Information Geometry of the EM and em Algorithms for Neural Networks

Shun-ichi Amari

Department of Mathematical Engineering and Information Physics

Faculty of Engineering, University of Tokyo

Bunkyo-ku, Tokyo 113, JAPAN

The present work is 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.

Requests for reprints should be sent to the author at the Department of Mathematical Engineering, the University of Tokyo, Bunkyo-ku, Hongo, Tokyo 113, Japan : fax +81-3-5689-5752.

running title: Geometry of EM Algorithm

Information Geometry of the EM and em Algorithms for Neural Networks

Shun-ichi Amari

Abstract

In order to realize an input-output relation given by noise-contaminated examples, it is effective to use a stochastic model of neural networks. A 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 forcussing on the EM and em algorithms, and proves a condition which guarantees their equivalence. Examples include 1) Boltzmann machines with hidden units, 2) mixtures of experts, 3) stochastic multilayer perceptron, 4) normal mixture model, 5) hidden Markov model, among others.

key words: 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 $p(\mathbf{y}|\mathbf{x})$ of the output $\mathbf{y}$ when $\mathbf{x}$ 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. 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 neural networks (see, e.g., Amari, 1990; Cheng and Titterington, 1994; Ripley, 1994; White, 1989).

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 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$. 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 $p(\mathbf{x};\boldsymbol{\theta})$ or a conditional probability distribution $p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$. 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 subject with new differential geometrical notions. It has been successfully applied to various information sciences such as statistical sciences (Amari [1982], [1985], Barndorff-Nielsen [1988], Barndorff-Nielsen, Cox and Reid [1986], Murray and Rice [1993], Kass [1989], Amari et al. [1987], etc.), information theory (Amari and Han [1989], Amari

[1989]), systems theory (Amari [1987a], Ohara and Amari [1992]) and many others. Applications to neural networks have already started in Amari [1991], 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 the credit assignment problem.

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 which are realizable by neural networks form a submanifold M of neural networks in S. On the other hand, observed data suggests some distribution (e.g., one given by the empirical distribution) in S. However, since observation or specification is incomplete, the observed or specified partial data defines many candidates of points in the manifold S which 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] and Neal and Hinton [1993]. This may be called the *em*-algorithm, because it is realized geometrically by the e-geodesic and m-geodesic

projections 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 is in D. The EM-algorithm has been applied to the Boltzmann machines (Byrne, [1992]), the hierarchical mixture of experts (Jordan and Jacobs [1994], see also Jacobs et al. [1991] for mixtures of experts) and others. A new formulation and acceleration method is proposed in this connection (Neal and Hinton [1993]). See also Jordan and Xu [1994], Xu, Jordan [1994].

The present paper elucidates the relation between the statistical EM-algorithm and the geometrical em-algorithm reported in short notes (Amari 1994). They are the same in most cases, and we prove a necessasry and sufficient condition which 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 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. Boltzmann machines with hidden units, where the hidden units improve the conditional probability distribution of output units conditioned on inputs.

2. 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 also Jordan and Xu [1994], Xu, Jordan and Hinton [1994].

3. Stochastic multilayer perceptron. Our theory gives a new learning algorithm different from the backpropagation method. The new algorithm seems more flexible with a better global convergence property than the backpropagation (an independent study on the stochastic perceptron by Rumelhart, personal communication).

4. 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.

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 aspects suggests various new algorithms including learning. The present paper does not study the computational aspect of the algorithms nor shows computer simulated results. These remain to be studied separately in future by applying the framework proposed in the present paper. Applications to hidden Markov random fields (Geman and Geman, 1984; Besag and Green, 1993; Günsch, Geman and Kehagias, 1994), dynamics of Boltzmann machines with asymmetric connections, etc. are also important subjects of future research.

The present paper is organized as follows. Sections 2 and 3 are devoted to introduction to the exponential family and curved exponential family, which are basic statistical models. Various important neural networks such as the Boltzmann machine, multilayer perceptron, mixture of expert networks are shown to be described in this statistical framework. Section 4 is a statistical preliminary where the maxmimum likelihood estimation is explained in terms of geometry. Readers familiar with neural networks and

statistics may skip these sections. Sectikon 5 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 6. They are analyzed in section 7, 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 8. Section 9 treats learning procedures and Section 10 studies a dual geodesic gradient flow in M and D. Appendix gives a short intuitive introduction to information geometry.

2 Exponential Families and Neural Networks

2.1 Exponential families

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 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$. When the probability density functions are written in the following form

$$p(x; \boldsymbol{\theta}) = \exp\{\sum \theta_i r_i(x) + k(x) - \psi(\boldsymbol{\theta})\}, \tag{2.1}$$

where $r_i(x)$, $i = 1, \cdots, n$, are functions of x, the family $S = \{p(x; \boldsymbol{\theta})\}$ is called an exponential family (Cox and Hinkley, 1974; Barndorff-Nielsen, 1976). We may treat $\mathbf{r} = (r_1, \cdots, r_n) \in \mathbf{R}^n$ as a new vector random variable, whose distribution is specified by

$$p(\mathbf{r}; \boldsymbol{\theta}) = \exp\{\sum \theta_i r_i - k(\mathbf{r}) - \psi(\boldsymbol{\theta})\} \tag{2.2}$$

with respect to a suitable measure $\mu(\mathbf{r})$ on $\mathbf{R}^n$. (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, and any point (that is, any distribution) in S is specified by one $\boldsymbol{\theta}$ (Amari, 1985). The $\boldsymbol{\theta}$ is called the natural or canonical parameter of the exponential family.

2.2 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

$$\begin{aligned} r_1 &= x, & r_2 &= x^2, \\ \theta_1 &= \frac{\mu}{\sigma^2}, & \theta_2 &= -\frac{1}{2\sigma^2}, \end{aligned}$$

the distribution is rewritten as

$$p(\mathbf{r}; \boldsymbol{\theta}) = \exp\{\theta_1 r_1 + \theta_2 r_2 + k(\mathbf{r}) - \psi(\boldsymbol{\theta})\}, \tag{2.3}$$

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

$$\begin{aligned} k(\mathbf{r}) &= 0, \\ \psi(\boldsymbol{\theta}) &= \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma) = -\frac{1}{4}(\theta_1)^2 \theta_2^{-1} - \frac{1}{2}\log(-\theta_2) + \frac{1}{2}\log\pi. \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} p(\mathbf{r}; \boldsymbol{\theta}) &= \exp\{\sum \theta_i r_i - \psi(\boldsymbol{\theta})\}, \\ \psi(\boldsymbol{\theta}) &= -\log p_0 = \log\{1 + \sum \exp(\theta_i)\}. \end{aligned} \tag{2.4}$$

Example 3. Stable distributions of Boltzmann machines. A Boltzmann machine is a stochastic neural network consisting of n neurons. The neurons are connected with symmetric connection weights $W = (w_{ij})$ (Ackley, Hinton and Sejnowski, 1985; Aarts and Korst, 1989). The self-connection is zero, $w_{ii} = 0$. Instead, we put $w_{ii} = s_i$ for the sake of notational convenience, where s_i is the bias term of the ith unit. Let $\mathbf{x} = (x_1, \cdots, x_n)$ be the state of a Boltzmann machine, where x_i takes on the binary values 0 and 1. The state transition of a Boltzmann machine is stochastic, forming a Markov chain. The stable distribution of the state $\mathbf{x}$ in the Boltzmann machine with connection matrix W is written as

$$p(\mathbf{x}; W) = \exp\{\frac{1}{2} \sum w_{ij} x_i x_j - \psi(W)\}. \tag{2.5}$$

Let B be the set of all the Boltzmann machines. Then, W plays the role of a coordinate system of B. There is a one-to-one correspondence between the Boltzmann machines and the probability distributions $p(\mathbf{x}; W)$ of the form (2.5), so that B is identified with the set of probability distributions $p(\mathbf{x}; W)$ of the form (2.5).

This set is an exponential family,

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

where $\boldsymbol{\theta} = (\theta_{ij})$, $\mathbf{r} = (r_{ij})$, $\boldsymbol{\theta} \cdot \mathbf{r} = \sum \theta_{ij} r_{ij}$,

$$r_{ij} = x_i x_j, \qquad i \le j,$$

$$\theta_{ij} = w_{ij}, \qquad i \le j,$$

and $\exp\{-\psi(W)\}$ is the normalization constant, $\psi(W)$ corresponding to the free energy in statistical physics. This ψ is the cummulant generating function in statistics.

Example 4. Normal mixture with hidden variables. Let us consider $k+1$ normal distributions subject to $N(\mu_i, \sigma_i^2)$, $i = 0, 1, \cdots, k$. Let z be a discrete random variable taking values on $\{0, 1, \cdots, k\}$ with probabilities $p_i = \text{Prob}\{z = i\}$, $i = 0, \cdots, k$. 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

$$p(x, z) = \sum_{i=0}^{k} \delta_i(z) p_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\}. \tag{2.6}$$

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

$$p(x) = \sum_i \frac{p_i}{\sqrt{2\pi}\sigma_i} \exp\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\}, \tag{2.7}$$

which is called the normal mixture distribution. This form can be regarded as a radial basis expansion of $p(x)$. The logarithm of the probability (2.6) is written, by eliminating $\delta_0(z)$ from $\sum_{i=0}^{k} \delta_i(z) = 1$, as

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

Hence, by putting

$$
\begin{aligned}
r_{11} &= x, & \theta_{11} &= \frac{\mu_0}{\sigma_0^2}, \\
r_{12} &= x^2, & \theta_{12} &= -\frac{1}{2\sigma_0^2}, \\
r_{2i} &= \delta_i(z), & \theta_{2i} &= \log\frac{p_i\sigma_0}{p_0\sigma_i} - \left(\frac{\mu_i^2}{2\sigma_i^2} - \frac{\mu_0^2}{2\sigma_0^2}\right), \\
r_{3i} &= x\delta_i(z), & \theta_{3i} &= \frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2}, \\
r_{4i} &= x^2\delta_i(z), & \theta_{4i} &= -\left(\frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_0^2}\right),
\end{aligned}
$$

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

$$p(x, z; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta} \cdot \mathbf{r} - \psi(\boldsymbol{\theta})\}, \tag{2.8}$$

where $\boldsymbol{\theta}\cdot\mathbf{r} = \sum \theta_{ij} r_{ij}$. The set of the normal mixtures (2.7) without z is not an exponential family.

Example 5. Markov chain. Let us consider an ergodic Markov chain over a finite state set $A = \{s_0, s_1, \cdots, s_k\}$. Let $\mathbf{y} = y_0 y_1 \cdots y_T$ be a sequence of states, where we assume that the initial state y_0 is in s_0 for the sake of simplicity.

The state transition matrix $P = (p_{ji})$ is

$$p_{ji} = \mathrm{Prob}\{y_{t+1} = s_j | y_t = s_i\}. \tag{2.9}$$

The probability of a random sequence $\mathbf{y}$ is given by

$$p(\mathbf{y}; P) = p_{y_0} \prod_{t=1}^{T} p_{y_t y_{t-1}} \tag{2.10}$$

and hence its logarithm is

$$l(\mathbf{y}; P) = \log p(y; P) = \sum_{t=1}^{T} \log p_{y_t y_{t-1}}. \tag{2.11}$$

Let n_{ji} be the numbers of the state transition from s_i to s_j in $\mathbf{y}$,

$$n_{ji} = \sharp\{y_t = s_i,\ y_{t+1} = s_j\}. \tag{2.12}$$

They are random variables depending on $\mathbf{y}$, and (2.11) is rewritten as

$$l(\mathbf{y}; P) = \sum_{i,j} n_{ji} \log p_{ji}.$$

Here, n_{ij}'s are not independent, because we have the constraints

$$n_i = \sum_{j=0}^{k} n_{ij} = \sum_{j=0}^{k} n_{ji}, \tag{2.13}$$

where n_i is the number that the state is at s_i in the process $\mathbf{y}$,

$$n_i = \sharp\{y_t = s_i, \quad t = 0, \cdots, T\}.$$

More precisely, (2.12) holds exactly if $y_T = s_0$. Otherwise, (2.12) holds within an error of ± 1, depending on the last state s_T. However, this does not matter when we treat asymptotic properties of the Markov chain where T is very large. (The exact theory can be constructed in the framework of the curved exponential family.) The variables n_{ij} $(i, j = 0, \cdots, k)$ can be rewritten in terms of n_i and n_{ij} $(i, j = 1, \cdots, k)$ by eliminating n_0, n_{0i}, n_{0j} according to

$$\begin{aligned} n_0 &= T - \sum_{i=1}^{k} n_i, \\ n_{0i} &= n_i - \sum_{j=1}^{k} n_{ij}, \\ n_{j0} &= n_j - \sum_{i=1}^{k} n_{ji}, \\ n_{00} &= n_0 - \sum_{j=1}^{k} n_{0j}. \end{aligned}$$

Then, the log probability is written as

$$l(\mathbf{y}; P) = \sum_{i=1}^{k} \theta_i n_i + \sum_{i,j=1}^{k} \theta_{ji} n_{ji} + T \log p_{00}, \tag{2.14}$$

where

$$\begin{aligned} \theta_{0i} &= \log p_{0i} + \log_{i0} - \log p_{00} \\ \theta_{ji} &= \log p_{ji} - \log p_{0i} - \log p_{j0} + \log p_{00}. \end{aligned} \tag{2.15}$$

Hence, by putting $\mathbf{r} = (n_i, n_{ji})$, $\boldsymbol{\theta} = (\theta_i, \theta_{ji})$ and

$$\psi(\boldsymbol{\theta}) = -T \log p_{00}, \tag{2.16}$$

the set of all the Markov chains on A is represented by an exponential family,

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

The set of random Markov fields is also represented in an exponential form.

2.3 Mixtures of expert neural nets

Example 6. Mixture of expert nets. A simplest example of mixtures of expert nets (Jacobs et al., 1991; Jordan and Jacobs, 1994) is presented here. There are various generalizations, including the hierarchical mixture (Jordan and Jacobs, 1994). Some generalization will be shown after this example. Let N_i $(i = 0, 1, \cdots, k)$ 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 probability of y_i given $\mathbf{x}$ is written as

$$\begin{aligned} p(y_i|\mathbf{x}, \mathbf{w}_i) &= \varphi(y_i, \mathbf{w}_i \cdot \mathbf{x}) \\ &= \exp\{y_i \mathbf{x} \cdot \mathbf{w}_i - \psi(\mathbf{w}_i \cdot \mathbf{x})\}, \end{aligned} \tag{2.17}$$

where φ is the sigmoidal function

$$\varphi(y, u) = \frac{\exp(yu)}{1 + \exp(u)} \tag{2.18}$$

and

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

The family (2.17) of conditional distributions is an exponential family where $r = y_i$ and $\theta = \mathbf{w}_i \cdot \mathbf{x}$ for a given $\mathbf{x}$.

The mixture of experts is composed of $k+1$ expert networks N_i with a gating network N which selects one of the expert nets for processing input $\mathbf{x}$. It receives the same input $\mathbf{x}$ and its output z is a random variable taking values on $\{0, 1, \cdots, k\}$. 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 (Fig.2). 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

$$g_i(\mathbf{x}) = p(z = i|\mathbf{x}) = \frac{\exp(\mathbf{v}_i \cdot \mathbf{x})}{\sum_{j=0}^{k} \exp\{\mathbf{v}_j \cdot \mathbf{x}\}}, \tag{2.19}$$

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 $j = 0, \cdots, k$.

The joint probability of (y, z) is

$$\begin{aligned} p(y, z|\mathbf{x}) &= \exp\{\sum_{i=1}^{k} \delta_i(z)(\mathbf{v}_i \cdot \mathbf{x} - \psi_i + \psi_0) \\ &\quad + \sum_{i=1}^{k} y\delta_i(z)(\mathbf{w}_i - \mathbf{w}_0) \cdot \mathbf{x} \\ &\quad + y\mathbf{w}_0 \cdot \mathbf{x} - \psi\}, \end{aligned} \tag{2.20}$$

where

$$\psi_i = \log\{1 + \exp(\mathbf{w}_i \cdot \mathbf{x})\}, \tag{2.21}$$

$$\psi = \log\{1 + \sum_{j=1}^{k} \exp(\mathbf{v}_j \cdot \mathbf{x})\} + \psi_0. \tag{2.22}$$

By putting

$$\begin{aligned} r_1 &= y, \\ r_{2i} &= \delta_i(z) y, \end{aligned} \tag{2.23}$$

$$r_{3i} = \delta_i(z),$$

$$\begin{aligned} \theta_1 &= \mathbf{w}_0 \cdot \mathbf{x}, \\ \theta_{2i} &= (\mathbf{w}_i - \mathbf{w}_0) \cdot \mathbf{x}, \\ \theta_{3i} &= \mathbf{v}_i \cdot \mathbf{x} - \psi_i + \psi_0, \end{aligned} \tag{2.24}$$

we have

$$p(y, z|\mathbf{x}) = \exp\{\boldsymbol{\theta} \cdot \mathbf{r} - \psi\}. \tag{2.25}$$

Therefore, the family $\{p(y, z|\mathbf{x})\}$ of the conditional distributions of the mixtures of experts is an exponential family, where $\mathbf{x}$ is fixed. We consider the case where various $\mathbf{x}$ are input in the next subsection.

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,

$$p(y_i|\mathbf{x}, \mathbf{w}_i) = \exp\{\theta(\mathbf{w}_i \cdot \mathbf{x})y_i + k - \psi(\theta)\}. \tag{2.26}$$

The binary neuron case (2.17) is given by putting

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

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

$$y_i = f(\mathbf{w}_i \cdot \mathbf{x}) + n, \tag{2.27}$$

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

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

The joint probability of the mixture of general N_i of (2.26) is again written as

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

where $\mathbf{r}$ is the same as (2.23) and

$$\begin{aligned} \theta_1 &= \theta(\mathbf{w}_0 \cdot \mathbf{x}), \\ \theta_{2i} &= \theta(\mathbf{w}_i \cdot \mathbf{x}) - \theta(\mathbf{w}_0 \cdot \mathbf{x}), \\ \theta_{3i} &= \mathbf{v}_i \cdot \mathbf{x} - \psi\{\theta(\mathbf{w}_i \cdot \mathbf{x})\} + \psi\{\theta(\mathbf{w}_0 \cdot \mathbf{x})\}. \end{aligned} \tag{2.28}$$

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

The mixture of experts is a stochastic model. It can be trained from examples based on stochastic techniques. However, it can be used deterministically in the excution 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 $z = i$.

Example 7. Mixture with hierarchical and general gating net. 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^*, $i = 0, \cdots, m$, be $(m+1)$ mixtures of experts nets, consisting of $k_i + 1$ expert nets $N_{i,j}$ $(j = 0, \cdots, k_i)$. 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

$$y = \sum_{i,j} \delta_i(z^*)\delta_j(z_i)y_{i,j}. \tag{2.29}$$

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, Jordan and Hinton [1994] proposed a different gating mechanism which is computationally tractable.

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

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

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

$$y = \sum_{\mu=0}^{K} \delta_\mu^*(z)y_\mu, \tag{2.30}$$

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 (2.20), by

$$\begin{aligned} p(y, z|\mathbf{x}) &= \exp\{\sum_{i=1}^{m} \delta_i(z)(\mathbf{v}_i \cdot \mathbf{x} - \psi_i + \psi_0) + \sum_{\mu=1}^{k} y\delta_\mu(z)(\mathbf{w}_\mu - \mathbf{w}_0) \cdot \mathbf{x} \\ &\quad + y\mathbf{w}_0 \cdot \mathbf{x} - \psi\}. \end{aligned}$$

The statistics $\mathbf{r}$ is

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

in this case and $\boldsymbol{\theta}$ is similar to (2.24) where θ_{2i} is replaced by $\theta_{2\mu}$.

The above gating 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

$$X = \bigcup X_i. \tag{2.31}$$

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 and Amari [1993]), which is a little more general than the Voronoi division. The hierarchical mixture divides each region further into Laguerre-Voronoi subregions

$$X_i = \bigcup X_{i,j}, \tag{2.32}$$

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, while the region X_μ^* on which signals are processed by N_μ, consists of unions of the corresponding subregions

$$X_\mu^* = \bigcup_{j \in I_\mu} X_j. \tag{2.33}$$

Each X_μ^* can be non-convex 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.

2.4 Product space

The conditional probability $p(y, z|\mathbf{x})$ depends on the input signal $\mathbf{x}$ in the mixture of expert nets (2.20). The set of such conditional distributions forms an exponential family,

but the canonical parameter $\boldsymbol{\theta}$ depends on $\mathbf{x}$ as is shown in (2.24) or

$$p\{\mathbf{r}|\boldsymbol{\theta}(\mathbf{x})\} = \exp\{\boldsymbol{\theta}(\mathbf{x}) \cdot \mathbf{r} - \psi\}, \tag{2.34}$$

where $\mathbf{r} = \mathbf{r}(y, z)$. A neural network is trained by giving various input signals $\mathbf{x}$ and corresponding output behaviors, where $\mathbf{x}$ is not fixed. When T independent data $(\mathbf{x}_1, y_1, z_1), \cdots, (\mathbf{x}_T, y_T, z_T)$ are given, the joint conditional distribution is written as

$$\begin{aligned} &p\{(y_1, z_1), \cdots, (y_T, z_T)|\mathbf{x}_1, \cdots, \mathbf{x}_T\} \\ &\quad = \prod_{t=1}^{T} p\{\mathbf{r}_t|\boldsymbol{\theta}(\mathbf{x}_t)\} = \exp\{\sum \boldsymbol{\theta}(\mathbf{x}_t) \cdot \mathbf{r}_t - \sum \psi_t\}. \end{aligned} \tag{2.35}$$

This again forms an exponential family of larger dimensions. Let

$$S_t = \{p(\mathbf{r}_t|\boldsymbol{\theta}(\mathbf{x}_t))\} \tag{2.36}$$

be the manifold of conditional probability distributions corresponding to the tth input $\mathbf{x}_t$. Then, the joint conditional distributions (2.35) are given by the product space

$$S_T^* = S_1 \times S_2 \times \cdots \times S_T. \tag{2.37}$$

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

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

the canonical parameter in S_T^* is

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

and (2.35) is rewritten as

$$p(\mathbf{r}_T^*|\boldsymbol{\theta}_T^*) = \exp\{\boldsymbol{\theta}_T^* \cdot \mathbf{r}_T^* - \psi\}, \tag{2.38}$$

where

$$\boldsymbol{\theta}_T^* \cdot \mathbf{r}_T^* = \sum_{t=1}^{T} \boldsymbol{\theta}_t \cdot \mathbf{r}_t.$$

However, $\boldsymbol{\theta}(\mathbf{x}_1), \cdots, \boldsymbol{\theta}(\mathbf{x}_T)$ cannot take arbitrary values, because all of them are determined not only by $\mathbf{x}_t$ but also restricted by the common network parameters $\mathbf{w}_i$'s and $\mathbf{v}_i$ in the form of (2.28) in the case of the mixture of experts. Hence, possible $\boldsymbol{\theta}_T^*$ is restricted to a subregion of S_T^*. In other words, the distributions realized by neural networks do not cover the entire exponential family S_T^* but are limited to some subregion (submanifold) of S_T^*. This is a curved exponential family explained in the next section.

3 Curved Exponential Families

3.1 Curved exponential families

A curved submanifold M of an exponential family S is called a curved exponential family. Let $S = \{p(\mathbf{r}; \boldsymbol{\theta})\}$ be an n-dimensional exponential family and let M be its m-dimensional submanifold. Let $\mathbf{u} = (u_1, \cdots, u_m)$, $m \leq n$, be a parameter or a coordinate system of M. A curved exponential family $M = \{p(\mathbf{r}; \mathbf{u})\}$ consists of the distributions

$$p(\mathbf{r}; \mathbf{u}) = \exp\{\boldsymbol{\theta}(\mathbf{u}) \cdot \mathbf{r} - k(\mathbf{r}) - \psi(\boldsymbol{\theta}(\mathbf{u}))\}, \tag{3.1}$$

where $\boldsymbol{\theta}(\mathbf{u})$ is a function of $\mathbf{u}$. So the distributions are parameterized by $\mathbf{u}$. Any distribution $p(\mathbf{r}; \mathbf{u})$ 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 (Fig. 3) composed of points written as

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{u}). \tag{3.2}$$

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

3.2 Simple examples

Example 8. 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 in order 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 which 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

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

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

$$\theta_1 = \frac{1}{u}, \qquad \theta_2 = -\frac{1}{2u^2}. \tag{3.3}$$

The shape of M is a parabolla in S.

Example 9. Multiple observations. Let us consider the multiple observation in § 2.4,

$$p(\mathbf{r}_T^* | \boldsymbol{\theta}_T^*) = \exp\{\boldsymbol{\theta}_T^* \cdot \mathbf{r}_T^* - \psi\}. \tag{3.4}$$

Here, $\boldsymbol{\theta}_T^* = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_T)$ 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 $\mathbf{v}_i$'s in the example of the mixture of expert nets,

$$\mathbf{u} = (\mathbf{w}_0, \mathbf{w}_1, \cdots, \mathbf{w}_k; \mathbf{v}_1, \cdots, \mathbf{v}_k).$$

Then, the distribution realized by the neural net of parameter $\mathbf{u}$ has the $\boldsymbol{\theta}_T^*$-coordinates given by

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u}), \quad t = 1, \cdots, T \tag{3.5}$$

in the form of (2.28). Hence the set M of such distributions 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 is the number of the network parameters $\mathbf{u}$ and is fixed.

3.3 Stable distributions of Boltzmann machines

Example 9. Boltzmann machine and general probability distributions. In order to study the behaviors of Boltzmann machines with hidden units, we need to introduce a larger manifold S consisting of all the probability distributions on the state space $X = \{\mathbf{x}\}$ (Amari, Kurata and Nagaoka, 1992). This is because not all the distributions on X are realizable by Boltzmann machines as their stable distributions. Let $q(\mathbf{x}) > 0$ be a distribution on X. Since x_i takes on $\{0, 1\}$, we can expand $q(\mathbf{x})$ as

$$\log q(\mathbf{x}) = \sum_{k=1}^{n} \sum_{i_1 < \cdots < i_k} \theta^k_{i_1 \cdots i_k} x_{i_1} \cdots x_{i_k} - \psi(\theta) \tag{3.6}$$

or

$$q(\mathbf{x}) = \exp\{\sum \sum \theta^k_{i_1 \cdots i_k} x_{i_1} \cdots x_{i_k} - \psi\} = \exp\{\boldsymbol{\theta} \cdot \mathbf{r} - \psi\}.$$

This shows that the set S of all the distributions over X is an exponential family, where the $\boldsymbol{\theta}$-coordinates are

$$\boldsymbol{\theta} = (\theta^1_i, \theta^2_{ij}, \cdots, \theta^n_{1 \cdots n})$$

and the corresonding random variables $\mathbf{r} = (r^1_i, r^2_{ij}, \cdots, r^n_{1 \cdots n})$ are given by

$$r^k_{i_1 \cdots i_k} = x_{i_1} \cdots x_{ik}, \qquad i_1 < i_2 < \cdots < i_k.$$

The distributions realized by Boltzmann machines are only a part of S, since the stable distributions of the Boltzmann machines are of the form

$$p(\mathbf{x}; W) = \exp\{\sum_{i \leq j} w_{ij} x_i x_j - \psi(W)\}.$$

By putting

$$\begin{aligned} \theta_i^1 &= \quad w_{ii} \quad = s_i \\ \theta_{ij}^2 &= \quad w_{ij}, \quad i < j, \end{aligned}$$

we see that the set M of the probability distributions realizable by Boltzmann machines is a submanifold of S defined by

$$\begin{aligned} &\theta_i^1 = s_i, \qquad &&\theta_{ij}^2 = w_{ij}, \\ &\theta_{i_1 \cdots i_k}^k = 0, \qquad &&k \geq 3. \end{aligned} \tag{3.7}$$

Here, the inner coordinates of M are $\mathbf{u} = (s_i, w_{ij})$, $i < j$. This is a linear submanifold of S, and in this case M itself forms an exponential family. So M is not a curved but a flat submanifold of S.

3.4 Stochastic multilayer perceptron

Example 10. Stochastic multilayer perceptron (cf. Amari, 1991). Let us consider a one hidden-layer perceptron with a single output unit (Fig.4). Let $\mathbf{x}$ be an input vector and let $\mathbf{z} = (z_i)$, $i = 1, \cdots, m$, 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 $\mathbf{x}$ is input, is written by using (2.17) as

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

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

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

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 $\mathbf{z}$, by

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

where $\mathbf{v}$ is the synaptic weight vector of the output unit. Here, a constant input $z_0 \equiv 1$ is added in $\mathbf{z} = (z_0, z_1, \cdots, z_m)$ 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 $\mathbf{x}$, the conditional probability of variables $(y, \mathbf{z})$ is given by

$$p(y, \mathbf{z}|\ \mathbf{x}; \mathbf{u}) = p(y|\mathbf{z}; \mathbf{u}) \prod_{i=1}^{m} p(z_i|\mathbf{x}; \mathbf{u}), \tag{3.8}$$

where $\mathbf{u}$ denotes the parameter summarizing all of $\mathbf{w}_1, \cdots, \mathbf{w}_m$ and $\mathbf{v}$,

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

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

$$p(y, \mathbf{z}, \mathbf{x}; \mathbf{u}) = q(\mathbf{x}) p(y, \mathbf{z}|\ \mathbf{x}; \mathbf{u}), \tag{3.9}$$

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

$$l(y, \mathbf{z}|\ \mathbf{x}; \mathbf{u}) = \log p(y, \mathbf{z}|\ \mathbf{x}; \mathbf{u}), \tag{3.10}$$

$$l(y, \mathbf{z}, \mathbf{x}; \mathbf{u}) = \log p(y, \mathbf{z}|\ \mathbf{x}; \mathbf{u}) + \log q(\mathbf{x}) \tag{3.11}$$

we see that maximizing the logarithm of the total probability (3.11) with respect to $\mathbf{u}$ is the same as maximizing the logarithm of the conditional probability (3.10). We have

$$\begin{aligned} l(y, \mathbf{z}, \mathbf{x}; \mathbf{u}) &= y \mathbf{v} \cdot \mathbf{z} + \sum_i z_i \mathbf{w}_i \cdot \mathbf{x} - \log\{1 + \exp(\mathbf{v} \cdot \mathbf{z})\} \\ &\quad - \sum_i \{1 + \exp(\mathbf{w}_i \cdot \mathbf{x})\} + \log q(\mathbf{x}). \end{aligned}$$

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

$$\delta_{\mathbf{k}}(\mathbf{z}) = \begin{cases} 1, & \mathbf{z} = \mathbf{k}, \\ 0, & \mathbf{z} \neq \mathbf{k}, \end{cases} \tag{3.12}$$

we have the following relations

$$\begin{aligned}
&\mathbf{v}\cdot\mathbf{z} = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z})\mathbf{k}\cdot\mathbf{v}, \\
&\log\{1+\exp(\mathbf{v}\cdot\mathbf{z})\} = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z})\log\{1+\exp(\mathbf{k}\cdot\mathbf{v})\} \\
&\sum_i z_i \mathbf{w}_i\cdot\mathbf{x} = \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z})\mathbf{w}_{\mathbf{k}}\cdot\mathbf{x},
\end{aligned}$$

where

$$\mathbf{w}_{\mathbf{k}} = \sum_i k_i \mathbf{w}_i, \qquad \mathbf{k} = (k_i). \tag{3.13}$$

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 2^m values.

The log likelihood is rewritten as

$$\begin{aligned}
l(y,\mathbf{z}|\mathbf{x},\mathbf{u}) &= \sum_{\mathbf{k}} \mathbf{k}\cdot\mathbf{v} y \delta_{\mathbf{k}}(\mathbf{z}) \\
&\quad - \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z})\log\{1+\exp(\mathbf{k}\cdot\mathbf{v})\} + \sum_{\mathbf{k}} \delta_{\mathbf{k}}(\mathbf{z})\mathbf{w}_{\mathbf{k}}\cdot\mathbf{x} \\
&\quad - \sum_i \log\{1+\exp(\mathbf{w}_i\cdot\mathbf{x})\}.
\end{aligned} \tag{3.14}$$

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

$$\begin{aligned}
r_{1,\mathbf{k}} &= y\delta_{\mathbf{k}}(\mathbf{z}), \\
\mathbf{r}_{2,\mathbf{k}} &= \mathbf{x}\delta_{\mathbf{k}}(\mathbf{z})
\end{aligned} \tag{3.15}$$

and

$$\begin{aligned}
\theta_{1,\mathbf{k}} &= \mathbf{k}\cdot\mathbf{v}, \\
\boldsymbol{\theta}_{2,\mathbf{k}} &= \mathbf{w}_{\mathbf{k}} + \log\{1+\exp(\mathbf{k}\cdot\mathbf{v})\}\mathbf{e}_0,
\end{aligned} \tag{3.16}$$

where $\mathbf{e}_0 = [1,0,\cdots,0]$ corresponding to the bias term $x_0 = 1$. Here, both $\mathbf{r}_{2,\mathbf{k}}$ and $\boldsymbol{\theta}_{2,\mathbf{k}}$ are vector-valued. We then have, from (3.14),

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

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

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

Here, $\boldsymbol{\theta}$ is not entirely free but is a non-linear vector function of $\mathbf{u} = (\mathbf{w}_1, \cdots, \mathbf{w}_m, \mathbf{v})$, $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{u})$ being given by (3.16). Hence, M is a curved exponential family.

In the case of multiple observations $(y_1, z_1, \mathbf{x}_1), \cdots, (y_T, z_T, \mathbf{x}_T)$, the joint distributions form a curved exponential family M embedded in the product space S_T^* as before. It should be noted that $\boldsymbol{\theta}(\mathbf{x}_t)$ does not depend on $\mathbf{x}_t$ in this special case. It is then possible to condense the product space S_T^* to a single S, as will be shown in the next section. It should be remarked that when learning takes place in the stochastic model of perceptron while examples are generated by a noiseless (deterministic) network, all of $|\mathbf{w}_i|$ and $|\mathbf{v}|$ diverge to infinity. So we need to control carefully the magnitudes of these parameters.

It should be remarked that 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 i-th hidden unit is

$$z_i = f(\mathbf{w}_i \cdot \mathbf{x}) + n_i \tag{3.17}$$

and the final output is

$$y = f(\mathbf{v} \cdot \mathbf{z}) + n, \tag{3.18}$$

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

$$p(\mathbf{z}|\ \mathbf{x}; \mathbf{u}) = c \exp\left[-\frac{1}{2\sigma^2} \sum \{z_i - f(\mathbf{w}_i \cdot \mathbf{x})\}^2\right] \tag{3.19}$$

$$p(y|\ \mathbf{z}; \mathbf{u}) = c \exp\left[-\frac{1}{2\sigma^2} \{y - f(\mathbf{v} \cdot \mathbf{z})\}^2\right]. \tag{3.20}$$

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

$$\begin{aligned}\sigma^2 l(y,\mathbf{z}|\mathbf{x};\mathbf{u}) &= \sum z_i f(\mathbf{w}_i\cdot\mathbf{x}) + yf(\mathbf{v}\cdot\mathbf{z}) - \frac{1}{2}\{f(\mathbf{v}\cdot\mathbf{z})\}^2 \\ &\quad + k(y,\mathbf{z}) - \psi. \end{aligned} \tag{3.21}$$

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, although we do not mention about it (see, e.g., Amari, 1987b; Barndorff-Nielsen and Jupp, 1989; Kawanabe and Amari, 1994). 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.

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

$$l(\mathbf{u}) = \sum_{t=1}^{T} \frac{1}{2}\{y_t - f(\mathbf{x}_t;\mathbf{u})\}^2, \tag{3.22}$$

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

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

is the same as the deterministic one. On the other hand, the paremeter $\mathbf{u}$ 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 (3.22). 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.

Since the joint probability of y and $\mathbf{z}$ is given from (3.19) and (3.20) as

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

we have the conditional probability of y by integrating the above with respect to $\mathbf{z}$. By using $\mathbf{n} = (n_i)$ in (3.17) 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 as

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

where

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

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

$$-\log p(y|\mathbf{x};\mathbf{u}) = \frac{1}{2}(y-f)^2\{1 + \frac{1}{2}\sigma^2 f'^2|\mathbf{v}|^2\}. \tag{3.23}$$

Therefore, the stochastic perceptron is expected to minimize the loss

$$l_{\text{stoch}}(\mathbf{u}) = \sum_{t=1}^{T}\frac{1}{2}\{y_t - f(\mathbf{x}_t;\mathbf{u})\}^2\{1 + \frac{1}{2}\sigma^2 f_t'^2|\mathbf{v}|^2\}. \tag{3.24}$$

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 has automatically the effect of decreasing loss at the range where $f(\mathbf{x};\mathbf{u})$ is almost saturated.

Learning based on the stochastic perceptron is shown to give a good performance by a preliminary computer simulation.

4 Geometry of Observation and Estimation

4.1 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 no hidden variables exist, as a preliminary of the hidden variable case studied in §5. A distribution 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 $p(\mathbf{r};\boldsymbol{\theta})$,

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\mathbf{r}] = \int \mathbf{r} p(\mathbf{r};\boldsymbol{\theta}) d\mu(\mathbf{r}), \tag{4.1}$$

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

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

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

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\psi(\boldsymbol{\theta}). \tag{4.2}$$

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 (4.2) is solved as

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta}). \tag{4.3}$$

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, (4.3) is given explicitly by

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

(see Amari [1985]).

Example 11. Examples of the η-coordinates. In the case of normal distributions (Example 1), the expectation parameter is

$$\eta_1 = E[x] = \mu = -\frac{\theta_1}{2\theta_2}, \tag{4.4}$$

$$\eta_2 = E[x^2] = \mu^2 + \sigma^2 = \left(\frac{\theta_1}{2\theta_2}\right)^2 - \frac{1}{2\theta_2}. \tag{4.5}$$

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

$$\eta_i = E[\delta_i(x)] = p_i = \frac{\exp(\theta_i)}{1 + \sum \exp(\theta_i)}. \tag{4.6}$$

In the case of $S = \{q(\mathbf{x})\}$ in Example 9,

$$\boldsymbol{\eta} = (\eta^1_i, \eta^2_{ij}, \cdots, \eta^n_{12\cdots n})$$

is given by

$$\eta^k_{i_1 \cdots i_k} = E[r^k_{i_1 \cdots i_k}] = E[x_{i_1} x_{i_2} \cdots x_{i_k}]. \tag{4.7}$$

Obviousrly, η^1_i is the expectation of x_i with respect to $q(\mathbf{x})$, η^2_{ij} are the second moments, and so on. Since $x_i = 0$ or 1, we have

$$\eta^k_{i_1 \cdots i_k} = \text{Prob}\{x_{i_1} = 1, x_{i_2} = 1, \cdots, x_{i_k} = 1\}.$$

The expectaion 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

$$l(\check{\mathbf{r}}; \boldsymbol{\theta}) = \boldsymbol{\theta} \cdot \check{\mathbf{r}} - \psi(\boldsymbol{\theta}). \tag{4.8}$$

By differentiating (4.8), 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. Let $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$ be the $\boldsymbol{\eta}$-coordinates of the m.l.e. point $\hat{\boldsymbol{\theta}}$. In terms of the $\boldsymbol{\eta}$-coordinates, the m.l.e. $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$ is directly given by

$$\hat{\boldsymbol{\eta}} = \check{\mathbf{r}}. \tag{4.9}$$

Hence, in an exponential family, when $\check{\mathbf{r}}$ is observed, it gives the m.l.e. in the $\boldsymbol{\eta}$-coordinates. If we want to obtain $\hat{\boldsymbol{\theta}}$, we need to calculate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\eta}})$ from (4.9). 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 useful in a curved exponential family and also in the case of repeated observations, as is shown in the next subsection.

4.2 Repeated observation and observed point

Let $\mathbf{r}_1, \cdots, \mathbf{r}_T$ 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

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

provided $\mathbf{r}_t$'s are independently generated. By introducing $\mathbf{r}^*_T = (\mathbf{r}_1, \cdots, \mathbf{r}_T)$, $\boldsymbol{\theta}^*_T = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T)$, the joint probability is written 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 $\boldsymbol{\eta}^*_T = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_T)$ are given by

$$\boldsymbol{\eta}_t = E_{\boldsymbol{\theta}_t}[\mathbf{r}_t]. \tag{4.10}$$

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

$$\hat{\boldsymbol{\eta}}_T^* = (\check{\mathbf{r}}_1, \cdots, \check{\mathbf{r}}_T). \tag{4.11}$$

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

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

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

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u}). \tag{4.12}$$

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

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}(\mathbf{x}_t, \mathbf{u}) = E_{\boldsymbol{\theta}_t}[\mathbf{r}]. \tag{4.13}$$

The equations (4.12) or (4.13) define a curved exponential family M^* in S_T^*.

In the i.i.d. case where $\mathbf{r}_1, \cdots \mathbf{r}_T$ are subject to the same distribution,

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

holds. In this case, the joint distribution is summarized in

$$p(\mathbf{r}_1, \cdots, \mathbf{r}_T; \boldsymbol{\theta}) = \exp\{\sum_{t=1}^{T} \mathbf{r}_t \cdot \boldsymbol{\theta} + \sum k(\mathbf{r}_t) - T\psi(\boldsymbol{\theta})\} \tag{4.14}$$

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,

$$\bar{\mathbf{r}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{r}_t, \tag{4.15}$$

which is the arithmetic mean of T random variables. The joint probability (4.14) can be transformed into the distribution of $\bar{\mathbf{r}}$,

$$p(\bar{\mathbf{r}}; \boldsymbol{\theta}) = \exp\{T(\bar{\mathbf{r}} \cdot \boldsymbol{\theta}) + \bar{k}(\bar{\mathbf{r}}) - \psi(\boldsymbol{\theta})\}, \tag{4.16}$$

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

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

the integration being taken over the region of $(\mathbf{r}_1, \cdots, \mathbf{r}_T)$ such that the arithmetic mean of $\mathbf{r}_1, \cdots, \mathbf{r}_T$ is $\bar{\mathbf{r}}$. However, (4.16) shows that the probability of $(\mathbf{r}_1, \cdots, \mathbf{r}_T)$ 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 and 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 and $\bar{k}(\bar{\mathbf{r}})$. 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 for the i.i.d. case, and we need to consider S_T^* in the general case.

The maximum likelihood estimator (m.l.e.) $\hat{\boldsymbol{\theta}}$ from the observed data $\mathbf{r}_1, \cdots \mathbf{r}_T$ 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

$$\bar{\mathbf{r}} = \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\hat{\boldsymbol{\theta}}). \tag{4.17}$$

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

$$\hat{\boldsymbol{\eta}} = \bar{\mathbf{r}} \tag{4.18}$$

when $\bar{\mathbf{r}}$ is observed. We call the distribution $p(\mathbf{r}; \hat{\boldsymbol{\theta}})$ or the point $\hat{\boldsymbol{\eta}}$ in S in the $\boldsymbol{\eta}$-coordinate system the observed point. The observed data $\mathbf{r}_1, \cdots, \mathbf{r}_T$ are simply summarized into the single observed point (that is, the m.l.e.) $\hat{\boldsymbol{\eta}}$ in S in the $\boldsymbol{\eta}$-coordinate system without any loss of information in the sense of Fisher.

4.3 Estimation in curved exponential family

Let us consider the maximum likelihood estimation in a curved exponential family M embedded in S. In the i.i.d. case, the observed data $\mathbf{r}_1, \cdots, \mathbf{r}_T$ are summarized in the observed point $\hat{\boldsymbol{\eta}} = \bar{\mathbf{r}}$ in S. This is a point of S but it does not necessarily belong to M. The m.l.e. $\hat{\mathbf{u}}$ or the corresponding distribution $\boldsymbol{\theta}(\hat{\mathbf{u}}) \in M$ is given by maximizing the log likelihood

$$l(\bar{\mathbf{r}}; \mathbf{u}) = \bar{\mathbf{r}} \cdot \boldsymbol{\theta}(\mathbf{u}) - \psi\{\boldsymbol{\theta}(\mathbf{u})\} + \bar{k}(\bar{\mathbf{r}}) \tag{4.19}$$

with respect to $\mathbf{u}$. We can neglect the term $\bar{k}$ since it does not depend on $\boldsymbol{\theta}$.

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

$$\begin{aligned} K(\boldsymbol{\theta}' || \boldsymbol{\theta}) &= \int p(\mathbf{r}; \boldsymbol{\theta}') \log \frac{p(\mathbf{r}; \boldsymbol{\theta}')}{p(\mathbf{r}; \boldsymbol{\theta})} d\mu(\mathbf{r}) \\ &= \varphi(\boldsymbol{\theta}') - \boldsymbol{\eta}' \cdot \boldsymbol{\theta} + \psi(\boldsymbol{\theta}), \end{aligned} \tag{4.20}$$

where

$$-\varphi(\boldsymbol{\theta}') = H(\boldsymbol{\theta}') = -\int p(\mathbf{r}; \boldsymbol{\theta}') \log p(\mathbf{r}; \boldsymbol{\theta}') d\mu(\mathbf{r}) \tag{4.21}$$

is the entropy of the distribution $p(\mathbf{r}; \boldsymbol{\theta}')$ and $\boldsymbol{\eta}'$ is the $\boldsymbol{\eta}$-coordinates of $\boldsymbol{\theta}'$. Therefore, maximizing the log likelihood (4.19) is equivalent to minimizing the KL-divergence $K(\hat{\boldsymbol{\theta}} || \boldsymbol{\theta}(\mathbf{u}))$ from the observed point $\hat{\boldsymbol{\theta}}$ to points $\boldsymbol{\theta}(\mathbf{u})$ belonging to M. This implies that the m.l.e. $\boldsymbol{\theta}(\hat{\mathbf{u}})$ is the point in M that is closest to the observed point $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\eta}}$ in S (Fig.5). Geometrically, the m.l.e. $\boldsymbol{\theta}(\hat{\mathbf{u}})$ is given by m-projecting $\hat{\boldsymbol{\theta}}$ to M, as is explained in Appendix. It is given by solving the likelihood equation

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

Because of

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

by introducing the matrix

$$B(\mathbf{u}) = \frac{\partial \boldsymbol{\theta}(\mathbf{u})}{\partial \mathbf{u}}, \tag{4.22}$$

the likelihood equation is given by

$$B(\mathbf{u})\{\bar{\mathbf{r}} - \boldsymbol{\eta}(\mathbf{u})\} = 0 \tag{4.23}$$

In S_T^*, the KL-divergence from $\boldsymbol{\theta}_T'^*$ to $\boldsymbol{\theta}_T^*$ is decomposed as

$$\begin{aligned} K(\boldsymbol{\theta}_T'^* \parallel \boldsymbol{\theta}_T^*) &= \sum_{t=1}^{T} K(\boldsymbol{\theta}_t' \parallel \boldsymbol{\theta}_t) \\ &= -\sum H(\boldsymbol{\theta}_t') - \sum \boldsymbol{\eta}_t \cdot \boldsymbol{\theta}_t + \sum \psi(\boldsymbol{\theta}_t). \end{aligned} \tag{4.24}$$

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

$$\sum_{t=1}^{T} B(\mathbf{x}_t, \mathbf{u})\{\mathbf{r}_t - \boldsymbol{\eta}(\mathbf{x}_t, \mathbf{u})\} = 0, \tag{4.25}$$

where

$$B(\mathbf{x}_t, \mathbf{u}) = \frac{\partial \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u})}{\partial \mathbf{u}}. \tag{4.26}$$

4.4 Examples of observed points in S

Example 12. Observed point in the normal mixture. This is the case of repeated observation. Let $\mathbf{r}_1, \cdots, \mathbf{r}_T$ be T independent i.i.d. data. They are summarized in $\bar{\mathbf{r}}$,

$$\begin{aligned} \bar{r}_{11} &= \tfrac{1}{T}\textstyle\sum x_t, \qquad \bar{r}_{12} = \tfrac{1}{T}\sum x_t^2, \\ \bar{r}_{2i} &= \tfrac{1}{T}\textstyle\sum \delta_i(z_t), \\ \bar{r}_{3i} &= \tfrac{1}{T}\textstyle\sum x_t \delta_i(z_t), \\ \bar{r}_{4i} &= \tfrac{1}{T}\textstyle\sum x_t^2 \delta_i(z_t), \end{aligned} \tag{4.27}$$

where $\mathbf{r}_t$ is a function of the tth observation (x_t, z_t), $t = 1, \cdots, T$. The $\boldsymbol{\eta}$-coordinates of

S are given by the expectation of $\mathbf{r}$ which is the same as that of $\bar{\mathbf{r}}$,

$$\begin{aligned}
\eta_{11} &= \sum_{i=0}^{k} p_i \mu_i, \quad \eta_{12} = \sum_{i=0}^{k} p_i (\mu_i^2 + \sigma_i^2), \\
\eta_{2i} &= p_i, \\
\eta_{3i} &= p_i \mu_i, \\
\eta_{4i} &= p_i (\mu_i^2 + \sigma_i^2).
\end{aligned} \tag{4.28}$$

Since S is an exponential family, when $\bar{\mathbf{r}}$ is given from the observed data, $\hat{\boldsymbol{\eta}} = \bar{\mathbf{r}}$ is the m.l.e. in the $\boldsymbol{\eta}$-coordinate system. The m.l.e. $\hat{\mathbf{u}}$ is obtained by solving

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

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

We next show the case of stochastic perceptron, where multiple observations $(\mathbf{r}_1, \cdots, \mathbf{r}_T)$ are also summarized in an observed point $\bar{\mathbf{r}}$ in S.

Example 13. Observed point in the multilayer perceptron. In the case of the stochastic multilayer perceptron, random variables $\mathbf{r}_1, \cdots, \mathbf{r}_T$ are summarized in the set of sufficient statistics $\bar{\mathbf{r}} = (\bar{\mathbf{r}}_{1,k}, \bar{\mathbf{r}}_{2,k})$,

$$\begin{aligned}
\bar{r}_{1,\mathbf{k}} &= \frac{1}{T} \sum_t y_t \delta_{\mathbf{k}}(\mathbf{z}_t), \\
\bar{\mathbf{r}}_{2,\mathbf{k}} &= \frac{1}{T} \sum_t \mathbf{x}_t \delta_{\mathbf{k}}(\mathbf{z}_t),
\end{aligned} \tag{4.29}$$

where $\mathbf{r}_t$ is composed of $(y_t, \mathbf{z}_t, \mathbf{x}_t)$. The $\boldsymbol{\theta}$-coordinates $\boldsymbol{\theta}(\mathbf{u})$ is given by (3.16). Since it does not depend on $\mathbf{x}_t$, it is common to all $\mathbf{r}_t$. The $\boldsymbol{\eta}$-coordinates $\boldsymbol{\eta}(\mathbf{u})$ are given by

$$\begin{aligned}
\eta_{1,\mathbf{k}} &= \mathrm{E}[y \delta_{\mathbf{k}}(\mathbf{z})] = \{\mathrm{E} \textstyle\prod_i \varphi(k_i, \mathbf{w}_i \cdot \mathbf{x}_)\} \varphi(1, \mathbf{v} \cdot \mathbf{k}), \\
\boldsymbol{\eta}_{2,\mathbf{k}} &= \mathrm{E}[\mathbf{x} \delta_{\mathbf{k}}(\mathbf{z})] = \mathrm{E}\{\mathbf{x} \textstyle\prod_i \varphi(k_i, \mathbf{w}_i \cdot \mathbf{x})\},
\end{aligned} \tag{4.30}$$

where E is the expectation with respect to $q(\mathbf{x})$. The observed point $\hat{\boldsymbol{\eta}}$ is given by the sufficient statistics $\bar{\mathbf{r}}$, and the m.l.e. $\hat{\mathbf{u}}$ is given by solving (4.23).

The following is the example of the mixture of expert nets where multiple observations cannot be summarized in the original simple S.

Example 14. Observed point in the mixture of experts. Let (y_t, z_t) be T independent observations generated from T independently generated inputs $\mathbf{x}_t$, $t = 1, \cdots, T$. In this case, we cannot summarize the corresponding T random variables $\mathbf{r}_1, \cdots, \mathbf{r}_T$ in the arithmetic mean $\bar{\mathbf{r}}$. This is because $\boldsymbol{\theta}_t = \boldsymbol{\theta}(\mathbf{x}_t, \mathbf{u})$ depends on $\mathbf{x}_t$, so that the conditional distribution $p(y_t, z_t|\boldsymbol{\theta}_t)$ is not identical. In the perceptron case, we have the identical distributions $p(y_t, z_t, \mathbf{x}_t)$ and summarize $\mathbf{r}_1, \cdots, \mathbf{r}_T$ in the sufficient statistics $\bar{\mathbf{r}}$. However, in the present case, the unconditional distributions $p(y_t, z_t, \mathbf{x}_t)$ do not form an exponential family. Therefore, we need to consider the larger direct product space of the exponential families

$$S_T^* = S_1 \times S_2 \times \cdots \times S_T,$$

where S_t is the manifold for $p(y_t, z_t|\mathbf{x}_t)$. The observations $\check{\mathbf{r}}_1, \cdots, \check{\mathbf{r}}_T$ define the observed point $\hat{\boldsymbol{\eta}}_T^* = (\check{\mathbf{r}}_1, \cdots, \check{\mathbf{r}}_T)$ in the product space S_T^*. Except for the large dimensionality of S_T^*, all the other discussions on estimation are the same.

5 Hidden Variables and Data Submanifold D

5.1 Partial observations

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 z in the case of neural networks. Some of them are observable or specifiable but some are not. For exmaple, $\mathbf{x}$ and y are observed but 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. Since $\mathbf{s}_h$ 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:

$$\begin{aligned} D \;=\; & \{\hat{\boldsymbol{\eta}} |\; \hat{\boldsymbol{\eta}} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h); \\ & \mathbf{s}_v \text{ being fixed at the observed value and } \mathbf{s}_h \\ & \text{taking arbitrary values}\}. \end{aligned} \tag{5.1}$$

In many cases, $\mathbf{r}$ is linear in $\mathbf{s}_h$, so that $t\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h^1) + (1-t)\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h^2) \in D$ 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, we linearize D by taking the linear combinations in the $\boldsymbol{\eta}$-coordinates. In these cases, D is a linear submanifold of S in the $\boldsymbol{\eta}$-coordinate system.

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

$$\mathbf{r} = (\mathbf{r}_v, \mathbf{r}_h), \tag{5.2}$$

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

$$D = \{\hat{\boldsymbol{\eta}} \mid \hat{\boldsymbol{\eta}}_v = \mathbf{r}_v,\ \hat{\boldsymbol{\eta}}_h : \text{ arbitrary}\}. \tag{5.3}$$

This forms a h-dimensional submanifold in S (Fig.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, 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 non-singular matrix

A

$$\tilde{\boldsymbol{\eta}} = A\boldsymbol{\eta}, \qquad \tilde{\mathbf{r}} = A\mathbf{r}, \tag{5.4}$$

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 while $\tilde{\boldsymbol{\eta}}_h$ 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 by

$$\tilde{\boldsymbol{\theta}} = A^{-1}\boldsymbol{\theta} \tag{5.5}$$

contravariantly to (5.4). 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 one example.

Example 15. Data submanifold in the normal mixture. For the sake of simplicity, we assume $\sigma_0 = \sigma_i = 1$ in the normal mixture of Example 4. The general case can be analyzed quite similarily. In this case, the log likelihood is written as

$$\begin{aligned} \log p(x; p_i, \mu_i) &= \mu_0 x + \sum_{i=1}^{k} \delta_i(z) \left\{ \log \frac{p_i}{p_0} - \frac{1}{2}(\mu_i^2 - \mu_0^2) \right\} \\ &\quad + \sum_{i=1}^{k} x\delta_i(z)(\mu_i - \mu_0) + \log p_0 - \frac{1}{2}\mu_0^2 - \frac{x^2}{2}. \end{aligned} \tag{5.6}$$

Therefore, this is a full (not curved) exponential family, i.e., $M = S$, 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

$$r_0 = x, \qquad r_{1i} = \delta_i(z), \qquad r_{2i} = x\delta_i(z), \tag{5.7}$$

$$\theta_0 = \mu_0, \qquad \theta_{1i} = \log \frac{p_i}{p_0} - \frac{1}{2}(\mu_i^2 - \mu_0^2), \qquad \theta_{2i} = \mu_i - \mu_0 \tag{5.8}$$

$$\eta_0 = \sum_{i=0}^{k} p_i \mu_i, \qquad \eta_{1i} = p_i, \qquad \eta_{2i} = p_i \mu_i, \tag{5.9}$$
$$(i = 1, \cdots, k).$$

When x is observed but z is not, $\mathbf{s}_v = x$ and $\mathbf{s}_h = \{\delta_i(z)\}$, and the observed data submanifold is given by

$$D = \{\hat{\boldsymbol{\eta}} |\ \hat{\eta}_0 = x,\ \hat{\eta}_{1i} = \alpha_i,\ \hat{\eta}_{2i} = x\alpha_i\} \tag{5.10}$$

where α_i are the free parameters corresponding to the unobserved $\delta_i(z)$. It should be noted that $\delta_i(t)$ is 0 or 1 but α_i takes any real values satisfying $0 \leq \alpha_i$, $\sum \alpha_i \leq 1$. Since D is a linear submanifold in $\boldsymbol{\eta}$, it is convenient to use another coordinate system $\tilde{\boldsymbol{\eta}}$,

$$\begin{aligned} \tilde{\eta}_0 &= \eta_0, \\ \tilde{\eta}_{1,i} &= \eta_{1,i}, \\ \tilde{\eta}_{2,i} &= \eta_{2,i} - x\eta_{1,i}. \end{aligned} \tag{5.11}$$

This is a linear transformation from $\boldsymbol{\eta}$ to $\tilde{\boldsymbol{\eta}}$ but the transformation matrix A depends on the observed value x. The data submanifold D is written as

$$D = \{\hat{\boldsymbol{\eta}} |\ \tilde{\eta}_0 = x, \tilde{\eta}_{2,i} = 0, \tilde{\eta}_{1,i} = \alpha_i = \text{ arbitrary }\} \tag{5.12}$$

in the new $\tilde{\boldsymbol{\eta}}$-coordinates. Hence, it is divided into $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\eta}}_v, \tilde{\boldsymbol{\eta}}_h)$,

$$\begin{aligned} \tilde{\boldsymbol{\eta}}_v &= (\tilde{\eta}_0, \eta_{2,i}), \\ \tilde{\boldsymbol{\eta}}_h &= (\tilde{\eta}_{1,i}). \end{aligned}$$

The corresponding $\tilde{\boldsymbol{\theta}}$-coordinates are

$$\begin{aligned} \tilde{\theta}_0 &= \theta_0, \\ \tilde{\theta}_{1,i} &= \theta_{1,i}, \\ \tilde{\theta}_{2,i} &= \theta_{2,i} + x\theta_{1,i}. \end{aligned} \tag{5.13}$$

It should be noted that the transformation depends on the observed x.

5.2 Data submanifold in the product space S_T^*

In the case of multiple or repeated observations, the partial observation determines the data submanifold D_T^*,

$$\begin{aligned} D_T^* &= \{\boldsymbol{\eta}_T^* |\ \boldsymbol{\eta}_T^* = (\mathbf{r}_1, \cdots, \mathbf{r}_T), \text{ where unobserved} \\ &\quad \text{random variables } \mathbf{s}_{h,t} \text{ remain as free parameters in each} \\ &\quad \mathbf{r}_t = \mathbf{r}(\mathbf{s}_{v,t}, \mathbf{s}_{h,t}),\ t = 1, \cdots, T\}. \end{aligned} \tag{5.14}$$

When the component D_t is linear, D_T^* is a $\boldsymbol{\eta}$-linear submanifold in the product space S_T^*.

In the i.i.d. case such as the normal mixture, a full observed data $\mathbf{r}_T^* = (\mathbf{r}_1, \cdots, \mathbf{r}_T)$ can be summarized into $\bar{\mathbf{r}} \in S$. However, we cannot summarize $D_T^* \subset S_T^*$ into $D \in S$ even in the i.i.d. case. This is because the transformation A of (5.4) in general depends on the observed data $\mathbf{s}_{v,t}$. Therefore, the $\tilde{\boldsymbol{\eta}}_t$-coordinates are diffenrent for different t, so that we cannot summarize the visual part $\tilde{\boldsymbol{\eta}}_{v,t}$ by taking their arithmetic mean. Hence, we need to treat D_T^* without summarizing it. The summarization is possible only when $\mathbf{r}_t$ is divided directly into the visible and hidden parts $\mathbf{r}_t = (\mathbf{r}_{v,t}, \mathbf{r}_{h,t})$.

Example 16. Data submanifold of the stochastic perceptron. The random variable $\mathbf{r}(t)$ at time t is composed of $(r_{1,\mathbf{k},t}, \mathbf{r}_{2,\mathbf{k},t})$, and the observed point $\hat{\boldsymbol{\eta}}_t = (r_{1,\mathbf{k},t}, \mathbf{r}_{2,\mathbf{k},t})$ is written as

$$\begin{aligned} \hat{\eta}_{1,\mathbf{k}} &= r_{1,\mathbf{k},t} = y_t \delta_{\mathbf{k}}(\mathbf{z}_t), \\ \hat{\boldsymbol{\eta}}_{2,\mathbf{k}} &= \mathbf{r}_{2,\mathbf{k},t} = \mathbf{x}_t \delta_{\mathbf{k}}(\mathbf{z}_t) \end{aligned}$$

where $\mathbf{x}_t$ and y_t are observable but $\delta_{\mathbf{k}}(\mathbf{z}_t)$ are not. Hence, the observed data submanifold D_t is

$$D_t = \{\boldsymbol{\eta} \mid \eta_{1,\mathbf{k}} = y_t \alpha_{\mathbf{k}},\ \boldsymbol{\eta}_{2,\mathbf{k}} = \mathbf{x}_t \alpha_{\mathbf{k}}\}, \tag{5.15}$$

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 $\boldsymbol{\eta}$-coordinates.

We take one component D_t in S_t, neglecting the suffix t. The $\tilde{\boldsymbol{\eta}}$-coordinates in which $\tilde{\mathbf{r}}$ is partitioned into $\tilde{\mathbf{r}}_v$ and $\tilde{\mathbf{r}}_h$ are given by

$$\begin{aligned}
\tilde{\eta}_0 &= \sum_{\mathbf{k}} \boldsymbol{\eta}_{1,\mathbf{k}} = y, \\
\tilde{\eta}_{2,0} &= \sum_{\mathbf{k}} \boldsymbol{\eta}_{2,\mathbf{k}} = \mathbf{x}, \\
\tilde{\eta}_{1,\mathbf{k}} &= x_1 \eta_{1,\mathbf{k}} - y \eta_{2,\mathbf{k},1} = 0, \quad k \neq 0 \\
\tilde{\eta}_{2,\mathbf{k},i} &= x_1 \eta_{2,\mathbf{k},i} - x_i \eta_{2,\mathbf{k},1} = 0, \quad \mathbf{k} \neq 0 \\
\tilde{\eta}_{2,\mathbf{k},1} &= \frac{1}{x_1} \eta_{2,\mathbf{k},1} = \alpha_{\mathbf{k}}, \quad \mathbf{k} \neq 0,
\end{aligned} \tag{5.16}$$

where $x_1 \neq 0$ is assumed and $\eta_{2,\mathbf{k},i}$ is the ith component of $\boldsymbol{\eta}_{2,\mathbf{k}}$. The first four components are observable $\tilde{\boldsymbol{\eta}}_v$, and the last $\tilde{\eta}_{2,\mathbf{k},1} = \alpha_{\mathbf{k}}$ are hidden, $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\eta}}_v, \tilde{\boldsymbol{\eta}}_h)$. The transformation A depends on $(\mathbf{x}, y)$, so that A_t depends on $(\mathbf{x}_t, y_t)$.

Example 17. Boltzmann machines with hidden units. In a Boltzmann machine, the neural units are divided into two parts, visible units and hidden units. The visible units include the input and output units whose behaviors we are interested in. The hidden units are connected with the visible units as well as with the hidden units, so that they can modify the behaviors of the input-output units, even though we do not have direct interest in the behaviors of the hidden units. Let us divide the state vector into the visible and hidden parts,

$$\mathbf{x} = (\mathbf{x}^V, \mathbf{x}^H).$$

The new random variables $\mathbf{r}^V$, $\mathbf{r}^H$, $\mathbf{r}^{VH}$ are introduced by

$$\begin{aligned}
r_{ij}^V &= x_i^V x_j^V, \qquad i < j \\
r_{ij}^{VH} &= x_i^V x_j^H, \\
r_{ij}^H &= x_i^H x_j^H, \qquad i < j.
\end{aligned} \tag{5.17}$$

We also denote the connection weights and bias terms by $\boldsymbol{\theta} = (\boldsymbol{\theta}^V, \boldsymbol{\theta}^H, \boldsymbol{\theta}^{VH})$,

$$\begin{aligned} \boldsymbol{\theta}^V &= (w_{ij}^{VV}), \quad i < j, \quad (w_{ii}^{VV} = s_i^V) \\ \boldsymbol{\theta}^H &= (w_{ij}^{HH}), \quad i < j, \quad (w_{ii}^{HH} = s_i^H) \\ \boldsymbol{\theta}^{VH} &= (w_{ij}^{VH}), \end{aligned} \tag{5.18}$$

where w_{ij}^{VH} is, for example, the connection weights from the jth hidden unit to the ith visible unit.

Let $q(\mathbf{x}^V)$ be the probability distribution which is to be realized on the visible units of a Boltzmann machine with hidden units. More precisely, by dividing the visible units into the input units and output units such that $\mathbf{x}^V = (\mathbf{x}^I, \mathbf{x}^O)$, it is required to realize the conditional distribution $q(\mathbf{x}^O|\mathbf{x}^I)$. Since the input distribution $q(\mathbf{x}^I)$ is given from the outside, the distribution $q(\mathbf{x}^V)$ is realized by $q(\mathbf{x}^O|\mathbf{x}^I)$ as

$$q(\mathbf{x}^V) = q(\mathbf{x}^I) q(\mathbf{x}^O|\mathbf{x}^I).$$

If $q(\mathbf{x}^V)$ is truly realizable by a Boltzmann machine, there is some W such that

$$q(\mathbf{x}^V) = \sum_{\mathbf{x}^H} p(\mathbf{x}^V, \mathbf{x}^H; W). \tag{5.19}$$

However, this is not necessarily the case, and it is required to approximate a given $q(\mathbf{x}^V)$ by a Boltzmann machine as far as possible.

When $q(\mathbf{x}^V)$ is specified explicitly from the outside, the data manifold D consists of all the distributions $q(\mathbf{x}^V, \mathbf{x}^H)$ whose marginal distribution is equal to the prescribed $q(\mathbf{x}^V)$,

$$D = \{q(\mathbf{x}^V, \mathbf{x}^H) \mid \sum_{\mathbf{x}^H} q(\mathbf{x}^V, \mathbf{x}^H) = q(\mathbf{x}^V)\}, \tag{5.20}$$

which is a submanifold of $S = \{q(\mathbf{x}^V, \mathbf{x}^H)\}$. We have shown the $\boldsymbol{\theta}$-coordinates and $\boldsymbol{\eta}$-coordinates of the manifold of probability distributions in (3.6) and (4.7). They hold even when $\mathbf{x}$ is divided into the $\mathbf{x}^V$ and $\mathbf{x}^H$ parts. The data submanifold D of (5.20) is rewritten in a linear form in the $\boldsymbol{\eta}$-coordinates, where the components corresponding

to $q(\mathbf{x}^V)$ are specified but the components corresponding to $q(\mathbf{x}^H|\mathbf{x}^V)$ remain as the free parameters in D. When $q(\mathbf{x}^V)$ is not explicitly given but independent examples $\mathbf{x}_1^V, \cdots, \mathbf{x}_T^V$ are observed, we can construct the $\boldsymbol{\eta}$-coordinates $\hat{\boldsymbol{\eta}}$ of $q(\mathbf{x}^V)$ from them. We need to treat D_T^* in this case.

Example 18. Hidden Markov chain. Let us consider a Markov chain with $k+1$ states $\{s_0, s_1, \cdots, s_k\}$. The state set is divided into $m+1$ subsets $A_0, \cdots, A_m$, where

$$A_0 = \{s_0, \cdots, s_{i_0}\}, \cdots, A_m = \{s_{i_{m-1}+1}, \cdots, s_{i_m}\}.$$

Assume that, when the state is in the set A_μ, the Markov chain outputs letter a_μ ($\mu = 0, \cdots, m$) from the alphabet $A = \{a_0, a_1, \cdots, a_m\}$. Then, corresponding to the state sequence $\mathbf{s} = s_0 s_1 \cdots s_T$, we can observe the output sequence $\mathbf{a} = a_0 a_1 \cdots a_T$, where $s_t \in A_t$. The probability of $\mathbf{a}$ is obtained by

$$p(a_0 a_1 \cdots a_T) = \sum p(s_0 s_1 \cdots s_T), \tag{5.21}$$

where the summation is taken over all s_t ($t = 0, \cdots, T$) such that $s_t \in A_t$.

In this case, a_t is a function of s_t. More generally, we can treat the case where a_t is stochastically determined from s_t subject to the conditional probability

$$p_{\mu i} = \text{Prob}\{a = \mu | s = i\}.$$

The sequence $a_0 a_1 a_2 \cdots$ forms a stochastic process, but is not necessary a Markov chain. Such a process is called a hidden Markov model, since a Markov chain underlies the process but we cannot observe its state. We can only observe the output letter sequence $\mathbf{a}$ which is a function of the state sequence $\mathbf{s}$ or stochastically generated from $\mathbf{s}$. This is a hidden model, where the sufficient statistic is a pair $(\mathbf{s}, \mathbf{a})$ but we can observe only $\mathbf{a}$ and $\mathbf{s}$ is hidden. It is also generalized to a hidden Markov random field (see, e.g., Geman and Geman, 1984; Besag and Green, 1993).

The class of hidden Markov models is much larger than that of Markov chains in spite of its simple mechanism of generation. It is widely used in applications, in particular in

speech recognition. The so-called Baum-Welch algorithm (Baum et al., 1970) is the same as the EM algorithm. Recently, Künsch, Geman and Kehagias (1994) tries to apply a hidden Markov random field to vision research. An old problem of state identification in the hidden Markov model was solved recently (Ito, Amari and Kobayashi, 1992). Although we do not analyze the hiden Markov model here, the information geometry is expected to provide it with fruitful perspectives.

Let S be the manifold of all the probability distributions of Ergodic time series on the state sequence $\mathbf{s}$. The set M of the Markov chains is an e-flat submanifold in S. On the other hand, an observed sequence $\mathbf{a}$ gives an observed submanifold D in S. The problem is to find the underlying Markov chain $\hat{\mathbf{u}} \in M$ that minimizes $K(D \parallel M)$. At the same time, the distribution $Q \in D$ that minimizes $K(D \parallel M)$ is obtained by the e-projection shown in the following. This is related to the conditional probability $p(\mathbf{s}|\mathbf{a};\hat{\mathbf{u}})$, which gives a Markov chain (Günsch, Geman and Kehagias, 1994). In the case of a random Markov field or hidden random Markov field, the calculation of the conditional probability or the e-projection corresponds to the image restoration.

6 EM Algorithm and em Algorithm

6.1 The EM algorithm

It is required to obtain the best parameter $\hat{\mathbf{u}}$ 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 of calculating the maximum likelihood estimator from partially observed data (Dempster, Laird and Rudin, 1977). It may be regarded as an iterative procedure to estimate 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 Rudin 1977).

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 is not available. It is one idea to estimate the unknown $\check{\mathbf{r}}$ based on the observed $\mathbf{s}_v$ by taking the conditional expectation

$$\hat{\mathbf{r}}(\mathbf{u}') = E[\mathbf{r}|\ \mathbf{s}_v;\boldsymbol{\theta}(\mathbf{u}')], \tag{6.1}$$

based on the temporal candidate distribtion $\boldsymbol{\theta}(\mathbf{u}')$ that is, the conditional expectation is taken based on the distribution $p\{\mathbf{r};\boldsymbol{\theta}(\mathbf{u}')\}$. Then, the log likelihood function is estimated by

$$E[\log p(\mathbf{r};\mathbf{u})|\ \mathbf{s}_v;\boldsymbol{\theta}(\mathbf{u}')] = \boldsymbol{\theta}(\mathbf{u}) \cdot \hat{\mathbf{r}}(\mathbf{u}') - \psi\{\boldsymbol{\theta}(\mathbf{u})\}, \tag{6.2}$$

where $\hat{\mathbf{r}}(\mathbf{u}')$ is given by (6.1). The estimated observed point $\hat{\mathbf{r}}$ or the estimated log likelihood function (6.2) 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 log likelihood (6.2) of the estimated $\hat{\mathbf{r}}$, or equivalently by minimizing the KL-divergence $K[\hat{\mathbf{r}}(\mathbf{u}') \parallel \boldsymbol{\theta}(\mathbf{u})]$ from the guessed data point $\hat{\boldsymbol{\eta}} = \hat{\mathbf{r}}(\mathbf{u}')$ 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 maximiaing (EM) algorithm, where the conditional expectation is used to obtain a candidate $\hat{\mathbf{r}}(\mathbf{u}')$ in D (Fig.7).

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

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,

$$\hat{\boldsymbol{\eta}}_{(i)} = E[\mathbf{r} \mid \mathbf{s}_v; \hat{P}_i]. \tag{6.3}$$

   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 candidate observed point $\hat{Q}_i \in D$. This is the point in M that minimizes $K(\hat{Q}_i \parallel P)$, $P \in M$ 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 $\bar{\mathbf{r}} = (1/T)\sum \mathbf{r}_t$, 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

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

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

$$\hat{\boldsymbol{\eta}}_{t,(i)} = E[\mathbf{r}_t \mid \mathbf{s}_{v,t}; \hat{P}_i]. \tag{6.4}$$

It is known that the likelihood is increased by one iteration of the E and M steps. Hence, the EM algorithm converges to the 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 $\hat{\mathbf{u}}_{m.l.e.}$, because 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.

6.2 Geometric *em*-algorithm

Since the best estimate $\hat{\mathbf{u}}$ in M minimizes the Kullback divergence from the observed data manifold D to the model M, it is natural to study the problem from the information geometrical point of view. When a complete data 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 $K(\bar{Q} \parallel P)$, $P \in M$. 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

$$K(\hat{P} \parallel \hat{Q}) = \min_{P \in M, Q \in D} K(Q \parallel P). \tag{6.5}$$

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 $K(Q \parallel P)$ was proposed by Csiszár and Tusnády [1984] in a general perspective. Amari, Kurata and Nagaoka (1992), Byrne (1992), Shimodaira (1993), Neal and Hinton (1994) also studied such problems in various fields. Information geometry proves that the point $\hat{P} \in M$ that minimizes $K(Q \parallel P)$ 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 (Appendix).

Dually to the above, the point $\hat{Q} \in D$ that minimizes $K(Q \parallel P)$ is given by the e-projection of P to D. The e-projection is given by the e-geodesic connecting P and $\hat{Q}$ which is orthogonal to D at $\hat{Q}$. 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 (Fig. 8).

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 $K(Q \parallel \hat{P}_i)$, $Q \in D$.

2. m-step:

m-project $\hat{Q}_i$ to M. This gives $\hat{P}_{i+1}$ that minimizes $K(\hat{Q}_i \parallel P)$, $P \in M$.

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

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

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

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

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 (4.27)

$$K(Q^* \parallel P^*) = \sum_{t=1}^{T} K(Q_t \parallel P_t). \tag{6.6}$$

This shows that the point $\hat{Q}^* = (\hat{Q}_1, \cdots, \hat{Q}_T)$ that minimizes $K(Q^* \parallel P^*)$, $Q^* \in D_T^*$, is composed of the points $\hat{Q}_t$ that minimizes $K(Q_t \parallel P_t)$, $Q_t \in D_t$. 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 very similar. The next section is devoted to elucidation of their relation.

7 Information Geometry of EM and em Algorithms

7.1 Properties of the e-projection

The present section establishes the geometry of the EM and em algorithms. It is intriguing to see if the two algorithms are equivalent or not. To show this, we elucidate the relation between the e-projection and the conditional expectation. In the beginning, we show 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 maximum likelihood estimator. We summarize it in the following theorem.

Theorem 1. Let $\bar{Q}$ 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,

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

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-projected point of P to D where 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 visual 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,

$$\boldsymbol{\theta}_h^P = \boldsymbol{\theta}_h^*. \tag{7.1}$$

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

$$p(\mathbf{r}_h | \ \check{\mathbf{r}}_v, P) = p(\mathbf{r}_h | \ \check{\mathbf{r}}_v, Q^*). \tag{7.2}$$

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

$$E_P[\mathbf{r}_h|\ \check{\mathbf{r}}_v] = E_{Q^*}[\mathbf{r}_h|\ \check{\mathbf{r}}_v]. \tag{7.3}$$

Proof. The property 1) is trivial, since $Q^* \in D$. 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 $K(Q \parallel P)$ is rewritten as

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

because of $E_Q[\mathbf{r}_v] = \check{\mathbf{r}}_v$. In the submanifold D, $\boldsymbol{\eta}_h$ is the free variable while $\boldsymbol{\eta}_v$ is fixed at the value $\check{\mathbf{r}}_v$. Since the e-projection minimizes $K(Q \parallel P)$, we have by differentiation

$$\begin{aligned} \frac{\partial K(Q \parallel P)}{\partial \boldsymbol{\eta}_h} &= \frac{\partial \boldsymbol{\theta}_v}{\partial \boldsymbol{\eta}_h} \cdot \bar{\mathbf{r}}_v + \frac{\partial \boldsymbol{\theta}_h}{\partial \boldsymbol{\eta}_h} \cdot \boldsymbol{\eta}_h + (\boldsymbol{\theta}_h - \boldsymbol{\theta}_h^P) \\ &\quad - \frac{\partial \psi(\boldsymbol{\theta}_Q)}{\partial \boldsymbol{\theta}_v} \frac{\partial \boldsymbol{\theta}_v}{\partial \boldsymbol{\eta}_h} - \frac{\partial \psi(\boldsymbol{\theta}_Q)}{\partial \boldsymbol{\theta}_h} \frac{\partial \boldsymbol{\theta}_h}{\partial \boldsymbol{\eta}_h} \\ &= \boldsymbol{\theta}_h - \boldsymbol{\theta}_h^P, \end{aligned} \tag{7.4}$$

because of

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

The right-hand side of (7.4) vanishes at the minimum point Q^*, so that (7.1) holds.

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

$$\begin{aligned} p(\mathbf{r}_h|\ \check{\mathbf{r}}_v, \boldsymbol{\theta}_P) &= \frac{p(\mathbf{r}_h, \check{\mathbf{r}}_v; \boldsymbol{\theta}_P)}{p(\check{\mathbf{r}}_v; \boldsymbol{\theta}_P)} \\ &= \frac{\exp\{\boldsymbol{\theta}_v^P \cdot \check{\mathbf{r}}_v + \boldsymbol{\theta}_h^P \cdot \mathbf{r}_h + k(\check{\mathbf{r}}_v, \mathbf{r}_h) - \psi\}}{\int \exp\{\boldsymbol{\theta}_v^P \cdot \check{\mathbf{r}}_v + \boldsymbol{\theta}_h^P \cdot \mathbf{r}_h + k(\check{\mathbf{r}}_v, \mathbf{r}_h) - \psi\} d\mu(\mathbf{r}_h)} \\ &= \exp\{\boldsymbol{\theta}_h^P \cdot \mathbf{r}_h - \tilde{\psi}\}, \end{aligned} \tag{7.5}$$

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

P and Q^*, and hence

$$E_P[\mathbf{r}_h|\ \check{\mathbf{r}}_v] = E_{Q^*}[\mathbf{r}_h|\ \check{\mathbf{r}}_v]. \tag{7.6}$$

7.2 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? In order 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

$$\mathbf{s}_Q(\mathbf{r}_v) = E_Q[\mathbf{r}_h|\mathbf{r}_v], \tag{7.7}$$

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

$$F : (\boldsymbol{\eta}_v, \boldsymbol{\eta}_h) \mapsto (\boldsymbol{\eta}_v, \mathbf{s}_Q(\boldsymbol{\eta}_v)). \tag{7.8}$$

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 $\boldsymbol{\eta}_v$ at an observed value, 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, since 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 $K\{F^{-1}(Q) \parallel P_i\}$, $Q \in D$.

2. M-step : Search for the point $\hat{P}_{i+1} \in M$ that minimizes $K(\hat{Q}_i \parallel P)$, $P \in M$.

It is clear that, when F is the identity map, 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 if, and only if, 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

$$\mathbf{s}_Q(\mathbf{r}_v) = \mathbf{a} + B\mathbf{r}_v \tag{7.9}$$

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

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

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

$$\boldsymbol{\eta}_h = E_Q[\mathbf{s}_Q(\mathbf{r}_v)] = \mathbf{s}_Q[E_Q(\mathbf{r}_v)] \tag{7.10}$$

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 $\boldsymbol{\eta}_v = \boldsymbol{\eta}_{v,0}$. 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, $0 \leq t \leq 1$, 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 $q(\mathbf{r}_h|\mathbf{r}_v)$ since the e-projection of Q_t is Q. Hence, the distribution $Q_t = \{q_t(\mathbf{r}_v, \mathbf{r}_h)\}$ is decomposed as

$$q_t(\mathbf{r}_v, \mathbf{r}_h) = p(\mathbf{r}_v; t)q(\mathbf{r}_h|\mathbf{r}_v). \tag{7.11}$$

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

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

Hence, (7.10) implies that

$$\int \mathbf{s}_Q(\mathbf{r}_v)p(\mathbf{r}_v; t)d\mathbf{r}_v = \mathbf{s}_Q\left[\int \mathbf{r}_v p(\mathbf{r}_v, t)d\mathbf{r}_v\right]. \tag{7.12}$$

We can choose any P under the condition that

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

Therefore, (7.12) immediately shows that $\mathbf{s}_Q(\mathbf{r}_v)$ is a linear function of $\mathbf{r}_v$ as in (7.9).

7.3 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 19. Geometry of normal multiplication model. Let $x_1, \cdots, x_T$ be independent normal random variables subject to $N(u, u^2)$. The sufficient statistics are $\bar{\mathbf{r}} = (r_v, r_h)$

$$\begin{aligned} r_v &= \frac{1}{T}\sum x_t, \\ r_h &= \frac{1}{T}\sum x_t^2, \end{aligned}$$

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 imbedded in $S = \{N(\mu, \sigma^2)\}$ given in Example 8. The corresponding $\boldsymbol{\eta}$-coordinates are

$$\begin{aligned} \eta_v = &\quad \mu &&= u, \\ \eta_h = &\quad \mu^2 + \sigma^2 &&= 2u^2. \end{aligned}$$

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

$$D = \{N(\mu, \sigma^2) | \quad \mu = \bar{r},\ \sigma^2 \text{ is arbitrary } \}$$

(see Fig.9). 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

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

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 reduces to $\sigma^2/2$. Obviously, x_2 has the same conditional distribution, although x_1 and x_2 are not conditionally indepenent. Hence, the conditional expectation of x_t^2 is

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

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

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

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

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

when $T = 2$, and since $\theta_h^* = \theta_h^P = -u^{-2}$ 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

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

since

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

This is confirmed from

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

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

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

that is the best candidate

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

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

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

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

while the EM-algorithm gives

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

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 differenece practically.

8 The EM and em Algorithms are Equivalent in Most Cases

8.1 Which is more natural, the EM or em?

When S is not the identity map, the EM- and em-algorithms are different, giving different estimates $\hat{\mathbf{u}}$. Which is more natural and which is better? Since the EM-algorithm gives the maximum likelihood estimator, and since it is known that the bias-corrected m.l.e. is higher-order efficient (Amari, 1995), the EM algorithm might look more natural from the statistical inference point of view. However, when we use a neural network model to approximates a given input-output relation, it is not a statistical inference. The

problem is to realize a neural network which approximate a given input-output behavior as good 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. Therefore, the em-algorithm that minimizes the divergence $K(D \parallel M)$ seems to give a more natural answer.

This becomes clearer when we consider the case where the data submanifold D is not m-flat, that is, not linear in the $\boldsymbol{\eta}$-coordinates. This is the case when the sufficient statistics

$$\mathbf{r} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h) \tag{8.1}$$

is nonlinear in $\mathbf{s}_h$. The data submanifold

$$D = \{\boldsymbol{\eta} |\ \boldsymbol{\eta} = \mathbf{r}(\mathbf{s}_v, \mathbf{s}_h),\ \mathbf{s}_v \text{ is fixed and } \mathbf{s}_h \text{ is free}\}$$

is then not m-flat, that is, it is not a linear submanifold in $\boldsymbol{\eta}$. The em-algorithm works well in this case, since the e-projection of P gives the point $\hat{Q}^*$ in D that minimizes $K(Q \parallel P)$, $Q \in D$. On the other hand, the point $\hat{Q}$ given by the conditional expectation

$$\hat{\boldsymbol{\eta}} = E[\mathbf{r}(\mathbf{s}_v, \mathbf{s}_h) | \mathbf{s}_v; P] \tag{8.2}$$

does not belong to D, that is the $\hat{Q}$ does not satisfy the constraint due to the observation of $\mathbf{s}_v$.

If we use the conditional expectation

$$\hat{\mathbf{s}}_h = E[\mathbf{s}_h | \mathbf{s}_v; P] \tag{8.3}$$

and define the point $\hat{Q}$ by

$$\hat{\boldsymbol{\eta}} = \mathbf{r}(\mathbf{s}_v, \hat{\mathbf{s}}_h), \tag{8.4}$$

the point $\hat{Q}$ belongs to D. However, the convergence to the m.l.e. is no more guaranteed.

We have already shown that the e-projection $\hat{Q}^*$ of P to D is the point whose conditional probability, and hence the conditional expectation, are the same as those of P, provided D is m-flat, that is, linear in the $\boldsymbol{\eta}$-coordinates. Usually, the conditional expectation is easily calculated, so that the e-projection is given through the conditional expectation. When D is curved, let D^* be an m-flat submanifold which is tangent to D at Q^*. Then, the conditional distribution at P conditioned on D^* is the same as that at Q^*. In this sense, the EM algorithm can be regarded as the linear approximation of the em-algorithm.

8.2 Linearization trick guaranteeing the equivalence of the EM and em algorithms

The two algorithms are equivalent in most cases. We show why they are equivalent. Let y be a random variable taking on the valules 0 and 1. Then, any function $f(y)$ is linear in y, since we have

$$f(y) = \{f(1) - f(0)\}y + f(0). \tag{8.5}$$

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 $\mathbf{m}$,

$$\mathbf{k}(\mathbf{r}) = \{k_{\mathbf{m}}(\mathbf{r}),\ \mathbf{m} \in K\} \tag{8.6}$$

by

$$k_{\mathbf{m}}(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{m}). \tag{8.7}$$

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

$$f(\mathbf{r}) = \sum_{\mathbf{m}} f(\mathbf{m}) k_{\mathbf{m}}(\mathbf{r}). \tag{8.8}$$

This shows that, when the visible random variable is binary, or when the visible random variable taking on a finite set is represented in the trivially 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.

Theorem 5. 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.

Proof. In the case of the binary stochastic perceptron and the mixture of binary expert nets, the component manifold S_t is the set of conditional probabilities $p(y, z|\mathbf{x}_t)$ or $p(y, \mathbf{z}, |\mathbf{x}_t)$. The visible variable is y in the both cases, so the $s_Q(y)$ is automatically linear in y. Hence the two algorithms are equivalent. Indeed, in the case of the mixture of binary experts, when y_t is observed in S_t, we have the new $\tilde{\boldsymbol{\eta}}$-coordinates

$$\tilde{\eta}_1 = \eta_1, \ \tilde{\eta}_{2,i} = y\eta_{3,i} - \eta_{2,i} = 0$$

which are the visible parts and

$$\tilde{\eta}_{3,i} = \eta_{3,i}$$

which are hidden parts. The corresponding random variables are

$$\tilde{r}_1 = y, \quad \tilde{r}_{2,i} = yr_{3,i} - r_{2,i} = 0$$

and

$$\tilde{r}_{3,i} = \delta_i(z).$$

Hence, the visible variable is y. In the case of the Boltzmann machine, the visible state is $\mathbf{x}^V$ of which components are binary. The visible variables are

$$r^V_{i_1 \cdots i_k} = x^V_{i_1} \cdots x^V_{i_k}$$

and their expectations give $\boldsymbol{\eta}^V$. Since $r^V_{i_1 \cdots i_k}$ are binary, the equivalence of the two algorithms is shown.

We next prove the case of the normal mixture model. For each component S_t in $S^* = S_1 \times \cdots \times S_T$, the observed data is x_t and z_t is missing. Hence, the data submanifold

D_t is given by

$$D_t = \{\boldsymbol{\eta} \mid \hat{\eta}_0 = x_t,\ \hat{\eta}_{1i} = \alpha_i^t,\ \hat{\eta}_{2i} = x_t\alpha_i^t\}, \tag{8.9}$$

where α_i is the parameter corresponding to $\delta_i(z_t)$. For a candidate point $\hat{P} \in M$, the conditional expectation of the missing variable $\delta_i(z_t)$ is

$$\begin{aligned} \alpha_i^t(x_t) &= E_{\hat{P}}[\delta_i(z_t) \mid x_t] \\ &= \frac{p_i \exp\{-\frac{1}{2}(x_t - \mu_i)^2\}}{\sum_j p_j \exp\{-\frac{1}{2}(x_t - \mu_j)^2\}}. \end{aligned} \tag{8.10}$$

This is the same as $E_Q[\delta_i(z_t)|x_t]$, where Q is the e-projection of $\hat{P}$ to D_t. When a distribution Q belongs to D_t, its $\boldsymbol{\eta}$-coordinates satisfy (8.8). This shows that any distribution in D_t is a trivial distribution satisfying

$$\mu_0 = \mu_i = x_t. \tag{8.11}$$

Hence, from (8.9)

$$E_Q E_Q[\delta_i(z_t)|x_t] = E_Q[\delta_i(z_t)] = \alpha_i$$

which does not depend on x_t. Hence, $s_Q(x_t)$ is linear (a constant in this case) and F is the identity map in the normal mixture model, providing the equivalence of the algorithms. We can also prove the equivalence in the case of the mixture of analog experts of (2.27) in a quite similar way.

8.3 Linearization in the function space

When $\mathbf{r}$ takes real vector values, the linearization trick still works,

$$f(\mathbf{r}) = \int f(\mathbf{m})\delta(\mathbf{r} - \mathbf{m})d\mathbf{m}. \tag{8.12}$$

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

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

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

$$p_{\text{emp}}(\mathbf{r}) = \frac{1}{T}\sum_{t=1}^{T}\delta(\mathbf{r} - \mathbf{r}_t). \tag{8.13}$$

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 Csiszár 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

$$M = \{p(\mathbf{r}_v, \mathbf{r}_h; \mathbf{u})\} \tag{8.14}$$

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

$$p_{\text{emp}}(\mathbf{r}_v) = \frac{1}{T}\sum_{t=1}^{T}\delta(\mathbf{r}_v - \mathbf{r}_{v,t}). \tag{8.15}$$

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

$$p(\mathbf{r}_v, \mathbf{r}_h) = p(\mathbf{r}_v)p(\mathbf{r}_h|\mathbf{r}_v). \tag{8.16}$$

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

$$D = \{p_{\text{emp}}(\mathbf{r}_v)p(\mathbf{r}_h|\mathbf{r}_v)\}. \tag{8.17}$$

The e-projection of $P = p(\mathbf{r}_v, \mathbf{r}_h; \mathbf{u})$ to D is the one that minimizes $K[q(\mathbf{r}_v, \mathbf{r}_h) \parallel p(\mathbf{r}_v, \mathbf{r}_h; \mathbf{u})]$, $q(\mathbf{r}_v, \mathbf{r}_h) \in D$. It is easy to show that the e-projection Q^* is given by

$$q^*(\mathbf{r}_v, \mathbf{r}_h) = p_{\text{emp}}(\mathbf{r}_v)p(\mathbf{r}_h|\mathbf{r}_v; \mathbf{u}) \tag{8.18}$$

when $T = 1$, and by

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

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 of the E-step is

$$\begin{aligned} E[\delta(\mathbf{r}_h - \mathbf{m})|\mathbf{r}_v; \mathbf{u}] &= \int \delta(\mathbf{r}_h - \mathbf{m}) p(\mathbf{m}|\mathbf{r}_v; \mathbf{u}) d\mathbf{m} \\ &= p(\mathbf{r}_h|\mathbf{r}_v; \mathbf{u}), \end{aligned} \tag{8.19}$$

giving the conditional probability. Hence, the two algorithms are the same, as is shown by Csiszár and Tusnády [1985] and Neal and Hinton [1993].

Theorem 6. 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

$$\begin{aligned} &E_{Q_i}[\log p(\mathbf{r}_h|\mathbf{r}_v; \mathbf{u})] \\ &\quad = \sum_{t=1}^{T} \int p(\mathbf{r}_{h,t}|\mathbf{r}_{v,t}; \hat{\mathbf{u}}_{(i)}) \log p(\mathbf{r}_{h,t}|\mathbf{r}_{v,t}; \mathbf{u}) d\mathbf{r}_{h,t}, \end{aligned} \tag{8.20}$$

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

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

8.4 Asymptotic equivalence of EM and em algorithms

When the number T of observations is large, the linear approximation holds asymptotically well, because of the law of large numbers. This is guaranteed by the large deviation theory in the case of the approximation problem or by the central limit theorem when the model is faithful. We show the following theorem.

Theorem 7. When the number T of observations is large, the EM- and em-algorithms are asymptotically equivalent.

Proof. We prove this only in the case of i.i.d. observations where observations are summarized in the arithmetic mean $\bar{\mathbf{r}} = (\bar{\mathbf{r}}_v, \bar{\mathbf{r}}_h)$. Let P be a point in M and let Q be a point in D which might not be m-flat. Let Q^* be the e-projction of P to D that minimizes $K(Q \parallel P)$, $Q \in D$. Let Q be a point in D whose $\boldsymbol{\eta}$-coordinates are denoted by $\boldsymbol{\eta}^Q = (\boldsymbol{\eta}_v^Q, \boldsymbol{\eta}_h^Q)$. By virtue of the large deviation theory (see, e.g., Saulis and Statulevicius, 1989), the probability density of $\bar{\mathbf{r}} = \boldsymbol{\eta}_Q$, that is, the probability of point Q being observed when the true distribution is P, is written as

$$p(\bar{\mathbf{r}}; P) \approx \exp\{-TK(Q \parallel P)\}, \tag{8.21}$$

asymptotically. Hence, the conditional distribution of $\bar{\mathbf{r}}_h = \boldsymbol{\eta}_{Q,h}$ is

$$p(\bar{\mathbf{r}}_h \mid \bar{\mathbf{r}}_v; P) \approx \exp[-T\{K(Q \parallel P) - K(Q^* \parallel P)\}], \tag{8.22}$$

because of

$$\begin{aligned} p(\bar{\mathbf{r}}_h; P) &= \int \exp\{-TK(Q \parallel P)\} d\bar{\mathbf{r}}_h \\ &\approx \exp\{-T(Q^* \parallel P)\}. \end{aligned}$$

This is because $K(Q \parallel P)$, $Q \in D$ defined by $\bar{\mathbf{r}}_v$, is minimized at Q^* and the so-called saddle point approximation (the Laplce method) is used. Hence, the conditional distribution (8.22) is concentrated at around Q^*, and the conditional expectation is

$$E[\bar{\mathbf{r}}_h \mid \bar{\mathbf{r}}_v; P] \approx \boldsymbol{\eta}_{Q,h}. \tag{8.23}$$

Hence, the conditional expectation gives asymptotically the same point as the e-projection.

The large deviation theory is used in the approximation problem where D is not necessarily close to M. When there exists a $P \in M$ from which the data $\mathbf{r}_1, \cdots, \mathbf{r}_T$ are generated, M and D are very close as $T \to \infty$. In this case, we can directly apply the central limit theorem to prove the result. In the non-i.i.d. case where the data submanifold is D_T^* in S_T^*, the joint probability density of $(\mathbf{r}_1, \cdots, \mathbf{r}_T)$ is written as

$$p(\mathbf{r}_1, \cdots, \mathbf{r}_T | P^*) = \exp\left\{-\sum_{t=1}^{T} K(Q_t \parallel P_t) - \sum_{t=1}^{T} H(Q_t)\right\}, \tag{8.24}$$

where $Q_T^* = (Q_1, \cdots, Q_T)$ is the point in D_T^* whose $\boldsymbol{\eta}$-coordinates are given by $\mathbf{r}_t$, $P^* = (P_1, \cdots, P_T) \in M \subset S_T^*$ and $H(Q_t)$ is the entropy of the distribution at Q_t. In the i.i.d. case, the entropy term is reduced when we transform the density $(\mathbf{r}_1, \cdots, \mathbf{r}_T)$ to their arthmetic mean $\bar{\mathbf{r}}$. In the present case, we need to apply the non-i.i.d. version of the theory of large deviations. In the general case, the e-projection might not be the same as the conditional expectation. However, the difference is asymptotically absorbed in the next M-step. This suggests a proof of the theorem in the general case, altough it might not be easy to construct a rigorous proof.

9 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 $\mathbf{r}_{v,t}$, $t = 1, \cdots, T$ 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 is summarized in a single observed point $\hat{\boldsymbol{\eta}}$. 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 (4.29). Let $\hat{\mathbf{u}}_t$ be the estimator at time t, and let $\hat{\boldsymbol{\eta}}_t$ be the guessed 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 $\mathbf{r}_{t+1}$ conditioned on $\mathbf{s}_{v,t+1}$ based on $\hat{\mathbf{u}}_t$). This gives a guess of the unobserved $\hat{\mathbf{r}}_{t+1}$. Then, modify the guessed observed point by

$$\hat{\boldsymbol{\eta}}_{t+1} = (1 - \varepsilon_t)\hat{\boldsymbol{\eta}}_t + \varepsilon_t \hat{\mathbf{r}}_{t+1}, \tag{9.1}$$

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 $\hat{P}_{t+1}$.

Since 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

$$\frac{\partial l}{\partial \mathbf{u}} = \mathbf{B}\{(\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\mathbf{u}))\}, \tag{9.2}$$

where $\mathbf{B} = \dfrac{\partial \boldsymbol{\theta}}{\partial \mathbf{u}}$ is a matrix. Hence, the gradient method applied to the incremental M-step is

2) incremental M-step

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \varepsilon_t \mathbf{B}(\hat{\mathbf{u}}_t)\{\hat{\boldsymbol{\eta}}_{t+1} - \boldsymbol{\eta}(\hat{\mathbf{u}}_t)\}. \tag{9.3}$$

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 $\hat{\boldsymbol{\eta}}$, we propose the following learning algorithm.

1) e-step (E-step)

Calculate the e-projection of P_t with the coordinates $\boldsymbol{\theta}(\mathbf{x}_{t+1}, \hat{\mathbf{u}}_t)$ 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 $\mathbf{s}_{v,t}$ 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 $\hat{\mathbf{r}}_1, \cdots, \hat{\mathbf{r}}_t$ may be discarded.

10 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,

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

or

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

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 $K(Q \parallel P)$ or $K[F^{-1}(Q) \parallel P]$. Let $Q \in D$ and $P \in M$ be a pair of points which move in the directions of decreasing $K(Q \parallel P)$, respectively. Then, we have the dual gradient flows $Q(t)$ and $P(t)$ (Fig.10). 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 a parameter of the dual gradient curves. The gradient flows are given in these coordinates by

$$\frac{d}{dt}\mathbf{u}(t) = -\varepsilon G_M^{-1} \frac{\partial}{\partial \mathbf{u}} K(\boldsymbol{\eta}_Q \parallel \boldsymbol{\theta}_P), \tag{10.1}$$

$$\frac{d}{dt}\boldsymbol{\eta}_{Q,h}(t) = -\varepsilon G_D^{-1} \frac{\partial}{\partial \boldsymbol{\eta}_h} K(\boldsymbol{\eta}_Q \parallel \boldsymbol{\theta}_P). \tag{10.2}$$

Here G_M and G_D are the Fisher information matrices of M and D, respectively. It is easy to prove the relations

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}_P} K(Q \parallel P) &= \boldsymbol{\eta}_P - \boldsymbol{\eta}_Q, \\ \frac{\partial}{\partial \boldsymbol{\eta}_Q} K(Q \parallel P) &= \boldsymbol{\theta}_Q - \boldsymbol{\theta}_P.\end{aligned}$$

Hence, the gradient flows are written as

$$\frac{d}{dt}\mathbf{u}(t) = -\varepsilon G_M^{-1} \frac{\partial \boldsymbol{\theta}(\mathbf{u})}{\partial \mathbf{u}} (\boldsymbol{\eta}_P(t) - \boldsymbol{\eta}_Q(\mathbf{u}(t))), \tag{10.3}$$

$$\frac{d}{dt}\boldsymbol{\eta}_{Q,h}(t) = -\varepsilon G_D^{-1} (\boldsymbol{\theta}_{Q,h}(\mathbf{u}(t)) - \boldsymbol{\theta}_{P,h}(t)), \tag{10.4}$$

where $\boldsymbol{\theta}_P$ and $\boldsymbol{\eta}_Q$ are $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$-coordinates of P and Q, respectively. The flows are e- and m-geodesic flows in D and M, respectively (Fujiwara and Amari, 1993).

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

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \varepsilon_t \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{u}} \{\hat{\boldsymbol{\eta}}_t - \boldsymbol{\eta}(\hat{\mathbf{u}}_t)\}, \tag{10.5}$$

$$\hat{\boldsymbol{\eta}}_{h,t+1} = \hat{\boldsymbol{\eta}}_{h,t} + \varepsilon'_t \{\boldsymbol{\theta}_h(\hat{\mathbf{u}}_t) - \boldsymbol{\theta}_h(\hat{\boldsymbol{\eta}}_t)\}. \tag{10.6}$$

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. Thus a unified geometrical framework is given to the EM and em algorithms. This makes it possible to formulate a learning algorithm version of the EM algorithm. The Boltzmann machine, mixture of experts, stochastic multilayer perceptron and normal mixture models are treated as examples. This framework opens a new area of research connecting neural networks, statistics and geometry.

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 and Rice, 1993, etc.) in order 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, 1987; Ohara and Amari, 1992), information theory (Amari and 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 and 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 about 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 are 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 and mixture straightness (geodesic) in an intuitive way. See §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 $\{0, 1, \cdots, n\}$. 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

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

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 curved 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

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

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

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

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

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

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 curved 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 itself 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 $\boldsymbol{\eta}_2$ are 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. Since 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. An exponential family is flat from the both criteria so that it is called a dually flat manifold. A dually flat manifold has deep differential geometrical structures.

A.2 Orthogonality and Fisher information

We explain the tangent space and the Riemannian metric in order to define the orthogonality of two curves. Let $\mathbf{e}_i (i = 1, \cdots, n)$ be the tangent vector along the coordinate curve θ_i, that is, the direction in which θ_i changes but no other θ_j $(j \neq i)$ change (Fig.11). (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 $\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$. 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 $\{\mathbf{e}_1^*, \cdots, \mathbf{e}_n^*\}$.

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

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

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

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

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

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

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

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

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 vectors 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

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

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

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

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 representing them in the dual bases as

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

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 (2.2) 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 and 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 Levi-Civita connection has non-zero 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

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

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

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

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 (Fig.12) (Nagaoka and Amari, 1981; 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.

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 (Fig.13). 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 connecting 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

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

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. Since $Q \in D$, 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 $\boldsymbol{\theta}_h^* = \boldsymbol{\theta}_h^P$. In order to obtain $\boldsymbol{\eta}^*$ or $\boldsymbol{\theta}^*$, we need to solve the following equations

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

or

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

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 manifold S, we have already defined the Riemannian metric by (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. Since $\nabla_{\mathbf{e}_i}\mathbf{e}_j$ 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 $\boldsymbol{\theta} = \boldsymbol{\theta}(t)$ is given by

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

or

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

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

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

The conservation of metric implies that

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

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,

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

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 $\mathbf{e}_i = \dfrac{\partial}{\partial\theta^i}$ and $\mathbf{e}^*_i = \dfrac{\partial}{\partial\eta^i}$ 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 and Kawanabe, 1994), and the conformal structure (Okamoto, Amari and 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 and Amari, 1994).

References

Aarts, E. H. L. and Korst, J. H. M. (1989). *Simulated Annealing and Boltzmann Machines*, Chichester: Wiley.

Ackley, D. , Hinton, G. and Sejnowski, T. (1985). A learning algorithm for Boltzmann machines, *Cognitive Science*, **9**, 147–169.

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**, pp53–82.

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 Cal: IMS.

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

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

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

Amari, S. (1994). The EM algorithm and information geometry in neural network learning, *Neural Computation*, to appear.

Amari, S., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L. and Rao, C. R. (1987). *Differential Geometry in Statistical Inferences*, IMS Lecture Notes Monograph Series, Vol.**10**, Hayward, California: IMS.

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

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

Amari, S. and Han, T. S. (1989). Statistical inference under multi-terminal rate restrictions — a differential geometrical approach, *IEEE Trans. on Information Theory*, **IT-35**, pp.217–227.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*, John Wiley: Chichester.

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. and 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 and Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, **41**, 164–171.

Besag, J. and 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 Trans. Neural Networks*, **3**, pp.612–620.

Chentsov, N. N. (1972). *Statistical Decision Rules and Optimal Inference* (in Russian), Moscow:Nauka, translated in English (1982), Rhode Island: AMS.

Cox, D. R. and 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. and Tusnády, G. (1984). Information geometry and alternating minimization procedures, In E. F. Dedewicz, et al. eds., *Statistics and Decisions*, Supplementary Issue, No.1, 205–237, Munich: Oldenburg Verlag.

Dempster, A. P., Laird, N. M. and Rudin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statist. Soc.*, **B-39**, 1–38.

Fujiwara, A. and Amari, S. (1994). Dualistic dynamical systems in the framework of information geometry, preprint.

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

Günsch, H. , Geman, S. and Kehagias, A. (1994). Hidden Markov random fields, to appear.

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

Jordan, M. I. and Jacobs, R. A. (1994). Higherarchical mixtures of experts and the EM-algorithm, *Neural Computation*, to appear.

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

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

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

Kawanabe, M. and Amari, S. (1994). Estimation of network parameters in semiparametric stochastic perceptron, *Neural Computation*, to appear.

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

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

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

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

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

Nomizu, K. and Simon U. (1991). Notes on Conjugate Connections, preprint.

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

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

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters, *Bull. Calcutta. Math. Soc.*, **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 clasification, *Journal of Royal Statistical Society*, **B56**, to appear.

Saulis, L. and Statulevicius, V. A. (1989). Limit Theorems for large deviations, Dordrecht: Kluwer Academic.

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

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

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

Zhuang, J. and Amari, S. (1993). Piecewise-linear division of signal space by a multilayer neural networks with the maximum detector (in Japanese), *Trans. Inst. Electr. Inf. Comm. Eng.*, **J76-D**, 1435–1440.

Figure Legends

Fig.1 Manifold of probability distributions

Fig.2 Mixture of expert nets

Fig.3 Curved exponential family

Fig.4 Stochastic perceptron

Fig.5 Maximum likelihood estimation

Fig.6 Data submanifold D : $\boldsymbol{\eta}$ = fixed

Fig.7 EM algorithm

Fig.8 em algorithm

Fig.9 Example where EM and em algorithms are different

Fig.10 Dual gradient flows

Fig.11 Tangent space

Fig.12 Generalied Pythagoras theorem

Fig.13 m-projection of P to M