Pergamon

Neural Networks, Vol. 8, No. 9, pp. 1379–1408, 1995

0893-6080/95 $9.50 + .00

0893-6080(95)00003-8

INVITED ARTICLE

Information Geometry of the EM and em Algorithms for Neural Networks

SHUN-ICHI AMARI

The Riken Frontier Research Program on the Brain Information Processing

(*Received* 31 *March* 1994; *revised and accepted* 9 *December* 1994)

Abstract—*To realize an input–output relation given by noise-contaminated examples, it is effective to use a stochastic model of neural networks. When the model network includes hidden units whose activation values are not specified nor observed, it is useful to estimate the hidden variables from the observed or specified input–output data based on the stochastic model. Two algorithms, the EM and em algorithms, have so far been proposed for this purpose. The EM algorithm is an iterative statistical technique of using the conditional expectation, and the em algorithm is a geometrical one given by information geometry. The em algorithm minimizes iteratively the Kullback–Leibler divergence in the manifold of neural networks. These two algorithms are equivalent in most cases. The present paper gives a unified information geometrical framework for studying stochastic models of neural networks, by focusing on the EM and em algorithms, and proves a condition that guarantees their equivalence. Examples include: (1) stochastic multilayer perceptron, (2) mixtures of experts, and (3) normal mixture model.*

Keywords—EM algorithm, Information geometry, Stochastic model of neural networks, Learning, Identification of neural network, e-Projection, m-Projection, Hidden variable.

1. INTRODUCTION

Neural networks have been remarked as universal approximators of nonlinear functions that can be trained from examples of input–output data. When the data includes noise, the input–output relation is described stochastically in terms of the conditional probability $p(y|x)$ of the output y when x is input. Some neural networks are stochastic in their own nature, like the Boltzmann machine, so that their behaviors are also described by probability distributions and stochastic dynamics (see Guan et al., 1994). Even when a network is deterministic, it is sometimes effective to train it as if it were a stochastic network, although it behaves deterministically in the execution mode. This is a stochastic model of (deterministic) neural networks. This suggests the usefulness of statistical ideas in neutral networks (see, e.g., Amari, 1990; Cheng & Titterington, 1994; Ripley, 1994; White, 1989, Guan et al., 1994).

Another quite different but important idea for developing a theory of neural networks originates from geometry. Let us consider a neural network including modifiable parameters (connection weights) summarized in a vector form $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. Then, the set of all the possible neural networks realized by changing $\boldsymbol{\theta}$ forms an n-dimensional manifold S, where $\boldsymbol{\theta}$ plays the role of a coordinate system of S. This is called the manifold of neural networks. Geometry of a neural manifold is useful for understanding the total capability of a class of networks. When a network is of a stochastic nature, each network is accompanied with a probability distribution $p(x; \boldsymbol{\theta})$ or a conditional probability distribution $p(y|x; \boldsymbol{\theta})$. Information geometry (Amari, 1985; Csiszár, 1975; Chentsov, 1972) connects these two sources of ideas. It originated from the information structure of a manifold of probability distributions and has been developed to be a new mathematical framework with new differential geometrical notions. It has been successfully applied to various information sciences such as statistical sciences (e.g., Amari, 1982, 1985; Barndorff-Nielsen, 1988; Barndorff-Nielsen, Cox, & Reid, 1986; Barndorff-Nielsen &

Acknowledgements: The present work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas on the Higher-Order Brain Functions from the Ministry of Education, Science, and Culture of Japan and by the Real World Computing Program, RWCP, Japan.

Requests for reprints should be sent to S. Amari, Department of Mathematical Engineering, University of Tokyo, Bunkyo-ku, Hongo, Tokyo 113, Japan.

Jupp, 1989; Murray & Rice, 1993; Kass, 1989; Amari et al., 1987), information theory (Amari & Han, 1989; Amari, 1989), systems theory (Amari, 1987a; Ohara & Amari, 1992), and many others. Applications to neural networks have already started with Amari (1991), Kawanabe and Amari (1994) and Amari, Kurata, and Nagaoka (1992).

The present paper unifies the statistical and geometrical ideas for studying neural networks including hidden units or unobservable variables, trying to establish a new information–geometric framework. In most learning neural networks, a desired input–output relation is specified by examples, but the activation values of hidden units are unobserved or left unspecified. It is convenient if they could be determined or filled in adequately such that the total network behavior is in a good agreement with the given input–output data. This is the hidden variable problem and can be regarded as a kind of credit assignment.

Two approaches have so far been proposed to solve this problem. One is information–geometric. Let us consider a manifold S of all the related (conditional) probability distributions, which might not be realizable by neural networks. Those that are realizable by neural networks form a submanifold M of neural networks in S. On the other hand, observed data suggest some distribution (e.g., one given by the empirical distribution) in S. However, because observation or specification is incomplete, the observed or specified partial data define many candidates of points in the manifold S that form a submanifold D. The best neural network is the one that minimizes the distance between the realizable M and the observed D. One can use the Kullback–Leibler divergence between D and M as the distance measure. This is a dual or alternative minimization problem, and the network in M that minimizes the divergence is selected as the optimal one. At the same time, the point in D that minimizes the divergence gives the estimated data complementing the partial observed data. This approach was proposed by Csiszár and Tusnády (1984). The same idea was used in Amari et al. (1992), Byrne (1992), and Shimodaira (1993) (see also Amari, 1991; Neal & Hinton, 1993). This may be called the *em* algorithm, because it is realized geometrically by the e-geodesic and m-geodesic projections to be explained later.

The other approach is statistical, using the *EM* (Expectation and Maximization) algorithm. This is an iterative algorithm of obtaining the maximum likelihood estimator in M, where the conditional expectation of the missing data is used to choose one candidate point in D. The *EM* algorithm has been applied to the Boltzmann machine (Byrne, 1992), the hierarchical mixture of experts (Jordan & Jacobs, 1994; see also Jacobs et al., 1991, for the mixture of experts), and others (Yuille, Stolerz, & Utans, 1994; Baldi & Chauvin, 1994; Streit & Luginbuhl, 1994). A new formulation and acceleration method is proposed in this connection (Neal & Hinton, 1993; see also Jordan & Xu, 1994; Xu, Jordan, & Hinton, 1994).

The present paper elucidates the relation between the statistical *EM* algorithm and the geometrical *em* algorithm reported in short notes (Amari, 1995). They are the same in most cases, and we prove a necessary and sufficient condition that guarantees their equivalence. We also give a simple example where the two algorithms give different solutions. By using the geometric method, it becomes much easier to understand the characteristics of the *EM* algorithm and its various versions. Moreover, we propose a learning version of the *EM* algorithm where the data are observed sequentially. A learning algorithm might be in general slower than the batch algorithm where all the accumulated data are available at a time. But the former is more flexible under a changing environment and its algorithm is much simpler.

We use a number of examples to explain our information–geometry approach. These examples are important by themselves:

1. Stochastic multilayer perceptron. Our theory gives a new learning algorithm different from the back-propagation method. The new algorithm seems more flexible with a better global convergence property than the back propagation (an independent study on the stochastic perceptron by Rumelhart, personal communication). This model is also related to that by Murray and Edwards (1994).
2. Normal mixture. This has been studied well in statistics. The radial basis function method is closely related to it. Neal and Hinton (1993) gave a new interpretation connecting it with geometry. Streit and Luginbuhl (1994) studied stochastic multilayer perceptron with the normal mixture and the *EM* algorithm.
3. Mixtures of expert networks, where the input signal space is automatically divided into regions such that signals in a region are processed by an expert network corresponding to this region. It is hidden in the input–output data, which signal should be processed by which expert network. This is a self-organizing network proposed by Jacobs et al. (1991), further generalized by Jordan and Jacobs (1994) (see Jordan & Xu, 1994; Xu, Jordan, & Hinton, 1994).

The present paper does not intend to present a detailed study of the above interesting models and related algorithms, but aims at theoretical elucidation of the geometrical structures underlying the *EM* and

em algorithms as well as an introduction to the information geometry and the *EM* algorithm. The geometrical aspect suggests various new algorithms including learning. The present paper does not study the computational aspect of the algorithms nor show computer simulated results. These remain to be studied separately in the future by applying the framework proposed in the present paper. Applications to hidden Markov random fields (Geman & Geman, 1984; Besag & Green, 1993; Künsch, Geman, & Kehagias, 1994; Baldi & Chauvin, 1994), dynamics of Boltzmann machines with asymmetric connections, etc., are also important subjects of future research.

The present paper is organized as follows. Section 2 is devoted to an introduction to the exponential family and curved exponential family, which are basic statistical models. Section 3 is a statistical preliminary where the maximum likelihood estimation is explained in terms of geometry. Readers familiar with neural networks and statistics may skip these sections. Section 4 introduces the main framework of the present theory. The set M of neural networks is shown to be embedded in the manifold S of related probability distributions as a submanifold. The partial observed data also defines a submanifold D in S. The problem is to find a network in M and filled-in data in D that minimize the Kullback–Leibler divergence between D and M. The statistical *EM* algorithm and the geometrical *em* algorithm are introduced in Section 5. They are analyzed in Section 6, the main part of the present paper, in terms of information geometry. The two algorithms are shown to be equivalent in most practical cases in Section 7. Section 8 treats learning procedures and Section 9 studies dual geodesic gradient flows in M and D. Section 10 studies the normal mixture and the mixture of expert nets. The Appendix gives a short intuitive introduction to information geometry.

2. EXPONENTIAL FAMILIES AND NEURAL NETWORKS

Exponential families of probability distributions are explained in the beginning. Let us consider a family of probability distributions of a random variable x (which may be a vector) whose probability density functions are specified by an n-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. When the probability density functions are written in the following form

$$p(x; \boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{n} \theta_i r_i(x) + k(x) - \psi(\boldsymbol{\theta})\right\}, \quad (1)$$

where $r_i(x)$, $i = 1, \ldots, n$, are functions of x, the family $S = \{p(x; \boldsymbol{\theta})\}$ is called an exponential family (Cox & Hinkley, 1974; Barndorff-Nielsen, 1978). We may treat $\mathbf{r} = (r_1, \ldots, r_n) \in \mathbf{R}^n$ as a new vector random variable whose distribution is specified

$$p(\mathbf{r}; \boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{n} \theta_i r_i - \psi(\boldsymbol{\theta})\right\} \quad (2)$$

with respect to a suitable measure $\mu(\mathbf{r})$ on $\mathbf{R}^n$. Here, the term $k(x)$ and the Jacobian density due to the transformation from x to $\mathbf{r}$ is absorbed in $\mu(\mathbf{r})$. The present paper treats such a case that x is a vector composed of a visible part $\mathbf{s}_v$ and a hidden part $\mathbf{s}_h$, $x = (\mathbf{s}_v, \mathbf{s}_h)$. In this case, we have $\mathbf{r} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h)$.

The set S of the probability distributions can be regarded as an n-dimensional manifold (space), where $\boldsymbol{\theta}$ plays the role of a coordinate system introduced in S (Amari, 1985). Any point (that is, any distribution) in S is specified by one $\boldsymbol{\theta}$. The $\boldsymbol{\theta}$ is called the natural or canonical parameter of the exponential family. To show that many important families of probability distributions are of exponential type, we give examples.

2.1. Simple Examples

Example 1. Normal distributions. Let x be a normal random variable subject to $N(\mu, \sigma^2)$, that is, with mean μ and variance σ^2,

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

The set of all the normal distributions forms a two-dimensional manifold with a coordinate system (μ, σ^2). This is an exponential family, because by putting

$$r_1 = x, \quad r_2 = x^2,$$
$$\theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = -\frac{1}{2\sigma^2},$$

the distribution is rewritten as

$$p(\mathbf{r}; \boldsymbol{\theta}) = \exp\{\theta_1 r_1 + \theta_2 r_2 - \psi(\boldsymbol{\theta})\}, \quad (3)$$

with respect to the delta measure $d\mu(\mathbf{r}) = \delta(r_2 - r_1^2) dr_1 dr_2$, where

$$\psi(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma)$$
$$= -\tfrac{1}{4}(\theta_1)^2 \theta_2^{-1} - \tfrac{1}{2}\log(-\theta_2) + \tfrac{1}{2}\log \pi.$$

Example 2. Discrete distributions. Let x be a discrete random variable taking values on the set $A = \{0, 1, \ldots, n\}$. Let $\mathbf{p} = (p_0, p_1, \ldots, p_n)$ be the

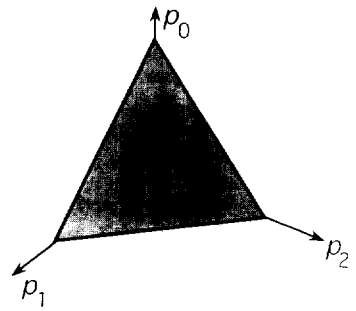

FIGURE 1. Manifold of probability distributions.

probability vector,

$$p_i = \text{Prob}\{x = i\}.$$

Then, any distribution on A is specified by a vector $\mathbf{p}$ satisfying $\Sigma p_i = 1$, $p_i > 0$. Hence, the set of all the discrete distributions $S = \{\mathbf{p}\}$ over A is an n-dimensional manifold. Figure 1 shows that S is a triangle in $\mathbf{R}^3$ when $n = 2$. This is an exponential family. Indeed, by putting

$$\theta_i = \log \frac{p_i}{p_0}, \quad i = 1, \ldots, n$$
$$r_i = \delta_i(x), \quad i = 1, \ldots, n$$

where $\delta_i(x) = 1$ when $x = i$ and $\delta_i(x) = 0$ otherwise, we have

$$p(\mathbf{r}; \boldsymbol{\theta}) = \exp\left\{\sum \theta_i r_i - \psi(\boldsymbol{\theta})\right\}, \tag{4}$$
$$\psi(\boldsymbol{\theta}) = -\log p_0 = \log\left\{1 + \sum \exp(\theta_i)\right\}.$$

Example 3. Normal mixture with hidden variables. Let us consider $k + 1$ normal distributions subject to $N(\mu_i, \sigma_i^2)$, $i = 0, 1, \ldots, k$. Let z be a discrete random variable taking values on $\{0, 1, \ldots, k\}$ with probabilities $p_i = \text{Prob}\{z = i\}$, $i = 0, \ldots, k$. Let x be a real random variable depending on z and is subject to the normal distribution $N(\mu_i, \sigma_i^2)$ when $z = i$. Then, the joint distribution of (x, z) is written as

$$p(x, z) = \sum_{i=0}^{k} \delta_i(z) p_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right\}. \tag{5}$$

Usually z is not observed, and the marginal distribution of x is

$$p(x) = \sum_i \frac{p_i}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right\}, \tag{6}$$

which is called the normal mixture distibution. This form can be regarded as a radial basis expansion of $p(x)$. The logarithm of the probability (6) is written, by eliminating $\delta_0(z)$ from $\sum_{i=1}^{k} \delta_i(z) = 1$, as

$$\begin{aligned} l(x, z) &= \log p(x, z) \\ &= \frac{\mu_o}{\sigma_0^2} x - \frac{1}{2\sigma_0^2} x^2 + \sum_{i=1}^{k} \delta_i(z) \\ &\quad \times \left\{\log \frac{p_i}{p_0} - \log \frac{\sigma_i}{\sigma_0} - \frac{\mu_i^2}{2\sigma_i^2} + \frac{\mu_0^2}{2\sigma_0^2}\right\} \\ &\quad + \sum_{i=1}^{k} x\delta_i(z)\left(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2}\right) \\ &\quad - \sum_{i=1}^{k} \delta_i(z) x^2 \left(\frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_0^2}\right) \\ &\quad + \log(p_0\sigma_0) - \frac{\mu_0^2}{2\sigma_0^2} - \log(\sqrt{2\pi}). \end{aligned}$$

Hence, by putting

$$r_{11} = x, \qquad \theta_{11} = \frac{\mu_0}{\sigma_0^2},$$

$$r_{12} = x^2, \qquad \theta_{12} = -\frac{1}{2\sigma_0^2},$$

$$r_{2i} = \delta_i(z), \qquad \theta_{2i} = \log\frac{p_i\sigma_0}{p_0\sigma_i} - \left(\frac{\mu_i^2}{2\sigma_i^2} - \frac{\mu_0^2}{2\sigma_0^2}\right),$$

$$r_{3i} = x\delta_i(z), \qquad \theta_{3i} = \frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2},$$

$$r_{4i} = x^2\delta_i(z), \qquad \theta_{4i} = -\left(\frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_0^2}\right),$$

where $i = 1, \ldots, k$, we can show that this model is an exponential family,

$$p(x, z; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \mathbf{r} - \psi(\boldsymbol{\theta})\}, \tag{7}$$

where $\boldsymbol{\theta} \cdot \mathbf{r} = \Sigma \theta_{ij} r_{ij}$. The set of the normal mixtures (6) without z is not an exponential family.

2.2. Stochastic Multilayer Perceptron

We use a stochastic perceptron to explain the general geometrical idea.

Example 4. Stochastic model of multilayer perceptron (cf. Amari, 1991). Let us consider a one hidden layer perceptron with a single output unit (Figure 2). Let x be an input vector and let $z = (z_i)$, $i = 1, \ldots, m$, be the outputs of m hidden units, z_i taking on the binary values 0 and 1. The probability of $z_i = 0, 1$, when x is input, is written as

$$p(z_i = 1 \,|\, \mathbf{x}) = \frac{\exp\{\mathbf{w}_i \cdot \mathbf{x}\}}{1 + \exp\{\mathbf{w}_i \cdot \mathbf{x}\}}$$

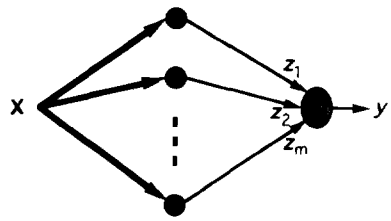

FIGURE 2. Stochastic multilayer perceptron.

where $\mathbf{w}_i$ is the synaptic weight vector of the ith hidden element. The threshold term is included in $\mathbf{w}_i$ by adding a constant input $x_0 \equiv 1$ so that $x = (x_0, x_1, \ldots, x_n)$,

$$\mathbf{w}_i \cdot \mathbf{x} = \sum_{j=1}^{n} w_{ij} x_j + w_{i0}.$$

The probability $p(z_i = 0 | \mathbf{x})$ is equal to $1 - p(z_i = 1 | \mathbf{x})$. We introduce the sigmoidal function

$$\varphi(z, u) = \frac{\exp(zu)}{1 + \exp(u)} \tag{8}$$

and summarize the probabilities as

$$p(z_i | \mathbf{x}) = \varphi(z_i, \mathbf{w}_i \cdot \mathbf{x}).$$

The output unit receives signal $\mathbf{z}$ from the hidden layer and emits a binary output y. Its probability is given, depending on the hidden signal z, by

$$p(y | \mathbf{z}) = \varphi(y, \mathbf{v} \cdot \mathbf{z}), \tag{9}$$

where $\mathbf{v}$ is the synaptic weight vector of the output unit. Here, a constant input $z_0 \equiv 1$ is added as $z = (z_0, z_1, \ldots, z_m)$ so that the bias term is also included in $\mathbf{v}$,

$$\mathbf{v} \cdot \mathbf{z} = \sum_{i=1}^{m} v_i z_i + v_0.$$

Given an input signal x, the conditional probability of variables $(y, \mathbf{z})$ is given by

$$p(y, \mathbf{z} | \mathbf{x}; \mathbf{u}) = p(y | \mathbf{z}; \mathbf{u}) \prod_{i=1}^{m} p(z_i | x; \mathbf{u}), \tag{10}$$

where the parameter $\mathbf{u}$ summarizes all of $w_1, \ldots, w_m$ and $\mathbf{v}$,

$$\mathbf{u} = (\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}).$$

The usual multilayer perceptron is a deterministic analog machine, where the outputs z_i and y are given by using the sigmoidal functions as

$$z_i = \varphi(1, \mathbf{w}_i \cdot \mathbf{x}), \quad y = \varphi(1, \mathbf{v} \cdot \mathbf{z}).$$

They are equal to the expected values of z_i and y in the stochastic model of perceptron. The stochastic model is used for the purpose of training the machine, but we can use the ordinary deterministic model for information processing after the training is completed.

When $\mathbf{x}$ is randomly generated subject to a probability distribution $q(\mathbf{x}) > 0$, $\mathbf{x} \in X$, the total probability distribution of $(y, \mathbf{z}, \mathbf{x})$ is written as

$$p(y, \mathbf{z}, \mathbf{x}; \mathbf{u}) = q(\mathbf{x}) p(y, \mathbf{z} | \mathbf{x}; \mathbf{u}), \tag{11}$$

where $q(\mathbf{x})$ may be unknown. However, by taking the logarithm,

$$l(y, \mathbf{z} | \mathbf{x}; \mathbf{u}) = \log p(y, \mathbf{z} | \mathbf{x}; \mathbf{u}), \tag{12}$$

$$l(y, \mathbf{z}, \mathbf{x}; \mathbf{u}) = \log p(y, \mathbf{z} | \mathbf{x}; \mathbf{u}) + \log q(\mathbf{x}), \tag{13}$$

we see that maximizing the logarithm of the total probability (13) with respect to $\mathbf{u}$ is the same as maximizing the logarithm of the condition probability (12).

To see if this family of distributions is of exponential family, we calculate the logarithm,

$$l(y, \mathbf{z}, \mathbf{x} | \mathbf{u}) = y \mathbf{v} \cdot \mathbf{z} + \sum_i z_i \mathbf{w}_i \cdot \mathbf{x} - \log\{1 + \exp(\mathbf{v} \cdot \mathbf{z})\} - \sum_i \{1 + \exp(\mathbf{w}_i \cdot \mathbf{x})\}.$$

Let us denote by $\mathbf{k}$ the values of $\mathbf{z}$, that is, $\mathbf{k}$ is an m-dimensional vector whose components take on $\{0, 1\}$. By introducing the delta function

$$\delta_{\mathbf{k}}(\mathbf{z}) = \begin{cases} 1, & \mathbf{z} = \mathbf{k}, \\ 0, & \mathbf{z} \neq \mathbf{k}, \end{cases} \tag{14}$$

we have the following relations

$$\mathbf{v} \cdot \mathbf{z} = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z}) \mathbf{k} \cdot \mathbf{v},$$

$$\log\{1 + \exp(\mathbf{v} \cdot \mathbf{z})\} = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z}) \log\{1 + \exp(\mathbf{k} \cdot \mathbf{v})\}$$

$$\sum_i z_i \mathbf{w}_i \cdot \mathbf{x} = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z}) \mathbf{w}_{\mathbf{k}} \cdot \mathbf{x},$$

where

$$\mathbf{w}_{\mathbf{k}} = \sum_i k_i \mathbf{w}_i, \quad \mathbf{k} = (k_i). \tag{15}$$

Note that k_0 is always equal to 1 corresponding to the

bias term $z_0 = 1$, and we put $\mathbf{w}_0 = 0$ for consistency. Here, we regard $\delta_{\mathbf{k}}(\mathbf{z})$ as 2^m random variables or a 2^m-dimensional vector random variable indexed by $\mathbf{k}$, $\mathbf{k}$ taking on 2^m values. Then, the log likelihood (the logarithm of probability considered as a function of parameter $\mathbf{u}$) is rewritten as

$$\begin{aligned} l(y, \mathbf{z}|\mathbf{x}, \mathbf{u}) = &\sum_{\mathbf{k}} \mathbf{k}\cdot\mathbf{v} y \delta_{\mathbf{k}}(\mathbf{z}) \\ &- \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z}) \log\{1 + \exp(\mathbf{k}\cdot\mathbf{v})\} \\ &+ \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z}) \mathbf{w}_{\mathbf{k}}\cdot\mathbf{x} \\ &- \sum_i \log\{1 + \exp(\mathbf{w}_i\cdot\mathbf{x})\}. \end{aligned} \tag{16}$$

We now show that the family S of the conditional distributions $\{p(y, \mathbf{z}|\mathbf{x}; \mathbf{u})\}$ of perceptron is an exponential family. To this end, we put

$$\begin{aligned} r_{1,\mathbf{k}} &= y\delta_{\mathbf{k}}(\mathbf{z}), \\ r_{2,\mathbf{k}} &= \delta_{\mathbf{k}}(\mathbf{z}), \end{aligned} \tag{17}$$

and

$$\begin{aligned} \theta_{1,\mathbf{k}} &= \mathbf{k}\cdot\mathbf{v}, \\ \theta_{2,\mathbf{k}} &= \mathbf{w}_{\mathbf{k}}\cdot\mathbf{x} - \log\{1 + \exp(\mathbf{k}\cdot\mathbf{v})\} \end{aligned} \tag{18}$$

We then have, from eqn (16),

$$p(\mathbf{r}|\mathbf{x}; \mathbf{u}) = \exp\{\mathbf{r}\cdot\boldsymbol{\theta} - \psi\},$$

where $\mathbf{r} = (r_{1,\mathbf{k}}, r_{2,\mathbf{k}})$, $\boldsymbol{\theta} = (\theta_{1,\mathbf{k}}, \theta_{2,\mathbf{k}})$,

$$\boldsymbol{\theta}\cdot\mathbf{r} = \sum_{\mathbf{k}} \theta_{1,\mathbf{k}} r_{1,\mathbf{k}} + \sum_{\mathbf{k}} \theta_{2,\mathbf{k}} r_{2,\mathbf{k}}$$

and

$$\varphi(\boldsymbol{\theta}) = \sum_i \log\{1 + \exp(\mathbf{w}_i\cdot\mathbf{x})\}.$$

Hence, S is an exponential family.

2.3. Repeated Observation and Product Space

Let $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T$ be T independent observations from the same distribution $p(\mathbf{r}; \boldsymbol{\theta})$. Their joint distribution is given by

$$p(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T; \boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{r}_t; \boldsymbol{\theta}).$$

When $p(\mathbf{r}; \boldsymbol{\theta})$ is of exponential type, we have from eqn (2),

$$p(\mathbf{r}_1, \ldots, \mathbf{r}_T; \boldsymbol{\theta}) = \exp\left\{\left(\sum \mathbf{r}_t\right)\cdot\boldsymbol{\theta} - T\varphi(\boldsymbol{\theta})\right\}.$$

Hence, this is again of the same exponential type distribution if we replace the random variable $\mathbf{r}$ by the sum $\Sigma\mathbf{r}_t$ of all the observations.

However, when conditional distributions $p(y, z|\mathbf{x})$ are considered, we encounter a different situation. Let $(y_t, z_t, \mathbf{x}_t)$, $t = 1, \ldots, T$, be T independent observations from the same network of exponential type. In this case, as seen from eqn (18), the canonical parameter $\boldsymbol{\theta}$ depends on $\mathbf{x}$ as is shown

$$p\{\mathbf{r}; \boldsymbol{\theta}(\mathbf{x})\} = \exp\{\boldsymbol{\theta}(\mathbf{x})\cdot\mathbf{r} - \psi\}, \tag{19}$$

where $\mathbf{r} = \mathbf{r}(y, \mathbf{z})$. A neural network is trained by giving various input signals $\mathbf{x}$ and corresponding outputs y, where $\mathbf{x}$ is not fixed. When T independent data $(\mathbf{x}_1, y_1, \mathbf{z}_1), \ldots, (\mathbf{x}_T, y_T, \mathbf{z}_T)$ are given, the joint conditional distribution is written as

$$\begin{aligned} &p\{(y_1, \mathbf{z}_1), \ldots, (y_T, \mathbf{z}_T)|\mathbf{x}_1, \ldots, \mathbf{x}_T\} \\ &= \prod_{t=1}^{T} p\{\mathbf{r}_t; \boldsymbol{\theta}(\mathbf{x}_t)\} = \exp\left\{\sum \boldsymbol{\theta}(\mathbf{x}_t)\cdot\mathbf{r}_t - \sum \psi_t\right\}. \end{aligned} \tag{20}$$

This forms an exponential family of larger dimensions. Let

$$S_t = \{p(\mathbf{r}_t; \boldsymbol{\theta}(\mathbf{x}_t))\} \tag{21}$$

be the manifold of conditional probability distributions corresponding to the tth input $\mathbf{x}_t$. Then, the joint conditional distributions (20) are given by the product space

$$S_T^* = S_1 \times S_2 \times \cdots \times S_T. \tag{22}$$

Here, the random variable $\mathbf{r}_T^*$ in S_T^* is

$$\mathbf{r}_T^* = (\mathbf{r}_1, \ldots, \mathbf{r}_T),$$

the canonical parameter in S_T^* is

$$\boldsymbol{\theta}_T^* = \{\boldsymbol{\theta}(\mathbf{x}_1), \ldots, \boldsymbol{\theta}(\mathbf{x}_T)\},$$

and eqn (20) is rewritten as

$$p(\mathbf{r}_T^*; \boldsymbol{\theta}_T^*) = \exp\{\boldsymbol{\theta}_T^*\cdot\mathbf{r}_T^* - \psi\theta_T^*\}, \tag{23}$$

where

$$\begin{aligned} \boldsymbol{\theta}_T^*\cdot\mathbf{r}_T^* &= \sum_{t=1}^{T} \boldsymbol{\theta}_t\cdot\mathbf{r}_t, \\ \varphi(\boldsymbol{\theta}_T^*) &= \sum_{t=1}^{T} \varphi\{\boldsymbol{\theta}(\mathbf{x}_t)\}. \end{aligned}$$

It should be remarked that $\boldsymbol{\theta}_T^* = \{\boldsymbol{\theta}(\mathbf{x}_1), \ldots, \boldsymbol{\theta}(\mathbf{x}_T)\}$ cannot take arbitrary values, because all of $\boldsymbol{\theta}(\mathbf{x}_t)$, $t = 1, \ldots, T$, are restricted by the common network parameters $\mathbf{w}_i$'s and $\mathbf{v}$ in the form of eqn (18). Hence, possible $\boldsymbol{\theta}_T^*$ is restricted in a subregion of S_T^*. This is a curved exponential family explained in the next section.

2.4. Curved Exponential Families

A smaller set of distributions that occupies a part of an exponential family S is called a curved exponential family when they form a submanifold M embedded in S. Let $S = \{p(\mathbf{r}; \boldsymbol{\theta})\}$ be an n-dimensional exponential family and let M be its m-dimensional submanifold. Let $\mathbf{u} = (u_1, \ldots, u_m)$, $m \leq n$, be a parameter or a coordinate system of M. A curved exponential family $M = \{p(\mathbf{r}; \boldsymbol{\theta}(\mathbf{u}))\}$ consists of the distributions

$$p(\mathbf{r}; \boldsymbol{\theta}(\mathbf{u})) = \exp\{\boldsymbol{\theta}(\mathbf{u}) \cdot \mathbf{r} - \psi(\boldsymbol{\theta}(\mathbf{u}))\}, \tag{24}$$

where $\boldsymbol{\theta}(\mathbf{u})$ is a function of $\mathbf{u}$. So the distributions are parameterized by $\mathbf{u}$. Any distribution $p(\mathbf{r}; \boldsymbol{\theta}(\mathbf{u}))$ in M belongs to S, and $\boldsymbol{\theta}(\mathbf{u})$ is the coordinates in S of the point $\mathbf{u}$ of M. In terms of geometry, M is a submanifold of S (Figure 3) composed of points written as

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{u}). \tag{25}$$

This is a parametric representation of M, where $\mathbf{u}$ is an inner coordinate system of M.

Example 5. Normal multiplication model. Let ε be a normal random variable subject to $N(0, 1)$, that is, the normal distribution with mean 0 and variance 1. We input a signal of the magnitude 1 to some unknown system to know how the signal is amplified or damped. The signal is contaminated by noise ε so that $s = 1 + \varepsilon$ is input. It is then multiplied by an unknown quantity u that we want to know. Therefore, the final response signal is $x = us = u(1 + \varepsilon)$. The problem is to estimate the amplification factor u. The observed

$$x = u(1 + \varepsilon)$$

is subject to $N(u, u^2)$. The set M of all such distributions is a curved exponential family embedded in $S = \{N(\mu, \sigma^2)\}$. Indeed, when the distribution is in M, its mean μ and variance σ^2 are not independent but are specified by a common u as

$$\mu = u, \quad \sigma^2 = u^2,$$

so that M forms a curve in S (Figure 4). Here u is the coordinate of M. In terms of the natural parameter $\boldsymbol{\theta}$, the coordinates of points in M satisfy

$$\theta_1 = \frac{1}{u}, \quad \theta_2 = -\frac{1}{2u^2}. \tag{26}$$

The shape of M is a parabola in S.

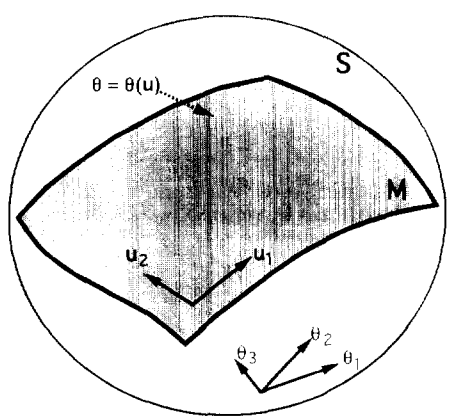

FIGURE 3. Curved exponential family.

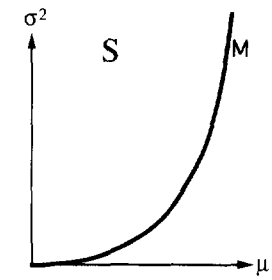

FIGURE 4. Example of curved exponential family.

Example 6. Multiple observation in stochastic perceptron. Let us consider multiple observations in the stochastic multilayer perceptron. The probability distributions are written as

$$p(\mathbf{r}_T^*; \boldsymbol{\theta}_T^*) = \exp\{\boldsymbol{\theta}_T^* \cdot \mathbf{r}_T^* - \psi\}. \tag{27}$$

Here, $\boldsymbol{\theta}_T^* = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T)$ is the coordinates of the product space S_T^*. Let $\mathbf{u}$ be the parameters of the underlying neural network, consisting of $\mathbf{w}_i$'s and $\mathbf{v}$. Then, the distribution realized by the neural net of parameter $\mathbf{u}$ has the $\boldsymbol{\theta}_T^*$ coordinates given by

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u}), \quad t = 1, \ldots, T. \tag{28}$$

When x_t, $t = 1, \ldots, T$, are given, $\mathbf{u}$ is the only free parameter in $\boldsymbol{\theta}_T^*$. Hence, the set M_T^* of probability distributions of stochastic perceptron forms a submanifold in S_T^*. This is a curved exponential family with the inner coordinate system $\mathbf{u}$. The dimension number of S_T^* increases without limit as

T tends to infinity, but that of M_T^* is the number of the network parameters $\mathbf{u}$ and is fixed.

3. GEOMETRY OF OBSERVATION AND ESTIMATION

3.1. Expectation Parameter

We consider the geometry of observation in an exponential family S or in a curved exponential family M embedded in S. Here it is presumed that all the random variables are visible, as a preliminary of the hidden variable case. A distribution $p(\mathbf{r}; \boldsymbol{\theta})$ is specified by parameter $\boldsymbol{\theta}$ in S, where $\mathbf{r}$ is the random variable. Let $\boldsymbol{\eta}$ be the expectation of the random variable $\mathbf{r}$ with respect to the distribution $p(\mathbf{r}; \boldsymbol{\theta})$,

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\mathbf{r}] = \int \mathbf{r} p(\mathbf{r}; \boldsymbol{\theta}) d\mu(\mathbf{r}), \tag{29}$$

where E_θ denotes the expectation with respect to $p(\mathbf{r}; \boldsymbol{\theta})$. By differentiating the identity

$$\int p(\mathbf{r}; \boldsymbol{\theta}) d\mu(\mathbf{r}) = \int \exp\{\boldsymbol{\theta} \cdot \mathbf{r} - \psi(\boldsymbol{\theta})\} d\mu(\mathbf{r}) = 1$$

with respect to $\boldsymbol{\theta}$, it is easily shown that $\boldsymbol{\eta}(\boldsymbol{\theta})$ is given by

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta}). \tag{30}$$

Moreover, it is known that the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is one-to-one. Therefore, this $\boldsymbol{\eta}$ can be used as another coordinate system (parameter) of S to specify the distributions. We call $\boldsymbol{\eta}$ the expectation parameter. Equation (30) is inverted as

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta}). \tag{31}$$

Let $\varphi(\boldsymbol{\eta})$ be the negative of the entropy of the distribution $p^*(\mathbf{r}; \boldsymbol{\eta})$ specified by $\boldsymbol{\eta}$,

$$\varphi(\boldsymbol{\eta}) = \int p^*(\mathbf{r}; \boldsymbol{\eta}) \log p^*(\mathbf{r}; \boldsymbol{\eta}) d\mu(\mathbf{r}),$$

where

$$p^*\{\mathbf{r}; \boldsymbol{\eta}(\boldsymbol{\theta})\} = p(\mathbf{r}; \boldsymbol{\theta}).$$

Then eqn (31) is given explicitly by

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \varphi(\boldsymbol{\eta})$$

(see Amari, 1985).

Example 7. Examples of the $\boldsymbol{\eta}$-coordinates. In the case of normal distributions (Example 1), the expectation parameter is

$$\eta_1 = E[x] = \mu = \frac{\theta_1}{2\theta_2}, \tag{32}$$

$$\eta_2 = E[x^2] = \mu^2 + \sigma^2 = \left(\frac{\theta_1}{2\theta_2}\right)^2 - \frac{1}{2\theta_2}. \tag{33}$$

In the case of the discrete distributions (Example 2),

$$\eta_i = E[\delta_i(x)] = p_i = \frac{\exp(\theta_i)}{1 + \sum \exp(\theta_i)}. \tag{34}$$

In the case of $S = \{p(y, \mathbf{z}; \boldsymbol{\theta})\}$ in Example 4 of stochastic perceptron,

$$\boldsymbol{\eta} = (\eta_{1,\mathbf{k}}, \eta_{2,\mathbf{k}})$$

is given by

$$\begin{aligned} \eta_{1,\mathbf{k}} &= E[y \delta_{\mathbf{k}}(\mathbf{z})] = \varphi(1, \mathbf{v} \cdot \mathbf{k}) \Pi \varphi(k_i, \mathbf{w}_i \cdot \mathbf{x}) \\ \eta_{2,\mathbf{k}} &= E[\delta_{\mathbf{k}}(\mathbf{z})] = \Pi \varphi(k_i, \mathbf{w}_i \cdot \mathbf{x}). \end{aligned} \tag{35}$$

The expectation parameter is useful for studying the maximum likelihood estimator (m.l.e.). Let $\check{\mathbf{r}}$ be an observed value of the random variable $\mathbf{r}$ in an exponential family $S = \{p(\mathbf{r}; \boldsymbol{\theta})\}$. In this section, it is assumed that all the random variables are observable. Let $\hat{\boldsymbol{\theta}}$ be the m.l.e. that maximizes the likelihood function $p(\check{\mathbf{r}}; \boldsymbol{\theta})$ or its logarithm

$$l(\check{\mathbf{r}}; \boldsymbol{\theta}) = \boldsymbol{\theta} \cdot \check{\mathbf{r}} - \psi(\boldsymbol{\theta}). \tag{36}$$

By differentiating eqn (36), the m.l.e. $\hat{\boldsymbol{\theta}}$ is proved to satisfy

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \check{\mathbf{r}} - \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\hat{\boldsymbol{\theta}}) = \check{\mathbf{r}} - \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}) = 0.$$

The m.l.e. $\hat{\boldsymbol{\theta}}$ is given by solving the above equation. When we use the $\boldsymbol{\eta}$-coordinates, the m.l.e. is directly given by the observed value itself,

$$\hat{\boldsymbol{\eta}} = \check{\mathbf{r}}. \tag{37}$$

If we want to obtain $\hat{\boldsymbol{\theta}}$, we need to calculate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\eta}})$,

$$\hat{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\eta}} \varphi(\hat{\boldsymbol{\eta}}).$$

The $\hat{\boldsymbol{\eta}}$ or the m.l.e. $\hat{\boldsymbol{\theta}}$ in the $\boldsymbol{\theta}$-coordinate determines a distribution $p(\mathbf{r}; \hat{\boldsymbol{\theta}})$, that is, a point in S. We call it the observed point. Its $\boldsymbol{\eta}$-coordinates are the observed

value of $\mathbf{r}$ itself. This looks rather trivial in an exponential family, but is not so in a curved exponential family, in particular, in repeated observations, as is shown in the next subsection.

3.2. Repeated Observation and Observed Point

Let $\mathbf{r}_1, \ldots, \mathbf{r}_T$ be T independent random variables, where $\mathbf{r}_t$ is subject to the distribution $p(\mathbf{r}_t; \boldsymbol{\theta}_t)$ in an exponential family S_t. The joint probability distribution is given by

$$p(\mathbf{r}_1, \ldots, \mathbf{r}_T; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T) = \prod_{t=1}^{T} p(\mathbf{r}_t; \boldsymbol{\theta}_t),$$

as

$$p(\mathbf{r}_T^*; \boldsymbol{\theta}_T^*) = \exp\{\boldsymbol{\theta}_T^* \cdot \mathbf{r}_T^* - \psi\}.$$

This is an extended exponential family S_T^*. Its expectation coordinates $\boldsymbol{\eta}_T^* = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)$ are given by

$$\boldsymbol{\eta}_t = E_{\boldsymbol{\theta}_t}[\mathbf{r}_t]. \tag{38}$$

When $\check{\mathbf{r}}_1, \ldots, \check{\mathbf{r}}_T$ are observed, we have the observed point in S_T^* whose expectation coordinates are

$$\hat{\boldsymbol{\eta}}_T^* = (\check{\mathbf{r}}_1, \ldots, \check{\mathbf{r}}_T). \tag{39}$$

The corresponding m.l.e. $\hat{\boldsymbol{\theta}}_T^*$ is given by solving

$$\hat{\boldsymbol{\theta}}_t = \frac{\partial \varphi(\hat{\boldsymbol{\eta}}_t)}{\partial \boldsymbol{\eta}}.$$

In the case of neural networks, $\boldsymbol{\theta}_t$ is determined by the tth input $\mathbf{x}_t$ and the common network parameter $\mathbf{u}$,

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u}). \tag{40}$$

The corresponding $\boldsymbol{\eta}$-coordinates are written as

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}(\mathbf{x}_t, \mathbf{u}) = E_{\boldsymbol{\theta}_t}[\mathbf{r}]. \tag{41}$$

The eqn (40) or (41) defines a curved exponential family M^* in S_T^*. The observed point $\hat{\boldsymbol{\eta}}_T^*$ or $\hat{\boldsymbol{\theta}}_T^*$ is not necessarily included in M because it does not satisfy eqn (40) or (41). Hence, we need to project it to M to obtain the m.l.e. $\hat{\mathbf{u}}$. This is studied in the next subsection. Before that, we show a simple case of i.i.d. observations.

Let us consider the case where all the $\mathbf{r}_t$ are subject to the same distribution $p(\mathbf{r}; \boldsymbol{\theta})$, so that

$$\boldsymbol{\theta} = \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \cdots = \boldsymbol{\theta}_T$$

holds. This is the case of repeated observations in statistics, and the joint distribution is summarized in

$$p(\mathbf{r}_1, \ldots, \mathbf{r}_T; \boldsymbol{\theta}) = \exp\left\{\sum_{t=1}^{T} \mathbf{r}_t \cdot \boldsymbol{\theta} - T\psi(\boldsymbol{\theta})\right\}. \tag{42}$$

It is then possible to reduce the extended direct product space S_T^* into a single S.

To this end, we introduce the new random variable,

$$\bar{\mathbf{r}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{r}_t, \tag{43}$$

which is the arithmetic mean of T random variables. The joint probability (42) can be transformed into the distribution of $\bar{\mathbf{r}}$,

$$p(\bar{\mathbf{r}}; \boldsymbol{\theta}) = \exp\{T(\bar{\mathbf{r}} \cdot \boldsymbol{\theta}) + k(\bar{\mathbf{r}}) - \psi(\boldsymbol{\theta})\}, \tag{44}$$

where $k(\bar{\mathbf{r}})$ is given rise to by the transformation of random variables from $(\mathbf{r}_1, \ldots, \mathbf{r}_T)$ to $\bar{\mathbf{r}}$,

$$\exp\{Tk(\bar{\mathbf{r}})\} = \int_{\bar{\mathbf{r}}} \prod d\mu(\mathbf{r}_t),$$

the integration being taken over the region of $(\mathbf{r}_1, \ldots, \mathbf{r}_T)$ where the arithmetic mean of $\mathbf{r}_1, \ldots, \mathbf{r}_T$ is $\bar{\mathbf{r}}$. However, eqn (44) shows that the probability of $(\mathbf{r}_1, \ldots, \mathbf{r}_T)$ is given through their arithmetic mean $\bar{\mathbf{r}}$, implying that $\bar{\mathbf{r}}$ is a sufficient statistic for estimating $\boldsymbol{\theta}$ or $\mathbf{u}$ (see standard textbooks on statistics, for example, Cox & Hinkley, 1974; Rao, 1973). Hence, the distributions of $\bar{\mathbf{r}}$ again form the same type of exponential family as S, except for the scale factor T. The term $k(\bar{\mathbf{r}})$ can be eliminated by using the dominating measure

$$d\bar{\mu}(\bar{\mathbf{r}}) = \exp\{Tk(\bar{\mathbf{r}})\} d\mu.$$

So it is possible to discuss repeated observations and estimation in the framework of the manifold S without referring to the product space S_T^*. However, this holds only in the i.i.d. case, and we need to consider S_T^* in the general case.

The m.l.e. $\hat{\boldsymbol{\theta}}$ from the observed data $\mathbf{r}_1, \ldots, \mathbf{r}_T$ is given by maximizing $p(\bar{\mathbf{r}}; \boldsymbol{\theta})$ or $\bar{\mathbf{r}} \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta})$ in an exponential family S without referring to S_T^*. By differentiation, it is given by the solution of

$$\bar{\mathbf{r}} = \frac{\partial}{\partial \boldsymbol{\theta}} (\hat{\boldsymbol{\theta}}). \tag{45}$$

In terms of the $\{\eta\}$-coordinates, this is simply given by

$$\hat{\eta} = \bar{r} \tag{46}$$

when $\bar{r}$ is observed. We call the distribution $p(r; \hat{\theta})$ or the point $\hat{\eta}$ in S in the η-coordinate system the observed point. The observed data $r_1, \ldots, r_T$ are simply summarized into the single observed point (i.e., the m.l.e.) $\hat{\eta}$ in S in the η-coordinate system without any loss of information in the sense of Fisher.

3.3. Estimation in Curved Exponential Family

Let us consider the m.l.e. in a curved exponential family M embedded in S. In the i.i.d. case, the observed data $r_1, \ldots, r_T$ are summarized in the observed point $\hat{\eta} = \bar{r}$ in S. This is a point of S but it does not necessarily belong to M. The m.l.e. $\hat{u}$ or the corresponding distribution $\theta(\hat{u}) \in M$ is given by maximizing the log likelihood

$$l(\bar{r}; u) = \bar{r} \cdot \theta(u) - \psi\{\theta(u)\} \tag{47}$$

with respect to u.

It is known that maximizing the likelihood is equivalent to minimizing the Kullback–Leibler divergence. The KL divergence from a distribution $p(r; \theta')$ to another distribution $p(r; \theta)$ is given by

$$\begin{aligned} K(\theta' \| \theta) &= \int p(r; \theta') \log \frac{p(r; \theta')}{p(r; \theta)} \, d\mu(r) \\ &= \varphi(\theta') - \eta' \cdot \theta + \psi(\theta), \end{aligned} \tag{48}$$

where

$$-\varphi(\theta') = H(\theta') = -\int p(r; \theta') \log p(r; \theta') d\mu(r) \tag{49}$$

is the entropy of the distribution $p(r; \theta')$ and η' is the η-coordinates of θ'. Therefore, maximizing the log likelihood (47) is equivalent to minimizing the KL divergence $K(\hat{\theta} \| \theta(u))$ from the observed point $\hat{\theta}$ to points $\theta(u)$ belonging to M. This implies that the m.l.e. $\theta(\hat{u})$ is the point in M that is closest to the observed point $\hat{\theta}$ or $\hat{\eta}$ in S (Figure 5). Geometrically, the m.l.e. $\theta(\hat{u})$ is given by m-projecting $\hat{\theta}$ to M, as is explained in the Appendix. It is given by solving the likelihood equation

$$\frac{\partial}{\partial u} \{\bar{r} \cdot \theta(u) - \psi(\theta(u))\} = 0.$$

Because of

$$\eta(u) = \frac{\partial \psi\{\theta(u)\}}{\partial \theta},$$

by introducing the matrix

$$B(u) = \frac{\partial \theta(u)}{\partial u} = \left(\frac{\partial \theta_i(u)}{\partial u^j}\right), \tag{50}$$

the likelihood equation is given by

$$B(u)\{\bar{r} - \eta(u)\} = 0. \tag{51}$$

In the direct space S_T^*, the KL divergence from $\theta_T'^*$ to θ_T^* is decomposed as

$$\begin{aligned} K(\theta_T'^* \| \theta_T^*) &= \sum_{t=1}^{T} K(\theta_t' \| \theta_t) \\ &= -\sum H(\theta_t') - \sum \eta_t' \cdot \theta_t + \sum \psi(\theta_t). \end{aligned} \tag{52}$$

Hence, the m.l.e. $\hat{u}$ is the point $\theta_T^*(\hat{u})$ in M^* that is closest to the observed point $\hat{\eta}_T^* = \bar{r}^* \in S_T^*$. The likelihood equation is

$$\sum_{t=1}^{T} B(x_t, u)\{\check{r}_t - \eta(x_t, u)\} = 0, \tag{53}$$

where

$$B(x_t, u) = \frac{\partial \theta(x_t, u)}{\partial u}. \tag{54}$$

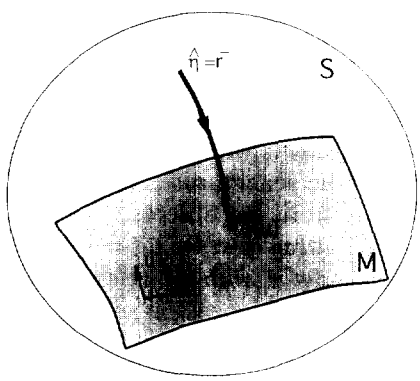

FIGURE 5. Maximum likelihood estimation.

4. HIDDEN VARIABLES AND DATA SUBMANIFOLD D

4.1. Partial Observation

The present paper treats the case where some parts of random variables cannot be observed. For example, the activations of hidden neurons might not be observed. Moreover, in neural learning, only a

desired input–output relation is specified without any specification on the desired outputs of hidden neurons. They should be determined adequately such that the input–output relation is approximated well. This subsection studies the simplest case where some constituents of the sufficient statistics $\mathbf{r}$ are not observed.

The sufficient statistic $\mathbf{r}$ is a vector function of basic random variables, say $\mathbf{x}$, y, and $\mathbf{z}$ in the case of neural networks. Some of them are observable or specifiable but some are not. For example, $\mathbf{x}$ and y are observable but $\mathbf{z}$ is not. In general, we represent the sufficient statistic $\mathbf{r}$ in terms of $\mathbf{r} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h)$, where $\mathbf{s}_v$ is observed like $\mathbf{x}$ and y but $\mathbf{s}_h$ is not. Because $\mathbf{s}_h$ is missing, we cannot identify the observed point $\hat{\boldsymbol{\eta}} = \mathbf{r}$ uniquely. Instead, given $\mathbf{s}_v$, we have candidates of the observed point $\hat{\boldsymbol{\eta}} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h)$, where $\mathbf{s}_h$ takes arbitrary values. The candidate points form a submanifold D called the observed data submanifold. It is defined as follows:

$$D = \{\hat{\boldsymbol{\eta}} \mid \hat{\boldsymbol{\eta}} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h); \ \mathbf{s}_v \text{ being fixed at the observed value and } \mathbf{s}_h \text{ taking arbitrary values}\}. \tag{55}$$

In many cases, $\mathbf{r}$ is linear in $\mathbf{s}_h$, so that $t\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h^1) + (1-t)\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h^2) \in D$ for $0 \le t \le 1$ when $\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h^1)$, $\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h^2) \in D$. When $\mathbf{s}_h$ is a discrete vector variable, D consists of a number of candidate points, where hidden $\mathbf{s}_h$ is subject to a probability distribution. In this case, we extend D into a linear submanifold in the $\boldsymbol{\eta}$-coordinates by taking all the linear combinations of the candidate points in it.

We first consider the simplest case, where $\mathbf{r}$ itself is divided into two parts, the visible and hidden parts,

$$\mathbf{r} = (\mathbf{r}_v, \mathbf{r}_h), \tag{56}$$

where $\mathbf{r}_v$ can be observed but $\mathbf{r}_h$ are hidden. So $\mathbf{r}_v = \mathbf{s}_v$ and $\mathbf{r}_h = \mathbf{s}_h$. Because $\mathbf{r}_h$ part is missing, the observed $\mathbf{r}_v$ gives the observed data submanifold in the following simple way,

$$D = \{\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_v, \hat{\boldsymbol{\eta}}_h) \mid \hat{\boldsymbol{\eta}}_v = \mathbf{r}_v, \hat{\boldsymbol{\eta}}_h : \text{arbitrary}\}. \tag{57}$$

This forms an h-dimensional submanifold in S (Figure 6), called the data submanifold or observed submanifold based on partial observation, where h is the dimension number of $\mathbf{r}_h$. It should be remarked that D is linear in the $\boldsymbol{\eta}$-coordinate system and $\boldsymbol{\eta}_h$ plays the role of an inner coordinate system of D. When all the data are observed, there are no hidden components and D reduces to the single observed point $\hat{\boldsymbol{\eta}}$.

In a general case where $\mathbf{r}$ is not divided in $(\mathbf{r}_v, \mathbf{r}_h)$ but D is still linear in the $\boldsymbol{\eta}$-coordinates, we have a linear transformation of $\boldsymbol{\eta}$ to $\tilde{\boldsymbol{\eta}}$ by using a nonsingular matrix A

$$\tilde{\boldsymbol{\eta}} = A\boldsymbol{\eta}, \quad \tilde{\mathbf{r}} = A\mathbf{r}, \tag{58}$$

such that $\tilde{\boldsymbol{\eta}}$ is divided into $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\eta}}_v, \tilde{\boldsymbol{\eta}}_h)$, where $\tilde{\boldsymbol{\eta}}_v$ is given from the observed variables whereas $\tilde{\boldsymbol{\eta}}_h$ is arbitrary. However, the $\boldsymbol{\eta}$-coordinates and $\boldsymbol{\theta}$-coordinates are dually coupled. Therefore, when we use $\tilde{\boldsymbol{\eta}}$-coordinates, we need to use the related $\tilde{\boldsymbol{\theta}}$-coordinates given

$$\tilde{\boldsymbol{\theta}} = A^{-1}\boldsymbol{\theta} \tag{59}$$

contravariantly to (58). Then, the probability distribution of $\tilde{\mathbf{r}}$ is written as

$$p(\tilde{\mathbf{r}}; \tilde{\boldsymbol{\theta}}) = \exp\{\tilde{\boldsymbol{\theta}} \cdot \tilde{\mathbf{r}} - \tilde{\psi}(\tilde{\boldsymbol{\theta}})\},$$

so that $\tilde{\boldsymbol{\theta}}$ is the new canonical parameter, and

$$\tilde{\boldsymbol{\eta}} = E_{\tilde{\boldsymbol{\theta}}}[\tilde{\mathbf{r}}] = \frac{\partial \tilde{\psi}}{\partial \tilde{\boldsymbol{\theta}}}$$

is the expectation parameter such that $\tilde{\mathbf{r}} = (\tilde{\mathbf{r}}_v, \tilde{\mathbf{r}}_h)$.

We show two examples.

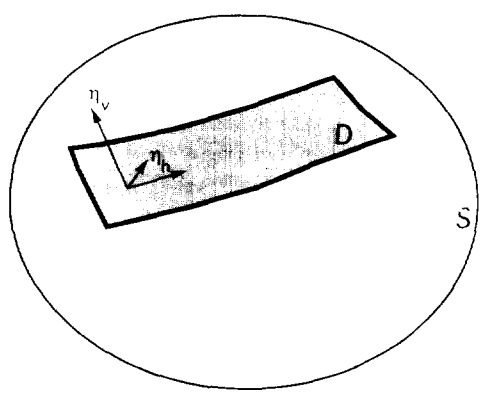

FIGURE 6. Data submanifold D : η =fixed.

Example 8. Partial observation in the normal multiplication model. Let $x_1, \ldots, x_T$ be independent normal random variables subject to $N(u, u^2)$. The sufficient statistics are $\tilde{\mathbf{r}} = (r_v, r_h)$

$$r_v = \frac{1}{T}\sum x_t,$$
$$r_h = \frac{1}{T}\sum x_t^2,$$

where we assume that r_v is observed but r_h is hidden. The statistical model $M = \{N(u, u^2)\}$ is a curved exponential family embedded in $S = \{N(\mu, \sigma^2)\}$

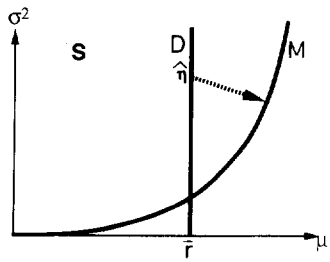

FIGURE 7. Example of data submanifold D and observed point $\hat{\eta}$.

given in Example 5. The corresponding η-coordinates are

$$\eta_v = \mu = \mu,$$
$$\eta_h = \mu^2 + \sigma^2 = 2u^2.$$

When $r_v = \tilde{r}$ is observed, the data manifold is given

$$D = \{N(\mu, \sigma^2) \mid \mu = \tilde{r},\ \sigma^2 \text{ is arbitrary}\}$$

(see Figure 7). When both Σx_t and Σx_t^2 are observed, we have a unique observed point $\hat{\eta} = \hat{\mathbf{r}}$. The best estimator is given by m-projecting $\hat{\eta}$ to M. In the hidden case, we do not know the exact observed point but we know only that it is in D.

Example 9. Data submanifold of the stochastic perceptron. The random variable $\mathbf{r}(t)$ at time t is composed of $(r_{1,\mathbf{k},t}, r_{2,\mathbf{k},t})$, and the observed point $\hat{\eta}_t = (r_{1,\mathbf{k},}, \mathbf{r}_{2,\mathbf{k},t})$ is written in terms of $(y_t, \mathbf{z}_t)$ as

$$\hat{\eta}_{1,\mathbf{k}} = r_{1,\mathbf{k},t} = y_t \delta_{\mathbf{k}}(\mathbf{z}_t),$$
$$\hat{\eta}_{2,\mathbf{k}} = r_{2,\mathbf{k},t} = \delta_{\mathbf{k}}(\mathbf{z}_t)$$

where y_t is observable but $\delta_{\mathbf{k}}(\mathbf{z}_t)$ are not. There are many candidate points depending on the value of $\mathbf{z}_t$. We take their linear combination in the η-coordinates. Let $\alpha_{\mathbf{k}}$ be the weight attached to $\delta_{\mathbf{k}}(\mathbf{z}_t)$ for forming their linear combinations. Then, the observed data submanifold D_t is

$$D_t = \{\eta \mid \eta_{1,\mathbf{k}} = y_t \alpha_{\mathbf{k}},\ \eta_{2,\mathbf{k}} = \alpha_{\mathbf{k}}\}, \tag{60}$$

where $\alpha_{\mathbf{k}}$'s are the free parameters satisfying

$$\sum_{\mathbf{k}} \alpha_{\mathbf{k}} = 1.$$

The data submanifold is

$$D_T^* = D_1 \times \cdots \times D_T,$$

which is a linear submanifold of S_T^* in the η-coordinates. We may interpret that $\alpha_{\mathbf{k}}$ denotes the probability of $\mathbf{z}_t = \mathbf{k}$, which is unobservable.

5. *EM* ALGORITHM AND *em* ALGORITHM

5.1. The *EM* Algorithm

It is required to obtain a good estimator $\hat{\mathbf{u}}$ from partially observed data D or D_T^*. Two algorithms are known to solve the hidden variable problem. One is the *EM* algorithm. The *EM* algorithm is a statistical technique for calculating the m.l.e. from partially observed data (Dempster, Laird, & Rubin, 1977). It may be regarded as an iterative procedure of estimating both the true parameter $\mathbf{u}$ and the missing data at the same time. We first show the original idea of the *EM* algorithm and its generalization called the *GEM* algorithm (Dempster, Laird, and Rubin 1977; Baum et al., 1970).

Let $M = \{p(\mathbf{r}; \boldsymbol{\theta}(\mathbf{u}))\}$ be a curved exponential family from which data $\mathbf{r}$ is generated. When the data $\check{\mathbf{r}}$ is observed, the m.l.e. $\hat{\mathbf{u}}$ is obtained by maximizing the log likelihood of the observed data,

$$\log p\{\check{\mathbf{r}}; \boldsymbol{\theta}(\mathbf{u})\} = \boldsymbol{\theta}(\mathbf{u}) \cdot \check{\mathbf{r}} - \psi\{\boldsymbol{\theta}(\mathbf{u})\}.$$

When the sufficient statistic $\mathbf{r} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h)$ includes hidden part $\mathbf{s}_h$, the complete data $\check{\mathbf{r}}$ is not available. It is one idea to estimate the unknown $\check{\mathbf{r}}$ based on the observed $\mathbf{s}_v$. If we have a candidate distribution specified by $\mathbf{u}'$, we can use the conditional expectation

$$\hat{\mathbf{r}}(\mathbf{u}') = E[\mathbf{r} \mid \mathbf{s}_v; \boldsymbol{\theta}(\mathbf{u}')] \tag{61}$$

to obtain a guess of $\hat{\mathbf{r}}$. This depends on the observed $\mathbf{s}_v$ and the candidate distribution $\boldsymbol{\theta}(\mathbf{u}')$, because the conditional expectation is taken with respect to the distribution $p\{\mathbf{r}; \boldsymbol{\theta}(\mathbf{u}')\}$. Then, the log likelihood/function is estimated by

$$E[\log p(\mathbf{r}; \mathbf{u}) \mid \mathbf{s}_v; \boldsymbol{\theta}(\mathbf{u}')] = \theta(\mathbf{u}) \cdot \hat{\mathbf{r}}(\mathbf{u}') - \psi\{\boldsymbol{\theta}(\mathbf{u})\} \tag{62}$$

where $\hat{\mathbf{r}}(\mathbf{u}')$ is given by eqn (61). The estimated observed point $\hat{\mathbf{r}}$ or the estimated log likelihood function (62) depends on the current candidate point $\boldsymbol{\theta}(\mathbf{u}') \in M$. We search for a better candidate $\boldsymbol{\theta}(\mathbf{u})$ by maximizing the estimated log likelihood (62), or equivalently by minimizing the KL divergence $K[\hat{\mathbf{r}}(\mathbf{u}') \| \boldsymbol{\theta}(\mathbf{u})]$ from the guessed data point $\hat{\eta} = \hat{\mathbf{r}}(\mathbf{u}')$ to M with respect to $\mathbf{u}$.

Thus, we have an algorithm of estimating a point $\mathbf{u}$ in M and a point $\hat{\mathbf{r}}$ in D iteratively, to search for the better candidates. This is the expectation and maximizing (*EM*) algorithm, where the conditional expectation is used to obtain a candidate $\hat{\mathbf{r}}(\mathbf{u}')$ in D (Figure 8).

The *EM* algorithm consists of the E-step and the M-step as follows.

0. Choose an arbitrary initial guess $\hat{\mathbf{u}}_0$. The initial

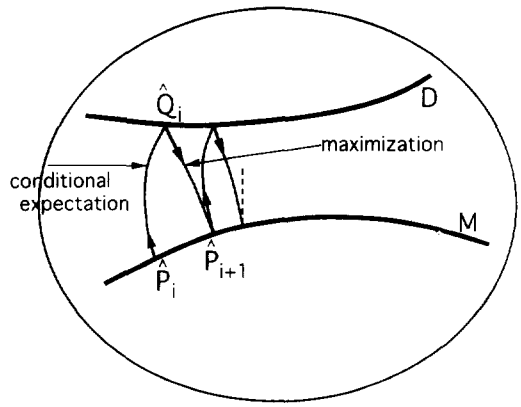

FIGURE 8. *EM* algorithm.

guessed distribution $\hat{P}_0 \in M$ is given by $\boldsymbol{\theta}(\mathbf{u}_0)$ in the $\boldsymbol{\theta}$-coordinates. Repeat the following for $i = 0, 1, 2, \ldots$.

1. *E*-step. Based on the candidate probability distribution $\hat{P}_i \in M$, calculate the conditional expectation of $\mathbf{r}$ under the condition that $\mathbf{s}_v$ is observed,

$$\hat{\boldsymbol{\eta}}_{(i)} = E[\mathbf{r} | \mathbf{s}_v; \hat{P}_i]. \tag{63}$$

This gives the ith candidate of the observed point $\hat{Q}_i \in D$, whose $\boldsymbol{\eta}$-coordinates are $\hat{\boldsymbol{\eta}}_{(i)}$.

2. *M*-step. Calculate the m.l.e. $\hat{\mathbf{u}}_{(i+1)}$ from the estimated observed point $\hat{Q}_i \in D$. This is the point in M that minimizes $K(\hat{Q}_i \| P)$, $P \in M$ or that maximizes the estimated log likelihood. This gives the $(i+1)$st candidate $\hat{P}_{i+1}$ given by $\boldsymbol{\theta}(\hat{\mathbf{u}}_{(i+1)})$.

When the partial observed data cannot be summarized in the single form $\bar{\mathbf{r}} = (1/T)\Sigma \mathbf{r}_t$, we have the data submanifold $D_T^* = D_1 \times \cdots \times D_T$ in the product space S_T^*. The *E*-step is to estimate $\mathbf{r}_T^*$ by

$$\hat{\mathbf{r}}_T^* = E[\mathbf{r}_T^* | \mathbf{s}_{v,1}, \ldots, s_{v,T}; \hat{P}_i].$$

Because $\mathbf{r}_t$ is correlated only to $\mathbf{s}_{v,t}$, $\hat{\mathbf{r}}_t$ is estimated separately from $\mathbf{s}_{v,t}$ in S_t, giving

$$\hat{\boldsymbol{\eta}}_{t,(i)} = E[\mathbf{r}_t | \mathbf{s}_{v,t}; \hat{P}_i]. \tag{64}$$

It is known that the likelihood is increased by one iteration of the *E*- and *M*-steps. Hence, the *EM* algorithm converges to the (local) maximum of the likelihood function of the visible variables where the hidden variables are eliminated. Hence, it converges to the (local) maximum of the likelihood equation. However, this does not necessarily mean that it converges to the maximum likelihood estimator $\hat{\mathbf{u}}_{\text{m.l.e.}}$, beause it may converge to a local maximum of the likelihood. It is also pointed out that the *EM* algorithm has a better global convergence property than the gradient or similar method of maximizing the log likelihood to obtain the m.l.e. directly.

5.2. Geometric *em* Algorithm

Because the true distribution is included in M and the observed data is in the manifold D, it is natural to study the problem from the information geometrical point of view. When a complete data point $\bar{Q}$ is observed, the maximum likelihood estimation searches for the point $\hat{P} \in M$ that is closest to the observed point $\bar{Q}$ in the sense of minimizing the divergence $K(\bar{Q} \| P)$, $P \in M$. When the observation is partial, we cannot identify the observed point $\bar{Q}$. Instead, the observed partial information gives a data submanifold D (or D_T^*). It is natural to search for the pair of points $\hat{P} \in M$, $\hat{Q} \in D$ that minimizes the divergence between D and M, that is,

$$K(\hat{P} \| \hat{Q}) = \min_{P \in M, Q \in D} K(Q \| P). \tag{65}$$

When the observation is complete, D reduces to the observed point $\bar{Q}$ so that $\hat{P}$ is the m.l.e.

Such dual minimization of $K(Q \| P)$ was proposed by Csiszár and Tusnády (1984) in a general perspective. Amari et al. (1992), Byrne (1992), Shimodaira (1993), Neal and Hinton (1994), etc., also studied such problems in various fileds. Information geometry proves that, for a given Q, the point $\hat{P} \in M$ that minimizes $K(Q \| P)$ is given by the m-projection of Q to M (see Appendix). The m-projection is given by the m-geodesic connecting $\hat{P}$ and Q, which is orthogonal to M at $\hat{P}$. In the present exponential family, a curve is an m-geodesic when it is linear in $\boldsymbol{\eta}$-coordinates. The orthogonality is defined in terms of the Fisher Riemannian information metric (see the Appendix).

Dually to the above, for a given P, the point $\hat{Q} \in D$ that minimizes $K(Q \| P)$ is given by the e-projection of P to D (see the Appendix). The e-projection is given by the e-geodesic connecting P and $\hat{Q}$, which is orthogonal to D to $\hat{Q}$. In the present exponential family, a curve is an e-geodesic when it is linear in $\boldsymbol{\theta}$-coordinates.

From the above considerations, we can formulate the geometric *em*-algorithm (e- and m-projection algorithm) as follows (Figure 9).

0. Choose an arbitrary initial guess $\hat{\mathbf{u}}_0$, which gives the initial distribution $\hat{P}_0 \in M$. From $i = 0$, repeat the following.
1. e-step. e-project $\hat{P}_i$ to D. This gives $\hat{Q}_i \in D$ that minimizes $K(Q \| \hat{P}_i)$, $Q \in D$.
2. m-step. m-project $\hat{Q}_i$ to M. This gives $\hat{P}_{i+1}$ that minimizes $K(\hat{Q}_i \| P)$, $P \in M$.

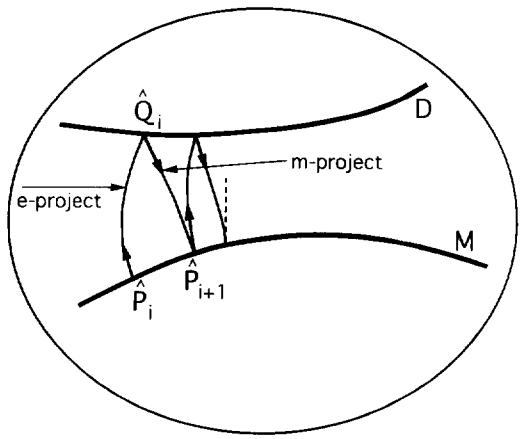

FIGURE 9. *em* algorithm.

When the observed submanifold is D_T^* in S_T^*, the situation is the same. Let

$$Q^* = (Q_1, Q_2, \ldots, Q_T)$$

be a point in D_T^*, where $Q_t \in D_t$, and let

$$P^* = (P_1, P_2, \ldots, P_T)$$

be a point in $M^* \subset S_T^*$. Here, the $\boldsymbol{\theta}$-coordinates $\boldsymbol{\theta}_t$ of P_t is given by

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u}).$$

We have from eqn (52)

$$K(Q^* \| P^*) = \sum_{t=1}^{T} K(Q_t \| P_t). \tag{66}$$

This shows that the point $\hat{Q}^* = (\hat{Q}_1, \ldots, \hat{Q}_T)$ that minimizes $K(Q^* \| P^*)$, $Q^* \in D_T^*$, is composed of the points $\hat{Q}_t$ that minimizes $K(Q_t \| P_t)$, $Q_t \in D_t$. Hence, the e-projection of P^* to D_T^* is given by e-projecting each component P_t to D_t. Therefore, the e-projection can be applied componentwise for each t.

The *EM* and *em* algorithms look quite similar. The next section is devoted to elucidation of their relation.

6. INFORMATION GEOMETRY OF *EM* AND *em* ALGORITHMS

6.1. Properties of the *e*-Projection

The present section elucidates the geometrical properties of the *EM* and *em* algorithms. It is intriguing to see if the two algorithms are equivalent or not. The key point is to show the relation between the e-projection and the conditional expectation. In the beginning, we study geometrical properties of the e- and m-projections. It is a well-known fact in statistics (Amari, 1985) that the m-projection of the observed point $\bar{Q}$ gives the m.l.e. We summarize it in the following theorem.

THEOREM 1. *Let $\hat{Q}$ be the observed point and M be a statistical model in S. The m-projection of $\bar{Q}$ to M gives the m.l.e. $\boldsymbol{\theta}(\hat{\mathbf{u}}) \in M$ that maximizes the likelihood. The m-projection is unique when M is e-flat.*

Properties of the e-projection are shown by the following theorem.

THEOREM 2. *When D is an m-flat submanifold, the e-projection of P to D is unique. Let D be represented by the separated form in the* $\boldsymbol{\eta}$-coordinates,

$$D = \{\boldsymbol{\eta} | \boldsymbol{\eta} = (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h), \quad \boldsymbol{\eta}_v = \check{\mathbf{r}}_v, \ \boldsymbol{\eta}_h \text{ is arbitrary}\}.$$

Let P be a point whose $\boldsymbol{\theta}$-coordinates are similarily partitioned as $\boldsymbol{\theta}_p = (\boldsymbol{\theta}_v^P, \boldsymbol{\theta}_h^P)$ and Q^* be the e-projection of P to D whose the $\boldsymbol{\eta}$- and $\boldsymbol{\theta}$-coordinates are denoted by $(\boldsymbol{\eta}_v^*, \boldsymbol{\eta}_h^*)$ and $(\boldsymbol{\theta}_v^*, \boldsymbol{\theta}_h^*)$, respectively. Then, the following properties hold:

1. *The visible part $\boldsymbol{\eta}_v^*$ of Q^* is given by $\boldsymbol{\eta}_v^* = \check{\mathbf{r}}_v$.*
2. *The hidden part $\boldsymbol{\theta}_h^*$ of the $\boldsymbol{\theta}$-coordinates of Q^* is kept invariant under the e-projection,*

$$\boldsymbol{\theta}_h^P = \boldsymbol{\theta}_h^*. \tag{67}$$

3. *The conditional probability of the hidden variable $\mathbf{r}_h$ at Q^* is equal to that at P,*

$$p(\mathbf{r}_h | \check{\mathbf{r}}_v, P) = p(\mathbf{r}_h | \check{\mathbf{r}}_v, Q^*). \tag{68}$$

4. *The conditional expectation of the hidden variables at P is equal to that at Q^*,*

$$E_P[\mathbf{r}_h | \check{\mathbf{r}}_v] = E_{Q^*}[r_h | \check{r}_v]. \tag{69}$$

Proof. The property 1. is trivial, because $Q^* \in D$. We first prove that the hidden part of the $\boldsymbol{\theta}$-coordinates $\boldsymbol{\theta}_h^*$ of Q^* is equal to the corresponding part $\boldsymbol{\theta}_h^P$ of P. The divergence $K(Q \| P)$ is rewritten as

$$\begin{aligned} K(Q \| P) &= E_Q\left[\log \frac{p(\mathbf{r}; \boldsymbol{\theta}_Q)}{p(\mathbf{r}; \boldsymbol{\theta}_P)}\right] \\ &= E_Q[(\boldsymbol{\theta}_v^Q - \boldsymbol{\theta}_v^P) \cdot \mathbf{r}_v + (\boldsymbol{\theta}_h^Q - \boldsymbol{\theta}_h^P) \cdot \mathbf{r}_h] \\ &\quad - \psi(\boldsymbol{\theta}_Q) + \psi(\boldsymbol{\theta}_P) \\ &= (\boldsymbol{\theta}_v^Q - \boldsymbol{\theta}_v^P) \cdot \check{\mathbf{r}}_v + (\boldsymbol{\theta}_h^Q - \boldsymbol{\theta}_h^P) \cdot \boldsymbol{\eta}_h^Q \\ &\quad - \psi(\boldsymbol{\theta}_Q) + \psi(\boldsymbol{\theta}_P), \end{aligned} \tag{70}$$

because of $E_Q[\mathbf{r}_v] = \check{\mathbf{r}}_v$ for $Q \in D$. In the submanifold D, $\boldsymbol{\eta}_h$ is the free variable whereas $\boldsymbol{\eta}_v$ is fixed at the

value $\check{\mathbf{r}}_r$. Because the e-projection minimizes $K(Q\|P)$, we have by differentiation

$$\frac{\partial K(Q\|P)}{\partial \boldsymbol{\eta}_h} = \frac{\partial \boldsymbol{\theta}_v}{\partial \boldsymbol{\eta}_h}\cdot \check{\mathbf{r}}_v + \frac{\partial \boldsymbol{\theta}_h}{\partial \boldsymbol{\eta}_h}\cdot \boldsymbol{\eta}_h + (\boldsymbol{\theta}_h - \boldsymbol{\theta}_h^P) - \frac{\partial \psi(\boldsymbol{\theta}_Q)}{\partial \boldsymbol{\theta}_v}\frac{\partial \boldsymbol{\theta}_v}{\partial \boldsymbol{\eta}_h} - \frac{\partial \psi(\boldsymbol{\theta}_Q)}{\partial \boldsymbol{\theta}_h}$$

because of

$$\frac{\partial \psi(\boldsymbol{\theta}_Q)}{\partial \boldsymbol{\theta}_v} = \boldsymbol{\eta}_v = \check{\mathbf{r}}_v, \quad \frac{\partial \psi(\boldsymbol{\theta}_Q)}{\partial \boldsymbol{\theta}_h} = \boldsymbol{\eta}_h. \tag{71}$$

The right-hand side of eqn (70) vanishes at the minimum point Q^*, so that eqn (67) holds.

The conditional distribution of $\mathbf{r}_h$ at P conditioned on $\mathbf{r}_v = \check{\mathbf{r}}_v$ is written as

$$\begin{aligned} p(\mathbf{r}_h|\check{\mathbf{r}}_v, \boldsymbol{\theta}_P) &= \frac{p(\mathbf{r}_h, \check{\mathbf{r}}_v; \boldsymbol{\theta}_P)}{p(\check{\mathbf{r}}_v; \boldsymbol{\theta}_P)} \\ &= \frac{\exp\{\boldsymbol{\theta}_v^P\cdot\check{\mathbf{r}}_v + \boldsymbol{\theta}_h^P\cdot\mathbf{r}_h - \psi\}}{\int \exp\{\boldsymbol{\theta}_v^P\cdot\check{\mathbf{r}}_v + \boldsymbol{\theta}_h^P\cdot\mathbf{r}_h - \psi\}d\mu(\mathbf{r}_h)} \\ &= \exp\{\boldsymbol{\theta}_h^P\cdot\mathbf{r}_h - \tilde{\psi}\}, \end{aligned}$$

where the normalization factor $\tilde{\psi}$ depends on $\boldsymbol{\theta}_h^P$ and $\check{r}_v$ but not on $\boldsymbol{\theta}_v^P$. Hence, the conditional expectation of $\mathbf{r}_h$ does not depend on $\boldsymbol{\theta}_v^P$. Therefore, because P and its e-projection Q^* have the same $\boldsymbol{\theta}_h$-coordinates, the conditional probabilities are equal at P and at Q^*, and hence

$$E_P[\mathbf{r}_h|\check{\mathbf{r}}_v] = E_{Q^*}[\mathbf{r}_h|\check{\mathbf{r}}_v]. \quad \blacksquare \tag{72}$$

6.2. Geometry of the *EM* Algorithm

Theorem 1 shows that the M-step and m-step are the same procedure giving the m.l.e. $\hat{\mathbf{u}}$. How are the E- and e-steps? To obtain a geometrical interpretation of the E-step, we define the following transformation $F: S \to S$. Let the $\boldsymbol{\eta}$-coordinates of Q be divided as $(\boldsymbol{\eta}_v, \boldsymbol{\eta}_h)$. We put

$$\mathbf{s}_Q(\mathbf{r}_v) = E_Q[\mathbf{r}_h|\mathbf{r}_v], \tag{73}$$

where E_Q is the conditional expectation at Q. This is a function of $\mathbf{r}_v$ and $Q = (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h)$. Let F be a mapping that maps a point $Q: (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h)$ to $Q: (\boldsymbol{\eta}_v, \mathbf{s}_Q(\boldsymbol{\eta}_v))$,

$$F: (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h) \mapsto (\boldsymbol{\eta}_v, \mathbf{s}_Q(\boldsymbol{\eta}_v)). \tag{74}$$

That is, F keeps $\boldsymbol{\eta}_v$ invariant and replaces the $\boldsymbol{\eta}_h$ (which is the unconditional expectation of $\mathbf{r}_h$ at Q) by the conditional expectation of $\mathbf{r}_h$ conditioned on $\mathbf{r}_v = \boldsymbol{\eta}_v$ (which is the expectation of $\mathbf{r}_v$ at Q). When D is defined by fixing $\boldsymbol{\eta}_v = \check{\mathbf{r}}_v$, F maps D to itself.

The E-step is interpreted as follows. Let P be a candidate point in M. The E-step gives the point $\hat{Q} \in D$ whose $\boldsymbol{\eta}_v$ part is equal to the observed $\check{\mathbf{r}}_v$ and $\boldsymbol{\eta}_h$ part is given by the conditional expectation $E_P[\mathbf{r}_h|\mathbf{r}_v = \check{\mathbf{r}}_v]$ of the hidden $\mathbf{r}_h$ at P. The conditional expectation at P is the same as that at the e-projected Q^* of P, because the conditional expectation is kept invariant under the e-projection (Theorem 2). Hence, the E-step implies to e-project P to D obtaining Q^* and then transform Q^* to $FQ*$ by replacing the $\boldsymbol{\eta}_h^*$ of Q^* by the conditional expectation of $\mathbf{r}_h$ at Q^*. So the E-step is rewritten as E-step: e-project $\hat{P}_i$ to D to obtain Q_i^*, and then transform it by F to give $\hat{Q}_i = FQ_i^*$. This also gives the following alternative description of the *EM* algorithm.

THEOREM 3. *The EM algorithm is formulated in the following dual minimization steps.*

1. *E-step. Search for the point* $\hat{Q}_i \in D$ *that minimizes* $K\{F^{-1}(Q)\|P_i\}$, $Q \in D$.
2. M-*step. Search for the point* $\hat{P}_{i+1} \in M$ *that minimizes* $K(\hat{Q}_i\|P)$, $P \in M$.

It is clear that, when F is the identity, the EM and *em* algorithms are the same. This holds when the conditional expectation of $\mathbf{r}_h$ conditioned on $\check{\mathbf{r}}_v$ at $Q \in D$ is equal to the unconditional expectation of $\mathbf{r}_h$ at $Q \in D$.

THEOREM 4. *The EM and em algorithms are equivalent iff, the conditional expectation* $\mathbf{s}_Q(\mathbf{r}_v) = E_Q[\mathbf{r}_h|\mathbf{r}_v]$ *is linear in* $\mathbf{r}_v$ *at any* $Q \in S$.

Proof. At $Q = (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h) \in D$, we have

$$\begin{aligned} \boldsymbol{\eta}_h &= E_Q[\mathbf{r}_h] = E_Q E_Q[\mathbf{r}_h|\mathbf{r}_v] = E_Q \mathbf{s}_Q(\mathbf{r}_v), \\ \boldsymbol{\eta}_v &= E_Q[\mathbf{r}_v]. \end{aligned}$$

When $s_Q(\mathbf{r}_v)$ is a linear function, it is written as

$$\mathbf{s}_Q(\mathbf{r}_v) = \mathbf{a} + B\mathbf{r}_v \tag{75}$$

for a constant $\mathbf{a}$ and a constant matrix B. We then have

$$\boldsymbol{\eta}_h = E_Q[\mathbf{s}_Q(\mathbf{r}_v)] = \mathbf{s}_Q(E_Q(\mathbf{r}_v)) = \mathbf{s}_Q(\boldsymbol{\eta}_v).$$

Hence, F is the identity and the two algorithms are equivalent. On the contrary, we assume that F is the identity, so that

$$\boldsymbol{\eta}_h = E_Q[\mathbf{s}_Q(\mathbf{r}_v)] = \mathbf{s}_Q[E_Q(\mathbf{r}_v)] \tag{76}$$

holds for any point $Q \in S$. Let $Q = \{q(\mathbf{r}_v, \mathbf{r}_h)\}$ be an arbitrary distribution belonging to D_0 defined by

$\eta_v = \eta_{v,0}$. Let $P = \{p(\mathbf{r}_v, \mathbf{r}_h)\}$ be an arbitrary distribution such that its e-projection to D_0 is equal to Q. Let Q_t, $0 \le t \le 1$, be the family of distributions on the e-geodesic connecting P and Q, where $Q_0 = Q$ and $Q_1 = P$. The family Q_t is an exponential family in S, and all the Q_t have the same conditional distribution $q(\mathbf{r}_h|\mathbf{r}_v)$ because the e-projection of Q_t is Q. Hence, the distribution $Q_t = \{q_t(\mathbf{r}_v, \mathbf{r}_h)\}$ is decomposed as

$$q_t(\mathbf{r}_v, \mathbf{r}_h) = p(r_v; t)q(\mathbf{r}_h|\mathbf{r}_v). \tag{77}$$

Because $\mathbf{s}_Q(\mathbf{r}_v)$ depends on Q only through its conditional distribution, we have

$$\mathbf{s}_{Q_t}(\mathbf{r}_v) = \mathbf{s}_Q(\mathbf{r}_v).$$

Hence, eqn (76) implies that

$$\int \mathbf{s}_Q(\mathbf{r}_v)p(\mathbf{r}_v; t)d\mathbf{r}_v = \mathbf{s}_Q\left[\int \mathbf{r}_v P(\mathbf{r}_v, t)d\mathbf{r}_v\right]. \tag{78}$$

We can choose any P under the condition that

$$\int p(\mathbf{r}_v; t)d\mathbf{r}_v = 1.$$

Therefore, eqn (78) immediately shows that $\mathbf{s}_Q(\mathbf{r}_v)$ is a linear function of $\mathbf{r}_v$ as in eqn (75). ■

6.3. An Example Where the *EM* and *em* Algorithms are Different

As will be shown in the next section, the two algorithms are equivalent in most important cases. Here, we show a simple example in which the two algorithms are different.

Example 10. The normal multiplication model (Example 8 continued). We first treat the case of $T = 2$ for simplicity's sake and search for the mapping F. Let $Q = N(\bar{r}, \sigma^2)$ be a point in D. When

$$r_v = \frac{1}{2}(x_1 + x_2) = \bar{r}$$

is observed, the conditional probability distribution $p_Q(x_t|\bar{r})$, $t = 1, 2$, is given as follows. The joint conditional distribution conditioned on $(x_1 + x_2)/2 = \bar{r}$ is

$$p_Q(x_1, x_2|\bar{r}) = c(\bar{r}) \exp\left\{-\frac{(x_1 - \bar{r})^2 + (x_2 - \bar{r})^2}{2\sigma^2}\right\} \times \delta\left(\frac{x_1 + x_2}{2} - \bar{r}\right),$$

where $c(\bar{r})$ is the normalization constant. By integrating this with respect to x_2, we have

$$p_Q(x_1|\bar{r}) = c \exp\left\{-\frac{(x_1 - \bar{r})^2}{\sigma^2}\right\}.$$

Thus, when we know $\bar{r}$, the variance of x_1 reduces to $\sigma^2/2$. Obviously, x_2 has the same conditional distribution, although x_1 and x_2 are not conditionally independent. Hence, the conditional expectation of x_t^2 is

$$E_Q[x_t^2|\bar{r}] = \bar{r}^2 + \frac{\sigma^2}{2}.$$

So the conditional expectation of r_h conditioned on $r_v = \bar{r}$ is given by

$$s_Q(\bar{r}) = E_Q\left[\frac{x_1^2 + x_2^2}{2}\,\middle|\,\bar{r}\right] = \bar{r}^2 + \frac{\sigma^2}{2}.$$

This is different from the unconditional expectation at Q

$$\eta_h = E_Q\left[\frac{x_1^2 + x_2^2}{2}\right] = \bar{r}^2 + \sigma^2.$$

This shows that F is not the identity and it maps $Q = N(\bar{r}, \sigma^2)$ to

$$FQ = N\left(\bar{r}, \bar{r}^2 + \frac{\sigma^2}{2}\right).$$

The e-projection Q^* of a point $P = N(u, u^2) \in M$ to D is given as follows. Because the $\boldsymbol{\theta}$-coordinates of a point $N(\mu, \sigma^2)$ is given by

$$\theta_v = \frac{2}{u}, \quad \theta_h = \frac{1}{u^2}$$

when $T = 2$, and because $\theta_h^* = \theta_h^P = -u^{-2}$ and $\eta_v^* = \bar{r}$, we have

$$Q^* = N(\bar{r}, u^2).$$

On the other hand, the E-step from the candidate $P \in M$ gives

$$\hat{Q} = N\left(\bar{r}, \frac{u^2}{2}\right),$$

because

$$E_p[r_h|r_v = \bar{r}] = \bar{r}^2 + \frac{u^2}{2}.$$

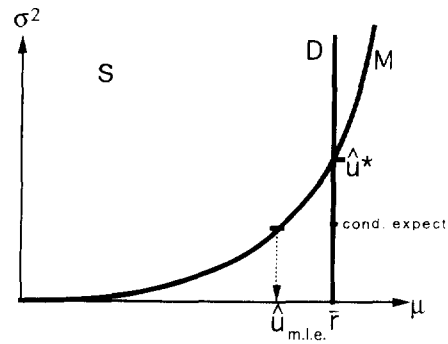

FIGURE 10. Example where *EM* and *em* algorithms are different.

This is confirmed from

$$FQ^* = \hat{Q}.$$

In the present problem, M and D intersect at a point $\hat{P}^* = N(\bar{r}, \bar{r}^2)$ at which $K(D \| \hat{P}^*) = 0$ (Figure 10). Hence the *em* algorithm converges to

$$\hat{P}^* = N(\bar{r}, \bar{r}^2),$$

that is, the best candidate

$$\hat{u}^* = \bar{r}$$

from the geometrical point of view. On the other hand, the *EM* algorithm converges to the m.l.e. $\hat{u}$,

$$\hat{u} = (\sqrt{3} - 1)\bar{r} \doteq 0.73\bar{r}.$$

Here, $K(D \| \hat{P})$ is not the minimum of $K(D \| M)$.

For general $T > 2$, we have, for $Q^* = N(\bar{r}, u^2)$,

$$FQ^* = N\left(\bar{r}, \frac{T-1}{T} u^2\right).$$

The solution of the *em* algorithm is

$$\hat{u}^* = \bar{r},$$

whereas the *EM* algorithm gives

$$\hat{u} = \frac{1}{2}\left(\sqrt{T^2 + 4T} - T\right)\bar{r} \approx \left(1 - \frac{1}{T}\right)\bar{r}.$$

Therefore, when T is large, $\hat{u}^* \doteq \hat{u}$ and the *em* algorithm and the *EM* algorithm are asymptotically equivalent. The asymptotic equivalence holds in general, so that we do not need to be bothered by their difference practically.

7. THE EQUIVALENCE OF THE *EM* AND *em* ALGORITHMS IN EXTENDED FRAMEWORK

7.1. Which is More Natural, the *EM* or *em*?

When F is not the identity, the *EM* and *em* algorithms are different, giving different estimates $\hat{\mathbf{u}}$. Which is more natural and which is better? Because the *EM* algorithm gives the m.l.e., and because it is known that the bias-corrected m.l.e. is higher-order efficient (Amari, 1985), the *EM* algorithm might look more natural from the statistical point of view. However, when we use a neural network model to approximate a given input–output relation, it is not statistical inference. The problem is to realize a neural network that approximates a given input–output behavior as faithfully as possible. The behavior is presented by examples and is summarized in the data submanifold D. Any point in D can equally explain the input–output data and their difference lies only in the hidden variables, which we do not care about. Therefore, the *em* algorithm that minimizes the divergence $K(D \| M)$ seems to give a more natural answer.

However, we show that the two algorithms are equivalent if we extend our framework from the exponential family to the function space. Even in the framework of the exponential family, they are proved to be asymptotically equivalent when T is large. When they are equivalent, the E-step of taking the conditional expectation may be computationally more tractable than the e-projection.

7.2. Linearization Trick Guaranteeing the Equivalence of the *EM* and *em* Algorithms

The two algorithms are equivalent when $s_Q(\mathbf{r}_h)$ is linear. We show that it is linear in the extended framework. Let y be a random variable taking on the values 0 and 1. Then, any function $f(y)$ is linear in y, because we have

$$f(y) = \{f(1) - f(0)\}y + f(0). \tag{79}$$

This trick can be used for any function $f(\mathbf{r})$, provided the variable $\mathbf{r}$ takes its value on a finite set K. Let $\mathbf{m}$ be an element of K, and we introduce a new vector variable $\mathbf{k}(\mathbf{r})$ indexed by $\mathbf{m}$,

$$\mathbf{k}(\mathbf{r}) = \{k_{\mathbf{m}}(\mathbf{r}), \mathbf{m} \in K\} \tag{80}$$

by

$$k_{\mathbf{m}}(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{m}). \tag{81}$$

Then, any function $f(\mathbf{r})$ is linear in the extended

random variable $\mathbf{k}(\mathbf{r})$, because

$$f(\mathbf{r}) = \sum_{\mathbf{m}} f(\mathbf{m}) k_{\mathbf{m}}(\mathbf{r}) \tag{82}$$

This shows that, when the visible random variable is binary, or when the visible random variable is represented in the extended vector form $\mathbf{k}(\mathbf{r}) = \{k_{\mathbf{m}}(\mathbf{r})\}$, the conditional expectation is linear in the visible variable $\mathbf{k}(\mathbf{r})$ and the *EM* and *em* algorithms are equivalent. Indeed, the *EM* and *em* algorithms are equivalent in the binary stochastic perceptron, the mixture of expert nets, the Boltzmann machine, and the normal mixture model.

The linearization trick still works in the continuous variable case because we have

$$f(\mathbf{r}) = \int f(\mathbf{m}) \delta(\mathbf{r} - \mathbf{m}) d\mathbf{m}. \tag{83}$$

In this case, the random variable $\mathbf{r}$ is extended to the random (generalized) function

$$\delta_{\mathbf{m}}(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{m}),$$

which is an element of the generalized function space. When T examples $\mathbf{r}_1, \ldots, \mathbf{r}_T$ are observed, they are summarized in the empirical distribution,

$$\hat{p}_{\text{emp}}(\mathbf{r}) = \frac{1}{T} \sum_{t=1}^{T} \delta(\mathbf{r} - \mathbf{r}_t). \tag{84}$$

Now we formulate the hidden variable problem in the function space, and prove that the *EM* and *em* algorithms are equivalent in the function space. Such a formulation is obtained by Cssiszár and Tusnády (1984), Shimodaira (1993), and Neal and Hinton (1993). Let $S = \{p(\mathbf{r}_v, \mathbf{r}_h)\}$ be the function space of all the density functions of random variables $\mathbf{r} = (\mathbf{r}_v, \mathbf{r}_h)$, where $\mathbf{r}_v$ is visible but $\mathbf{r}_h$ is hidden. Let

$$M = \{p(\mathbf{r}_v, \mathbf{r}_h; \mathbf{u})\} \tag{85}$$

be parametric model specified by parameter $\mathbf{u}$. When $\mathbf{r}^*_{v,T} = (\mathbf{r}_{v,1}, \ldots, \mathbf{r}_{v,T})$ is observed, we can summarize it in the function space as

$$\hat{p}_{\text{emp}}(\mathbf{r}_v) = \frac{1}{T} \sum_{t=1}^{T} \delta(\mathbf{r}_v - \mathbf{r}_{v,t}). \tag{86}$$

This is the empirical distribution based on the observed data. Let us decompose $p(\mathbf{r}_v, \mathbf{r}_h)$ as

$$p(\mathbf{r}_v, \mathbf{r}_h) = p(\mathbf{r}_v) p(\mathbf{r}_h | \mathbf{r}_v). \tag{87}$$

When $\mathbf{r}_v$'s are observed, $p(\mathbf{r}_v)$ is given by the empirical distribution but $p(\mathbf{r}_h | \mathbf{r}_v)$ remains free. Hence, the observed data submanifold D consists of the functions

$$D = \{\hat{p}_{\text{emp}}(\mathbf{r}_v) p(\mathbf{r}_h | \mathbf{r}_v)\} \tag{88}$$

in the function space, where $\hat{p}_{\text{emp}}$ is observed but $p(\mathbf{s}_h | \mathbf{r}_v)$ are free functions. The e-projection of $P = p(\mathbf{r}_v, \mathbf{r}_h; \mathbf{u})$ to D is the one that minimizes $K[q(r_v, \mathbf{r}_h) \| p(\mathbf{r}_v, \mathbf{r}_h; \mathbf{u})]$, $q(\mathbf{r}_v, \mathbf{r}_h) \in D$. The structure is transparent if we show that the e-projection Q^* is given by

$$q^*(\mathbf{r}_v, \mathbf{r}_h) = \hat{p}_{\text{emp}}(\mathbf{r}_v) p(\mathbf{r}_h | \mathbf{r}_v; \mathbf{u}) \tag{89}$$

when $T = 1$, and by

$$\prod_{t=1}^{T} q^*(\mathbf{r}_{v,t}, \mathbf{r}_{h,t}) = \prod_{t=1}^{T} \hat{p}_{\text{emp}}(\mathbf{r}_{v,t}) p(\mathbf{r}_{h,t}; \mathbf{u}),$$

for general T. On the other hand, by extending the hidden random variable to the function $\delta(\mathbf{r}_h - \mathbf{m})$, its conditional expectation is

$$\begin{aligned} E[\delta(\mathbf{r}_h - \mathbf{m}) | \mathbf{r}_v; \mathbf{u}] &= \int \delta(\mathbf{r}_h - \mathbf{m}) p(\mathbf{m} | \mathbf{r}_v; \mathbf{u}) d\mathbf{m} \\ &= p(\mathbf{r}_h | \mathbf{r}_v; \mathbf{u}), \end{aligned} \tag{90}$$

giving the conditional probability at $\mathbf{u}$. Hence, the two algorithms are the same, as is shown by Csiszár and Tusnády (1985) and Neal and Hinton (1993).

THEOREM 5. *The EM and em algorithms are equivalent in the function space. The E or e-step is*

E (e)-step. Obtain the conditional probability $p(\mathbf{r}_h | \mathbf{r}_v; \hat{\mathbf{u}}_{(i)})$.
M m-step. Maximize

$$E_{Q_i}[\log p(\mathbf{r}_h | \mathbf{r}_v; \mathbf{u})] = \sum_{t=1}^{T} \int p(\mathbf{r}_{h,t} | \mathbf{r}_{v,t}; \hat{\mathbf{u}}_{(i)}) \log p(\mathbf{r}_{h,t} | \mathbf{r}_{v,t}; \mathbf{u}) d\mathbf{r}_{h,t}, \tag{91}$$

with respect to $\mathbf{u}$, *which gives* $\hat{\mathbf{u}}_{(i+1)}$.

The structure is transparent if we consider in the function space. However, we need to keep all the data $\mathbf{r}_{v,1}, \ldots, \mathbf{r}_{v,T}$ without summarizing them into a sufficient statistics.

7.3. Analog Stochastic Perceptron — an Example

Not all the neural models are exponential or curved exponential families. We show this by using an

analog stochastic perceptron (see Amari, 1991). Let $\mathbf{x}$ be an input, $\mathbf{z}$ be the output of hidden units, and y be the final output. Let f be an analog sigmoidal function, for example,

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

The output of the ith hidden unit is

$$z_i = f(\mathbf{w}_i \cdot \mathbf{x}) + n_i \tag{92}$$

and the final output is

$$y = f(\mathbf{v} \cdot \mathbf{z}) + n, \tag{93}$$

where n_i and n are independent normal random noises subject to $N(0, \sigma^2)$. We then have

$$p(\mathbf{z}|\mathbf{x}; \mathbf{u}) = c \exp\left[-\frac{1}{2\sigma^2}\sum \{z_i - f(\mathbf{w}_i \cdot \mathbf{x})\}^2\right] \tag{94}$$

$$p(y|\mathbf{z}; \mathbf{u}) = c' \exp\left[-\frac{1}{2\sigma^2}\{y - f(\mathbf{v} \cdot \mathbf{z})\}^2\right]. \tag{95}$$

Hence, the logarithm of the joint conditional probability distribution is written as

$$\sigma^2 l(y, \mathbf{z}|\mathbf{x}; \mathbf{u}) = \sum z_i f(\mathbf{w}_i \cdot \mathbf{x}) + y f(\mathbf{v} \cdot \mathbf{z}) - \tfrac{1}{2}\{f(\mathbf{v} \cdot \mathbf{z})\}^2 + k(y, \mathbf{z}) - \psi. \tag{96}$$

This cannot be summarized in the form of $\boldsymbol{\theta} \cdot \mathbf{r}$, a bilinear form in an extended random variable $\mathbf{r}$ and a function $\boldsymbol{\theta}(\mathbf{u}, \mathbf{x})$ of parameters. Hence, the distributions $p(y, \mathbf{z}|\mathbf{x}; \mathbf{u})$ do not belong to a curved exponential family. However, it is possible to generalize the information geometry to be responsible for such cases by introducing the manifold of functions. The information geometry and the *EM* algorithm are also applicable to this case (Csiszár and Tusnády, 1984).

It should also be remarked that a stochastic perceptron reduces to the ordinary analog perceptron if the stochastic outputs z_i and y are replaced by their expected values. Hence, an ordinary multilayer perceptron can be trained by the stochastic method in the training phase.

The *EM* algorithm works as follows.

E-step. Calculate the conditional distribution of z_t based on observation $(y_t, \mathbf{x}_t)$ at time t. This is given by

$$p(\mathbf{z}_t|y_t, \mathbf{x}_t, \hat{\mathbf{u}}_i) = \frac{\exp\left\{-\frac{1}{2\sigma^2}\left[|\mathbf{z}_t - \mathbf{f}_t|^2 + \{y_t - f(\hat{\mathbf{v}}_i \cdot \mathbf{z}_t)\}^2\right]\right\}}{\int \exp\left\{-\frac{1}{2\sigma^2}\left[|\mathbf{z}_t - \mathbf{f}_t|^2 + \{y_t - f(\hat{\mathbf{v}}_i \cdot \mathbf{z}_t)\}^2\right]\right\} d\mathbf{z}_t}$$

where $\mathbf{f}_t = (f_{1,t}, \ldots, f_{k,t})$,

$$f_{j,t} = f(\hat{\mathbf{w}}_j \cdot \mathbf{x}_t).$$

M-step. Calculate the $\mathbf{u}$ that maximizes

$$\sum_t \int p(\mathbf{z}_t|y_t, x_t, \hat{\mathbf{u}}_i) \log p(\mathbf{z}_t|y_t, \mathbf{x}_t, \mathbf{u}) d\mathbf{z}_t$$

and put it as $\hat{\mathbf{u}}_{i+1}$. This process converges to the maximum likelihood estimate. The learning version of this procedure is also easily given. However, it is not clear how good is the m.l.e. We compare the behavior of $\hat{\mathbf{u}}_{\text{m.l.e.}}$ with the back-prop solution in the special case where σ^2 is very small.

In the case of the analog multilayer perceptron, the conventional back-propagation learning rule is designed to minimize the empirical error

$$l(\mathbf{u}) = \sum_{t=1}^{T} \frac{1}{2}\{y_t - f(\mathbf{x}_t; \mathbf{u})\}^2, \tag{97}$$

where y_t is the specified output and $f(\mathbf{x}_t; \mathbf{u})$ is the output from the network with parameter $\mathbf{u}$. In the present analog stochastic perceptron, $f(\mathbf{x}_t; \mathbf{u})$ is given in the execution mode by the expected value of the output, so that the output

$$f(\mathbf{x}_t; \mathbf{u}) = f\left\{\sum v_i f(\mathbf{w}_i \cdot \mathbf{x}_t)\right\}$$

is the same as the deterministic one. On the other hand, the paramter $\mathbf{u}$ is modified to maximize the likelihood function of examples in the learning mode. It is in general difficult to write down the likelihood function explicitly and compare it with the square loss (97). We calculate here the likelihood function when the noise σ is very small, to show the difference between the conventional least square loss and the statistical loss given from the stochastic perceptron.

Because the joint probability of y and $\mathbf{z}$ is given from eqns (94) and (95) as

$$p(y, \mathbf{z}|\mathbf{x}) = c \exp\left\{-\frac{1}{2\sigma^2}\left[\sum \{z_i - f(\mathbf{w}_i \cdot \mathbf{x})\}^2 + \{y - f(\mathbf{v} \cdot \mathbf{z})\}^2\right]\right\},$$

we have the conditional probability of y by

integrating the above with respect to $\mathbf{z}$. By using $\mathbf{n} = (n_i)$ in eqn (94) and putting $\mathbf{n}' = \sigma^{-1}\mathbf{n}$, we have

$$p(y|\mathbf{x}) = c\int \exp\left\{-\frac{|\mathbf{n}'|^2}{2} - \frac{1}{2}[y - f(\mathbf{v}\cdot\{f(\mathbf{w}_i\cdot\mathbf{x}) + \sigma\mathbf{n}'\})]^2\right\} d\mathbf{n}'$$

where

$$\mathbf{v}\cdot f(\mathbf{w}_i\cdot\mathbf{x}) = \sum v_i f((\mathbf{w}_i\cdot\mathbf{x}).$$

When σ is small, we expand

$$f[\mathbf{v}\cdot\{f(\mathbf{w}_i\cdot\mathbf{x}) + \sigma\mathbf{n}'\}] = f + \sigma f'\mathbf{v}\cdot\mathbf{n}' + \frac{1}{2}\sigma^2 f''(\mathbf{v}\cdot\mathbf{n}')^2,$$

where

$$f = f\left\{\sum v_i f(\mathbf{w}_i\cdot\mathbf{x})\right\} = f(\mathbf{x};\mathbf{u})$$

is the output of the network without noise and the prime is differentiation. We finally obtain

$$-\log p(y|\mathbf{x};\mathbf{u}) = \frac{1}{2}(y-f)^2\left\{1 + \frac{1}{2}\sigma^2 f'^2|\mathbf{v}|^2\right\}. \quad (98)$$

Therefore, the stochastic perceptron is expected to minimize the loss

$$l_{\text{stoch}}(u) = \sum_{t=1}^{T}\frac{1}{2}\{y_t - f(x_t;u)\}^2\left\{1 + \frac{1}{2}\sigma^2 f'^2|\mathbf{v}|^2\right\}. \quad (99)$$

This shows that the factor

$$1 + \frac{1}{2}\sigma^2 f'^2\mathbf{v}^2$$

is multiplied to the ordinary squared error. This is very reasonable because it automatically has the effect of decreasing loss at the range where $f(\mathbf{x};\mathbf{x})$ is almost saturated. Learning of the stochastic perceptron is shown to give a good performance by a preliminary computer simulation. This is also related to the recent finding of the enhanced performance of the multilayer performance with noise (Murray & Edwards, 1994).

8. LEARNING PROCEDURES

We can rewrite the *EM* or *em* algorithm in the on-line learning form when a partial observation $\mathbf{r}_{v,t}$ or $\mathbf{s}_{v,t}$ is available one at each time t, while the batch algorithm uses all the data $\mathbf{r}_{v,t}$, $t = 1,\ldots,T$ stored together. In general, the convergence might be slower but the algorithm is much simpler in learning. Moreover, learning is robust against any changes in the environmental structures.

We propose the following learning algorithm when all the data are summarized in a single observed point $\hat{\boldsymbol{\eta}}$ in S. The algorithm is essentially the same as that of Neal and Hinton (1993) and Jordan and Jacob (1993). The algorithm is applicable to the stochastic perceptron. Let $\hat{\mathbf{u}}_t$ be the estimator at time t, and let $\hat{\boldsymbol{\eta}}_t$ be the guess of the observed point at time t. Given a partial observation $\mathbf{s}_{v,t+1}$ at time $t+1$, the learning procedure is as follows.

1. e-step (E-step). Calculate the e-projection of the present $\hat{P}_t$ to D_{t+1} (the conditional expectation of $\hat{\mathbf{r}}_{t+1}$ conditioned on $s_{v,t+1}$ based on $\hat{\mathbf{u}}_t$). This gives a guess of the unobserved $\hat{\mathbf{r}}_{t+1}$. Then, modify the guess of the observed point by

$$\hat{\boldsymbol{\eta}}_{t+1} = (1-\varepsilon_t)\hat{\boldsymbol{\eta}}_t + \varepsilon_t\hat{\mathbf{r}}_{t+1}, \quad (100)$$

   where ε_t is a constant or a decreasing sequence such as $\varepsilon_t \sim c/t$.
2. M-step. Calculate the m.l.e. from $\hat{\boldsymbol{\eta}}_{t+1}$ and put it as $\hat{\mathbf{u}}_{t+1}$ which gives $\hat{P}_{t+1}$.

Because calculation of the m.l.e. is usually not easy, we can use the gradient method to obtain the next $\hat{\mathbf{u}}_{t+1}$. The log likelihood for $\hat{\boldsymbol{\eta}}_{t+1}$ is

$$l = \boldsymbol{\theta}(\mathbf{u})\cdot\hat{\boldsymbol{\eta}}_{t+1} - \psi(\mathbf{u}),$$

and its gradient is

$$\frac{\partial l}{\partial \mathbf{u}} = B\{(\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\mathbf{u}))\}, \quad (101)$$

where

$$B = \frac{\partial\boldsymbol{\theta}}{\partial\mathbf{u}}$$

is a matrix. Hence, the gradient method applied to the incremental M-step is

2′. incremental M-step

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \varepsilon_t\mathbf{B}\{\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\hat{\mathbf{u}}_t)\}.$$

Any acceleration method is applicable to the above. The scoring method gives

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \varepsilon_t G_t^*\mathbf{B}\{\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\hat{\mathbf{u}}_t)\}$$

where G_t^* is the inverse of the Fisher information matrix of M at $\hat{\mathbf{u}}_t$.

When the observed data cannot be summarized into a single $\hat{\boldsymbol{\eta}}$ so that we need to consider the

product space S_T^*, we propose the following learning algorithm.

1. *e*-step (*E*-step). Calculate the *e*-projection of P_t whose $\boldsymbol{\theta}$-coordinates are $\boldsymbol{\theta}(\mathbf{x}_{t+1}, \hat{\mathbf{u}}_t)$ to D_{t+1} (the conditional expectation of $\mathbf{r}_{t+1}$). This gives a guessed $\hat{\mathbf{r}}_{t+1}$.
2. Incremental *M*-step

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \varepsilon \mathbf{B}(\mathbf{x}_{t+1}, \hat{\mathbf{u}}_t)\{\hat{\mathbf{r}}_{t+1} - \boldsymbol{\eta}(\hat{\mathbf{u}}_t, \mathbf{x}_t)\}.$$

In this case, the log likelihood is written as

$$l = \sum_t \{\boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u}) \cdot \mathbf{r}_t - \psi(\mathbf{x}_t, \mathbf{u})\}.$$

In the *e*-step, we guess $\hat{\mathbf{r}}_{t+1}$ from the observed variable $\mathbf{s}_{v,t+1}$ and $\mathbf{x}_{t+1}$ based on $\hat{\mathbf{u}}_t$. In the incremental *M*-step, we use only the newest data $\hat{\mathbf{r}}_{t+1}$ to calculate the gradient, where old data $\hat{\mathbf{r}}, \ldots, \hat{\mathbf{r}}_t$ may be discarded.

Example 11. Learning of the stochastic perceptron. The *EM* and *em* algorithms are the same in this case. The $E(e)$-step works as follows. The conditional expectation conditioned on the observed y_{t+1} is

$$\begin{aligned} \hat{r}_{2,\mathbf{k}} &= E[\delta_{\mathbf{k}}(\mathbf{z})] = \text{Prob}\{z = \mathbf{k} \,|\, y_{t+1}, \mathbf{u}_t\} \\ &= \frac{\varphi(y_{t+1}, \mathbf{k} \cdot \mathbf{v}_t)\Pi\, \varphi(k_i, \mathbf{x} \cdot \mathbf{w}_{i,t})}{\sum_{\mathbf{k}} \varphi(y_{t+1}, \mathbf{k} \cdot \mathbf{v}_t)\Pi\, \varphi(k_i, \mathbf{x} \cdot \mathbf{w}_{i,t})} \\ \hat{r}_{1,\mathbf{k}} &= \begin{cases} 0, & y_{t+1} = 0 \\ \hat{r}_{2,\mathbf{k}}, & y_{t+1} = 1. \end{cases} \end{aligned} \tag{102}$$

In the learning phase, only current $\hat{\mathbf{r}}_{t+1}$ is evaluated by the above conditional expectation. In the batch processing, old $\hat{\mathbf{r}}_t$ is also modified based on $\hat{\mathbf{u}}_t$ which was renewed from $\hat{\mathbf{u}}_{t+1}$. We can use the incremental algorithm by calculating the gradient of $\hat{\mathbf{r}}$ with respect to $\mathbf{u}$.

The *M*-step is calculation of the m.l.e. The incremental algorithm uses the gradient of

$$l = \log p = \boldsymbol{\theta}(\mathbf{u}) \cdot \hat{\mathbf{r}} - \psi\{\boldsymbol{\theta}(\mathbf{u})\}.$$

We can calculate $\partial l/\partial \mathbf{u}$ by using eqn (16), giving

$$\begin{aligned} \frac{\partial l}{\partial v} &= \sum_{\mathbf{k}} \hat{r}_{2,\mathbf{k}}\{y - \varphi(1 \cdot \mathbf{k} \cdot \mathbf{v})\}\mathbf{k}, \\ \frac{\partial l}{\partial \mathbf{w}_i} &= \sum_{\mathbf{k}} \delta(k_i)\hat{r}_{2,\mathbf{k}}\mathbf{x} - \frac{\mathbf{x}\exp(\mathbf{w}_i \cdot \mathbf{x})}{1 + \exp(\mathbf{w}_i \cdot \mathbf{x})}. \end{aligned}$$

The new $\hat{\mathbf{u}}_{t+1}$ is given by

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \varepsilon_t \frac{\partial l}{\partial \mathbf{u}}.$$

IMPORTANT REMARK 1. When examples are generated by a noiseless (deterministic) network with $y = 0$ or 1, all of $|\mathbf{w}_i|$ and $|\mathbf{v}|$ diverge to infinity by learning. In such a case, we need to control their magnitude carefully by normalization. Then, learning works in such a case.

IMPORTANT REMARK 2. When an analog input–output relation $\bar{y} = f(\mathbf{x})$ is learned by using the stochastic perceptron with $y = 0, 1$ output, the values of $\bar{y}_t$ in the examples $(\bar{y}_1, \mathbf{x}_1), \ldots, (\bar{y}_T, \mathbf{x}_T)$ are not binary but real. In such a case $\bar{y}_t$ is interpreted that, when $\mathbf{x}_t$ is input, $\bar{y}_t$ is the probability of $y_t = 1$. In other words, when $\mathbf{x}_t$ is input in many times, $y_t = 1$ occurs with relative frequency $\bar{y}_t$. The conditional expectations are given by

$$\begin{aligned} \hat{r}_{1,k} &= \bar{y}_{t+1} E[\delta_{\mathbf{k}}(\mathbf{z}) \,|\, y_{t+1} = 1, \mathbf{u}_t], \\ \hat{r}_{2,k} &= \bar{y}_{t+1} E[\delta_{\mathbf{k}}(\mathbf{z}) \,|\, y_{t+1} = 1, \mathbf{u}_t] \\ &\quad + (1 - \bar{y}_{t+1}) E[\delta_{\mathbf{k}}(\mathbf{z}) \,|\, y_{t+1} = 0, \mathbf{u}_t]. \end{aligned} \tag{103}$$

9. DIFFERENTIAL AND INCREMENTAL FORMS OF THE *EM* ALGORITHM

From the geometrical point of view, the *EM* and *em* algorithms search for the (local) minimum of the divergence from D to M,

$$K(D \,\|\, M) = \min_{Q \in D,\, P \in M} K(Q \,\|\, P),$$

or

$$K^*(D \,\|\, M) = \min_{Q \in D,\, P \in M} K\{F^{-1}(Q) \,\|\, P\},$$

where Q moves in D and P moves in M. Therefore, we can associate dual gradient flows in D and in M of the single function $K(Q \,\|\, P)$ or $K[F^{-1}(Q) \,\|\, P]$. Let $Q \in D$ and $P \in M$ be a pair of points that move in the directions of decreasing $K(Q \,\|\, P)$, respectively. Then, we have the dual gradient flows $Q(t)$ and $P(t)$ (Figure 11). Let $\mathbf{u}(t)$ be the parameter to specify a point $P \in M$ whose $\boldsymbol{\theta}$-coordinates are $\boldsymbol{\theta}_P = \boldsymbol{\theta}(\mathbf{u}(t))$ and let $\boldsymbol{\eta}_Q(t) = (\boldsymbol{\eta}_{Q,v}, \boldsymbol{\eta}_{Q,h})$ be the coordinates of $Q \in D$ so that the observed part is restricted to $\boldsymbol{\eta}_{Q,v} = \bar{\mathbf{r}}_v$, where t is the parameter of the dual gradient curves. The gradient flows are given in these coordinates by

$$\frac{d}{dt}\mathbf{u}(t) = -\varepsilon G_M^{-1} \frac{\partial}{\partial \mathbf{u}} K(\boldsymbol{\eta}_Q \,\|\, \boldsymbol{\theta}_P), \tag{104}$$

$$\frac{d}{dt}\boldsymbol{\eta}_{Q,h}(t) = -\varepsilon G_D^{-1} \frac{\partial}{\partial \boldsymbol{\eta}_h} K(\boldsymbol{\eta}_Q \,\|\, \boldsymbol{\theta}_P). \tag{105}$$

Here G_M and G_D are the Fisher information matrices

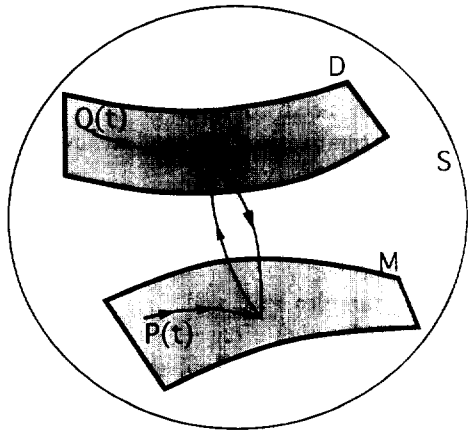

FIGURE 11. Dual gradient flows.

of M and D, respectively. It is easy to prove the relations

$$\frac{\partial}{\partial \theta_P} K(Q\|P) = \eta_P - \eta_Q,$$
$$\frac{\partial}{\partial \eta_Q} K(Q\|P) = \theta_Q - \theta_P.$$

Hence, the gradient flows are written as

$$\frac{d}{dt}\mathbf{u}(t) = -\varepsilon G_M^{-1} \frac{\partial \theta(\mathbf{u})}{\partial \mathbf{u}} \{\eta_P(t) - \eta_Q(\mathbf{u}(t))\}, \quad (106)$$
$$\frac{d}{dt}\eta_{Q,h}(t) = -\varepsilon G_D^{-1} \{\theta_Q, h(\mathbf{u}(t)) - \theta_{P,h}\theta_{Q,h}(t)\}, \quad (107)$$

where θ_P and η_Q are θ- and η-coordinates of P and Q, respectively. The flows are e-and m-geodesic flows in D and M, respectively (Fujiwara & Amari, 1995).

By discretizing the above geodesic flows and neglecting the metric tensor terms G^{-1}, we have the following incremental algorithm: For $t = 1, 2, 3, \ldots$,

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t - \varepsilon_t \frac{\partial \theta}{\partial \mathbf{u}} \{\hat{\eta}_t - \eta(\hat{\mathbf{u}}_t)\}, \quad (108)$$
$$\hat{\eta}_{h,t+1} = \hat{\eta}_{h,t} - \varepsilon_t' \{\theta_h(\hat{\mathbf{u}}_t) - \theta_h(\hat{\eta}_t)\}. \quad (109)$$

10. EXAMPLES — NORMAL MIXTURE, RADIAL BASIS EXPANSION AND MIXTURE OF EXPERT NETS

Here we give two important examples, briefly, to show the applicability of the stochastic method. More detailed studies are necessary for these example as well as the stochastic perceptron. Hidden Markov models (higher-order), Boltzmann machines, and Boltzmann machines with asymmetric connections are also good examples to be treated in the framework (see Amari et al., 1992; Ito, Amari, & Kobayashi, 1992).

10.1. Normal Mixture and Radial Basis Function

The normal mixture model (see Example 3) has been studied by Neal and Hinton (1993) from the point of view of the *EM* algorithm. We study the problem from the point of view of information geometry. The problem is originally concerned with the density estimation or approximation, but the same technique is applicable to the radial basis type function approximation where desired output signals are provided from a teacher. We have shown in eqn (8) that the normal mixture with hidden variable z is an exponential family. Here the parameters of the model are summarized as

$$\mathbf{u} = (p_1, \ldots, p_k; \mu_0, \ldots, \mu_k; \sigma_0^2, \ldots, \sigma_k^2).$$

Hence, we have $\theta = \theta(\mathbf{u})$, and this belongs to a curved exponential family M. Let us consider the case of repeated observations. Let $\mathbf{r}_1, \ldots, \mathbf{r}_T$ be T independent i.i.d. data. They are summarized in $\bar{\mathbf{r}}$,

$$\begin{aligned} \bar{r}_{11} &= \frac{1}{T}\sum x_t, \quad \bar{r}_{12} = \frac{1}{T}\sum x_t^2, \\ \bar{r}_{2i} &= \frac{1}{T}\sum \delta_i(z_t), \\ \bar{r}_{3i} &= \frac{1}{T}\sum x_t \delta_i(z_t), \\ \bar{r}_{4i} &= \frac{1}{T}\sum x_t^2 \delta_i(z_t), \end{aligned} \quad (110)$$

where $\mathbf{r}_t$ is a function of the tth observation (x_t, z_t), $t = 1, \ldots, T$.

The η-coordinates of S are given in terms of $\mathbf{u}$ by the expectation of $\mathbf{r}$, which is the same as that of $\bar{\mathbf{r}}$, $\eta = \eta(\mathbf{u})$,

$$\begin{aligned} \eta_{11} &= \sum_{i=0}^{k} p_i \mu_i, \quad \eta_{12} = \sum_{i=0}^{k} p_i(\mu_i^2 + \sigma_i^2), \\ \eta_{2i} &= p_i, \\ \eta_{3i} &= p_i \mu_i, \\ \eta_{4i} &= p_i(\mu_i^2 + \sigma_i^2). \end{aligned} \quad (111)$$

The M is given by $\eta = \eta(\mathbf{u})$ in the η-coordinates. Because S is an exponential family, when all the data $(\mathbf{x}_t, z_t)$ are observed, the observed point is given by $\hat{\eta} = \bar{\mathbf{r}}$ in S. The m.l.e. $\hat{\mathbf{u}}$ is given by projecting $\hat{\eta}$ to M. This is obtained by solving

$$\bar{\mathbf{r}} = \eta(\hat{\mathbf{u}}),$$

where $\mathbf{u} = (p_i, \mu_i, \sigma_i^2)$, $i = 0, \ldots, k$. The m.l.e. $\hat{\mathbf{u}}$ minimizes $K[\bar{\mathbf{r}}\|\eta(\mathbf{u})]$.

When the hidden variables z_t are not observed, we

cannot summarize $\mathbf{r}_t (t = 1, \ldots, T)$ into a single $\bar{\mathbf{r}}$. We need to treat the product space $S_T^* = S_1 \times \cdots \times S_T$. Partial observation then defines a data submanifold D_t, and hence $D_T^* = D_1 \times \cdots \times D_T$. We show this by using a simpler case of $\sigma_0 = \sigma_i = 1$. The general case can be analyzed quite similarily. In this special case, the log likelihood is written as

$$\log p(x; p_i, \mu_i) = \mu_0 x + \sum_{i=1}^{k} \delta_i(z) \left\{ \log \frac{p_i}{p_0} - \frac{1}{2}(\mu_i^2 - \mu_0^2) \right\} + \sum_{i=1}^{k} x \delta_i(z)(\mu_i - \mu_0) + \log p_0 - \frac{1}{2}\mu_0^2 - \frac{x^2}{2}. \tag{112}$$

Therefore, this is a full (not curved) exponential family, that is $M = S$, where the random variable $\mathbf{r} = (r_0, r_{1i}, r_{2i})$, the $\boldsymbol{\theta}$-coordinates and the $\boldsymbol{\eta}$-coordinates are given, respectively, by

$$r_0 = x, \quad r_{1i} = \delta_i(z), \quad r_{2i} = x\delta_i(z), \tag{113}$$

$$\theta_0 = \mu_0, \quad \theta_{1i} = \log (p_i/p_0) - \frac{1}{2}(\mu_i^2 - \mu_0^2), \quad \theta_{2i} = \mu_i - \mu_0 \tag{114}$$

$$\eta_0 = \sum_{i=0}^{k} p_i \mu_i, \quad \eta_{1i} = p_i, \quad \eta_{2i} = p_i \mu_i, \quad (i = 1, \ldots, k). \tag{115}$$

When x_t is observed but z_t is not at time t, $\mathbf{s}_v = x_t$ and $\mathbf{s}_h = \{\delta_i(z_t)\}$, and the observed data submanifold D_t is given by

$$D_t = \{\hat{\boldsymbol{\eta}} \mid \hat{\eta}_0 = x_t, \hat{\eta}_{1i} = \alpha_i, \hat{\eta}_{2i} = x_t \alpha_i\} \tag{116}$$

where α_i are the free parameters corresponding to the unobserved $\delta_i(z_t)$. It should be noted that $\delta_i(z_t)$ is 0 or 1 but α_i takes any real values satisfying $0 \leq \alpha_i$, $\Sigma \alpha_i \leq 1$. The D_t is a linear submanifold in $\boldsymbol{\eta}$, but it depends on x_t. Hence, D_t is different for each t, so that we cannot summarize them into a single submanifold D but we need to treat the product D_T^*. The model M_T^* is simply given by $\boldsymbol{\theta}_t = \boldsymbol{\theta}$. Hence, M_T^* is a submanifold of S_T^*.

Now we discuss the E-step. For a candidate point $\hat{P} \in M$, the conditional expectation of the missing variable $\delta_i(z_t)$ is calculated as

$$\alpha_i^t(x_t) = E_{\hat{P}}[\delta_i(z_t) \mid x_t] = \frac{\hat{p}_i \exp\{-\frac{1}{2}(x_t - \hat{\mu}_i)^2\}}{\sum_j \hat{p}_j \exp\{-\frac{1}{2}(x_t - \hat{\mu}_j)^2\}}. \tag{117}$$

This is the same as $E_Q[\delta_i(z_t) \mid x_t]$, where Q is the e-projection of $\hat{P}$ to D_t. It is easy to show that the e-projection gives the same answer.

We can write down the learning or incremental form of the *EM* algorithm. This can easily be generalized to the multidimensional normal mixture,

$$p(\mathbf{x}, z; \mathbf{u}) = \sum_{i=0}^{k} \delta_i(z) p_i \frac{1}{(\sqrt{2\pi})^d |\sum_i|^{1/2}} \times \exp\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \textstyle\sum_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\},$$

where $\mathbf{u}$ consists of p_i, $\boldsymbol{\mu}_i$, and Σ_i.

The probability $p(\mathbf{x}; \mathbf{u})$ looks like the radial basis expansion of the density function. The radial basis expansion is used to approximate a function $y = f(\mathbf{x})$ in the expanded form

$$f(\mathbf{x}) \sim \sum p_i \frac{1}{(\sqrt{2\pi})^d |\sum_i|^{1/2}} \times \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \sum_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\}.$$

When $f(\mathbf{x}) > 0$ for all $\mathbf{x}$ in a domain R and

$$\int_R f(\mathbf{x}) d\mathbf{x} = 1,$$

we can interpret y as the probability density of $\mathbf{x}$, which is unknown. When $(\mathbf{x}_t, y_t)$ is observed but z_t is not, we interpret that $\mathbf{x}_t$ is observed a number of times proportional to the teacher signal y_t. When $\mathbf{x}$ is uniformly distributed over R, we can use the *EM* algorithm for supervised learning of $y = f(\mathbf{x})$. When $\mathbf{x}$ is not uniform but has an unknown density $q(\mathbf{x})$, we estimate $q(\mathbf{x})$ by another network. The result of the former network gives an approximation to $f(\mathbf{x})/q(\mathbf{x})$, so that we can calculate an approximation to $f(\mathbf{x})$.

10.2. Mixture of Expert Neural Nets

A simplest case of mixtures of expert nets (Jacobs et al., 1991; Jordan & Jacobs, 1994) is presented here. There are various generalizations, including the hierarchical mixture (Jordan and Jacobs, 1994; Xu et al., 1994). Another generalization will be shown in the end of this section. Let $N_i (i = 0, 1, \ldots, k)$ be $k + 1$ stochastic neural networks called experts, receiving a common input $\mathbf{x}$ and emitting a binary output y_i. In the present simplest case, we assume that N_i is a simple stochastic neuron such that it emits a binary output y_i depending on the weighted sum $u_i = \mathbf{w}_i \cdot \mathbf{x}$ of the input $\mathbf{x}$ (Amari, 1991). The

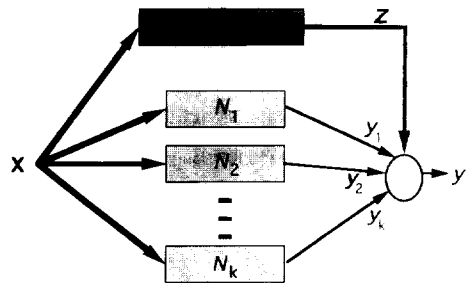

FIGURE 12. Mixture of expert nets.

probability of y_i given $\mathbf{x}$ is written as

$$p(y_i|\mathbf{x}, \mathbf{w}_i) = \varphi(y_i, \mathbf{w}_i \cdot \mathbf{x}) = \exp\{y_i \mathbf{x} \cdot \mathbf{w}_i - \psi(\mathbf{w}_i \cdot \mathbf{x})\}, \quad (118)$$

The mixture of experts is composed of $k+1$ expert networks N_i with a gating network N that selects one of the expert nets for processing input $\mathbf{x}$. It receives the same input $\mathbf{x}$ and its output $\mathbf{z}$ is a random variable taking values on $\{0, 1, \ldots, k\}$. When the output of the gating network is $z = i$, the gating network decides that the signal $\mathbf{x}$ should be processed by network N_i. Hence, the final output y is equal to y_i in this case (Figure 12). In general, the output y is written as

$$y = \sum_{i=0}^{k} \delta_i(z) y_i.$$

The output z of the gating network is determined stochastically by the softmax function

$$g_i(\mathbf{x}) = p(z = i|\mathbf{x}) = \frac{\exp(\mathbf{v}_i \cdot \mathbf{x})}{\sum_{j=0}^{k} \exp\{\mathbf{v}_j \cdot \mathbf{x}\}}, \quad (119)$$

where $\mathbf{v}_i$ is the connection weight vector of the ith gating output. Without loss of generality, we may put $\mathbf{v}_0 = 0$, because g_i is invariant under the transformation $\mathbf{v}_j \to \mathbf{v}_j - \mathbf{a}$ for any $\mathbf{a}$ and $j = 0, \ldots, k$.

The joint probability of (y, z) is

$$p(y, z|\mathbf{x}) = \exp\left\{\sum_{i=1}^{k} \delta_i(z)(\mathbf{v}_i \cdot \mathbf{x} - \psi_i + \psi_0) + \sum_{i=1}^{k} y\delta_i(z)(\mathbf{w}_i - \mathbf{w}_0) \cdot \mathbf{x} + y\mathbf{w}_0 \cdot \mathbf{x} - \psi\right\}, \quad (120)$$

where

$$\psi_i = \log\{1 + \exp(\mathbf{w}_i \cdot \mathbf{x})\}, \quad (121)$$

$$\psi = \log\left\{1 + \sum_{j=1}^{k} \exp(\mathbf{v}_j \cdot \mathbf{x})\right\} + \psi_0. \quad (122)$$

By putting

$$r_1 = y, \quad r_{2i} = \delta_i(z) y, \quad r_{3i} = \delta_i(z), \quad (123)$$

$$\theta_1 = \mathbf{w}_0 \cdot \mathbf{x}, \quad \theta_{2i} = (\mathbf{w}_i - \mathbf{w}_0) \cdot \mathbf{x}, \quad \theta_{3i} = \mathbf{v}_i \cdot \mathbf{x} - \psi_i + \psi_0, \quad (124)$$

we have

$$p(y, z|\mathbf{x}) = \exp\{\boldsymbol{\theta} \cdot \mathbf{r} - \psi\}. \quad (125)$$

Therefore, the family $\{p(y, z|\mathbf{x})\}$ of the conditional distributions of the mixtures of experts is an exponential family, where $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{x})$.

We have shown the simplest case where N_i is a binary stochastic neuron. Jacobs et al. (1991) and Jordan and Jacobs (1994) treated a more general case that the conditional distribution of the output y_i of N_i belongs to a general exponential family, for example,

$$p(y_i|\mathbf{x}, \mathbf{w}_i) = \exp\{\theta(\mathbf{w}_i \cdot \mathbf{x}) y_i + k - \psi(\theta)\}. \quad (126)$$

The binary neuron case (17) is given by putting

$$\theta(\mathbf{w}_i \cdot \mathbf{x}) = \mathbf{w}_i \cdot \mathbf{x}, \quad \psi = \log\{1 + \exp(\mathbf{w}_i \cdot \mathbf{x})\}.$$

Another typical example of N_i is an analog stochastic neuron where the output y_i is written as

$$y_i = f(\mathbf{w}_i \cdot \mathbf{x}) + n, \quad (127)$$

where f is the sigmoidal function and n is a random noise subject to $N(0, \sigma^2)$. In this case

$$\theta = \frac{1}{\sigma^2} f(\mathbf{w}_i \cdot \mathbf{x}), \quad k = \frac{y_i^2}{2\sigma^2}, \quad \psi = \frac{1}{2\sigma^2}\{f(\mathbf{w}_i \cdot \mathbf{x})\}^2 + \log(\sqrt{2\pi}\sigma).$$

The joint probability of the mixture of general N_i of eqn (126) is again written as

$$p(y, z|\mathbf{x}) = \exp\left\{\sum \theta_{ij} r_{ij} - \psi\right\},$$

where $\mathbf{r}$ is the same as eqn (123) and

$$\begin{aligned}\theta_1 &= \theta(\mathbf{w}_0 \cdot \mathbf{x}),\\ \theta_{2i} &= \theta(\mathbf{w}_i \cdot \mathbf{x}) - \theta(\mathbf{w}_0 \cdot \mathbf{x}),\\ \theta_{3i} &= \mathbf{v}_i \cdot \mathbf{x} - \psi\{\theta(\mathbf{w}_i \cdot \mathbf{x})\} + \psi\{\theta(\mathbf{w}_0 \cdot \mathbf{x})\}.\end{aligned} \tag{128}$$

The conditional probability of the total net is again an exponential family.

The mixture of experts is of stochastic model. It can be trained from examples based on stochastic techniques. However, it can be used deterministically in the execution mode. In this case, an expert N_i emits the expectation $E[y_i]$ of the stochastic y_i, and the gating network gives weights

$$g_i = E[\delta_i(z)] = \text{Prob}\{z = z_i\}$$

such that the final output $\bar{y}$ is the weighted sum,

$$\bar{y} = \sum_{i=0}^{k} g_i E[y_i].$$

It is also possible that the gating network N plays the winner-take-all rule, choosing one candidate network N_i that has the largest probability of $z = i$.

It is important that the mixtures of exponential family expert nets are again an exponential family. Jordan and Jacobs (1994) showed that a hierarchical mixture can be successively constructed by using mixtures of expert nets as component expert nets. This is again an exponential family, so that a higher-order hierarchical mixture is constructed in this manner. We show one hierarchical step. Let N_i^*, $i = 0, \ldots, m$, be $(m+1)$ mixtures of experts nets, consisting of $k_i + 1$ expert nets N_i, $(j = 0, \ldots, k_i)$. Let z_i be the output of the gating network of N_i^*. We then construct a mixture of N_i^*'s where z^* is the variable of the total gating network. When $z^* = i$, the mixture N_i^* is selected, and then a network $N_{i,j}$ is selected from N_i^* when $z_i = j$. Thus, the final output y is equal to the output $y_{i,j}$ of $N_{i,j}$. Hence, we have

$$y = \sum_{i,j} \delta_i(z^*)\delta_j(z_i)y_{i,j}. \tag{129}$$

This shows that the hierarchical structure is used to generalize the gating mechanism, one of $N_{i,j}$ being selected hierarchically depending on $\mathbf{x}$. Xu et al. (1994) proposed a different gating mechanism that is computationally tractable.

We propose another gating mechanism. Let $N_0, \ldots, N_k$ be component expert nets. We denote them by N_μ, $\mu = 0, \ldots, k$. The variable z of gating network takes values on the set $A = \{0, 1, \ldots, m\}$, where $m \geq k$. Now the set A is partitioned into $k+1$ subsets,

$$A = \{I_0, I_1, \ldots, I_k\}$$

such that $I_0 = \{0, 1, \ldots, i_0\}$, $I_1 = \{i_0 + 1, \ldots, i_0 + i_1\}, \ldots$. When z takes its value in I_μ, $\mu = 0, 1, \ldots, k$, that is, $z \in I_\mu$, the component network N_μ is selected, so that

$$y = \sum_{\mu=0}^{K} \delta_\mu^*(z)y_\mu, \tag{130}$$

where $\delta_\mu^*(z) = 1$ when $z \in I_\mu$ and is otherwise 0.

This is again an exponential family, because the conditional distribution of (y, z) is obtained, in a similar way as eqn (20), by

$$p(y, z \mid \mathbf{x}) = \exp\left\{\sum_{i}^{m} = \delta_i(z)(\mathbf{v}_i \cdot \mathbf{x} - \psi_{i+0}) + \sum_{\mu=1}^{k} y\delta\mu(z)(\mathbf{w}_\mu - \mathbf{w}_0) \cdot \mathbf{x} + y\mathbf{w}_0 \cdot \mathbf{x} - \psi\right\}.$$

The statistics $\mathbf{r}$ is

$$\begin{aligned}r_1 &= y,\\ r_{2\mu} &= \delta_\mu^*(z), \quad \mu = 1, \ldots, k,\\ r_{3i} &= \delta_i(z), \quad i = 1, \ldots, m,\end{aligned}$$

in this case and $\boldsymbol{\theta}$ is similar to eqn (124) where θ_{2i} is replaced by $\theta_{2\mu}$.

The above grating net has no hierarchical structure but is very simple and flexible. We explain this by using the deterministic limit where softmax function is replaced by the max function or the winner-take-all mechanism. A gating network with the maximum selector divides the set $X = \{\mathbf{x}\}$ of input signals into subregions X_i

$$X = \bigcup X_i. \tag{131}$$

A signal $\mathbf{x}$ is processed by N_i when $\mathbf{x} \in X_i$. It is known that gating network realizes a Laguerre-Voronoi division (Zhuang & Amari, 1993), which is a little more general than the Voronoi division. The hierarchical mixture divides each region further into Laguerre–Voronoi subregions

$$X_i = \bigcup X_{i,j}, \tag{132}$$

depending on the values of z_i. When $\mathbf{x}$ belongs to $X_{i,j}$, it is processed by $N_{i,j}$. Thus, the hierarchical structure is capable of forming a more complex division than the simple Laguerre–Voronoi one, although each $X_{i,j}$ is still convex.

On the other hand, the proposed net divides X into

$m+1$ Laguerre–Voronoi regions X_j, whereas the region X_μ^* on which signals are processed by N_μ, consists of unions of the corresponding subregions

$$X_\mu^* = \bigcup_{j \in I_\mu} X_j. \tag{133}$$

Each X_μ^* can be nonconvex in this case, and it is universal in the sense that any division can be approximated by this method if m is sufficiently large. Hence, this gating network realizes a more flexible and complex division in a simple manner.

We can write down the *EM* and *em* algorithm and the incremental or learning version explicitly.

11. CONCLUSIONS

The information geometry of the *EM* algorithm and the *em* algorithm is constructed. It is proved that they are equivalent in most practical cases, and are equivalent asymptotically and also in the extended function space. Thus, a unified geometrical framework is given to the *EM* and *em* algorithms. This makes it possible to formulate a learning version of the *EM* algorithm. The stochastic multilayer perceptron, normal mixture models, and mixture of experts are treated as examples. This framework opens a new area of research connecting neural networks, statistics, and geometry.

REFERENCES

Amari, S. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Annals of Statistics*, **10**, pp. 357–385.

Amari, S. (1985). *Differential geometrical methods in statistics*, Springer Lecture Notes in Statistics, 28, Springer.

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

Amari, S. (1987b). Differential geometrical theory of statistics. In Amari, S. et al. (Eds.), *Differential geometry in statistical inference*, IMS Monograph Series (Vol. **10**, Chap. 2, pp. 19–94). Heyward, CA: IMS.

Amari, S. (1989). Fisher information under restriction of Shannon information in multiterminal situations. *Annals of Institute of Statistical Mathematics*, **41**, 623–648.

Amari, S. (1990). Mathematical foundations of neurocomputing. *Proceedings of the IEEE*, **78**, 1443–1463.

Amari, S. (1991). Dualistic geometry of the manifold of higher-order neurons. *Neural Networks*, **4**, 443–451.

Amari, S. (1995). The EM algorithm and information geometry in neural network learning. *Neural Computation*, **7**, 13–18.

Amari, S., Barndorff–Nielsen, O. E., Kass, R. E., Lauritzen, S. L., & Rao, C. R. (1987). *Differential geometry in statistical inferences*, IMS Lecture Notes Monograph Series (Vol. **10**). Hayward, CA: IMS.

Amari, S., & Han, T. S. (1989). Statistical inference under multiterminal rate restrictions—a differential geometrical approach. *IEEE Transactions on Information Theory*, **IT-35**, 217–227.

Amari, S., & Kawanabe, M. (1994). Information geometry of estimating functions in semiparametric statistical models. *METR*, **94-1**, University of Tokyo.

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

Baldi, P., & Chauvin, Y. (1994). Smooth on-line learning algorithm for hidden Markov models. *Neural Computation*, **6**, 307–318.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*, Chichester: John Wiley.

Barndorff-Nielsen, O. E. (1988). *Parametric statistical model and likelihood*, Springer Lecture Notes in Statistics, Vol. 50, Springer.

Barndorff-Nielsen, O. E., Cox, R. D., & Reid, N. (1986). The role of differential geometry in statistical theory. *International Statistical Review*, **54**, 83–96.

Barndorff-Nielsen, O. E., Jupp, P. E. (1989). Approximating exponential models. *Annals of Institute of Statistical Mathematics*, **41**, 247–267.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Besag, J., & Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of Royal Statistical Society*, **B-55**, 25–37.

Byrne, W. (1992). Alternating minimization and Boltzmann machine learning. *IEEE Transactions on Neural Networks*, **3**, 612–620.

Cheng, B., & Titterington, D. M. (1994). Neural networks—a review from statistical perspectives—comments and rejoinders. *Statistical Science*, **9**, 31–54.

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

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*, London: Chapman and Hall.

Csiszár, I. (1975) I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, **3**, 146–158.

Csiszár, I., & Tusnády, G. (1984). Information geometry and alternating minimization procedures. In E. F. Dedewicz, et al. (Eds), *Statistics and decisions* (Supplementary Issue, no. 1, pp. 205–237), Munich: Oldenburg Verlag.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, **B-39**, 1–38.

Fujiwara, A., & Amari, S. (1995). Dualistic dynamical systems in the framework of information geometry. *Physica D*, **80**, 317–327.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721–741.

Guan, Y., Clarkson, T. G., Taylor, J. G., Gorse, D. (1994). Noisy reinforcement training for pRAM nets. *Neural Networks*, **7**, 523–538.

Ito, H., Amari, S., & Kobayashi, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Transactions on Information Theory*, **IT-38**, 324–333.

Jordan, M. I., & Jacobs, R. A. (1994). Higherarchical mixtures of experts and the *EM*-algorithm. *Neural Computation*, **6**, 181–214.

Jordan, M. I., & Xu, L. (1994). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks* (in press).

Jacobs, R. A., Jordan, M. I., Nolwan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.

Kass, R. E. (1989). The geometry of asymptotic inference (with discussions). *Statistical Science*, **4**, 188–234.

Kawanabe, M., & Amari, S. (1994). Estimation of network parameters in semiparametric stochastic perceptron. *Neural Computation*, **6**, 1244–1261.

Künsch, H., Geman, S., & Kehagias, A. (1993). Hidden Markov random field, to appear.

Kurose, T. (1990). Dual connections and affine geometry. *Mathematische Zeitshrift*, **203**.

Murray, A. F., & Edwards, P. J. (1994). Enhanced MLP performance and fault tolerance resulting from synaptic weight noise during training. *IEEE Transactions*, **NN-5**, 792–802.

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

Nagaoka, H., & Amari, S. (1982). Differential geometry of smooth families of probability distributions. *METR*, **82-7**.

Neal, R. N., & Hinton, G. E. (1993). A new view of the EM algorithm that justifies incremental and other variants. To appear.

Nomizu, K., & Simon, U. (1991). Notes on conjugate connections. Preprint.

Ohara, A., & Amari S. (1992). Differential geometric structures of stable feedback systems with dual connections. in *Proceedings of the 2nd Workshop on Systems, Structure and Control*, Prague, pp. 176–179.

Okamoto, I., Amari, S., & K. Takeuchi (1991). Asymptotic theory of sequential estimation: Differential geometrical approach. *Annals of Statistics*, **19**, 961–981.

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

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: John Wiley.

Ripley, B. D. (1994). Neural networks and related method for classification. *Journal of Royal Statistical Society*, **B56**, 409–456.

Shimodaira, H. (1993). A new criterion for selecting models from partially observed data. *Proceedings of 4th International Workshop on Artificial Intelligence and Statistics*, pp. 381–386.

Streit, R. L., & Luginjbuhl, T. E. (1994). Maximum likelihood training of probabilistic neural networks. *IEEE Transactions*, **NN-5**, 764–783.

White, H. (1989). Learning in artificial networks: A statistical perspective. *Neural Computation*, **1**, 425–464.

Xu, L., Jordan, M. I., & Hinton, J. (1994). New gating net for mixture of experts, EM algorithm and piecewise function approximations. Preprint.

Yuille, A. L., Stolorz, P., & Uttans, J. (1994). Statistical physics, mixtures of distributions and the EM algorithm. *Neural Computation*, **6**, 334–340.

Zhuang, J., & Amari, S. (1993). Piecewise-linear division of signal space by a multilayer neural networks with the maximum detector (in Japanese). *Transactions of the Institute of Electronics, Information and Communication Engineers*. **J76-D**, 1435–1440.

APPENDIX: INFORMATION GEOMETRY

A.1. Dual Geometry of Exponential Family

Invariant geometrical structures of a general manifold S of probability distributions have been studied in detail (Amari, 1985; Murray & Rice, 1993, etc.) to obtain intrinsic properties of a statistical model. The geometrical theory has successfully been applied to various fields of information sciences such as statistics (Amari, 1985; Kass, 1989), systems theory (Amari, 1987a; Ohara & Amari, 1992), information theory (Amari & Han, 1989; Amari, 1989), neural networks (Amari, 1991; Amari et al. 1992), and many others. Mathematicians are studying this new geometrical structure of differential geometry (Nomizu & Simon, 1991; Kurose, 1990). It is a Riemannian manifold equipped with a couple of dual affine connections. The duality in affine connections is a new notion introduced in differential geometry originated from information science. However, we do not mention the differential geometrical concepts such as the Riemannian metric, affine connection, curvature, etc., in detail. Instead, we show two types of straightness of a curve (geodesic) in a dually flat manifold. The orthogonality of two curves is also shown intuitively. We then show fundamental theorems on a dually flat manifold. Readers are asked to refer to Amari (1985), Murray and Rice (1993), and related works.

Here we explain the exponential straightness (geodesic) and mixture straightness (geodesic) in an intuitive way. See Section A.4 for a more formal definition based on the covariant derivative and affine connection. We first treat the manifold S of discrete probabilities of Example 2, where x takes on $\{0, 1, \ldots, n\}$. A probability distribution in S is denoted by $p(x)$ or $\mathbf{p} = (p_i)$, where $p_i = \text{Prob}\{x = i\}$. For two probability distributions $p_1(x)$ and $p_2(x)$, there are two special curves connecting them in the manifold S. When we connect $\log p_1(x)$ and $\log p_2(x)$ linearly, we have the exponential family $\{p(x; t)\}$ given by

$$\log p(x; t) = (1 - t)\log p_1(x) + t \log p_2(x) - \psi(t) \qquad \text{(A.1)}$$

or

$$p(x; t) = \exp\{tr(x) + \log p_1(x) - \psi(t)\},$$

where

$$r(x) = \log \frac{p_2(x)}{p_1(x)}$$

is a new random variable, $\psi(t)$ is the normalization factor, and $0 \leq t \leq 1$ is the parameter of the curve. This curve is regarded as a "straight line" (geodesic) connecting $p_1(x)$ and $p_2(x)$ in S from the exponential family standpoint. In terms of the $\boldsymbol{\theta}$-coordinate system, the coordinates $\boldsymbol{\theta}(t)$ of $p(x, t)$ are written as

$$\boldsymbol{\theta}(t) = (1 - t)\boldsymbol{\theta}_1 + t\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1), \qquad \text{(A.2)}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the $\boldsymbol{\theta}$-coordinates of $p_1(x)$ and $p_2(x)$, respectively. Therefore, the exponential geodesic is a linear curve in the $\boldsymbol{\theta}$-coordinates.

The other is the mixture family $\{p^*(x, t)\}$ connecting the two distributions by the curve

$$p^*(x, t) = (1 - t)p_1(x) + tp_2(x). \qquad \text{(A.3)}$$

This curve is regarded as a "straight line" (geodesic) from the mixture standpoint. Both standpoints have their own proper meanings. It is easy to show that the $\boldsymbol{\eta}$-coordinates of $p^*(x, t)$ is written as

$$\boldsymbol{\eta}^*(t) = \boldsymbol{\eta}_1 + t(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1). \qquad \text{(A.4)}$$

Hence, the mixture geodesic is a linear curve in the $\boldsymbol{\eta}$-coordinates.

By generalizing this idea, we give the following definitions of straightness or flatness in the manifold S of any exponential family, where we have two coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Let us consider a curve $\boldsymbol{\theta}(t)$ connecting two points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ linearly in the $\boldsymbol{\theta}$-coordinates,

$$\boldsymbol{\theta}(t) = t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) + \boldsymbol{\theta}_1.$$

This curve is regarded as a straight line from the exponential standpoint and is called an exponential geodesic or e-geodesic. In particular, each coordinate curve $\theta_i = t$, $\theta_j = c_j (j \neq i)$ of the $\boldsymbol{\theta}$-coordinate system is an e-geodesic. This implies that the manifold S is e-flat, having an affine coordinate system $\boldsymbol{\theta}$, from the e-flatness point of view. The $\boldsymbol{\theta}$ is called the e-coordinate system.

On the other hand, when a curve connecting two distributions $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ is linear

$$\boldsymbol{\eta}^*(t) = t(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) + \boldsymbol{\eta}_1$$

in the $\boldsymbol{\eta}$-coordinate system, the curve is said to be the mixture geodesic or m-geodesic connecting $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. The coordinate curves η_i of the $\boldsymbol{\eta}$-coordinate system are m-geodesics by themselves. Because the coordinate transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is in general nonlinear, an e-geodesic is not an m-geodesic in general. Therefore, we have two different criteria of flatness. The manifold of an exponential family is flat from both criteria so that it is called a dually flat manifold. A dually flat manifold has rich differential geometrical structures.

A.2. Orthoganlity and Fisher Information

We explain the tangent space and the Riemannian metric to define the orthogonality of two curves. Let $\mathbf{e}_i (i = 1, \ldots, n)$ be the tangent vector along the coordinate curve θ_i, that is, the direction in which θ_i changes but no other $\theta_j (j \neq i)$ change (Figure A.1). (Mathematicians denote it symbolically by $\partial/\partial\theta_i$.) The tangent space T_p of S at a point $P = (\boldsymbol{\theta})$ is the vector space spanned by $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. When we use the m-coordinate $\boldsymbol{\eta}$, the tangent direction of the coordinate curve η_i is denoted by $\mathbf{e}_i^*$. The same tangent space T_p is spanned also by $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*\}$.

We now introduce the inner product in tangent space T_p. To this end, we write the inner product of $\mathbf{e}_i$ and $\mathbf{e}_j$ as

$$g_{ij}(\boldsymbol{\theta}) = \langle \mathbf{e}_i, \mathbf{e}_j \rangle, \tag{A.5}$$

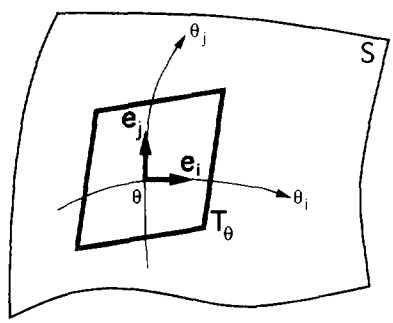

FIGURE A.1. Tangent space.

where $G = (g_{ij})$ is an $n \times n$ positive-definite matrix. It is natural to define it by

$$g_{ij}(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial\theta_i} \log p(x; \boldsymbol{\theta}) \frac{\partial}{\partial\theta_j} \log p(x; \boldsymbol{\theta})\right], \tag{A.6}$$

where E denotes the expectation with respect to $p(x; \boldsymbol{\theta})$ (Rao, 1945). This $G = (g_{ij})$ is called the Fisher information matrix, which plays a central role in theoretical statistics. If we use the m-coordinate system, the same inner product should be given in terms of the basis vectors $\mathbf{e}_i^*$ as

$$\begin{aligned} g_{ij}^* &= \langle \mathbf{e}_i^*, \mathbf{e}_j^* \rangle \\ &= E\left[\frac{\partial}{\partial\eta_i} \log p(x; \boldsymbol{\theta}(\boldsymbol{\eta})) \frac{\partial}{\partial\eta_j} \log p(x; \boldsymbol{\theta}(\boldsymbol{\eta}))\right]. \end{aligned} \tag{A.7}$$

It is proved that $G^* = (g_{ij}^*)$ is the inverse matrix of G.

A manifold S is said to be Riemannian when the inner product $G(\boldsymbol{\theta})$ is defined on the tangent space T_p at each point P. Our manifold of probability distributions is a Riemannian manifold having two different but dually coupled criteria of "flatness."

Let $\boldsymbol{\theta} = \boldsymbol{\theta}(t)$ or $\boldsymbol{\eta} = \boldsymbol{\eta}(t)$ be a curve. The tangent of the curve is a vector given by

$$\dot{\boldsymbol{\theta}}(t) = \sum \dot{\theta}_i(t) \mathbf{e}_i \tag{A.8}$$

where $\dot{}$ denotes d/dt. If the same curve is denoted by $\boldsymbol{\eta} = \boldsymbol{\eta}(t)$ in the m-coordinate system, the same tangent vector is written in terms of the $\{\mathbf{e}_i^*\}$ as

$$\dot{\boldsymbol{\eta}}(t) = \sum \eta_i(t) \mathbf{e}_i^*(t). \tag{A.9}$$

When two curves $\boldsymbol{\theta}_1(t)$ and $\boldsymbol{\theta}_2(t)$ intersect at $\boldsymbol{\theta}$, they are orthogonal if the inner product of the two tangent vector vanishes,

$$0 = \langle \dot{\boldsymbol{\theta}}_1(t), \dot{\boldsymbol{\theta}}_2(t) \rangle = \sum g_{ij}(\boldsymbol{\theta}) \dot{\theta}_{1i} \dot{\theta}_{2j}.$$

This is rewritten in the m-coordinates as

$$\sum g_{ij}^*(\boldsymbol{\eta}) \dot{\eta}_{1i} \dot{\eta}_{2j} = 0.$$

Now we show the important duality relation between the bases $\{\mathbf{e}_i\}$ and $\{\mathbf{e}_i^*\}$. It is proved from $G^* = G^{-1}$ that

$$\mathbf{e}_i = \sum g_{ij} \mathbf{e}_j^*, \quad \mathbf{e}_j^* = \sum g_{ij}^* \mathbf{e}_i. \tag{A.10}$$

This implies the following theorem.

THEOREM A1. *The two bases $\{\mathbf{e}_i\}$ and $\{\mathbf{e}_i^*\}$ are dual or reciprocal in the sense that*

$$\langle \mathbf{e}_i, \mathbf{e}_j^* \rangle = \delta_{ij} \tag{A.11}$$

holds at any point of M, where δ_{ij} is the Kronecker delta.

The inner product of two vectors $\mathbf{a}$ and $\mathbf{b}$ can easily be obtained by representig them in the dual bases as

$$\mathbf{a} = \sum a_i \mathbf{e}_i, \quad \mathbf{b} = \sum b_j^* \mathbf{e}_j^*.$$

Then, the inner product is given by

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum a_i b_i^*. \tag{A.12}$$

This shows the usefulness of the dual bases.

The set S of probability distributions of an exponential family (2) is regarded as an n-dimensional manifold having two coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ among others. It is a dually flat Riemannian manifold, where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ play a special dual role. In such a manifold, it is shown (Amari, 1985; Murray & Rice, 1993) that these two coordinate systems are connected by the Legendre transformation

$$\eta_i = \frac{\partial}{\partial\theta^i} \psi(\boldsymbol{\theta}), \tag{A.13}$$

$$\theta_i = \frac{\partial}{\partial\eta^i} \varphi(\boldsymbol{\eta}), \tag{A.14}$$

where $\varphi(\boldsymbol{\eta})$ is defined by the relation

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \sum \theta_i \eta_i = 0. \tag{A.15}$$

It is known that $-\varphi(\boldsymbol{\eta})$ is the entropy $H(\boldsymbol{\eta})$ of the distribution specified by $\boldsymbol{\eta}$.

A.3. Information Geometry of S

The manifold S of an exponential family is a dually flat Riemannian space. More precisely, S has two dually coupled affine connections with respect to the Riemannian metric $G(\boldsymbol{\theta})$ and the Riemann–Christoffel curvatures vanish with respect to these connections but the Levii–Civita connection has nonzero curvature. When a manifold is dually flat, it is proved that an invariant divergence measure $K(P, Q)$ is defined between two points $P, Q \in S$.

The divergence $K(P, Q)$ is derived from the geometric structure of the underlying manifold. In the case of the manifold of an exponential family, it is written as

$$K(\boldsymbol{\theta}_P, \boldsymbol{\eta}_Q) = \psi(\boldsymbol{\theta}_P) + \varphi(\boldsymbol{\eta}_Q) - \theta_{Pi}\eta_{Qi}, \tag{A.16}$$

where $\boldsymbol{\theta}_P$ and $\boldsymbol{\eta}_Q$ are the $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$-coordinates of P and Q, respectively, and is proved to be equal to the KL divergence

$$K(P\|Q) = E_{\boldsymbol{\theta}_P}\left[\log \frac{p(\mathbf{r}; \boldsymbol{\theta}_P)}{p(\mathbf{r}; \boldsymbol{\theta}_Q)}\right]. \tag{A.17}$$

The divergence is not symmetric, that is, $K(P, Q) \neq K(Q, P)$ in general, but $K(P, Q) \geq 0$ and the equality holds when and only when $P = Q$. Moreover, when P and $Q = P + dP$ are close, $K(P, P + dP)$ is a half of the square of the Riemannian distance,

$$K(P, P + dP) = \frac{1}{2} \sum g_{ij}(\boldsymbol{\theta}) d\theta_i d\theta_j, \tag{A.18}$$

where the coordinates of P and $P + dP$ are $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + d\boldsymbol{\theta}$, respectively. The essential role of the divergence is shown by the following generalized Pythagoras theorem (Figure A.2) (Nagaoka & Amari, 1982; Amari, 1985).

THEOREM A2. *Let P, Q, and R be three points in a dually flat manifold such that the m-geodesic connecting P and Q is orthogonal at Q to the e-geodesic connecting Q and R. Then,*

$$K(P, Q) + K(Q, R) = K(P, R). \tag{A.19}$$

When S is self-dual and flat, it reduces to the Euclidean space. The divergence $K(P, Q)$ is a half of the squared Euclidean distance in this case, so that this theorem reduces to the ordinary Pythagoras theorem. Two important corollaries follow.

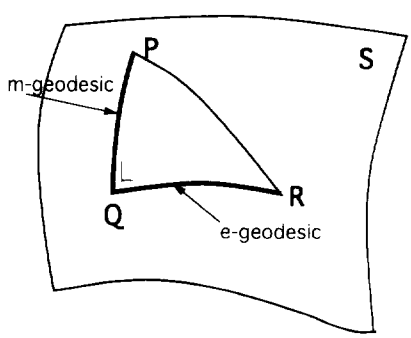

FIGURE A.2.. Generalized Pythagoras theorem.

Let M be a submanifold in S and let P be a point in S. We search for the point $\hat{Q}_P$ in M that is closest to P in the sense of the divergence (Figure A.3) When S is Euclidean, $\hat{Q}_P$ is given by the orthogonal projection of P to M. In a dually flat manifold, $K(P, Q) \neq K(Q, P)$, so that we have two criteria of closeness and two solutions to this problem.

To solve the problem, we explain the concept of the e- and m-projections. When the e (m-)-geodesic connection P and $Q \in M$ is orthogonal at Q to M, the point Q is said to be the e (m-)-projection of P to M. The generalized Pythagoras theorem shows that the e (m-)-projection $\hat{Q}_P(\hat{Q}^*_P)$ of P to Q gives the extremal point of $K(Q, P)$ ($K(P, Q)$).

COROLLARY 1. *The point $\hat{Q}_P \in M$ that minimizes $K(P, Q)$, $Q \in M$, is given by the m-projection of P to M. The m-projection is unique when M is an e-flat submanifold.*

COROLLARY 2. *The point $\hat{Q}^*_P \in D$ that minimizes $K(Q, P)$, $Q \in D$, is given by the e-projection of P to D. The e-projection is unique when D is an m-flat manifold.*

The e- and m-projections are given in the coordinate forms as follows. Let D be an m-flat submanifold. By taking an adequate $\boldsymbol{\eta}$-coordinate system, we can divide $\boldsymbol{\eta} = (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h)$ such that D is defined by

$$D = \{\boldsymbol{\eta} | \boldsymbol{\eta}_v = \mathbf{c};\ \boldsymbol{\eta}_h \text{ is arbitrary}\}.$$

Correspondingly, $\boldsymbol{\theta}$ is divided into $\boldsymbol{\theta} = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_h)$. Let $P \in S$ be a point with coordinates $\boldsymbol{\theta}^P = (\boldsymbol{\theta}^P_v, \boldsymbol{\theta}^P_h)$ and let Q^* be the e-projection of P to D. Let $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_v, \boldsymbol{\theta}^*_h)$ and $\boldsymbol{\eta}^* = (\boldsymbol{\eta}^*_v, \boldsymbol{\eta}^*_h)$ be the $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$-coordinates of Q^*, respectively. Because $Q \in D$, we have $\boldsymbol{\eta}^*_v = \mathbf{c}$. On the other hand, Theorem 2 shows that $\boldsymbol{\theta}_h$ part is invariant under the e-projection. Therefore, we have $\boldsymbol{\theta}^*_h = \boldsymbol{\theta}^P_h$. To obtain $\boldsymbol{\eta}^*$ or $\boldsymbol{\theta}^*$, we need to solve the following equations

$$\begin{cases} \boldsymbol{\eta}^*_h = \dfrac{\partial}{\partial \boldsymbol{\theta}_h} \psi(\boldsymbol{\theta}^*_v, \boldsymbol{\theta}^P_h) \\ \boldsymbol{\eta}^*_v = \mathbf{c} \end{cases} \tag{A.20}$$

or

$$\begin{cases} \boldsymbol{\theta}^*_v = \dfrac{\partial}{\partial \boldsymbol{\eta}_v} \varphi(\boldsymbol{\eta}^*_v, \mathbf{c}), \\ \boldsymbol{\theta}^*_h = \boldsymbol{\theta}^P_h. \end{cases} \tag{A.21}$$

We have a similar dual formulation for the m-projection when M is e-flat.

A.4. Dual Differential Geometry

We briefly describe the dual differential geometry of probability distributions for readers familiar with differential geometry. In a

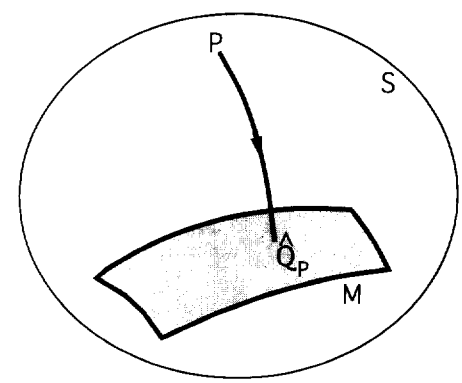

FIGURE A.3. m-projection of P to M.

manifold S, we have already defined the Riemannian metric by eqn (A.5). An affine connection is defined by a covariant derivative $\nabla_X Y$ for two vector fields X and Y. This is determined if we give the covariant derivative $\nabla_{\mathbf{e}_i}\mathbf{e}_j$ for the vector fields $\mathbf{e}_i$ and $\mathbf{e}_j$ of the natural basis connected with a coordinate system. Because $\nabla_{\mathbf{e}_i}\mathbf{e}_j$ is a vector field, it is determined in the component form by

$$\Gamma_{ijk}(P) = \langle \nabla_{\mathbf{e}_i}\mathbf{e}_j, \mathbf{e}_k \rangle. \tag{A.22}$$

This three-index quantity Γ_{ijk} is called the coefficients of the underlying affine connection.

In a Riemannian manifold S, we have the Levi–Civita or Riemannian connection defined by

$$\Gamma_{ijk} = \frac{1}{2}\left(\frac{\partial}{\partial\theta^i} g_{jk} + \frac{\partial}{\partial\theta^j} g_{ik} - \frac{\partial}{\partial\theta^k} g_{ij}\right). \tag{A.23}$$

This is the only torsion-free affine connection preserving the Riemannian metric. When an affine connection is defined, the geodesic curve $\theta = \theta(t)$ is given by

$$\nabla_{\dot{\boldsymbol{\theta}}}\dot{\boldsymbol{\theta}} = 0 \tag{A.24}$$

or

$$\sum g_{ik}\ddot{\theta}^k + \sum \Gamma_{jki}\dot{\theta}^j\dot{\theta}^k = 0, \tag{A.25}$$

where $\dot{}$ denotes d/dt. A geodesic is a minimum length curve connecting two points when the Riemannian connection is used. Let $c : \boldsymbol{\theta} = \boldsymbol{\theta}(t)$ be a curve connecting two points P and Q. When tangent vector field $A(t)$ along the curve satisfies

$$\nabla_{\dot{\boldsymbol{\theta}}} A = 0, \tag{A.26}$$

a vector $A(P)$ in T_P is said to be transported to $A(Q) \in T_Q$ in parallel along the curve by the affine connection. We write the parallel transport as

$$A(Q) = \Pi_C A(P). \tag{A.27}$$

The conservation of metric implies that

$$\langle A(P), A(P)\rangle_P = \langle \Pi_C A(P), \Pi_C A(P)\rangle_Q. \tag{A.28}$$

The dual geometry defines new torsion-free affine connections different from the Riemannian one. In a manifold S of probability distributions, we can define an invariant tensor

$$T_{ijk} = E\left[\frac{\partial}{\partial\theta^i}\log p(x,\boldsymbol{\theta})\frac{\partial}{\partial\theta^j}\log p(x,\boldsymbol{\theta})\frac{\partial}{\partial\theta^k}\log p(x,\boldsymbol{\theta})\right]. \tag{A.29}$$

The α-connection is defined by

$$\Gamma^{(\alpha)}_{ijk} = \Gamma_{ijk} - \frac{\alpha}{2}T_{ijk}. \tag{A.30}$$

The $\alpha = 0$-connection is the Riemannian connection. The $\alpha = 1$ connection is called the exponential connection and the $\alpha = -1$ connection is called the mixture connection. They are dual in the sense that they together preserve the metric,

$$\langle A(P), A(P)\rangle_P = \left\langle \Pi^{(1)}_C A(P), \Pi^{(-1)}_C A(P)\right\rangle_Q, \tag{A.31}$$

where $\Pi^{(1)}_c$ and $\Pi^{(-1)}_c$ are the parallel transports with respect to $\alpha = 1$ and $\alpha = -1$ connections, respectively.

An exponential family is special in the sense that they are dually flat, that is, the Riemann–Christoffel curvatures vanish for $\alpha = \pm 1$. In such a case, we have an e-affine coordinate system $\boldsymbol{\theta}$ for which $\Gamma^{(1)}_{ijk}$ vanishes. In this case, the geodesic equation reduces to

$$\ddot{\boldsymbol{\theta}} = 0$$

or

$$\boldsymbol{\theta}(t) = \mathbf{a} + \mathbf{b}t.$$

This is an e-geodesic. Dually to the above we have another m-affine coordinate system $\boldsymbol{\eta}$ for which $\Gamma^{(-1)}_{ijk}$ vanishes. The m-geodesic is then written as

$$\boldsymbol{\eta}(t) = \mathbf{a} + \mathbf{b}t$$

in this coordinate system.

In a dually flat manifold, the following theorem holds.

THEOREM A3. *When S is dually flat, there exist two affine-coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ and two potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$ such that the metric is given by*

$$g_{ij}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\psi(\boldsymbol{\theta}), \tag{A.32}$$

$$g^*_{ij}(\boldsymbol{\eta}) = \frac{\partial^2}{\partial\eta_i\partial\eta_j}\varphi(\boldsymbol{\eta}). \tag{A.33}$$

The two coordinate systems are connected by the Legendre transformation

$$\eta_i = \frac{\partial}{\partial\theta^i}\psi(\boldsymbol{\theta}),$$
$$\theta_i = \frac{\partial}{\partial\eta_i}\varphi(\boldsymbol{\eta}),$$
$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \sum \theta_i\eta_i = 0.$$

The natural bases $\mathbf{e}_i = \partial/\partial\theta^i$ and $\mathbf{e}^*_i = \partial/\partial\eta^i$ are dual,

$$\langle \mathbf{e}_i, \mathbf{e}^*_j\rangle = \delta_{ij}.$$

As we stated before, an invariant divergence is defined by

$$K(P, Q) = \psi(P) + \varphi(Q) - \sum \theta^P_i\eta^Q_i,$$

and the generalized Pythagoras theorem holds.

Information geometry studies new geometrical structures existing in manifolds of probability distributions. It is generalized to the fibre bundle structure (Amari & Kawanabe, 1994) and the conformal structure (Okamoto, Amari, & Takeuchi, 1988). It has been applied successfully to various fields of information sciences as is mentioned earlier. It is also related to completely integrable dynamical systems (Fujiwara & Amari, 1995).