Information geometry of rotor Boltzmann machines

Masaki Kobayashi

1 University of Yamanashi
4-3-11 Takeda, Kofu, Yamanashi 400-8511, Japan

a) k-masaki@yamanashi.ac.jp

Received September 26, 2015; Revised November 15, 2015; Published April 1, 2016

Abstract: A complex-valued Hopfield neural network is a useful model for processing multi-level data. A rotor Hopfield network is an extension of a complex-valued Hopfield neural network but much more flexible. In addition, a rotor Hopfield neural network has excellent storage capacity and noise robustness characteristics. In the present work, we investigate the rotor Boltzmann machine (RoBM), a stochastic model of a rotor Hopfield neural network, through information geometry, which is a useful tool for analyzing stochastic models. We discuss RoBM through concepts of information geometry, such as the Fisher metric, parameters and potential functions. Moreover, we provide natural gradient descent learning and em-algorithms for RoBM as applications of information geometry.

Key Words: complex-valued neural networks, rotor Boltzmann machine, information geometry, em-algorithm

1. Introduction

A Boltzmann machine is a stochastic model of a Hopfield neural network (HNN) that has symmetric recurrent connections [1, 8, 9]. HNNs have been expanded to some advanced models such as complex-valued Hopfield neural networks (CHNN) and rotor Hopfield neural networks (RHNNs) [7, 10, 19–21]. Their stochastic models have been studied as well [22, 27, 28].

CHNNs are constructed based on complex numbers and are multi-valued models, unlike ordinary HNNs. They have often been applied to processing multi-level data such as grayscale image data [5, 6, 26]. An RHNN is a high-dimensional model of HNNs. In particular, a two-dimensional RHNN includes a CHNN and has been applied to processing multi-level data instead of CHNNs [11, 12, 16, 25]. A CHNN is highly restricted by rotation invariance when rotated patterns of training patterns are stable. Rotation invariance causes too many spurious states and reduces the noise robustness of CHNNs [11, 13, 14]. Two-dimensional RHNNs can avoid rotation invariance and improve noise robustness [12, 15, 16]. Especially, chaotic associative memories by CHNNs have serious problems as a result of rotation invariance. In the recall process of chaotic associative memories by CHNNs, many spurious patterns appear. On the other hand, chaotic associative memories by RHNNs do not recall rotated patterns other than reversed patterns [16]. Moreover, chaotic associative memories by RHNNs with periodic activation functions can avoid recalling reversed patterns [25].

Complex-valued Boltzmann machines (CBMs) and Rotor Boltzmann machines (RoBMs) are stochas-
tic models of CHNNs and RoBM, respectively. CBMs are restricted by rotation invariance. Rotation invariance forces CBMs to have the same distribution for all rotated patterns [27, 28]. This is a severe restriction. On the other hand, two-dimensional RoBMs have twice as many parameters as CBMs and do not suffer from rotation invariance. Thus RoBMs realize more flexible distributions than CBMs.

Amari has investigated Boltzmann machines through information geometry [2, 3]. Information geometry is an excellent method for analyzing stochastic models [4]. He revealed that Boltzmann machines form an exponential family. The Fisher metric is a Riemannian metric of a stochastic manifold and determines the distance between two distributions. Amari determined the Fisher metric of Boltzmann manifolds, which consist of Boltzmann machines. The natural and expectation parameters are useful parameters of the Boltzmann manifold, and they form dual coordinate systems. There exists a Legendre transformation between natural and expectation parameters. Their potential functions were also determined. He proposed natural gradient descent learning for Boltzmann machines, as well as the em-algorithm for Boltzmann machines with hidden neurons. Kobayashi has investigated complex-valued Boltzmann machines, which are stochastic complex-valued Hopfield neural networks, through information geometry [17, 18]. In particular, he proposed quantized CBMs and investigated along the same line of research as Amari [3].

In the present study, we investigate the information geometry of RoBMs along the line of research by Amari [3]. First, we reveal that RoBMs form an exponential family and determine the Fisher metric, natural and expectation coordinate systems, and the potential functions of RoBMs. Next, we propose quantized two-dimensional RoBMs (QTRoBMs) and investigate the information geometry of QTRoBMs. We discuss Boltzmann learning and natural gradient descent learning for RoBMs. Finally, we investigate QTRoBM with hidden neurons and present the em-algorithm.

2. Rotor Boltzmann machines

A rotor neuron is a high-dimensional neuron model and an RoBM is a high-dimensional model of Boltzmann machines. A two-dimensional RoBM is an extension of CBMs.

Let \( S^N \) be the set of \( N \)-dimensional vectors whose lengths are one. The set of rotor neuron states is \( S^N \). An input to a rotor neuron is an \( N \)-dimensional vector. For an input \( \mathbf{I} \), the probability \( p(\mathbf{x}; \mathbf{I}) \) that the rotor neuron state is \( \mathbf{x} \in S^N \) is defined as

\[
p(\mathbf{x}; \mathbf{I}) = \frac{\exp(\mathbf{x} \cdot \mathbf{I})}{\int \exp(\mathbf{x} \cdot \mathbf{I}) d\mathbf{x}},
\]

where \( \mathbf{x} \cdot \mathbf{I} \) is the inner product of the vectors \( \mathbf{x} \) and \( \mathbf{I} \). In other words, \( p(\mathbf{x}; \mathbf{I}) \) is proportional to \( \exp(\mathbf{x} \cdot \mathbf{I}) \).

Next, we describe the connection weights of RoBM. Connection weights are \( N \times N \) matrices. Let \( W_{ij} \) be the connection weight from the neuron \( j \) to the neuron \( i \). The connection weights of Boltzmann machines are generally restricted. For example, those of ordinary Boltzmann machines and CBMs are symmetric and complex-conjugate, respectively. Those of RoBMs are restricted by \( W_{i,j} = W^T_{j,i} \), where superscript \( T \) represents a transpose. Figure 1 shows the connection weights of two-dimensional RoBMs. Let \( \mathbf{x}_i \) be the state of the neuron \( i \). Then, the weighted sum input to the neuron \( i \) is \( \mathbf{I}_i = \sum_{j \neq i} W_{ij} \mathbf{x}_j \).
Third, we define the energy function of RoBMs. We denote the state of RoBM by \( X = (x_1, x_2, \ldots, x_L) \), where \( L \) is the number of neurons. The energy \( H \) of \( X \) is defined as

\[
H(X) = -\frac{1}{2} \sum_i \sum_{j \neq i} x_i^T W_{ij} x_j.
\]  

When the state of the neuron \( k \) is updated to \( x_k' \), the energy gap \( \Delta H \) is as follows:

\[
\Delta H = \frac{1}{2} \sum_{j \neq k} x_k^T W_{kj} x_j - \frac{1}{2} \sum_{j \neq k} x_k'^T W_{kj} x_j + \frac{1}{2} \sum_{i \neq k} x_i^T W_{ik} x_k - \frac{1}{2} \sum_{i \neq k} x_i'^T W_{ik} x_k
\]

\[
= \frac{1}{2} \sum_{j \neq k} x_k^T W_{kj} x_j - \frac{1}{2} \sum_{j \neq k} x_k'^T W_{kj} x_j + \frac{1}{2} \sum_{i \neq k} x_i^T W_{ik} x_i - \frac{1}{2} \sum_{i \neq k} x_i'^T W_{ik} x_i
\]

\[
= x_k^T I_k - x_k'^T I_k
\]

\[
= x_k \cdot I_k - x_k' \cdot I_k.
\]

We denote the stationary distributions of the state \( X \) and the transition probabilities from the state \( X \) to the state \( X' \) as \( p(X) \) and \( p(X \rightarrow X') \), respectively. The states \( X \) and \( X' \) differ only in the state of the neuron \( k \). Further, we can describe the states \( X \) and \( X' \) as follows:

\[
X = (x_1, x_2, \ldots, x_k, \ldots, x_L),
\]

\[
X' = (x_1, x_2, \ldots, x'_k, \ldots, x_L).
\]

The relation

\[
p(X)p(X \rightarrow X') = p(X')p(X' \rightarrow X)
\]

holds according to the stationary condition. Therefore, we obtain the following equations:

\[
\frac{p(X')}{p(X)} = \frac{p(X \rightarrow X')}{p(X' \rightarrow X)}
\]

\[
= \frac{\exp(x_k' \cdot I_k)}{\exp(x_k \cdot I_k)}
\]

\[
= \exp(-\Delta H).
\]

Thus, we obtain the following equations:

\[
p(X) = \frac{\exp(-H(X))}{Z},
\]

\[
Z = \int \exp(-H(X)) \, dX.
\]

3. Information geometry of exponential family

In this section, we briefly describe the information geometry of an exponential family. See [4] for the detail of information geometry. An exponential family is a probability distribution family defined by the following form using functions \( C(x) \) and \( F_i(x) \):

\[
p(x; \theta) = \exp(C(x) + \sum_i F_i(x) \theta_i - \psi(\theta)),
\]

\[
\psi(\theta) = \log \int \exp(C(x) + \sum_i F_i(x) \theta_i) \, dx,
\]

where \( x \) is a random variable and \( \theta = (\theta_1, \theta_2, \ldots, \theta_n) \) is the parameter. The parameter \( \theta \) is referred to as a natural parameter. For a non-singular \( n \times n \) matrix \( A \) and an \( n \)-dimensional vector \( b \), we define \( \bar{\theta} \) as
\[ \overline{\theta} = (\overline{\theta}_1, \overline{\theta}_2, \ldots, \overline{\theta}_n) = \theta A + b. \]  

(18)

Then, \( \overline{\theta} \) is another natural parameter of the exponential family. Conversely, all natural parameters are written in the form (18). We denote the logarithmic likelihood function of \( p(x; \theta) \) as \( l(x; \theta) = \log p(x; \theta) \).

We denote the expectation \( E_\theta[f(x)] \) of \( f(x) \) with respect to \( p(x; \theta) \) as

\[ E_\theta[f(x)] = \int f(x)p(x; \theta)dx. \]  

(19)

We may also write \( E_\theta[f(x)] \) instead of \( E_\theta[f(x)] \). For two tangent vectors \( D_1 \) and \( D_2 \), the Fisher metric \( g(D_1, D_2) \) is defined as

\[ g(D_1, D_2) = E_\theta[D_1l(x; \theta)D_2l(x; \theta)]. \]  

(20)

In particular, we denote \( g_{ij} = g(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j}) \), and \( g_{ij} = g_{ji} \) is definitely true. In general, the Fisher metric \( g_{ij} \) may also be written as

\[ g_{ij} = -E_\theta \left[ \frac{\partial^2 l(x; \theta)}{\partial \theta_i \partial \theta_j} \right]. \]  

(21)

**Theorem 1**

For an exponential family, the equation

\[ g_{ij} = E_\theta[F_i(x)F_j(x)] - E_\theta[F_i(x)]E_\theta[F_j(x)]. \]  

(22)

holds.

**Proof**

See the Appendix for details.

We denote the expectation of \( F_i(x) \) by \( \eta_i = E_\theta[F_i(x)] \). \( \eta = (\eta_1, \eta_2, \ldots, \eta_n) \) is another parameter referred to as an expectation parameter. \( \eta \) is determined by \( \theta \) and is referred to as the conjugate parameter of \( \theta \). If we take another natural parameter, then we obtain a different conjugate parameter.

We define a function \( \phi(\eta) \) as

\[ \phi(\eta) = \sum_i \theta_i \eta_i - \psi(\theta). \]  

(23)

Then, the following theorem holds.

**Theorem 2**

\[ \frac{\partial \psi(\theta)}{\partial \theta_i} = \eta_i \]  

(24)

\[ \frac{\partial \phi(\eta)}{\partial \eta_i} = \theta_i \]  

(25)

\[ \frac{\partial \eta_j}{\partial \theta_i} = g_{ij} \]  

(26)

**Proof**

See the Appendix for details.

Theorem 2 implies the Legendre transformations between the parameters \( \theta \) and \( \eta \). The functions \( \psi \) and \( \phi \) are referred to as their potential functions. From \( g_{ij} = g_{ji} \), we obtain the following corollary.

**Corollary 1**

\[ \frac{\partial \eta_i}{\partial \theta_j} = \frac{\partial \eta_j}{\partial \theta_i} \]  

(27)

We denote the Fisher metric with respect to the parameter \( \eta \) as

\[ g^{ij} = E_\theta \left[ \frac{\partial l(x; \theta)}{\partial \eta_i} \frac{\partial l(x; \theta)}{\partial \eta_j} \right]. \]  

(28)

The equation \( g^{ij} = g^{ji} \) is definitely true. The following theorem holds.
Theorem 3

\[ \frac{\partial \theta_i}{\partial \eta_j} = g^{ij} \] (29)

**Proof** From Eq. (26), we obtain the following equation.

\[ g^{ij} = E_\theta \left[ \frac{\partial l(x; \theta)}{\partial \eta_i} \frac{\partial l(x; \theta)}{\partial \eta_j} \right] \] (30)

\[ = \sum_{h,k} \frac{\partial \theta_h}{\partial \eta_i} \frac{\partial \theta_k}{\partial \eta_j} E_\theta \left[ \frac{\partial l(x; \theta)}{\partial \theta_h} \frac{\partial l(x; \theta)}{\partial \theta_k} \right] \] (31)

\[ = \sum_{h,k} \frac{\partial \theta_h}{\partial \eta_i} \frac{\partial \eta_k}{\partial \eta_j} g_{hk} \] (32)

\[ = \sum_{h,k} \frac{\partial \theta_h}{\partial \eta_i} \frac{\partial \eta_k}{\partial \eta_j} \frac{\partial \theta_k}{\partial \eta_h} \] (33)

\[ = \sum_{h,k} \frac{\partial \theta_h}{\partial \eta_i} \frac{\partial \eta_k}{\partial \eta_j} \] (34)

\[ = \frac{\partial \theta_i}{\partial \eta_j} \] (35)

From \( g^{ij} = g^{ji} \), we obtain the following corollary.

**Corollary 2**

\[ \frac{\partial \theta_i}{\partial \eta_j} = \frac{\partial \theta_j}{\partial \eta_i} \] (36)

Theorem 4

\[ \sum_k g^{ik} g_{kj} = \delta_{ij} \] (37)

**Proof**

\[ \sum_k g^{ik} g_{kj} = \sum_k \frac{\partial \theta_i}{\partial \eta_k} \frac{\partial \eta_k}{\partial \eta_j} = \frac{\partial \theta_i}{\partial \eta_j} = \delta_{ij} \] (38)

We denote \( g_i^j = g(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \eta_j}) \). Then the following theorem holds.

Theorem 5

\[ g_i^j = \delta_{ij} \] (39)

**Proof** From Theorems 3 and 4, we obtain the following equation.

\[ g_i^j = g(\frac{\partial}{\partial \theta_i}, \sum_k \frac{\partial \theta_k}{\partial \eta_j} \frac{\partial}{\partial \theta_k}) \] (40)

\[ = \sum_k \frac{\partial \theta_k}{\partial \eta_j} g(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_k}) \] (41)

\[ = \sum_k g^{ik} g_{ki} \] (42)

\[ = \delta_{ij} \] (43)
4. Information geometry of rotor Boltzmann machines

The number of parameters between two neurons of RoBM is $N^2$; the number of connections is $L(L-1)$. Since there exists the relation $W_{ij} = W_{ji}^T$, the number of parameters of RoBM is $\frac{1}{2} N^2 L(L-1)$. Let $x_i^a$ and $w_{ij}^{a,b}$ be the $a$th element of $x_i$ and the $(a, b)$th element of $W_{ij}$, respectively. Then $W = (w_{ij}^{a,b})_{i<j}$ is the vector of all parameters of RoBM. We also write $p(X; W)$ in place of $p(X)$ in order to clarify the parameters. Since $Z$ is a function of $W$, we may substitute $\psi(W) = \log Z$. We can describe $p(X; W)$ as follows:

$$p(X; W) = \exp(-H(X) - \psi(W))$$ (44)

Therefore, RoBM is an exponential family and $\psi(W)$ is a natural parameter. We put $\alpha = (i, j, a, b)$ and write $w_{ij}^{a,b}$ for simplicity, then we can describe expression (45) as

$$p(X; W) = \exp(\sum_{i<j} \sum_{a,b} x_i^a x_j^b w_{ij}^{a,b} - \psi(W)).$$ (45)

We can describe $l(X; W)$ as follows:

$$l(X; W) = \log p(X; W)$$ (47)

$$= \sum_{i<j} x_i^a x_j^b w_{ij}^{a,b} - \psi(W).$$ (48)

From Theorem 1, the Fisher metric for tangent vectors $\frac{\partial}{\partial w_{ij}^{a,b}}$ and $\frac{\partial}{\partial w_{kl}^{c,d}}$ is as follows, where $\beta = (k, l, c, d)$:

**Theorem 6**

$$g_{\alpha\beta} = E \left[ x_i^a x_j^b x_k^c x_l^d \right] - E \left[ x_i^a x_j^b \right] E \left[ x_k^c x_l^d \right].$$ (49)

The expectation parameter $W^* = (v_\alpha)$, which is a conjugate parameter of the natural parameter, is given as

$$v_\alpha = E \left[ x_i^a x_j^b \right].$$ (50)

We define

$$\phi(W) = E \left[ l(X; W) \right].$$ (51)

**Theorem 7**

$$\psi(W) + \phi(W) = \sum_{\alpha} w_\alpha v_\alpha$$ (52)

**Proof**

$$\phi(W) = E \left[ \sum_{\alpha} x_i^a x_j^b w_\alpha - \psi(W) \right]$$ (53)

$$= \sum_{\alpha} w_\alpha E \left[ x_i^a x_j^b \right] - \psi(W)$$ (54)

$$= \sum_{\alpha} w_\alpha v_\alpha - \psi(W)$$ (55)

We obtain the conclusion.

Thus, we found that $\psi(W)$ and $\phi(W)$ are potential functions of Legendre transformations between the parameters $w_\alpha$ and $v_\alpha$.

We can describe the Fisher metric $g^{\alpha\beta}$ of $\frac{\partial}{\partial w_{ij}^{a,b}}$ and $\frac{\partial}{\partial w_{kl}^{c,d}}$ as

$$g^{\alpha\beta} = E \left[ \frac{\partial l(X; W)}{\partial w_\alpha} \frac{\partial l(X; W)}{\partial w_\beta} \right].$$ (56)

From Theorems 2, 3, and 4, we obtain the following theorem.
Theorem 8

\[
\frac{\partial \psi(W)}{\partial w_\alpha} = v_\alpha \quad (57)
\]

\[
\frac{\partial \phi(W)}{\partial v_\alpha} = w_\alpha \quad (58)
\]

\[
\frac{\partial v_\alpha}{\partial w_\beta} = g_{\alpha\beta} \quad (59)
\]

\[
\frac{\partial w_\alpha}{\partial v_\beta} = g_{\alpha\beta} \quad (60)
\]

\[
\sum_\gamma g^{\gamma\alpha} g_{\gamma\beta} = \delta_{\alpha\beta} \quad (61)
\]

5. Quantized two-dimensional rotor Boltzmann machines

A two-dimensional RoBM (TRoBM) is an extension of CBMs. If the connection weight \(w_{ij}\) is restricted by the form

\[
w_{ij} = \begin{pmatrix} u_{ij} & -v_{ij} \\ v_{ij} & u_{ij} \end{pmatrix}, \quad (62)
\]
then the equation $w_{ij}^T = w_{ji}$ holds. This restricted TRoBM is equivalent to a CBM. When we regard $x_i$ and $w_{ij}$ as $x_i^1 + x_i^2 \sqrt{-1}$ and $u_{ij} + v_{ij} \sqrt{-1}$, respectively, we can identify this restricted TRoBM and CBM. Figure 2(a) shows neurons and connection weights of CBM and TRoBM. Figure 2(b) shows a TRoBM equivalent to a CBM.

Quantized complex-valued neurons have been readily available. We construct a QTRoBM based on QCBM. First, we define the set $S_K = \{s_k\}_{k=0}^{K-1}$ of quantized two-dimensional rotor neuron states as

$$s_k = \left( \frac{\cos \frac{2k\pi}{K}}{\sin \frac{2k\pi}{K}} \right).$$

(63)

$K$ is the resolution factor. Figure 3 shows a quantized two-dimensional rotor neuron in case of $K = 8$. If we regard complex numbers as two-dimensional vectors, complex-valued neurons and rotor neurons are identical. Next, we define quantized two-dimensional rotor neurons. For an input $I$, which is a two-dimensional vector, the probability $p(s_k; I)$ that the rotor neuron state is $s_k$ is described as

$$p(s_k; I) = \frac{\exp(s_k \cdot I)}{\sum_{k=0}^{K-1} \exp(s_k \cdot I)}.$$  

(64)

We can define the energy $H(X)$ and determine the probability $p(X; W)$ in a manner similar to that used for RoBMs.

$$p(X; W) = \frac{\exp(-H(X))}{Z}$$

(65)

$$Z = \sum_{X \in V} \exp(-H(X))$$

(66)

$V$ is the $L$ direct product of $S_K$.

6. Information geometry of finite discrete distribution

Let $X$ be a finite set $\{x_0, x_1, \ldots, x_n\}$. Consider the whole set $S$ of probability distributions such that $p(x = x_i) > 0$ for all $x_i \in X$. This distribution is referred to as discrete distribution. If we substitute $p_i = p(x = x_i)$ for $i \neq 0$, then $p(x = x_0) = 1 - \sum_{i=1}^{n} p_i$. $p = (p_1, p_2, \cdots, p_n)$ is a parameter of the discrete distribution family. $S$ is an exponential family as expressed by the following form.

$$p(x) = \exp \left( \sum_{j=1}^{n} F_j(x) \theta_j - \psi(\theta) \right)$$

(67)

$$F_i(x) = \begin{cases} 1 & (x = x_i) \\ 0 & (x \neq x_i) \end{cases}$$

(68)
\[
\psi(\theta) = \log(1 + \sum_{j=1}^{n} \exp \theta_j)
\]  

(69)

In fact, substituting \( \theta_i = \log p_i - \log(1 - \sum_{j=1}^{n} p_j) \), we obtain the following equation for \( i \neq 0 \).

\[
\psi(\theta) = -\log(1 - \sum_{j=1}^{n} p_j)
\]  

(70)

\[
p(x = x_i) = \exp(\log \frac{p_i}{1 - \sum_{j=1}^{n} p_j} + \log(1 - \sum_{j=1}^{n} p_j))
\]  

(71)

\[
= \exp(\log p_i)
\]  

(72)

\[
= p_i
\]  

(73)

\[
p(x = x_0) = \exp(\log(1 - \sum_{j=1}^{n} p_j))
\]  

(74)

\[
= 1 - \sum_{j=1}^{n} p_j
\]  

(75)

Let \( \theta(p) \) and \( \theta(q) \) be the parameters corresponding to two distributions \( p(x) \) and \( q(x) \) in \( S \), respectively. For any real number \( t (0 \leq t \leq 1) \), \( (1-t)\theta(p) + t\theta(q) \) belongs to \( S \). The line \( (1-t)\theta(p) + t\theta(q) \) on \( S \) is referred to as \( e \)-geodesic between \( p(x) \) and \( q(x) \) (Fig. 4(a)). The distribution \( r(x; t) \) corresponding to \( (1-t)\theta(p) + t\theta(q) \) is expressed as follows:

\[
r(x; t) = p(x)^{1-t}q(x)^tf(t),
\]  

(76)

\[
f(t) = \frac{1}{\int p(x)^{1-t}q(x)^tf(x)dx}.
\]  

(77)

For any real number \( t (0 \leq t \leq 1) \), \( (1-t)p(x) + tq(x) \) also belongs to \( S \). The curve \( r(x; t) = (1-t)p(x) + tq(x) \) on \( S \) is referred to as m-geodesic. The expectation of \( r(x; t) \) is determined as

\[
E_\theta[r(x; t)] = (1-t)E_\theta[p(x)] + tE_\theta[q(x)].
\]  

(78)

Using the expectation parameter, we can express the m-geodesic between \( p(x) \) and \( q(x) \) as

\[
\eta(r(x; t)) = (1-t)\eta(p) + t\eta(q).
\]  

(79)

Therefore, m-geodesic is a line with respect to an expectation parameter (Fig. 4(b)).

Let \( M \) be a subset of \( S \). If the \( e \)-geodesic between \( p(x) \) and \( q(x) \) belongs to \( M \) for any distributions \( p(x) \) and \( q(x) \) in \( M \), \( M \) is referred to as e-flat. \( M \) is e-flat if and only if \( M \) is an exponential family. If \( M \) is e-flat, \( M \) is an affine subset of \( S \) with regard to a natural parameter. Hence we can take a natural parameter \( (\theta_1, \theta_2, \cdots, \theta_n) \) of \( S \) satisfying

\[
M = \{ (\theta_1, \theta_2, \cdots, \theta_n, 0, \cdots, 0) \}.
\]  

(80)

If the m-geodesic between \( p(x) \) and \( q(x) \) belongs to \( M \) for any distributions \( p(x) \) and \( q(x) \) in \( M \), \( M \) is referred to as m-flat. \( M \) is m-flat if and only if \( M \) is an affine subset of \( S \) with regard to an expectation parameter. If \( M \) is m-flat, we can take an expectation parameter \( (\eta_1, \eta_2, \cdots, \eta_n) \) of \( S \) satisfying

\[
M = \{ (\eta_1, \eta_2, \cdots, \eta_n, 0, \cdots, 0) \}.
\]  

(81)

We define the KL divergence between distributions \( p \) and \( q \in S \) as follows:

\[
KL(p, q) = E_p \left[ \log \frac{p(x)}{q(x)} \right]
\]  

(82)

\[
= E_p \left[ l(x; \theta(p)) - l(x; \theta(q)) \right].
\]  

(83)
KL(p, q) is always non-negative and vanishes if and only if p(x) = q(x). However, KL(p, q) and KL(q, p) are generally different. KL divergence is a measure of the difference between two distributions and is also known as the pseudo-distance. By the definition of exponential family and Eq. (23), we obtain the following equations.

\[ KL(p, q) = E_p \left[ \sum_i F_i(x) \{ \theta_i(p) - \theta_i(q) \} \right] + \psi(\theta(q)) - \psi(\theta(p)) \tag{84} \]

\[ = \sum_i \eta_i(p) \{ \theta_i(p) - \theta_i(q) \} + \psi(\theta(q)) + \phi(\eta(p)) - \sum_i \theta_i(p) \eta_i(p) \tag{85} \]

\[ = \psi(\theta(q)) + \phi(\eta(p)) - \sum_i \theta_i(q) \eta_i(p) \tag{86} \]

**Theorem 9** Let p, q, and r belong to S. Let \( \gamma_{pr} \) and \( \gamma_{qr} \) be m-geodesic between p and r and e-geodesic between q and r. If \( \gamma_{pr} \) and \( \gamma_{qr} \) are orthogonal, then we obtain

\[ KL(p, q) = KL(p, r) + KL(r, q). \tag{87} \]

**Proof** See the Appendix for details.

Suppose that M is e-flat. For p in S and q in M, if the m-geodesic between p and q is orthogonal to M, then q is referred to as the m-projection of p onto M. From Theorem 9, for any r in M, we obtain the inequality

\[ KL(p, r) = KL(p, q) + KL(q, r) \geq KL(p, q). \tag{88} \]

Therefore, the m-projection q of p minimizes KL(p, q). Suppose that M is m-flat. For p in S and q in M, if the e-geodesic between p and q is orthogonal to M, then q is referred to as the e-projection of p onto M. In a manner similar to that for the m-projection, we find that the e-projection q of p minimizes KL(q, p).

A combination \((\eta_1, \eta_2, \ldots, \eta_r, \theta_{r+1}, \ldots, \theta_n)\) of the natural parameter and expectation parameter forms another parameter, referred to as a mixed parameter. Suppose that a subset M of S is e-flat and its dimension is r. Then, M is expressed by \((\theta_1, \theta_2, \ldots, \theta_r, 0, \ldots, 0)\).

**Theorem 10** Suppose that M is e-flat and expressed by \((\theta_1, \ldots, \theta_r, 0, \ldots, 0)\) with regard to a natural parameter, and p in S is expressed by \((\eta_1(p), \ldots, \eta_n(p))\) with regard to the expectation parameter. Then, \( q = (\eta_1(p), \ldots, \eta_r(p), 0, \ldots, 0) \) in the mixed parameter \((\eta_1, \ldots, \eta_r, \theta_{r+1}, \ldots, \theta_n)\) is the m-projection of p onto M.

**Proof** See the Appendix for details.

We can prove the following theorem in the same manner.

**Theorem 11** Suppose that M is m-flat and expressed by \((\eta_1, \ldots, \eta_r, 0, \ldots, 0)\) with regard to an expectation parameter, and p in S is expressed by \((\theta_1(p), \ldots, \theta_n(p))\) with regard to the expectation parameter. Then \( q = (\theta_1(p), \ldots, \theta_r(p), 0, \ldots, 0) \) in the mixed parameter \((\theta_1, \ldots, \theta_r, \eta_{r+1}, \ldots, \eta_n)\) is the e-projection of p onto M.

7. Information geometry of quantized two-dimensional rotor Boltzmann machines

We can also obtain all the theorems in section 4 for QTRoBMs in the same manner. In addition, we can prove some theorems on the information geometry of QTRoBMs. Let A be the set of all distributions on V. If \( q \in A \), then q satisfies the following conditions:

\[ q(x) > 0 \quad (x \in V), \tag{89} \]
\sum_{X \in V} q(X) = 1. \tag{90}

Since $V$ is a finite set, $A$ is an exponential family and has $K^L - 1$ parameters. Let $M$ be the set of all QTRoBMs with resolution factor $K$ and $L$ neurons. $M$ has parameter $\mathbf{W} = (w_{ij}^{ab})_{i<j}$, where $a$ and $b$ are 1 or 2 and may be described as
\begin{align*}
p(X; \mathbf{W}) &= \exp(\sum_{\alpha} x_i^a x_j^b w_{\alpha} - \psi(\mathbf{W})), \tag{91}
\end{align*}

where $w_{\alpha} = w_{ij}^{ab}$ for $\alpha = (i, j, a, b)$. The parameter $\mathbf{W}$ has $2L(L - 1)$ elements and is a natural parameter of $M$. Since $A$ and $M$ are exponential families and $M$ is a subset of $A$, we can append the parameters $\bar{\mathbf{W}} = (\bar{w}_1, \cdots, \bar{w}_{KL - 2L(L-1)})$ such that $(\mathbf{W}, \bar{\mathbf{W}})$ is a natural parameter, and there exists $F_k(\mathbf{X})$ and $\bar{\psi}(\mathbf{W}, \bar{\mathbf{W}})$ such that the exponential family $A$ can be written as
\begin{align*}
p(X; \mathbf{W}, \bar{\mathbf{W}}) &= \exp(\sum_{\alpha} x_i^a x_j^b w_{\alpha} + \sum_k F_k(\mathbf{X})\bar{w}_k - \bar{\psi}(\mathbf{W}, \bar{\mathbf{W}})). \tag{92}
\end{align*}

Since $M$ is $p(X; \mathbf{W}, \mathbf{0})$, we obtain $\bar{\psi}(\mathbf{W}, \mathbf{0}) = \psi(\mathbf{W})$. We define the conjugate expectation parameter $(U, \bar{U})$ of $(\mathbf{W}, \bar{\mathbf{W}})$ as follows:
\begin{align*}
U &= (u_{\alpha}), \tag{93} \\
\bar{U} &= (\bar{u}_k), \tag{94} \\
u_{\alpha} &= E_p [x_i^a x_j^b], \tag{95} \\
\bar{u}_k &= E_p [F_k(\mathbf{X})]. \tag{96}
\end{align*}

We give a learning rule for QTRoBMs. The purpose of learning is to find a QTRoBM $q(X; \mathbf{W})$ in $M$ to approximate a given distribution $p(X)$ in $A$. Boltzmann learning uses KL divergence as a measure of approximation and produces a gradient descent learning for KL divergence. We determine the Boltzmann learning for QTRoBMs. Let $q(X; \mathbf{W})$ be a current QTRoBM. We solve the following equation to move $q$ such that the KL divergence decreases:
\begin{align*}
\Delta w_{\alpha} &= -\epsilon \frac{\partial KL(p, q)}{\partial w_{\alpha}}, \tag{97}
\end{align*}

where $\epsilon$ is a learning rate.
\begin{align*}
KL(p, q) &= E_p \left[ \log \frac{p(X)}{q(X; \mathbf{W})} \right], \tag{98} \\
\frac{\partial KL(p, q)}{\partial w_{\alpha}} &= -E_p \left[ \frac{\partial \log q(X; \mathbf{W})}{\partial w_{\alpha}} \right] \tag{99} \\
&= -E_p \left[ x_i^a x_j^b - \frac{\partial \psi(\mathbf{W})}{\partial w_{\alpha}} \right] \tag{100} \\
&= -E_p \left[ x_i^a x_j^b - E_q [x_i^a x_j^b] \right] - E_q [x_i^a x_j^b] \tag{101} \\
&= E_q [x_i^a x_j^b] - E_p [x_i^a x_j^b] \tag{102}
\end{align*}

We obtained the following Boltzmann learning rule.
\begin{align*}
\Delta w_{\alpha} &= \epsilon \left( E_p [x_i^a x_j^b] - E_q [x_i^a x_j^b] \right) \tag{103}
\end{align*}

Boltzmann learning depends on the parameter. We give another learning rule independent of the parameter, referred to as the natural gradient descent rule. If $q$ moves in the direction $p - q$, $q$ trails along the m-geodesic through $p$ and $q$, which forms a line in an expectation parameter. The natural gradient learning rule tries to move $q$ along the line as follows:
\begin{align*}
\Delta u_{\alpha}(q) &= \epsilon (u_{\alpha}(p) - u_{\alpha}(q)), \tag{104} \\
\Delta \bar{u}_k(q) &= \epsilon (\bar{u}_k(p) - \bar{u}_k(q)). \tag{105}
\end{align*}
We may rewrite the left side of Eq. (104) as follows:
\[
\Delta u_\alpha = \sum_\beta \frac{\partial u_\alpha}{\partial w_\beta} \Delta w_\beta + \sum_l \frac{\partial u_\alpha}{\partial \tilde{w}_l} \Delta \tilde{w}_l \\
= \sum_\beta g_{\alpha\beta} \Delta w_\beta + \sum_l g_{\alpha l} \Delta \tilde{w}_l. \tag{106}
\]
Since \(\tilde{w}_l\) is fixed at 0 and \(\Delta \tilde{w}_l = 0\) holds, we obtain
\[
\Delta u_\alpha = \sum_\beta g_{\alpha\beta} \Delta w_\beta. \tag{107}
\]
Therefore, we obtain natural gradient learning rule as follows:
\[
\Delta w_\alpha(q) = \sum_\beta g^{\alpha\beta} \Delta u_\beta \tag{108}
\]
\[
= \epsilon \sum_\beta g^{\alpha\beta} (u_\beta(p) - u_\beta(q)) \tag{109}
\]
\[
= \epsilon \sum_\beta g^{\alpha\beta} (E_p[x_i^a x_j^b] - E_q[x_i^a x_j^b]). \tag{110}
\]
Since \(\tilde{w}_k\) is fixed, the natural gradient descent rule does not realize the move in the direction \(p - q\) and provides its projection to \(M\) (Fig. 5).

8. Quantized two-dimensional rotor Boltzmann machine with hidden neurons

In the present section, we consider QTRoBM with hidden neurons (Fig. 6). Let \(X\) and \(Y\) be the visible and hidden neurons, respectively. We denote the set of distribution \(p(X,Y)\) realizable by a QTRoBM as \(M\). Then, \(M\) is e-flat. Since two states \((X,Y)\) and \((X,Y')\) differ only with respect to hidden neurons, they are identical. We define \(p(X)\) as
\[
p(X) = \sum_Y p(X,Y). \tag{111}
\]
We denote the set of distribution \(p(X)\) realizable by a QTRoBM as \(\tilde{M}\). The purpose of learning is to approximate a given \(p(X)\) by \(q(X)\) realizable by a QTRoBM. We define the following set \(D\) for a given \(p(X)\):
\[
D = \{q(X,Y) | q(X) = p(X)\}. \tag{112}
\]
Then, \(D\) is m-flat. We define KL-divergence between \(D\) and \(M\) as
\[
KL(D,M) = \min_{\begin{subarray}{c}p(X,Y) \in D \\ q(X,Y) \in M\end{subarray}} KL(p(X,Y), q(X,Y)). \tag{113}
\]
**Fig. 6.** (a) QTRoBM consists of only visible neurons. (b) QTRoBM with hidden neurons includes hidden neurons. Two states that are different only in hidden neurons are identical.

**Theorem 12**

\[ KL(D, M) = \min_{q(X) \in M} KL(p(X), q(X)) \]  

**Proof** See the Appendix for details.

We have to minimize only \( KL(D, M) \). We provide the em-algorithm to minimize \( KL(D, M) \). We start with an initial QTRoBM \( q_0(X, Y) \) and iteratively create better approximations \( q_i(X, Y) \). Let \( p_{i+1}(X, Y) \) be the e-projection of \( q_i(X, Y) \) to \( D \). This process is referred to as e-step. We take \( q_i(X, Y) \) by the m-projection of \( p_i(X, Y) \) to \( M \). This process is referred to as the m-step. Then, the inequality

\[ KL(p_{i+1}, q_{i+1}) \leq KL(p_i, q_i) \leq KL(p_{i+1}, q_i). \]

holds. This inequality shows that this algorithm iteratively produces better approximations. Boltzmann learning and natural gradient descent learning realize the m-step. We have only to minimize \( KL(p_{i+1}(X, Y), q_i(X, Y)) \) to realize the e-step. From \( p_{i+1}(X, Y) = p(X)p_{i+1}(Y|X) \), we obtain the following equation:

\[ KL(p_{i+1}(X, Y), q_i(X, Y)) = E_{p_{i+1}(X, Y)} \left[ \log \frac{p(X)p_{i+1}(Y|X)}{q_i(X)q_i(Y|X)} \right] \]

\[ = KL(p(X), q_i(X)) + E_{p(X)} \left[ KL(p_{i+1}(Y|X), q_i(Y|X)) \right]. \]

Therefore, we can minimize \( KL(p_{i+1}(X, Y), q_i(X, Y)) \) by \( p_{i+1}(Y|X) = q_i(Y|X) \). We summarize the em-algorithm as follows.

1. e-step is the e-projection of \( q_i \) to \( D \) and is realizable by \( p_{i+1}(X, Y) = p(X)q_i(Y|X) \).

2. m-step is the m-projection of \( p_i \) to \( M \) and is realizable by Boltzmann learning or natural gradient descent learning.

**9. Conclusions**

An RoBM is a high-dimensional model of neural networks. A TRoBM is an extension of CBMs and is extremely important. In the present work, we have investigated RoBMs through information geometry. We revealed that RoBMs form an exponential family and determined the Fisher metric, natural
and expectation parameters, and their potential functions. In addition, we proposed QTRoBMs and provided Boltzmann learning, natural gradient descent learning, and the em-algorithm for them. We need infinite dimensional information geometry in order to discuss the em-algorithm for continuous RoBMs [23, 24]. In the future, we should develop approximate learning algorithms, such as mean field theory, for pragmatic learning algorithms.

Appendix

A. Proof of Theorem 1

\[ \frac{\partial \psi(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \log \int \exp(C(x) + \sum_j F_j(x)\theta_j)dx \] (A-1)

\[ = \int F_i(x)\exp(C(x) + \sum_j F_j(x)\theta_j)dx \] (A-2)

\[ = \int F_i(x)\exp(C(x) + \sum_j F_j(x)\theta_j - \psi(\theta))dx \] (A-3)

\[ = \int F_i(x)p(x; \theta)dx \] (A-4)

\[ l(x; \theta) = C(x) + \sum_j F_j(x)\theta_j - \psi(\theta) \] (A-5)

\[ \frac{\partial l(x; \theta)}{\partial \theta_i} = F_i(x) - \frac{\partial \psi(\theta)}{\partial \theta_i} \] (A-6)

\[ = F_i(x) - E_\theta[F_i(x)] \] (A-7)

\[ g_{ij} = E_\theta \left[ \frac{\partial l(x; \theta)}{\partial \theta_i} \frac{\partial l(x; \theta)}{\partial \theta_j} \right] \] (A-8)

\[ = E_\theta \left[ (F_i(x) - E_\theta[F_i(x)])(F_j(x) - E_\theta[F_j(x)]) \right] \] (A-9)

\[ = E_\theta [F_i(x)F_j(x)] - E_\theta [F_i(x)] E_\theta [F_j(x)] \] (A-10)

B. Proof of Theorem 2

Equation (24) was proven in the proof of Theorem 1.

\[ \frac{\partial \phi(\eta)}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} \left( \sum_j \theta_j \eta_j - \psi(\theta) \right) \] (B-1)

\[ = \sum_j \frac{\partial \theta_j}{\partial \eta_i} \eta_j + \theta_i - \sum_j \frac{\partial \theta_j}{\partial \eta_i} \frac{\partial \psi(\theta)}{\partial \theta_j} \] (B-2)

\[ = \sum_j \frac{\partial \theta_j}{\partial \eta_i} \eta_j + \theta_i - \sum_j \frac{\partial \theta_j}{\partial \eta_i} \eta_j \] (B-3)

\[ = \theta_i \] (B-4)

Thus, we obtained Eq. (25).
C. Proof of Theorem 9

The tangent vector of $\gamma_{pr}$ at $r$ is $\sum_i (\eta_i(r) - \eta_i(p)) \frac{\partial}{\partial \eta_i}$. The tangent vector of $\gamma_{qr}$ at $q$ is $\sum_i (\theta_i(r) - \theta_i(q)) \frac{\partial}{\partial \theta_i}$. Since they are orthogonal, we obtain the following equations:

$$g(\sum_i (\eta_i(r) - \eta_i(p)) \frac{\partial}{\partial \eta_i}, \sum_i (\theta_i(r) - \theta_i(q)) \frac{\partial}{\partial \theta_i}) = \sum_{i,j} \langle \eta_i(r) - \eta_i(p), (\theta_j(r) - \theta_j(q)) \rangle$$

Therefore, the following equations hold:

$$KL(p, r) + KL(r, q) - KL(p, q) = \psi(\theta(r)) + \phi(\eta(p)) - \sum_i \theta_i(r) \eta_i(p) + \psi(\theta(q)) + \phi(\eta(r))$$

$$- \sum_i \theta_i(q) \eta_i(r) - \psi(\theta(q)) - \phi(\eta(p)) + \sum_i \theta_i(q) \eta_i(p)$$

$$KL(p, r) + KL(r, q) - KL(p, q) = \psi(\theta(r)) + \phi(\eta(r)) - \sum_i \theta_i(r) \eta_i(p) - \sum_i \theta_i(q) \eta_i(r) + \sum_i \theta_i(q) \eta_i(p)$$

$$KL(p, r) + KL(r, q) - KL(p, q) = \sum_i (\theta_i(r) - \theta_i(q))(\eta_i(r) - \eta_i(p))$$

$$= 0.$$  

Therefore, tangent vectors $D_m$ and $D_m$ are orthogonal. We find that $q$ is the m-projection.

D. Proof of Theorem 10

$q$ can be expressed in the expectation parameter as

$$(\eta(p), \cdots, \eta_r(q), \cdots, \eta_n(q)).$$

(E-1)

The tangent vector of the m-geodesic through $p$ and $q$ is parallel to the vector $D_m$:

$$D_m = \sum_{i=r+1}^{n} (\eta_i(p) - \eta_i(q)) \frac{\partial}{\partial \eta_i}.$$  

(D-2)

A tangent vector at $q$ of $M$ can be expressed by the vector $D_M$:

$$D_M = \sum_{i=1}^{r} \alpha_i \frac{\partial}{\partial \theta_i}.$$  

(D-3)

Further, we can calculate $g(D_M, D_m)$ as follows:

$$g(D_M, D_m) = \sum_{i=1}^{r} \sum_{j=r+1}^{n} \alpha_i \langle \eta_j(p) - \eta_j(q) \rangle g(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \eta_j})$$

$$= \sum_{i=1}^{r} \sum_{j=r+1}^{n} \alpha_i \langle \eta_j(p) - \eta_j(q) \rangle g_i^j$$

Therefore, tangent vectors $D_M$ and $D_m$ are orthogonal. We find that $q$ is the m-projection.

E. Proof of Theorem 12

From $p(X, Y) \in D$, $p(X)$ is fixed.
\begin{align}
KL(p(X, Y), q(X, Y)) &= KL(p(X)p(Y|X), q(X)q(Y|X)) \\
 &= E_{p(X,Y)} \left[ \log \frac{p(X)p(Y|X)}{q(X)q(Y|X)} \right] \\
 &= E_{p(X,Y)} \left[ \log \frac{p(X)}{q(X)} + \log \frac{p(Y|X)}{q(Y|X)} \right] \\
 &= KL(p(X), q(X)) + E_{p(X)} [KL(p(Y|X), q(Y|X))].
\end{align}

Hence, we have the following equation:

\begin{equation}
KL(D, M) = \min_{p(Y|X), q(X)} KL(p(X, Y), q(X, Y))
\end{equation}

\begin{equation}
= \min_{p(Y|X), q(X)} (KL(p(X), q(X)) + E_{p[X]} [KL(p(Y|X), q(Y|X))]).
\end{equation}

If \( p(Y|X) = q(Y|X) \), the second term of (E-6) vanishes. We obtain the following equation:

\begin{align}
KL(D, M) &= \min_{q(X,Y) \in M} KL(p(X), q(X)) \\
&= \min_{q(X) \in M} KL(p(X), q(X)).
\end{align}

References

memory model,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer
abilities,” Proceedings of the National Academy of Sciences of the United States of America,
[9] J.J. Hopfield, “Neurons with graded response have collective computational properties like those
of two-state neurons,” Proceedings of the National Academy of Sciences of the United States of
of the 9th IEEE/ACIS International Conference on Computer and Information Science, pp. 497–502,
2010.
2011.

282