Dualistic Dynamical Systems in the Framework of Information Geometry

Akio Fujiwara and Shun-ichi Amari

METR 93-17 November 1993

Dualistic Dynamical Systems in the Framework of Information Geometry

Akio Fujiwara and Shun-ichi Amari,
University of Tokyo[*]

Abstract

Dualistic properties of a gradient flow on a manifold M associated with a dualistic structure (g, ∇, ∇^*) is studied from an information geometrical viewpoint. Some useful applications are also investigated.

Keywords : information geometry, gradient flow

[*] Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

1 Introduction

Recently, dynamical systems that solve some optimization problems have been proposed and studied actively, which seems to be motivated mainly by the hope of analog parallel computing. Since Karmarkar [1] proposed the projective scaling algorithm, a kind of interior point method for solving linear programming problems, its modifications and analyses have been made by many researchers. Further, some gradient flows on a Lie group are usefully applied for solving optimization problems, such as sorting of lists (Brockett [2]), diagonalization of matrices (Brockett [2], Nakamura [3]), matching problems (Brockett [4]), etc. In analyzing these dynamical systems, differential geometry played an essential role (Bayer and Lagarias [5], Brockett [6]), which is the same as in classical mechanics [7] or in nonequilibrium statistical physics [8].

On the other hand, infomation geometry, originated from the geometric study of the manifold of probability distributions (Amari [9][10], Nagaoka and Amari [11], Amari *et al.* [12], Chentsov [13], etc), has been successfully applied to many fields, such as statistical inference (Amari [10], Kumon and Amari [14], Amari and Kumon [15], Okamoto *et al.* [16]), control systems theory (Amari [17], Ohara and Amari [18]), multiterminal information theory (Amari and Han [19], Amari [20]), and neural networks (Amari *et al.* [21]), etc. It is tempting to apply information geometry to the analysis of dynamical systems mentioned above. Obata *et al.* [22] applied it, though very naively, to some nonequilibrium processes, to find that the Uhlenbeck–Ornstein process is a geodesic motion with respect to the exponential connection on a Gaussian model. Nakamura [23] also pointed out that certain gradient flows on Gaussian and multinomial distributions can be characterized as completely integrable Hamiltonian systems. It seems that these results are important suggestions toward the connection between two seemingly unrelated fields, information geometry and integrable dynamical systems.

In this paper, general dualistic properties of a gradient flow on a manifold M associated with a dualistic structure (g, ∇, ∇^*) is studied from an information geometrical point of view. In order to demonstrate the importance of the information geometrical viewpoint in analyzing dynamical systems, some applications are given to such fields as statistical inference, neural networks, EM algorithms, linear programming problems, and mathematical biology.

2 Dualistic geometry

We first give a brief summary of dualistic geometry. For details, consult [10][24]. Mathematicians are also studying the dualistic geometry (see Nomizu and Simon [25], Kurose [26]). Let M be a Riemannian manifold with metric g. Two affine connections represented by covariant derivatives ∇ and ∇^* on M are said to be *dual* with respect to g if, for any vector field A, B, and C on M,

$$Ag(B|C) = g(\nabla_A B|C) + g(B|\nabla^*_A C), \tag{2.1}$$

where $g(B|C)$ denotes the inner product of B and C with respect to the metric g. If the torsions and the Riemannian curvatures of M with respect to the connections ∇ and ∇^* vanish, M is said to be *dually flat* or simply *flat*. This does not imply that the manifold is Euclidean, because the Riemannian curvature due to the Levi–Civita connection does not necessarily vanish. For a dually flat manifold M, a pair of divergences are defined in the following way. We first construct mutually dual affine coordinates on M. Since M is flat with respect to ∇ and ∇^*, M has ∇–affine coordinate $\theta = [\theta^i]$ and ∇^*–affine coordinate $\eta = [\eta_i]$ such that $\nabla_A \partial_i = 0$ for the natural basis vector field $\partial_i = \partial/\partial\theta^i$ and that $\nabla^*_A \partial^j = 0$ for the natural basis vector field $\partial^j = \partial/\partial\eta_j$. Moreover, we can choose θ and η such that

$$g(\partial_i \mid \partial^j) = \delta_i^j \tag{2.2}$$

holds at any point in M. Then it was proved that there exist such potential functions $\psi(\theta)$, $\phi(\eta)$ on M satisfying

$$\theta^i = \partial^i \phi(\eta), \qquad \eta_i = \partial_i \psi(\theta), \qquad \psi(\theta) + \phi(\eta) - \theta \cdot \eta = 0,$$

where $\theta \cdot \eta = \theta^i \eta_i$ and the Einstein's summation convention $\theta^i \eta_i = \sum_i \theta^i \eta_i$ is assumed hereafter. By using these potentials, we define the ∇–divergence D from $p_1 \in M$ to $p_2 \in M$ as

$$D(p_1 \parallel p_2) = \psi(\theta_2) + \phi(\eta_1) - \theta_2 \cdot \eta_1,$$

where η_1 and θ_2 are the η– and θ–coordinates of points p_1 and p_2, respectively. According to the duality, the ∇^*–divergence D^* is given as

$$D^*(p_1 \parallel p_2) = D(p_2 \parallel p_1).$$

Note that our nomenclature of divergences is different from the convention in Amari's book [10], where the ∇–divergence D from $p_1 \in M$ to $p_2 \in M$ is defined in a converse manner as

$$D(p_1 \parallel p_2) = \psi(\theta_1) + \phi(\eta_2) - \theta_1 \cdot \eta_2.$$

The statistical manifold, which is a family of probability distributions, is a good example of such a structure. In this case, the ∇–divergence is proved to be equal to the Kullback–Leibler information.

Next, we tackle the converse problem. When a potential $U(\theta)$ is given on a manifold M, where $\theta = [\theta^i]$ is a local coordinate system of M, we construct a natural dually flat structure on it. In the following, we restrict ourselves to a domain Θ in which the potential $U(\theta)$ is a convex function with respect to θ. We first define another coordinate system $\eta = [\eta_i]$ and the corresponding potential $V(\eta)$ by a Legendre transformation as

$$\eta_j = \partial_j U(\theta), \qquad V(\eta) = \max_{\theta\in\Theta} \{\theta^i \eta_i - U(\theta)\}.$$

Then $\theta^j = \partial^j V(\eta)$ holds, and the pair (θ, η) satisfies the identity

$$U(\theta) + V(\eta) - \theta \cdot \eta = 0.$$

The metric h on M is defined by

$$h_{ij} = \partial_i \partial_j U(\theta).$$

This definition can be rewritten as

$$h_{ij} = \frac{\partial \eta_j}{\partial \theta^i},$$

which readily leads to the relation

$$h^{ij} = \frac{\partial \theta^j}{\partial \eta_i} = \partial^i \partial^j V(\eta),$$

where (h^{ij}) is the inverse of matrix (h_{ji}). This indicates that the coordinate systems θ and η are mutually dual with respect to h in the sense of Eq. (2.2). Further let us set

$$T_{ijk} = \partial_i \partial_j \partial_k U(\theta),$$

and define the α–connection $\nabla^{(\alpha)}$ by its parameters

$$\Gamma^{(\alpha)}_{ijk} = h(\nabla^{(\alpha)}_{\partial_i} \partial_j \mid \partial_k) = [ij; k] - \frac{\alpha}{2} T_{ijk},$$

where $[ij;k]$ is the Levi–Civita connection of the metric h. Then the corresponding covariant derivatives $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual with respect to h, so that M has a dualistic structure $(h, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$. In particular, when $\alpha = \pm 1$, M is dually flat, and θ and η become $\alpha = +1$ and -1 affine coordinates respectively, which can be affirmed by a straightforward computation. In this way, a dualistic structure $(h, \nabla^{(+1)}, \nabla^{(-1)})$ on M is derived in a natural manner from the potential $U(\theta)$. The $(+1)$–divergence is defined as follows:

$$\begin{aligned} D^{(+1)}(p_1 \parallel p_2) &= U(\theta_2) + V(\eta_1) - \theta_2 \cdot \eta_1 \\ &= U(\theta_2) - U(\theta_1) - (\theta_2 - \theta_1) \cdot \partial_\theta U(\theta_1), \end{aligned}$$

where (θ_1, η_1) and (θ_2, η_2) are the dual affine coordinates of the points $p_1, p_2 \in M$, respectively. Note that the point whose η coordinates vanish corresponds to the minimum of the potential $U(\theta)$.

3 Dualistic Dynamical Systems on a Flat Manifold

In this section, dualistic structures of a gradient system on a dually flat manifold is studied.

Theorem 1 *Let M be a dually flat manifold with respect to the dualistic structure (g, ∇, ∇^*), in which a potential function $U(p)$ on M is defined by*

$$U(p) = D(q \parallel p),$$

where $D(q \parallel p)$ is the ∇–divergence and $q \in M$ is an arbitrarily prefixed point. Then the gradient flow [27, p. 205]

$$\dot{\theta}^i = -g^{ij}\partial_j U(\theta) \tag{3.1}$$

converges to the point q along the ∇^*–geodesic, where θ is the ∇–affine coordinates of point p, and $U(\theta) = U(p(\theta))$.

Proof Let us denote the mutually dual affine coordinates of the points p and q by (θ, η) and (θ', η'), respectively. Since ∇–divergence $D(q \parallel p)$ becomes

$$D(q \parallel p) = \psi(\theta) + \phi(\eta') - \theta \cdot \eta',$$

the gradient flow can be expressed in the form

$$\dot{\theta}^i = -g^{ij}\{\partial_j \psi(\theta) - \eta'_j\}.$$

By multiplying g_{ji} to both sides and using the identity

$$g_{ji}\dot{\theta}^i = \frac{\partial \eta_j}{\partial \theta^i}\frac{d\theta^i}{dt} = \frac{d\eta_j}{dt},$$

we have

$$\dot{\eta}_j = -\{\eta_j - \eta'_j\},$$

which can readily be integrated to obtain

$$\eta_j(t) = \eta'_j + \{\eta_j(0) - \eta'_j\}e^{-t}.$$

This proves the theorem. □

Here we give some examples of Theorem 1. Let us consider the family of Gaussian probability distributions with mean μ and variance σ^2:

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

This is a typical example of exponential family since it can be represented in the form

$$\log p(x;\mu,\sigma^2) = \theta^1 f_1(x) + \theta^2 f_2(x) - \psi(\theta)$$

where

$$\theta^1 = \frac{\mu}{\sigma^2}, \qquad \theta^2 = \frac{1}{2\sigma^2}$$

and

$$f_1(x) = x, \qquad f_2(x) = -x^2, \qquad \psi(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi}\sigma.$$

Since the cumulant generating function

$$\psi(\theta) = \log\int \exp\left\{\theta^1 f_1(x) + \theta^2 f_2(x)\right\}dx$$

is a convex function in $\theta = (\theta^1, \theta^2)$, the dually flat structure is naturally defined. The metric

$$g_{ij} = \partial_i\partial_j\psi(\theta)$$

is the Fisher information metric, $\nabla^{(1)}$ is the exponential or e–connection, and $\nabla^{(-1)}$ is the mixture or m–connection [10]. Throughout this example, g is the Fisher metric, θ_q^j are the e-affine coordinates of the point q, and $p_\theta(x) = p(x; \mu, \sigma^2)$.

We first let ∇ and ∇^* be exponential and mixture connections, respectively. Further let us set q as a δ–distribution concentrated on the origin. Then the potential becomes

$$U(\theta) = D^{(e)}(q \parallel p_\theta) = K(q, p_\theta) = \psi(\theta) + \text{const.}, \tag{3.2}$$

where K is the Kullback–Leibler information defined by

$$K(p_1, p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx.$$

The corresponding gradient flow coincides with Nakamura's dynamics [23], which converges to the δ–distribution q along an m–geodesic. Strictly speaking, some suitable renormalization is needed in (3.2) since the constant diverges in this case.

Conversely, let ∇ and ∇^* be mixture and exponential connections, respectively. Further let us set q as a uniform distribution, then θ_q^2 vanishes and θ_q^1 remains indefinite. In this case, ∇–affine parameters are the expectation parameters $\eta_i = E_\theta[f_i(x)]$ where $E_\theta[\ \cdot\ ]$ denotes expectation at p_θ, and the dynamics takes the form

$$\dot{\eta}_i = -g_{ij} \partial^j U(\eta). \tag{3.3}$$

Since the potential becomes (after some renormalization)

$$U(\eta) = D^{(m)}(q \parallel p_\theta) = K(p_\theta, q) = -[\text{ entropy of } p_\theta\ ] - \theta_q^i \eta_i + \text{const.},$$

the dynamics is a steepest ascent flow of entropy, which converges to the uniform distribution q along an e–geodesic. Moreover, if we rescale the time logarithmically such as

$$(t + \tau)\, \dot{\eta}_i = -g_{ij} \partial^j U(\eta), \qquad (t \geq 0,\ \tau > 0), \tag{3.4}$$

then the dynamics, which traces the same trajectory but in a different speed, can be integrated easily in the e–affine parameters and is given by

$$\theta^j(t) = \theta_q^j + \frac{\tau}{t + \tau}(\theta^j(0) - \theta_q^j),$$

where $\theta^j(0)$ is the e–affine coordinates of the initial point. This solution can be expressed also in the (μ, σ) space as

$$\begin{aligned} \mu(t) &= \frac{\theta^1(t)}{2\theta^2(t)} = \frac{\theta^1(0) - \theta_q^1}{2\theta^2(0)} + \frac{\theta_q^1}{2\tau\theta^2(0)}\,(t+\tau), \\ \sigma^2(t) &= \frac{1}{2\theta^2(t)} = \frac{1}{2\tau\theta^2(0)}\,(t+\tau). \end{aligned}$$

Here we used the relation $\theta_q^2 = 0$. If we set

$$\mu_0 = \frac{\theta^1(0) - \theta_q^1}{2\theta^2(0)}, \qquad v = \frac{\theta_q^1}{2\tau\theta^2(0)}, \qquad D = \frac{1}{4\tau\theta^2(0)},$$

then we have

$$\mu(t) = \mu_0 + v(t+\tau), \qquad \sigma^2(t) = 2D(t+\tau),$$

which shows that the dynamics (3.4) is nothing but a Uhlenbeck–Ornstein process [22].

We next consider an example in the linear programming problem of the form

$$\left\{ \begin{array}{l} \text{minimize} \quad \mathbf{c}\cdot\mathbf{x}, \\ \text{subject to} \quad A\mathbf{x} \le \mathbf{b}. \end{array} \right. \tag{3.5}$$

In order for the constraints $A\mathbf{x} < \mathbf{b}$ to be satisfied automatically, we introduce a convex potential function on $\mathbf{R}^n$ by

$$U(\mathbf{x}) = -\sum_i \log\,(b_i - A_{ij}x^j)$$

(see also Lagarias [28]). From this potential, we can derive a dualistic structure $(h, \nabla^{(+1)}, \nabla^{(-1)})$, in which $\mathbf{x} = (x^i)$ forms a $\nabla^{(+1)}$–affine coordinate system. The dual $\nabla^{(-1)}$–affine coordinate system $\mathbf{y} = (y_j)$ is

$$y_j = \sum_k \frac{A_{kj}}{b_k - A_{kl}x^l},$$

and the metric h is

$$h_{ij} = \sum_k \frac{A_{ki}A_{kj}}{(b_k - A_{kl}x^l)^2}.$$

Let us consider a gradient flow with respect to the potential $\psi(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x}$:

$$\dot{x}^i = -h^{ij}\partial_j\psi(\mathbf{x}) = -h^{ij}c_j.$$

Then it is shown that the flow converges to the optimum point for the problem (3.5) along (-1)–geodesic. Indeed,

$$\frac{d}{dt}\psi(\mathbf{x}) = c_i \dot{x}^i = -h^{ij} c_i c_j < 0.$$

The reason why the potential $\psi(\mathbf{x}) = \mathbf{c} \cdot \mathbf{x}$ is used here instead of the divergence is clarified in section 5.3.

4 Constrained Dynamics on a Submanifold

In this section, we investigate a dynamical system which is induced on a submanifold $N = \{p_\xi \ ; \ \xi \in \Xi \subset \mathbf{R}^n\}$ embedded in a flat manifold M with respect to a dualistic structure (g, ∇, ∇^*). Theorem 1 indicates that the gradient flow in M with respect to the potential $U(p) = D(q \parallel p)$ is a dynamical system whose gradient vector is the tangent vector along ∇^*–geodesic connecting the two points p and q (for short, ∇^*–tangent vector). Therefore we can construct a constrained dynamics on N by projecting the gradient ∇^*–tangent vector onto the tangent space $T_p(N)$ with respect to the metric g, see Fig.1. Such a dynamical system on N induced from a gradient flow on M is also a gradient flow on N of the form

$$\dot{\xi}^i = -\tilde{g}^{ij} \frac{\partial}{\partial \xi^j} D(q \parallel p_\xi), \tag{4.1}$$

where ξ is an arbitrary local coordinate system of N, and

$$\tilde{g}_{ij} = g(\frac{\partial}{\partial \xi^i} | \frac{\partial}{\partial \xi^j})$$

is an induced metric on N and $(\tilde{g}^{ij})$ is the inverse matrix of $(\tilde{g}_{ij})$. Let us examine geodesic characterization of such restricted dynamics. In general, the dualistic structure (g, ∇, ∇^*) on a manifold M naturally induces a dualistic structure $(\tilde{g}, \tilde{\nabla}, \tilde{\nabla}^*)$ on a submanifold N of M by projection with respect to the metric g onto the tangent space $T(N)$, i.e., for every vector fields A, B, C on N, $\tilde{\nabla}$ and $\tilde{\nabla}^*$ are defined by

$$\begin{aligned} g(\nabla_A B | C) &= g(\tilde{\nabla}_A B | C) = \tilde{g}(\tilde{\nabla}_A B | C), \\ g(\nabla^*_A B | C) &= g(\tilde{\nabla}^*_A B | C) = \tilde{g}(\tilde{\nabla}^*_A B | C). \end{aligned}$$

Duality of $\tilde{\nabla}$ and $\tilde{\nabla}^*$ follows directly from the definition (2.1). A submanifold N is called autoparallel with respect to the connection ∇ if for all $A, B \in T(N)$, $\nabla_A B \in T(N)$ holds.

Theorem 2 *If N is ∇–autoparallel, then the gradient flow (4.1) converges to a unique stationary point independent of the initial point along $\tilde{\nabla}^*$–geodesic.*

Proof Since N is ∇–autoparallel, Pythagorian theorem [10] assures that there exists a unique point $p^{(0)}$ on N satisfying

$$D(q \parallel p_\xi) = D(q \parallel p^{(0)}) + D(p^{(0)} \parallel p_\xi).$$

Then

$$\frac{\partial}{\partial \xi^j} D(q \parallel p_\xi) = \frac{\partial}{\partial \xi^j} D(p^{(0)} \parallel p_\xi).$$

Moreover, the following lemma indicates that

$$D(p^{(0)} \parallel p_\xi) = \tilde{D}(p^{(0)} \parallel p_\xi),$$

where $\tilde{D}$ denotes the $\tilde{\nabla}$–divergence on N. Therefore, the theorem can be proved in the same way as Theorem 1 by taking ξ as an $\tilde{\nabla}$–affine coordinate system without loss of generality. □

Lemma 1 *Given a flat manifold M with respect to (g, ∇, ∇^*). If a submanifold N embedded in M is ∇ or ∇^*–autoparallel, then N is also intrinsically flat with respect to $(\tilde{g}, \tilde{\nabla}, \tilde{\nabla}^*)$ and for every two points $p_1, p_2 \in N$,*

$$D(p_1 \parallel p_2) = \tilde{D}(p_1 \parallel p_2),$$

where D and $\tilde{D}$ are ∇–divergence on M and $\tilde{\nabla}$–divergence on N, respectively.

Proof First, we show the flatness of N. A straightforward calculation shows that the connection coefficients of $\tilde{\nabla}$ and $\tilde{\nabla}^*$ with respect to an arbitrary coordinate system of N become $\tilde{\Gamma}_{ijk} = \Gamma_{ijk}$ and $\tilde{\Gamma}^*_{ijk} = \Gamma^*_{ijk}$. Then if M is torsion free with respect to ∇ and ∇^*, N is automatically torsion free with respect to $\tilde{\nabla}$ and $\tilde{\nabla}^*$. To evaluate the Riemanian curvatures of N, we first let N be ∇–autoparallel. Then the following commutative diagram

$$\begin{tikzcd}
T_{p_1}(M) \arrow[r, "\nabla"] & T_{p_2}(M) \arrow[d, "id."] \\
T_{p_1}(N) \arrow[u, "id."] \arrow[r, "\tilde{\nabla}"] & T_{p_2}(N)
\end{tikzcd}$$

indicates that the parallel transports of tangent vectors of N with respect to $\tilde{\nabla}$ does not depend on the choice of the path connecting two points p_1 and p_2 since M is ∇–curvature free. Therefore, N is $\tilde{\nabla}$–curvature free. Moreover, in general, when one of the dual connections is curvature free, the other is automatically curvature free. Hence N is flat. In the same way, ∇^*–autoparallel case can be proved.

Next, we derive the divergence equality. Let the dual affine coordinate systems of M be (θ, η) and the corresponding dual potentials $(\psi(\theta), \phi(\eta))$. Suppose N is ∇–autoparallel, then N can be parameterized with a $\tilde{\nabla}$–affine parameter $\xi = [\xi^i]$ of N as

$$\theta^i(\xi) = a^i_j \xi^j + b^i, \qquad ({}^\exists a^i_j, {}^\exists b^i \in \mathbf{R}).$$

The accompanying dual $\tilde{\nabla}^*$–affine parameter $\zeta = [\zeta_i]$ satisfies

$$\delta^j_i = g(\frac{\partial}{\partial \xi^i} \mid \frac{\partial}{\partial \zeta_j}) = \frac{\partial \theta^k}{\partial \xi^i}\frac{\partial \eta_l}{\partial \zeta_j} g(\frac{\partial}{\partial \theta^k} \mid \frac{\partial}{\partial \eta_l}) = a^k_i \frac{\partial \eta_k}{\partial \zeta_j}.$$

Integration of this equation yields

$$a^k_i \eta_k(\zeta) = \zeta_i + c_i, \qquad ({}^\exists c_i \in \mathbf{R}).$$

Further, let us define two functions on N by

$$\begin{aligned} \tilde{\psi}(\xi) &= \psi(\theta(\xi)) - c_i \xi^i, \\ \tilde{\phi}(\zeta) &= \phi(\eta(\zeta)) - b^i \eta_i(\zeta). \end{aligned}$$

Then we have

$$\begin{aligned} \tilde{\psi}(\xi) + \tilde{\phi}(\zeta) - \xi \cdot \zeta &= \{\psi(\theta(\xi)) - c_i \xi^i\} + \{\phi(\eta(\zeta)) - b^i \eta_i(\zeta)\} - \xi^i \zeta_i \\ &= \psi(\theta(\xi)) + \phi(\eta(\zeta)) - (c_i + \zeta_i)\xi^i - b^i \eta_i(\zeta) \\ &= \psi(\theta(\xi)) + \phi(\eta(\zeta)) - (a^k_i \xi^i + b^k)\eta_k(\zeta) \\ &= \psi(\theta(\xi)) + \phi(\eta(\zeta)) - \theta(\xi) \cdot \eta(\zeta) \\ &= 0, \end{aligned}$$

and

$$\frac{\partial}{\partial \xi^i}\tilde{\psi}(\xi) = \frac{\partial \psi}{\partial \theta^k}\frac{\partial \theta^k}{\partial \xi^i} - c_i = \eta_k a^k_i - c_i = \zeta_i,$$

$$\frac{\partial}{\partial \zeta_i}\tilde{\phi}(\zeta) = \left(\frac{\partial \phi}{\partial \eta_k} - b^k\right)\frac{\partial \eta_k}{\partial \zeta_i} = (a^k_j \xi^j)\frac{\partial \eta_k}{\partial \zeta_i} = \xi^i.$$

These relations imply that the two functions $\tilde{\psi}(\xi)$ and $\tilde{\phi}(\zeta)$ form the dual potentials on N with respect to the dualistic structure $(\tilde{g}, \tilde{\nabla}, \tilde{\nabla}^*)$. Then

$$\begin{aligned}
\tilde{D}(p_1 \parallel p_2) &= \tilde{\psi}(\xi_2) + \tilde{\phi}(\zeta_1) - \xi_2 \cdot \zeta_1 \\
&= \{\psi(\theta(\xi_2)) - c_i \xi_2^i\} + \{\phi(\eta(\zeta_1)) - b^i \eta_i(\zeta_1)\} - \xi_2^i \zeta_{1i} \\
&= \psi(\theta(\xi_2)) + \phi(\eta(\zeta_1)) - (c_i + \zeta_{1i})\xi_2^i - b^i \eta_i(\zeta_1) \\
&= \psi(\theta(\xi_2)) + \phi(\eta(\zeta_1)) - (a_i^k \xi_2^i + b^k)\eta_k(\zeta_1) \\
&= \psi(\theta(\xi_2)) + \phi(\eta(\zeta_1)) - \theta(\xi_2) \cdot \eta(\zeta_1) \\
&= D(p_1 \parallel p_2).
\end{aligned}$$

On the other hand, when N is ∇^*–autoparallel, then the above proof leads to

$$D^*(p_1 \parallel p_2) = \tilde{D}^*(p_1 \parallel p_2),$$

which shows $D(p_2 \parallel p_1) = \tilde{D}(p_2 \parallel p_1)$ according to the duality. □

Note that Theorem 1 can be derived directly from Theorem 2 by setting $M = N$.

5 Examples

In this section, we give some examples which can be viewed as gradient systems on submanifolds embedded in manifolds having dualistic structures.

5.1 Maximum likelihood estimation

We first give a rather trivial example from statistical inference. Let $M = \{p_\theta\}_{\theta \in \Theta}$ be a parametric statistical model embedded in the set of all the positive probability distributions $\mathcal{P}$ on a finite set $\mathcal{X}$. The maximum likelihood estimation is a method to estimate the parameter θ from n observed data $(x_1, \cdots, x_n) \in \mathcal{X}^n$ such that

$$\max_\theta \prod_{i=1}^n p_\theta(x_i).$$

It is easy to derive another equivalent formulation which takes the form

$$\min_\theta \ K(q_n, p_\theta),$$

where K is the Kullback–Leibler information which is identical to the divergence with respect to the exponential connection, and q_n is the empirical distribution defined by

$$q_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x), \qquad \delta_{x_i}(x) = \begin{cases} 1, & \text{if } x = x_i, \\ 0, & \text{otherwise.} \end{cases}$$

Consider the dynamics of the form

$$\dot{\theta}^i = -g^{ij}\partial_j K(q_n, p_\theta), \tag{5.1}$$

where g^{ij} is the inverse of the Fisher metric on M.

Corollary 3 *The gradient flow (5.1) converges to a local maximum likelihood estimate on M. If M is an exponential family, then it converges to the unique maximum likelihood estimate independent of the initial point along an m-geodesic.*

5.2 Boltzmann Machines and EM algorithms

We next consider a neural network called Boltzmann machine with n visible units and m hidden units [29]. Let $\mathbf{x}_V = (x_1^V, \cdots, x_n^V) \in \{0,1\}^n$ and $\mathbf{x}_H = (x_1^H, \cdots, x_m^H) \in \{0,1\}^m$ be the states of visible and hidden units, respectively. The stationary distribution of the Boltzmann machine is written as

$$p(\mathbf{x}_V, \mathbf{x}_H; w) = \exp\{-E(\mathbf{x}_V, \mathbf{x}_H; w) - \psi(w)\},$$

where $\psi(w)$ is the normalization factor and

$$E(\mathbf{x}_V, \mathbf{x}_H; w) = \sum_{i>j} w_V^{ij} x_i^V x_j^V + \sum_{i>j} w_H^{ij} x_i^H x_j^H + \sum_{i,j} w_{VH}^{ij} x_i^V x_j^H$$

is the "energy" of the state $\mathbf{x} = (\mathbf{x}_V, \mathbf{x}_H)$ and $w = (w_V^{ij}, w_H^{ij}, w_{VH}^{ij})$. The coefficients $w_V^{ij}, w_H^{ij}, w_{VH}^{ij}$ are connection weights among the visible units, among the hidden units, and between the visible and the hidden units, respectively. The thresholds are included in the terms w_V^{i0} and w_H^{i0} by emvisaging virtual units $x_0^V = x_0^H \equiv 1$. Let us denote the set of all the probability distributions on $\mathbf{x} = (\mathbf{x}_V, \mathbf{x}_H)$ by $\mathcal{P}$. Then $\mathcal{P}$ is a manifold having a dualistic structure as a statistical manifold with $\dim \mathcal{P} = 2^{n+m} - 1$. The Boltzmann machine is trained by examples given from a stationary distribution $Q(\mathbf{x}_V)$ on the visible units. The purpose of this learning is to approximate $Q(\mathbf{x}_V)$

by $p(\mathbf{x}_V) = \sum_{\mathbf{x}_H} p(\mathbf{x}_V, \mathbf{x}_H; w)$ of the stationary distribution of the Boltzmann machine by modifying the parameter w. Amari *et al.* [21] proved that learning of Boltzmann machines is equivalent to the divergence minimization problem of the form

$$\min_{p \in B,\ q \in E_Q} K(q,p). \tag{5.2}$$

Here, K is the Kullback–Leibler information, B is the set of all the probability distributions in $\mathcal{P}$ realizable by the Boltzmann machine, i.e.,

$$B = \{p(\mathbf{x}_V, \mathbf{x}_H; w) \in \mathcal{P}\ ;\ w^{ij} \in \mathbf{R}\},$$

and E_Q is the set of all the probability distributions in $\mathcal{P}$ that have the same marginal distribution $Q(\mathbf{x}_V)$ on the visible units, i.e.,

$$E_Q = \{q(\mathbf{x}_V, \mathbf{x}_H) \in \mathcal{P}\ ;\ \sum_{\mathbf{x}_H} q(\mathbf{x}_V, \mathbf{x}_H) = Q(\mathbf{x}_V)\}.$$

It is easy to see that B is an exponential family with $\dim B = \frac{1}{2}(n+m)(n+m+1)$ and E_Q is a mixture family, since E_Q is defined by a linear constraint, with $\dim E_Q = 2^n(2^m - 1)$.

We propose simultaneous gradient flows to solve the minimization problem (5.2),

$$\left\{ \begin{array}{lcl} \dot{w}^{ij} & = & -g^{ij,kl} \dfrac{\partial}{\partial w^{kl}} K(q,p), \\ \\ \dot{q}_k & = & -g_{kl} \dfrac{\partial}{\partial q_l} K(q,p), \end{array} \right. \tag{5.3}$$

where w^{ij} are e–affine parameters of $p \in B$, q_k are m–affine parameters of $q \in E_Q$, $g^{ij,kl}$ is the inverse of the Fisher metric on B, and g_{kl} is the Fisher metric on E_Q. Since $K(q,p) = D^{(e)}(q \parallel p) = D^{(m)}(p \parallel q)$ where $D^{(e)}$ and $D^{(m)}$ are divergences with respect to the exponential and the mixture connections, the dynamical system (5.3) has a geometrical interpretation in the sense of section 4: The dynamics for p is a restricted gradient flow on B under the influence of the "force" from q (more precisely, m–tangent vector), while the dynamics for q is a restricted gradient flow on E_Q under the "force" from p (e–tangent vector), each flow being coupled through the equation (5.3) to form a two body dymanics, see Fig. 2.

If there are no hidden units, then $\dim E_Q = 0$ and the gradient system (5.3) is reduced to the same situation as in the maximum likelihood estimation. In the general case, the dynamics converges to a local minimum.

Observing the fact that the dynamics (5.3) can be solved simultaneously, it is expected that its convergence is much faster than the conventional nested iterative EM algorithm or its modifications [30].

5.3 Linear Programming

Let us consider a general linear programming problem in a canonical form

$$\left\{\begin{array}{ll} \text{minimize} & \mathbf{c}\cdot\mathbf{x}, \\ \text{subject to} & A\mathbf{x}=\mathbf{b},\ \mathbf{x}\geq\mathbf{0}. \end{array}\right. \tag{5.4}$$

The problem considered in section 3 is of course included. In this formulation, the domain of the problem can be regarded as a submanifold M of the enveloping space $\mathbf{R}^n_+$, where

$$\mathbf{R}^n_+ = \{\mathbf{x}\in\mathbf{R}^n\ ;\ \mathbf{x}\geq\mathbf{0}\}, \qquad M=\{\mathbf{x}\in\mathbf{R}^n_+\ ;\ A\mathbf{x}=\mathbf{b}\}.$$

Let us construct a dualistic structure on $\mathbf{R}^n_+$ in which $\mathbf{x}$ becomes a $(+1)$–affine coordinate system. In order for the nonnegativity constraints $\mathbf{x}\geq\mathbf{0}$ to be satisfied automatically, we introduce a convex potential function on $\mathbf{R}^n_+$ by

$$U(\mathbf{x}) = -\sum_{i=1}^{n}\log x^i.$$

By using the general scheme presented in section 2, we can derive a natural dualistic structure $(h, \nabla^{(+1)}, \nabla^{(-1)})$. For instance, the metric is

$$h_{ij}(\mathbf{x}) = \partial_i\partial_j U(\mathbf{x}) = \frac{\delta_{ij}}{(x^i)^2},$$

the (-1)–affine coordinate system

$$y_j = \partial_j U(\mathbf{x}) = -\frac{1}{x^j},$$

and $(+1)$–divergence

$$D^{(+1)}(p' \parallel p) = \sum_{i=1}^{n}\left[\frac{x^i}{x'^i} - \log\frac{x^i}{x'^i}\right] - n.$$

Setting aside the constraints $A\mathbf{x}=\mathbf{b}$, we have the trivial explicit solution of the problem (5.4) as

$$x^i_0 = \left\{\begin{array}{ll} 0, & (c_i\geq 0), \\ +\infty, & (c_i<0). \end{array}\right. \tag{5.5}$$

Let p_0 be this optimal point. We then have a (formal) gradient system

$$\dot{x}^i = -h^{ij}\partial_j D^{(+1)}(p_0 \parallel p), \tag{5.6}$$

which converges to the p_0 along a $\nabla^{(-1)}$–geodesic. Returning to the original constrained problem, the submanifold M determined by the affine constraints $A\mathbf{x} = \mathbf{b}$ is $\nabla^{(+1)}$–autoparallel since $\mathbf{x}$ is $(+1)$–affine coordinate system. So we have

Corollary 4 *Let an arbitrary coordinate system on M be $\xi = [\xi^i]$. Then the constrained dynamics*

$$\dot{\xi}^i = -\tilde{h}^{ij}\frac{\partial}{\partial \xi^j} D^{(+1)}(p_0 \parallel p_\xi)$$

on M converges to the optimum solution of the linear programming problem (5.4) along a $\tilde{\nabla}^{(-1)}$–geodesic.

However, since (5.5) is singular, this result seems quite formal. A realistic modification of the dynamics (5.6) may be

$$\dot{x}^i = -h^{ij}\partial_j\{\mathbf{c}\cdot\mathbf{x}\} = -h^{ij}c_j,$$

whose gradient vectors are parallel to those of (5.6) but always finite. This is a steepest descent flow of the potential $\psi(\mathbf{x}) = \mathbf{c}\cdot\mathbf{x}$ with respect to the metric h in $\mathbf{R}^n_+$ and is also a $\nabla^{(-1)}$–geodesic motion. Therefore, the corresponding gradient flow constrained on the submanifold M

$$\dot{\xi}^i = -\tilde{h}^{ij}\frac{\partial}{\partial \xi^j}\{\mathbf{c}\cdot\mathbf{x}_\xi\}$$

has the same properties as in Corollary 2. This is the continuous version of the affine scaling method [31] famous in the linear programming problem.

5.4 Replicator Equation in Mathematical Biology

It is well-known that the dynamical system on a simplex $S_n = \{\mathbf{x} \in \mathbf{R}^{n+1}_+ \ ; \ \sum_{i=0}^n x^i = 1\}$ of the form

$$\dot{x}^i = x^i(a_{ij}x^j - a_{rs}x^r x^s) \tag{5.7}$$

is often encountered in mathematical biology, which is called the replicator equation [32]. In social biology, this is a game dynamical equation that describes the evolution of competing phenotypes. In the symmetric case

$a_{ij} = a_{ji}$, it describes a continuous version of the selection equation in population genetics. Further, there exists a close connection between (5.7) and the Lotka–Volterra equation, i.e., there exists a natural diffeomorphism on S_n to $\mathbf{R}_+^n$ which maps the trajectory of (5.7) to the solution of the Lotka–Volterra equation [32]. Let us examine (5.7) in view of information geometry. We first define a potential function on $\mathbf{R}_+^{n+1}$ by

$$U(\mathbf{x}) = \sum_{i=1}^{n} x^i \log x^i.$$

From this potential, we derive a natural dualistic structure $(h, \nabla^{(+1)}, \nabla^{(-1)})$ in that $\mathbf{x}$ is $(+1)$–affine coordinate system. In particular, the metric

$$h_{ij} = \frac{\delta_{ij}}{x^i}$$

is called Shahshahani metric [33]. It can be shown that if $a_{ij} = a_{ji}$, then (5.7) is a steepest ascent flow of the potential $\psi(\mathbf{x}) = \frac{1}{2} a_{ij} x^i x^j$ with respect to the Shahshahani metric restricted on the simplex S_n [32]. In this case, however, the dynamics does not trace a $\tilde{\nabla}^{(-1)}$–geodesic in general since $\partial_j \psi(\mathbf{x}) = a_{jk} x^k$ is not a $\nabla^{(-1)}$–tangent vector, i.e., it is not parallel to $\partial_j D^{(+1)}(p_0 \parallel p) = \log(x^j / x_0^j)$.

6 Other Related Topics

It is possible to characterize the dualistic gradient flow as a completely integrable Hamiltonian system, which gives a generalization of Nakamura's work [23].

Theorem 5 *If N is ∇–autoparallel and $\dim N$ is even, say $2k$, then the dynamical system (4.1) can be regarded as a compeletely integrable Hamiltonian system with generalized positions $Q_i = \partial_{2i} D(q \parallel p_\xi)$, generalized momentums $P^i = -1/\partial_{2i-1} D(q \parallel p_\xi)$, and Hamiltonian $\mathcal{H} = -Q_i P^i$, provided ξ is taken, without loss of generality, a $\tilde{\nabla}$–affine coordinate system. The k quantities $\mathcal{H}_i = \partial_{2i} D(q \parallel p_\xi) / \partial_{2i-1} D(q \parallel p_\xi)$ are mutually independent constants of motion.*

Proof As was mentioned in the proof of Theorem 3, the dynamical system (4.1) can be rewritten in the $\tilde{\nabla}^*$–coordinate system ζ as

$$\dot{\zeta}_j = -(\zeta_j - \zeta_j^{(0)}), \tag{6.1}$$

where $\zeta_j^{(0)}$ is the ζ–coordinate of $p^{(0)}$. Further, in this dual coordinate system,

$$Q_i = \zeta_{2i} - \zeta_{2i}^{(0)}, \quad P^i = -1/(\zeta_{2i-1} - \zeta_{2i-1}^{(0)}), \quad \mathcal{H}_i = \frac{\zeta_{2i} - \zeta_{2i}^{(0)}}{\zeta_{2i-1} - \zeta_{2i-1}^{(0)}}.$$

Then

$$\dot{\mathcal{H}}_i = \frac{1}{(\zeta_{2i-1} - \zeta_{2i-1}^{(0)})^2} \left\{ (\zeta_{2i} - \zeta_{2i}^{(0)})\dot{\zeta}_{2i-1} - \dot{\zeta}_{2i}(\zeta_{2i-1} - \zeta_{2i-1}^{(0)}) \right\} = 0.$$

Involutivity is shown as follows.

$$\begin{aligned} \{\mathcal{H}_i, \mathcal{H}_j\} &= \sum_{l=1}^{k} \left\{ \frac{\partial \mathcal{H}_i}{\partial P^l}\frac{\partial \mathcal{H}_j}{\partial Q_l} - \frac{\partial \mathcal{H}_i}{\partial Q_l}\frac{\partial \mathcal{H}_j}{\partial P^l} \right\} \\ &= \frac{\partial \mathcal{H}_i}{\partial P^i}\frac{\partial \mathcal{H}_j}{\partial Q_i} - \frac{\partial \mathcal{H}_i}{\partial Q_i}\frac{\partial \mathcal{H}_j}{\partial P^i} = 0. \end{aligned}$$

Independency of $\mathcal{H}_i$ is trivial. By straightforward computation, Hamilton's equation

$$\frac{dQ_i}{dt} = \frac{\partial \mathcal{H}}{\partial P^i}, \qquad \frac{dP^i}{dt} = -\frac{\partial \mathcal{H}}{\partial Q_i},$$

are reduced to

$$\dot{\zeta}_{2i} = -(\zeta_{2i} - \zeta_{2i}^{(0)}), \qquad \dot{\zeta}_{2i-1} = -(\zeta_{2i-1} - \zeta_{2i-1}^{(0)}),$$

which reproduce the original dynamical equation (6.1). □

The condition for $\dim N$ to be even is not essential. Indeed, if N is ∇–autoparallel and $\dim N$ is odd, then the dynamical system (4.1) can be regarded as a subdynamics of a higher dimensional completely integrable Hamiltonian system by combining it with an independent odd dimensional gradient system.

In a naive sense, a $2k$ dimensional Hamiltonian system is equivalent to a k dimensional Lagrangian system. From this analogy, we can imagine a 2nd order dynamics of the form

$$\ddot{\theta}^l + \left\{ \begin{matrix} l \\ ij \end{matrix} \right\} \dot{\theta}^i \dot{\theta}^j = -g^{lj}\partial_j U(\theta).$$

This is the equation of motion of a particle constrained on a submanifold N associated with a potential $U(\theta)$. It is well known that this dynamics can be derived by the variational principle with Lagrangian

$$\mathcal{L} = \frac{1}{2} g_{ij} \dot{\theta}^i \dot{\theta}^j - U(\theta).$$

In the same way, if we consider a dynamical system of the form

$$\ddot{\theta}^l + \Gamma^{(-1)l}_{ij} \dot{\theta}^i \dot{\theta}^j = -h^{lj} \partial_j U(\theta),$$

where θ is a $(+1)$–affine coordinate system of N, then we have

$$\ddot{\eta}_i = -\eta_i$$

in the dual affine coordinate system, which indicates that the system is composed of k independent harmonic oscillators and can be regarded as a completely integrable Hamiltonian system. In this case, however, it is not yet clear whether the system can be derived by a certain variational principle.

7 Conclusions

We first constructed the gradient flow on a flat manifold M with respect to a dualistic structure (g, ∇, ∇^*) which converges to an arbitrarily prefixed point along the ∇–geodesic. We next derived a constrained dynamics on a submanifold N embedded in a flat manifold M, and investigated its geodesic characterization. Some examples followed, which are collected from various regions of optimization problems such as statistical inference, neural networks, EM algorithms, linear programming problems, and mathematical biology. Finally, we showed that the dualistic gradient flow can be characterized as a completely integrable Hamiltonian system.

Acknowledgment

The authors are grateful to Professor Y. Nakamura and Professor H. Nagaoka for fruitful discussions and comments.

References

[1] N. Karmarkar, "A new polynomial-time algorithm for linear programming," Combinatorica **4** (1984) 373.

[2] R. W. Brockett, "Dynamical systems that sort lists, diagonalize matrices and solve linear prigramming problems," Proc. 27th IEEE Conf. on Decision and Control, IEEE, (1988) 799.

[3] Y. Nakamura, "A new nonlinear dynamical system that leads to eigenvalues," Jap. J. Industrial and Appl. Math **9** (1992) 133.

[4] R. W. Brockett, "Least squares matching problems," Linear Algebra and its Appl. **122/123/124** (1989) 761.

[5] D. A. Bayer and J. C. Lagarias, "The nonlinear geometry of linear programming, I and II," Trans. American Math. Soc. **314** (1989) 499, 527.

[6] R. W. Brockett, "Differential geometry and the design of gradient algorithms," in R. E. Greens and S. T. Yau, eds., "Differential geometry," Amer. Math. Soc. 1993.

[7] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed. (Benjamin, New York, 1985).

[8] R. Balian, Y. Alhassid, and H. Reinhardt, "Dissipation in many–body systems : A geometric approach based on information theory," Physics Reports **131** (1986) 1.

[9] S. Amari, "Differential geometry of curved exponential families — Curvatures and information loss," Ann. Statist. **10** (1982) 357.

[10] S. Amari, *Differential-Geometrical Methods in Statistics*, Lec. Notes in Statist. **28** (Springer, Berlin, 1985).

[11] H. Nagaoka and S. Amari, "Differential geometry of smooth families of probability distributions," METR **82-7**, Univ. Tokyo (1982).

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

[13] N. N. Chentsov, *Optimal Decision Rules and Optimal Inference* (in Russian), (Nauka, Moskow, 1972) : Translated into English (AMS, Rhode Island, 1982).

[14] M. Kumon and S. Amari, "Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference," in Proc. R. Soc. London **A387** (1983) 429.

[15] S. Amari and M. Kumon, "Estimation in the presence of infinitely many nuisance parameters — Geometry of estimating functions," Ann. Statist. **16** (1988) 1044.

[16] I. Okamoto, S. Amari, and K. Takeuchi, "Asymptotic theory of sequential estimation : Differential geometrical approach," Ann. Stst. **19** (1991) 961.

[17] S. Amari, "Differential geometry of a parametric family of invertible linear systems — Riemannian metric, dual affine connections and divergence," Math. Systems Theory **20** (1987) 53.

[18] A. Ohara and S. Amari, "Differential geometric structures of stable state feedback systems with dual connections," Kybernetika (to appear).

[19] S. Amari and T. S. Han, "Statistical inference under multi-terminal rate restrictions — A differential geometrical approach," IEEE Trans. Information Theory **IT-35** (1989) 217.

[20] S. Amari, "Fisher information under restriction of Shannon information in multiterminal situations," Ann. Inst. Statist. Math. **41** (1989) 623.

[21] S. Amari, K. Kurata, and H. Nagaoka, "Information geometry of Boltzmann machines," IEEE Trans. Neural Networks **3** (1992) 260.

[22] T. Obata, H. Hara, and K. Endo "Differential geometry of nonequilibrium processes," Phys. Rev. **A 45** (1992) 6997.

[23] Y. Nakamura, "Completely integrable gradient systems on the manifolds of Gaussian and multinomial distributions," Jap. J. Industrial and Appl. Math. **10** (1993) 179.

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

[25] K. Nomizu and U. Simon, "Notes on conjugate connections," preprint.

[26] T. Kurose, "Dual connections and affine geometry," Math. Z. **203** (1990) 115.

[27] M. H. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Pure and Appl. Math., Vol. 5 (Academic, New York, 1974).

[28] J. C. Lagarias, "The nonlinear geometry of linear programming. III Projective Legendre transform coordinates and Hilbert geometry," Trans. American Math. Soc. **320** (1990) 193.

[29] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," Cognitive Science **9** (1985) 147.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via EM algorithm," J. Royal Stat. Soc. **B39** (1977) 1.

[31] E. R. Barns, "A variation on Karmarkar's algorithm for solving linear programing problems," Math. Programming **36** (1986) 174.

[32] J. Hofbauer and K. Sigmund, *The theory of Evolution and Dynamical Systems* (Cambridge Univ. Press, Cambridge, 1988).

[33] E. Akin, *The Geometry of Population Genetics*, Lec. Notes in Biomath. **31** (Springer, Berlin, 1979).