数理解析研究所講究録
第 822 巻 1993 年 32-42

Dynamical Systems on Statistical Models

Akio Fujiwara (藤原 彰夫)
Department of Mathematical Engineering and Information Physics
University of Tokyo, Tokyo 113, Japan

Abstract

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

1 Introduction

Motivated mainly by classical mechanics, completely integrable dynamical systems have been investigated by many researchers. Furthermore, some authors have sought contacts with other fields such as linear programming [4] and eigenvalue problems of matrices [5], see also [6] and the references cited therein.

On the other hand, some physicists have studied nonequilibrium or dissipative processes from a geometrical viewpoint [3]. Obata et al. also examined some nonequilibrium processes by using information geometry [9]. They showed the Uhlembeck–Ornstein process is a geodesic motion with respect to the exponential connection on a Gaussian model.

Quite recently, Nakamura pointed out that certain gradient flows on Gaussian and multinomial distributions can be characterized as completely integrable Hamiltonian systems [8]. This is the first suggestion of the connection between two seemingly unrelated fields, i.e., information geometry and completely 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. Statistical significance of the gradient flow is also investigated.

2 Dualistic geometry

We first give a brief summary of dualistic geometry. For details, consult [2]. Let M be a Riemannian manifold with metric g. Two affine connections ∇ and ∇^* on M are said to be *dual* with respect to g if for any vector field A, B, and C on M,

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

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 *flat*, and a pair of divergences on M are defined in the following way. We first construct mutually dual affine coordinates on M, i.e., ∇–affine coordinate $\theta = [\theta^i]$ and ∇^*–affine coordinate $\eta = [\eta_i]$ which satisfy

$$g(\partial_i \mid \partial^j) = \delta_i^j, \tag{1}$$

where $\partial_i = \partial/\partial\theta^i$ and $\partial^j = \partial/\partial\eta_j$. Then there exist such potential functions $\psi(\theta)$, $\phi(\eta)$ on M satisfying

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

where $\theta \cdot \eta = \theta^i\eta_i$. By using these potentials, we define the ∇–divergence D 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

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

For instance, let M be a set of positive probability distributions on a set $\mathcal{X}$, g the Fisher metric, ∇ and ∇^* the exponential and mixture connections, respectively. Then the exponential divergence D is given by

$$D(p_1 \parallel p_2) = \int_{\mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)} dx,$$

which is identical to the Kullback–Leibler divergence $K(p_1, p_2)$. Note that our manner of naming of divergences is different from Amari's one.

Next, we tackle the converse problem, i.e., let us construct a natural dualistic structure for an arbitrary manifold M on which a potential $U(\theta)$ is given, where $\theta = [\theta^i]$ is a local coordinate system of M. 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) satisfy the identity

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

The metric $\hat{g}$ on M is defined by

$$\hat{g}_{ij} = \partial_i \partial_j U(\theta).$$

This definition can be rewritten as

$$\hat{g}_{ij} = \frac{\partial \eta_j}{\partial \theta^i},$$

which readily leads to the relation

$$\hat{g}^{ij} = \frac{\partial \theta^j}{\partial \eta_i} = \partial^i \partial^j V(\eta).$$

This indicates that the coordinate systems θ and η are mutually dual with respect to $\hat{g}$ in the sense of Eq. (1). Further let us set

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

and define the α–connection by

$$\Gamma^{(\alpha)}_{ijk} = [ij;k] - \frac{\alpha}{2} T_{ijk},$$

with $[ij;k]$ the Levi–Civita connection, then 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 $(\hat{g}, \nabla^{(+1)}, \nabla^{(-1)})$ on M is derived in a natural manner from the potential $U(\theta)$. The $(+1)$–divergence is defined as follows:

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

where (θ_1, η_1) and (θ_2, η_2) are the dual affine coordinates of 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

In this section, we examine dualistic structures of a gradient system on a flat manifold.

Theorem 3.1 *Let M be a flat manifold with respect to the dualistic structure (g, ∇, ∇^*), $U(p)$ a potential function on M with respect to an arbitrarily prefixed point $q \in M$ defined by*

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

where $D(q \parallel p)$ is the ∇-divergence. Then the gradient flow [7, p. 205]

$$\dot{\theta}^i = -g^{ij} \partial_j U(\theta) \tag{2}$$

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

Proof Since ∇–divergence $D(q \parallel p)$ is rewritten as

$$\begin{aligned} D(q \parallel p) &= \psi(\theta(p)) + \phi(\eta(q)) - \theta(p) \cdot \eta(q) \\ &= \psi(\theta(p)) + \{-\psi(\theta(q)) + \theta(q) \cdot \eta(q)\} - \theta(p) \cdot \eta(q) \\ &= \psi(\theta(p)) - \psi(\theta(q)) + \{\theta(q) - \theta(p)\} \cdot \eta(q), \end{aligned}$$

the gradient flow can be expressed in the form

$$\dot{\theta}^i(p) = -g^{ij}\{\partial_j \psi(\theta(p)) - \eta_j(q)\}.$$

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

$$\dot{\eta}_j(p) = -\{\eta_j(p) - \eta_j(q)\},$$

which can readily be integrated to obtain

$$\eta_j(p(t)) = \eta_j(q) + \{\eta_j(p(0)) - \eta_j(q)\}e^{-t}.$$

This proves the proposition. ■

Example 3.1 *Here we give two examples of Theorem 3.1. Let us consider a Gaussian family with mean μ and variance σ^2:*

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

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

$$\log p_\theta(x) = \theta^1 f_1(x) + \theta^2 f_2(x) - \psi(\theta)$$

where

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

are e–affine parameters and

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

Throughout this example, g is the Fisher metric.

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

$$U(\theta) = D^{(e)}(q \,\|\, p_\theta) = K(q, p_\theta) = \psi(\theta),$$

and the corresponding gradient flow coincides with Nakamura's dynamics [8]*, which converges to the δ–distribution q along an m–geodesic.*

Conversely, let ∇ and ∇^* be mixture and exponential connections, respectively. Further let us set q as a uniform distribution on $\mathcal{X}$, then $\theta^2(q)$ vanishes and $\theta^1(q)$ 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

$$\dot{\eta}_i = -g_{ij}\partial^j U(\eta). \tag{3}$$

Since the potential becomes

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

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

$$t\,\dot{\eta}_i = -g_{ij}\partial^j U(\eta), \tag{4}$$

then the dynamics can be integrated easily and expressed in the e–affine parameters as

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

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

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

Here we used the relation $\theta^2(q) = 0$. If we set

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

then we have

$$\mu(t) = \mu_0 + vt, \qquad \sigma^2(t) = 2Dt,$$

which shows that the dynamics (4) is nothing but a Uhlembeck–Ornstein process [9].

Next, we consider another situation. Given a manifold M and a locally convex potential $U(\theta)$, then we can induce a natural dualistic structure $(\hat{g}, \nabla^{(+1)}, \nabla^{(-1)})$ on M by the procedure mentioned in the previous section. Let us examine a gradient flow on M of the form

$$\dot{\theta}^i = -\hat{g}^{ij}\partial_j U(\theta), \tag{5}$$

which can be reexpressed in the dual affine coordinates as

$$\dot{\eta}_i = -\eta_i. \tag{6}$$

In case $\dim M$ is even, this dynamical system can be characterized as a completely integrable Hamiltonian system [1, p. 392], which is a generalization of Nakamura's results [8], as follows.

Theorem 3.2 *If* $\dim M$ *is even, say* $2m$*, then the dynamical system* (6) *is a completely integrable Hamiltonian system with position* $Q_k = \eta_{2k}$*, momentum* $P^k = -1/\eta_{2k-1}$*, and Hamiltonian* $\mathcal{H} = -Q_k P^k$*,* $(k = 1, \cdots, m)$*. The* m *quantities* $\mathcal{H}_k = \eta_{2k}/\eta_{2k-1}$ *are mutually independent constants of motion.*

Proof By using (6), we have

$$\dot{\mathcal{H}}_k = \frac{1}{\eta_{2k-1}^2}(\dot{\eta}_{2k}\eta_{2k-1} - \eta_{2k}\dot{\eta}_{2k-1}) = 0.$$

Independency and involutiveness of $\{\mathcal{H}_k\}_{k=1}^m$ are trivial. By straightforward computation, Hamilton's equations

$$\frac{dQ}{dt} = \frac{\partial \mathcal{H}}{\partial P^k}, \qquad \frac{dP}{dt} = -\frac{\partial \mathcal{H}}{\partial Q_k}$$

are reduced to

$$\dot{\eta}_{2k} = -\eta_{2k}, \qquad \dot{\eta}_{2k-1} = -\eta_{2k-1},$$

which reproduce the original gradient flow (6). ■

Note that if $\dim M$ is odd, then the dynamical system (6) can be regarded as a subdynamics of a higher dimensional completely integrable Hamiltonian system by combining it with an independent odd dimensional gradient system.

4 Constrained Dynamics on a Parametric Model

In this section, we examine a dynamical system which is induced on a parametric statistical model. Let $M = \{p_\theta\}_{\theta \in \Theta}$ be a parametric model embedded in the set of probability distributions $\mathcal{P}$ on $\mathcal{X}$. Theorem 3.1 indicates that the gradient flow in $\mathcal{P}$ with respect to the potential $U(p) = K(q_n, p)$ with q_n the empirical distribution is a dynamical system whose gradient vector is m–tangent vector from the point p toward the empirical distribution q_n which in general falls out of the model M. Therefore we can construct a constrained dynamics on the model M by projecting the gradient m–tangent vector onto the tangent space $T_p(M)$ of the model with respect to the Fisher metric.

Theorem 4.1 *Such an induced dynamical system is also a gradient flow on M of the form*

$$\dot{\theta}^i = -g^{ij}\partial_j K(q_n, p_\theta), \tag{7}$$

where g is the Fisher metric on M. This flow converges to a locally maximum likelihood estimate.

Proof Let us define a bilinear form $\langle \cdot, \cdot \rangle$ on $T_p(M)$ by

$$\langle f(x), g(x) \rangle = \int_{\mathcal{X}} f(x) g(x) dx,$$

where $f(x)$ and $g(x)$ are an m–tangent vector and an e–tangent vector, respectively. Note that this value is identical to the conventional Fisher inner product in information geometry. Then the projection of the m–tangent vector from the point p toward the empirical distribution q_n onto the tangent space $T_p(M)$, expressed as $a^i \partial_i p_\theta(x)$, satisfies

$$\langle q_n(x) - p_\theta(x),\ \partial_j \log p_\theta(x) \rangle = \langle a^i \partial_i p_\theta(x),\ \partial_j \log p_\theta(x) \rangle.$$

This leads to

$$\begin{aligned} a^i &= g^{ij} \langle q_n(x) - p_\theta(x),\ \partial_j \log p_\theta(x) \rangle \\ &= g^{ij} \langle q_n(x),\ \partial_j \log p_\theta(x) \rangle \\ &= -g^{ij} \partial_j K(q_n, p_\theta). \end{aligned}$$

Hence the induced dynamical system becomes

$$\dot{p}_\theta(x) = \dot{\theta}^i \partial_i p_\theta(x) = a^i \partial_i p_\theta(x),$$

or

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

Every equilibrium point of this flow satisfies $a^i = 0$ for all i, which is nothing but foots of the m–geodesic perpendiculars from q_n onto the model M. ∎

Lemma 4.1 *Suppose the potential $U(\theta)$ on M is given by the Kullback-Leibler divergence, i.e., $U(\theta) = K(q_n, p_\theta)$. The induced metric $\hat{g}_{ij}(\theta)$ is identical to the Fisher metric $g_{ij}(\theta)$ for every $q_n \in \mathcal{P}$ iff the model M is an exponential family.*

Proof If $\hat{g}_{ij}(\theta)$ is identical to $g_{ij}(\theta)$ for every $q_n \in \mathcal{P}$, then

$$g_{ij}(\theta) - \hat{g}_{ij}(\theta) = -\int_{\mathcal{X}} \{p_\theta(x) - q_n(x)\}\partial_i\partial_j \log p_\theta(x) dx = 0.$$

This shows that $\partial_i\partial_j \log p_\theta(x)$ does not depend on x, i.e., there exists a function $\psi(\theta)$ such that

$$\partial_i\partial_j \log p_\theta(x) = -\partial_i\partial_j\psi(\theta)$$

holds. This equation can readily be integrated to yield

$$\log p_\theta(x) = c(x) + \theta^i f_i(x) - \psi(\theta),$$

which shows that the model $\{p_\theta(x)\}$ is an exponential family. The converse statement is evident from the calculations above. ∎

Theorem 4.2 *If model M is an exponential family, then the induced gradient flow* (7) *converges to the unique maximum likelihood estimate with respect to the empirical distribution q_n along m–geodesic. Moreover, if* dim$M$ *is even, say $2m$, then the flow is a completely integrable Hamiltonian system with position $Q_k = \partial_{2k}K(q_n, p_\theta)$, momentum $P^k = -1/\partial_{2k-1}K(q_n, p_\theta)$, and Hamiltonian $\mathcal{H} = -Q_k P^k$, $k = 1, \cdots, m$.*

Proof Straightforward from Theorems 3.1, 3.2, 4.1, and Lemma 4.1. ∎

5 Concluding Remarks

We have constructed a gradient flows on a flat manifold M with respect to a dualistic structure (g, ∇, ∇^*) which converges to an arbitrarily prefixed point along ∇–geodesic. If $\dim M$ is even, this flow can be also characterized as a completely integrable Hamiltonian flow.

We have also derived a constrained dynamics on a submanifold M embedded in a statistical manifold $\mathcal{P}$, which converges to the locally maximum likelihood estimate. If M is an exponential family, then the flow evolves along m–geodesics. In case $\dim M$ is even, the flow can also be considered as a completely integrable Hamiltonian system. However, ststistical meaning of such characterization as a Hamiltonian system is not clear.

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

$$\ddot{\theta}^k + \left\{ {k \atop ij} \right\} \dot{\theta}^i \dot{\theta}^j = -g^{kj} \partial_j U(\theta),$$

which is the equation of motion of a particle constrained on a manifold M associated with a potential $U(\theta)$. It is well known that this dynamics can be derived by the variational principle with Lagrangean

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

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

$$\ddot{\theta}^k + \Gamma_{ij}^{(-1)k} \dot{\theta}^i \dot{\theta}^j = -\hat{g}^{kj} \partial_j U(\theta),$$

then we have

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

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

Acknowledgments

Gratitude is expressed to Prof. S. Amari for his critical reading of the manuscript and valuable comments. Thanks are due to Prof. Y. Nakamura for fruitful discussions.

References

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

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

[3] R. Balian, Y. Alhassid, and H. Reinhardt, " Dissipation in Many–Body Systems : A Geometric Approach Based on Information Theory," Physics Reports, **131**, pp. 1–146 (1986).

[4] D. A. Bayer and J. C. Lagarias, "The Nonlinear Geometry of Linear Programming, I and II," Trans. American Math. Soc., **314**, pp. 499–526, pp. 527–581 (1989).

[5] R. W. Brockett, "Dynamical Systems That Sort Lists, Diagonalize Matrices and Solve Linear Problems," Proc. 27th IEEE Conf. on Decision and Control, IEEE, pp. 799–803 (1988).

[6] R. W. Brockett, "Differential Geometry and the Design of Gradient Algorithms," preprint.

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

[8] Y. Nakamura, "Completely Integrable Gradient Systems on the Manifolds of Gaussian and Multinomial Distributions," Japan J. Industrial and Applied Math. (to appear).

[9] T. Obata, H. Hara, and K. Endo "Differential Geometry of Nonequilibrium Processes," Phys. Rev. **A 45**, pp. 6997–7001 (1992).