Doubly autoparallelism on the space of probability distributions

Atsumi Ohara
University of Fukui

Joint work with Prof. Ishi (Nagoya Univ.)

統計多様体の幾何学とその周辺(9)
October 15-16, 2017 @北大

Introduction

- Information geometry is a branch of differential geometry with Riemannian metric and a pair of affine connections.
- It originates from the study of geometric structure for the family of probability densities in 80's, and is now developing in many ways.
- Widely related to information science, mathematics and statistical physics.

Information geometry on $\mathcal{M}$

Def. Statistical manifold: $(\mathcal{M}, g, \nabla, \nabla^*)$

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

X, Y and Z : arbitrary vector fields on $\mathcal{M}$

★ g : Riemannian metric

★ (∇, ∇^*) : torsion-free affine connections

$R^{\nabla} = 0,\ R^{\nabla^*} = 0$ ➡ dually flat

★ $\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^*$: α-connections

Definition

- Def. Let (S, g, ∇, ∇^*) be a statitical manifold and *M* be its submanifold. We call *M* a doubly autoparallel submanifold in *S* when the followings hold:
  - $\forall X, Y \in \mathcal{X}(M),\ \nabla_X Y \in \mathcal{X}(M)$
  - $\forall X, Y \in \mathcal{X}(M),\ \nabla^*_X Y \in \mathcal{X}(M)$

DA on symmetric cones Ω

- Thm. [UO04] The α-connection is represented by the mutation of Jordan algebra:

$$\left(\nabla^*_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j}\right)_x = -2u_i \perp_{x^{-1}} u_j .$$

- Thm. [OI] Submanifolds $M = (W + p) \cap \Omega$ in a symmetric cone Ω iff the subspace W is a Jordan subalgebra.

Related facts or applications

- MLE for structured covariance matrices is tractable (cast to convex program: inversely linear structure)
  [Anderson70, Malley94]
- Explicitly solvable SDP problems [O99]
- Structure of α-power means on symmetric cones [O04]

- The self-similar (*Barenblatt–Pattle*) solution for the porous medium equation [OW10]

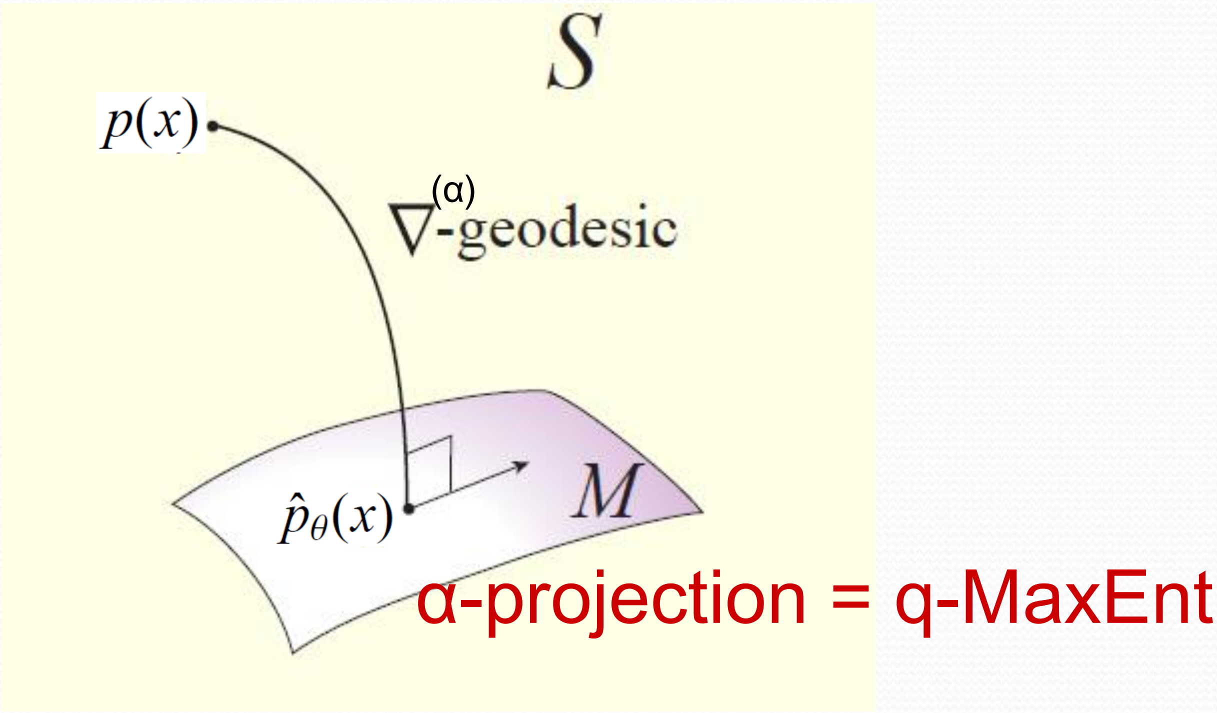
Important Properties

<u>Proposition</u> The following statements are equivalent:

- A submanifold M is doubly autoparallel (DA)
- M is autoparallel w.r.t. the α -connections
  $$\nabla^{(\alpha)} = \{(1+\alpha)\nabla + (1-\alpha)\nabla^*\}/2$$
  for two different α 's.
- M is autoparallel w.r.t. all the α -connections.
- all the α -geodesics connecting two points on M lay in M (if it is simply connected).
- M is affinely constrained in both ∇- and ∇^*-affine coordinates.

Furthermore,

- If M is DA, then α -projections (q-MaxEnt) from p to M are unique for all α if they exist.



The purpose of this talk

Study of DA submanifolds in the space of probability distributions

- Probability simplex = the space of distributions on finite samples

$\Rightarrow$ Linear algebraic approach

Outline

- Preliminaries & examples
- Characterization of DA on the probability simplex
- Classification of DA
- Concluding remarks

Preliminaries

- Positive orthant

$$\mathbf{R}_{++}^{n+1} := \{p \in \mathbf{R}^{n+1} | p_i > 0,\ i = 1, \cdots, n+1\},$$

- Probability simplex

$$\mathcal{S}^n := \left\{ p \in \mathbf{R}_{++}^{n+1} \middle| \sum_{i=1}^{n+1} p_i = 1 \right\}$$

- The set of finite events Ω ={1,2, … , n+1 }

- Discrete probability distributions:

$$p(X = i) = p_i > 0, \ i = 1, \cdots, n+1$$

$$p(X) = \sum_{i=1}^{n+1} p_i \delta_i(X), \quad p = (p_i) \in \mathcal{S}^n \quad \delta_i(j) = \delta_{ij}$$

- statistical model (p_i is parameterized by ξ)

$$p(X;\xi) = \sum_{i=1}^{n+1} p_i(\xi)\delta_i(X)$$

  - Ex: full model $\mathcal{P}_n$: $p_i = \xi^i, \ i = 1, \cdots, n$

$$p(X;\xi) = \sum_{i=1}^{n} \xi^i \delta_i(X) + \left(1 - \sum_{i=1}^{n} \xi^i\right) \delta_{n+1}(X)$$

IG on the simplex (full model)

- Riemannian metric g

 (=Fisher information matrix) $\partial_i = \dfrac{\partial}{\partial p_i}$

$$g_{ij}(p) = \sum_{X \in \Omega} p(X)(\partial_i \log p(X))(\partial_j \log p(X)) \quad i,j = 1,\cdots,n$$

- mutually dual affine connections $\nabla^{(e)}$ and $\nabla^{(m)}$
  - $\nabla^{(e)}$: exponential connection (α=1)

$$\Gamma^{(e)}_{ij,k}(p) = \Gamma^{(1)}_{ijk}(p) = \sum_{X \in \Omega} p(X)(\partial_i \partial_j \log p(X))(\partial_k \log p(X)) \quad i,j,k = 1,\cdots,n$$

  - $\nabla^{(m)}$: mixture connection (α=-1)

$$\Gamma^{(m)}_{ij,k}(p) = \Gamma^{(-1)}_{ijk}(p) = \sum_{X \in \Omega} p(X)(\partial_i \partial_j p(X))(\partial_k \log p(X)) \quad i,j,k = 1,\cdots,n$$

affine coordinates

- $\nabla^{(m)}$ -affine coordinates: (η_i)

$$\eta_i = \sum_{X \in \Omega} p(X)\delta_i(X) = p_i$$

Each p_i is affine w.r.t. $\xi \Leftrightarrow$ the model is $\nabla^{(m)}$-autoparallel

- $\nabla^{(e)}$-affine coordinates: (θ^i)

$$\theta^i = \log\left(\frac{p_i}{1 - \sum_{i=1}^{n} p_i}\right)$$

$$p(X) = \exp\left\{\sum_{i=1}^{n} \theta^i \delta_i(X) - \psi(\theta)\right\} \qquad \psi(\theta) := \log\left(1 + \sum_{i=1}^{n} \exp\theta^i\right)$$

Each θ^i is affine w.r.t. $\xi \Leftrightarrow$ the model is $\nabla^{(e)}$ -autoparallel

Example 1 (1)

- S^n : the probability simlex in $\mathbf{R}^{n+1}$
- $(S^n, g, \nabla^{(e)}, \nabla^{(m)})$, *g*: the Fisher metric.
- *W*: a subspace spanned by
  - *d* (<*n*) vertices $v^{(k)} = (\delta_i^k) \in \mathbf{R}^{n+1}$ in $\overline{S^n}$, and
  - non-vertex point $v^{(0)}$ in $\overline{S^n}$ linearly independent of $\{v^{(k)}\}_{k=1}^d$

$$W = \text{span}\{v^{(0)}, v^{(1)}, \cdots, v^{(d)}\}$$

- $M = W \cap S^n$ is doubly autoparallel.

Example 1 (2)

Proof for *d*=2 (Similar arguments hold for arbitrary *d*.)

- The m-affine coordinates $\eta = (\eta_i)$ of $v^{(i)},\ i = 0, 1, 2$:

$$v^{(1)} = (1\ 0\ \cdots\ 0)^T, \quad v^{(2)} = (0\ 1\ 0\ \cdots\ 0)^T,$$

$$v^{(0)} = (0\ 0\ p_3 \cdots p_{n+1}),\ \sum_{i=3} p_i = 1, \quad p_i > 0$$

- The m-affine coordinates of *p* in *M*:

$$p = \xi_1 v^{(1)} + \xi_2 v^{(2)} + (1 - \xi_1 - \xi_2) v^{(0)} \in M = W \cap \mathcal{S}^n$$

$$\eta_1 = \xi_1,\ \eta_2 = \xi_2,\ \eta_i = (1 - \xi_1 - \xi_2) p_i,\ i = 3, \cdots, n+1,$$

$$(\xi_1 > 0,\ \xi_2 > 0,\ \xi_1 + \xi_2 < 1).$$

affine in ξ_i, *i*=1,2

Example 1 (3)

- The e-affine coordinates of p in M:

$$\theta^i = \log\left(\frac{p_i}{1-\sum_{i=1}^{n} p_i}\right)$$

$$\theta^1 = \zeta_1,\ \theta^2 = \zeta_2,\ \theta^i = \log p_i + c,\ i = 3, \cdots, n+1,$$
$$(\zeta_i = \log\{\xi_i/(1-\xi_1-\xi_2)\},\ i = 1, 2,\ c = -\log p_{n+1}).$$

affine in ζ_i, $i=1,2$

IG on the positive orthant $\mathbf{R}^{n+1}_{++}$

- Riemannian metric $\tilde{g}$

$$\partial_i = \frac{\partial}{\partial p_i}$$

$$\tilde{g}_{ij}(p) = \sum_{X\in\Omega} p(X)(\partial_i \log p(X))(\partial_j \log p(X)) = \frac{\delta_{ij}}{p_i}$$

- $\tilde{\nabla}^{(m)}$: m-connection (α=-1)

$$\tilde{\Gamma}^{(m)}_{ij,k}(p) = \sum_{X\in\Omega} p(X)(\partial_i\partial_j p(X))(\partial_k \log p(X)) = 0$$

(p_i):$\tilde{\nabla}^{(m)}$-affine coordinates

- $\tilde{\nabla}^{(e)}$: e-connection (α=1)

$$\tilde{\Gamma}^{(e)}_{ij,k}(p) = \sum_{X\in\Omega} p(X)(\partial_i\partial_j \log p(X))(\partial_k \log p(X)) = -\frac{\delta^k_{ij}}{p_i}$$

$$\log p(X) = \sum_{X\in\Omega} (\log p_i)\delta_i(X)$$

$(\log p_i)$:$\tilde{\nabla}^{(e)}$-affine coordinates

Denormalization

- Def. denormalization of a submanifold M in $\mathcal{S}^n$

$$\tilde{M} = \{\tau p \in \mathbf{R}_{++}^{n+1} | p \in M,\ \tau > 0\}$$

Lem [Amari & Nagaoka 2000]

The following statements are equivalent:

- A submanifold M is $\nabla^{(\pm 1)}$- autoparallel in $\mathcal{S}^n$,
- A denormalization $\tilde{M}$ is $\tilde{\nabla}^{(\pm 1)}$ -autoparallel in $\mathbf{R}_{++}^{n+1}$.

observations

- W: a subspace in $\mathbf{R}^{n+1}$

  $M = W \cap S^n \Leftrightarrow \tilde{M} = W \cap \mathbf{R}^{n+1}_{++}$ is $\tilde{\nabla}^{(m)}$-autoparallel

  $\Leftrightarrow M$ is $\nabla^{(m)}$ -autoparallel

- $\log \tilde{M} = b + W'$, $\begin{cases} W' \text{ is another subspace of the same dim.} \\ b \text{ is a constant vector in } \mathbf{R}^{n+1}. \end{cases}$

  $\Leftrightarrow$ $\tilde{M}$ is $\tilde{\nabla}^{(e)}$ -autoparallel

  $\Leftrightarrow$ M is $\nabla^{(e)}$ -autoparallel

where $\log W = \{\log p | p \in W\}, \quad \log p = (\log p_i) \in \mathbf{R}^{n+1}$

Main results

- Thm Assume $a \in \tilde{M} = W \cap \mathbf{R}^{n+1}_{++} \ (= W_{++})$.

$$\exists W', \ \log(a+W)_{++} = \log a + W'$$
$$\Leftrightarrow \ 1) W' = a^{-1} \circ W, \quad 2) \forall u, v \in W, \ u \circ a^{-1} \circ w \in W$$

  - Here, $(\mathbf{R}^{n+1}, \circ)$ is defined by Hadamard product $\circ$, i.e.,

$$x \circ y = (x^i) \circ (y^i) = (x^i y^i), \quad e = \mathbf{1}, \quad x^{-1} = \left(\frac{1}{x^i} \right)$$

- Rem 2) implies *W* should be a subalgebra in $(\mathbf{R}^{n+1}, \circ_{a^{-1}})$.
  where $\circ_{a^{-1}} := \circ a^{-1} \circ$ is a mutation of $\circ$ by *a*.

Main results

Characterization of DA

- Cor A $\nabla^{(m)}$-autoparallel submanifold $M = W \cap \mathcal{S}^n$ is DA iff the subspace *W* is a subalgebra of $(\mathbf{R}^{n+1}, \circ_{a^{-1}})$ with the identity element $a \in \tilde{M}$.

Main results

Classification for *W*

- <u>Thm</u> (Classification for *W*)

  W is a subalgebra in $(\mathbf{R}^{n+1}, \circ_{a^{-1}})$ with $a \in \tilde{M}$

iff *W* is of the form:

$$W = \mathbf{R}^q \times \mathbf{R}a_1 \times \cdots \times \mathbf{R}a_r$$

i.e.,

$$W = \{x = (y^T \ t_1 a_1^T \ \cdots \ t_r a_r^T)^T \mid y \in \mathbf{R}^q, \ a_i \in \mathbf{R}^{n_i}_{++}, \ t_i \in \mathbf{R}, i = 1, \cdots, r\}$$

$$\text{, where } q + \sum_{i=1}^{r} n_i = n + 1, \quad q \geqq 0, \quad r > 0, \quad 2 \leq n_1 \leq \cdots \leq n_r$$

up to permutations of elements.

<u>Rem</u> dim $W = q+r$.

Example (continued)

- For $a \in \tilde{M} = W \cap \mathbf{R}^{n+1}$, we set

$$a = (1\ 2\ p_3\ \cdots\ p_{n+1})^T,\ a_0 = (1\ 2)^T,\ a_1 = (p_3\ \cdots\ p_{n+1})^T.$$

$$\left(\sum_{i=3}^{n+1} p_i = 1,\ p_i > 0,\ i = 3, \cdots, n+1\right)$$

- ***q*=2, *r*=1,** $n_1 = n - 1$

$$V = a^{-1} \circ W = \{(z^T\ t\mathbf{1}^T)^T \in \mathbf{R}^{n+1} |\ \forall z \in \mathbf{R}^2,\ t\mathbf{1} \in \mathbf{R}^{n-1},\ \forall t \in \mathbf{R}\}$$

- *W*={ $w = (\xi_1\ \xi_2\ tp_3\ \cdots\ tp_{n+1})^T$ }

- Every elements in $M = W \cap \mathcal{S}^n$ is represented by

$$w = (\xi_1\ \xi_2\ tp_3\ \cdots\ tp_{n+1})^T, \quad t := 1 - \xi_1 - \xi_2,$$

Conclusions

- Characterization of DA submanifolds for the space of discrete probability distributions
- Its classification
  - Algebraic structure is closely related.
- Applications (future work)
  - Statistical modeling
    - Stochastic reasoning (Belief Propagation [Ikeda et al 04])?
    - Explicitly solvable LP problems
- Relation with Markov embeddings [Nagaoka 2017]

Ref. A. Ohara and H. Ishi, arXiv:1711.11456v1 (2017)