Information geometry of divergences and means on the space of all probability measures having positive density function

Hiroyasu Satoh (Nippon Institute of Technology, hiroyasu@nit.ac.jp)

Abstract The space of all probability measures having positive density function on a measure space (M, λ) carries a Riemannian metric G, called the Fisher metric. By using divergences which are distance-like functions, we can define a family of torsion-free affine connections $\{\nabla^{(\alpha)}\}_{\alpha\in\mathbb{R}}$ which satisfies that $\nabla^{(-\alpha)}$ is the dual connection of $\nabla^{(\alpha)}$ with respect to G and $\nabla^{(0)}$ is the Levi-Civita connection of G. We define the normalized power mean of two probability measures and give characterizations of geodesic segments of $\nabla^{(\alpha)}$, $\alpha = -1, 0, 1$ in terms of means of its endpoints. Moreover, we show that integrations of kinetic energy of (± 1)-geodesic segments are equal to the symmetrized Kullback-Leibler divergence of its endpoints. This is based on joint work [5] with Mitsuhiro Itoh.

1 Fisher metric

Let (M, λ) be a measure space with fixed probability measure λ and $\mathcal{P}^+(M)$ be the space of all probability measures on M;

$$\mathcal{P}^+(M) := \left\{ \mu \;\middle|\; \int_M d\mu = 1,\ \mu \ll \lambda,\ \frac{d\mu}{d\lambda} > 0 \right\}.$$

We can regard $\mathcal{P}^+(M)$ as an infinite dimensional manifold whose tangent space at μ is

$$T_\mu\mathcal{P}^+(M) = \left\{ \tau \;\middle|\; \int_M d\tau = 0,\ \int_M \left(\frac{d\tau}{d\mu}\right)^2 d\mu < \infty \right\}.$$

Definition 1. The *Fisher metric* G on $\mathcal{P}^+(M)$ is defined by

$$G_\mu(\tau_1, \tau_2) = \int_M \frac{d\tau_1}{d\mu} \cdot \frac{d\tau_2}{d\mu}\, d\mu, \qquad \tau_1, \tau_2 \in T_\mu\mathcal{P}^+(M).$$

Theorem 2 ([3])**.** (i) The Levi-Civita connection ∇^G of G is given by

$$\nabla^G_{\tau_1}\tau_2 = \frac{1}{2}\left(\frac{d\tau_1}{d\mu} \cdot \frac{d\tau_2}{d\mu} - G_\mu(\tau_1, \tau_2)\right)\mu,$$

where $\tau_1 \in T_\mu\mathcal{P}^+(M)$ and τ_2 is regarded as a constant vector field.

(ii) $(\mathcal{P}^+(M), G)$ is of constant sectional curvature $1/4$.

(iii) the geodesic $\gamma(t)$ satisfying $\gamma(0) = \mu$ and $\dot\gamma(0) = \tau$ is given by

$$\gamma(t) = \left(\cos\frac{t}{2} + \sin\frac{t}{2} \cdot \frac{d\tau}{d\mu}\right)^2 \mu.$$

2 Geodesics and normalized geometric means

Definition 3. We define the *normalized k-power mean* $\varphi^{(k)}(\mu_1, \mu_2)$ of $\mu_1, \mu_2 \in \mathcal{P}^+(M)$ by

$$\varphi^{(k)}(\mu_1, \mu_2) = \frac{1}{C}\left\{1 + \left(\frac{d\mu_2}{d\mu_1}\right)^k\right\}^{1/k} \mu_1 \in \mathcal{P}^+(M),$$

where C is a normalization constant. In particular, we call $\varphi^{(1)}$ and $\varphi^{(0)}$ the arithmetic mean and the normalized geometric mean, respectively.

Theorem 4 ([4, 5])**.** If M is connected, then for any $\mu_1, \mu_2 \in T_\mu\mathcal{P}^+(M)$ there exists a unique geodesic segment $\gamma : [0, \ell] \to \mathcal{P}^+(M)$ joining these two points. Here

(i) $\ell = \ell(\mu_1, \mu_2) := 2\arccos\left(\int_M \sqrt{\frac{d\mu_2}{d\mu_1}}\, d\mu_1\right) \in [0, \pi)$ and ℓ is the distance function of $(\mathcal{P}^+(M), G)$.

(ii) γ is given by $\gamma(t) = a_1(t)\,\mu_1 + a_2(t)\,\mu_2 + a_3(t)\,\varphi^{(0)}(\mu_1, \mu_2)$, where $\{a_i\}_{i=1,2,3}$ are functions on $[0, \ell]$ satisfying $\sum_i a_i(t) = 1,\ a_i(t) \geq 0$.

(iii) $\dot\gamma(0) = \cot(\ell/2)\ \left(\varphi^{(0)}(\mu_1, \mu_2) - \mu_1\right)$ (see Figure, left).

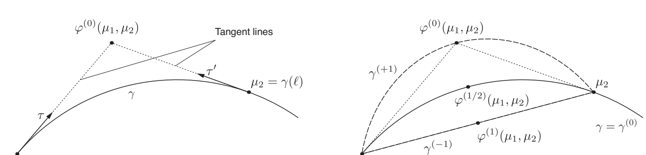

Figure: Geodesic segments and Means

3 A family of affine connections induced by divergences

Definition 5. A *divergence* on $\mathcal{P}^+(M)$ is a function $D : \mathcal{P}^+(M) \times \mathcal{P}^+(M) \to \mathbb{R}$ satisfying the following properties;

(i) $D[\mu : \mu_1] \geq 0$ for $\forall \mu, \mu_1 \in \mathcal{P}^+(M)$ and equality holds iff $\mu_1 = \mu$.

(ii) $\tau_\mu D[\mu : \mu_1]|_{\mu_1=\mu} = \tau_\mu D[\mu_1 : \mu]|_{\mu_1=\mu} = 0$.

(iii) $-\tau_\mu\tau_{\mu_1} D[\mu : \mu_1]|_{\mu_1=\mu} > 0$ for any tangent vector τ.

Example 6 ([1, 2])**.** (i) $D_{\mathrm{KL}}[\mu_1 : \mu_2] := -\int_M \log\left(\frac{d\mu_2}{d\mu_1}\right) d\mu_1$ is called the *Kullback-Leibler divergence.*

(ii) A convex function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(1) = 0,\ f''(0) = 1$ gives $D_f[\mu_1 : \mu_2] := \int_M f\left(\frac{d\mu_2}{d\mu_1}\right) d\mu_1$ which is called the *f-divergence.*

(iii) The f-divergence given by a function

$$f^{(\alpha)}(u) = \begin{cases} u\log u & (\alpha = 1) \\ -\log u & (\alpha = -1) \\ \dfrac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) & (\alpha \neq \pm 1) \end{cases}$$

is called the *α-divergence*, denoted by $D^{(\alpha)}$. $D^{(-1)} = D_{\mathrm{KL}}$ (see (i)).

Remark 7 ([2])**.** A divergence induces a torsion-free dualistic structure, i.e., a metric g and two torsion-free affine connections ∇, ∇^* satisfying

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z).$$

In particular, a dualistic structure on $\mathcal{P}^+(M)$ induced by an f-divergence consists of the Fisher metric and the connection induced by α-divergence;

$$g = G, \quad \nabla = \nabla^{(\alpha)}, \quad \nabla^* = \nabla^{(-\alpha)}, \qquad \alpha = 2f'''(1) + 3.$$

Theorem 8 ([5])**.** (i) The affine connection $\nabla^{(\alpha)}$ at μ is given by

$$\nabla^{(\alpha)}_{\tau_1}\tau_2(\mu) = -\frac{1+\alpha}{2}\left(\frac{d\tau_1}{d\mu}\frac{d\tau_2}{d\mu} - G_\mu(\tau_1, \tau_2)\right)\mu$$

where $\tau_1 \in T_\mu\mathcal{P}^+(M)$ and τ_2 is regarded as a constant vector field.

(ii) For any $\mu_1, \mu_2 \in T_\mu\mathcal{P}^+(M)$ there exists a unique geodesic segment $\gamma^{(\pm 1)} : [0, 1] \to \mathcal{P}^+(M)$ of $\nabla^{(\pm 1)}$ joining these two points, given by

$$\gamma^{(1)}(t) = \left\{\int_M \left(\frac{d\mu_2}{d\mu_1}\right)^t d\mu_1\right\}^{-1} \left(\frac{d\mu_2}{d\mu_1}\right)^t \mu_1, \quad \gamma^{(-1)}(t) = (1-t)\mu_1 + t\mu_2$$

and their midpoints are $\varphi^{(1)}(\mu_1, \mu_2)$ and $\varphi^{(0)}(\mu, \mu_1)$, respectively (see Figure, right).

(iii)
$$\int_0^1 G(\dot\gamma^{(1)}(t), \dot\gamma^{(1)}(t))\, dt = \int_0^1 G(\dot\gamma^{(-1)}(t), \dot\gamma^{(-1)}(t))\, dt = \frac{1}{2}\left(D_{\mathrm{KL}}[\mu_1 : \mu_2] + D_{\mathrm{KL}}[\mu_2 : \mu_1]\right).$$

References

[1] S.-I. Amari, *Information geometry and Its applications*, Applied Mathematical Sciences **194**, Springer, 2016.

[2] S.-I. Amari and H. Nagaoka, Methods of information geometry, Trans. Math. Monogr. **191**, AMS, 2000.

[3] T. Friedrich, Die Fisher-Information und symplektische Strukturen, Math. Nach. **153** (1991), 273-296.

[4] M. Itoh and H. Satoh, Entropy **17** (2015),1814-1849.

[5] M. Itoh and H. Satoh, in preparation.

Acknowledgment This work was supported by JSPS KAKENHI Grant No.15K17545.