Information Geometry of Boltzmann Machines

Shun-ichi Amari, *Member, IEEE,* Koji Kurata, and Hiroshi Nagaoka, *Member, IEEE*

Abstract— **A Boltzmann machine is a network of stochastic neurons. The set of all the Boltzmann machines with a fixed topology forms a geometrical manifold of high dimension, where modifiable synaptic weights of connections play the role of a coordinate system to specify networks. A learning trajectory, for example, is a curve in this manifold. It is important to study the geometry of the neural manifold, rather than the behavior of a single network, in order to know the capabilities and limitations of neural networks of a fixed topology. Using the new theory of information geometry, the present paper establishes a natural invariant Riemannian metric and a dual pair of affine connections on the Boltzmann neural network manifold. The meaning of geometrical structures is elucidated from the stochastic and the statistical point of view. This leads to a natural modification of the Boltzmann machine learning rule.**

I. Introduction

THE behavior of a single neural network may be studied by mathematical analysis or by computer simulation. However, it is important to study not a specific network but a whole family of networks to elucidate the capabilities and limitations of the family of networks of a fixed architecture. When neural networks of a fixed architecture are specified by N free parameters, such as connection weights and threshold values, the set of such neural networks forms an N-dimensional manifold in the sense of geometry. Each neural network is regarded as a point in the manifold.

A neural manifold may be embedded, as a submanifold, in the manifold of a more complex neural architecture or in the manifold of a general (nonneural) information processing system. It is interesting and important to know the intrinsic geometry of a neural manifold. There arise two different types of problems. One is the inner geometry of a neural manifold, such as the intrinsic distance between two networks and the curvature of the neural manifold and their roles in information processing. The other is the relative geometry, which shows how and where the neural manifold is embedded in a larger manifold. These geometrical considerations help to discover the ability of a family of neural networks to approximate a given general information processing behavior, by finding the most appropriate neural network approximation in it. These considerations are also useful for understanding the learning behavior of neural networks [1], because learning is a motion in the neural manifold.

Manuscript received March 21, 1991; revised October 2, 1991.

S. Amari is with the Department of Mathematical Engineering, University of Tokyo, Tokyo 113, Japan.

K. Kurata is with the Department of Biophysical Engineering, Osaka University, Toyonaka 560, Japan.

H. Nagaoka is with the Department of Information Engineering, Hokkaido University, Sapporo 060, Japan.

IEEE Log Number 9104762.

The present paper studies the geometry of the manifold of Boltzmann machines (Ackley, Hinton and Sejnowski [2], Aarts and Korst [3]). Since a Boltzmann machine is uniquely related to the stationary Gibbs distribution on the state, the manifold of Boltzmann machines can be identified with the manifold of stationary probability distributions, and this manifold is embedded in the manifold of all the probability distributions. A new differential geometry has been constructed on the manifold of probability distributions (Amari [4], [5], Nagaoka and Amari [6], Amari *et al.* [7], Chentsov [8], etc.). This is a Reimannian manifold having dually coupled affine connections. Such a dual structure has been proved to play a fundamental role in statistical inference (Amari [5], Kumon and Amari [9], Amari and Kuman [10], Okamoto *et al.* [11]), in control systems theory (Amari [12]), and in multiterminal information theory (Amari and Han [13], Amari [14]).

We apply this powerful new geometry to the manifold of Boltzmann machines. It will be proved that the manifold of Boltzmann machines without hidden units is an e-flat manifold, so that an invariant divergence measure is introduced into it. We solve the following approximation problem: Given an information source from an environment, find the Boltzmann machine that realizes, as faithfully as possible, the given probability distribution as its stationary distribution. It will be shown that the best approximation is given by the m-geodesic projection, and is unique in the case of no hidden units. Furthermore, a generalized Pythagorian theorem makes it possible to decompose the approximation error in an invariant manner. However, for Boltzmann machines with hidden units, the corresponding manifold is not e-flat, but it has some interesting properties. A Riemannian metric is also introduced into the manifold of Boltzmann machines and its meaning is explained in connection with learning.

The present paper tries to show the importance and usefulness of a geometric approach to the manifold of neural networks. It is also expected that the present approach will provide a useful method for analyzing random Markov fields (Geman and Geman [15]), and the physicist's theory of neural networks (see, e.g., Amit [16], Levin, Tishby, and Solla [17], and Seung and Sompolinsky [18]). This theory also adds a new method to the mathematical foundations of neurocomputing (Amari [19]; see also a related paper by Amari [20]).

II. Boltzmann Machines—Stochastic Neural Networks

A. The Stationary Distribution of a Boltzmann Machine

A Boltzmann machine is a neural network with symmetric recurrent connections, where each neural element fires stochas-

tically and the firing depends on the weighted sum of inputs. More specifically, let us consider a neural network consisting of n neurons, and let $w_{ij} = w_{ji}$ be the symmetric connection weight between the ith neuron and the jth neuron. The self-recurrent connection w_{ii} is assumed to be 0. Let h_i be the threshold of the ith neuron.

Each neuron takes on two values: 1 (excited) and 0 (quiescent). The potential of the ith neuron, u_i, is defined by

$$u_i = \sum_j w_{ij}x_j - h_i$$

where x_j is the state of the jth neuron. In order to make the notation simpler, we introduce the zeroth neuron whose output is always $x_0 = 1$, and we let the connection weight from the zeroth neuron to the ith neuron be $w_{i0} = -h_i$. Then u_i can be written

$$u_i = \sum_{j=0}^{n} w_{ij}x_j. \tag{1}$$

Each neuron changes its state asynchronously depending on u_i, where the new state, x_i, of the ith neuron is equal to 1 with probability $f(u_i)$ and is equal to 0 with probability $1 - f(u_i)$, and

$$f(u) = \frac{1}{1 + \exp\{-u\}}. \tag{2}$$

The vector $\boldsymbol{x} = (x_1, \cdots, x_n)$ is called the state of the Boltzmann machine. The state transition is mathematically described by a Markov chain with 2^n states $\boldsymbol{x}$. When all the neurons are connected, it forms an ergodic Markov chain, having a unique stationary distribution $p(\boldsymbol{x})$. The stationary distribution is given by

$$p(\boldsymbol{x}) = c \exp\{-E(\boldsymbol{x})\}, \tag{3}$$

where c is a normalizing constant and

$$E(\boldsymbol{x}) = -\sum_{i>j} w_{ij}x_ix_j \tag{4}$$

is called the energy of the state $\boldsymbol{x}$. Whatever initial state of Boltzmann machine starts from, the probability of a state $\boldsymbol{x}$ converges to $p(\boldsymbol{x})$, and state $\boldsymbol{x}$ appears with relative frequency $p(\boldsymbol{x})$ over a long course of time.

B. Learning of Boltzmann Machines

We first explain the simple case where no hidden elements exist. Let $q(\boldsymbol{x})$ be a probability distribution over $\boldsymbol{x}$ with which an environmental information source emits a signal $\boldsymbol{x}$. Signals are generated independently subject to $q(\boldsymbol{x})$ and presented to a Boltzmann machine. This machine is required to modify its connection weights and thresholds so that it will simulate the environmental information source. More precisely, it is required that the stationary distribution $p(\boldsymbol{x})$ of the Boltzmann machine become as close to $q(\boldsymbol{x})$ as possible.

Ackley *et al.* [2] proposed a learning rule in which the average adjustment of w_{ij} is given by

$$\Delta w_{ij} = \Delta w_{ji} = \varepsilon(q_{ij} - p_{ij}), \tag{5}$$

where ε is a small constant, q_{ij} is the relative frequency that both x_i and x_j are jointly excited under the probability distribution $q(\boldsymbol{x})$, and p_{ij} is the relative frequency that both x_i and x_j are jointly excited under $p(\boldsymbol{x})$, that is, when the Boltzmann machine is running freely. These quantities can be written as

$$q_{ij} = E_q[x_ix_j] = \sum_x q(\boldsymbol{x})x_ix_j. \tag{6}$$

and

$$p_{ij} = E_p[x_ix_j] = \sum p(\boldsymbol{x})x_ix_j, \tag{7}$$

where E_q and E_p denote, respectively, the expectation with respect to the probability distributions $q(\boldsymbol{x})$ and $p(\boldsymbol{x})$.

The learning rule is realized by the Hebbian synaptic modification method in two phases. In the first phase, the input learning phase, the connection weight w_{ij} is increased by a small amount whenever both x_i and x_j are exited by an input $\boldsymbol{x}$; hence on average the increment of w_{ij} is proportional to q_{ij}. In the second phase, the free or antilearning phase, w_{ij} is decreased by the same small amount whenever both x_i and x_j are excited by a free state transition; hence the decrement of w_{ij} is proportional to p_{ij} on average.

Ackley *et al.* [2] proved that the learning rule realizes a stochastic descent of the Kullback information $I(q:p)$ between the two distributions $q(\boldsymbol{x})$ and $p(\boldsymbol{x})$. That is, on average,

$$\Delta w_{ij} = -\varepsilon\frac{\partial I(q:p)}{\partial w_{ij}}, \tag{8}$$

where

$$I(q:p) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}. \tag{9}$$

We may consider a more general situation where neurons are divided into two parts, namely visible neurons and hidden neurons. Visible neurons are divided further into input and output neurons. In the learning phase inputs are applied directly to visible neurons. Inputs are represented by a vector $\boldsymbol{x}_V = (\boldsymbol{x}_I, \boldsymbol{x}_O)$ on visible neurons, where $\boldsymbol{x}_I$ and $\boldsymbol{x}_O$ correspond to the states on the input and output neurons, respectively, and components on hidden neurons have no meaning in this phase. A visible input $\boldsymbol{x}_V$ is generated from the environmental information source. Its joint probability distribution is denoted by $q(\boldsymbol{x}_I, \boldsymbol{x}_O)$. The conditional distribution of $\boldsymbol{x}_O$ given $\boldsymbol{x}_I$ is written as

$$q(\boldsymbol{x}_O|\boldsymbol{x}_I) = \frac{q(\boldsymbol{x}_I, \boldsymbol{x}_O)}{q(\boldsymbol{x}_I)}, \tag{10}$$

where

$$q(\boldsymbol{x}_I) = \sum_{\boldsymbol{x}_O} q(\boldsymbol{x}_I, \boldsymbol{x}_O).$$

In the working or recalling phase, only input part $\boldsymbol{x}_I$ of the $\boldsymbol{x}_V$ is applied. The stochastic state transitions take place under this condition of fixed $\boldsymbol{x}_I$, so that the conditional stationary distribution $p(\boldsymbol{x}_H, \boldsymbol{x}_O|\boldsymbol{x}_I)$ is realized, where $\boldsymbol{x}_H$ denotes the

states on the hidden neurons. The distribution can be calculated from the connection weights, thresholds, and fixed $\boldsymbol{x}_I$ similar to (3).

In this more general case the Boltzmann machine is required to realize the conditional probability distribution $q(\boldsymbol{x}_O|\boldsymbol{x}_I)$ of the environment as faithfully as possible by learning. The distribution of the state in the hidden neurons is of no concern, but $p(\boldsymbol{x}_O, \boldsymbol{x}_I)$ should be as close to $q(\boldsymbol{x}_O, \boldsymbol{x}_I)$ as possible.

It can also be shown (Ackley *et al.* [2]) that the learning rule (5) gives a stochastic gradient descent of the conditional Kullback information,

$$I[q:p] = \sum q(\boldsymbol{x}_I) q(\boldsymbol{x}_O|\boldsymbol{x}_I) \log \frac{q(\boldsymbol{x}_I, \boldsymbol{x}_O)}{p(\boldsymbol{x}_I, \boldsymbol{x}_O)}, \tag{11}$$

where q_{ij} denotes the relative frequency of $x_i = x_j = 1$ under the condition that $\boldsymbol{x}_I$ and $\boldsymbol{x}_O$ be fixed and p_{ij} is the relative frequency in the restricted free run when only $\boldsymbol{x}_I$ is fixed.

III. The Manifold S of All the Probability Distributions

Let $S = \{q(\boldsymbol{x})\}$ be the set of all the probability distributions over the state space X consisting of 2^n states $\boldsymbol{x}$. Since

$$\sum_{\boldsymbol{x}} q(\boldsymbol{x}) = 1,$$

then a probability distribution is specified by $2^n - 1$ numbers $q(\boldsymbol{x}_1), q(\boldsymbol{x}_2), \cdots, q(\boldsymbol{x}_{N-1})$, where $N = 2^n$ and $\boldsymbol{x}_i$ is the ith state among the 2^N possible states. This shows that S has $2^n - 1$ degrees of freedom; hence S forms a manifold of $(2^n - 1)$ dimensions.

In fact, the additional restriction

$$0 < q(\boldsymbol{x}) < 1$$

implies S is an open simplex of $2^n - 1$ dimensions. Here we exclude the boundary on which $q(\boldsymbol{x}) = 0$ for some $\boldsymbol{x}$ so that S is an open set.

The differential geometry of manifolds of probability distributions has been studied in detail (Amari [5], Chentsov [8], Amari *et al.* [7]). It has a natural Riemannian metric given by the Fisher information matrix (Rao [21]), together with a dual pair of affine connections (Nagaoka and Amari [6], Amari [5]). Duality of affine connections is a new concept introduced in differential geometry from an information theory point of view (Amari [5], Nagaoka and Amari [6]), and has been proved to be useful in various information sciences (statistical inference, Amari [5], Amari and Kumon [10], Okamoto *et al.* [11], etc.; see also review papers by Barndorff-Nielsen, *et al.* [22] and by Kass [23]; control systems theory, Amari [12]; multiterminal information theory, Amari and Han [13], Amari [14]). It's rather surprising to see that the theory is also useful for elucidating Karmarkar's inner method in linear programming.

It has been proved that the set of all the probability distributions over a finite set forms a dually flat manifold, which enjoys a number of good properties. The present manifold S of all $q(\boldsymbol{x})$ is such a manifold. Its properties are stated in this section intuitively without proofs (see Amari [5], Amari and Han [13]).

Before showing them, let us introduce adequate coordinate systems in S by expanding the logarithm of a probability distribution $q(\boldsymbol{x})$ as

$$\log q(\boldsymbol{x}) = \sum \theta_1^{i_1} x_{i_1} + \sum \theta_2^{i_1 i_2} x_{i_1} x_{i_2} + \sum \theta_3^{i_1 i_2 i_3} x_{i_1} x_{i_2} x_{i_3} + \cdots + \theta_n^{i_1 \cdots i_n} x_{i_1} \cdots x_{i_n} - \psi, \tag{12}$$

where the indices in the coefficients $\theta_m^{i_1 \cdots i_m}$ are assumed to satisfy

$$i_1 < i_2 < \cdots < i_m.$$

This condition eliminates redundant expressions. There are $2^n - 1$ coefficients or parameters $\left(\theta_1^i, \theta_2^{i_1 i_2}, \cdots, \theta_n^{12\cdots n}\right)$ that specify one $q(\boldsymbol{x})$, and ψ is a normalizing constant depending on the parameters. Therefore, the set of θ's is a coordinate system of S.

Let us introduce a new notation: Let I denote any subset $(i_1, i_2, \cdots, i_m)$, $i_1 < i_2 < \cdots < i_m$, $1 \leqq m \leqq n$, of indices $\{1, \cdots, n\}$. The above parameters are then summarized in a $(2^n - 1)$-dimensional vector

$$\theta = (\theta^I),$$

where the index I runs over all combinations $(i_1, i_2, \cdots, i_m)$, $i_1 < i_2 \cdots i_m$, $1 \leqq m \leqq n$, and

$$\theta^I = \theta_m^{i_1 \cdots i_m}$$

for $I = (i_1, \cdots, i_m)$. Similarly, for $\boldsymbol{x}$, we define an extended input vector $\boldsymbol{X}(\boldsymbol{x})$:

$$\boldsymbol{X} = (X_I),$$

where

$$X_I = x_{i_1} x_{i_2} \cdots x_{i_m}$$

for $I = (i_1, i_2, \cdots, i_m)$. The probability distribution $q(\boldsymbol{x}; \theta)$ specified by θ is then written as

$$\log q(\boldsymbol{x}; \theta) = \sum_I \theta^I X_I - \psi(\theta) \tag{13}$$

or

$$q(\boldsymbol{x}; \theta) = \exp\left\{\sum \theta^I X_I - \psi(\theta)\right\}. \tag{14}$$

This shows that the set S of probability distributions is an exponential family, where the vector parameter θ specifying the distribution is called the natural or canonical parameter in statistics.

Let $\eta = (\eta_I)$ be the expectation of the extended vector $\boldsymbol{X}$ under the distribution $q(\boldsymbol{x}; \theta)$:

$$\eta_I = E_\theta[X_I] = \mathrm{Prob}_\theta\{x_{i_1} = x_{i_2} = \cdots = x_{i_m} = 1 | I = (i_1, \cdots, i_m)\}, \tag{15}$$

where E_θ is the expectation with respect to $q(\boldsymbol{x}; \theta)$. The vector η is called the expectation parameter in statistics, and it can be used as a parameter specifying the distribution uniquely,

because the relation between θ and η is bijective (one-to-one). They can be proved to be connected by the Legendre transformation,

$$\theta^I = \frac{\partial}{\partial \eta_I}\psi(\theta), \qquad \eta_I = \frac{\partial}{\partial \theta^I}\phi(\eta), \tag{16}$$

where $\phi(\eta)$ is defined by the identity

$$\psi(\theta) + \phi(\eta) - \sum \theta^I \eta_I = 0. \tag{17}$$

The function $\psi(\theta)$ is the cumulant generating function,

$$\psi(\theta) = \log\left(\sum_{\boldsymbol{x}} \exp\left\{\sum_I \theta^I X_I(\boldsymbol{x})\right\}\right), \tag{18}$$

and the function $\phi(\eta)$ is the negative of entropy,

$$\phi(\eta) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}; \theta(\eta)) \log q(\boldsymbol{x}; \theta(\eta)). \tag{19}$$

We can introduce two dually coupled criteria of linearity or flatness in the manifold S. To explain them, let us consider a curve $q(\boldsymbol{x}, t)$ connecting two distributions $q_1(\boldsymbol{x})$ and $q_2(\boldsymbol{x})$ (Fig. 1), where t is a scalar parameter such that

$$q(\boldsymbol{x}, 0) = q_1(\boldsymbol{x})$$
$$q(\boldsymbol{x}, 1) = q_2(\boldsymbol{x}).$$

One criterion is called the exponential criterion, which declares that a curve is e-linear when $\log q(\boldsymbol{x}, t)$ is a linear sum of $\log q_1(\boldsymbol{x})$ and $\log q_2(\boldsymbol{x})$, namely

$$\log q(\boldsymbol{x}, t) = (1-t)\log q_1(\boldsymbol{x}) + t \log q_2(x) + c_t$$

where c_t is a normalizing constant. The set $\{q(\boldsymbol{x}, t)\}$ of distributions forms an exponential family in statistics. Since the coordinates θ^I are coefficients of the logarithm of a probability distribution as shown in (12), it gives a linear coordinate system in the exponential criterion. Any curve, or submanifold of S, whose parametric representation is linear in θ is said to be e-linear or e-flat. In particular, a parametric curve $\theta = \theta(t)$ which is linear in θ,

$$\theta^I(t) = a^I t + b^I$$

is called an e-geodesic or e-straight line.

The other criterion is called the mixture criterion, which declares that a curve is a m-geodesic or a m-straight line when

$$q(\boldsymbol{x}, t) = (1-t)q_1(\boldsymbol{x}) + t q_2(\boldsymbol{x}).$$

This is because this set $\{q(\boldsymbol{x}, t)\}$ of distributions forms a mixture family in statistics. Since the coordinates η_I are linear in the distribution function $q(\boldsymbol{x})$ as is seen from (15), it gives a linear coordinate system in the mixture criterion. Any curve or submanifold is m-linear or m-flat when its parametric representation is linear in η. In particular, a curve $\eta = \eta(t)$ is m-geodesic when it is given by

$$\eta_I(t) = a_I t + b_I.$$

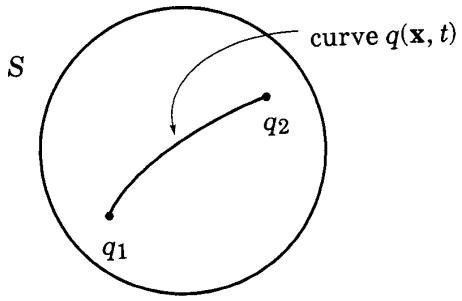

Fig. 1. A curve connecting $q_1(\boldsymbol{x})$ and $q_2(\boldsymbol{x})$.

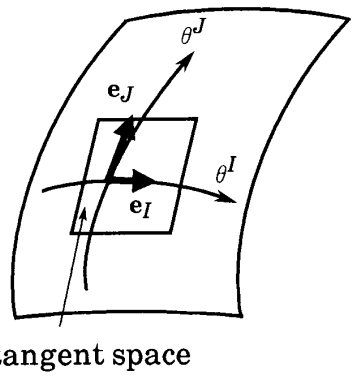

Fig. 2. Coordinate curves, tangent space, and basis vectors.

The concept of linearity or flatness should be defined first by introducing invariant affine connections in S. Two invariant affine connections can be introduced in the manifold of smooth probability distributions (see Amari [5], Amari and Han [13]). By analyzing their properties, it is found that the manifold S is special in the sense that it is dually flat in the sense that the e- and m-Riemann–Christoffel curvatures vanish. This guarantees the existence of two special linear coordinate systems, which indeed are proved to be the present θ and η. We give a brief introduction to the dual differential geometry in the Appendix.

It is also shown that there exists a unique invariant metric in S. Let $\boldsymbol{e}_I$ be the tangent vector along the coordinate curve θ^I (Fig. 2). The tangent space at θ is a linear space spanned by these tangent vectors. In other words, the set $\{\boldsymbol{e}_I\}$ forms a basis of the tangent space at each point of S. Their inner products

$$g_{IJ}(\theta) = \boldsymbol{e}_I \cdot \boldsymbol{e}_J$$

define a Riemannian metric tensor g_{IJ} which forms a matrix $g = (g_{IJ})$.

Since the tangent space is a local linearization of the manifold, g_{IJ} gives a local metric. The square of the distance between two nearby distributions $Q = q(\boldsymbol{x}, \theta)$ and $Q' = q(\boldsymbol{x}, \theta + d\theta)$ specified by θ and $\theta + d\theta$ is given by the quadratic form

$$ds^2 = d\theta \cdot d\theta = \sum g_{IJ} d\theta^I d\theta^J, \tag{20}$$

because

$$d\theta = \sum d\theta^I \boldsymbol{e}_I = \overline{QQ'}$$

is regarded as an (infinitesimally small) vector belonging to the tangent space. Two different vectors $d_1\theta$ and $d_2\theta$ are orthogonal when

$$d_1\theta \cdot d_2\theta = \sum g_{IJ} d_1\theta^I d_2\theta^J = 0.$$

What is then the metric in our manifold S? The Fisher information matrix can be proved to be the only invariant metric that can be introduced in the manifold of probability distributions (Rao [21], Chentsov [8]). It is defined by

$$g_{IJ} = E\left[\frac{\partial}{\partial\theta^I}\log q(\boldsymbol{x};\theta)\frac{\partial}{\partial\theta^J}\log q(\boldsymbol{x};\theta)\right] \tag{21}$$

in a general coordinate system, and in the present e-coordinate system, it is calculated simply by the following formula:

$$g_{IJ} = \frac{\partial^2}{\partial\theta^I\partial\theta^J}\psi(\theta). \tag{22}$$

The metric tensor g^{IJ} in the dual η-coordinate system is given by

$$g^{IJ}(\eta) = \frac{\partial^2}{\partial\eta_I\partial\eta_J}\phi. \tag{23}$$

Moreover, $\left(g^{IJ}\right)$ and (g_{IJ}) can be proved to be mutually inverse matrices,

$$\sum g^{IJ} g_{JK} = \delta^I_K,$$

where δ^I_K is the Kronecker delta (i.e., $\delta^I_K = 1$ when $I = K$ and is 0 otherwise). This implies that $\{\boldsymbol{e}_I\}$ and $\{\boldsymbol{e}^I\}$ are mutually dual or reciprocal bases of the tangent space at every point of S,

$$\boldsymbol{e}_I \cdot \boldsymbol{e}^J = \delta^J_I. \tag{24}$$

It should be noted that S is a curved Riemannian manifold from the point of view of the Riemannian metric g_{IJ}. However, it admits a pair of dually flat connections or dual criteria by which S is flat. A dually flat manifold in general admits a unique divergence measure between two points P and Q from P to Q. It is defined by

$$D(P,Q) = \psi(\theta_P) + \phi\left(\eta^Q\right) - \sum \theta^I_P \eta^Q_I, \tag{25}$$

where $\theta_P = \left(\theta^I_P\right)$ is the θ-coordinates of $P \in S$ and $\eta^Q = \left(\eta^Q_I\right)$ is the η-coordinates of $Q \in S$. It is not symmetric, but satisfies the inequality

$$D(P,Q) \geqq 0$$

with the equality if and only if $P = Q$. The divergence is a generalization of the squared distance. When two points are close, it gives half the square of the Riemannian distance,

$$D(P, P+dP) = \frac{1}{2}\sum g_{IJ}\, d\theta^I\, d\theta^J. \tag{26}$$

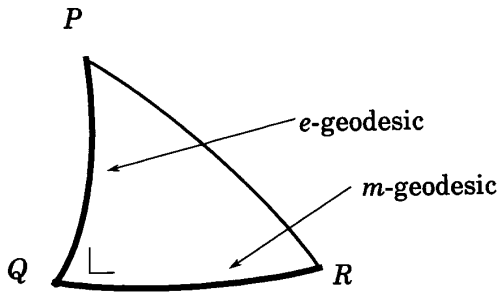

Fig. 3. Pythagorian theorem: $D(P,Q) + D(Q,R) = D(P,R)$.

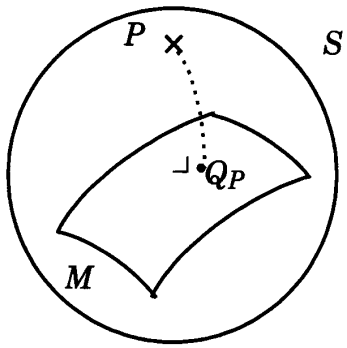

Fig. 4. Projection Q_P from P to M.

In the present manifold S of probability distributions the divergence between two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ turns out to be the Kullback information,

$$D(p(\boldsymbol{x}), q(\boldsymbol{x})) = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} = I(p:q). \tag{27}$$

The divergence satisfies the following two theorems (Nagaoka and Amari [6], Amari [5]) (Figs. 3 and 4). They show that a dually flat manifold is indeed a dualistic generalization of the Euclidean space.

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

$$D(P,Q) + D(Q,R) = D(P,R). \tag{28}$$

Projection Theorem: Let M be a smooth submanifold in S. Given a point $P \in S$, let Q_P be the point that belongs to M and is closest to P in the sense of the divergence,

$$Q_P = \text{argmin}_{Q\in M} D(P,Q) \tag{29}$$

The point Q_P is given by the dual geodesic projection of P to M. Furthermore, the projection is unique when M is e-flat.

Here, Q_P is said to be the geodesic projection of P to M, when the geodesic connecting P and $Q_P \in M$ is orthogonal to M, implying that the tangent vector of the geodesic is orthogonal to the tangent space of M at Q_P.

IV. The Boltzmann Without Hidden Units

A Boltzmann machine without hidden units realizes a probability distribution $p(\boldsymbol{x}; w)$ depending on the connection matrix $w = (w_{ij})$,

$$\log p(\boldsymbol{x}; w) = \sum_{i>j} w_{ij} x_i x_j - \psi(w). \tag{30}$$

where $\psi(w)$ is the normalization constant. Conversely, a probability distribution of the form (30) is realized by a Boltzmann machine with connections w_{ij} as its stationary distribution. Let B be the set of all the probability distributions realized by Boltzmann machines. Since each distribution is specified uniquely by the connection matrix w, B can also be regarded as the set of Boltzmann machines. We call B the Boltzmann manifold, where w plays a role of its coordinate system.

Clearly, B is a submanifold of S. From (12) and (30) the probability distributions realized by B are given in S in the θ-coordinates by

$$\begin{aligned} \theta_1^i &= w_{i0} \\ \theta_2^{ij} &= w_{ij} \\ \theta_m^{i_1 \cdots i_m} &= 0, \qquad m > 2. \end{aligned}$$

The above relations are linear in θ. This implies that B is an e-linear submanifold of S, because B is characterized by the e-linear constraints

$$\theta_m^{i_1 i_2 \cdots i_m} = 0, \qquad m = 3, \cdots, n \tag{31}$$

in the θ-coordinate system of S.

Therefore, B inherits a dually linear geometrical structure from S. Moreover, B itself forms an exponential family of probability distributions. Its e-coordinate system is w^{ij}, where we use upper indices in w^{ij} in order to emphasize that it is the e-affine coordinate system.

The dual coordinate system is given by

$$\begin{aligned} \eta_{ij} = p_{ij} = E_w[x_i x_j] = \mathrm{Prob}\{x_i = x_j = 1\}, \\ i < j, \qquad i = 0, 1, \cdots, n. \end{aligned}$$

The dual coordinates are given by the corresponding part η_i^1 and $\eta_{i_1 i_2}^2$ of the η-coordinates in S. The other parts $\eta_{i_1 \cdots i_m}^m$ are not equal to 0 but are nonlinear functions in η_{ij} when the distribution is in B. This implies that B is not an m-linear submanifold of S. However, since B is e-linear, we can introduce the dually related m-linear structure in B, such that p_{ij} give the m-linear coordinates in B.

The dually flat geometry of B elucidates the following structure of the set of Boltzmann machines.

A. Stochastic Meanings of the Connection Weights and Thresholds

The coordinates w^{ij} obviously consist of the connection weights w_{ij} and thresholds $-w^{i0}$. This is a biological interpretation. However, for an identical mathematical formulation, physicists endow w^{ij} and $-w^{i0}$ with the meaning of the exchange energy of spins and of the local magnetization; this is the spin-glass analogy of neural networks. We now have the third interpretation, the stochastic interpretation. Define the conditional probability

$$p^*(x_i = x_j = 1|A) = \mathrm{Prob}\{x_i = x_j = 1|A\},$$

where A denotes a fixed firing pattern of all the neurons other than the ith and jth neurons. This implies, for example, that $p^*(x_i = x_j = 1|A)$ is the conditional probability that both of the ith and the jth neuron fire when all the other neurons are in the fixed state specified by the condition A,

Theorem 1: The following relations explain the stochastic meaning of $-w^{i0}$ and w^{ij}:

$$-w^{i0} = \log \frac{p^*(x_i = 1|A)}{p^*(x_i = 0|A)} \tag{32}$$

and

$$w^{ij} = \log \frac{p^*(x_i = x_j = 1|A) p^*(x_i = x_j = 0|A)}{p^*(x_i = 1, x_j = 0|A) p^*(x_i = 0, x_j = 1|A)}. \tag{33}$$

The relations hold for any conditions A on the other neurons.

Proof: We prove the simplest case of when the condition A states that all the other neurons are in their quiescent state, that is, $x_k = 0$ for $k \neq i, j$. The other cases can be proved in a similar way. The probability of $x_i = x_j = 1$ and all the other x_k being equal to 0 is given, form (3) and (4), by

$$p(x_i = 1, x_j = 1; A) = \exp\{w_{ij} + w_{i0} + w_{j0} - \psi(w)\}.$$

Similarly, the probability that $x_i = 1$ but all the other x_k being equal to 0 is

$$p(x_i = 1, x_j = 0; A) = \exp\{w_{i0} - \psi(w)\}.$$

On the other hand, the probability that no neurons fire, i.e., $\boldsymbol{x} = 0$, is given by

$$p(x_i = x_j = 0; A) = \exp\{-\psi(w)\}.$$

From these relations, we have

$$\exp(w_{i0}) = \frac{\mathrm{Prob}\{x_i = 1|A\}}{\mathrm{Prob}\{x_i = 0|A\}},$$

and

$$\exp\{w^{ij}\} = \frac{\mathrm{Prob}\{x_i = x_j = 1|A\}\,\mathrm{Prob}\{x_i = x_j = 0|A\}}{\mathrm{Prob}\{x_i = 1, x_j = 0|A\}\,\mathrm{Prob}\{x_j = 1, x_i = 0|A\}}.$$

This proves the theorem.

The theorem shows that w_{ij} represents the degree of conditional dependence of the firing of the ith and jth neurons, given any fixed conditions on the states of the other neurons. When $w_{ij} = 0$, the firing of the ith and jth neurons is conditionally independent. The fact that this does not depend on the condition A implies that there are no mutual correlations of more than two neurons in the stationary distribution realized by a Boltzmann machine:

$$\exp\{w^{ij}\} = \frac{\mathrm{Prob}\{x_i = x_j = 1|A\}\,\mathrm{Prob}\{x_i = x_j = 0|A\}}{\mathrm{Prob}\{x_i = 1, x_j = 0|A\}\,\mathrm{Prob}\{x_j = 1, x_i = 0|A\}}.$$

B. The Divergence Between Two Boltzmann Machines

The divergence between two Boltzmann machines specified by w_1^{ij} and w_2^{ij} is measured invariantly by

$$D(w_1, w_2) = \sum_{\boldsymbol{x}} p(\boldsymbol{x}; \boldsymbol{w}_1) \log \frac{p(\boldsymbol{x}; \boldsymbol{w}_1)}{p(\boldsymbol{x}; \boldsymbol{w}_2)} = \psi(\boldsymbol{w}_1) + \phi(\boldsymbol{w}_2) - \sum_{i>j} w_1^{ij} p_{2ij}, \quad (34)$$

which is the Kullback information of the two distributions. In particular, when the two machines are infinitesimally close, then $\boldsymbol{w}_1 = \boldsymbol{w}$ and $\boldsymbol{w}_2 = \boldsymbol{w} + d\boldsymbol{w}$ and the divergence is half the square of the Riemannian distance,

$$D(\boldsymbol{w}, \boldsymbol{w} + d\boldsymbol{w}) = \frac{1}{2} \sum g_{ij,kl}\, dw^{ij}\, dw^{kl}.$$

Here the Riemannian metric is calculated from

$$g_{ij,kl} = \frac{\partial^2}{\partial w^{ij} \partial w^{kl}} \psi(w) = \frac{\partial p_{ij}}{\partial w^{kl}}. \quad (35)$$

Theorem 2: The Riemannian metric tensor is given by

$$g_{ij,kl} = p_{ijkl} - p_{ij} p_{kl}, \qquad i < j, \qquad k < l \quad (36)$$

where

$$p_{ijkl} = \text{Prob}\{x_i = x_j = x_k = x_l = 1\}.$$

Proof: We prove the case where all the indices i, j, k, and l are different. Other cases can be proved similarly, where p_{ijjk}, for example, is interpreted as

$$p_{ijjk} = \text{Prob}\{x_i = x_j = x_k = 1\}.$$

We have

$$p_{ij} = \sum_{\boldsymbol{x}} x_i x_j \exp\Big\{\sum w^{st} x_s x_t - \psi(w)\Big\}.$$

By differentiating it with respect to w^{kl}, we easily have

$$\frac{\partial p_{ij}}{\partial w^{kl}} = \sum_{\boldsymbol{x}} x_i x_j (x_k x_l - p_{kl}) \exp\Big\{\sum w^{st} x_s x_t - \psi(w)\Big\}$$

because of

$$\frac{\partial \psi}{\partial w^{kl}} = p_{kl}.$$

This completes the proof.

C. The Statistical Meaning of the Metric Tensor

The metric tensor $g_{ij,kl}$ is the Fisher information matrix, which represents the amount of information in one observed state $\boldsymbol{x}$ that can be used to estimate the structural parameters w^{ij} or p_{ij}. Indeed, let $\hat{p}_{ij}$ be the observed frequency of the joint firing of the ith and jth neurons when the Boltzmann machine is running freely and let $\hat{w}^{ij}$ be the maximum likelihood estimator of w^{ij} based on $\hat{p}_{ij}$. Then it can be shown that the estimation error is given by the following theorem.

Theorem 3: The expected squared error (covariance matrix of the estimator) is asymptotically given by

$$E\big[(\hat{w}^{ij} - w^{ij})(\hat{w}^{kl} - w^{kl})\big] = \frac{1}{N} g^{ij,kl}, \quad (37)$$

where N is the number of observations and $\big(g^{ij,kl}\big)$ is the inverse matrix of $(g_{ij,kl})$.

V. Projection and Learning

Given an environmental probability distribution $q(\boldsymbol{x})$ we consider the problem of realizing it by a Boltzmann machine as faithfully as possible. Here the divergence is used as the criterion of approximation. Since B is an e-flat submanifold of S, then the best approximation $p(\boldsymbol{x}) \in B$ is obtained by the m-projection of q onto B,

$$p(\boldsymbol{x}) = \prod_B q(\boldsymbol{x}), \quad (38)$$

where $\prod_B$ denotes the m-projection onto B. Since B is e-flat, the projection is unique. In other words there are no local minima other than the global minimum in the divergence function.

Let the θ- and η-coordinates of the given $q(\boldsymbol{x})$ be

$$\theta^I = \Big(w^{ij}, w_3^{ijk}, \cdots, w_n^{1\cdots n}\Big) \quad (39)$$

$$\eta_I = \big(p_{ij}, p_{ijk}^3, \cdots, p_{1\cdots n}^n\big) \quad (40)$$

respectively. Note that w^{ij} here stand for w_1^{0i} and w_2^{ij} and that p_{ij} stands for p_{0i}^1 and p_{ij}^2. In order to obtain an explicit expression of the m-projection, we use the following mixed coordinates:

$$\xi = \big(p_{ij}; w^{*I}\big), \quad (41)$$

where the first part of ξ consists of the first part p_{ij} of the η-coordinates and the latter part, w^{*I}, consists of that of the θ-coordinates $w_m^{i_1 \cdots i_m}$ $(m \geqq 3)$.

Theorem 4: The mixed coordinates of the projection $P = \prod_B Q$ of Q are given by

$$\xi_P = (p_{ij}; 0). \quad (42)$$

Proof: Since the w^{*I} part of the θ-coordinates of the above ξ_P are zero, then P is in B. The η-coordinates of P are

$$(p_{ij}; \eta_I^*),$$

so that both P and Q are included in the submanifold

$$M = \{\eta | \text{the first part of the } \eta\text{-coordinates is fixed to be equal to } p_{ij}\}.$$

This manifold M is an m-flat manifold, because it is defined by linear constraints in the η-coordinates. Therefore, the m-geodesic connection P and Q is in M.

The tangent space of M at P is spanned by the vectors $\{e^J\}$, where J denotes the latter part of the η-coordinates. This is because M is defined by fixing the former part of the η-coordinates to be equal to p_{ij}, while the latter part is free.

On the other hand, the tangent space of B at P is spanned by $\{\boldsymbol{e}_I\}$, where I denotes the former part of the θ-coordinates. This is because B is defined by fixing the latter part of the θ-coordinates equal to 0 while the former part is free. Because of the reciprocity of the two coordinate systems,

$$\boldsymbol{e}_I \cdot \boldsymbol{e}^J = 0. \qquad I \neq J.$$

and hence M and B are orthogonal at P. Hence, the m-geodesic connecting P and Q is orthogonal to B, proving that the m-projection of Q onto B is P.

The connection weights w^{ij} of P can be calculated in practice by solving

$$p_{ij} = \frac{\partial}{\partial w^{ij}} \psi\left(w^{ij}; 0\right).$$

These weights can also be obtained by learning as follows. Let R be the current state of a Boltzmann machine which is being adapted to an environment Q. Let P be the m-projection of Q. Since M and B are orthogonal, the Pythagorian theorem gives

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

This shows that the error $D(Q, R)$ of the current machine R is decomposed into two terms: $D(P, R)$, denoting the divergence between R and the optimal machine P, and $D(Q, P)$, denoting the unavoidable divergence since Q does not belong to B. The latter, $D(Q, P)$, does not depend on the current R.

Let w^{ij} be the connection weights of R. In order to improve R we calculate the gradient of $D(Q, R)$. It is easily calculated from (34) to be

$$\frac{\partial D(Q, R)}{\partial w^{ij}} = \frac{\partial D(P, R)}{\partial w^{ij}} = p_{ij}^Q - p_{ij}^R.$$

Therefore, the gradient learning method is given by

$$\Delta w^{ij} = \varepsilon\left(p_{ij}^R - p_{ij}^Q\right). \tag{43}$$

However, this does not have an invariant meaning. The true steepest descent method in a Riemannian manifold should be

$$\Delta w^{ij} = \varepsilon \sum g^{ij,kl}\left(p_{kl}^R - p_{kl}^Q\right), \tag{44}$$

where $\left(g^{ij,kl}\right)$ is the inverse of the metric tensor. The merit of the new learning rule is given by the following theorem.

Theorem 5: The trajectory of learning (44) is the m-geodesic connecting the present point R and the optimal point P in B.

Proof: Since the corresponding change in p_{ij}^R is rewritten as $\Delta p_{ij}^R = \sum g_{ij,kl} \Delta w^{kl}$, then the learning equation gives an m-geodesic equation in B.

The revised learning algorithm gives an m-geodesic trajectory, which is a linear equation in the η-coordinates. This suggests that the convergence speed is much higher and that a large learning constant ε can be used to accelerate the convergence. These effects are remarkable as the present machine R lies close to the boundary of B, as shown by computer simulation.

The connection weights w^{ij} sometimes diverge to infinity by learning. Therefore, we need to study when divergence occurs. We first prove the following lemma.

Lemma: The e-coordinates θ^I diverge to infinity at the boundary of S. The e-coordinates w^{ij} diverge to infinity at the boundary of B. Otherwise they are finite.

Proof: The e-coordinates θ are given by the gradient of the negative of the entropy function defined on S and B. The gradients of the entropy functions diverge at the boundary of S and B. This proves the lemma.

Since S is a $(2^n - 1)$-dimensional simplex, its boundary satisfies

$$p(\boldsymbol{x}) = 0$$

for some $\boldsymbol{x}$. When the number of such $\boldsymbol{x}$ is 1 we have a $(2^n - 2)$-dimensional face of the closed simplex $\overline{S}$, and when the number is k, we have a $(2^n - k - 1)$-dimensional face of $\overline{S}$. A constraint ($p(\boldsymbol{x}) = 0$) can be rewritten in terms of η_I as

$$\sum_I C^I(\boldsymbol{x}) \eta_I = 0, \tag{45}$$

where $C^I(\boldsymbol{x})$ is defined by the inclusion–exclusion relation.

By combining some of the constraints, we have constraints of the form

$$\sum D_\alpha^{ij} p_{ij} = 0, \tag{46}$$

which are written only in terms of p_{ij}'s. They define the boundary of B, which is included in the boundary of S, that is,

$$\partial B \subset \partial S.$$

The constraints include the following:

$$1 - \sum_{i \in Q} p_i + \sum_{i,j \in Q} p_{ij} = 0 \tag{47}$$

where Q is any subset of indices $\{1, \cdots, n\}$.

Theorem 6: An inner point of S is mapped to an inner point of B by the m-projection. Some of the boundary points in S are mapped to inner points of B.

VI. Boltzmann Machines with Hidden Units

The manifold of Boltzmann machines with hidden units is treated in this section. Let $\boldsymbol{x}_V = (x_{V,i})$ be the state of visible units, and let $\boldsymbol{x}_H = (x_{H,i})$ be the state of hidden units, where the entire state of a Boltzmann machine is represented by $\boldsymbol{x} = (\boldsymbol{x}_V, \boldsymbol{x}_H)$. Let $q(\boldsymbol{x}_V)$ be the probability distribution of an environmental information source, which the Boltzmann machine is to adopt to by receiving inputs.

The "energy" of the state $\boldsymbol{x} = (\boldsymbol{X}_V, \boldsymbol{x}_H)$ in a Boltzmann machine is given by

$$E(\boldsymbol{x}_V, \boldsymbol{x}_H) = -\{w_V \cdot X_V + w_H \cdot X_H + w_{VH} \cdot X_{VH}\}, \tag{48}$$

where the following notation is used:

$$w_V \cdot X_V = \sum_{i>j} w_V^{ij} x_i^V x_j^V$$

$$w_H \cdot X_H = \sum_{i>j} w_H^{ij} x_i^H x_j^H$$

$$w_{VH} \cdot X_{VH} = \sum w_{VH}^{ij} x_i^V x_j^H.$$

Here w_V^{ij}, w_H^{ij}, and w_{VH}^{ij} are connection weights among the visible units, among the hidden units, and between the visible and hidden units, respectively. The thresholds are included in the terms w_V^{i0} and w_H^{i0}. The components of $\boldsymbol{x}_V$ and $\boldsymbol{x}_H$ are written as x_i^V and x_i^H, respectively.

The stationary distribution of a Boltzmann machine is written

$$p(\boldsymbol{x}_V, \boldsymbol{x}_H; w) = \exp\{-E(\boldsymbol{x}_V, \boldsymbol{x}_H) - \psi(w)\}, \quad (49)$$

where potential function ψ is a function of $w = (w_V, w_H, w_{VH})$.

The problem is to approximate a given $q(\boldsymbol{x}_V)$ by $p(\boldsymbol{x}_V; w)$ realizable by a Boltzmann machine. The set of probability distributions $\{p(\boldsymbol{x}_V; w)\}$ parameterized by w is a curved submanifold in the set S_V of all the probability distributions on the visible units. Therefore, it does not have the good properties which are possessed by the manifold B of Boltzmann machines with no hidden units. The best approximation is given by the Boltzmann machine with hidden units that minimizes the divergence

$$D[q(\boldsymbol{x}_V), p(\boldsymbol{x}_V; w)]. \quad (50)$$

This is given by the m-projection of $q \in S_V$ to B_V. However, the projection is not unique, implying the existence of a number of local minima. Also it can be shown [2] that the gradient of the divergence is again given by

$$\frac{\partial D}{\partial w_{ij}} = -(q_{ij} - p_{ij}). \quad (51)$$

Since a Boltzmann machine produces a stationary distribution over the visible and hidden elements, it is more helpful to treat the problem in the manifold $S_{V,H}$ of the distributions over $\boldsymbol{x} = (\boldsymbol{x}_V, \boldsymbol{x}_H)$ realized by Boltzmann machines with hidden units. The submanifold $S_{V,H}$ forms an e-flat submanifold of S as before. However, the environment specifies only the marginal distribution $q(\boldsymbol{x}_V)$ on the visible elements, and we do not care about the distribution on the hidden elements.

Given $q(\boldsymbol{x}_V)$, let E_q be the set of probability distributions in $S_{V,H}$, that have the same marginal distribution $q(\boldsymbol{x}_V)$ on the visible units, that is,

$$E_q = \left\{ q(\boldsymbol{x}_V, \boldsymbol{x}_H) \;\middle|\; \sum_{\boldsymbol{x}_H} q(\boldsymbol{x}_V, \boldsymbol{x}_H) = q(\boldsymbol{x}_V) \right\}. \quad (52)$$

It is easy to prove that E_q forms an m-flat submanifold in S. Therefore, the problem is restated as a problem in S, that is; search for a Boltzmann machine in B that minimizes the divergence from E_q to B,

$$D(E_q, B) = \min_{q \in E_q, p \in B} D(q, p). \quad (53)$$

We have the following theorem in connection with this restated problem.

Theorem 7: The minimum divergence $D(E_q, B)$ in the entire manifold is equal to the minimum divergence $D[q(\boldsymbol{x}_V), B_V]$ in the visible manifold.

Proof: By using the conditional distributions and marginal distributions, we have the following relation:

$$\begin{aligned} &D[q(\boldsymbol{x}_V, \boldsymbol{x}_H), p(\boldsymbol{x}_V, \boldsymbol{x}_H)] \\ &\quad = D[q(\boldsymbol{x}_V)q(\boldsymbol{x}_H|\boldsymbol{x}_V), p(\boldsymbol{x}_V)p(\boldsymbol{x}_H|\boldsymbol{x}_V)] \\ &\quad = E_q\left[\log \frac{q(\boldsymbol{x}_V)}{p(\boldsymbol{x}_V)} + \log \frac{q(\boldsymbol{x}_H|\boldsymbol{x}_V)}{p(\boldsymbol{x}_H|\boldsymbol{x}_V)}\right] \\ &\quad = D[q(\boldsymbol{x}_V), p(\boldsymbol{x}_V)] \\ &\qquad + E_q(\boldsymbol{x}_V) D[q(\boldsymbol{x}_H|\boldsymbol{x}_V), p(\boldsymbol{x}_H|\boldsymbol{x}_V)]. \end{aligned} \quad (54)$$

From this we have

$$\begin{aligned} D(E_q, B) &= \min_{p \in B, q \in E_q} D(q, p) \\ &= \min_{w, q(\boldsymbol{x}_H|\boldsymbol{x}_V)} \{D[q(\boldsymbol{x}_V), p(\boldsymbol{x}_V)] \\ &\qquad + E_q D[q(\boldsymbol{x}_H|\boldsymbol{x}_V), p(\boldsymbol{x}_H|\boldsymbol{x}_V)]\} \\ &= \min_w D[q(\boldsymbol{x}_V), p(\boldsymbol{x}_V)] = D[q, B_V]. \end{aligned}$$

Given $q(\boldsymbol{x}_V, \boldsymbol{x}_H)$, its best approximation is given by the m-geodesic projection of q to B, denoted by $\prod_{Bq}$. On the other hand, given $p(\boldsymbol{x}_V, \boldsymbol{x}_H)$, the distribution $q^*(\boldsymbol{x}_V, \boldsymbol{x}_H)$ in E_q that minimizes $D[q(\boldsymbol{x}_V, \boldsymbol{x}_H), p(\boldsymbol{x}_V, \boldsymbol{x}_H)]$ is given by the e-geodesic projection of p to E_q,

$$q^*(\boldsymbol{x}_V, \boldsymbol{x}_H) = \prod_E p(\boldsymbol{x}_V, \boldsymbol{x}_H). \quad (55)$$

The projection is unique, because E_q is m-flat, and is given by

$$q^*(\boldsymbol{x}_V, \boldsymbol{x}_H) = q(\boldsymbol{x}_V) p(\boldsymbol{x}_H|\boldsymbol{x}_V), \quad (56)$$

which follows easily from relation (54).

This suggests the following iterative algorithm for obtaining the best approximation $p(\boldsymbol{x}_V, \boldsymbol{x}_H; w^*) \in B$ to a given $q(\boldsymbol{x}_V)$: Let the initial Boltzmann machine be $p(\boldsymbol{x}_V, \boldsymbol{x}_H; w_0)$ and put

$$p_0(\boldsymbol{x}_V, \boldsymbol{x}_H) = p(\boldsymbol{x}_V, \boldsymbol{x}_H; w_0).$$

For $i = 1, 2, \cdots$,

1) Put $q_{i+1}(\boldsymbol{x}_V, \boldsymbol{x}_H) = \prod_E p_i(\boldsymbol{x}_V, \boldsymbol{x}_H; w_i)$.
2) Put $p_{i+1}(\boldsymbol{x}_V, \boldsymbol{x}_H; w_{i+1}) = \prod_B q_{i+1}(\boldsymbol{x}_V, \boldsymbol{x}_H)$.

Here, $\prod_E p$ denotes the e-geodesic projection of p to E_q, and $\prod_B q$ denotes the m-geodesic projection of q to B.

Theorem 8: The monotonic relation

$$D[q_i, p_i] \geqq D[q_{i+1}, p_i] \geqq D[q_{i+1}, p_{i+1}]$$

holds, where the equality holds only for the fixed points of the projections; that is,

$$\prod_E p^* = q^* \qquad \prod_B q^* = p^*. \quad (57)$$

VII. Conclusions

Differential geometry provides a new method for treating the set of all neural networks of a fixed topology as a whole. Information geometry, which originates from the intrinsic properties of a smooth family of probability distributions, is also appropriate to the study of the manifold of Boltzmann machines. This is because a Boltzmann machine is identified with its stationary probability distribution on the state space.

The present paper introduces the method of information geometry to neural network research. The manifold of simple Boltzmann machines with no hidden units is proved to be e-flat and m-flat, so that it possesses nice properties. The projection and approximation theorems are proved by using the geometrical structures. The theorems are extended to the manifold of Boltzmann machines with hidden units.

The present paper, together with Amari [20], is a first step in constructing a geometrical theory of manifolds of neural networks,[1] and furthers the geometrical theory of nonlinear systems.

APPENDIX

A. Riemannian Manifold

A set M is said to be a manifold of dimension n when, roughly speaking, it is locally topologically isomorphic to the n-dimensional Euclidean space. A point of M is specified by coordinates $\xi = (\xi^1, \xi^2, \cdots, \xi^n)$ at least locally, and we discuss a part covered by this coordinate system. There are an infinite number of admissible coordinate systems in M but geometrical structures of M are invariant under whatever coordinate system we use.

A tangent space T_ξ of M at point ξ is a linear space spanned by n vectors $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n$, which are tangent vectors along coordinate axes ξ^i. (Note that $\boldsymbol{e}_i$ is denoted by $\partial/\partial\xi^i$ in a standard textbook of differential geometry and is called the natural basis vector of T_ξ.) Therefore, T_ξ can be regarded as the local linearization of M at point ξ, although mathematicians are reluctant in stating in this way.

A manifold is said to be Riemannian when it has locally a Euclidian metric structure, that is, its tangent space has an inner product. Let

$$g_{ij}(\xi) = \boldsymbol{e}_i \cdot \boldsymbol{e}_j$$

be the inner product of $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ in T_ξ. The matrix $g = (g_{ij})$ determines the inner product of any vectors in T_ξ, and is called the metric tensor. Obviously, $g_{ij}(\xi)$ changes as position ξ changes. A Euclidean space is such a special manifold that $g_{ij}(\xi)$ becomes the identity matrix $\delta = (\delta_{ij})$ at any ξ by taking an adequate coordinate system (which is called the orthogonal Cartesian system). When the manifold itself is "curved," no such coordinate system exists.

B. Affine Connections

A Riemannian manifold is covered by tangent spaces T_ξ attached at all the points $\xi \in M$. In order to recover the entire structure from the set of linear approximations T_ξ, we need to connect neighboring T_ξ's. This can be done by defining a correspondence between vectors belonging to different nearby tangent spaces. More definitely, this can be done by defining how a vector field $\boldsymbol{e}_j(\xi)$ defined at each ξ changes "intrinsically" as point ξ moves. This change is represented by the covariant derivative

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

which is a vector showing how $\boldsymbol{e}_j(\xi)$ changes "intrinsically" as point ξ moves in the direction of the coordinate axis ξ^i (that is, in the direction of $\boldsymbol{e}_i$).

Since this is a vector, its components are given by

$$\Gamma_{ijk}(\xi) = (\nabla_{\boldsymbol{e}_i} \boldsymbol{e}_j) \cdot \boldsymbol{e}_k$$

which is a quantity having n^3 components and represents the components of the affine connection. In other words, when we define an affine connection, we can recover and understand an entire curved structure of the original manifold M.

When an affine connection is defined, we have a geodesic which is a curve whose tangent direction never changes as its position changes. Let $\xi(t)$ be a curve and let $\dot{\xi}(t)$ be its tangent vector. Then, the geodesic (direction-invariant curve) is defined by the differential equation

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

or

$$\sum_i g_{ki} \frac{d^2\xi^i}{dt^2} + \sum_{i,j} \Gamma_{kij} \frac{d\xi^i}{dt} \frac{d\xi^j}{dt} = 0.$$

There is a very natural affine connection called the Levit–Civita or Riemannian connection, defined by

$$\Gamma_{ijk} = \frac{1}{2} (\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}),$$

where $\partial_i = \partial/\partial\xi^i$. It is interesting to note that a geodesic is not only a direction-invariant line but is also the shortest line when the Levi–Civita connection is used. This nice coincidence never occurs if another connection is introduced.

C. Parallel Transport of Vectors

When an affine connection is defined, we can transport a vector $\boldsymbol{A} = \sum A^i \boldsymbol{e}_i$ belonging to T_ξ "parallelly" to another tangent space $T_{\xi'}$ along a curve $\xi(t)$ connecting ξ and ξ'. This is because we have correspondences between the basis vectors $\boldsymbol{e}_i$ along the curve. This can indeed be done by solving the differential equation

$$\nabla_{\dot{\xi}} \boldsymbol{A}(t) = 0.$$

where $\boldsymbol{A}(0) = \boldsymbol{A} \in T_\xi$ and

$$\boldsymbol{A}(1) = \prod_{\xi(t)} \boldsymbol{A} \in T_{\xi'}$$

is the $\boldsymbol{A}$ transported in parallel fashion along $\xi(t)$.

The parallel transport operation in general depends on the path $\xi(t)$ along which a vector is transported from ξ to ξ'. When, and only when, the Riemann–Chirstoffel curvature calculated from the affine connection vanishes, the parallel transport does not depend on the path connecting the start and destination points. The manifold is said to be flat in this case.

[1] *Note Added in Proof:* A paper entitled "Alternating Minimization and Boltzmann Machine Learning," by W. Byrne (to be published in this journal), has come to the authors' attention. This paper is closely related to Section VI of the present paper. It should be remarked that the present geometry of the manifold of neural networks has little to do with the geometric analysis of the behavior of a single neural network by A. Pellionisz.

When a manifold is flat, we have an affine coordinate system $\theta = (\theta^i)$ such that

$$\Gamma_{ijk}(\theta) = 0$$

in this coordinate system. This implies that the coordinate curves are geodesics and a linear curve in θ is also a geodesic. It should be noted that a geodesic is not necessarily a minimum length line unless we use the Levi–Civita connection.

D. Duall Affine Connections

When the parallel transport does not change the inner product, that is, for two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ in T_ξ, and their parallel transports $\prod \boldsymbol{A}$ and $\prod \boldsymbol{B}$ along any curve satisfy

$$\boldsymbol{A} \cdot \boldsymbol{B} = \prod \boldsymbol{A} \cdot \prod \boldsymbol{B},$$

the affine connection is said to be metric. The Levi–Civita connection is the only metric connection without torsion.

Assume that there are two different connections in M, and let $\prod$ and $\prod^*$ be the two corresponding parallel transports. When the inner product is kept invariant in the sense that

$$\boldsymbol{A} \cdot \boldsymbol{B} = \prod \boldsymbol{A} \cdot \prod^* \boldsymbol{B},$$

we say that the two affine connections are dually coupled with respect to the metric. The Levi–Civita connection is the self-dual connection satisfying $\prod = \prod^*$, that is, the connection dual to itself. The concept of dual connections originates from the structure of statistical inference [5], and mathematicians are now studying its consequences.

When one dual connection is curvature free, that is, flat, the other is automatically flat. Such a manifold is said to be dually flat. A dually flat Riemannian manifold possesses a number of beautiful properties. There exist two dually coupled affine coordinate systems θ and η connected to each other by the Legendre transformation. Moreover, an invariant divergence measure is defined in M such that the generalized Phythagorian theorem and the generalized projection theorem hold.

E. Manifold of Probability Distributions

Let x be a random variable whose probability distribution is specified by $p(x;\xi)$, where ξ is an n-dimensional parameter. The set $M = \{p(x;\xi)\}$ of all such distributions forms an n-dimensional manifold, where ξ plays the role of a coordinate system. What is the natural geometry to be introduced in M? We pose two invariancy principles such that the geometrical structure itself should be invariant under the parameter ξ we use to specify the distributions and under the nonlinear measure (or naming) we use to specify x.

It is proved that the Fisher information

$$g_{ij}(\xi) = E[\partial_i l(x,\xi)\partial_j l(x,\xi)]$$

is the only invariant metric, where E denotes the expectation with respect to $p(x,\xi)$, $l(x,\xi) = \log p(x,\xi)$, and

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

On the other hand, we have a family of invariant connections parameterized by a scalar a, which are given by

$$\Gamma^{(\alpha)}_{ijk} = E\left[\left\{\partial_i\partial_j l(x,\xi) + \frac{1-\alpha}{2}\partial_i l(x,\xi)\partial_j l(x,\xi)\right\} \cdot \partial_k l(x,\xi)\right].$$

This is called the α-connection. In particular, the $\alpha = 1$ connection is called the exponential or e-connection, and the $\alpha = -1$ connection is called the mixture or m-connection.

The important fact is that α- and $-\alpha$-connections are dual with respect to the Fisher information metric. A wide class of important probability distributions is proved to be dually flat. This leads to a wide area of applications.

ACKNOWLEDGMENT

The authors would like to thank Dr. K. Judd not only for his help with the English but also for valuable comments on the manuscript.

REFERENCES

[1] S. Amari, N. Fujita, and S. Shinomoto, "Four types of learning curves," *Neural Computation,* to appear.
[2] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science,* vol. 9, pp. 147–169, 1985.
[3] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines.* New York: Wiley, 1989.
[4] S. Amari, "Differential geometry of curved exponential families—Curvatures and information loss," *Annals of Statistics,* vol. 10, no. 2, pp. 357–385, 1982.
[5] S. Amari, *Differential-Geometrical Methods in Statistics* (Springer Lecture Notes in Statistics vol. 28). New York: Springer, 1985.
[6] H. Nagaoka and S. Amari, "Differential geometry of smooth families of probability distributions," *METR 82-7,* Univ. Tokyo, 1982.
[7] S. Amari, O. E. Barndorff-Hielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, *Differential Geometry in Statistical Inferences* (IMS Lecture Notes Monograph Series, vol. 10). IMS CA: Hayward, 1987.
[8] N .N. Chentsov, *Optimal Decission Rules and Optimal Inference* (in Russian). Moscow: Nauka, 1972. Translated into English, AMS: Rhode Island, 1982.
[9] M. Kumon and S. Amari, "Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference," in *Proc. R. Soc. London,* A387, pp. 429–458, 1983.
[10] S. Amari and M. Kumon, "Estimation in the presence of infinitely many nuisance parameters—Geometry of estimating functions," *Annals of Statistics,* vol. 16, pp. 1044–1068, 1988.
[11] I. Okamoto, S. Amari and K. Takeuchi, "Asymptotic theory of sequential estimation: Differential geometrical approach," *Annals of Statistics,* vol. 19, pp. 961–981, 1991.
[12] S. Amari "Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections and divergence," *Mathematical Systems Theory,* vol. 20, pp. 53–82, 1987.
[13] S. Amari and T. S. Han, "Statistical inference under multi-terminal rate restrictions—A differential geometrical approach," *IEEE Trans. Inform. Theory,* vol. IT-35, pp. 217–227, 1989.
[14] S. Amari, "Fisher information under restriction of Shannon information in multiterminal situations," *Ann. Inst. Statist. Math.* 41, no. 4, pp. 623–648, 1989.
[15] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayestan restoration of images," *IEEE Trans., Pattern Anal. Machine Intell.,* vol. PAMI-6, pp. 721–741, 1984.
[16] A. Amit, *Modeling Brain Function.* Cambridge University Press, 1989.
[17] E. Levin, N. Tishby and S. A. Solla, "A statistical approach to learning and generalization in layered neural networks," *Proc. IEEE,* vol. 78, pp. 1568–1574, 1990.
[18] H. Seung and H. E. Sompolinski, "Statistical mechanics of learning from examples," *Phys. Rev. A,* to appear.
[19] S. Amari, "Mathematical foundations of neurocomputing," *Proc. IEEE,* vol. 78, pp. 1443–1463, 1990.

[20] S. Amari, "Dualistic geometry of the manifold of higher-order neurons," *Neural Networks,* vol. 4, pp. 443–451, 1991.
[21] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81–91, 1945.
[22] O. Barndorf-Nielsen, D. R. Cox, and N. Reid, "The role of differential geometry in statistical theory," *Int. Statist. Rev.*, vol. 54, pp. 83–96, 1986.
[23] R. E. Kass, "The geometry of asymptotic inference," *Statistical Science,* vol. 4, pp. 188–234, 1989.

Shun-ichi Amari (M'88) 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,* and a member of the editorial or advisory editorial boards of more than 15 international journals.

Koji Kurata was born in Japan in 1958. He received the B.S., M.S., and Ph.D. degrees in mathematical engineering from the University of Tokyo, Japan, in 1981, 1983, and 1990, respectively.

From 1984 to 1990 he was a Research Associate at the University of Tokyo. He has been an Assistant Professor at Osaka University since 1990. His research interests include theoretical aspects of neural networks, especially those of topological mapping organization. He is also interested in pattern formations in homogeneous fields.

Hiroshi Nagaoka (M'88) was born in Tokyo in 1955. He received the B.Eng. and M.Eng. degrees in 1980 and 1982, respectively, from the University of Tokyo and the Dr.Eng degree from Osaka University in 1987.

He was a lecturer at the Tokyo Engineering University from 1986 to 1989 and is currently a lecturer at Hokkaido University. His research interests include the differential geometry of statistical inference, quantum state estimation, linear system theory, and neural networks.