recursively Regularized Across Frequencies. *IEEE Transactions on Audio, Speech, and Language Processing* **19**:3, 624-639. [CrossRef]
19. Francesco Nesta, Ted S Wada, Biing-Hwang Juang. 2011. Batch-Online Semi-Blind Source Separation Applied to Multi-Channel Acoustic Echo Cancellation. *IEEE Transactions on Audio, Speech, and Language Processing* **19**:3, 583-599. [CrossRef]
20. Masashi Sugiyama, Makoto Yamada, Paul von Bünau, Taiji Suzuki, Takafumi Kanamori, Motoaki Kawanabe. 2011. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks* **24**:2, 183-198. [CrossRef]
21. Model Design and Selection Considerations **20113128**, 37-63. [CrossRef]
22. Taiji Suzuki, Masashi Sugiyama. 2011. Least-Squares Independent Component Analysis. *Neural Computation* **23**:1, 284-301. [Abstract] [Full Text] [PDF] [PDF Plus]

23. Takuya Yoshioka, Tomohiro Nakatani, Masato Miyoshi, Hiroshi G. Okuno. 2011. Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization. *IEEE Transactions on Audio, Speech, and Language Processing* **19**:1, 69-84. [CrossRef]

24. Simone Fiori. 2011. Solving Minimal-Distance Problems over the Manifold of Real-Symplectic Matrices. *SIAM Journal on Matrix Analysis and Applications* **32**:3, 938. [CrossRef]

25. Tao Yu, John H L Hansen. 2010. Discriminative Training for Multiple Observation Likelihood Ratio Based Voice Activity Detection. *IEEE Signal Processing Letters* **17**:11, 897-900. [CrossRef]

26. Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, Andrew Cotter. 2010. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming* . [CrossRef]

27. Dijun Luo, Heng Huang, Chris Ding, Feiping Nie. 2010. On the eigenvectors of p-Laplacian. *Machine Learning* **81**:1, 37-51. [CrossRef]

28. Zhao Zhang, Man Jiang, Ning Ye. 2010. Effective multiplicative updates for non-negative discriminative learning in multimodal dimensionality reduction. *Artificial Intelligence Review* **34**:3, 235-260. [CrossRef]

29. Shun-ichi Amari. 2010. Information geometry in optimization, machine learning and statistical inference. *Frontiers of Electrical and Electronic Engineering in China* **5**:3, 241-260. [CrossRef]

30. Sven Hoffmann, Michael Falkenstein. 2010. Independent component analysis of erroneous and correct responses suggests online response control. *Human Brain Mapping* **31**:9, 1305-1315. [CrossRef]

31. Ying Tang, Jianping Li. 2010. Normalized natural gradient in independent component analysis. *Signal Processing* **90**:9, 2773-2777. [CrossRef]

32. Hirofumi Nakajima, Kazuhiro Nakadai, Yuji Hasegawa, Hiroshi Tsujino. 2010. Blind Source Separation With Parameter-Free Adaptive Step-Size Method for Robot Audition. *IEEE Transactions on Audio, Speech, and Language Processing* **18**:6, 1476-1485. [CrossRef]

33. Simone Fiori. 2010. Learning by Natural Gradient on Noncompact Matrix-Type Pseudo-Riemannian Manifolds. *IEEE Transactions on Neural Networks* **21**:5, 841-852. [CrossRef]

34. Byungchan Kim, Jooyoung Park, Shinsuk Park, Sungchul Kang. 2010. Impedance Learning for Robotic Contact Tasks Using Natural Actor-Critic Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **40**:2, 433-443. [CrossRef]

35. Francisco das Chagas de Souza, Orlando José Tobias, Rui Seara, Dennis R. Morgan. 2010. A PNLMS Algorithm With Individual Activation Factors. *IEEE Transactions on Signal Processing* **58**:4, 2036-2047. [CrossRef]

36. Y. Panagakis, C. Kotropoulos, G.R. Arce. 2010. Non-Negative Multilinear Principal Component Analysis of Auditory Temporal Modulations for Music Genre Classification. *IEEE Transactions on Audio, Speech, and Language Processing* **18**:3, 576-588. [CrossRef]

37. V. Zarzoso, P. Comon. 2010. Robust Independent Component Analysis by Iterative Maximization of the Kurtosis Contrast With Algebraic Optimal Step Size. *IEEE Transactions on Neural Networks* **21**:2, 248-261. [CrossRef]

38. Csaba Szepesvári. 2010. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **4**:1, 1-103. [CrossRef]

39. Addisson Salazar, Luis Vergara, Arturo Serrano, Jorge Igual. 2010. A general procedure for learning mixtures of independent component analyzers. *Pattern Recognition* **43**:1, 69-85. [CrossRef]

40. Chun-Nan Hsu, Han-Shen Huang, Yu-Ming Chang, Yuh-Jye Lee. 2009. Periodic step-size adaptation in second-order gradient descent for single-pass on-line structured learning. *Machine Learning* **77**:2-3, 195-224. [CrossRef]

41. Yunfeng Xue, Feng Ju, Yujia Wang, Jie Yang. 2009. A source adaptive independent component analysis algorithm through solving the estimating equation. *Expert Systems with Applications* **36**:10, 12306-12313. [CrossRef]

42. Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, Mark Lee. 2009. Natural actor–critic algorithms#. *Automatica* **45**:11, 2471-2482. [CrossRef]

43. Manuele Bicego, El#bieta Pe¸kalska, David M.J. Tax, Robert P.W. Duin. 2009. Component-based discriminative classification for hidden Markov models. *Pattern Recognition* **42**:11, 2637-2648. [CrossRef]

44. Nizar Bouguila, Ola Amayri. 2009. A discrete mixture-based kernel for SVMs: Application to spam and image categorization. *Information Processing & Management* **45**:6, 631-642. [CrossRef]

45. Yoshitatsu Matsuda, Kazunori Yamaguchi. 2009. Linear Multilayer ICA Using Adaptive PCA. *Neural Processing Letters* **30**:2, 133-144. [CrossRef]

46. Jing Sui, Tülay Adali, Godfrey D. Pearlson, Vincent P. Clark, Vince D. Calhoun. 2009. A method for accurate group difference detection by constraining the mixing coefficients in an ICA framework. *Human Brain Mapping* **30**:9, 2953-2970. [CrossRef]

47. Zhenning Zhang, Huafei Sun, Fengwei Zhong. 2009. Natural gradient-projection algorithm for distribution control. *Optimal Control Applications and Methods* **30**:5, 495-504. [CrossRef]

48. Ken-ichi Tamura, Miho Komiya, Masato Inoue, Yoshiyuki Kabashima. 2009. Decoding Algorithm of Low-density Parity-check Codes based on Bowman-Levin Approximation. *New Generation Computing* **27**:4, 347-363. [CrossRef]

49. Makoto Miyakoshi, Moyoko Tomiyasu, Epifanio Bagarinao, Shumei Murakami, Toshiharu Nakai. 2009. A Phantom Study On Component Segregation for MR Images Using ICA. *Academic Radiology* **16**:8, 1025-1028. [CrossRef]

50. Haiting Tian, Jing Jin, Chunxi Zhang, Ningfang Song. 2009. Informax-Based Data Fusion for Sensor Network. *IEEE Sensors Journal* **9**:7, 820-827. [CrossRef]

51. Frank Nielsen, Richard Nock. 2009. Sided and Symmetrized Bregman Centroids. *IEEE Transactions on Information Theory* **55**:6, 2882-2904. [CrossRef]

52. Anand Oka, Lutz Lampe. 2009. Incremental Distributed Identification of Markov Random Field Models in Wireless Sensor Networks. *IEEE Transactions on Signal Processing* **57**:6, 2396-2405. [CrossRef]

53. L.R. Vega, H. Rey, J. Benesty, S. Tressens. 2009. A Family of Robust Algorithms Exploiting Sparsity in Adaptive Filters. *IEEE Transactions on Audio, Speech, and Language Processing* **17**:4, 572-581. [CrossRef]

54. Jin-Sung Yoon, Gye-Young Kim, Hyung-Il Choi. 2009. Development of an Adult Image Classifier using Skin Color. *The Journal of the Korea Contents Association* **9**:4, 1-11. [CrossRef]

55. Jae-Min Kim, Hyun-Soo Kang. 2009. Efficient Coding Technique for 4X4 Intra Prediction Modes using the Statistical Distribution of Intra Modes of Adjacent Intra Blocks. *The Journal of the Korea Contents Association* **9**:4, 12-18. [CrossRef]

56. Sang-Hoon Oh. 2009. Comparisons of Linear Feature Extraction Methods. *The Journal of the Korea Contents Association* **9**:4, 121-130. [CrossRef]

57. Jian-Qiang Liu, Da-Zheng Feng, Wei-Wei Zhang. 2009. Adaptive Improved Natural Gradient Algorithm for Blind Source Separation. *Neural Computation* **21**:3, 872-889. [Abstract] [Full Text] [PDF] [PDF Plus]

58. Masahiro Yukawa. 2009. Krylov-Proportionate Adaptive Filtering Techniques Not Limited to Sparse Systems. *IEEE Transactions on Signal Processing* **57**:3, 927-943. [CrossRef]

59. Sam McKennoch, Thomas Voegtlin, Linda Bushnell. 2009. Spike-Timing Error Backpropagation in Theta Neuron Networks. *Neural Computation* **21**:1, 9-45. [Abstract] [Full Text] [PDF] [PDF Plus]

60. Jingyu Liu, Godfrey Pearlson, Andreas Windemuth, Gualberto Ruano, Nora I. Perrone-Bizzozero, Vince Calhoun. 2009. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human Brain Mapping* **30**:1, 241-255. [CrossRef]

61. Elizabeth Hoppe, Michael Roan. 2009. Non-linear, adaptive array processing for acoustic interference suppression. *The Journal of the Acoustical Society of America* **125**:6, 3835. [CrossRef]

62. Tetsuro Morimura, Eiji Uchibe, Kenji Doya. 2008. Natural actor-critic with baseline adjustment for variance reduction. *Artificial Life and Robotics* **13**:1, 275-279. [CrossRef]

63. Wai Yie Leong, Danilo P. Mandic. 2008. Post-Nonlinear Blind Extraction in the Presence of Ill-Conditioned Mixing. *IEEE Transactions on Circuits and Systems I: Regular Papers* **55**:9, 2631-2638. [CrossRef]

64. Sam McKennoch, Thomas Voegtlin, Linda Bushnell. 2008. Spike-Timing Error Backpropagation in Theta Neuron Networks. *Neural Computation*, ahead of print080804143617793-37. [CrossRef]

65. U. Manmontri, P.A. Naylor. 2008. A Class of Frobenius Norm-Based Algorithms Using Penalty Term and Natural Gradient for Blind Signal Separation. *IEEE Transactions on Audio, Speech, and Language Processing* **16**:6, 1181-1193. [CrossRef]

66. F. Cousseau, T. Ozeki, S.-i. Amari. 2008. Dynamics of Learning in Multilayer Perceptrons Near Singularities. *IEEE Transactions on Neural Networks* **19**:8, 1313-1328. [CrossRef]

67. B. Baddeley. 2008. Reinforcement Learning in Continuous Time and Space: Interference and Not Ill Conditioning Is the Main Problem When Using Distributed Function Approximators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **38**:4, 950-956. [CrossRef]

68. Zhishun Wang, Bradley S. Peterson. 2008. Partner#matching for the automated identification of reproducible ICA components from fMRI datasets: Algorithm and validation. *Human Brain Mapping* **29**:8, 875-893. [CrossRef]

69. R HORIE. 2008. An optimization framework of biological dynamical systems. *Journal of Theoretical Biology* **253**:1, 45-54. [CrossRef]

70. Anand Oka, Lutz Lampe. 2008. Energy Efficient Distributed Filtering With Wireless Sensor Networks. *IEEE Transactions on Signal Processing* **56**:5, 2062-2075. [CrossRef]
71. J PETERS, S SCHAAL. 2008. Reinforcement learning of motor skills with policy gradients. *Neural Networks* **21**:4, 682-697. [CrossRef]
72. Simone Fiori. 2008. A Study on Neural Learning on Manifold Foliations: The Case of the Lie Group SU(3). *Neural Computation* **20**:4, 1091-1117. [Abstract] [PDF] [PDF Plus]
73. Zhaoshui He, Shengli Xie, Liqing Zhang, Andrzej Cichocki. 2008. A Note on Lewicki-Sejnowski Gradient for Learning Overcomplete Representations. *Neural Computation* **20**:3, 636-643. [Abstract] [PDF] [PDF Plus]
74. Haikun Wei, Jun Zhang, Florent Cousseau, Tomoko Ozeki, Shun-ichi Amari. 2008. Dynamics of Learning Near Singularities in Layered Networks. *Neural Computation* **20**:3, 813-843. [Abstract] [PDF] [PDF Plus]
75. Ah Chung Tsoi, Liangsuo Ma. 2008. A Balanced Approach to Multichannel Blind Deconvolution. *IEEE Transactions on Circuits and Systems I: Regular Papers* **55**:2, 599-613. [CrossRef]
76. Hualiang Li, T. Adali. 2008. A Class of Complex ICA Algorithms Based on the Kurtosis Cost Function. *IEEE Transactions on Neural Networks* **19**:3, 408-420. [CrossRef]
77. Nuo Zhang, Jianming Lu, Takashi Yahagi. 2008. A method of independent component analysis based on radial basis function networks using noise estimation. *Electronics and Communications in Japan* **91**:3, 45-52. [CrossRef]
78. Traian E. Abrudan, Jan Eriksson, Visa Koivunen. 2008. Steepest Descent Algorithms for Optimization Under Unitary Matrix Constraint. *IEEE Transactions on Signal Processing* **56**:3, 1134-1147. [CrossRef]
79. Z YANG, J LAAKSONEN. 2008. Principal whitened gradient for information geometry#. *Neural Networks* **21**:2-3, 232-240. [CrossRef]
80. Hualiang Li, Tülay Adal#. 2008. Complex-Valued Adaptive Signal Processing Using Nonlinear Functions. *EURASIP Journal on Advances in Signal Processing* **2008**, 1-10. [CrossRef]
81. David N. Levin. 2008. Using state space differential geometry for nonlinear blind source separation. *Journal of Applied Physics* **103**:4, 044906. [CrossRef]
82. Rafik Zayani, Ridha Bouallegue, Daniel Roviras. 2008. Adaptive Predistortions Based on Neural Networks Associated with Levenberg-Marquardt Algorithm for Satellite Down Links. *EURASIP Journal on Wireless Communications and Networking* **2008**, 1-16. [CrossRef]
83. Liangsuo Ma, Ah Chung Tsoi. 2007. A unified balanced approach to multichannel blind deconvolution. *Signal, Image and Video Processing* **1**:4, 369-384. [CrossRef]
84. Y NAKAMURA, T MORI, M SATO, S ISHII. 2007. Reinforcement learning for a biped robot based on a CPG-actor-critic method. *Neural Networks* **20**:6, 723-735. [CrossRef]
85. Masatoshi Funabashi, Kazuyuki Aihara. 2007. Modeling birdsong learning with a chaotic Elman network. *Artificial Life and Robotics* **11**:2, 162-166. [CrossRef]
86. A GONZALEZ, J DORRONSORO. 2007. Natural learning in NLDA networks. *Neural Networks* **20**:5, 610-620. [CrossRef]
87. Jani Even, Kenji Sugimoto. 2007. An ICA approach to semi-blind identification of strictly proper systems based on interactor polynomial matrix. *International Journal of Robust and Nonlinear Control* **17**:8, 752-768. [CrossRef]
88. Fuliang Yin, Tiemin Mei, Jun Wang. 2007. Blind-Source Separation Based on Decorrelation and Nonstationarity. *IEEE Transactions on Circuits and Systems I: Regular Papers* **54**:5, 1150-1158. [CrossRef]
89. Evaldo Arajo de Oliveira. 2007. The Rosenblatt Bayesian Algorithm Learning in a Nonstationary Environment. *IEEE Transactions on Neural Networks* **18**:2, 584-588. [CrossRef]
90. N. Hironaga, A.A. Ioannides. 2007. Localization of individual area neuronal activity. *NeuroImage* **34**:4, 1519-1534. [CrossRef]
91. Bin Xia, Liqing Zhang. 2007. Blind Deconvolution in Nonminimum Phase Systems Using Cascade Structure. *EURASIP Journal on Advances in Signal Processing* **2007**, 1-11. [CrossRef]
92. Stefano Squartini, Andrea Arcangeli, Francesco Piazza. 2007. Stability Analysis of Natural Gradient Learning Rules in Complete ICA: A Unifying Perspective. *IEEE Signal Processing Letters* **14**:1, 54-57. [CrossRef]
93. Tadahiro Azetsu, Eiji Uchino, Noriaki Suetake. 2007. Blind Separation and Sound Localization by Using Frequency-domain ICA. *Soft Computing* **11**:2, 185-192. [CrossRef]
94. S CHOI. 2006. Differential learning algorithms for decorrelation and independent component analysis. *Neural Networks* **19**:10, 1558-1567. [CrossRef]

95. H. Sawada, S. Araki, R. Mukai, S. Makino. 2006. Blind Extraction of Dominant Target Sources Using ICA and Time-Frequency Masking. *IEEE Transactions on Audio, Speech and Language Processing* **14**:6, 2165-2173. [CrossRef]

96. T. Mei, J. Xi, F. Yin, A. Mertins, J.F. Chicharo. 2006. Blind Source Separation Based on Time-Domain Optimization of a Frequency-Domain Independence Criterion. *IEEE Transactions on Audio, Speech and Language Processing* **14**:6, 2075-2085. [CrossRef]

97. S JIN, Y KWON, J JEONG, S KWON, D SHIN. 2006. Increased information transmission during scientific hypothesis generation: Mutual information analysis of multichannel EEG. *International Journal of Psychophysiology* **62**:2, 337-344. [CrossRef]

98. Keiji Miura, Masato Okada, Shun-ichi Amari. 2006. Estimating Spiking Irregularities Under Changing Environments. *Neural Computation* **18**:10, 2359-2386. [Abstract] [PDF] [PDF Plus]

99. Jimin Ye, Xianda Zhang, Xiaolong Zhu. 2006. Blind source separation with unknown and dynamically changing number of source signals. *Science in China Series F: Information Sciences* **49**:5, 627-638. [CrossRef]

100. Jayanta Basak. 2006. Online Adaptive Decision Trees: Pattern Classification and Function Approximation. *Neural Computation* **18**:9, 2062-2101. [Abstract] [PDF] [PDF Plus]

101. Shin Ishii, Wako Yoshida. 2006. Part 4: Reinforcement learning: Machine learning and natural learning. *New Generation Computing* **24**:3, 325-350. [CrossRef]

102. Arthur C. Tsai, Michelle Liou, Tzyy-Ping Jung, Julie A. Onton, Philip E. Cheng, Chien-Chih Huang, Jeng-Ren Duann, Scott Makeig. 2006. Mapping single-trial EEG records on the cortical surface through a spatiotemporal modality. *NeuroImage* **32**:1, 195-207. [CrossRef]

103. Makoto Terumitsu, Yukihiko Fujii, Kiyotaka Suzuki, Ingrid L. Kwee, Tsutomu Nakada. 2006. Human primary motor cortex shows hemispheric specialization for speech. *NeuroReport* **17**:11, 1091-1095. [CrossRef]

104. Shun-ichi Amari , Hyeyoung Park , Tomoko Ozeki . 2006. Singularities Affect Dynamics of Learning in Neuromanifolds. *Neural Computation* **18**:5, 1007-1065. [Abstract] [PDF] [PDF Plus]

105. Xiaolong Zhu, Xianda Zhang. 2006. A signal-adaptive algorithm for blind separation of sources with mixed kurtosis signs. *Journal of Electronics (China)* **23**:3, 399-403. [CrossRef]

106. Xiao-Long Zhu , Xian-Da Zhang , Ji-Min Ye . 2006. A Generalized Contrast Function and Stability Analysis for Overdetermined Blind Separation of Instantaneous Mixtures. *Neural Computation* **18**:3, 709-728. [Abstract] [PDF] [PDF Plus]

107. Hyeyong Park. 2006. Part 2: Multilayer perceptron and natural gradient learning. *New Generation Computing* **24**:1, 79-95. [CrossRef]

108. W. Wan. 2006. Implementing Online Natural Gradient Learning: Problems and Solutions. *IEEE Transactions on Neural Networks* **17**:2, 317-329. [CrossRef]

109. Xiao-Long Zhu, Xian-Da Zhang, Zi-Zhe Ding, Ying Jia. 2006. Adaptive nonlinear PCA algorithms for blind source separation without prewhitening. *IEEE Transactions on Circuits and Systems I: Regular Papers* **53**:3, 745-753. [CrossRef]

110. Sheng Wan, Larry E. Banta. 2006. Parameter Incremental Learning Algorithm for Neural Networks. *IEEE Transactions on Neural Networks* **17**:6, 1424-1438. [CrossRef]

111. X. Shen, H. Xu, F. Cong, J. Lei, K. Huang, Y. Zhang, G. Meng. 2006. Blind equalisation algorithm of FIR MIMO system in frequency domain. *IEE Proceedings - Vision, Image, and Signal Processing* **153**:5, 703. [CrossRef]

112. Nuo Zhang, Jianming Lu, Takashi Yahagi. 2006. A Method of Independent Component Analysis Based on Radial Basis Function Networks Using Noise Estimation. *IEEJ Transactions on Electronics, Information and Systems* **126**:6, 780-787. [CrossRef]

113. P. Gao, W.L. Woo, S.S. Dlay. 2006. Non-linear independent component analysis using series reversion and Weierstrass network. *IEE Proceedings - Vision, Image, and Signal Processing* **153**:2, 115. [CrossRef]

114. Wai Yie Leong, John Homer, Danilo P. Mandic. 2006. An Implementation of Nonlinear Multiuser Detection in Rayleigh Fading Channel. *EURASIP Journal on Wireless Communications and Networking* **2006**, 1-10. [CrossRef]

115. Ryo Mukai, Hiroshi Sawada, Shoko Araki, Shoji Makino. 2006. Frequency-Domain Blind Source Separation of Many Speech Signals Using Near-Field and Far-Field Models. *EURASIP Journal on Advances in Signal Processing* **2006**, 1-14. [CrossRef]

116. W.Y. Leong, J. Homer. 2006. Blind multiuser receiver for DS-CDMA wireless system. *IEE Proceedings - Communications* **153**:5, 733. [CrossRef]

117. P. Gao, W.L. Woo, S.S. Dlay. 2006. Weierstrass approach to blind source separation of multiple nonlinearly mixed signals. *IEE Proceedings - Circuits, Devices and Systems* **153**:4, 332. [CrossRef]

118. A ELEUTERI, R TAGLIAFERRI, L MILANO. 2005. A novel information geometric approach to variable selection in MLP networks. *Neural Networks* **18**:10, 1309-1318. [CrossRef]

119. C.-T. Lin, W.-C. Cheng, S.-F. Liang. 2005. A 3-D Surface Reconstruction Approach Based on Postnonlinear ICA Model. *IEEE Transactions on Neural Networks* **16**:6, 1638-1650. [CrossRef]

120. S. Fiori. 2005. Formulation and Integration of Learning Differential Equations on the Stiefel Manifold. *IEEE Transactions on Neural Networks* **16**:6, 1697-1701. [CrossRef]

121. Hirokazu Asano, Hiroya Nakao. 2005. Independent Component Analysis of Spatiotemporal Chaos. *Journal of the Physics Society Japan* **74**:6, 1661-1665. [CrossRef]

122. Shun-ichi Amari , Hiroyuki Nakahara . 2005. Difficulty of Singularity in Population Coding. *Neural Computation* **17**:4, 839-858. [Abstract] [PDF] [PDF Plus]

123. T. Tanaka. 2005. Generalized weighted rules for principal components tracking. *IEEE Transactions on Signal Processing* **53**:4, 1243-1253. [CrossRef]

124. D. Erdogmus, O. Fontenla-Romero, J.C. Principe, A. Alonso-Betanzos, E. Castillo. 2005. Linear-Least-Squares Initialization of Multilayer Perceptrons Through Backpropagation of the Desired Response. *IEEE Transactions on Neural Networks* **16**:2, 325-337. [CrossRef]

125. Kun Zhang , Lai-Wan Chan . 2005. Extended Gaussianization Method for Blind Separation of Post-Nonlinear Mixtures. *Neural Computation* **17**:2, 425-452. [Abstract] [PDF] [PDF Plus]

126. C. Xiang, S. Ding, T.H. Lee. 2005. Geometrical Interpretation and Architecture Selection of MLP. *IEEE Transactions on Neural Networks* **16**:1, 84-96. [CrossRef]

127. Eiji Uchino, Noriaki Suetake, Morihiko Sakano. 2005. Blind deconvolution by using phase spectral constraints and natural gradient. *IEICE Electronics Express* **2**:9, 316-320. [CrossRef]

128. A.J. Caamano, R. Boloix-Tortosa, J. Ramos, J.J. Murillo-Fuentes. 2004. Hybrid Higher-Order Statistics Learning in Multiuser Detection. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)* **34**:4, 417-424. [CrossRef]

129. N. Bouguila, D. Ziou, J. Vaillancourt. 2004. Unsupervised Learning of a Finite Mixture Model Based on the Dirichlet Distribution and Its Application. *IEEE Transactions on Image Processing* **13**:11, 1533-1543. [CrossRef]

130. M SATO, T YOSHIOKA, S KAJIHARA, K TOYAMA, N GODA, K DOYA, M KAWATO. 2004. Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage* **23**:3, 806-826. [CrossRef]

131. J PELTONEN, A KLAMI, S KASKI. 2004. Improved learning of Riemannian metrics for exploratory analysis. *Neural Networks* **17**:8-9, 1087-1100. [CrossRef]

132. D.T. Pham. 2004. Fast Algorithms for Mutual Information Based Independent Component Analysis. *IEEE Transactions on Signal Processing* **52**:10, 2690-2700. [CrossRef]

133. Jayanta Basak . 2004. Online Adaptive Decision Trees. *Neural Computation* **16**:9, 1959-1981. [Abstract] [PDF] [PDF Plus]

134. Shiro Ikeda , Toshiyuki Tanaka , Shun-ichi Amari . 2004. Stochastic Reasoning, Free Energy, and Information Geometry. *Neural Computation* **16**:9, 1779-1810. [Abstract] [PDF] [PDF Plus]

135. H. Sawada, R. Mukai, S. Araki, S. Makino. 2004. A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation. *IEEE Transactions on Speech and Audio Processing* **12**:5, 530-538. [CrossRef]

136. Ji-Min Ye, Xiao-Long Zhu, Xian-Da Zhang. 2004. Adaptive Blind Separation with an Unknown Number of Sources. *Neural Computation* **16**:8, 1641-1660. [Abstract] [PDF] [PDF Plus]

137. M. Welling, R.S. Zemel, G.E. Hinton. 2004. Probabilistic Sequential Independent Components Analysis. *IEEE Transactions on Neural Networks* **15**:4, 838-849. [CrossRef]

138. S.A. Cruces-Alvarez, A. Cichocki, S. Amari. 2004. From Blind Signal Extraction to Blind Instantaneous Signal Separation: Criteria, Algorithms, and Stability. *IEEE Transactions on Neural Networks* **15**:4, 859-873. [CrossRef]

139. H CHEN, D YAO. 2004. Discussion on the choice of separated components in fMRI data analysis by spatial independent component analysis. *Magnetic Resonance Imaging* **22**:6, 827-833. [CrossRef]

140. N. Xu, X. Gao, B. Hong, X. Miao, S. Gao, F. Yang. 2004. BCI Competition 2003&#8212;Data Set IIb: Enhancing P300 Wave Detection Using ICA-Based Subspace Projections for BCI Applications. *IEEE Transactions on Biomedical Engineering* **51**:6, 1067-1072. [CrossRef]

141. Masato Inoue, Hyeyoung Park, Masato Okada. 2004. Dynamics of the adaptive natural gradient descent method for soft committee machines. *Physical Review E* **69**:5. . [CrossRef]

142. Xiuwen Liu, A. Srivastava, K. Gallivan. 2004. Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**:5, 662-666. [CrossRef]

143. L. Zhang, A. Cichocki, S. Amari. 2004. Multichannel Blind Deconvolution of Nonminimum-Phase Systems Using Filter Decomposition. *IEEE Transactions on Signal Processing* **52**:5, 1430-1442. [CrossRef]

144. S Makeig. 2004. Mining event-related brain dynamics. *Trends in Cognitive Sciences* **8**:5, 204-210. [CrossRef]

145. L. Zhang, A. Cichocki, S. Amari. 2004. Self-Adaptive Blind Source Separation Based on Activation Functions Adaptation. *IEEE Transactions on Neural Networks* **15**:2, 233-244. [CrossRef]

146. A Ossadtchi. 2004. Automated interictal spike detection and source localization in magnetoencephalography using independent components analysis and spatio-temporal clustering. *Clinical Neurophysiology* **115**:3, 508-522. [CrossRef]

147. Hyeyoung Park , Noboru Murata , Shun-ichi Amari . 2004. Improving Generalization Performance of Natural Gradient Learning Using Optimized Regularization by NIC. *Neural Computation* **16**:2, 355-382. [Abstract] [PDF] [PDF Plus]

148. Scott Makeig, Arnaud Delorme, Marissa Westerfield, Tzyy-Ping Jung, Jeanne Townsend, Eric Courchesne, Terrence J. Sejnowski. 2004. Electroencephalographic Brain Dynamics Following Manually Responded Visual Targets. *PLoS Biology* **2**:6, e176. [CrossRef]

149. B Lu. 2003. Converting general nonlinear programming problems into separable programming problems with feedforward neural networks. *Neural Networks* **16**:7, 1059-1074. [CrossRef]

150. E Mizutani. 2003. On structure-exploiting trust-region regularized nonlinear least squares algorithms for neural-network learning. *Neural Networks* **16**:5-6, 745-753. [CrossRef]

151. Shun-ichi Amari, Tomoko Ozeki, Hyeyoung Park. 2003. Learning and inference in hierarchical models with singularities. *Systems and Computers in Japan* **34**:7, 34-42. [CrossRef]

152. Shun-Tian Lou, Xian-Da Zhang. 2003. Fuzzy-based learning rate determination for blind source separation. *IEEE Transactions on Fuzzy Systems* **11**:3, 375-383. [CrossRef]

153. M.D. Plumbley. 2003. Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks* **14**:3, 534-543. [CrossRef]

154. Jianting Cao, N. Murata, S.-i. Amari, A. Cichocki, T. Takeda. 2003. A robust approach to independent component analysis of signals with high-level noise measurements. *IEEE Transactions on Neural Networks* **14**:3, 631-645. [CrossRef]

155. S Fiori. 2003. Overview of independent component analysis technique with an application to synthetic aperture radar (SAR) imagery processing. *Neural Networks* **16**:3-4, 453-467. [CrossRef]

156. Masato Inoue, Hyeyoung Park, Masato Okada. 2003. On-Line Learning Theory of Soft Committee Machines with Correlated Hidden Units -- Steepest Gradient Descent and Natural Gradient Descent --. *Journal of the Physics Society Japan* **72**:4, 805-810. [CrossRef]

157. A. Bortoletti, C. Di Fiore, S. Fanelli, P. Zellini. 2003. A new class of quasi-newtonian methods for optimal learning in mlp-networks. *IEEE Transactions on Neural Networks* **14**:2, 263-273. [CrossRef]

158. S. Araki, R. Mukai, S. Makino, T. Nishikawa, H. Saruwatari. 2003. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing* **11**:2, 109-116. [CrossRef]

159. Yoshio Onozaka, Masahiro Nakagawa. 2003. Back propagation learning with periodic chaos neurons. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* **86**:3, 11-19. [CrossRef]

160. Kiyotaka SUZUKI, Hitoshi MATSUZAWA, Hironaka IGARASHI, Masaki WATANABE, Naoki NAKAYAMA, Ingrid L. KWEE, Tsutomu NAKADA. 2003. All-phase MR Angiography Using Independent Component Analysis of Dynamic Contrast Enhanced MRI Time Series: .PHI.-MRA. *Magnetic Resonance in Medical Sciences* **2**:1, 23-27. [CrossRef]

161. Tadej Kosel, Igor Grabec, Franc Kosel. 2003. Intelligent location of two simultaneously active acoustic emission sources: Part II. *Aircraft Engineering and Aerospace Technology* **75**:2, 137-142. [CrossRef]

162. Peixun Luo, K. Michael Wong. 2003. Dynamical and stationary properties of on-line learning from finite training sets. *Physical Review E* **67**:1. . [CrossRef]

163. H.S. Sahambi, K. Khorasani. 2003. A neural-network appearance-based 3-D object recognition using independent component analysis. *IEEE Transactions on Neural Networks* **14**:1, 138-149. [CrossRef]

164. Nihat Ay . 2002. Locality of Global Stochastic Interaction in Directed Acyclic Networks. *Neural Computation* **14**:12, 2959-2980. [Abstract] [PDF] [PDF Plus]

165. Xiao-Long Zhu, Xian-Da Zhang. 2002. Adaptive RLS algorithm for blind source separation using a natural gradient. *IEEE Signal Processing Letters* **9**:12, 432-435. [CrossRef]

166. M. Ibnkahla. 2002. Natural gradient learning neural networks for adaptive inversion of Hammerstein systems. *IEEE Signal Processing Letters* **9**:10, 315-317. [CrossRef]

167. H. Abdulkader, F. Langlet, D. Roviras, F. Castanie. 2002. Natural gradient algorithm for neural networks applied to non-linear high power amplifiers. *International Journal of Adaptive Control and Signal Processing* **16**:8, 557-576. [CrossRef]

168. Katsuyuki Hagiwara . 2002. On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario. *Neural Computation* **14**:8, 1979-2002. [Abstract] [PDF] [PDF Plus]

169. Minami Mihoko , Shinto Eguchi . 2002. Robust Blind Source Separation by Beta Divergence. *Neural Computation* **14**:8, 1859-1886. [Abstract] [PDF] [PDF Plus]

170. A. Taleb. 2002. A generic framework for blind source separation in structured nonlinear models. *IEEE Transactions on Signal Processing* **50**:8, 1819-1830. [CrossRef]

171. R.K. Martin, W.A. Sethares, R.C. Williamson, C.R. Johnson. 2002. Exploiting sparsity in adaptive filters. *IEEE Transactions on Signal Processing* **50**:8, 1883-1894. [CrossRef]

172. Nicol N. Schraudolph . 2002. Fast Curvature Matrix-Vector Products for Second-Order Gradient Descent. *Neural Computation* **14**:7, 1723-1738. [Abstract] [PDF] [PDF Plus]

173. N Murata. 2002. On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks* **15**:4-6, 743-760. [CrossRef]

174. S. Fiori. 2002. A theory for learning based on rigid bodies dynamics. *IEEE Transactions on Neural Networks* **13**:3, 521-531. [CrossRef]

175. Tadej Kosel, Igor Grabec. 2002. Location of two simultaneously active continuous acoustic emission sources on an aluminum beam. *Aircraft Engineering and Aerospace Technology* **74**:1, 4-8. [CrossRef]

176. Kiyotaka Suzuki, Tohru Kiryu, Tsutomu Nakada. 2002. Fast and precise independent component analysis for high field fMRI time series tailored using prior information on spatiotemporal structure. *Human Brain Mapping* **15**:1, 54-66. [CrossRef]

177. W.L. Woo, S. Sali. 2002. General multilayer perceptron demixer scheme for nonlinear blind signal separation. *IEE Proceedings - Vision, Image, and Signal Processing* **149**:5, 253. [CrossRef]

178. S Fiori. 2002. Hybrid independent component analysis by adaptive LUT activation function neurons. *Neural Networks* **15**:1, 85-94. [CrossRef]

179. S Choi. 2002. Equivariant nonstationary source separation. *Neural Networks* **15**:1, 121-130. [CrossRef]

180. T Chen. 2001. Unified stabilization approach to principal and minor components extraction algorithms. *Neural Networks* **14**:10, 1377-1387. [CrossRef]

181. Mark Zlochin , Yoram Baram . 2001. Manifold Stochastic Dynamics for Bayesian Learning. *Neural Computation* **13**:11, 2549-2572. [Abstract] [PDF] [PDF Plus]

182. Shotaro Akaho, Shinji Umeyama. 2001. Multimodal independent component analysis?A method of feature extraction from multiple information sources. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* **84**:11, 21-28. [CrossRef]

183. N Ampazis. 2001. A dynamical model for the analysis and acceleration of learning in feedforward networks. *Neural Networks* **14**:8, 1075-1088. [CrossRef]

184. Dinh-Tuan Pham, J.-F. Cardoso. 2001. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing* **49**:9, 1837-1848. [CrossRef]

185. Simone Fiori . 2001. A Theory for Learning by Weight Flow on Stiefel-Grassman Manifold. *Neural Computation* **13**:7, 1625-1647. [Abstract] [PDF] [PDF Plus]

186. Masa-aki Sato . 2001. Online Model Selection Based on the Variational Bayes. *Neural Computation* **13**:7, 1649-1681. [Abstract] [PDF] [PDF Plus]

187. S.-I. Amari. 2001. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory* **47**:5, 1701-1711. [CrossRef]

188. S. Kaski, J. Sinkkonen, J. Peltonen. 2001. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks* **12**:4, 936-947. [CrossRef]

189. T.-P. Jung, S. Makeig, M.J. McKeown, A.J. Bell, T.-W. Lee, T.J. Sejnowski. 2001. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE* **89**:7, 1107-1122. [CrossRef]

190. Thomas Wachtler, Te-Won Lee, Terrence J. Sejnowski. 2001. Chromatic structure of natural scenes. *Journal of the Optical Society of America A* **18**:1, 65. [CrossRef]

191. M Sugiyama. 2001. Properties of incremental projection learning. *Neural Networks* **14**:1, 67-78. [CrossRef]

192. M Sugiyama. 2001. Incremental projection learning for optimal generalization. *Neural Networks* **14**:1, 53-66. [CrossRef]

193. L. Castedo-Ribas, A. Cichocki, S. Cruces-Alvarez. 2000. An iterative inversion approach to blind source separation. *IEEE Transactions on Neural Networks* **11**:6, 1423-1437. [CrossRef]

194. Te-Won Lee, M.S. Lewicki, T.J. Sejnowski. 2000. ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**:10, 1078-1089. [CrossRef]

195. Shun-ichi Amari . 2000. Estimating Functions of Independent Component Analysis for Temporally Correlated Signals. *Neural Computation* **12**:9, 2083-2107. [Abstract] [PDF] [PDF Plus]

196. H Park. 2000. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks* **13**:7, 755-764. [CrossRef]

197. T Nakada. 2000. Independent component-cross correlation-sequential epoch (ICS) analysis of high field fMRI time series: direct visualization of dual representation of the primary motor cortex in human. *Neuroscience Research* **37**:3, 237-244. [CrossRef]

198. Shun-ichi Amari , Tian-Ping Chen , Andrzej Cichocki . 2000. Nonholonomic Orthogonal Learning Algorithms for Blind Source Separation. *Neural Computation* **12**:6, 1463-1484. [Abstract] [PDF] [PDF Plus]

199. Shun-ichi Amari , Hyeyoung Park , Kenji Fukumizu . 2000. Adaptive Method of Realizing Natural Gradient Learning for Multilayer Perceptrons. *Neural Computation* **12**:6, 1399-1409. [Abstract] [PDF] [PDF Plus]

200. A. Navia-Vázquez , A. R. Figueiras-Vidal . 2000. Efficient Block Training of Multilayer Perceptrons. *Neural Computation* **12**:6, 1429-1447. [Abstract] [PDF] [PDF Plus]

201. Tom Heskes . 2000. On “Natural” Learning and Pruning in Multilayered Perceptrons. *Neural Computation* **12**:4, 881-901. [Abstract] [PDF] [PDF Plus]

202. K Fukumizu. 2000. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks* **13**:3, 317-327. [CrossRef]

203. J.J. Murillo-Fuentes, F.J. Gonza#lez-Serrano. 2000. Improving stability in blind source separation with stochastic median gradient. *Electronics Letters* **36**:19, 1662. [CrossRef]

204. Dominik Endres, Peter Riegler. 1999. *Journal of Physics A: Mathematical and General* **32**:49, 8655-8663. [CrossRef]

205. Shun-ichi Amari . 1999. Natural Gradient Learning for Over- and Under-Complete Bases in ICA. *Neural Computation* **11**:8, 1875-1883. [Abstract] [PDF] [PDF Plus]

206. L.-Q. Zhang, A. Cichocki, S. Amari. 1999. Natural gradient algorithm for blind separation of overdetermined mixture with additive noise. *IEEE Signal Processing Letters* **6**:11, 293-295. [CrossRef]

207. A. Taleb, C. Jutten. 1999. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing* **47**:10, 2807-2820. [CrossRef]

208. J. Basak, S. Amari. 1999. Blind separation of uniformly distributed signals: a general approach. *IEEE Transactions on Neural Networks* **10**:5, 1173-1185. [CrossRef]

209. H.H. Yang. 1999. Serial updating rule for blind separation derived from the method of scoring. *IEEE Transactions on Signal Processing* **47**:8, 2279-2285. [CrossRef]

210. Bruno Apolloni, Egidio Battistini, Diego de Falco. 1999. *Journal of Physics A: Mathematical and General* **32**:30, 5529-5538. [CrossRef]

211. Silvia Scarpetta, Magnus Rattray, David Saad. 1999. *Journal of Physics A: Mathematical and General* **32**:22, 4047-4059. [CrossRef]

212. F AIRES, A CHEDIN, J NADAL. 1999. Analyse de séries temporelles géophysiques et théorie de l'information: L'analyse en composantes indépendantes. *Comptes Rendus de l'Académie des Sciences - Series IIA - Earth and Planetary Science* **328**:9, 569-575. [CrossRef]

213. Jayanta Basak , Shun-ichi Amari . 1999. Blind Separation of a Mixture of Uniformly Distributed Source Signals: A Novel Approach. *Neural Computation* **11**:4, 1011-1034. [Abstract] [PDF] [PDF Plus]

214. Piërre van de Laar , Tom Heskes . 1999. Pruning Using Parameter and Neuronal Metrics. *Neural Computation* **11**:4, 977-993. [Abstract] [PDF] [PDF Plus]

215. S.C. Douglas. 1999. Equivariant adaptive selective transmission. *IEEE Transactions on Signal Processing* **47**:5, 1223-1231. [CrossRef]

216. Peter Dayan . 1999. Recurrent Sampling Models for the Helmholtz Machine. *Neural Computation* **11**:3, 653-677. [Abstract] [PDF] [PDF Plus]

217. Magnus Rattray, David Saad. 1999. Analysis of natural gradient descent for multilayer neural networks. *Physical Review E* **59**:4, 4523-4532. [CrossRef]

218. S. Amari. 1999. Superefficiency in blind source separation. *IEEE Transactions on Signal Processing* **47**:4, 936-944. [CrossRef]

219. Te-Won Lee, M.S. Lewicki, M. Girolami, T.J. Sejnowski. 1999. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters* **6**:4, 87-90. [CrossRef]

220. Te-Won Lee , Mark Girolami , Terrence J. Sejnowski . 1999. Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources. *Neural Computation* **11**:2, 417-441. [Abstract] [PDF] [PDF Plus]

221. Magnus Rattray, David Saad, Shun-ichi Amari. 1998. Natural Gradient Descent for On-Line Learning. *Physical Review Letters* **81**:24, 5461-5464. [CrossRef]

222. Howard Hua Yang , Shun-ichi Amari . 1998. Complexity Issues in Natural Gradient Descent Method for Training Multilayer Perceptrons. *Neural Computation* **10**:8, 2137-2157. [Abstract] [PDF] [PDF Plus]

223. Mark Girolami . 1998. An Alternative Perspective on Adaptive Independent Component Analysis Algorithms. *Neural Computation* **10**:8, 2103-2114. [Abstract] [PDF] [PDF Plus]

224. Magnus Rattray, David Saad. 1998. Analysis of on-line training with optimal learning rates. *Physical Review E* **58**:5, 6379-6391. [CrossRef]

225. S. Amari, A. Cichocki. 1998. Adaptive blind signal processing-neural network approaches. *Proceedings of the IEEE* **86**:10, 2026-2048. [CrossRef]

226. J.-F. Cardoso. 1998. Blind signal separation: statistical principles. *Proceedings of the IEEE* **86**:10, 2009-2025. [CrossRef]

227. Richard Hahnloser. 1998. Learning algorithms based on linearization. *Network: Computation in Neural Systems* **9**:3, 363-380. [CrossRef]

228. Siegfried Bös. 1998. *Journal of Physics A: Mathematical and General* **31**:22, L413-L417. [CrossRef]

229. T Chen. 1998. A unified algorithm for principal and minor components extraction. *Neural Networks* **11**:3, 385-390. [CrossRef]

230. Csaba SzepesváriReinforcement Learning Algorithms for MDPs . [CrossRef]