

GEOMETRY OF LEARNING IN MULTILAYER PERCEPTRONS

Shun-ichi Amari, Hyeyoung Park and Tomoko Ozeki

Key words: Learning, information geometry, singular statistical model, neural networks.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: Neural networks provide a good model of learning from statistical data. Multilayer perceptron is regarded as a statistical model in which a nonlinear input-output relation is realized. The set of multilayer perceptrons forms a statistical manifold in which learning and estimation takes place. This is a Riemannian manifold with Fisher information metric. However, such a hierarchical model includes algebraic singularities at which the Fisher information matrix degenerates. This causes various difficulties in learning and statistical estimation. The present paper elucidates the structure of singularities, and how they influence the behavior of learning. The paper describes a new learning algorithm, named the natural gradient method, to overcome such difficulties. Various statistical problems in singular models are discussed, and the models selection criteria (AIC and MDL) are studied in this framework.

1 Introduction

The multilayer perceptron is a simple feedforward model of neural networks, which transforms input signals to output signals nonlinearly. It is a universal approximator in the sense that any nonlinear transformation is approximated sufficiently well by an adequate perceptron, if the number of hidden units is large.

In order to realize a good approximator, examples of input-output pairs are used. On-line learning receives a series of training examples one by one, and modifies the parameters of a perceptron each time when one example is given. Usually old examples are then discarded. Batch learning keeps all the examples and modifies the parameters in a batch mode.

A multilayer perceptron is an old model of learning machines, and the error-correcting learning algorithm was established for simple perceptrons in the sixties. Amari proposed a gradient descent learning method for multilayer perceptrons [2], which was rediscovered later independently and became popular under the name of backpropagation [23].

We study the set of multilayer perceptrons of a fixed architecture, which include a number of modifiable parameters called connection weights and biases. The set forms a multi-dimensional manifold, where all these parameters play a role of admissible coordinate systems. Learning takes place in the manifold, drawing a trajectory.

It is important to study the geometrical structure of the manifold which we call a neuromanifold. We will show by statistical considerations that the neuromanifold is Riemannian whose metric is specified by the Fisher information matrix [3]. Moreover, it has a pair of affine connections [4], but we do not state them in the present paper. The neuromanifold has singularities where the Fisher information (or the Riemannian metric) degenerates [5]. This is an interesting statistical model, because the conventional Cramér-Rao paradigm excludes such a model, assuming the existence and non-degeneracy of the Fisher information matrix as regularity conditions.

It is known that the convergence speed of a multilayer perceptron is usually very slow. This is caused by the Riemannian character in particular by its degeneracy, because the conventional backprop learning method does not take the Riemannian nature into account. The state of a network is often attracted by singularities by the conventional algorithm and takes long time before getting rid of them. The natural gradient learning algorithm was proposed to overcome the flaw, which takes the Riemannian gradient instead of the conventional gradient [3]. We show in the present paper the reasons why it works so well. We also explain an adaptive method of implementing the natural gradient [8]. In the case of the squared error criterion under Gaussian noises, the natural gradient algorithm coincides with the adaptive version of the Gauss-Newton method, but they differ in more general models (see [17]).

We finally study the dynamics of learning and the nature of singularities and explain the reason why learning trajectories are attracted to and stay longer in a neighborhood of singularities. The statistical analysis of behaviors of estimators in a neighborhood of singularities is another important problem to be studied. We show the conventional criteria of model selection such as AIC and MDL fail in this case.

2 Neuromanifold of multilayer perceptrons

Let us consider a multilayer perceptron of h hidden units and one output unit, which receives n-dimensional input signals $\boldsymbol{x} = (x_1, \cdots, x_n)$. A hidden unit, say the i-th unit receives $\boldsymbol{x}$, and takes its weighted sum, resulting in the potential

$$u_i = \boldsymbol{w}_i \cdot \boldsymbol{x}. \tag{1}$$

Here $\boldsymbol{w}_i = (w_{i1}, \cdots, w_{in})$ is the weight vector of the i-th unit, and we neglect the bias term for the sake of simplifying the notation. The unit calculates the nonlinear transform of the potential, $\varphi(u)$, where the nonlinear function is

$$\varphi(u) = \tanh u. \tag{2}$$

The final units collects all the outputs of the hidden units, and its final output is their weighted sum, if no noise intervenes. We put

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum v_i \varphi(u_i) = \sum v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}), \tag{3}$$

where we summarized all the modifiable parameters in a large vector $\boldsymbol{\theta} = (v_1, \cdots, v_m\ ;\ \boldsymbol{w}_1, \cdots, \boldsymbol{w}_m)$. The final output of the perceptron is disturbed by noise, so that

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \varepsilon, \tag{4}$$

where we assume that ε is a Gaussian noise with mean 0 and variance 1. Therefore, its behavior is represented by the conditional probability of y given $\boldsymbol{x}$,

$$p\left(y|\boldsymbol{x}, \boldsymbol{\theta}\right) = c \exp\left\{-\frac{1}{2}\left(y - f(\boldsymbol{x}, \boldsymbol{\theta})\right)^2\right\} \tag{5}$$

or the joint probability distribution of $(\boldsymbol{x}, y)$,

$$p(y, \boldsymbol{x}; \boldsymbol{\theta}) = q(\boldsymbol{x}) p\left(y|\boldsymbol{x}, \boldsymbol{\theta}\right) \tag{6}$$

where $q(\boldsymbol{x})$ is the probability distribution of inputs $\boldsymbol{x}$. The set of all the perceptrons is a manifold called a neuromanifold M where $\boldsymbol{\theta}$ plays the role of the coordinate system. Each point of the neuromanifold corresponds to the probability distribution (5) or (6).

3 Fisher information matrix and the Riemannian metric

The Fisher information matrix G is given by

$$G(\boldsymbol{\theta}) = E\left[\nabla \log p(y, \boldsymbol{x}; \boldsymbol{\theta}) \nabla \log p(y, \boldsymbol{x}; \boldsymbol{\theta})^T\right] \tag{7}$$

which is further calculated as

$$G(\boldsymbol{\theta}) = E\left[\nabla f(\boldsymbol{x}, \boldsymbol{\theta}) \nabla f(\boldsymbol{x}, \boldsymbol{\theta})^T\right], \tag{8}$$

where E denotes expectation, $\nabla = (\partial/\partial\theta_i)$ is the gradient and T denotes transpose of a vector. Let us define the square of the distance between two nearby perceptrons whose parameters are $\boldsymbol{\theta}$ and $\boldsymbol{\theta}+d\boldsymbol{\theta}$. Information geometry gives the squared distance by the quadratic form

$$ds^2 = d\boldsymbol{\theta}^T G(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{9}$$

This is the Riemannian metric, where the Fisher information metric is used as the Riemannian metric tensor [19]. This is the only invariant metric to be introduced in the manifold of probability distributions.

Given a (large) number N of independently generated input-output pairs $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$, the maximum likelihood estimator (or any other first order efficient estimator) satisfies the Cramér-Rao bound. Hence, the distance is large when two perceptrons are well separated in the sense that their estimation can be done precisely. However, different from the ordinary statistical model, the neuromanifold includes points at which the Fisher information degenerates and its inverse diverges. This is related to the unidentifiability of network parameters.

4 Identifiability of perceptrons and singularity

The behavior of a perceptron is invariant under the following two operations [10]:

1. Change of signs of v_i and $\boldsymbol{w}_i$ at the same time.
2. Permutation of the hidden units, which causes permutation of the weight vectors $\{\boldsymbol{w}_i\}$ and the output weight $\{v_i\}$ at the same time.

This causes the following unidentifiability:

1. When $v_i = 0$ or $\boldsymbol{w}_i = 0$, the behavior is the same whatever value $\boldsymbol{w}_i$ or v_i takes.
2. When $\boldsymbol{w}_i = \boldsymbol{w}_j$ (or $\boldsymbol{w}_i = -\boldsymbol{w}_j$), the behavior is the same when

$$v_i + v_j = v_i' + v_j' \qquad \left(v_i - v_j = v_i' - v_j'\right) \tag{10}$$

holds for two perceptrons $\{\boldsymbol{w}_i, v_i\}$ and $\{\boldsymbol{w}_i, v_i'\}$.

We call the set

$$C = \{\boldsymbol{\theta} |\ v_i\, |\boldsymbol{w}_i| = 0 \quad \text{or} \quad \boldsymbol{w}_i = \pm \boldsymbol{w}_j\} \tag{11}$$

the critical set on which unidentifiability takes place. The Fisher information degenerates on the critical set, because the unidentifiability implies that the estimation error does not converge to 0 even when N goes to infinity. Hence the statistical model is non-regular, and the Riemannian metric is singular. See also [7], [8], [9], [15]; [24], [27].

Let us introduce the equivalence relation $\approx$, by which two perceptrons with different parameters are equivalent when their input-output behaviors are the same. Then the set

$$\tilde{M} = M/\approx \tag{12}$$

includes algebraic singularities and dimensions are reduced on the critical set. The conventional theory of statistical estimation does not hold in a neighborhood of singularities.

5 Natural gradient learning algorithm

Let

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\} \tag{13}$$

be the set of input-output examples, which we call the training set. Here we assume that the examples are generated independently by using the true perceptron whose parameters are given by $\boldsymbol{\theta}_0$.

Given the training set, we want to obtain the estimated parameters $\hat{\boldsymbol{\theta}}$ which is closest to the true one. The performance of the estimator $\hat{\boldsymbol{\theta}}$ is

measured by the generalization error, which is the expectation of the squared error for a new example $(\boldsymbol{x}, y)$,

$$L\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{2}E\left[\left\{y - f\left(\boldsymbol{x}, \hat{\boldsymbol{\theta}}\right)\right\}^2\right]. \tag{14}$$

The conventional on-line learning algorithm uses the gradient of the instantaneous error at time t,

$$\nabla e\left(\boldsymbol{x}_t, y_t, \hat{\boldsymbol{\theta}}_t\right) = \frac{1}{2}\nabla\left\{y_t - f\left(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_t\right)\right\}^2 \tag{15}$$

to update the current parameters $\hat{\boldsymbol{\theta}}_t$ to the new one,

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - c\nabla e. \tag{16}$$

The gradient of a function is believed to be the steepest direction of change. This is true only when the coordinate system $\boldsymbol{\theta}$ is orthonormal in a Euclidean space. The steepest direction of e is given by

$$\tilde{\nabla} e = G^{-1}\nabla e \tag{17}$$

in a Riemannian space, where G^{-1} is the inverse of the Riemannian metric matrix.

Amari [3] proposed to use the Riemannian gradient for learning,

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - cG^{-1}\nabla e, \tag{18}$$

which is called the natural gradient method. The natural gradient method is proved to give an Fisher efficient estimator, even though examples are used only once when they are observed, and then discarded.

The performance of the natural gradient method is largely different from the conventional method, when the Riemannian structure is very different from the Euclidean one. It will be seen that this is indeed the case with multilayer perceptrons, because they include singularities where the Riemannian metric degenerates.

It is known that the learning trajectory is often trapped in the so called plateaus, at which the parameters change so slowly, and it takes long time to get rid of. The statistical physical approach made it clear that the parameters are once attracted to the critical set of the neuromanifold, so that the set becomes plateaus of learning [25], [21], [18]. Rattray, Saad and Amari [20] analyzed the dynamics of the natural gradient learning method, and showed that it has an idealistic characteristic for avoiding plateaus. See also [14].

6 Implementation of natural gradient—Adaptive natural gradient method

In order to implement the natural gradient method, one needs to use the inverse G^{-1} of the Fisher information matrix. However, it is in general difficult

to calculate the Fisher information matrix, because it uses the expectation with respect to the unknown distribution $q(\boldsymbol{x})$ of inputs. Moreover, it is computationally heavy to invert the matrix G when the number of parameters is large.

Amari, Park and Fukumizu [8] proposed an adaptive method to obtain an estimate of the inverse of the Fisher matrix. It is an iterative method, and the estimate $\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right)$ is calculated by

$$\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right) = (1+c')\,\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_{t-1}\right) - c'\tilde{\nabla} f_t \left(\tilde{\nabla} f_t\right)^T, \tag{19}$$

where $f_t = f(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$ and c' is another learning constant which may depend on t. One should choose c and c' carefully. By using this estimate $\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right)$, we can obtain the update rule of the adaptive natural gradient method of the form,

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - c\,\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right)\nabla e, \tag{20}$$

Park, Amari and Fukumizu [17] generalized the idea to be applicable to more general cost functions.

7 Dynamics of learning in the neighborhood of the critical set

In order to see the dynamics of learning, let us consider the special case of perceptrons consisting of two hidden units. Let us consider the set $Q(\boldsymbol{w}, v)$

$$Q(\boldsymbol{w}, v) = \{\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}, v_1 + v_2 = v\} \tag{21}$$

which is a part of the critical set. This corresponds to the set of all the perceptrons which have only one hidden unit, where the weight vector is $\boldsymbol{w}$ and the output weight is v. Let the true parameters be $\boldsymbol{\theta}_0 = \{\boldsymbol{w}_1, \boldsymbol{w}_2, v_1, v_2\}$, where $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$ so that it needs two hidden units.

Let $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{w}}, \bar{v})$ be the best perceptron with one hidden unit that approximates the input-output function $f(\boldsymbol{x}, \boldsymbol{\theta}_0)$ of the true perceptron. Then, all the perceptrons of two hidden units on the line:

$$\boldsymbol{w}_1 = \boldsymbol{w}_2 = \bar{\boldsymbol{w}}, \quad v_1 + v_2 = \bar{v} \tag{22}$$

corresponds to the best approximation by one hidden unit perceptron. Let us transform the two weights as

$$\boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2), \quad \boldsymbol{u} = \frac{1}{2}(\boldsymbol{w}_1 - \boldsymbol{w}_2). \tag{23}$$

Then, the derivative of $L(\boldsymbol{\theta})$ along the line is 0, because all the perceptrons are equivalent along the line. The derivative in the direction of changing $\bar{\boldsymbol{w}}$

and $\bar{v}$ are zero, because they are the best approximator. The derivative in the direction of $\boldsymbol{u}$ is again 0, because the perceptrons having $\boldsymbol{u}$ is equivalent to that having $-\boldsymbol{u}$ that is derived by changing the two hidden units. Hence the line forms critical points of the cost function. This implies that it is very difficult to get rid of it once the parameters are attracted to $Q(\bar{\boldsymbol{w}}, \bar{v})$.

Fukumizu and Amari [12] calculated the Hessian of L. When it is positive definite, the line is really attracting. When it includes the negative eigenvalues, the state is escaping in these directions eventually. They showed that, in some cases, a part of the line is really attracting in some region, while it is really a saddle having directions of escape (although the derivative is 0). In such a case, the perceptron is once truly attracted to the line, and stays inside the line fluctuating around it because of random noise until it finds the place from which it can escape from the line. This is clearly a plateau.

This explains the plateau phenomenon. In order to show why the natural gradient works well, we need to evaluate the natural gradient in the neighborhood of the critical points. We can then prove that the natural gradient has a large magnitude in the neighborhood of the critical set, so that the plateau phenomena will disappear. Computer simulations confirm this observation.

8 Estimation and testing in the neighborhood of the critical set

The Fisher information matrix degenerates on the critical set. Therefore, the Cramér-Rao paradigm cannot be valid in the neighborhood of the critical set. Let us consider the statistical test

$$H_0 \ : \ \boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{24}$$

against

$$H_1 \ : \ \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \tag{25}$$

The likelihood ratio statistics is given by

$$\lambda = \frac{1}{2} \log \frac{\sum p\left(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)}{\sum p\left(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0\right)}, \tag{26}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator. When the true point $\boldsymbol{\theta}_0$ is a regular point, that is, it is not in the critical region, the mle (maximum likelihood estimator) is asymptotically subject to the Gaussian distribution with mean 0 and the variance-covariance matrix $G^{-1}\left(\boldsymbol{\theta}_0\right)/N$, where N is the number of observations. In such a case, the log likelihood-ratio statistics is expanded in the Taylor series, giving

$$\lambda = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T G^{-1}\left(\boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right). \tag{27}$$

Hence this is due to the χ^2-distribution of the degrees of freedom equal to the number k of parameters. Its expectation is

$$E[\lambda] = \frac{k}{N}. \tag{28}$$

However, when the true distribution $\boldsymbol{\theta}_0$ lies on the critical set, the situation changes. The Fisher information matrix degenerates, and G^{-1} diverges, so that the expansion is no more valid. The expectation of the log likelihood estimator is asymptotically written as

$$E[\lambda] = \frac{c(N)}{N} k \tag{29}$$

where the term $c(N)$ takes various forms depending on the nature of singularities. Fukumizu [11] showed that

$$c(N) = \log N \tag{30}$$

in the case of multilayer perceptrons under a certain condition. In the case of the Gaussian mixture,

$$c(N) = \log \log N \tag{31}$$

holds [13], [16].

Since the parameters are not identifiable, we cannot estimate the parameters when the true one is on the critical set. However, we can estimate its equivalence class, and the consistency holds with the order of $\sqrt{N}$. When the true one is close to the singular point, the estimation of the parameters suffers from similar difficulty. For fixed N, the variance of the estimators diverges in inversely proportional to the square of the distance from the critical set. We need a new framework to analyze such singular cases.

9 Bayesian estimator

The Bayesian estimator is used in many cases where an adequate prior distribution is assumed for the purpose of penalizing complex models based on data. It is empirically known that the Bayesian posterior distribution or its maximizer behaves well in the case of large scale neural networks. In such a case, one uses a non-zero smooth prior on the neuromanifold.

However, a smooth prior is not regular in the equivalence class $\tilde{M}$ of the neuromanifold, because a point in the equivalence class includes infinitely many equivalent parameters when it is in the critical point. This implies that the Bayesian smooth prior is in favor of singular points (perceptrons with a smaller number of hidden units) with an infinitely large factor. Hence the Bayesian method works well in such a case to avoid overfitting. One may use a very large perceptrons with a smooth Bayesian prior, and an adequate smaller model is selected.

The Bayesian estimator of singular models was studied by Watanabe [28], [29] by using the method of algebraic geometry, in particular Hironaka's theory of resolution of singularity and Sato's formula in the theory of algebraic analysis.

10 Model selection

In order to obtain an adequate model, one should select a good class of models based on data, that is, one should determine the number of hidden units. This is the problem of model selection. AIC, BIC and MDL have been widely used as criteria of model selection.

AIC [1] is the criterion to minimize the generalization error. The model that minimizes

$$\text{AIC} = \text{training error} + \frac{k}{N} \tag{32}$$

is selected by this criterion. This is derived from the asymptotic statistical analysis, where the mle estimator $\hat{\boldsymbol{\theta}}$ is subject to the Gaussian distribution asymptotically.

MDL [22] is the criterion to minimize the length of encoding the observed data by using a family of parametric models. It is given asymptotically by the minimizer of

$$\text{MDL} = \text{training error} + \frac{\log N}{2N} k \tag{33}$$

The Bayesian criterion BIC [26] gives the same criterion as MDL.

However, in the case of multilayer perceptrons, the neuromanifold of perceptrons with a smaller number of hidden units are included in that with a larger number, but the former is the critical set of the larger neuromanifold. Therefore, the maximum likelihood estimator (or any other efficient estimators) is no more subject to the Gaussian distribution even asymptotically. Model selection is required when the estimator is close to the critical set, and hence the validity of AIC and MDL fails to hold. One should evaluate the log likelihood-ratio statistics more carefully in such a case [6].

There have been reported many computer simulations of applications of AIC and MDL. Sometimes AIC works better, while MDL does better in other cases. Such confusing reports seem to be given rise to by the difference of regular and singular models and also the different nature of singularities.

11 Conclusions

Multilayer perceptrons are popular nonlinear models for nonlinear regression analysis of observed data. A class of perceptrons is specified by the number of hidden units, and a smaller class is included in a larger class. A class of multilayer perceptrons forms a manifold named the neuromanifold, where modifiable parameters play the role of the coordinate system.

The neuromanifold is a Riemannian space, where the Fisher information matrix plays the role of the Riemannian metric. A remarkable point is that it is singular in the sense that the Riemannian metric degenerates on a subset of the manifold, in which the neuromanifold of a smaller hidden units are embedded.

We proposed the natural gradient learning method, which takes the Riemannian nature into account. It works well because it avoids the plateaus existing in to the critical set corresponding to the neuromanifold of a smaller number of hidden units.

Conventional statistical analysis assumes the existence and non-degeneracy of the Fisher information matrix. However in the case of multilayer perceptrons, as well as other similar hierarchical models, the singularity is unavoidable in its nature. The criteria of model selection such as AIC and MDL fail their validity under such circumstances.

The present paper reviews such aspects of learning in neural networks, which require a new statistical analysis of singular models. Geometry will be useful for this purpose.

References

[1] Akaike H. (1974). *A new look at the statistical model identification.* IEEE Trans. Automatic Control AC-19, 716–723.

[2] Amari S. (1965). *Theory of adaptive pattern classifiers.* IEEE Trans. Elect. Comput. EC-16, 299–307.

[3] Amari S. (1998). *Natural gradient works efficiently in learning.* Neural Computation **10**, 251–276.

[4] Amari S., Nagaoka H. (2000). *Information geometry.* AMS and Oxford University Press, New York.

[5] Amari S., Ozeki T. (2001). *Differential and algebraic geometry of multilayer perceptrons.* IEICE Trans., E84-A, 31–38.

[6] Amari S. (2003). *New consideration on criteria of model selection.* Neural Networks and Soft Computing (Proceedings of the Sixth International Conference on Neural Networks and Soft Computing), L. Rutkowski and J. Kacprzyk (eds.), 25–30.

[7] Amari S., Ozeki T., Park H. (2003). *Learning and inference in hierarchical models with singularities.* Systems and Computers in Japan **34** (7), 701–708.

[8] Amari S., Park H., Fukumizu K. (2000). *Adaptive method of realizing natural gradient learning for multilayer perceptrons.* Neural Computation **12**, 1399–1409.

[9] Amari S., Park H., Ozeki T. (2002). *Geometrical singularities in the neuromanifold of multilayer perceptrons.* Advances in Neural Information Processing Systems,T.G. Dietterich, S. Becker, and Z. Ghahramani (eds.) **14**, 343–350.

[10] Chen A.M., Liu H., Hecht-Nielsen R. (1993). *On the geometry of feedforward neural network error surfaces.* Neural Computation **5**, 910–927.

[11] Fukumizu K. (2003). *Likelihood ratio of unidentifiable models and multilayer neural networks.* The Annals of Statistics **31** (3), 833–851.

[12] Fukumizu K., Amari S. (2000). *Local minima and plateaus in hierarchical structures of multilayer perceptrons.* Neural Networks **13**, 317–327.

[13] Hartigan J.A. (1985). *A failure of likelihood asymptotics for normal mixtures.* Proc. Barkeley Conf. in Honor of J. Neyman and J. Kiefer **2**, 807–810.

[14] Inoue M., Park H., Okada M. (2003). *On-line learning theory of soft committee machines with correlated hidden units - Steepest gradient descent and natural gradient descent -.* J. Phys. Soc. Jpn **72** (4), 805–810.

[15] Kůrková V., Kainen P.C. (1994). *Functionally equivalent feedforward neural networks.* Neural Computation **6**, 543–558.

[16] Lin X., Shao Y., (2003). *Asymptotics for likelihood ratio tests under loss of identifiability.* The Annals of Statistics **31** (3), 807–832.

[17] Park H., Amari S., Fukumizu K. (2000). *Adaptive natural gradient learning algorithms for various stochastic models.* Neural Networks **13**, 755–764.

[18] Park H., Inoue M., Okada M. (2003). *Learning dynamics of multilayer perceptrons with unidentifiable parameters.* J. Phys. A: Mathe. Gen. **36** (47), 11753–11764.

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

[20] Rattray M., Saad D., Amari S. (1998). *Natural gradient descent for on-line learning.* Physical Review Letters **81**, 5461–5464.

[21] Riegler P., Biehl M. (1995). *On-line backpropagation in two-layered neural networks.* J. Phys. A; Mathe. Gen. **28**, L507–L513.

[22] Rissanen J. (1978). *Modelling by shortest data description.* Automata **14**, 465–471.

[23] Rumelhart D.E., Hinton G.E., Williams R.J. (1986). *Learning internal representations by error propagation.* In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (eds.), Parallel distributed processing (Vol. **1**, 318–362), Cambridge, MA:MIT Press.

[24] Rüger S.M., Ossen A. (1995). *The metric of weight space.* Neural Processing Letters **5**, 63–72.

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

[26] Schwarz G. (1978). *Estimating the dimension of a model.* The Annals of Statistics **6**, 461–464.

[27] Sussmann H.J. (1992). *Uniqueness of the weights for minimal feedforward nets with a given input-output map.* Neural Networks **5**, 589–593.

[28] Watanabe S. (2001a). *Algebraic analysis for non-identifiable learning machines.* Neural Computation **13**, 899 – 933.

[29] Watanabe S. (2001b). *Algebraic geometrical methods for hierarchical learning machines.* Neural Networks **14** (8), 1409 – 1060.

Address: S. Amari, T. Ozeki, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Satitma, 351-0198, Japan
H. Park, Dept. of Computer Science, College of Natural Science, Kyungpook National University, Sangyuk-dong, Buk-gu, Daegu, 702-701, Korea

E-mail: {amari,tomoko}@brain.riken.jp, {hypark}@knu.ac.kr