INVITED PAPER Special Section on Nonlinear Dynamics—Adaptive, Learning and Neural Systems—

Information Geometry of Neural Networks

Shun-ichi AMARI†, *Member*

SUMMARY Information geometry is a new poweful method of information sciences. Information geometry is applied to manifolds of neural networks of various architectures. Here is proposed a new theoretical approach to the manifold consisting of feedforward neural networks, the manifold of Boltzmann machines and the manifold of neural networks of recurrent connections. This opens a new direction of studies on a family of neural networks, not a study of behaviors of single neural networks.

key words: *neural networks, information geometry, non-linear system, probability distribution*

1. Introduction

The brain indeed has rich capability and flexibility to perform various types of information processing. Information is distributed in the brain over a vast number of neurons as a time-developing excitation pattern and is processed in parallel by mutual interaction of neurons. There might be strong information principles which guarantee surprising capability of information processing through parallel and distributed processing of information. The brain can be regarded as a biological realization of these principles through a long history of evolution.

Recently, researchers have interests in modeling information processing in terms of neural networks. Models are sometimes faithful but simplified copies of the actual neural networks, but sometimes are far from the real brain. However, even if models are different from the actual brain, they could be an engineering realization of the above mentioned information principles that the brain also has used in different styles. From this point of view, the recent research(1) on neural networks provides a new powerful method for elucidating non-linear systems with large degrees of freedom, and hence is enthusiastically welcome in various fields of engineering sciences.

Most of research on neural networks or systems are concerned with behaviors of networks or systems, and little attention has been paid to the mutual relation of two networks or more generally a geometrical structure among a family of networks. It is important to study the geometry of a family of networks in order to elucidate the capability and limitation of, not a single network, but a family of networks of a given architecture as a whole.

Information geometry(2) is a new method of approach to families of information systems. More precisely, it originated from the study of intrinsic geometrical structures of a continuous family of probability distributions which forms a geometrical manifold(3)-(5). It gives a new concepts of a manifold of dual connections to differential geometry. The theory is not only useful in statistical inference but has been proved to be fruitful in various information sciences, providing them with a new powerful method(6)-(9). The present paper tries to establish applications of information geometry to various architectures of neural networks. We show that information geometry is applicable to manifolds of feedforward neural networks, a mixture model of networks, recurrent neural networks such as Boltzmann machines and more general neural networks of asymmetric recurrent connections. It should be remarked that the same argument holds for a family of general linear and nonlinear systems other than neural networks.

We give here a general approach. Some examples(8),(9) have already been studied in detail, but most of the subjects are expected to be studied in future.

2. Manifolds of Neural Networks

Various manifolds of neural networks are presented in this section. They have different architectures and different behaviors. Since we use the method of information geometry originated from the study of families of probability distributions(3), each network is associated with a probability distribution on a respective sample space.

2.1 Feedforward Networks

Let us consider a memoryless stochastic system which receives an m-dimensional vector input $\boldsymbol{x}=(x_1, \cdots, x_m)$ and emits a k-dimensional output $\boldsymbol{z}=(z_1, \cdots, z_k)$. When the system is memoryless and stochastic, its behavior is given by a conditional probability (density) function $p(\boldsymbol{z}|\boldsymbol{x})$, which gives the probability of emitting $\boldsymbol{z}$ when $\boldsymbol{x}$ is input. Here, the input x_i and

Manuscript received December 20, 1991.

† The author is with the Faculty of Engineering, The University of Tokyo, Tokyo, 113 Japan.

output z_j may be either a continuous variable or a discrete variable.

A feedforward multilayer neural network is a typical example of such systems. When neural elements behave stochastically, the network emits output $\boldsymbol{z}$ stochastically depending on the current input $\boldsymbol{x}$. In this case, a network directly corresponds to a conditional probability distribution $p(\boldsymbol{z}|\boldsymbol{x})$. Even when a network is deterministic, it is possible to interprete a deterministic analog output o_i of the i-th output element as the probability that the output element fires, provided $0<o_i<1$. In this case, the network is regarded as emitting a stochastic output z_i taking on binary values 0 and 1, and the analog network output o_i specifies the firing probability $p(z_i=1|\boldsymbol{x})=o_i$. It is then expected that the actual pulse frequency is close to the network output o_i.

When a system is parameterized by an n-dimensional real parameter $\xi=(\xi^1, \xi^2, \cdots, \xi^n)$, its conditional probability is denoted by $p(\boldsymbol{z}|\boldsymbol{x}; \xi)$ as a function of ξ. For example, when such systems are realized by a prescribed architecture of neural networks, the parameter ξ is composed of all the modifiable synaptic weights of connections and threshold values.

Let M be the set of all such neural networks parameterized by ξ. Then each network in M is specified by the coordinates ξ. When the manifold M consists of neural networks, it is called the manifold of neural networks[8]. We study the geometric properties of the manifold M.

We first show an intuitive reason why the geometry is useful in studying non-linear systems. It should be noted that the present theory does not directly study the behaviors of single neural networks, but studies a family of networks of the same architecture. Therefore, this is a comparative study of various networks. We can define an intrinsic distance or divergence measure between two networks or between their behaviors. The geometry also provides an overall view of M, which elucidates the capabilities and limitations of the system architecture of M.

We explain this by using a simple example of multilayer deterministic neural networks, which emits a continuous output

$$\boldsymbol{z}=\boldsymbol{f}(\boldsymbol{x}, \xi)$$

when $\boldsymbol{x}$ is input, where ξ is the parameter specifying each network. Let M be the set of all such networks. Each network realizes a mapping from $X=\{\boldsymbol{x}\}$ to $Z=\{\boldsymbol{z}\}$. Let S be the set of all the continuous functions from X to Z,

$$S=\{\boldsymbol{h}(\boldsymbol{x})\}.$$

Then, $\boldsymbol{f}(\boldsymbol{x}, \xi)$ is also a member of S. Hence, the manifold M of neural networks is included in S, occupying a part of S as an n-dimensional submanifold (Fig. 1).

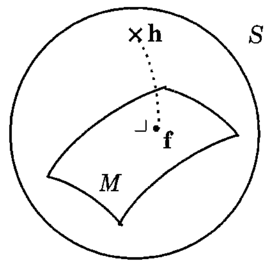

Fig. 1 Neural manifold M imbedded in the function space S.

Let $\boldsymbol{h}(\boldsymbol{x})$ be a function which does not belong to M, that is, not realizable by a neural network. We want to approximate it by a suitable function realizable by a neural network. In order to answer this problem we need a distance or divergence measure. When S has a Euclidean structure, the best approximation is given by projecting $\boldsymbol{h}(\boldsymbol{x})$ to M. When S and M are Riemannian, we need to project it by using a geodesic curve. Moreover, it is interesting to know which part M occupies in S. In order that any function $\boldsymbol{h}(\boldsymbol{x})$ in S is well approximated by a function in M, it is necessary for M to be highly curved like the Peano curve. This guarantees a large approximation capacity of M. However, this contradicts learnability. All of these properties are geometrical, explaining the necessity of a geometrical theory.

2.2 Manifold of Boltzmann Machines

A Boltzmann machine is a recurrently connected neural network with stochastic neurons. Let z_i be the output of the i-th neuron. The connection w_{ij} is assumed to be symmetric, that is $w_{ij}=w_{ji}$, and $w_{ii}=0$ is also assumed. Let

$$u_i=\sum w_{ij}z_j-h_i$$

be the weighted sum of the output from other neurons to the i-th neuron where h_i is a threshold. The i-th neuron then takes $z_i=1$ with probability $f(u_i)$ and $z=0$ with probability $1-f(u_i)$, where

$$f(u)=\frac{1}{1+\exp\{-u\}}.$$

Each neuron changes its state asynchronously. The state of the network $\boldsymbol{z}=(z_i)$ is changed stochastically in this manner. Mathematically speaking, a Boltzmann machine defines an ergodic Markov chain, where the state space $Z=\{\boldsymbol{z}\}$ consists of 2^n elements when n is the number of neurons. The function of the state

$$E(\boldsymbol{z})=-\frac{1}{2}\sum w_{ij}z_iz_j+\sum h_iz_i$$

is called the "energy" of the state. It is easy to show that the Markov chain satisfies the condition of detailed balance and that the stable distribution is given by

$$p(\boldsymbol{z}) = c \exp\{-E(\boldsymbol{z})\},$$

where c is the normalization constant.

Let $\xi = (\xi^i)$ be the set of parameters specifying a Boltzmann machine, namely w_{ij}'s and h_i's. The set of all the Boltzmann machines with n neurons forms a manifold M_B with an admissible coordinate system $\xi = (\xi^i)^{(9)}$. A Boltzmann machine specified by ξ corresponds uniquely to the stationary probability distribution

$$p(\boldsymbol{z}; \xi) = c \exp\{-E(\boldsymbol{z}; \xi)\},$$

where $E(\boldsymbol{z}; \xi)$ is the energy function.

Let S be the set of all the probability distributions $q = \{q(\boldsymbol{z})\}$, $\boldsymbol{z} \in Z$, over 2^n atoms $\boldsymbol{z}$, satisfying

$$0 < q(\boldsymbol{z}) < 1,$$

$$\sum_{\boldsymbol{z}} q(\boldsymbol{z}) = 1.$$

Since a distribution q is specified by $2^n - 1$ numbers $q(\boldsymbol{z})$'s, S is an (2^n-1)-dimentional manifold. The Boltzmann manifold M_B is an $n(n+1)/2$-dimentional submanifold imbedded in S. Therefore, only a special type of probability distributions are in M_B, that is, are realizable by Boltzmann machines. In order to approximate a given q by a Boltzmann machine, we need geometrical analysis of the manifolds S and M_B.

2.3 Manifold of General Recurrent Networks

A Boltzmann machine is a recurrent network with the restriction of the symmetry $w_{ij} = w_{ji}$ and the off-diagonality $w_{ii} = 0$. However, it is important to study more general networks of recurrent connections without these constraints. The set M_R of all the networks with general recurrent connections forms another manifold, where the coordinate system $\xi = (\xi^i)$ includes $n(n+1)$ components. The Boltzmann manifold M_B can be regarded as a submanifold of M_R specified by the constraints $w_{ij} = w_{ji}$ and $w_{ii} = 0$.

In order to apply the method of information geometry, we need to associate a network with a probability distribution. However, the stable distribution in this case does not have a one-to-one correspondence with the network, because a number of networks have the same stable distribution in common. Therefore, we associate a network with the Markov chain that the network defines, the Markov chain over the state space $Z = \{\boldsymbol{z}\}$. A Markov chain is defined by the state transition matrix $p(\boldsymbol{z}'|\boldsymbol{z})$, the probability that the next state is $\boldsymbol{z}'$ when the present state is $\boldsymbol{z}$. The Markov chain of the network specified by ξ is denoted by $p(\boldsymbol{z}'|\boldsymbol{z}; \xi)$.

Let S_M be the set of all the Markov chains over the state space Z. It is a higher dimensional manifold, in which M_R is imbedded as a submanifold, because not all the Markov chains are realizable by recurrent neural networks. We can then study the geometry of the manifold of recurrent neural networks by the method of information geometry. This elucidates dynamical properties of neural networks.

3. Method of Information Geometry

We present here a brief introduction to information geometry. Refere to a monograph[2], a review paper[1], etc. for details.

Let us consider a family M of probability distributions $p(\boldsymbol{x}; \xi)$, parameterized by ξ, of random variable $\boldsymbol{x}$, which may be discrete, continuous or even infinite-dimensional (in the case of stochastic processes). It forms a geometrical manifold with an admissible coordinate system ξ. We first introduce a tangent vector space to each point of M, which may be regarded as a local linearization of M at each point.

Let T_ξ be the tangent space at point ξ of M (Fig. 2). It is spanned by basis vectors $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n$, where $\boldsymbol{e}_i$ is the tangent vector along the coordinate curve ξ^i. Mathematicians denote $\boldsymbol{e}_i$ by $\partial/\partial\xi^i$, calling $\{\boldsymbol{e}_i\}$ the natural basis associated with the coordinate system $\xi = (\xi^i)$. The tangent space T_ξ represents the local structure of M by linearizing it. Let us introduce a metric structure in T_ξ by defining the inner product $\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle$ of basis vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$. The quantity $g = (g_{ij})$,

$$g_{ij}(\xi) = \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle$$

is called the Riemannian metric of M. Information geometry defines the metric by

$$g_{ij}(\xi) = E[\partial_i \log p(\boldsymbol{x}; \xi)\, \partial_j \log p(\boldsymbol{x}; \xi)],$$

where E denotes the expectation with respect to $p(\boldsymbol{x}; \xi)$,

$$E[a(\boldsymbol{x})] = \int p(\boldsymbol{x}; \xi)\, a(\boldsymbol{x})\, d\boldsymbol{x},$$

and

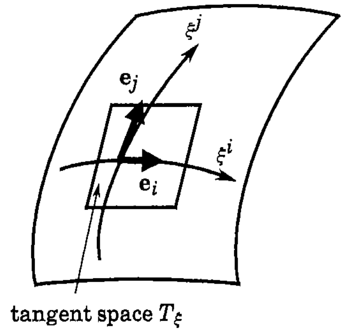

Fig. 2 Tangent space T_ξ at point ξ.

$$\partial_i = \frac{\partial}{\partial \xi^i}.$$

The matrix (tensor) g is called the Fisher information matrix in statistics.

The metric defines the local distance between two nearby systems specified by coordinates ξ and $\xi + d\xi$. Their difference can be identified with an element

$$d\xi = \sum d\xi^i \boldsymbol{e}_i$$

in the tangent space T_ξ. Therefore, the square of their distance is given by the quadratic form

$$ds^2 = \langle d\xi, d\xi \rangle = \sum g_{ij} d\xi^i d\xi^j.$$

The meaning of the distance can be understood from the theory of statistics. Let $\boldsymbol{x}_i$, $i = 1, \cdots, N$, be N independent observations of random variable $\boldsymbol{x}$. When the true parameter ξ is unknown, we need to estimate it. Let $\hat{\xi}$ be the maximum likelihood estimator based on N observations. Its mean square error is asymptotically given by

$$E[(\hat{\xi}^i - \xi^i)(\hat{\xi}^j - \xi^j)] = \frac{1}{N} g^{ij},$$

where (g^{ij}) is the inverse matrix of (g_{ij}). This implies that the distance ds of two distributions specified by ξ and $\xi + d\xi$ is defined by the degree of their undistinguishability.

A manifold M is said to be Riemannian, when it is equipped with the metric tensor $g_{ij}(\xi)$, that is, when the inner product is defined in every tangent space T_ξ. Therefore, our manifold M is Riemannian. Each tangent space T_ξ is regarded as the local linearization at each point of the manifold M. In order to recover its curved structure, we need to connect nearby tangent spaces. This can be done by introducing an affine connection, which establishes an affine correspondence between two nearby tangent spaces.

We cannot simply say that the i-th basis vectors $\boldsymbol{e}_i(\xi)$ are the same or correspond to each other at different tangent spaces. This is because the manifold may be curved and the coordinate lines may also be curved. Therefore, $\boldsymbol{e}_i(\xi)$ changes as point ξ changes. Its intrinsic change is defined by the covariant derivative

$$\nabla_{\boldsymbol{e}_j} \boldsymbol{e}_i,$$

which shows the intrinsic change of $\boldsymbol{e}_i(\xi)$ as ξ moves in the direction of $\boldsymbol{e}_j$. This is a vector, so that it is represented by the components

$$\Gamma_{jik}(\xi) = \langle \nabla_{\boldsymbol{e}_j} \boldsymbol{e}_i, \boldsymbol{e}_k \rangle,$$

which are called the coefficients of the affine connection.

By using the affine connection, we can define a geodesic which is the generalization of a straight line. Let

$$\xi = \xi(t)$$

be a curve in M parameterized by a scalar t. The tangent vector of the curve is given by

$$\dot{\xi}(t) = \sum \dot{\xi}^i(t) \boldsymbol{e}_i,$$

where $\cdot$ denotes d/dt. When $\dot{\xi}(t)$ does not change intrinsically along the curve, that is

$$\nabla_{\dot{\xi}} \dot{\xi} = 0,$$

or in the component form

$$\sum g_{ij} \ddot{\xi}^j + \sum \Gamma_{ijk} \dot{\xi}^j \dot{\xi}^k = 0,$$

the curve is called a geodesic. Here, the covariant derivative $\nabla_{\dot{\xi}} \dot{\xi}$ is naturally extended from that of the basis vectors, $\nabla_{\boldsymbol{e}_j} \boldsymbol{e}_j$.

Two different affine connections are naturally introduced in the manifold of probability distributions. One is the exponential connection (e-connection) defined by

$$\Gamma^{(e)}_{ijk}(\xi) = E[\partial_i \partial_j \log p(\boldsymbol{x}; \xi)\, \partial_k \log p(\boldsymbol{x}; \xi)]$$

and the other is the mixture connection (m-connection) defined by

$$\Gamma^{(m)}_{ijk}(\xi) = \Gamma^{(e)}_{ijk} + T_{ijk}$$

where

$$T_{ijk}(\xi) = E[\partial_i \log p \;\; \partial_j \log p \;\; \partial_k \log p].$$

These two connections are dually coupled, and this provides a new dual theory of differential geometry[(1),(2)].

In some special but important cases, a manifold of dual connections becomes dually flat in the sense that the e- and m-Riemann-Christoffel curvatures vanish. In such a case, an e-affine and m-affine coordinate systems, denoted by $\theta = (\theta^i)$ and $\eta = (\eta_i)$ exist for which $\Gamma^{(e)}_{ijk} = 0$ in the θ-coordinate system and $\Gamma^{(m)}_{ijk} = 0$ in the η-coordinate system. This implies that the θ-coordinate lines are e-geodesics and the η-coordinate lines are m-geodesics.

A dually flat manifold has interesting properties. There are two potential functions $\psi(\theta)$ and $\varphi(\eta)$ for which

$$g_{ij}(\theta) = \frac{\partial^2 \psi(\theta)}{\partial \theta^i \partial \theta^j}, \quad g^{ij}(\eta) = \frac{\partial^2 \varphi(\eta)}{\partial \eta_i \partial \eta_j}$$

hold. Moreover, the two coordinates are connected by the Legendre transformation

$$\psi(\theta) + \varphi(\eta) - \sum \theta^i \eta_i = 0.$$

An invariant divergence measure $D(P, Q)$ can be introduced between two points P and Q in a dually flat manifold M. It is given by

$$D(P, Q) = \psi(\theta_P) + \varphi(\eta_Q) - \sum \theta^i_P \eta^Q_i,$$

where θ_P and η_Q are the θ- and η-coordinates of P and

Q, respectively. The divergence is half a square of the Riemannian distance

$$D(P, P+dP)=\frac{1}{2}\sum g_{ij}d\theta^i d\theta^j,$$

when $Q=P+dP$ is infinitesimally close to P. The divergence turns out to be the Kullback information

$$D(P, Q)=\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

in many cases.

The remarkable properties of a dually flat manifold is that a generalized Pythagorus theorem holds in it.

Generalized Pythagorus Theorem 1: Let P, Q, R be three points such that the m-geodesic connecting P and Q is orthogonal at Q to the e-geodesic connecting Q and R. Then,

$$D(P, Q)+D(Q, R)=D(P, R).$$

This gives the following projection theorem.

Generalized Projection Theorem 2: Let V be a smooth closed region in M, and P be a point not belonging to V. The best approximation $\hat{Q}$ of P by a point belonging to V in the sense of minimizing the divergence $D(P, Q)$, $Q\in V$, is given by projecting P to ∂V by an m-geodesic,

$$\min_{Q\in V} D(P, Q)=D(P, \hat{Q}).$$

The projection is unique when V is e-convex.

The theorems are of fundamental use in elucidating the geometric properties of manifolds of neural networks.

4. Properties of Neural Network Manifolds

4.1 Manifold of Feedforward Networks

Each network of the manifold M of feedforward type is identified with a conditional distribution $p(\boldsymbol{z}|\boldsymbol{x}; \xi)$. Let $p(\boldsymbol{x})$ be the distribution of input signals. We then have the joint probability distribution of input-output pairs

$$p(\boldsymbol{x}, \boldsymbol{z}; \xi)=p(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x}; \xi).$$

Therefore, the neural manifold is identified with the manifold of probability distributions $\{p(\boldsymbol{x}, \boldsymbol{z}; \xi)\}$, where the random variable is the pair $(\boldsymbol{x}, \boldsymbol{z})$. The metric, and e- and m-connections are defined in the usual way. The inverse (g^{ij}) of the metric represents the estimation errors when the network parameter is estimated from observed input-output data pairs.

A neural manifold M is in general curved in the manifold S of functions. However, there are some interesting cases where the neural manifold is dually flat. One example is the manifold of higher-order neurons. This case is analyzed by Amari[8], giving an invariant decomposition of approximation processes. The meaning of the potential functions as well as the divergence are elucidated.

Another example is the manifold of a mixture model, in which the output probability $p(\boldsymbol{z}|\boldsymbol{x}; \xi)$ is given by a linear combination of component distributions $a_i(\boldsymbol{z}|\boldsymbol{x})$.

$$p(\boldsymbol{z}|\boldsymbol{x}; \xi)=\sum\xi^i a_i(\boldsymbol{z}|\boldsymbol{x}).$$

Such a model is remarked recently[10], but has not yet been analyzed in detail. It is fruitful to study such a model by information geometry.

4.2 Approximation of Probability by Boltzmann Machine

Given a probability distribution $q(\boldsymbol{z})$, we can find a Boltzmann machine which approximates $q(\boldsymbol{z})$. This is given by the learning algorithm proposed by Ackley et al.[11]. The geometry of the Boltzmann manifold M_B is studied in detail by Amari et al.[9]. It is proved that M_B is dually flat. This implies that the best approximation is given by the m-projection of q to M_B. A new learning rule is also proposed such that its trajectory is an e-geodesic. This has much quicker convergence than the conventional one.

However, the manifold of Boltzmann machines with hidden units is not dually flat. Only preliminary studies have been done, and we need deeper research on this manifold.

4.3 Manifold of Markov Chains

Let S be the manifold of all the Markov chains over the state space $Z=\{\boldsymbol{z}\}$. It is proved that S is a dually flat manifold. The manifold M_R of the recurrent neural networks forms a dually flat submanifold imbedded in S. Therefore, we can approximate any Markov chain by a recurrent network. The best approximation is given by projecting it to M_R by an m-geodesic. This projection is unique, since M_R is e-flat.

A Boltzmann machine also belongs to M_R, because it is a special example of the general recurrent network. It satisfies the constraints,

$$w_{ij}=w_{ji},\quad w_{ii}=0.$$

The Boltzmann manifold M_B may form an e-flat submanifold of M_R. Therefore, the manifold M_R of recurrent neural networks is invariantly decomposed into a foliation of flat manifolds. We can decompose a network into two factors, a Boltzmann type stationary component and a cyclic component.

Information geometry is expected to open a new theory of Markov chains and of recurrent neural networks. This is also an important subject to be studied further.

5. Conclusions

Research of neural network models is not only useful for understanding the brain functioning, but also is important as a method of analyzing non-linear systems. Behaviors of neural networks have so far been studied intensively, but little attention has been paid to the study of a family of neural networks as a whole. Such studies are important for understanding the capabilities and limitations of neural networks of a given architecture.

Information geometry gives a new powerful method to information sciences. It has been applied to manifolds of neural networks of various architectures, opening a new direction of study.

References

(1) Amari S.: "Mathematical Foundation of Neurocomputing", Proc. IEEE, **78**, pp. 1443-1463 (1990).

(2) Amari S.: "Information Geometry", Ouyousuri (Mathematical Sciences), **2**, pp. 37-56 (1992).

(3) Amari S.: "Differential-Geometrical Methods in Statistics", Springer Lecture Notes, **28**, Springer (1985).

(4) Kumon M. and Amari S.: "Geometrical Theory of Higher-Order Asymptotics of Test, Interval Estimator and Conditional Inference", Proc. R. Soc. Lond., **A387**, pp. 429-458 (1983).

(5) Amari S. and Kumon M.: "Estimation in the Presence of Infinitely Many Nuisance Parameters—Geometry of Estimating Functions", Annals of Statistics, **16**, pp. 1044-1068 (1988).

(6) Amari S.: "Differential Geometry of a Parametric Family of Invertible Linear Systems—Riemannian Metric, Dual Affine Connections and Divergence", Mathematical Systems Theory, **20**, pp. 53-82 (1987).

(7) Amari S. and Han T. S.: "Statistical Inference under Multi-Terminal Rate Restrictions—a Differential Geometrical Approach", IEEE Trans. Inf. Theory, **IT-35**, pp. 217-227 (1989).

(8) Amari S.: "Dualistic Geometry of the Manifold of Higher-Order Neurons", Neural Networks, **4**, pp. 443-451 (1990).

(9) Amari S., Kurata K. and Nagaoka H.: "Information Geometry of Boltzmann Machines", IEEE Trans. Neural Networks, **3** (March 1992).

(10) Jacobs R. A., Jordan M. I., Nolman S. J. and Hinton G. E.: "Adaptive Mixtures of Local Experts", Neural Computation, **3**, pp. 79-87 (1991).

(11) Ackley D. H., Hinton G. E. and Sejnowski T. J.: "A Learning Algorithm for Boltzmann Machine", Cognitive Science, **9**, pp. 147-169 (1985).

Shun-ichi Amari was born in Tokyo, Japan, on January 3, 1936. He graduated from the University of Tokyo in 1958 majoring in mathematical engineering, and received the Dr.Eng. degree from the University of Tokyo in 1963. He was a Professor at Kyushu University and is now a Professor at the University of Tokyo. He has been engaged in research in many areas of mathematical engineering and applied mathematics, among them topological network theory, differential geometry of continuum mechanics, pattern recognition, and information sciences. In particular, he has devoted himself to mathematical foundations of neural network theory. Another main subject of his research is the information geometry that he himself proposed. Dr. Amari is a founding governing board member of the International Neural Network Society, a Coeditor-in-Chief of *Neural Networks*, Editor-in-Chief of *Mathematical Sciences*, and a member of the editorial or advisory editorial boards of more than 15 international journals.