Front. Electr. Electron. Eng. China 2010, 5(3): 241–260
DOI 10.1007/s11460-010-0101-3

RESEARCH ARTICLE

Shun-ichi AMARI

Information geometry in optimization, machine learning and statistical inference

Abstract The present article gives an introduction to information geometry and surveys its applications in the area of machine learning, optimization and statistical inference. Information geometry is explained intuitively by using divergence functions introduced in a manifold of probability distributions and other general manifolds. They give a Riemannian structure together with a pair of dual flatness criteria. Many manifolds are dually flat. When a manifold is dually flat, a generalized Pythagorean theorem and related projection theorem are introduced. They provide useful means for various approximation and optimization problems. We apply them to alternative minimization problems, Ying-Yang machines and belief propagation algorithm in machine learning.

Keywords information geometry, machine learning, optimization, statistical inference, divergence, graphical model, Ying-Yang machine

1 Introduction

Information geometry [1] deals with a manifold of probability distributions from the geometrical point of view. It studies the invariant structure by using the Riemannian geometry equipped with a dual pair of affine connections. Since probability distributions are used in many problems in optimization, machine learning, vision, statistical inference, neural networks and others, information geometry provides a useful and strong tool to many areas of information sciences and engineering.

Many researchers in these fields, however, are not familiar with modern differential geometry. The present article intends to give an understandable introduction to information geometry without modern differential geometry. Since underlying manifolds in most applications are dually flat, the dually flat structure plays a fundamental role. We explain the fundamental dual structure and related dual geodesics without using the concept of affine connections and covariant derivatives.

We begin with a divergence function between two points in a manifold. When it satisfies an invariance criterion of information monotonicity, it gives a family of f-divergences [2]. When a divergence is derived from a convex function in the form of Bregman divergence [3], this gives another type of divergence, where the Kullback-Leibler divergence belongs to both of them. We derive a geometrical structure from a divergence function [4]. The Fisher information Riemannian structure is derived from an invariant divergence (f-divergence) (see Refs. [1,5]), while the dually flat structure is derived from the Bregman divergence (convex function).

The manifold of all discrete probability distributions is dually flat, where the Kullback-Leibler divergence plays a key role. We give the generalized Pythagorean theorem and projection theorem in a dually flat manifold, which plays a fundamental role in applications. Such a structure is not limited to a manifold of probability distributions, but can be extended to the manifolds of positive arrays, matrices and visual signals, and will be used in neural networks and optimization problems.

After introducing basic properties, we show three areas of applications. One is application to the alternative minimization procedures such as the expectation-maximization (EM) algorithm in statistics [6–8]. The second is an application to the Ying-Yang machine introduced and extensively studied by Xu [9–14]. The third one is application to belief propagation algorithm of stochastic reasoning in machine learning or artificial intelligence [15–17]. There are many other applications in analysis of spiking patterns of the brain, neural networks, boosting algorithm of machine learning, as well as wide range of statistical inference, which we do not

Received January 15, 2010; accepted February 5, 2010

Shun-ichi AMARI (✉)
RIKEN Brain Science Institute, Saitama 351-0198, Japan
E-mail: amari@brain.riken.jp

mention here.

2 Divergence function and information geometry

2.1 Manifold of probability distributions and positive arrays

We introduce divergence functions in various spaces or manifolds. To begin with, we show typical examples of manifolds of probability distributions. A one-dimensional Gaussian distribution with mean μ and variance σ^2 is represented by its probability density function

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

It is parameterized by a two-dimensional parameter $\boldsymbol{\xi} = (\mu,\sigma)$. Hence, when we treat all such Gaussian distributions, not a particular one, we need to consider the set S_{G} of all the Gaussian distributions. It forms a two-dimensional manifold

$$S_{\text{G}} = \{p(x;\boldsymbol{\xi})\}, \tag{2}$$

where $\boldsymbol{\xi} = (\mu,\sigma)$ is a coordinate system of S_{G}. This is not the only coordinate system. It is possible to use other parameterizations or coordinate systems when we study S_{G}.

We show another example. Let x be a discrete random variable taking values on a finite set $X = \{0,1,\ldots,n\}$. Then, a probability distribution is specified by a vector $\boldsymbol{p} = (p_0,p_1,\ldots,p_n)$, where

$$p_i = \text{Prob}\,\{x = i\}. \tag{3}$$

We may write

$$p(x;\boldsymbol{p}) = \sum p_i\delta_i(x), \tag{4}$$

where

$$\delta_i(x) = \begin{cases} 1, \, x = i, \\ 0, \, x \neq i. \end{cases} \tag{5}$$

Since $\boldsymbol{p}$ is a probability vector, we have

$$\sum p_i = 1, \tag{6}$$

and we assume

$$p_i > 0. \tag{7}$$

The set of all the probability distributions is denoted by

$$S_n = \{\boldsymbol{p}\}, \tag{8}$$

which is an n-dimensional simplex because of (6) and (7). When $n = 2$, S_n is a triangle (Fig. 1). S_n is an n-dimensional manifold, and $\boldsymbol{\xi} = (p_1,\ldots,p_n)$ is a coordinate system. There are many other coordinate systems. For example,

$$\theta_i = \log\frac{p_i}{p_0}, \qquad i = 1,\ldots,n, \tag{9}$$

is an important coordinate system of S_n, as we will see later.

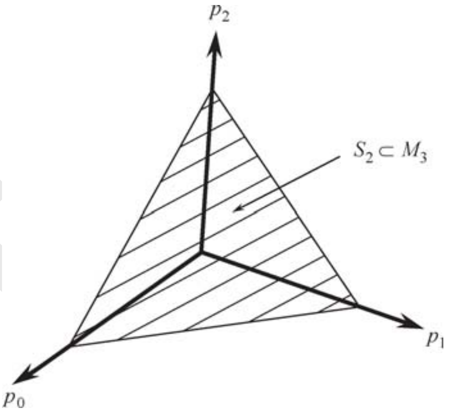

Fig. 1 Manifold S_2 of discrete probability distributions

The third example deals with positive measures, not probability measures. When we disregard the constraint $\sum p_i = 1$ of (6) in S_n, keeping $p_i > 0$, $\boldsymbol{p}$ is regarded as an $(n+1)$-dimensional positive arrays, or a positive measure where $x = i$ has measure p_i. We denote the set of positive measures or arrays by

$$M_{n+1} = \{\boldsymbol{z},\ z_i > 0\,;\ i = 0,1,\ldots,n\}. \tag{10}$$

This is an $(n+1)$-dimensional manifold with a coordinate system $\boldsymbol{z}$. S_n is its submanifold derived by a linear constraint $\sum z_i = 1$.

In general, we can regard any regular statistical model

$$S = \{p(x,\boldsymbol{\xi})\} \tag{11}$$

parameterized by $\boldsymbol{\xi}$ as a manifold with a (local) coordinate system $\boldsymbol{\xi}$. It is a space M of positive measures, when the constraint $\int p(x,\boldsymbol{\xi})\mathrm{d}x = 1$ is discarded. We may treat any other types of manifolds and introduce dual structures in them. For example, we will consider a manifold consisting of positive-definite matrices.

2.2 Divergence function and geometry

We consider a manifold S having a local coordinate system $\boldsymbol{z} = (z_i)$. A function $D[\boldsymbol{z}:\boldsymbol{w}]$ between two points $\boldsymbol{z}$ and $\boldsymbol{w}$ of S is called a divergence function when it satisfies the following two properties:

1) $D[\boldsymbol{z}:\boldsymbol{w}] \geqslant 0$, with equality when and only when $\boldsymbol{z} = \boldsymbol{w}$.
2) When the difference between $\boldsymbol{w}$ and $\boldsymbol{z}$ is infinitesimally small, we may write $\boldsymbol{w} = \boldsymbol{z} + \mathrm{d}\boldsymbol{z}$ and Taylor expansion gives

$$D\left[\boldsymbol{z}:\boldsymbol{z}+\mathrm{d}\boldsymbol{z}\right] = \sum g_{ij}(\boldsymbol{z})\mathrm{d}z_i\mathrm{d}z_j, \tag{12}$$

where

$$g_{ij}(\boldsymbol{z}) = \frac{\partial^2}{\partial z_i \partial z_j}D[\boldsymbol{z}:\boldsymbol{w}]_{|\boldsymbol{w}=\boldsymbol{z}} \tag{13}$$

is a positive-definite matrix.

A divergence does not need to be symmetric, and

$$D\left[\boldsymbol{z}:\boldsymbol{w}\right]\neq D\left[\boldsymbol{w}:\boldsymbol{z}\right] \tag{14}$$

in general, nor does it satisfy the triangular inequality. Hence, it is not a distance. It rather has a dimension of square of distance as is seen from (12). So (12) is considered to define the square of the local distance ds between two nearby points $Z = (\boldsymbol{z})$ and $Z + \mathrm{d}Z = (\boldsymbol{z} + \mathrm{d}\boldsymbol{z})$,

$$\mathrm{d}s^2 = \sum g_{ij}(\boldsymbol{z})\mathrm{d}z_i\mathrm{d}z_j. \tag{15}$$

More precisely, d$\boldsymbol{z}$ is regarded as a small line element connecting two points Z and $Z + \mathrm{d}Z$. This is a tangent vector at point Z (Fig. 2). When a manifold has a positive-definite matrix $g_{ij}(\boldsymbol{z})$ at each point $\boldsymbol{z}$, it is called a Riemannian manifold, and (g_{ij}) is a Riemannian metric tensor.

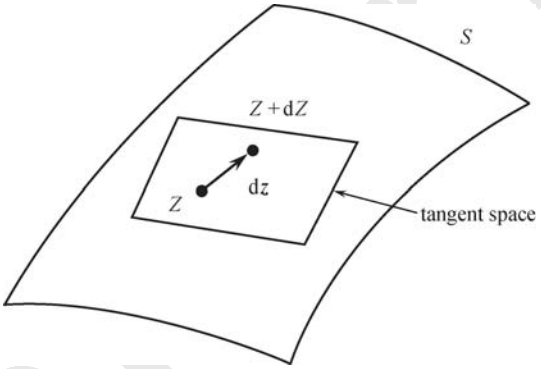

Fig. 2 Manifold, tangent space and tangent vector

An affine connection defines a correspondence between two nearby tangent spaces. By using it, a geodesic is defined: A geodesic is a curve of which the tangent directions do not change along the curve by this correspondence (Fig. 3). It is given mathematically by a covariant derivative, and technically by the Christoffel symbol, $\Gamma_{ijk}(\boldsymbol{z})$, which has three indices.

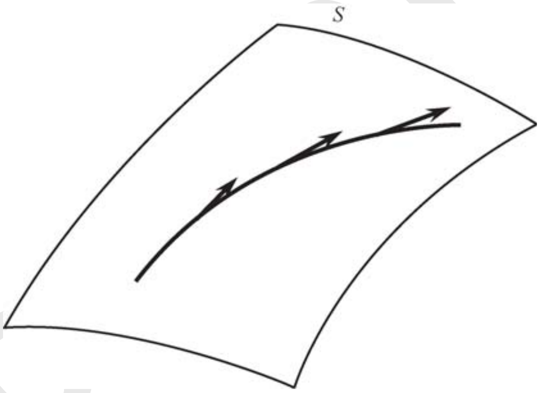

Fig. 3 Geodesic, keeping the same tangent direction

One may skip the following two paragraphs, since technical details are not used in the following. Eguchi [4] proposed the following two connections:

$$\Gamma_{ijk}(\boldsymbol{z}) = -\frac{\partial^3}{\partial z_i \partial z_j \partial w_k} D[\boldsymbol{z}:\boldsymbol{w}]_{|\boldsymbol{w}=\boldsymbol{z}}, \tag{16}$$

$$\Gamma^*_{ijk}(\boldsymbol{z}) = -\frac{\partial^3}{\partial w_i \partial w_j \partial z_k} D[\boldsymbol{z}:\boldsymbol{w}]_{|\boldsymbol{w}=\boldsymbol{z}}, \tag{17}$$

derived from a divergence $D[\boldsymbol{z}:\boldsymbol{w}]$. These two are dually coupled with respect to the Riemannian metric g_{ij} [1]. The meaning of dually coupled affine connections is not explained here, but will become clear in later sections, by using specific examples.

The Euclidean divergence, defined by

$$D_{\mathrm{E}}[\boldsymbol{z}:\boldsymbol{w}] = \frac{1}{2}\sum (z_i - w_i)^2, \tag{18}$$

is a special case of divergence. We have

$$g_{ij} = \delta_{ij}, \tag{19}$$

$$\Gamma_{ijk} = \Gamma^*_{ijk} = 0, \tag{20}$$

where δ_{ij} is the Kronecker delta. Therefore, the derived geometry is Euclidean. Since it is self-dual ($\Gamma_{ijk} = \Gamma^*_{ijk}$), the duality does not play a role.

2.3 Invariant divergence: f-divergence

2.3.1 Information monotonicity

Let us consider a function $\boldsymbol{t}(\boldsymbol{x})$ of random variable $\boldsymbol{x}$, where $\boldsymbol{t}$ and $\boldsymbol{x}$ are vector-valued. We can derive probability distribution $\bar{p}(\boldsymbol{t}, \boldsymbol{\xi})$ of $\boldsymbol{t}$ from $p(\boldsymbol{x}, \boldsymbol{\xi})$ by

$$\bar{p}(\boldsymbol{t}, \boldsymbol{\xi})\mathrm{d}\boldsymbol{t} = \int_{\boldsymbol{t}=\boldsymbol{t}(\boldsymbol{x})} p(\boldsymbol{x}, \boldsymbol{\xi})\mathrm{d}\boldsymbol{x}. \tag{21}$$

When $\boldsymbol{t}(\boldsymbol{x})$ is not reversible, that is, $\boldsymbol{t}(\boldsymbol{x})$ is a many-to-one mapping, there is loss of information by summarizing observed data $\boldsymbol{x}$ into reduced $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$. Hence, for two divergences between two distributions specified by $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$,

$$D = D\left[p\left(\boldsymbol{x}, \boldsymbol{\xi}_1\right) : p\left(\boldsymbol{x}, \boldsymbol{\xi}_2\right)\right], \tag{22}$$

$$\bar{D} = D\left[\bar{p}\left(\boldsymbol{t}, \boldsymbol{\xi}_1\right) : \bar{p}\left(\boldsymbol{t}, \boldsymbol{\xi}_2\right)\right], \tag{23}$$

it is natural to require

$$\bar{D} \leqslant D. \tag{24}$$

The equality holds, when and only when $\boldsymbol{t}(\boldsymbol{x})$ is a sufficient statistics. A divergence function is said to be invariant, when this requirement is satisfied. The invariance is a key concept to construct information geometry of probability distributions [1,5].

Here, we use a simplified version of invariance due to Csiszár [18,19]. We consider the space S_n of all probability distributions over $n+1$ atoms $X = \{x_0, x_1, \ldots, x_n\}$. A probability distributions is given by $\boldsymbol{p} = (p_0, p_1, \ldots, p_n)$, $p_i = \mathrm{Prob}\left\{x = x_i\right\}$.

Let us divide X into m subsets, $T_1, T_2, \ldots, T_m$ ($m < n+1$), say

$$T_1 = \{x_1, x_2, x_5\}, \quad T_2 = \{x_3, x_8, \ldots\}, \quad \ldots \tag{25}$$

This is a partition of X,

$$X = \cup T_i, \tag{26}$$
$$T_i \cap T_j = \phi. \tag{27}$$

Let t be a mapping from X to $\{T_1, \ldots, T_m\}$. Assume that we do not know the outcome x directly. Instead we can observe $t(x) = T_j$, knowing the subset T_j to which x belongs. This is called coarse-graining of X into $T = \{T_1, \ldots, T_m\}$.

The coarse-graining generates a new probability distributions $\bar{\boldsymbol{p}} = (\bar{p}_1, \ldots, \bar{p}_m)$ over $T_1, \ldots, T_m$,

$$\bar{p}_j = \text{Prob}\{T_j\} = \sum_{x \in T_j} \text{Prob}\{x\}. \tag{28}$$

Let $\bar{D}[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}]$ be an induced divergence between $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{q}}$. Since coarse-graining summarizes a number of elements into one subset, detailed information of the outcome is lost. Therefore, it is natural to require

$$\bar{D}[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}] \leqslant D[\boldsymbol{p} : \boldsymbol{q}]. \tag{29}$$

When does the equality hold? Assume that the outcome x is known to belong to T_j. This gives some information to distinguish two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. If we know further detail of x inside subset T_j, we obtain more information to distinguish the two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. Since x belongs to T_j, we consider the two conditional probability distributions

$$p(x_i | x_i \in T_j), \quad q(x_i | x_i \in T_j) \tag{30}$$

under the condition that x is in subset T_j. If the two distributions are equal, we cannot obtain further information to distinguish $\boldsymbol{p}$ from $\boldsymbol{q}$ by observing x inside T_j. Hence,

$$\bar{D}[\bar{\boldsymbol{p}} : \bar{\boldsymbol{q}}] = D[\boldsymbol{p} : \boldsymbol{q}] \tag{31}$$

holds, when and only when

$$p(x_i | T_j) = q(x_i | T_j) \tag{32}$$

for all T_j and all $x_i \in T_j$, or

$$\frac{p_i}{q_i} = \lambda_j \tag{33}$$

for all $x_i \in T_j$ for some constant λ_j.

A divergence satisfying the above requirements is called an invariant divergence, and such a property is termed as information monotonicity.

2.3.2 f-divergence

The f-divergence was introduced by Csiszár [2] and also by Ali and Silvey [20]. It is defined by

$$D_f[\boldsymbol{p} : \boldsymbol{q}] = \sum p_i f\left(\frac{q_i}{p_i}\right), \tag{34}$$

where f is a convex function satisfying

$$f(1) = 0. \tag{35}$$

For a function cf with a constant c, we have

$$D_{cf}[\boldsymbol{p} : \boldsymbol{q}] = cD_f[\boldsymbol{p} : \boldsymbol{q}]. \tag{36}$$

Hence, f and cf give the same divergence except for the scale factor c. In order to standardize the scale of divergence, we may assume that

$$f''(1) = 1, \tag{37}$$

provided f is differentiable. Further, for $f_c(u) = f(u) - c(u-1)$ where c is any constant, we have

$$D_{f_c}[\boldsymbol{p} : \boldsymbol{q}] = D_f[\boldsymbol{p} : \boldsymbol{q}]. \tag{38}$$

Hence, we may use such an f that satisfies

$$f'(1) = 0 \tag{39}$$

without loss of generality. A convex function satisfying the above three conditions (35), (37), (39) is called a standard f function.

A divergence is said to be decomposable, when it is a sum of functions of components

$$D[\boldsymbol{p} : \boldsymbol{q}] = \sum_i D[p_i : q_i]. \tag{40}$$

The f-divergence (34) is a decomposable divergence.

Csiszár [18] found that any f-divergence satisfies information monotonicity. Moreover, the class of f-divergences is unique in the sense that any decomposable divergence satisfying information monotonicity is an f-divergence.

Theorem 1 Any f-divergence satisfies the information monotonicity. Conversely, any decomposable information monotonic divergence is written in the form of f-divergence.

A proof is found, e.g., in Ref. [21].

The Riemannian metric and affine connections derived from an f-divergence has a common invariant structure [1]. They are given by the Fisher information metric and $\pm\alpha$-connections, which are shown in a later section.

An extensive list of f-divergences is given in Cichocki et al. [22]. Some of them are listed below.

1) The Kullback-Leibler (KL-) divergence: $f(u) = u \log u - (u-1)$:

$$D_{\text{KL}}[\boldsymbol{p} : \boldsymbol{q}] = \sum p_i \log \frac{q_i}{p_i}. \tag{41}$$

2) Squared Hellinger distance: $f(u)=(\sqrt{u}-1)^2$:

$$D_{\mathrm{Hel}}[\boldsymbol{p}:\boldsymbol{q}]=\sum\left(\sqrt{p_i}-\sqrt{q_i}\right)^2. \tag{42}$$

3) The α-divergence:

$$f_\alpha(u)=\frac{4}{1-\alpha^2}\left(1-u^{\frac{1+\alpha}{2}}\right)-\frac{2}{1-\alpha}(u-1), \tag{43}$$

$$D_\alpha[\boldsymbol{p}:\boldsymbol{q}]=\frac{4}{1-\alpha^2}\sum\left(1-p_i^{\frac{1+\alpha}{2}}q_i^{\frac{1-\alpha}{2}}\right). \tag{44}$$

The α-divergence was introduced by Havrda and Charvát [23], and has been studied extensively by Amari and Nagaoka [1]. Its applications were described earlier in Chernoff [24], and later in Matsuyama [25], Amari [26], etc. to mention a few. It is the squared Hellinger distance for $\alpha=0$. The KL-divergence and its reciprocal are obtained in the limit of $\alpha\to\pm 1$. See Refs. [21,22,26] for the α-structure.

2.3.3 f-divergence of positive measures

We now extend the f-divergence to the space of positive measures M_n over $X=\{x_1,\ldots,x_n\}$, whose points are given by the coordinates $\boldsymbol{z}=(z_1,\ldots,z_n)$, $z_i>0$. Here, z_i is the mass (measure) of x_i where the total mass $\sum z_i$ is positive and arbitrary. In many applications, $\boldsymbol{z}$ is a non-negative array, and we can extend it to a non-negative double array $\boldsymbol{z}=(z_{ij})$, etc., that is, matrices and tensors. We first derive an f-divergence in M_n: For two positive measures $\boldsymbol{z}$ and $\boldsymbol{y}$, an f-divergence is given by

$$D_f[\boldsymbol{z}:\boldsymbol{y}]=\sum z_i f\left(\frac{y_i}{z_i}\right), \tag{45}$$

where f is a standard convex function. It should be noted that an f-divergence is no more invariant under the transformation from $f(u)$ to

$$f_c(u)=f(u)-c(u-1). \tag{46}$$

Hence, it is absolutely necessary to use a standard f in the case of M_n, because the conditions of divergence are violated otherwise.

Among all f-divergences, the α divergence [1,21,22] is given by

$$D_\alpha[\boldsymbol{z}:\boldsymbol{y}]=\sum z_i f_\alpha\left(\frac{y_i}{z_i}\right), \tag{47}$$

where

$$f_\alpha(u)=\begin{cases}\dfrac{4}{1-\alpha^2}\left(1-u^{\frac{1+\alpha}{2}}\right)-\dfrac{2}{1-\alpha}(u-1), & \alpha\neq\pm1,\\ u\log u-(u-1), & \alpha=1,\\ -\log u+(u-1), & \alpha=-1,\end{cases} \tag{48}$$

and plays a special role in M_n. Here, we use a simple power function $u^{(1+\alpha)/2}$, changing it by adding a linear and a constant term to become the standard convex function $f_\alpha(u)$. It includes the logarithm as a limiting case.

The α-divergence is explicitly given in the following form [1,21,22]:

$$D_\alpha[\boldsymbol{z}:\boldsymbol{y}]=\begin{cases}\dfrac{2}{1-\alpha^2}\sum\left(\dfrac{1-\alpha}{2}z_i+\dfrac{1+\alpha}{2}y_i-z_i^{\frac{1+\alpha}{2}}y_i^{\frac{1-\alpha}{2}}\right), & \alpha\neq\pm1,\\ \sum\left(z_i-y_i+y_i\log\dfrac{y_i}{z_i}\right), & \alpha=1,\\ \sum\left(y_i-z_i+z_i\log\dfrac{z_i}{y_i}\right), & \alpha=-1.\end{cases} \tag{49}$$

2.4 Flat divergence: Bregman divergence

A divergence $D[\boldsymbol{z}:\boldsymbol{w}]$ gives a Riemannian metric g_{ij} by (13), and a pair of affine connections by (16), (17). When (20) holds, the manifold is flat. We study divergence functions which give flat structure in this subsection. In this case, coordinate system $\boldsymbol{z}$ is flat, that is affine, although the metric $g_{ij}(\boldsymbol{z})$ is not Euclidean. When $\boldsymbol{z}$ is affine, all the coordinate curves are regarded as straight lines, that is, geodesics. Here, we separate two concepts, flatness and metric. Mathematically speaking, we use an affine connection Γ_{ijk} to define flatness, which is not necessarily derived from the metric g_{ij}. Note that the Levi-Civita affine connection is derived from the metric in Riemannian geometry, but our approach is more general.

2.4.1 Convex function

We show in this subsection that a dually flat geometry is derived from a divergence due to a convex function in terms of the Bregman divergence [3]. Conversely a dually flat manifold always has a convex function to define its geometry. The divergence derived from it is called the canonical divergence of a dually flat manifold [1].

Let $\psi(\boldsymbol{z})$ be a strictly convex differentiable function. Then, the linear hyperplane tangent to the graph

$$y=\psi(\boldsymbol{z}) \tag{50}$$

at $\boldsymbol{z}_0$ is

$$y = \nabla\psi\left(\boldsymbol{z}_0\right) \cdot \left(\boldsymbol{z} - \boldsymbol{z}_0\right) + \psi\left(\boldsymbol{z}_0\right), \tag{51}$$

and is always below the graph $y = \psi(\boldsymbol{z})$ (Fig. 4). Here, $\nabla\psi$ is the gradient,

$$\nabla\psi = \left(\frac{\partial}{\partial z_1}\psi(\boldsymbol{z}), \ldots, \frac{\partial}{\partial z_n}\psi(\boldsymbol{z})\right). \tag{52}$$

The difference of $\psi(\boldsymbol{z})$ and the tangent hyperplane is

$$D_\psi\left[\boldsymbol{z} : \boldsymbol{z}_0\right] = \psi(\boldsymbol{z}) - \psi\left(\boldsymbol{z}_0\right) - \nabla\psi\left(\boldsymbol{z}_0\right) \cdot \left(\boldsymbol{z} - \boldsymbol{z}_0\right) \geqslant 0. \tag{53}$$

See Fig. 4. This is called the Bregman divergence induced by $\psi(\boldsymbol{z})$.

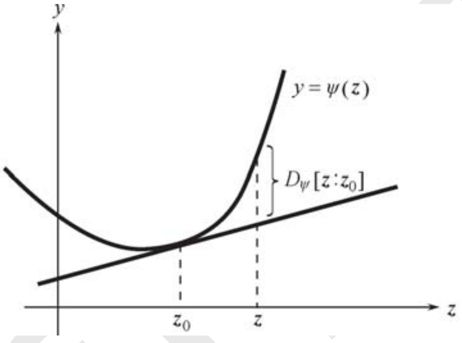

Fig. 4 Bregman divergence due to convex function

This satisfies the conditions of a divergence. The induced Riemannian metric is

$$g_{ij}(\boldsymbol{z}) = \frac{\partial^2}{\partial z_i \partial z_j}\psi(\boldsymbol{z}), \tag{54}$$

and the coefficients of the affine connection vanish luckily,

$$\Gamma_{ijk}(\boldsymbol{z}) = 0. \tag{55}$$

Therefore, $\boldsymbol{z}$ is an affine coordinate system, and a geodesic is always written as the linear form

$$\boldsymbol{z}(t) = \boldsymbol{a}t + \boldsymbol{b} \tag{56}$$

with parameter t and constant vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ (see Ref. [27]).

2.4.2 Dual structure

In order to give a dual description related to a convex function $\psi(\boldsymbol{z})$, let us define its gradient,

$$\boldsymbol{z}^* = \nabla\psi(\boldsymbol{z}). \tag{57}$$

This is the Legendre transformation, and the correspondence between $\boldsymbol{z}$ and $\boldsymbol{z}^*$ is one-to-one. Hence, $\boldsymbol{z}^*$ is regarded as another (nonlinear) coordinate system of S. In order to describe the geometry of S in terms of the dual coordinate system $\boldsymbol{z}^*$, we search for a dual convex function $\psi^*\left(\boldsymbol{z}^*\right)$. We can obtain a dual potential function by

$$\psi^*\left(\boldsymbol{z}^*\right) = \max_{\boldsymbol{z}}\left\{\boldsymbol{z} \cdot \boldsymbol{z}^* - \psi(\boldsymbol{z})\right\}, \tag{58}$$

which is convex in $\boldsymbol{z}^*$. The two coordinate systems are dual, since the pair $\boldsymbol{z}$ and $\boldsymbol{z}^*$ satisfies the following relation:

$$\psi(\boldsymbol{z}) + \psi^*\left(\boldsymbol{z}^*\right) - \boldsymbol{z} \cdot \boldsymbol{z}^* = 0. \tag{59}$$

The inverse transformation is given by

$$\boldsymbol{z} = \nabla\psi^*\left(\boldsymbol{z}^*\right). \tag{60}$$

We can define the Bregman divergence $D_{\psi^*}\left[\boldsymbol{z}^* : \boldsymbol{w}^*\right]$ by using $\psi^*\left(\boldsymbol{z}^*\right)$. However, it is easy to prove

$$D_{\psi^*}\left[\boldsymbol{z}^* : \boldsymbol{w}^*\right] = D_\psi\left[\boldsymbol{w} : \boldsymbol{z}\right]. \tag{61}$$

Hence, they are essentially the same, that is the same when $\boldsymbol{w}$ and $\boldsymbol{z}$ are interchanged. It is enough to use only one common divergence.

By simple calculations, we see that the Bregman divergence can be rewritten in the dual form as

$$D[\boldsymbol{z} : \boldsymbol{w}] = \psi(\boldsymbol{z}) + \psi^*\left(\boldsymbol{w}^*\right) - \boldsymbol{z} \cdot \boldsymbol{w}^*. \tag{62}$$

We have thus two coordinate systems $\boldsymbol{z}$ and $\boldsymbol{z}^*$ of S. They define two flat structures such that a curve with parameter t,

$$\boldsymbol{z}(t) = t\boldsymbol{a} + \boldsymbol{b}, \tag{63}$$

is a ψ-geodesic, and

$$\boldsymbol{z}^*(t) = t\boldsymbol{c}^* + \boldsymbol{d}^* \tag{64}$$

is a ψ^*-geodesic, where $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}^*$ and $\boldsymbol{d}^*$ are constants.

A Riemannian metric is defined by two metric tensors $G = \left(g_{ij}\right)$ and $G^* = \left(g^*_{ij}\right)$,

$$g_{ij} = \frac{\partial^2}{\partial z_i \partial z_j}\psi(\boldsymbol{z}), \tag{65}$$

$$g^*_{ij} = \frac{\partial^2}{\partial z^*_i \partial z^*_j}\psi^*\left(\boldsymbol{z}^*\right), \tag{66}$$

in the two coordinate systems, and they are mutual inverses, $G^* = G^{-1}$ [1,27]. Because the squared local distance of two nearby points $\boldsymbol{z}$ and $\boldsymbol{z} + \mathrm{d}\boldsymbol{z}$ is given by

$$D\left[\boldsymbol{z} : \boldsymbol{z} + \mathrm{d}\boldsymbol{z}\right] = \frac{1}{2}\sum g_{ij}(\boldsymbol{z})\mathrm{d}z_i\mathrm{d}z_j, \tag{67}$$

$$= \frac{1}{2}\sum g^*_{ij}\left(\boldsymbol{z}^*\right)\mathrm{d}z^*_i\mathrm{d}z^*_j, \tag{68}$$

they give the same Riemannian distance. However, the two geodesics are different, which are also different from the Riemannian geodesic derived from the Riemannian metric. Hence, the minimality of the curve length does not hold for ψ- and ψ^*-geodesics.

Two geodesic curves (63) and (64) intersect at $t = 0$, when $\boldsymbol{b} = \boldsymbol{d}^*$. In such a case, they are orthogonal if their tangent vectors are orthogonal in the sense of the Riemannian metric. The orthogonality condition can be represented in a simple form in terms of the two dual coordinates as

$$\langle \boldsymbol{a}, \boldsymbol{b}^* \rangle = \sum a_i b_i^* = 0. \tag{69}$$

2.4.3 Pythagorean theorem and projection theorem

A generalized Pythagorean theorem and projection theorem hold for a dually flat space [1]. They are highlights of a dually flat manifold.

Generalized Pythagorean Theorem. Let $\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{q}$ be three points in S such that the ψ^*-geodesic connecting $\boldsymbol{r}$ and $\boldsymbol{s}$ is orthogonal to the ψ-geodesic connecting $\boldsymbol{s}$ and $\boldsymbol{q}$ (Fig. 5). Then,

$$D_\psi[\boldsymbol{r} : \boldsymbol{s}] + D_\psi[\boldsymbol{s} : \boldsymbol{q}] = D_\psi[\boldsymbol{r} : \boldsymbol{q}]. \tag{70}$$

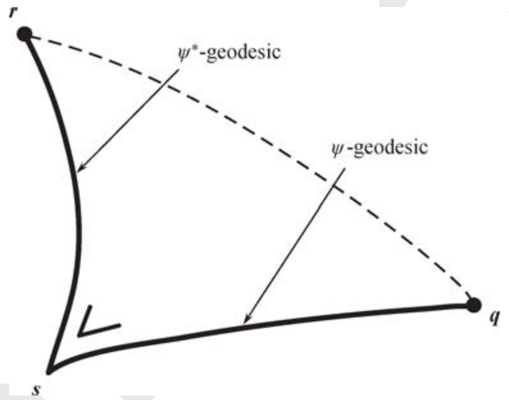

Fig. 5 Pythagorean theorem

When S is a Euclidean space with

$$D[\boldsymbol{r} : \boldsymbol{s}] = \frac{1}{2} \sum (r_i - s_i)^2, \tag{71}$$

(70) is the Pythagorean theorem itself.

Now, we state the projection theorem. Let M be a submanifold of S (Fig. 6). Given $\boldsymbol{p} \in S$, a point $\boldsymbol{q} \in M$ is said to be the ψ-projection (ψ^*-projection) of $\boldsymbol{p}$ to M when the ψ-geodesic (ψ^*-geodesic) connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ is orthogonal to M with respect to the Riemannian metric g_{ij}.

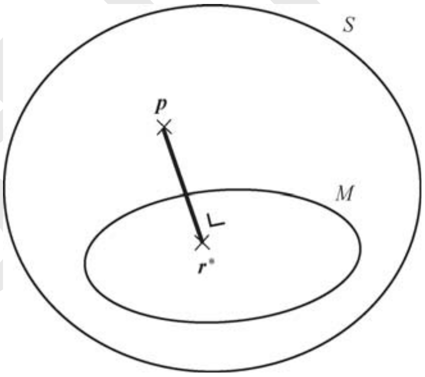

Fig. 6 Projection of $\boldsymbol{p}$ to M

Projection Theorem. Given $\boldsymbol{p} \in S$, the minimizer $\boldsymbol{r}^*$ of $D_\psi(\boldsymbol{p}, \boldsymbol{r})$, $\boldsymbol{r} \in M$, is the ψ^*-projection of $\boldsymbol{p}$ to M, and the minimizer $\boldsymbol{r}^{**}$ of $D_\psi(\boldsymbol{r}, \boldsymbol{p})$, $\boldsymbol{r} \in M$, is the ψ-projection of $\boldsymbol{p}$ to M.

The theorem is useful in various optimization problems searching for the closest point $\boldsymbol{r} \in M$ to a given $\boldsymbol{p}$.

2.4.4 Decomposable convex function and generalization

Let $U(z)$ be a convex function of a scalar z. Then, we have a simple convex function of $\boldsymbol{z}$,

$$\psi(\boldsymbol{z}) = \sum U(z_i). \tag{72}$$

The dual of U is given by

$$U^*(z^*) = \max_z \{zz^* - U(z)\}, \tag{73}$$

and the dual coordinates are

$$z_i^* = U'(z_i). \tag{74}$$

The dual convex function is

$$\psi^*(\boldsymbol{z}^*) = \sum U^*(z_i^*). \tag{75}$$

The ψ-divergence due to U between $\boldsymbol{z}$ and $\boldsymbol{w}$ is decomposable and given by the sum of components [28–30],

$$D_\psi[\boldsymbol{z} : \boldsymbol{w}] = \sum \{U(z_i) + U^*(w_i^*) - z_i w_i^*\}. \tag{76}$$

We have discussed the flat structure given by a convex function due to $U(z)$. We now consider a more general case by using a nonlinear rescaling. Let us define by $r = k(z)$ a nonlinear scale r in the z-axis, where k is a monotone function. If we have a convex function $U(r)$ of the induced variable r, there emerges a new dually flat structure with a new convex function $U(r)$. Given two functions $k(z)$ and $U(r)$, we can introduce a new dually flat Riemannian structure in S, where the flat coordinates are not $\boldsymbol{z}$ but $\boldsymbol{r} = k(\boldsymbol{z})$. There are infinitely many such structures depending on the choice of k and U. We show two typical examples, which give the α-divergence and β-divergence [22] in later sections. Both of them use a power function of the type u^q including the log and exponential function as the limiting cases.

2.5 Invariant flat structure of S

This section focuses on the manifolds which are equipped with both invariant and flat geometrical structure.

2.5.1 Exponential family of probability distributions

The following type of probability distributions of random variable $\boldsymbol{y}$ parameterized by $\boldsymbol{\theta}$,

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = \exp\left\{\sum \theta_i s_i(\boldsymbol{y}) - \psi(\boldsymbol{\theta})\right\}, \tag{77}$$

under a certain measure $\mu(\boldsymbol{y})$ on the space $Y = \{\boldsymbol{y}\}$ is called an exponential family. We denote it now by S. We introduce random variables $\boldsymbol{x} = (x_i)$,

$$x_i = s_i(\boldsymbol{y}), \tag{78}$$

and rewrite (77) in the standard form

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\sum \theta_i x_i - \psi(\boldsymbol{\theta})\right\}, \tag{79}$$

where $\exp\{-\psi(\boldsymbol{\theta})\}$ is the normalization factor satisfying

$$\psi(\boldsymbol{\theta}) = \log \int \exp\left\{\sum \theta_i x_i\right\} \mathrm{d}\mu(\boldsymbol{x}). \tag{80}$$

This is called the free energy in the physics community and is a convex function.

We first give a few examples of exponential families.

1. Gaussian distributions S_{G}

A Gaussian distribution is given by

$$p\left(y; \mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}. \tag{81}$$

Hence, all Gaussian distributions form a two-dimensional manifold S_{G}, where (μ, σ) plays the role of a coordinate system. Let us introduce new variables

$$x_1 = y, \tag{82}$$

$$x_2 = y^2, \tag{83}$$

$$\theta_1 = \frac{\mu}{2\sigma^2}, \tag{84}$$

$$\theta_2 = -\frac{1}{2\sigma^2}. \tag{85}$$

Then, (81) can be rewritten in the standard form

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\theta_1 x_1 + \theta_2 x_2 - \psi(\boldsymbol{\theta})\right\}, \tag{86}$$

where

$$\psi(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} - \log\left(\sqrt{2\pi}\sigma\right). \tag{87}$$

Here, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is a new coordinate system of S_{G}, called the natural or canonical coordinate system.

2. Discrete distributions S_n

Let y be a random variable taking values on $\{0, 1, \ldots, n\}$. By defining

$$x_i = \delta_i(y), \quad i = 1, 2, \ldots, n, \tag{88}$$

and

$$\theta_i = \log \frac{p_i}{p_0}, \tag{89}$$

(4) is rewritten in the standard form

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\sum \theta_i x_i - \psi(\boldsymbol{\theta})\right\}, \tag{90}$$

where

$$\psi(\boldsymbol{\theta}) = -\log p_0. \tag{91}$$

From

$$p_0 = 1 - \sum p_i = 1 - p_0 \sum e^{\theta_i}, \tag{92}$$

we have

$$\psi(\boldsymbol{\theta}) = \log\left(1 + \sum e^{\theta_i}\right). \tag{93}$$

2.5.2 Geometry of exponential family

The Fisher information matrix is defined in a manifold of probability distributions by

$$g_{ij}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial\theta_i} \log p(\boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial}{\partial\theta_j} p(\boldsymbol{x}, \boldsymbol{\theta})\right], \tag{94}$$

where E is expectation with respect to $p(\boldsymbol{x}, \boldsymbol{\theta})$. When S is an exponential family (79), it is calculated as

$$g_{ij}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \psi(\boldsymbol{\theta}). \tag{95}$$

This is a positive-definite matrix and $\psi(\boldsymbol{\theta})$ is a convex function. Hence, this is the Riemannian metric induced from a convex function ψ.

We can also prove that the invariant Riemannian metric derived from an f-divergence is exactly the same as the Fisher metric for any standard convex function f. It is an invariant metric.

A geometry is induced to $S = \{p(\boldsymbol{x}, \boldsymbol{\theta})\}$ through the convex function $\psi(\boldsymbol{\theta})$. Here, $\boldsymbol{\theta}$ is an affine coordinate system, which we call e-affine or e-flat coordinates, "e" representing the exponential structure of (79). By the Legendre transformation of ψ, we have the dual coordinates (we denote them by $\boldsymbol{\eta}$ instead of $\boldsymbol{\theta}^*$),

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}). \tag{96}$$

By calculating the expectation of x_i, we have

$$\mathrm{E}\left[x_i\right] = \eta_i = \frac{\partial}{\partial\theta_i}\psi(\boldsymbol{\theta}). \tag{97}$$

By this reason, the dual parameters $\boldsymbol{\eta} = (\eta_i)$ are called the expectation parameters of an exponential family. We call $\boldsymbol{\eta}$ the m-affine or m-flat coordinates, where "m" implying the mixture structure [1].

The dual convex function is given by the negative of entropy

$$\psi^*(\boldsymbol{\eta}) = \int p\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{\eta})\right) \log p\left(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{\eta})\right) \mathrm{d}\boldsymbol{x}, \tag{98}$$

where $p(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{\eta}))$ is considered as a function of $\boldsymbol{\eta}$. Then, the Bregman divergence is

$$D\left[\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2\right] = \psi\left(\boldsymbol{\theta}_1\right) + \psi^*\left(\boldsymbol{\eta}_2\right) - \boldsymbol{\theta}_1 \cdot \boldsymbol{\eta}_2, \tag{99}$$

where $\boldsymbol{\eta}_i$ is the dual coordinates of $\boldsymbol{\theta}_i$, $i = 1, 2$.

We show that manifolds of exponential families have both invariant and flat geometrical structure. However, they are not unique and it is possible to introduce both invariant flat structure to a mixture family of probability distributions.

Theorem 2 The Bregman divergence is the Kullback-Leibler divergence given by

$$\mathrm{KL}\left[p\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right) : p\left(\boldsymbol{x}, \boldsymbol{\theta}_2\right)\right] = \int p\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right) \log \frac{p\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right)}{p\left(\boldsymbol{x}, \boldsymbol{\theta}_2\right)} \mathrm{d}\boldsymbol{x}. \tag{100}$$

This shows that the KL-divergence is the canonical divergence derived from the dual flat structure of $\psi(\boldsymbol{\theta})$. Moreover, it is proved that the KL-divergence is the unique divergence belonging to both the f-divergence and Bregman divergence classes for manifolds of probability distributions.

For a manifold of positive measures, there are divergences other than the KL-divergence, which are invariant and dually flat, that is, belonging to the classes f-divergences and Bregman divergences at the same time. See the next subsection.

2.5.3 Invariant and flat structure in manifold of positive measures

We consider a manifold M of positive arrays $\boldsymbol{z}, z_i > 0$, where $\sum z_i$ can be arbitrary. We define an invariant and flat structure in M. Here, $\boldsymbol{z}$ is a coordinate system, and consider the f-divergence given by

$$D_f\left[\boldsymbol{z} : \boldsymbol{w}\right] = \sum_{i=1}^{n} z_i f\left(\frac{w_i}{z_i}\right). \tag{101}$$

We now introduce a new coordinate system defined by

$$r_i^{(\alpha)} = k_\alpha\left(z_i\right), \tag{102}$$

where the α-representation of z_i is given by

$$k_\alpha(u) = \begin{cases} \dfrac{2}{1-\alpha}\left(u^{\frac{1-\alpha}{2}} - 1\right), & \alpha \neq 1, \\ \log u, & \alpha = 1. \end{cases} \tag{103}$$

Then, the α-coordinates $\boldsymbol{r}_\alpha$ of M are defined by

$$r_i^{(\alpha)} = k_\alpha\left(z_i\right). \tag{104}$$

We use a convex function $U_\alpha(r)$ of r defined by

$$U_\alpha(r) = \frac{2}{1+\alpha} k_\alpha^{-1}(r). \tag{105}$$

In terms of z, this is a linear function,

$$U_\alpha\left\{r(z)\right\} = \frac{2}{1+\alpha} z, \tag{106}$$

which is not a (strictly) convex function of z but is convex with respect to r. The α-potential function defined by

$$\begin{aligned} \psi_\alpha(\boldsymbol{r}) &= \frac{2}{1+\alpha} \sum k_\alpha^{-1}\left(r_i\right) \\ &= \frac{2}{1+\alpha} \sum \left(1 + \frac{1-\alpha}{2} r_i\right)^{\frac{2}{1-\alpha}} \end{aligned} \tag{107}$$

is a convex function of $\boldsymbol{r}$.

The dual potential is simply given by

$$\psi_\alpha^*\left(\boldsymbol{r}^*\right) = \psi_{-\alpha}\left(\boldsymbol{r}^*\right), \tag{108}$$

and the dual affine coordinates are

$$\boldsymbol{r}^{*(\alpha)} = \boldsymbol{r}^{(-\alpha)} = k_{-\alpha}(\boldsymbol{z}). \tag{109}$$

We can then prove the following theorem.

Theorem 3 The α-divergence of M is the Bregman divergence derived from the α-representation k_α and the linear function U_α of z.

Proof We have the Bregman divergence between $\boldsymbol{r} = k(\boldsymbol{z})$ and $\boldsymbol{s} = k(\boldsymbol{y})$ based on ψ_α as

$$D_\alpha[\boldsymbol{r} : \boldsymbol{s}] = \psi_\alpha(\boldsymbol{r}) + \psi_{-\alpha}\left(\boldsymbol{s}^*\right) - \boldsymbol{r} \cdot \boldsymbol{s}^*, \tag{110}$$

where $\boldsymbol{s}^*$ is the dual of $\boldsymbol{s}$. By substituting (107), (108) and (109) in (110), we see that $D_\alpha[\boldsymbol{r}(\boldsymbol{z}) : \boldsymbol{s}(\boldsymbol{y})]$ is equal to the α-divergence defined in (49).

This proves that the α-divergence for any α belongs to the intersection of the classes of f-divergences and Bregman divergences in M. Hence, it possesses information monotonicity and the induced geometry is dually flat at the same time. However, the constraint $\sum z_i = 1$ is not linear in the flat coordinates $\boldsymbol{r}^\alpha$ or $\boldsymbol{r}^{-\alpha}$ except for the case $\alpha = \pm 1$. Hence, although M is flat for any α, the manifold P of probability distributions is not dually flat except for $\alpha = \pm 1$, and it is a curved submanifold in M. Hence, the α-divergences ($\alpha \neq \pm 1$) do not belong to the class of Bregman divergences in P. This proves that the KL-divergence belongs to both classes of divergences in P and is unique.

The α-divergences are used in many applications, e.g., Refs. [21,25,26].

Theorem 4 The α-divergence is the unique class of divergences sitting at the intersection of the f-divergence and Bregman divergence classes.

Proof The f-divergence is of the form (45) and the decomposable Bregman divergence is of the form

$$D[\boldsymbol{z} : \boldsymbol{y}] = \sum \tilde{U}\left(z_i\right) + \sum \tilde{U}^*\left(y_i\right) - \sum r_i\left(z_i\right) r_i^*\left(y_i\right), \tag{111}$$

where

$$\tilde{U}(z) = U(k(z)), \tag{112}$$

and so on. When they are equal, we have f, r, and r^* that satisfy

$$z f\left(\frac{y}{z}\right) = r(z) r^*(y) \tag{113}$$

except for additive terms depending only on z or y. Differentiating the above by y, we get

$$f'\left(\frac{y}{z}\right) = r(z) {r^*}'(y). \tag{114}$$

By putting $x = y$, $y = 1/z$, we get

$$f'(xy) = r\left(\frac{1}{y}\right) r^{*'}(x). \tag{115}$$

Hence, by putting $h(u) = \log f'(u)$, we have

$$h(xy) = s(x) + t(y), \tag{116}$$

where $s(x) = \log r^{*'}(x)$, $t(y) = r(1/y)$. By differentiating the above equation with respect to x and putting $x = 1$, we get

$$h'(y) = \frac{c}{y}. \tag{117}$$

This proves that f is of the form

$$f(u) = \begin{cases} cu^{\frac{1+\alpha}{2}}, & \alpha \neq \pm 1, \\ cu \log u, & \alpha = 1, \\ c \log u, & \alpha = -1. \end{cases} \tag{118}$$

By changing the above into the standard form, we arrive at the α-divergence.

2.6 β-divergence

It is useful to introduce a family of β-divergences, which are not invariant but dually flat, having the structure of Bregman divergences. This is used in the field of machine learning and statistics [28,29].

Use $\boldsymbol{z}$ itself as a coordinate system of M; that is, the representation function is $k(z) = z$. The β-divergence [30] is induced by the potential function

$$U_\beta(z) = \begin{cases} \dfrac{1}{\beta+1}(1+\beta z)^{\frac{\beta+1}{\beta}}, & \beta > 0, \\ \exp z, & \beta = 0. \end{cases} \tag{119}$$

The β-divergence is thus written as

$$D_\beta[\boldsymbol{z} : \boldsymbol{y}] = \begin{cases} \dfrac{1}{\beta+1}\sum\left(y_i^{\beta+1} - z_i^{\beta+1}\right) - \dfrac{1}{\beta}\sum z_i^\beta (y_i - z_i), & \beta > 0, \\ \sum\left(z_i \log\dfrac{z_i}{y_i} + y_i - z_i\right), & \beta = 0. \end{cases} \tag{120}$$

It is the KL-divergence when $\beta = 0$, but it is different from the α-divergence when $\beta > 0$.

Minami and Eguchi [30] demonstrated that statistical inference based on the β-divergence ($\beta > 0$) is robust. Such idea has been applied to machine learning in Murata et al. [29]. The β-divergence induces a dually flat structure in M. Since its flat coordinates are $\boldsymbol{z}$, the restriction

$$\sum z_i = 1 \tag{121}$$

is a linear constraint. Hence, the manifold P of the probability distributions is also dually flat, where $\boldsymbol{z}$, $\sum z_i = 1$, is its flat coordinates. The dual flat coordinates are

$$z_i^* = \begin{cases} (1+\beta z_i)^{\frac{1}{\beta}}, & \beta \neq 0, \\ \exp z_i, & \beta = 0, \end{cases} \tag{122}$$

depending on β.

3 Information geometry of alternative minimization

Given a divergence function $D[\boldsymbol{z} : \boldsymbol{w}]$ on a space S, we encounter the problem of minimizing $D[\boldsymbol{z} : \boldsymbol{w}]$ under the constraints that $\boldsymbol{z}$ belongs to a submanifold $M \subset S$ and $\boldsymbol{w}$ belongs to another submanifold $E \subset S$. A typical example is the EM algorithm [6], where M represents a manifold of partially observed data and E represents a statistical model to which the true distribution is assumed to belong [8]. The EM algorithm is popular in statistics, machine learning and others [7,31,32].

3.1 Geometry of alternative minimization

When a divergence function $D[\boldsymbol{z} : \boldsymbol{w}]$ is given in a manifold S, we consider two submanifolds M and E, where we assume $\boldsymbol{z} \in M$ and $\boldsymbol{w} \in E$. The divergence of two submanifolds M and E is given by

$$D[M : E] = \min_{\boldsymbol{z} \in M, \boldsymbol{w} \in E} D[\boldsymbol{z} : \boldsymbol{w}] = D\left[\boldsymbol{z}^* : \boldsymbol{w}^*\right], \tag{123}$$

where $\boldsymbol{z}^* \in M$ and $\boldsymbol{w}^* \in E$ are a pair of the points that attain the minimum. When M and E intersect, $D[M : E] = 0$, and $\boldsymbol{z}^* = \boldsymbol{w}^*$ lies at the intersection.

An alternative minimization is a procedure to calculate $(\boldsymbol{z}^*, \boldsymbol{w}^*)$ iteratively (see Fig. 7):

1) Take an initial point $\boldsymbol{z}^0 \in M$ for $t = 0$. Repeat steps 2 and 3 for $t = 0, 1, 2, \ldots$ until convergence.

2) Calculate

$$\boldsymbol{w}^t = \arg\min_{\boldsymbol{w} \in E} D\left[\boldsymbol{z}^t : \boldsymbol{w}\right]. \tag{124}$$

3) Calculate

$$\boldsymbol{z}^{t+1} = \arg\min_{\boldsymbol{z} \in M} D\left[\boldsymbol{z} : \boldsymbol{w}^{t+1}\right]. \tag{125}$$

When the direct minimization with respect to one

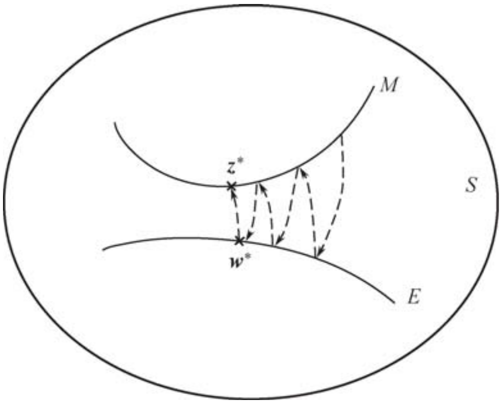

Fig. 7 Alternative minimization

variable is computationally difficult, we may use a decremental procedure:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta_t \nabla_{\boldsymbol{w}} D\left[\boldsymbol{z}^t : \boldsymbol{w}^t\right], \tag{126}$$

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \eta_t \nabla_{\boldsymbol{z}} D\left[\boldsymbol{z}^{t+1} : \boldsymbol{w}^t\right], \tag{127}$$

where η_t is a learning constant and $\nabla_{\boldsymbol{w}}$ and $\nabla_{\boldsymbol{z}}$ are gradients with respect to $\boldsymbol{w}$ and $\boldsymbol{z}$, respectively.

It should also be noted that the natural gradient (Riemannian gradient of D) [33] is given by

$$\tilde{\nabla}_{\boldsymbol{z}} D[\boldsymbol{z} : \boldsymbol{w}] = G^{-1}(\boldsymbol{z}) \frac{\partial}{\partial \boldsymbol{z}} D[\boldsymbol{z} : \boldsymbol{w}], \tag{128}$$

where

$$G = \frac{\partial^2}{\partial \boldsymbol{z} \partial \boldsymbol{z}} D[\boldsymbol{z} : \boldsymbol{w}]_{|\boldsymbol{w}=\boldsymbol{z}}. \tag{129}$$

Hence, this becomes the Newton-type method.

3.2 Alternative projections

When $D[\boldsymbol{z} : \boldsymbol{w}]$ is a Bregman divergence derived from a convex function ψ, the space S is equipped with a dually flat structure. Given $\boldsymbol{w}$, the minimization of $D[\boldsymbol{z} : \boldsymbol{w}]$ with respect to $\boldsymbol{z} \in M$ is given by the ψ-projection of $\boldsymbol{w}$ to M,

$$\arg\min_{\boldsymbol{z}} D[\boldsymbol{z} : \boldsymbol{w}] = \prod_M^{(\psi)} \boldsymbol{w}. \tag{130}$$

This is unique when M is ψ^*-flat. On the other hand, given $\boldsymbol{z}$, the minimization with respect to $\boldsymbol{w}$ is given by the ψ^*-projection,

$$\arg\min_{\boldsymbol{w}} D[\boldsymbol{z} : \boldsymbol{w}] = \prod_M^{(\psi^*)} \boldsymbol{z}. \tag{131}$$

This is unique when M is ψ-flat.

3.3 EM algorithm

Let us consider a statistical model $p(\boldsymbol{z}; \boldsymbol{\xi})$ with random variables $\boldsymbol{z}$, parameterized by $\boldsymbol{\xi}$. The set of all such distributions $E = \{p(\boldsymbol{z}; \boldsymbol{\xi})\}$ forms a submanifold included in the entire manifold of probability distributions

$$S = \{p(\boldsymbol{z})\}. \tag{132}$$

Assume that vector random variable $\boldsymbol{z}$ is divided into two parts $\boldsymbol{z} = (\boldsymbol{h}, \boldsymbol{x})$ and the values of $\boldsymbol{x}$ are observed but $\boldsymbol{h}$ are not. Non-observed random variables $\boldsymbol{h} = (h_i)$ are called missing data or hidden variables. It is possible to eliminate unobservable $\boldsymbol{h}$ by summing up $p(\boldsymbol{h}, \boldsymbol{x})$ over all $\boldsymbol{h}$, and we have a new statistical model,

$$\tilde{p}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{\xi}), \tag{133}$$

$$\tilde{E} = \{\tilde{p}(\boldsymbol{x}; \boldsymbol{\xi})\}. \tag{134}$$

This $\tilde{p}(\boldsymbol{x}, \boldsymbol{\xi})$ is the marginal distribution of $p(\boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\xi})$. Based on observed $\boldsymbol{x}$, we can estimate $\hat{\boldsymbol{\xi}}$, for example, by maximizing the likelihood,

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} \tilde{p}(\boldsymbol{x}, \boldsymbol{\xi}). \tag{135}$$

However, in many problems $p(\boldsymbol{z}; \boldsymbol{\xi})$ has a simple form while $\tilde{p}(\boldsymbol{x}; \boldsymbol{\xi})$ does not. The marginal distribution $\tilde{p}(\boldsymbol{x}, \boldsymbol{\xi})$ has a complicated form and it is computationally difficult to estimate $\boldsymbol{\xi}$ from $\boldsymbol{x}$. It is much easier to calculate $\hat{\boldsymbol{\xi}}$ when $\boldsymbol{z}$ is observed. The EM algorithm is a powerful method used in such a case [6].

The EM algorithm consists of iterative procedures to use values of the hidden variables $\boldsymbol{h}$ guessed from observed $\boldsymbol{x}$ and the current estimator $\hat{\boldsymbol{\xi}}$. Let us assume that n iid data $\boldsymbol{z}_t = (\boldsymbol{h}_t, \boldsymbol{x}_t)$, $t = 1, \ldots, n$, are generated from a distribution $p(\boldsymbol{z}, \boldsymbol{\xi})$. If data $\boldsymbol{h}_t$ are not missing, we have the following empirical distribution:

$$\bar{p}(\boldsymbol{h}, \boldsymbol{x} \left| \bar{\boldsymbol{h}}, \bar{\boldsymbol{x}} \right.) = \frac{1}{n} \sum_t \delta\left(\boldsymbol{x} - \boldsymbol{x}_t\right) \delta\left(\boldsymbol{h} - \boldsymbol{h}_t\right), \tag{136}$$

where $\bar{\boldsymbol{h}}$ and $\bar{\boldsymbol{x}}$ represent the sets of data $(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots)$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$ and δ is the delta function. When S is an exponential family, $\bar{p}(\boldsymbol{h}, \boldsymbol{x} \left| \bar{\boldsymbol{h}}, \bar{\boldsymbol{x}} \right.)$ is the distribution given by the sufficient statistics composed of $\left(\bar{\boldsymbol{h}}, \bar{\boldsymbol{x}}\right)$.

The hidden variables are not observed in reality, and the observed part $\bar{\boldsymbol{x}}$ cannot determine the empirical distribution $\bar{p}(\boldsymbol{x}, \boldsymbol{y})$ (136). To overcome this difficulty, let us use a conditional probability $q(\boldsymbol{h}|\boldsymbol{x})$ of $\boldsymbol{h}$ when $\boldsymbol{x}$ is observed. We then have

$$p(\boldsymbol{h}, \boldsymbol{x}) = q(\boldsymbol{h}|\boldsymbol{x}) p(\boldsymbol{x}). \tag{137}$$

In the present case, the observed part $\bar{\boldsymbol{x}}$ determines the empirical distribution $\bar{p}(\boldsymbol{x})$,

$$\bar{p}(\boldsymbol{x}) = \frac{1}{n} \sum_t \delta\left(\boldsymbol{x} - \boldsymbol{x}_t\right), \tag{138}$$

but the conditional part $q(\boldsymbol{h}|\boldsymbol{x})$ remains unknown. Taking this into account, we define a submanifold based on the observed data $\bar{\boldsymbol{x}}$,

$$M\left(\bar{\boldsymbol{x}}\right) = \left\{p(\boldsymbol{h}, \boldsymbol{x}) \left| p(\boldsymbol{h}, \boldsymbol{x}) = q(\boldsymbol{h}|\boldsymbol{x}) \bar{p}(\boldsymbol{x}) \right.\right\}, \tag{139}$$

where $q(\boldsymbol{h}|\boldsymbol{x})$ is arbitrary.

In the observable case, we have an empirical distribution $\bar{p}(\boldsymbol{h}, \boldsymbol{x}) \in S$ from the observation, which does not usually belong to our model E. The maximum likelihood estimator $\hat{\boldsymbol{\xi}}$ is the point in E that minimizes the KL-divergence from $\bar{p}$ to E,

$$\hat{\boldsymbol{\xi}} = \arg\min \mathrm{KL}\left[\bar{p}(\boldsymbol{h}, \boldsymbol{x}) : p(\boldsymbol{x}, \boldsymbol{\xi})\right]. \tag{140}$$

In the present case, we do not know $\bar{p}$ but know $M(\bar{\boldsymbol{x}})$ which includes the unknown $\bar{p}(\boldsymbol{h}, \boldsymbol{x})$. Hence, we search for $\hat{\boldsymbol{\xi}} \in E$ and $\hat{q}(\boldsymbol{h}|\boldsymbol{x}) \in M(\bar{\boldsymbol{x}})$ that minimizes

$$\mathrm{KL}[M : E]. \tag{141}$$

The pair of minimizers gives the points which define the divergence of two submanifolds $M(\bar{\boldsymbol{x}})$ and E. The maximum likelihood estimator is a minimizer of (141).

Given partial observations $\bar{\boldsymbol{x}}$, the EM algorithm consists of the following interactive procedures, $t = 0, 1, \ldots$. We begin with an arbitrary initial guess $\boldsymbol{\xi}_0$ and $q_0(\boldsymbol{h}|\boldsymbol{x})$, and construct a candidate $p_t(\boldsymbol{h}, \boldsymbol{x}) = q_t(\boldsymbol{h}|\boldsymbol{x})\bar{p}(\boldsymbol{x}) \in M = M(\bar{\boldsymbol{x}})$ for $t = 0, 1, 2, \ldots$. We search for the next guess $\boldsymbol{\xi}_{t+1}$ that minimizes $\mathrm{KL}\left[\boldsymbol{p}_t(\boldsymbol{h}, \boldsymbol{x}) : p(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{\xi})\right]$. This is the same as maximizing the pseudo-log-likelihood

$$L_t(\boldsymbol{\xi}) = \sum_{\boldsymbol{h}, \boldsymbol{x}_i} p_t(\boldsymbol{h}, \boldsymbol{x}_i) \log p(\boldsymbol{h}, \boldsymbol{x}_i; \boldsymbol{\xi}). \tag{142}$$

This is called the M-step, that is the m-projection of $p_t(\boldsymbol{h}, \boldsymbol{x}) \in M$ to E. Let the maximum value be $\boldsymbol{\xi}_{t+1}$, which is the estimator at $t + 1$.

We then search for the next candidate $p_t(\boldsymbol{h}, \boldsymbol{x}) \in M$ that minimizes $\mathrm{KL}\left[p(\boldsymbol{h}, \boldsymbol{x}) : p\left(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{\xi}_{t+1}\right)\right]$. The minimizer is denoted by $p_{t+1}(\boldsymbol{h}, \boldsymbol{x})$. This distribution is used to calculate the expectation of the new log likelihood function $L_{t+1}(\boldsymbol{\xi})$. Hence, this is called the E-step. The following theorem [8] plays a fundamental role in calculating the expectation with respect to $p_t(\boldsymbol{h}, \boldsymbol{x})$.

Theorem 5 Along the e-geodesic of projecting $p(\boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\xi}_t)$ to $M(\bar{\boldsymbol{x}})$, the conditional probability distribution $q_t(\boldsymbol{h}|\boldsymbol{x})$ is kept invariant. Hence, by the e-projection of $p(\boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\xi}_t)$ to M, the resultant conditional distribution is given by

$$q_t(\boldsymbol{h}\,|\boldsymbol{x}_i) = p(\boldsymbol{h}\,|\boldsymbol{x}_i; \boldsymbol{\xi}_t). \tag{143}$$

This shows that the e-projection is given by

$$q_t(\boldsymbol{h}, \boldsymbol{x}) = p(\boldsymbol{h}\,|\boldsymbol{x}; \boldsymbol{\xi}_t)\, \delta(\boldsymbol{x} - \bar{\boldsymbol{x}}). \tag{144}$$

Hence, the theorem implies that the next expected log likelihood is calculated by

$$L_{t+1}(\boldsymbol{\xi}) = \sum_{\boldsymbol{h}, \boldsymbol{x}_i} p\left(\boldsymbol{h}\,|\boldsymbol{x}_i; \boldsymbol{\xi}_{t+1}\right) \log p(\boldsymbol{h}, \boldsymbol{x}_i, \boldsymbol{\xi}), \tag{145}$$

which is to be maximized.

The EM algorithm is formally described as follows.

1) Begin with an arbitrary initial guess $\boldsymbol{\xi}_0$ at $t = 0$, and repeat the E-step and M-step for $t = 0, 1, 2, \ldots$, until convergence.
2) E-step: Use the e-projection of $p(\boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\xi}_t)$ to $M(\bar{\boldsymbol{x}})$ to obtain
$$q_t(\boldsymbol{h}|\boldsymbol{x}) = p(\boldsymbol{h}\,|\boldsymbol{x}; \boldsymbol{\xi}_t), \tag{146}$$
and calculate the conditional expectation
$$L_t(\boldsymbol{\xi}) = \sum_{\boldsymbol{h}, \boldsymbol{x}_i} p(\boldsymbol{h}\,|\boldsymbol{x}_i; \boldsymbol{\xi}_t) \log p(\boldsymbol{h}, \boldsymbol{x}_i; \boldsymbol{\xi}). \tag{147}$$
3) M-step: Calculate the maximizer of $L_t(\boldsymbol{\xi})$, that is the m-projection of $p_t(\boldsymbol{h}, \boldsymbol{x})$ to E,
$$\boldsymbol{\xi}_{t+1} = \arg\max L_t(\boldsymbol{\xi}). \tag{148}$$

4 Information geometry of Ying-Yang machine

4.1 Recognition model and generating model

We study the Ying-Yang mechanism proposed by Xu [9–14] from the information geometry point of view. Let us consider a hierarchical stochastic system, which consists of two vector random variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Let $\boldsymbol{x}$ be a lower level random variable representing a primitive description of the world, while let $\boldsymbol{y}$ be a higher level random variable representing a more conceptual or elaborated description. Obviously, $\boldsymbol{x}$ and $\boldsymbol{y}$ are closely correlated. One may consider them as information representations in the brain. Sensory information activates neurons in a primitive sensory area of the brain, and its firing pattern is represented by $\boldsymbol{x}$. Components x_i of $\boldsymbol{x}$ can be binary variables taking values 0 and 1, or analog values representing firing rates of neurons. The conceptual information activates a higher-order area of the brain, with a firing pattern $\boldsymbol{y}$.

A probability distribution $p(\boldsymbol{x})$ of $\boldsymbol{x}$ reflects the information structure of the outer world. When $\boldsymbol{x}$ is observed, the brain processes this primitive information and activates a higher-order area to recognize its higher-order representation $\boldsymbol{y}$. Its mechanism is represented by a conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$. This is possibly specified by a parameter $\boldsymbol{\xi}$, in the form of $p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\xi})$. Then, the joint distribution is given by

$$p(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\xi}) = p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\xi})p(\boldsymbol{x}). \tag{149}$$

The parameters $\boldsymbol{\xi}$ would be synaptic weights of neurons to generate pattern $\boldsymbol{y}$. When we receive information $\boldsymbol{x}$, it is transformed to higher-order conceptual information $\boldsymbol{y}$. This is the bottom-up process and we call this a recognition model.

Our brain can work in the reverse way. From conceptual information pattern $\boldsymbol{y}$, a primitive pattern $\boldsymbol{x}$ will be generated by the top-down neural mechanism. It is represented by a conditional probability $q(\boldsymbol{x}|\boldsymbol{y})$, which will be parameterized by $\boldsymbol{\zeta}$, as $q(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\zeta})$. Here, $\boldsymbol{\zeta}$ will represent the top-down neural mechanism. When the probability distribution of $\boldsymbol{y}$ is $q(\boldsymbol{y};\boldsymbol{\zeta})$, the entire process is represented by another joint probability distribution

$$q(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\zeta}) = q(\boldsymbol{y};\boldsymbol{\zeta})q(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\zeta}). \tag{150}$$

We have thus two stochastic mechanisms $p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\xi})$ and $q(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\zeta})$. Let us denoted by S the set of all the joint probability distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$. The recognition model

$$M_R = \{p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\xi})\} \tag{151}$$

forms its submanifold, and the generative model

$$M_G = \{q(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\zeta})\} \tag{152}$$

forms another submanifold.

This type of information mechanism is proposed by Xu [9,14], and named the Ying-Yang machine. The Yang machine (male machine) is responsible for the recognition model $\{p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\xi})\}$ and the Ying machine (female machine) is responsible for the generative model $\{q(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\zeta})\}$. These two models are different, but should be put in harmony (Ying-Yang harmony).

4.2 Ying-Yang harmony

When a divergence function $D\left[p(\boldsymbol{x},\boldsymbol{y}):q(\boldsymbol{x},\boldsymbol{y})\right]$ is defined in S, we have the divergence between the two models M_R and M_G,

$$D\left[M_R:M_G\right] = \min_{\boldsymbol{\xi},\boldsymbol{\zeta}} D\left[p(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\xi}):q(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\zeta})\right]. \tag{153}$$

When M_R and M_G intersect,

$$D\left[M_R:M_G\right] = 0 \tag{154}$$

at the intersecting points. However, they do not intersect in general,

$$M_R \cap M_G = \phi. \tag{155}$$

It is desirable that the recognition and generating models are closely related. The minimizer of D gives such a harmonized pair called the Ying-Yang harmony [14].

A typical divergence is the Kullback-Leibler divergence KL$[p:q]$, which has merits of being invariant and generating dually flat geometrical structure. In this case, we can define the two projections, e-projection and m-projection. Let $p^*(\boldsymbol{y},\boldsymbol{x})$ and $q^*(\boldsymbol{y},\boldsymbol{x})$ be the pair of minimizers of $D[p:q]$. Then, the m-projection of q^* to M_R is p^*,

$$\prod\nolimits_{M_R}^{(m)} q^* = p^*, \tag{156}$$

and the e-projection of p^* to M_G is q^*,

$$\prod\nolimits_{M_G}^{(e)} p^* = q^*. \tag{157}$$

These relations hold for any dually flat divergence, generated by a convex function ψ over S.

An iterative algorithm to realize the Ying-Yang harmony is, for $t=0,1,2,\ldots$,

$$p_{t+1} = \prod\nolimits_{M_R}^{(m)} q_t, \tag{158}$$

$$q_{t+1} = \prod\nolimits_{M_G}^{(e)} p_{t+1}. \tag{159}$$

This is a typical alternative minimization algorithm.

4.3 Various models of Ying-Yang structure

A Yang machine receives signal $\boldsymbol{x}$ from the outside, and processes it by generating a higher-order signal $\boldsymbol{y}$, which is stochastically generated by using the conditional probability $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi})$. The performance of the machine is improved by learning, where the parameter $\boldsymbol{\xi}$ is modified step-by-step. We give two typical examples of information processing; pattern classification and self-organization.

Learning of a pattern classifier is a supervised learning process, where teacher signal z is given.

1. pattern classifier

Let $C=\{C_1,\ldots,C_m\}$ denotes the set of categories, and each $\boldsymbol{x}$ belongs to one of them. Given an $\boldsymbol{x}$, a machine processes it and gives an answer $\boldsymbol{y}_\kappa$, where $\boldsymbol{y}_\kappa$, $\kappa=1,\ldots,m$, correspond to one of the categories. In other words, $\boldsymbol{y}_\kappa$ is a pattern representing C_κ. The output $\boldsymbol{y}$ is generated from $\boldsymbol{x}$:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\xi}) + \boldsymbol{n}, \tag{160}$$

where $\boldsymbol{n}$ is a zero-mean Gaussian noise with covariance matrix $\sigma^2 I$, I being the identity matrix. Then, the conditional distribution is given by

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\xi}) = c\exp\left\{-\frac{1}{2\sigma^2}\left|\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\xi})\right|^2\right\}, \tag{161}$$

where c is the normalization constant.

The function $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\xi})$ is specified by a number of parameters $\boldsymbol{\xi}$. In the case of multilayer perceptron, it is given by

$$y_i = \sum_j v_{ij}\varphi\left\{\sum_k w_{jk}x_k - h_j\right\}. \tag{162}$$

Here, w_{jk} is the synaptic weight from input x_k to the jth hidden unit, h_j is its threshold, and φ is a sigmoidal function. The output of the jth hidden unit is hence $\varphi\left(\sum w_{jk}x_k - h_j\right)$ and is connected linearly to the ith output unit with synaptic weight v_{ij}.

In the case of supervised learning, a teacher signal $\boldsymbol{y}_\kappa$ representing the category C_κ is given, when the input pattern is $\boldsymbol{x} \in C_\kappa$. Since the current output of the Yang machine is $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\xi})$, the error of the machine is

$$l\left(\boldsymbol{\xi}, \boldsymbol{x}, \boldsymbol{y}_\kappa\right) = \frac{1}{2}\left|\boldsymbol{y}_\kappa - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\xi})\right|^2. \tag{163}$$

Therefore, the stochastic gradient learning rule modifies the current $\boldsymbol{\xi}$ to $\boldsymbol{\xi} + \Delta\boldsymbol{\xi}$ by

$$\Delta\boldsymbol{\xi} = -\varepsilon\frac{\partial l}{\partial \boldsymbol{\xi}}, \tag{164}$$

where ε is the learning constant and $\nabla l = \left(\frac{\partial l}{\partial \xi_1}, \ldots, \frac{\partial l}{\partial \xi_m}\right)$ is the gradient of l with respect to $\boldsymbol{\xi}$. Since the space of $\boldsymbol{\xi}$ is Riemannian, the natural gradient learning rule [33] is given by

$$\Delta\boldsymbol{\xi} = -\varepsilon G^{-1}\nabla l, \tag{165}$$

where G is the Fisher information matrix derived from the conditional distribution.

2. Non-supervised learning and clustering

When signals $\boldsymbol{x}$ in the outside world are divided into a number of clusters, it is expected that Yang machine reproduces such a clustering structure. An example of clustered signals is the Gaussian mixture

$$p(\boldsymbol{x}) = \sum_{\kappa=1}^{m} w_\kappa \exp\left\{-\frac{1}{2\sigma^2}\left|\boldsymbol{x} - \boldsymbol{y}_\kappa\right|^2\right\}, \tag{166}$$

where $\boldsymbol{y}_\kappa$ is the center of cluster C_κ. Given $\boldsymbol{x}$, one can see which cluster it belongs to by calculating the output $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\xi}) + \boldsymbol{n}$. Here, $\boldsymbol{y}$ is an inner representation of the clusters and its calculation is controlled by $\boldsymbol{\xi}$. There are a number of clustering algorithms. See, e.g., Ref. [9]. We can use them to modify $\boldsymbol{\xi}$. Since no teacher signals exist, this is unsupervised learning.

Another typical non-supervised learning is implemented by the self-organization model given by Amari and Takeuchi [34].

The role of the Ying machine is to generate a pattern $\tilde{\boldsymbol{x}}$ from a conceptual information pattern $\boldsymbol{y}$. Since the mechanism of generating $\tilde{\boldsymbol{x}}$ is governed by the conditional probability $q(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\zeta})$, its parameter is modified to fit well the corresponding Yang machine $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})$. As $\boldsymbol{\xi}$ changes by learning, $\boldsymbol{\zeta}$ changes correspondingly. The Ying machine can be used to improve the primitive representation $\boldsymbol{x}$ by filling missing information and removing noise by using $\tilde{\boldsymbol{x}}$.

4.4 Bayesian Ying-Yang machine

Let us consider joint probability distributions $p(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{x}, \boldsymbol{y})$ which are equal. Here, the Ying machine and the Yang machine are identical, and hence the machines are in perfect matching. Now consider $\boldsymbol{y}$ as the parameters to specify the distribution of $\boldsymbol{x}$, and we have a parameterized statistical model

$$M_{\text{Ying}} = \{p(\boldsymbol{x}|\boldsymbol{y})\}, \tag{167}$$

generated by the Ying mechanism. Here, the parameter $\boldsymbol{y}$ is a random variable, and hence the Bayesian framework is introduced. The prior probability distribution of $\boldsymbol{y}$ is

$$p(\boldsymbol{y}) = \int p(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{x}. \tag{168}$$

The Bayesian framework to estimate $\boldsymbol{y}$ from observed $\boldsymbol{x}$, or more rigorously, a number of independent observations, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, is as follows. Let $\bar{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be observed data and

$$p\left(\bar{\boldsymbol{x}}|\boldsymbol{y}\right) = \prod_{i=1}^{n} p\left(\boldsymbol{x}_i|\boldsymbol{y}\right). \tag{169}$$

The prior distribution of $\boldsymbol{y}$ is $p(\boldsymbol{y})$, but after observing $\boldsymbol{x}$, the posterior distribution is

$$p(\boldsymbol{y}|\bar{\boldsymbol{x}}) = \frac{p(\bar{\boldsymbol{x}}, \boldsymbol{y})}{\int p(\bar{\boldsymbol{x}}, \boldsymbol{y})\mathrm{d}\boldsymbol{y}}. \tag{170}$$

This is the Yang mechanism, and the

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\bar{\boldsymbol{x}}) \tag{171}$$

is the Bayesian posterior estimator.

A full exponential family $p(\boldsymbol{x}, \boldsymbol{y}) = \exp\{\sum y_i x_i + k(\boldsymbol{x}) + w(\boldsymbol{y}) - \psi(\boldsymbol{x}, \boldsymbol{y})\}$ is an interesting special case. Here, $\boldsymbol{y}$ is the natural parameter for the exponential family of distributions,

$$q(\boldsymbol{x}|\boldsymbol{y}) = \exp\left\{\sum y_i x_i + k(\boldsymbol{x}) - \psi(\boldsymbol{y})\right\}. \tag{172}$$

The family of distributions $M_{\text{Ying}} = \{q(\boldsymbol{x}|\boldsymbol{y})\}$ specified by parameters $\boldsymbol{y}$ is again an exponential family. It has a dually flat Riemannian structure.

Dually to the above, the family of the posterior distributions form a manifold $M_{\text{Yang}} = \{p(\boldsymbol{y}|\boldsymbol{x})\}$, consisting of

$$p(\boldsymbol{y}|\boldsymbol{x}) = \exp\left\{\sum x_i y_i + w(\boldsymbol{y}) - \psi^*(\boldsymbol{y})\right\}. \tag{173}$$

It is again an exponential family, where $\boldsymbol{x}$ $(\bar{\boldsymbol{x}} = \sum \boldsymbol{x}_i/n)$ is the natural parameter. It defines a dually flat Riemannian structure, too.

When the probability distribution $p(\boldsymbol{y})$ of the Yang machine includes a hyper parameter $p(\boldsymbol{y}) = p(\boldsymbol{y}, \boldsymbol{\zeta})$, we have a family $q(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\zeta})$. It is possible that the Yang machine have a different parametric family $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\xi})$. Then, we need to discuss the Ying mechanism and the Yang mechanism separately, keeping the Ying-Yang harmony.

This opens a new perspective to the Bayesian framework to be studied in future from the Ying-Yang standpoint. Information geometry gives a useful tool for analyzing it.

5 Geometry of belief propagation

A direct application of information geometry to machine learning is studied here. When a number of correlated random variables $z_1, \ldots, z_N$ exist, we need to treat their joint probability distribution $q(z_1, \ldots, z_N)$. Here, we assume for simplicity's sake that z_i is binary, taking values 1 and -1. Among all variables z_i, some are observed and the others are not. Therefore, z_i are divided into two parts, $\{x_1, \ldots, x_n\}$ and $\{y_{n+1}, \ldots, y_N\}$, where the values of $\boldsymbol{y} = (y_{n+1}, \ldots, y_N)$ are observed. Our problem is to estimate the values of unobserved variables $\boldsymbol{x} = (x_1, \ldots, x_n)$ based on observed $\boldsymbol{y}$. This is a typical problem of stochastic inference. We use the conditional probability distribution $q(\boldsymbol{x}|\boldsymbol{y})$ for estimating $\boldsymbol{x}$.

The graphical model or the Bayesian network is used to represent stochastic interactions among random variables [35]. In order to estimate $\boldsymbol{x}$, the belief propagation (BP) algorithm [15] uses a graphical model or a Bayesian network, and it provides a powerful algorithm in the field of artificial intelligence. However, its performance is not clear when the underlying graph has loopy structure. The present section studies this problem by using information geometry due to Ikeda, Tanaka and Amari [16,17]. The CCCP algorithm [36,37] is another powerful procedure for finding the BP solution. Its geometry is also studied.

5.1 Stochastic inference

We use a conditional probability $q(\boldsymbol{x}|\boldsymbol{y})$ to estimate the values of $\boldsymbol{x}$. Since $\boldsymbol{y}$ is observed and fixed, we hereafter denote it simply by $q(\boldsymbol{x})$, suppressing observed variables $\boldsymbol{y}$, but it always means $q(\boldsymbol{x}|\boldsymbol{y})$, depending on $\boldsymbol{y}$.

We have two estimators of $\boldsymbol{x}$. One is the maximum likelihood estimator that maximizes $q(\boldsymbol{x})$:

$$\hat{\boldsymbol{x}}_{\text{mle}} = \arg\max_{\boldsymbol{x}} q(\boldsymbol{x}). \tag{174}$$

However, calculation of $\hat{\boldsymbol{x}}_{\text{mle}}$ is computationally difficult when n is large, because we need to search for the maximum among 2^n candidates $\boldsymbol{x}$'s. Another estimator $\tilde{\boldsymbol{x}}$ tries to minimize the expected number of errors of n component random variables $\tilde{x}_1, \ldots, \tilde{x}_n$. When $q(\boldsymbol{x})$ is known, the expectation of x_i is

$$\mathrm{E}[x_i] = \sum_{\boldsymbol{x}} x_i q(\boldsymbol{x}) = \sum x_i q_i(x_i). \tag{175}$$

This depends only on the marginal distribution of x_i,

$$q_i(x_i) = \sum{}' q(x_1, \ldots, x_n), \tag{176}$$

where summation $\sum'$ is taken over all $x_1, \ldots, x_n$ except for x_i. The expectation is denoted by

$$\begin{aligned}\eta_i &= \mathrm{E}[x_i] = \mathrm{Prob}[x_i = 1] - \mathrm{Prob}[x_i = -1] \\ &= q_i(1) - q_i(-1).\end{aligned} \tag{177}$$

When $\eta_i > 0$, $\mathrm{Prob}[x_i = 1] > \mathrm{Prob}[x_i = -1]$, so that, $\tilde{x}_i = 1$, otherwise $\tilde{x}_i = -1$. Therefore,

$$\tilde{x}_i = \mathrm{sgn}(\eta_i) \tag{178}$$

is the estimator that minimizes the number of errors in the components of $\tilde{\boldsymbol{x}}$, where

$$\mathrm{sgn}(u) = \begin{cases} 1, & u \geqslant 0, \\ -1, & u < 0. \end{cases} \tag{179}$$

The marginal distribution $q_i(x_i)$ is given in terms of η_i by

$$\mathrm{Prob}\{x_i = 1\} = \frac{1 + \eta_i}{2}. \tag{180}$$

Therefore, our problem reduces to calculation of η_i, the expectation of x_i. However, exact calculation is computationally heavy when n is large, because (176) requires summation over 2^{n-1} $\boldsymbol{x}$'s. Physicists use the mean field approximation for this purpose [38]. The belief propagation is another powerful method of calculating it approximately by iterative procedures.

5.2 Graphical structure and random Markov structure

Let us consider a general probability distribution $p(\boldsymbol{x})$. In the special case where there are no interactions among $x_1, \ldots, x_n$, they are independent, and the probability is written as the product of component distributions $p_i(x_i)$. Since we can always write the probabilities, in the binary case, as

$$p_i(1) = \frac{e^{h_i}}{\psi}, \quad p_i(-1) = \frac{e^{-h_i}}{\psi}, \tag{181}$$

where

$$\psi = e^{h_i} + e^{-h_i}, \tag{182}$$

we have

$$p_i(x_i) = \exp\{h_i x_i - \psi(h_i)\}. \tag{183}$$

Hence,

$$p(\boldsymbol{x}) = \exp\left\{\sum h_i x_i - \psi(\boldsymbol{h})\right\}, \tag{184}$$

where

$$\psi(\boldsymbol{h}) = \sum \psi(h_i), \tag{185}$$

$$\psi_i(h_i) = \log\{\exp(h_i) + \exp(-h_i)\}. \tag{186}$$

When $x_1, \ldots, x_n$ are not independent but their interactions exist, we denote the interaction among k variables $x_{i_1}, \ldots, x_{i_k}$ by a product term $x_{i_1} \cdot \cdots \cdot x_{i_k}$,

$$c(\boldsymbol{x}) = x_{i_1} \cdot \cdots \cdot x_{i_k}. \tag{187}$$

When there are many interaction terms, $p(\boldsymbol{x})$ is written as

$$p(\boldsymbol{x}) = \exp\left\{\sum_i h_i x_i + \sum_r s_r c_r(\boldsymbol{x}) - \psi\right\}, \tag{188}$$

where ψ is the normalization factor, $c_r(\boldsymbol{x})$ represents a monomial interaction term such as

$$c_r(\boldsymbol{x}) = x_{i_1} \cdot \cdots \cdot x_{i_k}, \tag{189}$$

where r is an index showing a subset $\{i_1, \ldots, i_k\}$ of indices $\{1, 2, \ldots, n\}$, $k \geqslant 2$ and s_r is the intensity of such interaction.

When all interactions are pairwise, $k = 2$, we have $c_r(\boldsymbol{x}) = x_i x_j$ for $r = (i, j)$. To represent the structure of pairwise interactions, we use a graph $G = \{N, B\}$, in which we have n nodes $N_1, \ldots, N_n$ representing random variables $x_1, \ldots, x_n$ and b branches $B_1, \ldots, B_b$. A branch B_r connects two nodes N_i and N_j, where r denotes the pair (x_i, x_j) of interaction. G is a non-directed graph having n nodes and b branches. When a graph has no loops, it is a tree graph, otherwise, it is a loopy graph. There are many physical and engineering problems having a graphical structure. They are, for example, spin glass, error-correcting codes, etc.

Interactions are not necessarily pairwise, and there may exist many interactions among more than two variables. Hence, we consider a general case (188), where r includes $\{i_1, \ldots, i_k\}$, and b is the number of interactions. This is an example of the random Markov field, where each set $r = \{i_1, \ldots, i_k\}$ forms a clique of the random graph.

We consider the following statistical model on graph G or a random Markov field specified by two vector parameters $\boldsymbol{\theta}$ and $\boldsymbol{v}$,

$$M = \{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{v})\}, \tag{190}$$

$$p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{v}) = \exp\left\{\sum \theta_i x_i + \sum v_r c_r(\boldsymbol{x}) - \psi(\boldsymbol{\theta}, \boldsymbol{v})\right\}. \tag{191}$$

This is an exponential family, forming a dually flat manifold, where $(\boldsymbol{\theta}, \boldsymbol{v})$ is its e-affine coordinates. Here, $\psi(\boldsymbol{\theta}, \boldsymbol{v})$ is a convex function from which a dually flat structure is derived. The model M includes the true distribution $q(\boldsymbol{x})$ in (13), i.e.,

$$q(\boldsymbol{x}) = \exp\{\boldsymbol{h} \cdot \boldsymbol{x} + \boldsymbol{s} \cdot \boldsymbol{c}(\boldsymbol{x}) - \psi(\boldsymbol{h}, \boldsymbol{s})\}, \tag{192}$$

which is given by $\boldsymbol{\theta} = \boldsymbol{h}$ and $\boldsymbol{v} = \boldsymbol{s} = (s_r)$. Our job is to find

$$\boldsymbol{\eta}^* = \mathrm{E}_q[\boldsymbol{x}], \tag{193}$$

where expectation E_q is taken with respect to $q(\boldsymbol{x})$, and $\boldsymbol{\eta}$ is the dual (m-affine) coordinates corresponding to $\boldsymbol{\theta}$.

We consider $b+1$ e-flat submanifolds, $M_0, M_1, \ldots, M_b$ in M. M_0 is the manifold of independent distributions given by $p_0(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{h} \cdot \boldsymbol{x} + \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi(\boldsymbol{\theta})\}$, and its e-coordinates are $\boldsymbol{\theta}$.

For each branch or clique r, we also consider an e-flat submanifold M_r,

$$M_r = \{p(\boldsymbol{x}, \boldsymbol{\zeta}_r)\}, \tag{194}$$

$$p(\boldsymbol{x}, \boldsymbol{\zeta}_r) = \exp\{\boldsymbol{h} \cdot \boldsymbol{x} + s_r c_r(\boldsymbol{x}) + \boldsymbol{\zeta}_r \cdot \boldsymbol{x} - \psi_r(\boldsymbol{\zeta}_r)\}, \tag{195}$$

where $\boldsymbol{\zeta}_r$ is e-affine coordinates. Note that a distribution in M_r includes only one nonlinear interaction term $c_r(\boldsymbol{x})$. Therefore, it is computationally easy to calculate $\mathrm{E}_r[\boldsymbol{x}]$ or $\mathrm{Prob}\{x_i = 1\}$ with respect to $p(\boldsymbol{x}, \boldsymbol{\zeta}_r)$. All of M_0 and M_r play proper roles for finding $\mathrm{E}_q[\boldsymbol{x}]$ of the target distribution (192).

5.3 m-projection

Let us first consider a special submanifold M_0 of independent distributions.

We show that the expectation $\boldsymbol{\eta}^* = \mathrm{E}_q[\boldsymbol{x}]$ is given by the m-projection of $q(\boldsymbol{x}) \in M$ to M_0 [39,40]. We denote it by

$$\hat{p}(\boldsymbol{x}) = \prod\nolimits_0 q(\boldsymbol{x}). \tag{196}$$

This is the point in M_0 that minimizes the KL-divergence from q to M_0,

$$\hat{p}(\boldsymbol{x}) = \underset{p \in M_0}{\arg\min}\, \mathrm{KL}[q : p]. \tag{197}$$

It is known that $\hat{p}$ is the point in M_0 such that the m-geodesic connecting q and $\hat{p}$ is orthogonal to M_0, and is unique, since M_0 is e-flat.

The m-projection has the following property.

Theorem 6 The m-projection of $q(\boldsymbol{x})$ to M_0 does not change the expectation of $\boldsymbol{x}$.

Proof Let $\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta}^*) \in M_0$ be the m-projection of $q(\boldsymbol{x})$. By differentiating $\mathrm{KL}[q : p(\boldsymbol{x}, \boldsymbol{\theta})]$ with respect to $\boldsymbol{\theta}$, we have

$$\sum \boldsymbol{x} q(\boldsymbol{x}) - \frac{\partial}{\partial \boldsymbol{\theta}} \psi_0(\boldsymbol{\theta}^*) = 0. \tag{198}$$

The first term is the expectation of $\boldsymbol{x}$ with respect to $q(\boldsymbol{x})$ and the second term is the $\boldsymbol{\eta}$-coordinates of $\boldsymbol{\theta}^*$ which is the expectation of $\boldsymbol{x}$ with respect to $p_0(\boldsymbol{x}, \boldsymbol{\theta}^*)$. Hence, they are equal.

5.4 Clique manifold M_r

The m-projection of $q(\boldsymbol{x})$ to M_0 is written as the product of the marginal distributions

$$\prod\nolimits_0 q(\boldsymbol{x}) = \prod_{i=1}^{n} q_i(x_i), \tag{199}$$

which is what we want to obtain. Its coordinates are given by $\boldsymbol{\theta}^*$, from which we easily have $\boldsymbol{\eta}^* = \mathrm{E}_{\boldsymbol{\theta}^*}[\boldsymbol{x}]$, since M_0 consists of independent distributions. However, it is difficult to calculate the m-projection, or to obtain $\boldsymbol{\theta}^*$ or $\boldsymbol{\eta}^*$ directly. Physicists use the mean field approximation for this purpose. The belief propagation is another method of obtaining an approximation of $\boldsymbol{\theta}^*$. Here, the nodes and branches (cliques) play an important role.

Since the difficulty of calculations is given rise to by the existence of a large number of branches or cliques B_r, we consider a model M_r which includes only one branch or clique B_r. Since a branch (clique) manifold M_r includes only one nonlinear term $c_r(\boldsymbol{x})$, it is written as

$$p(\boldsymbol{x}, \boldsymbol{\zeta}_r) = \exp\left\{\boldsymbol{h} \cdot \boldsymbol{x} + s_r c_r(\boldsymbol{x}) + \boldsymbol{\zeta}_r \cdot \boldsymbol{x} - \psi_r\right\}, \tag{200}$$

where $\boldsymbol{\zeta}_r$ is a free vector parameter. Comparing this with $q(\boldsymbol{x})$, we see that the sum of all the nonlinear terms

$$\sum_{r' \neq r} s_{r'} c_{r'}(\boldsymbol{x}) \tag{201}$$

except for $s_r c_r(\boldsymbol{x})$ is replaced by a linear term $\boldsymbol{\zeta}_r \cdot \boldsymbol{x}$. Hence, by choosing an adequate $\boldsymbol{\zeta}_r$, $p(\boldsymbol{x}, \boldsymbol{\zeta}_r)$ is expected to give the same expectation $\mathrm{E}_q[\boldsymbol{x}]$ as $q(\boldsymbol{x})$ or its very good approximation. Moreover, M_r includes only one nonlinear term so that it is easy to calculate the expectation of $\boldsymbol{x}$ with respect to $p(\boldsymbol{x}, \boldsymbol{\zeta}_r)$. In the following algorithm, all branch (clique) manifolds M_r cooperate iteratively to give a good entire solution.

5.5 Belief propagation

We have the true distribution $q(\boldsymbol{x})$, b clique distributions $p(\boldsymbol{x}, \boldsymbol{\zeta}_r) \in M_r$, $\boldsymbol{r} = 1, \ldots, b$, and an independent distribution $p_0(\boldsymbol{x}, \boldsymbol{\theta}) \in M_0$. All of them join force to search for a good approximation $\hat{p}(\boldsymbol{x}) \in M_0$ of $q(\boldsymbol{x})$.

Let $\boldsymbol{\zeta}_r$ be the current solution which M_r believes to give a good approximation of $q(\boldsymbol{x})$. We then project it to M_0, giving the equivalent solution $\boldsymbol{\theta}_r$ in M_0 having the same expectation,

$$p(\boldsymbol{x}, \boldsymbol{\theta}_r) = \prod\nolimits_0 p_r(\boldsymbol{x}, \boldsymbol{\zeta}_r). \tag{202}$$

We abbreviate this as

$$\boldsymbol{\theta}_r = \prod\nolimits_0 \boldsymbol{\zeta}_r. \tag{203}$$

The $\boldsymbol{\theta}_r$ specifies the independent distribution equivalent to $p_r(\boldsymbol{x}, \boldsymbol{\zeta}_r)$ in the sense that they give the same expectation of $\boldsymbol{x}$. Since $\boldsymbol{\theta}_r$ includes both the effect of the single nonlinear term $s_r c_r(\boldsymbol{x})$ specific to M_r and that due to the linear term $\boldsymbol{\zeta}_r$ in (200),

$$\boldsymbol{\xi}_r = \boldsymbol{\theta}_r - \boldsymbol{\zeta}_r \tag{204}$$

represents the effect of the single nonlinear term $s_r c_r(\boldsymbol{x})$. This is the linearized version of $s_r c_r(\boldsymbol{x})$. Hence, M_r knows that the linearized version of $s_r c_r(\boldsymbol{x})$ is $\boldsymbol{\xi}_r$, and broadcasts to all the other models M_0 and $M_{r'}$ $(r' \neq r)$ that the linear counterpart of $s_r c_r(\boldsymbol{x})$ is $\boldsymbol{\xi}_r$. Receiving these messages from all M_r, M_0 guesses that the equivalent linear term of $\sum s_r c_r(\boldsymbol{x})$ will be

$$\boldsymbol{\theta} = \sum \boldsymbol{\xi}_r. \tag{205}$$

Since M_r in turn receives messages $\boldsymbol{\xi}_{r'}$ from all other $M_{r'}$, M_r uses them to form a new $\boldsymbol{\zeta}'_r$,

$$\boldsymbol{\zeta}'_r = \sum_{r' \neq r} \boldsymbol{\xi}_r, \tag{206}$$

which is the linearized version of $\sum_{r' \neq r} s_{r'} c_{r'}(\boldsymbol{x})$. This process is repeated.

The above algorithm is written as follows, where the current candidates $\boldsymbol{\theta}^t \in M_0$ and $\boldsymbol{\zeta}^t_r \in M_r$ at time t, $t = 0, 1, \ldots$, are renewed iteratively until convergence [17].

Geometrical BP Algorithm:

1) Put $t = 0$, and start with initial guesses $\boldsymbol{\zeta}^0_r$, for example, $\boldsymbol{\zeta}^0_r = 0$.

2) For $t = 0, 1, 2, \ldots$, m-project $p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}^t_r\right)$ to M_0, and obtain the linearized version of $s_r c_r(\boldsymbol{x})$,

$$\boldsymbol{\xi}^{t+1}_r = \prod_0 p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}^t_r\right) - \boldsymbol{\zeta}^t_r. \tag{207}$$

3) Summarize all the effects of $s_r c_r(\boldsymbol{x})$, to give

$$\boldsymbol{\theta}^{t+1} = \sum_r \boldsymbol{\xi}^{t+1}_r. \tag{208}$$

4) Update $\boldsymbol{\zeta}_r$ by

$$\boldsymbol{\zeta}^{t+1}_r = \sum_{r' \neq r} \boldsymbol{\xi}^{t+1}_{r'} = \boldsymbol{\theta}^{t+1} - \boldsymbol{\xi}^{t+1}_r. \tag{209}$$

5) Stop when the algorithm converges.

5.6 Analysis of the solution of the algorithm

Let us assume that the algorithm has converged to $\{\boldsymbol{\zeta}^*_r\}$ and $\boldsymbol{\theta}^*$. Then, our solution is given by $p_0(\boldsymbol{x}, \boldsymbol{\theta}^*)$, from which

$$\eta^*_i = \mathrm{E}_0[x_i] \tag{210}$$

is easily calculated. However, this might not be the exact solution but is only an approximation. Therefore, we need to study its properties. To this end, we study the relations among the true distribution $q(\boldsymbol{x})$, the converged clique distributions $p^*_r = p_r(\boldsymbol{x}, \boldsymbol{\zeta}^*_r)$ and the converged marginalized distribution $p^*_0 = p_0(\boldsymbol{x}, \boldsymbol{\theta}^*)$. They are written as

$$q(\boldsymbol{x}) = \exp\left\{\boldsymbol{h} \cdot \boldsymbol{x} + \sum s_r c_r(\boldsymbol{x}) - \psi\right\}, \tag{211}$$

$$p_0(\boldsymbol{x}, \boldsymbol{\theta}^*) = \exp\left\{\boldsymbol{h} \cdot \boldsymbol{x} + \sum \boldsymbol{\xi}^*_r \cdot \boldsymbol{x} - \psi_0\right\}, \tag{212}$$

$$p_r(\boldsymbol{x}, \boldsymbol{\zeta}^*_r) = \exp\left\{\boldsymbol{h} \cdot \boldsymbol{x} + \sum_{r' \neq r} \boldsymbol{\xi}^*_{r'} \cdot \boldsymbol{x} + s_r c_r(\boldsymbol{x}) - \psi_r\right\}. \tag{213}$$

The convergence point satisfies the two conditions:

$$1.\quad m\text{-condition}:\ \boldsymbol{\theta}^* = \prod\nolimits_0 p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^*\right); \tag{214}$$

$$2.\quad e\text{-condition}:\quad \boldsymbol{\theta}^* = \frac{1}{b-1}\sum_{r=1}^{b}\boldsymbol{\zeta}_r^*. \tag{215}$$

Let us consider the m-flat submanifold M^* connecting $p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right)$ and all of $p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^*\right)$,

$$M^* = \Big\{p(\boldsymbol{x})\,\Big|p(\boldsymbol{x}) = d_0 p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right) + \sum d_r p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^*\right); \quad d_0 + \sum d_r = 1\Big\}. \tag{216}$$

This is a mixture family composed of $p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right)$ and $p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^*\right)$.

We also consider the e-flat submanifold E^* connecting $p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right)$ and $p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^*\right)$,

$$E^* = \Big\{p(\boldsymbol{x})\big|\log p(\boldsymbol{x}) = v_0 \log p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right) + \sum_r v_r \log p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r\right) - \psi\Big\}. \tag{217}$$

Theorem 7 The m-condition implies that M^* intersects all of M_0 and M_r orthogonally. The e-condition implies that E^* include the true distribution $q(\boldsymbol{x})$.

Proof The m-condition guarantees that m-projection of p_r^* to M_0 is p_0^*. Since the m-geodesic connecting p_r^* and p_0 is included in M^* and the expectations are equal,

$$\boldsymbol{\eta}_r^* = \boldsymbol{\eta}_0^*, \tag{218}$$

and M^* is orthogonal to M_0 and M_r. The e-condition is easily shown by putting $c_0 = -(b-1)$, $c_r = 1$, because

$$q(\boldsymbol{x}) = c p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right)^{-(b-1)} \prod p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^*\right) \tag{219}$$

from (211)–(213), where $c = e^{-\psi_q}$.

The algorithm searches for $\boldsymbol{\theta}^*$ and $\boldsymbol{\zeta}_r^*$ until both the m- and e-conditions are satisfied. If M^* includes $q(\boldsymbol{x})$, its m-projection to M_0 is $p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^*\right)$. Hence, $\boldsymbol{\theta}^*$ gives the true solution, but this is not guaranteed. Instead, E^* includes $q(\boldsymbol{x})$.

The following theorem is known and easy to prove.

Theorem 8 When the underlying graph is a tree, both E^* and M^* include $q(\boldsymbol{x})$, and the algorithm gives the exact solution.

5.7 CCCP procedure for belief propagation

The above algorithm is essentially the same as the BP algorithm given by Pearl [15] and studied by many followers. It is iterative search procedures for finding M^* and E^*. In steps $2) - 5)$, it uses m-projection to obtain new $\boldsymbol{\zeta}_r$'s, until the m-condition is satisfied eventually. The m-projections of all M_r become identical when the m-condition is satisfied. On the other hand, in each step of 3), we obtain $\boldsymbol{\theta} \in M_0$ such that the e-condition is satisfied. Hence, throughout the procedures, the e-condition is satisfied.

We have a different type of algorithm, which searches for $\boldsymbol{\theta}$ that eventually satisfies the e-condition, while the m-condition is always satisfied. Such an algorithm is proposed by Yuille [36] and Yuille and Rangarajan [37]. It consists of the two steps, beginning with initial guess $\boldsymbol{\theta}^0$:

CCCP Algorithm:

Step 1 (inner loop): Given $\boldsymbol{\theta}^t$, calculate $\left\{\boldsymbol{\zeta}_r^{t+1}\right\}$ by solving

$$\prod\nolimits_0 p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^{t+1}\right) = b\boldsymbol{\theta}^t - \sum \boldsymbol{\zeta}_r^{t+1}. \tag{220}$$

Step 2 (outer loop): Given $\left\{\boldsymbol{\zeta}_r^{t+1}\right\}$, calculate $\boldsymbol{\theta}^{t+1}$ by

$$\boldsymbol{\theta}^{t+1} = b\boldsymbol{\theta}^t - \sum_r \boldsymbol{\zeta}_r^{t+1}. \tag{221}$$

The inner loop use an iterative procedures to obtain new $\left\{\boldsymbol{\zeta}_r^{t+1}\right\}$ in such a way that they satisfy the m-condition together with old $\boldsymbol{\theta}^t$. Hence, the m-condition is always satisfied, and we search for new $\boldsymbol{\theta}^{t+1}$ until the e-condition is satisfied.

We may simplify the inner loop by

$$\prod\nolimits_0 p_r\left(\boldsymbol{x}, \boldsymbol{\zeta}_r^{t+1}\right) = p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^t\right), \tag{222}$$

which can be solved directly. This gives a computationally easier procedure. There are some difference in the basin of attraction for the original and simplified procedures.

We have formulated a number of algorithms for stochastic reasoning in terms of information geometry. It not only clarifies the procedures intuitively but make it possible to analyze the stability of the equilibrium and speed of convergence. Moreover, we can estimate the error of estimation by using the curvatures of E^* and M^*. The new type of free energy is also defined. See Refs. [16,17] for more details.

6 Conclusions

We have given the dual geometrical structure in manifolds of probability distributions, positive measures or arrays, matrices, tensors and others. The dually flat structure is given from a convex function in general, which gives a Riemannian metric and a pair of dual flatness criteria. The information invariancy and dual flat structure are explained by using divergence functions without rigorous differential geometrical terminology. The dual geometrical structure is applied to various engineering problems, which include statistical inference, machine learning, optimization, signal processing and neural networks. Applications to alternative optimization, Ying-Yang machine and belief propagations are shown from the geometrical point of view.

References

1. Amari S, Nagaoka H. Methods of Information Geometry. New York: Oxford University Press, 2000
2. Csiszár I. Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica, 1967, 2: 299–318
3. Bregman L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 1967, 7(3): 200–217
4. Eguchi S. Second order efficiency of minimum contrast estimators in a curved exponential family. The Annals of Statistics, 1983, 11(3): 793–803
5. Chentsov N N. Statistical Decision Rules and Optimal Inference. Rhode Island, USA: American Mathematical Society, 1982 (originally published in Russian, Moscow: Nauka, 1972)
6. Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B, 1977, 39(1): 1–38
7. Csiszár I, Tusnády G. Information geometry and alternating minimization procedures. Statistics and Decisions, 1984, Supplement Issue 1: 205–237
8. Amari S. Information geometry of the EM and em algorithms for neural networks. Neural Networks, 1995, 8(9): 1379–1408
9. Xu L. Bayesian Ying-Yang machine, clustering and number of clusters. Pattern Recognition Letters, 1997, 18(11–13): 1167–1178
10. Xu L. RBF nets, mixture experts, and Bayesian Ying-Yang learning. Neurocomputing, 1998, 19(1–3): 223–257
11. Xu L. Bayesian Kullback Ying-Yang dependence reduction theory. Neurocomputing, 1998, 22(1–3): 81–111
12. Xu L. BYY harmony learning, independent state space, and generalized APT financial analyses. IEEE Transactions on Neural Networks, 2001, 12(4): 822–849
13. Xu L. Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models. International Journal of Neural Systems, 2001, 11(1): 43–69
14. Xu L. BYY harmony learning, structural RPCL, and topological self-organizing on mixture models. Neural Networks, 2002, 15(8–9): 1125–1151
15. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann, 1988
16. Ikeda S, Tanaka T, Amari S. Information geometry of turbo and low-density parity-check codes. IEEE Transactions on Information Theory, 2004, 50(6): 1097–1114
17. Ikeda S, Tanaka T, Amari S. Stochastic reasoning, free energy, and information geometry. Neural Computation, 2004, 16(9): 1779–1810
18. Csiszár I. Information measures: A critical survey. In: Transactions of the 7th Prague Conference. 1974, 83–86
19. Csiszár I. Axiomatic characterizations of information measures. Entropy, 2008, 10(3): 261–273
20. Ali M S, Silvey S D. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society. Series B, 1966, 28(1): 131–142
21. Amari S. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Transactions on Information Theory, 2009, 55(11): 4925–4931
22. Cichocki A, Adunek R, Phan A H, Amari S. Nonnegative Matrix and Tensor Factorizations. John Wiley, 2009
23. Havrda J, Charvát F. Quantification method of classification process: Concept of structural α-entropy. Kybernetika, 1967, 3: 30–35
24. Chernoff H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics, 1952, 23(4): 493–507
25. Matsuyama Y. The α-EM algorithm: Surrogate likelihood maximization using α-logarithmic information measures. IEEE Transactions on Information Theory, 2002, 49(3): 672–706
26. Amari S. Integration of stochastic models by minimizing α-divergence. Neural Computation, 2007, 19(10): 2780–2796
27. Amari S. Information geometry and its applications: Convex function and dually flat manifold. In: Nielsen F ed. Emerging Trends in Visual Computing. Lecture Notes in Computer Science, Vol 5416. Berlin: Springer-Verlag, 2009, 75–102
28. Eguchi S, Copas J. A class of logistic-type discriminant functions. Biometrika, 2002, 89(1): 1–22
29. Murata N, Takenouchi T, Kanamori T, Eguchi S. Information geometry of U-boost and Bregman divergence. Neural Computation, 2004, 16(7): 1437–1481
30. Minami M, Eguchi S. Robust blind source separation by beta-divergence. Neural Computation, 2002, 14(8): 1859–1886
31. Byrne W. Alternating minimization and Boltzmann machine learning. IEEE Transactions on Neural Networks, 1992, 3(4): 612–620
32. Amari S, Kurata K, Nagaoka H. Information geometry of Boltzmann machines. IEEE Transactions on Neural Networks, 1992, 3(2): 260–271
33. Amari S. Natural gradient works efficiently in learning. Neural Computation, 1998, 10(2): 251–276
34. Amari S, Takeuchi A. Mathematical theory on formation of category detecting nerve cells. Biological Cybernetics, 1978, 29(3): 127–136
35. Jordan M I. Learning in Graphical Models. Cambridge, MA: MIT Press, 1999
36. Yuille A L. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. Neural Computation, 2002, 14(7): 1691–1722
37. Yuille A L, Rangarajan A. The concave-convex procedure. Neural Computation, 2003, 15(4): 915–936
38. Opper M, Saad D. Advanced Mean Field Methods—Theory and Practice. Cambridge, MA: MIT Press, 2001

39. Tanaka T. Information geometry of mean-field approximation. Neural Computation, 2000, 12(8): 1951–1968
40. Amari S, Ikeda S, Shimokawa H. Information geometry and mean field approximation: The α-projection approach. In: Opper M, Saad D, eds. Advanced Mean Field Methods—Theory and Practice. Cambridge, MA: MIT Press, 2001, 241–257

Shun-ichi Amari was born in Tokyo, Japan, on January 3, 1936. He graduated from the Graduate School of the University of Tokyo in 1963 majoring in mathematical engineering and received Degree of Doctor of Engineering.

He worked as an Associate Professor at Kyushu University and the University of Tokyo, and then a Full Professor at the University of Tokyo, and is now Professor-Emeritus. He moved to RIKEN Brain Science Institute and served as Director for five years and is now Senior Advisor. He has been engaged in research in wide areas of mathematical science and engineering, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, and information sciences. In particular, he has devoted himself to mathematical foundations of neural networks, including statistical neurodynamics, dynamical theory of neural fields, associative memory, self-organization, and general learning theory. Another main subject of his research is information geometry initiated by himself, which applies modern differential geometry to statistical inference, information theory, control theory, stochastic reasoning, and neural networks, providing a new powerful method to information sciences and probability theory.

Dr. Amari is past President of International Neural Networks Society and Institute of Electronic, Information and Communication Engineers, Japan. He received Emanuel R. Piore Award and Neural Networks Pioneer Award from the IEEE, the Japan Academy Award, C&C Award and Caianiello Memorial Award. He was the founding co-editor-in-chief of Neural Networks, among many other journals.