Information Geometry of α-Projection in Mean Field Approximation

Shun-ichi Amari*, Shiro Ikeda† and Hidetoshi Shimokawa‡

Abstract

Information geometry is applied to mean field approximation for elucidating its properties in the spin glass model or the Boltzmann machine. The α-divergence is used for approximation, where α-geodesic projection plays an important role. The naive mean field approximation and TAP approximation are studied from the point of view of information geometry, which treats the intrinsic geometric structures of a family of probability distributions. The bifurcation of the α-projection is studied, at which the uniqueness of the α-approximation is broken.

1 Introduction

Mean field approximation uses a simple tractable family of probability distributions to calculate quantities related to a complex probability distribution including mutual interactions. Information geometry, on the other hand, studies intrinsic geometrical structure existing in the manifold of probability distributions (Chentsov, 1982; Amari, 1985; Amari and Nagaoka, 2000). It was used for analyzing performances of learning in the Boltzmann machine (Amari, Kurata and Nagaoka, 1992), the EM algorithm (Amari, 1995), multilayer perceptrons (Amari, 1998), etc. A number of new works appeared recently

*RIKEN, BSI
†PRESTO, JST
‡Univ. of Tokyo

which treated mean field approximation from the point of view of information geometry (Tanaka [this volume], Tanaka [2000], Kappen [this volume], and Bhattacharyya and Keerthi [1999, 2000]).

The present article studies the relation between mean field approximation and information geometry in more detail. We treat a simple spin model like the SK model or the Boltzmann machine, and study how the mean values of spins can be approximated. It is known (Tanaka, this volume) that, given a probability distribution including complex mutual interactions of spins or neurons, the mean value of each spin is kept constant when it is projected by the m-geodesic to the submanifold consisting of independent probability distributions. On the other hand, the m-projection is computationally intractable for a large system. Instead, its projection by the e-geodesic is easy to calculate. However, the mean value is changed by this, so that only an approximate value is obtained. This gives the naive mean field approximation.

A family of divergence measures named the α-divergence is defined invariantly in the manifold of probability distributions (Amari, 1985; Amari and Nagaoka, 2000). The $\alpha = -1$ divergence is known as the Kullback-Leibler divergence or cross entropy, $\alpha = 1$ as the reverse of the K-L divergence, and $\alpha = 0$ as the Hellinger distance. This concept is closely related to the Rényi entropy (R'enyi, 1961; see also Chernoff, 1952; the f-divergence of Csiszár, 1975). These divergence functions give a unique Riemannian metric to the manifold of probability distributions. It moreover gives a family of invariant affine connections named the α-connections where α- and $-\alpha$-affine connections are dually coupled to each other with respect to the Riemannian metric (Nagaoka and Amari, 1982; Amari, 1985; Amari and Nagaoka, 2000). The α-geodesic is defined in this context. It should be remarked that the Tsallis entropy (Tsallis, 1988) is closely connected to the α-geometry.

We use the α-geodesic projection to elucidate various mean field approximations. The α-projection is the point in the tractable subspace consisting of independent distributions that minimizes the α-divergence from a given true distribution to the subspace. It is given by the α-geodesic that is orthogonal to the subspace at the α-projected point. This gives

a family of the α-approximations, where $\alpha = -1$ is the distribution giving the true mean value and $\alpha = 1$ is the naive mean approximation. Therefore, it is interesting to know how the α-approximation depends on α. We also study the TAP approximation in this framework.

We can prove that the m-projection ($\alpha = -1$) is unique, while α-projection ($\alpha \neq -1$) is not necessarily so. The e-projection ($\alpha = 1$), that is the naive mean field approximation, is not necessarily uniquely solved. Therefore, it is interesting to see how the α-approximation bifurcates depending on α. We calculate the Hessian of the α-approximation which shows the stability or the local minimality of the projection.

2 Geometry of Mean Field Approximation

Let us consider a system of spins or Boltzmann machine, where $\boldsymbol{x} = (x_1, \cdots, x_n)\,; x_i = \pm 1$ denotes the values of n spins. The equilibrium probability distribution is given by

$$q(\boldsymbol{x}\,;\, W, \boldsymbol{h}) = \exp\left\{W \cdot X + \boldsymbol{h} \cdot \boldsymbol{x} - \psi_q\right\}, \tag{1}$$

where $W = (w_{ij})$ and $X = (x_i x_j)$ are symmetric matrices and

$$W \cdot X = \frac{1}{2}\sum w_{ij} x_i x_j \quad (w_{ii} = 0)\,, \tag{2}$$

$$\boldsymbol{h} \cdot \boldsymbol{x} = \sum h_i x_i. \tag{3}$$

Here, W denotes the mutual interactions of the spins, $\boldsymbol{h}$ the outer field, and $e^{-\psi_q}$ is the normalization constant, $Z = e^{\psi_q}$ is the partition function, and $\psi_q = \psi_q(W, \boldsymbol{h})$ is called the free energy in physics or the cumulant generating function in statistics.

Let $\boldsymbol{S}$ be the family of probability distributions of the above form (1), where $(W, \boldsymbol{h})$ forms a coordinate system to specify each distribution in $\boldsymbol{S}$. Let E_q be the expectation operator with respect to q. Then, the expectations of X and $\boldsymbol{x}$,

$$K_q = E_q[X] = \left(E_q\left[x_i x_j\right]\right), \quad \boldsymbol{m}[q] = E_q[\boldsymbol{x}] = \left(E_q\left[x_i\right]\right). \tag{4}$$

form another coordinate system of $\boldsymbol{S}$. Our theory is based on information geometry and is applicable to many other general cases, but for simplicity's sake, we stick on this simple problem of obtaining a good approximation of $\boldsymbol{m}[q]$ for a given q.

Let us consider the subspace $\boldsymbol{E}$ of $\boldsymbol{S}$ such that x_i's are independent or $W = 0$. A distribution $p \in \boldsymbol{E}$ is written as

$$p(x, \boldsymbol{h}) = \exp\left\{\boldsymbol{h} \cdot \boldsymbol{x} - \psi_p\right\}. \tag{5}$$

This is a submanifold of $\boldsymbol{S}$ specified by $W = 0$, and $\boldsymbol{h}$ is its coordinates. The expectation

$$\boldsymbol{m} = E_p[\boldsymbol{x}] \tag{6}$$

is another coordinate system of $\boldsymbol{E}$.

Physicists know that it is computationally difficult to calculate $\boldsymbol{m}[q]$ from $q(\boldsymbol{x}, W, \boldsymbol{h})$. It is given by

$$\boldsymbol{m}[q] = \frac{\partial}{\partial \boldsymbol{h}} \psi_q(W, \boldsymbol{h}) \tag{7}$$

but the partition function $Z_q = e^{-\psi_q}$ is difficult to obtain when the system size is large. On the other hand, for $p \in \boldsymbol{E}$, it is easy to obtain $\boldsymbol{m} = E_p[\boldsymbol{x}]$ because x_i are independent. Hence, the mean field approximation tries to use quantities obtained in the form of expectation with respect to some relevant $p \in \boldsymbol{E}$.

Physicists established the method of approximation, called the mean field theory, including the TAP approximation. The problem is formulated more naturally in the framework of information geometry (Tanaka [this volume], Kappen [this volume] and Bhattacharyya and Keerthi [2000]). The present paper tries to give another way to elucidate this problem by information geometry of the α-connections introduced by Nagaoka and Amari [1982], Amari [1985] and Amari and Nagaoka [2000].

3 Concepts from Information Geometry

Here, we introduce some concepts of information geometry without entering in details. Let y be a discrete random variable taking values on a finite set $\{0, 1, \cdots, N-1\}$, and let $p(y)$ and $q(y)$ be two probability distributions. In the case of spins, y represents $2^n \boldsymbol{x}$'s where $N = 2^n$.

α-divergence

The α-divergence from q to p is defined by

$$D_\alpha[q\ :\ p] = \frac{4}{1-\alpha^2}\left(1 - \sum_y q^{\frac{1-\alpha}{2}} p^{\frac{1+\alpha}{2}}\right), \quad \alpha \neq \pm 1 \tag{8}$$

and for $\alpha = \pm 1$

$$D_{-1}[q:p] = \sum q \log \frac{q}{p}, \tag{9}$$

$$D_1[q:p] = \sum p \log \frac{p}{q}. \tag{10}$$

The latter two are the Kullback-Leibler divergence and its reverse. When $\alpha = 0$, it is the Hellinger distance,

$$D_0[q:p] = 2\sum \left(\sqrt{p} - \sqrt{q}\right)^2. \tag{11}$$

The divergence satisfies

$$D_\alpha[q:p] \geq 0, \tag{12}$$

with equality when and only when $q = p$. However, it is not symmetric except for $\alpha = 0$, and it satisfies

$$D_\alpha[q:p] = D_{-\alpha}[p:q]. \tag{13}$$

The α-divergence may be calculated in the following way. Let us consider a curve of probability distributions parameterized by t,

$$p(y,t) = e^{-\psi(t)} q^{\frac{1-t}{2}} p^{\frac{1+t}{2}} = \exp\left\{\frac{1}{2}\log pq + \frac{t}{2}\log\frac{p}{q} - \psi(t)\right\}, \tag{14}$$

which is an exponential family connecting q and p. Here, $e^{-\psi(t)}$ is the normalization constant. We then have

$$D_\alpha[q\ :\ p] = \frac{4}{1-\alpha^2}\left(1 - e^{\psi(\alpha)}\right), \quad \alpha \neq \pm 1. \tag{15}$$

We also have

$$\psi'(\alpha) = \frac{1}{2} E_\alpha\left[\log\frac{p}{q}\right], \tag{16}$$

$$\psi''(\alpha) = E_\alpha\left[\frac{1}{4}\left(\log\frac{p}{q}\right)^2\right] - \left\{\psi'(\alpha)\right\}^2, \tag{17}$$

$$\psi'''(\alpha) = \frac{1}{8} E_\alpha\left[\left\{\left(\log\frac{p}{q}\right) - E_\alpha\left[\log\frac{p}{q}\right]\right\}^3\right], \tag{18}$$

where E_α denotes expectation with respect to $p(y, \alpha)$. For $\alpha = \pm 1$, $\psi(\pm 1) = 0$. By taking the limit, we have

$$D_\alpha = 2\alpha\psi'(\alpha), \quad \alpha = \pm 1. \tag{19}$$

The family of the α-divergences gives invariant measures provided by information geometry.

4 Parametric model and Fisher information

When $p(y)$ is specified by a coordinate system ξ, it is written as $p(y, \xi)$. The $N-1$ quantities

$$p_i = \text{Prob}\{y = i\}, \quad i = 1, \cdots, N-1 \tag{20}$$

form coordinates of $p(y)$. There are many other coordinate systems. For example

$$\theta_i = \log \frac{p_i}{p_0} \tag{21}$$

is another coordinate system. Let

$$\delta_i(y) = \begin{cases} 1, & y = i, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Then, we have

$$p(y, \xi) = \sum p_i \delta_i(y) + p_0 \delta_0(y), \tag{23}$$

$$p(y, \theta) = \exp\left\{ \sum \theta_i \delta_i(y) + \log p_0 \right\}. \tag{24}$$

The α-divergence for two nearby distributions $p(y, \xi)$ and $p(y, \xi + d\xi)$ is expanded as

$$D_\alpha \left[p(y, \xi) : p(y, \xi + d\xi) \right] = \frac{1}{2} \sum_{i,j} g_{ij}(\xi) d\xi_i d\xi_j, \tag{25}$$

where the right-hand side does not depend on α. The matrix $G(\xi) = (g_{ij}(\xi))$ is positive-definite and symmetric, given by

$$g_{ij}(\xi) = E\left[\frac{\partial \log p(y, \xi)}{\partial \xi_i} \frac{\partial \log p(y, \xi)}{\partial \xi_j} \right]. \tag{26}$$

This is called the Fisher information. It gives the unique invariant Riemannian metric to the manifold of probability distributions.

α-projection

Let $\boldsymbol{M}$ be a submanifold in $\boldsymbol{S}$. Given $q \in \boldsymbol{S}$, the point $p^* \in \boldsymbol{M}$ is called the α-projection of q to $\boldsymbol{M}$, when function $D_\alpha[q, p]$, $p \in \boldsymbol{M}$ takes a critical value at p^*, that is

$$\frac{\partial}{\partial \xi} D_\alpha \left[q : p(y, \xi) \right] = 0, \tag{27}$$

at p^* where ξ is a coordinate system of $\boldsymbol{M}$. The minimizer of $D_\alpha[q : p]$, $p \in M$, is the α-projection of q to $\boldsymbol{M}$. We denote it by

$$p^* = \prod_\alpha q. \tag{28}$$

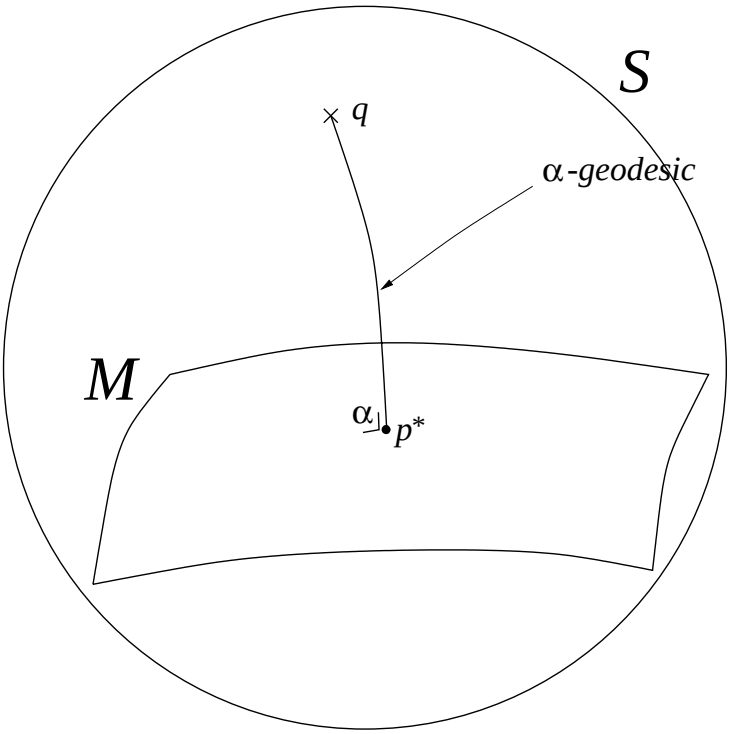

Figure 1: α-projection

In order to characterize the α-projection, we need to define the α-affine connection and α-geodesic derived therefrom. We do not explain them here (see Amari, 1985; Amari and Nagaoka, 2000). We show the following fact. See Fig.1.

Theorem 1. A point $p^* \in \boldsymbol{M}$ is the α-projection of q to $\boldsymbol{M}$, when and only when the α-geodesic connecting q and p^* is orthogonal to $\boldsymbol{M}$ in the sense of the Fisher Riemannian metric G.

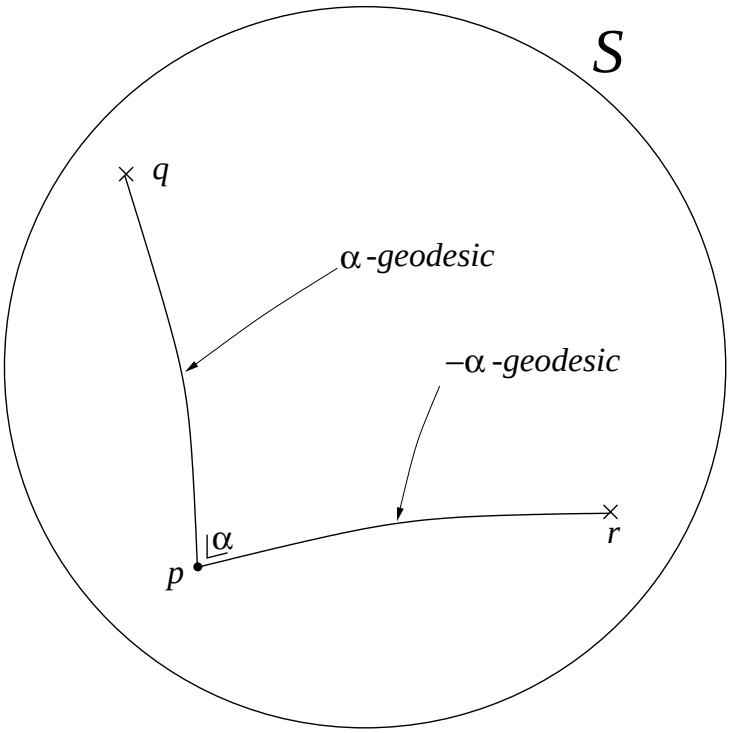

Figure 2: Pythagoras' theorem

Exponential family

A family of distributions is called an exponential family, when its probability distributions are written as

$$p(y, \theta) = \exp\left\{\sum \theta_i k_i(y) - \psi(\theta)\right\} \tag{29}$$

by using an appropriate coordinate system θ, where $\boldsymbol{k} = k_i(y)$ are adequate functions of y. The spin system or Boltzmann machine (1) is an exponential family, where

$$\theta = (W, \boldsymbol{h}) \tag{30}$$

and $\boldsymbol{k}$ consists of

$$\boldsymbol{k} = (X, \boldsymbol{x}). \tag{31}$$

The exponential family forms an $\alpha = \pm 1$ flat manifold, that is, $\alpha = \pm 1$ Riemann-Christoffel curvatures vanish identically, but this is a non-Euclidean space. There exist $\alpha = \pm 1$ affine coordinate systems in such a manifold. The above θ is an $\alpha = 1$ affine coordinate system, called the e-affine (exponential-affine), because the log probability is

linear in θ. An e-geodesic ($\alpha = 1$ geodesic) is linear in θ. More generally, for any two distributions $p(y)$ and $q(y)$, the e-geodesic connecting them is given by

$$\log p(y,t) = (1-t)\log p(y) + t \log q(y) - \psi(t). \tag{32}$$

Let us denote the expectation of $\boldsymbol{k}$ by η,

$$\eta = E[\boldsymbol{k}]. \tag{33}$$

It is known that this η forms another coordinate system of an exponential family. This is an $\alpha = -1$ affine coordinate system, or m-affine (mixture affine) coordinate system. The two coordinate systems are connected by the Legendre transformation,

$$\eta = \frac{\partial}{\partial\theta}\psi(\theta), \tag{34}$$

$$\theta = \frac{\partial}{\partial\eta}\varphi(\eta), \tag{35}$$

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

$$\psi(\theta) + \varphi(\eta) - \theta \cdot \eta = 0 \tag{36}$$

holds. Any linear curve in η is an m-geodesic.

An important property is given by the following theorem. See Fig.2.

Theorem 2. Let p, q, r be three probability distributions in an $\pm\alpha$-flat manifold $\boldsymbol{S}$. When the α-geodesic connecting p and q is orthogonal at q with respect to the Riemannian metric to the $-\alpha$-geodesic connecting q and r,

$$D_\alpha[p:q] + D_\alpha[q:r] = D_\alpha[p:r]. \tag{37}$$

From this follows

Theorem 3. Let $\boldsymbol{M}$ be a smooth submanifold in an $\pm\alpha$-flat manifold $\boldsymbol{S}$, and let p^* be the α-projection from q to $\boldsymbol{M}$. Then, the α-geodesic connecting q and p^* is orthogonal to $\boldsymbol{M}$ and vice versa.

5 Geometry of $\boldsymbol{E}$

Since $\boldsymbol{E}$ consists of all the independent distributions, it is easy to show the geometry of $\boldsymbol{E}$. Moreover, $\boldsymbol{E}$ itself is an exponential family,

$$p(\boldsymbol{x}, \bar{\boldsymbol{h}}) = \exp\left\{\bar{\boldsymbol{h}} \cdot \boldsymbol{x} - \psi_0(\bar{\boldsymbol{h}})\right\}, \tag{38}$$

where

$$e^{\psi_0(\bar{\boldsymbol{h}})} = \prod_i \left(e^{\bar{h}_i} + e^{-\bar{h}_i}\right) \tag{39}$$

or

$$\psi_0(\bar{\boldsymbol{h}}) = \sum_i \log\left(e^{\bar{h}_i} + e^{-\bar{h}_i}\right). \tag{40}$$

This $\bar{\boldsymbol{h}} = \left(\bar{h}_1, \cdots, \bar{h}_n\right)$ is the e-affine coordinates of $\boldsymbol{E}$. Its m-affine coordinates are given by

$$\boldsymbol{m} = E_p[\boldsymbol{x}] = \frac{\partial}{\partial \bar{\boldsymbol{h}}} \psi_0(\bar{\boldsymbol{h}}), \tag{41}$$

which is easily calculated as

$$m_i = E_p[x_i] = \frac{\partial}{\partial \bar{h}_i} \psi_0(\bar{\boldsymbol{h}}) = \frac{e^{\bar{h}_i} - e^{-\bar{h}_i}}{e^{\bar{h}_i} + e^{-\bar{h}_i}} = \tanh\ \bar{h}_i. \tag{42}$$

This is solved as

$$e^{\bar{h}_i} = \sqrt{\frac{1+m_i}{1-m_i}}. \tag{43}$$

In terms of $\boldsymbol{m}$, the probability is written as

$$p(\boldsymbol{x}, \boldsymbol{m}) = \prod \frac{1 + m_i x_i}{2}, \quad x_i = \pm 1. \tag{44}$$

The Riemannian metric or the Fisher information $G = (g_{ij})$ is

$$\begin{aligned} g_{ij} &= \frac{\partial m_i}{\partial \bar{h}_j} \\ &= \left(1 - m_i^2\right) \delta_{ij}. \end{aligned} \tag{45}$$

Its inverse $\bar{G} = G^{-1} = (\bar{g}_{ij})$ is

$$\bar{g}_{ij} = \left(\frac{1}{1 - m_i^2}\right) \delta_{ij}. \tag{46}$$

Let $l(\boldsymbol{x}, \boldsymbol{m}) = \log p(\boldsymbol{x}, \boldsymbol{m})$. We then have

$$\partial_{m_i} l = \frac{x_i - m_i}{1 - m_i^2}, \tag{47}$$

$$\partial_{m_i}^2 l = -\left(\partial_{m_i} l\right)^2. \tag{48}$$

6 The α-projection and mean field approximation

Given $q \in S$, its α-projection to $\boldsymbol{E}$ is given by

$$\bar{p}_\alpha = \prod_\alpha q = \arg\min_{p \in \boldsymbol{E}} D_\alpha[q \ : \ p]. \tag{49}$$

We denote by $\boldsymbol{m}_\alpha[q]$ the expectation of $\boldsymbol{x}$ with respect to $\bar{p}_\alpha$, that is $E_{\bar{p}_\alpha}[x]$. Then it is given by

$$\frac{\partial}{\partial \boldsymbol{m}} D_\alpha \left[q \ : \ p\left(\boldsymbol{x}, \boldsymbol{m}_\alpha\right)\right] = 0. \tag{50}$$

From the point of view of information geometry, $\prod_\alpha q = p\left(\boldsymbol{x}, \boldsymbol{m}_\alpha\right) \in \boldsymbol{E}$ is the α-geodesic projection of q to $\boldsymbol{E}$ in the sense that the α-geodesic connecting q and p is orthogonal to $\boldsymbol{E}$ at $p = p\left(\boldsymbol{x}, \boldsymbol{m}_\alpha\right)$.

When $\alpha = -1$, $\bar{p}_{-1}$ is the m-projection ($\alpha = -1$-projection) of q to $\boldsymbol{E}$. We have

$$\boldsymbol{m}_{-1} = \boldsymbol{m}[q] \tag{51}$$

which is the quantity we want to obtain. This relation is directly calculated by solving

$$\frac{\partial}{\partial \boldsymbol{m}} D_{-1} \left[q \ : \ p\left(\boldsymbol{m}_{-1}\right)\right] = 0, \tag{52}$$

because this is equivalent to

$$\frac{\partial}{\partial \boldsymbol{m}} \int q \log p(\boldsymbol{x}, \boldsymbol{m}) dx = 0, \tag{53}$$

or

$$\frac{\partial}{\partial \boldsymbol{m}} E_q \left[l(\boldsymbol{x}, \boldsymbol{x})\right] = 0 = \text{const} E_q \left[\boldsymbol{x} - \boldsymbol{m}\right]. \tag{54}$$

Hence,

$$\boldsymbol{m}_{-1} = E_q[\boldsymbol{x}] \tag{55}$$

which is the quantity we have searched for. But we cannot calculate $E_q[\boldsymbol{x}]$ explicitly, due to the difficulty in calculating Z or ψ_q for q.

Physicists tried to obtain $\boldsymbol{m}_{-1}$ by mean field approximation in an intuitive way. If we use the e-projection of q to $\boldsymbol{E}$ instead of the m-projection, we have the naive mean field

approximation (Tanaka, 2000). To show this, we calculate the e-projection (1-projection) of q to $\boldsymbol{E}$. For $\alpha = 1$,

$$\begin{aligned} D_1[q \ : \ p] &= D_{-1}[p : q] = E_p\left[\bar{\boldsymbol{h}} \cdot \boldsymbol{x} - \psi_p - (WX + \boldsymbol{h} \cdot \boldsymbol{x} - \psi_q)\right] \\ &= \bar{\boldsymbol{h}} \cdot \boldsymbol{m} - \psi_p - W \cdot M - \boldsymbol{h} \cdot \boldsymbol{m} + \psi_q \end{aligned} \tag{56}$$

because of $M = E_p[X] = \boldsymbol{m}\boldsymbol{m}^T$. Hence,

$$\begin{aligned} \frac{\partial D_1}{\partial \boldsymbol{m}} &= \frac{\partial \bar{\boldsymbol{h}}}{\partial \boldsymbol{m}} \frac{\partial}{\partial \bar{\boldsymbol{h}}} \left(\bar{\boldsymbol{h}} \cdot \boldsymbol{m} - \psi_p\right) - W\boldsymbol{m} - h \\ &= \tanh^{-1}\boldsymbol{m} - W\boldsymbol{m} - \boldsymbol{h}. \end{aligned} \tag{57}$$

This gives

$$\boldsymbol{m}_1[q] = \tanh\left[W\boldsymbol{m}_1[q] + \boldsymbol{h}\right], \tag{58}$$

known as the "naive" mean-field approximation. In the component form, this is

$$m_i = \tanh\left(\sum w_{ij} m_j + h_i\right). \tag{59}$$

This equation can have a number of solutions. It is necessary to check which solution minimizes $D_1[q : p]$. The minimization may be attained at the boundary of $m_i = \pm 1$.

Similarly, we have the α-projection $\boldsymbol{m}_\alpha[q]$ by solving

$$\frac{\partial}{\partial \boldsymbol{m}} D_\alpha\left[q \ : \ p(\boldsymbol{x}, \boldsymbol{m})\right] = 0. \tag{60}$$

However, it is not tractable to obtain $\boldsymbol{m}_\alpha$ explicitly except for $\alpha = 1$.

7 α-trajectory

For a fixed q, its α-projection $\boldsymbol{m}_\alpha[q]$ is considered as a path in $\boldsymbol{E}$ connecting the true $\boldsymbol{m}_{-1} = \boldsymbol{m}[q]$ and the mean-field approximation $\boldsymbol{m}_1[q]$. This is called the α-trajectory of q in $\boldsymbol{E}$. See Fig.3.

The tangent direction of the trajectory is given by $d\boldsymbol{m}_\alpha/d\alpha$. This is given from

$$\partial_m D_\alpha\left[q \ : \ p\left(\boldsymbol{m}_\alpha\right)\right] = 0, \tag{61}$$

$$\partial_m D_{\alpha + d\alpha}\left[q \ : \ p\left(\boldsymbol{m}_\alpha + d\boldsymbol{m}_\alpha\right)\right] = 0 \tag{62}$$

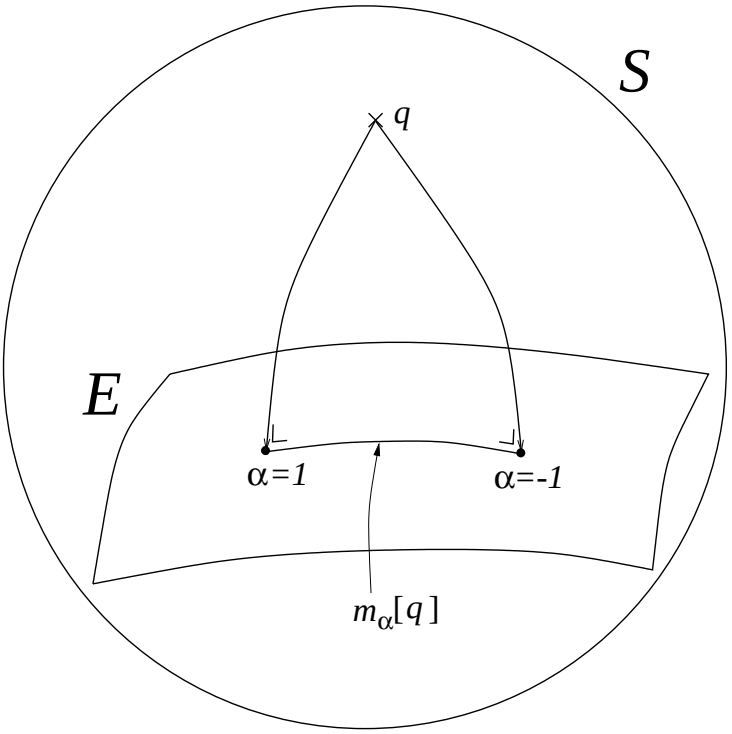

Figure 3: α-trajectory

so that, by Taylor expansion,

$$\partial_m \partial_\alpha D_\alpha d\alpha + \partial_m^2 D_\alpha d\boldsymbol{m}_\alpha = 0. \tag{63}$$

Here, $\partial_m = \partial/\partial \boldsymbol{m}$ and $\partial_\alpha = d/d\alpha$. We then have

$$\frac{d\boldsymbol{m}_\alpha}{d\alpha} = -\left\{\partial_m^2 D_\alpha\right\}^{-1} \partial_m \partial_\alpha D_\alpha. \tag{64}$$

Starting from the naive approximation, we may improve it by the expansion

$$\boldsymbol{m}_\alpha[q] = \boldsymbol{m}_1[q] + \frac{d\boldsymbol{m}_\alpha}{d\alpha}(\alpha - 1) + \frac{1}{2}\frac{d^2\boldsymbol{m}_\alpha}{d\alpha^2}(\alpha-1)^2 + \cdots \tag{65}$$

where the derivatives are evaluated at $\alpha = 1$, provided they are calculated easily. Another idea is to integrate $d\boldsymbol{m}_\alpha/d\alpha$, provided the derivative at α can be calculated. We cannot solve these methods now.

In order to study the α-trajectory, we show some preliminary calculations. The α-divergence from q to $p(x, \boldsymbol{m}) \in \boldsymbol{E}$ is written as

$$D_\alpha\left[q : p(\boldsymbol{m})\right] = \frac{4}{1-\alpha^2}\left(1 - e^{\psi(\alpha, \boldsymbol{m})}\right) \tag{66}$$

where

$$\psi(\alpha, \boldsymbol{m}) = \log \sum q^{\frac{1-\alpha}{2}} p^{\frac{1+\alpha}{2}} = \log E_p \left[\left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right]. \tag{67}$$

We first calculate $\partial_m \psi$, since $\boldsymbol{m}_\alpha$ is given by

$$\partial_m \psi \left(\alpha, \boldsymbol{m}_\alpha \right) = 0. \tag{68}$$

We have

$$\begin{aligned} \partial_m \psi &= \frac{1+\alpha}{2} e^{-\psi(\alpha, \boldsymbol{m})} \sum q^{\frac{1-\alpha}{2}} p^{-\frac{1-\alpha}{2}} \partial_m p \\ &= \frac{1+\alpha}{2} e^{-\psi} E_p \left[\partial_m l \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right]. \end{aligned} \tag{69}$$

We then have

$$\partial_m^2 \psi = - \left(\partial_m \psi \right)^2 + \frac{1+\alpha}{2} e^{-\psi} \partial_m E_p \left[\partial_m l \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right], \tag{70}$$

where $\partial_m^2 \psi$ is a matrix and $\left(\partial_m \psi \right)^2$ implies $\left(\partial_m \psi \right) \left(\partial_m \psi \right)^T$. At $\boldsymbol{m} = \boldsymbol{m}_\alpha$, the first term vanishes and

$$\partial_m^2 \psi = \frac{1+\alpha}{2} e^{-\psi} E_p \left[\left(\left(\partial_m l \right)^2 + \partial_m^2 l \right) \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} - \frac{1-\alpha}{2} \left(\partial_m l \right)^2 \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right]. \tag{71}$$

We also have

$$\begin{aligned} \partial_m \partial_\alpha \psi(\alpha, \boldsymbol{m}) &= \frac{1+\alpha}{4} e^{-\psi} E_p \left[\partial_m l \log \frac{p}{q} \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right] + \frac{1}{1+\alpha} \partial_m \psi \\ &= \frac{1+\alpha}{4} e^{-\psi} E_p \left[\partial_m l \log \frac{p}{q} \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right]. \end{aligned} \tag{72}$$

From this we have

$$\frac{d \boldsymbol{m}_\alpha}{d \alpha} = -\frac{1}{2} A^{-1} E_p \left[\partial_m l \log \frac{p}{q} \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right] \tag{73}$$

$$A = E_p \left[\left(\left(\partial_m l \right)^2 + \partial_m^2 l \right) \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} - \frac{1-\alpha}{2} \left(\partial_m l \right)^2 \left(\frac{q}{p} \right)^{\frac{1-\alpha}{2}} \right]. \tag{74}$$

For $\alpha = 1$, we have

$$\frac{d \boldsymbol{m}_\alpha}{d \alpha} = -\frac{1}{2} \frac{E_p \left[\partial_m l \left\{ \log(p/q) \right\}^2 \right]}{E_p \left[\left\{ \left(\partial_m l \right)^2 + \partial_m^2 l \right\} \log(p/q) + \left(\partial_m l \right)^2 \right]}. \tag{75}$$

8 α-Hessian

The α-projection $\boldsymbol{m}_\alpha[q]$ is given by the point $\bar{p}_\alpha$ in $\boldsymbol{E}$ that is the orthogonal projection of q to $\boldsymbol{E}$ by the α-geodesic. Such projection is not necessarily unique. The projection is not necessarily the minimizer of $D_\alpha[q:p]$ but is a saddle or even the maximizer of D_α. To elucidate this we calculate the Hessian $H^\alpha = \left(H^\alpha_{ij}\right)$

$$H^\alpha_{ij} = \frac{\partial^2}{\partial m_i \partial m_j} D_\alpha[q : p(\boldsymbol{m})] \tag{76}$$

at the α-projection $\boldsymbol{m}_\alpha[q]$. When H^α is positive-definite, the α-projection $\boldsymbol{m}_\alpha$ gives a local minimum, but it is otherwise a saddle or local maximum. We have

$$\begin{aligned} H^\alpha_{ij} &= -\frac{2}{1+\alpha} E_p\left[\left\{\frac{1+\alpha}{2}\partial_i l \partial_j l + \partial_i \partial_j l\right\} f_\alpha\right] \\ &= -E_p\left[\left\{\partial_i l \partial_j l + \frac{2}{1+\alpha}\partial_i \partial_j l\right\} f_\alpha\right], \end{aligned} \tag{77}$$

where $\partial_i = \partial/\partial m_i$ and $f_\alpha = \frac{2}{1-\alpha}\left(\frac{q}{p}\right)^{\frac{1-\alpha}{2}}$. From

$$\partial_i l = \frac{1}{1-m_i^2}\left(x_i - m_i\right), \tag{78}$$

$$\partial_i \partial_j l = -\delta_{ij}\frac{1}{\left(1-m_i^2\right)^2}\left(x_i - m_i\right)^2, \tag{79}$$

we finally have

$$\begin{aligned} H^\alpha_{ij} &= \frac{-1}{\left(1-m_i^2\right)\left(1-m_j^2\right)} E_p\left[\left(x_i - m_i\right)\left(x_j - m_j\right) f_\alpha\right] \\ &= -\bar{g}_{ii}\bar{g}_{jj}\left\{E_p\left[x_i x_j f_\alpha\right] - m_i m_j\right\}, \quad i \neq j \end{aligned} \tag{80}$$

because of

$$m^\alpha_i[q] = E_p\left[x_i f_\alpha\right] \tag{81}$$

and for $i \neq j$

$$H^\alpha_{ii} = \frac{1-\alpha}{1+\alpha}\left(\bar{g}_{ii}\right)^2 E_p\left[\left(x_i - m_i\right)^2 f_\alpha\right]. \tag{82}$$

We calculate the two special cases $\alpha = \pm 1$. For $\alpha = -1$,

$$\begin{aligned} H^{-1}_{ij} &= \partial_i \partial_j \int q \log \frac{q}{p} dx \\ &= \left(\bar{g}_{ii}\right)^2 \delta_{ij} E_q\left[\left(x_i - m_i\right)^2\right] \\ &= \bar{g}_{ii}\delta_{ij} = \bar{G}, \end{aligned} \tag{83}$$

where $\bar{G}$ is the inverse of the Fisher information matrix of $\boldsymbol{E}$. This is diagonal and positive-definite. Because $\boldsymbol{E}$ is e-flat, we know that $\alpha = -1$-projection gives the global minimum, is unique (that is no other critical points) and gives the true solution $\boldsymbol{m}_{-1}[q] = E_q[\boldsymbol{x}]$.

For $\alpha = 1$, we have

$$\begin{aligned} H^1_{ij} &= \partial_i \partial_j \int p \log \frac{p}{q} dx \\ &= \bar{g}_{ii}\delta_{ij} - E_p\left[(\partial_i\partial_j l + \partial_i l \partial_j l)\log q\right]. \end{aligned} \tag{84}$$

Hence,

$$H^1_{ii} = \bar{g}_{ii} = \frac{1}{1-m_i^2}, \tag{85}$$

$$\begin{aligned} H^1_{ij} &= -\bar{g}_{ii}\bar{g}_{jj} E_p\left[(x_i - m_j)(x_j - m_j)\log q\right] \\ &= -\bar{g}_{ii}\bar{g}_{jj} E_p\Bigg[(x_i - m_i)(x_j - m_j)\Bigg\{\sum w_{kl}x_k x_l \\ &\qquad + \sum h_k x_k - \psi_q\Bigg\}\Bigg] \\ &= -w_{ij}. \end{aligned} \tag{86}$$

This shows that H^1 is not necessarily positive-definite. This fact is related to the α-curvature of $\boldsymbol{E}$. When $n = 2$ (two neurons), it is positive definite when and only when $w = w_{12}$ satisfies

$$w^2 < \frac{1}{1-m_1^2}\frac{1}{1-m_2^2}. \tag{87}$$

Otherwise, it is a saddle. When $h_1 = h_2$, there exist one or two local minima other than this.

This fact implies that the naive mean field approximation might give a pathological solution in some cases. When $|w|$ is large, the above two-spin system is dynamically bistable, having two stable solutions $x_1 = x_2 = 1$, $x_1 = x_2 = -1$ ($w > 0$) or $x_1 = -x_2$ ($w < 0$).

When $\alpha = -1$, the α-projection $\boldsymbol{m}_{-1}$ is unique. Starting from $\alpha = -1$, the α-trajectory $\boldsymbol{m}_\alpha$ bifurcates at some α, and then bifurcates further, depending on W. See Fig.4. When W is large, the naive mean field approximation (59) may have an exponentially large number of solutions. It is interesting to study the diagram of bifurcation for the α-trajectory.

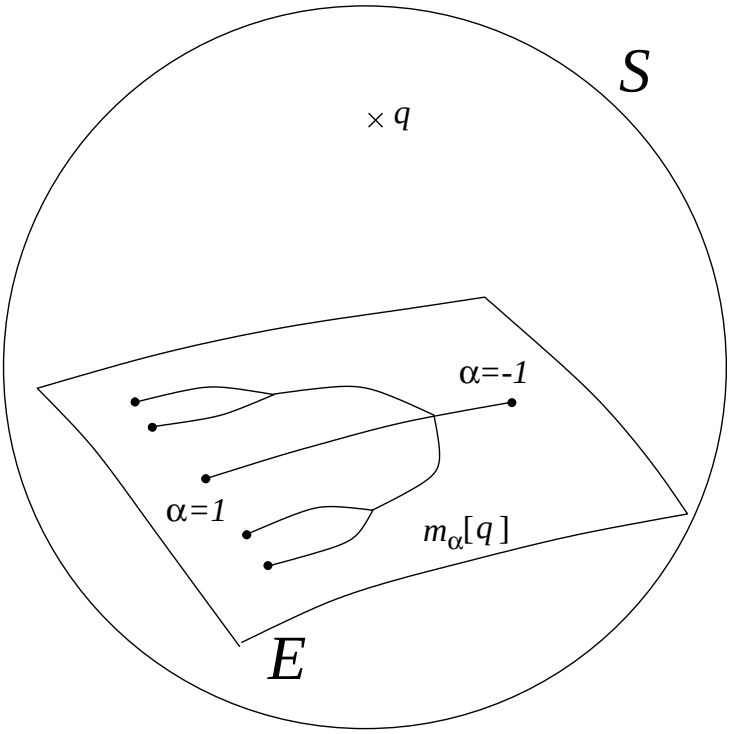

Figure 4: Bifurcation

9 Small w approximation of the α-trajectory

We give an explicit formula for the α-projection $\boldsymbol{m}_\alpha$ assuming that w_{ij} are small. The TAP solution corresponds to $\boldsymbol{m}_{-1}[q]$ under this approximation.

Let us consider an exponential family $\{p(\boldsymbol{x}, \theta)\}$. For two nearby distributions $q = p(\boldsymbol{x}, \theta + d\theta)$ and $p = p(\boldsymbol{x}, \theta)$, we have

$$D_\alpha(q, p) = D_\alpha(\theta + d\theta, \theta) = \frac{1}{2} \sum g_{ij} d\theta^i d\theta^j + \frac{3-\alpha}{12} \sum T_{ijk} d\theta^i d\theta^j d\theta^k, \tag{88}$$

where

$$g_{ij} = \frac{\partial^2}{\partial\theta^i \partial\theta^j} \psi(\theta), \quad T_{ijk} = \frac{\partial^3}{\partial\theta^i \partial\theta^j \partial\theta^k} \psi(\theta) = \frac{\partial}{\partial\theta^k} g_{ij}. \tag{89}$$

In our case, $\theta = (W, \boldsymbol{h})$, and for $q = q(\boldsymbol{x}, dW, \boldsymbol{h})$ and $p = p(\boldsymbol{x}, 0, \bar{\boldsymbol{h}})$, we have

$$d\theta = (dW, d\boldsymbol{h}), \tag{90}$$

where $W = dW$ is assumed to be small and $d\boldsymbol{h} = \boldsymbol{h} - \bar{\boldsymbol{h}}$. We can calculate g_{ij} and T_{ijk} at $p \in \boldsymbol{E}$, for example, for $I = (i, j)$ and $j = (k, l)$,

$$g_{IJ} = E_p \left[(x_i x_j - m_i m_j) (x_k x_l - m_k m_l) \right]. \tag{91}$$

The metric G consists of the three parts g_{IJ}, g_{Ik}, g_{kl} where I, J etc are index pairs corresponding to $dW^I = w^{ij}$, and small letter indices i, j etc refer to $dh^i = h^i - \theta^i$.

Note that

$$g_{ij} = E\left[\left(x_i - m_i\right)\left(x_j - m_j\right)\right] \tag{92}$$

$$T_{ijk} = E\left[\left(x_i - m_i\right)\left(x_j - m_j\right)\left(x_k - m_k\right)\right] \tag{93}$$

etc. By using this, we have the following expansion,

$$\begin{aligned} D_\alpha[q,p] = \frac{1}{2}g_{IJ}dw^I dw^J + \frac{1}{2}g_{ij}dh^i dh^j + g_{Ik}dw^I dh^k \\ + \frac{3-\alpha}{12}\Big\{T_{IJK}dw^I dw^J dw^K + 3T_{Ijk}dw^I dh^j dh^k \\ + 3T_{IJk}dw^I dw^J dh^k + T_{ijk}dh^i dh^j dh^k\Big\}, \end{aligned} \tag{94}$$

where the summation convention is used for repeated indices. In order to obtain the α-projection, we solve

$$\frac{\partial D_\alpha}{\partial \bar{h}_i} = 0, \tag{95}$$

where indices i of $d\theta_i$ are decomposed into indices pairs $I = (i, j)$, etc. for $W^I = w^{ij}$ and single indices $i, j, \cdots$ for h_i. For example, we have

$$\begin{aligned} 0 = -g_{il}dh^i - g_{Il}dw^I - \frac{1-\alpha}{4}\Big(T_{IJl}dw^I dw^J \\ + T_{ijl}dh^i dh^j + 2T_{Ikl}dw^I dh^k\Big) + O\left(w^3\right), \end{aligned} \tag{96}$$

where we used

$$\frac{\partial}{\partial \bar{h}_l}dh^i = -\delta_{il}. \tag{97}$$

The first-order solution to (96) does not depend on α,

$$h^i - \bar{h}_i = g^{il}g_{Il}dw^I \tag{98}$$

or

$$\theta^i = h^i + \sum w_{ij}m_j \tag{99}$$

which is the naive mean field approximation.

In order to calculate higher-order corrections, we note

$$g_{ij} = \delta_{ij}\left(1-m_i^2\right) \tag{100}$$

$$g_{Ik} = \left(1-m_i^2\right)\left\{\delta_{ki}m_j+\delta_{kj}m_i\right\}, \quad I=(j,j) \tag{101}$$

$$g_{IJ} = E\left[\left(x_ix_j-m_im_j\right)\left(x_kx_l-m_km_l\right)\right], \quad I=(i,j),\ J=(k,l) \tag{102}$$

etc. Quantities T can be calculated similarly. The second-order correction term A_l, which is given by the second term of (96), is obtained after painful calculations as

$$A_l = m_l\sum\left(w_{lk}\right)^2\left(1-m_k^2\right). \tag{103}$$

Some easy terms are

$$T_{ijl}dh^idh^j = -2m_l\left(1-m_l^2\right)\left(dw^{lk}m_k\right)^2 \tag{104}$$

$$T_{Ikl}dw^Idh^l = 2m_l\left(1-m_l^2\right)\left(dw^{lk}m_k\right)^2 - \left(1-m_l^2\right)\left(1-m_k^2\right)dw^{lk}dw^{ks}m_s. \tag{105}$$

After all, we have

$$\bar{h}_l = m_l+\sum w_{lk}m_k+\frac{1-\alpha}{2}m_l\sum\left(w_{lk}\right)^2\left(1-m_k^2\right). \tag{106}$$

or

$$m_l^\alpha = \tanh\left(h_l+\sum w_{ek}m_k^\alpha+\frac{1-\alpha}{2}m_l^\alpha\sum\left(w_{lk}\right)^2\left(1-m_k^{\alpha 2}\right)\right). \tag{107}$$

This gives the α-projections in terms of parameter α, where $\alpha=1$ is for the naive approximation and $\alpha=-1$ is for the TAP approximation. This is small w approximation of the α-trajectory and is valid for small w.

Conclusions

The present article studies the geometrical structure underlying mean field approximation. Information geometry is used for this purpose which has the Riemannian metric together with dual pairs of affine connections. Information geometry gives the α-structure to the manifold of probability distributions of the SK-spin glass system or the Boltzmann

machine. The α-divergence is defined in the manifold which is invariant under a certain criterion.

The mean field approximation is a method of calculating quantities related to a complex probability distribution, by using a simple tractable model such as the family of independent distributions. The $\alpha = -1$ projection of the distribution to submanifold consisting of independent distributions is known to give the correct answer, but it is intractable. The $\alpha = 1$ approximation is tractable, but it gives only an approximation.

We search for possibility of using the α-approximation that minimizes the α-divergence. It is unfortunately difficult to calculate. But its properties are studied for future study. We have also shown the information-geometric meaning of the TAP approximation.

We have elucidated the fact that $\alpha = -1$ projection is unique, giving the true solution but α-approximation ($\alpha \neq -1$) is not necessarily unique. When we study the trajectory consisting of the α-projections, there are a number of bifurcations where the α-projection bifurcates. It is an interesting problem to study the properties of such bifurcation and its implications.

The present article is preliminary to further studies on interesting problems connecting information geometry and statistical physics.

References

[1] S. Amari, *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics 28, Springer-Verlag, 1985.

[2] S. Amari, Dualistic geometry of the manifold of higher-order neurons, *Neural Networks*, 4:443–451, 1991.

[3] S. Amari, Information geometry of EM and em algorithms for neural networks, Neural Networks, 8:1379–1408, 1995.

[4] S. Amari, Natural gradient works efficiently in learning, *Neural Computation*, 10:251–276, 1998.

[5] S. Amari, K. Kurata, and H. Nagaoka, Information geometry of Boltzmann machines, *IEEE Transactions on Neural Networks*, 3:26—271, 1992.

[6] S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS and Oxford University Press, 2000.

[7] C. Bhattacharyya and S.S. Keerthi, Mean-field methods for Stochastic Connectionist Networks, TR No. IIsc-CSA-00-03.

[8] C. Bhattacharyya and S.S. Keerthi, Plefka's mean-field theory from a Variational viewpoint, TR NO. IISc-CSA-00-02.

[9] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations, *Annals of Mathematical Statistics*, 23:493–507, 1952.

[10] N.N. Chentsov (Čencov), *Statistical Decision Rules and Optimal Inference*, American Mathematical Society, Rhode Island, U.S.A., 1982, (Originally published in Russian, Nauka, Moscow, 1972).

[11] I. Csiszár, I-divergence geometry of probability distributions and minimization problems, *The Annals of Probability*, 3:146–158, 1975.

[12] H. Nagaoka and S. Amari, Differential geometry of smooth families of probability distributions, Technical Report METR 82-7, Dept. of Math. Eng. and Instr. Phys, Univ. of Tokyo, 1982.

[13] A. Rényi, On measures of entropy and information, In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, University of California Press, 1961.

[14] T. Tanaka, Information geometry of mean field approximation, *Neural Computation*, to appear.

[15] C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics, *Journal of Statistical Physics*, 52:479–, 1988.