IGAIA IV, Liblice, June 2016

In honor of Amari-sensei

Revisit to the Autoparallelity
and the Canonical Divergence
for Dually Flat Spaces

Hiroshi Nagaoka

The University of Electro-Communications, Tokyo.

- A talk on a similar subject was given in a workshop held in Nara, March 2012 without details.

- A related result (but in a different context) will be presented in the forthcoming IEEE ISIT, July 2016, with the title ”A characterization of statistical manifolds on which the relative entropy is a Bregman divergence”.

First, I would like to review some of the fundamental concepts of information geometry, and then state the main result and explain the proof of some nontrivial parts.

A dually flat space

A manifold S equipped with (g, ∇, ∇^*) such that

(1) g is a Riemannian metric.

(2) ∇ and ∇^* are flat affine connections which are dual w.r.t. g:

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

(3) S is covered by a global ∇-affine chart θ and a global ∇^*-affine chart η.

The canonical divergence

When θ and η are chosen to satisfy

$$g\left(\partial_i, \partial^j\right) = \delta^i_j, \qquad \partial_i := \frac{\partial}{\partial \theta^i}, \quad \partial^j := \frac{\partial}{\partial \eta_j}$$

the canonical divergence D: $S \times S \to \mathbb{R}$ is defined by

$$(1) \qquad D(p\|q) = \varphi(p) + \psi(q) - \sum_i \eta_i(p)\theta^i(q)$$

by functions $\psi, \varphi : M \to \mathbb{R}$ satisfying

$$(2) \qquad \eta_i = \partial_i \psi,$$

$$(3) \qquad \theta^i = \partial^i \varphi,$$

$$(4) \qquad \varphi + \psi = \sum_i \eta_i \, \theta^i$$

$\boxed{\text{Another characterization of the canonical divergence}}$

The canonical divergence is also defined as a function $D : S \times S \to \mathbb{R}$ such that for any $p, q, r \in S$

(1) $D(p\|q) \geq 0$

(2) $D(p\|q) = 0 \;\Leftrightarrow p = q$

(3) $D(p\|q) + D(q\|r) - D(p\|r) = \sum_i \{\eta_i(p) - \eta_i(q)\} \left\{\theta^i(r) - \theta^i(q)\right\}$

Note: (3) is the necessary and sufficient condition for mappings $\theta : S \to \mathbb{R}^{\dim S}$ and $\eta : S \to \mathbb{R}^{\dim S}$ to be a pair of dual affine charts.

Autoparallelity

- For a manifold S with an affine connection ∇ and a submanifold $M \subset S$,

$$\begin{aligned} & M \text{ is } \nabla\text{-autoparallel} \ \ (\text{in } S). \\ \overset{\text{def}}{\Longleftrightarrow}\ & \forall X, Y\colon \text{ vector fields on } M,\ \nabla_X Y \text{ is a vector field on } M. \\ \Longleftrightarrow\ & \text{The restriction } \nabla|_M \text{ becomes an affine connection on } M. \\ \Longleftrightarrow\ & M \text{ is } \nabla\text{-totally geodesic (when } \nabla \text{ is torsion-free).} \end{aligned}$$

- We denote

$$M \overset{\nabla}{\subset} S \overset{\text{def}}{\Longleftrightarrow} M \text{ is } \nabla\text{-autoparallel in } S$$

- When S is ∇-flat with a ∇-affine chart $\theta : S \to \mathbb{R}^{\dim S}$,

$$M \overset{\nabla}{\subset} S \Longleftrightarrow \theta(M) \text{ forms (an open subset of) an affine subspace of } \mathbb{R}^{\dim S}.$$

$$\Longrightarrow M \text{ is } \nabla|_M\text{-flat. (or we simply say } M \text{ is } \nabla\text{-flat.)}$$

Autoparallel submanifolds of a dually flat space

When

(1) (S, g, ∇, ∇^*) is dually flat with canonical divergence D and

(2) either $M \overset{\nabla}{\subset} S$ or $M \overset{\nabla^*}{\subset} S$,

then

(3) M is dually flat w.r.t. $(g|_M, \nabla|_M, \pi_M(\nabla^*))$ or $(g|_M, \pi_M(\nabla), \nabla^*|_M)$

(where π_M is the g-orthogonal projection onto M)

(4) The canonical divergence of M is $D|_{M\times M}$.

$\boxed{\mathscr{P}(\mathscr{X}) \text{ and its autoparallel submanifolds}}$

— as a representative example —

- $\mathscr{X}$: an arbitrary finite set
- $S = \mathscr{P}(\mathscr{X}) := \{p \,|\, p : \mathscr{X} \to (0,1), \;\; \sum_x p(x) = 1\}$
- $g = g^{(\mathrm{F})}$: Fisher metric
- $\nabla = \nabla^{(\mathrm{e})}$: exponential connection, e-connection
- $\nabla^* = \nabla^{(\mathrm{m})}$: mixture connection, m-connection
- $D = D_{\mathrm{KL}}$: $D_{\mathrm{KL}}(p\|q) = \sum_x p(x) \log \dfrac{p(x)}{q(x)}$
- We abbreviate $\nabla^{(\mathrm{e})}$- and $\nabla^{(\mathrm{m})}$- as e- and m-, respectively:
  e.g., e-geodesic, m-autoparallel , etc.

$\boxed{\mathscr{P}(\mathscr{X}) \text{ and its autoparallel submanifolds; cont.}}$

- $M \overset{\mathrm{e}}{\subset} \mathscr{P}(\mathscr{X}) \iff M$ is an exponential family on $\mathscr{X}$;
  (e-family for short)

  $$\log p_\theta(x) = C(x) + \sum_i \theta^i F_i(x) - \psi(\theta)$$

  $\theta = (\theta^i)$: an e-affine chart

  $$\eta_i(p) = \mathrm{E}_p[F_i(X)] = \sum_x F_i(x) p(x)$$

  $\eta = (\eta_i)$: an m-affine chart

- $M \overset{\mathrm{m}}{\subset} \mathscr{P}(\mathscr{X}) \iff M$ is a mixture family on $\mathscr{X}$;
  (m-family for short)

  $$p_\eta(x) = C(x) + \sum_i \eta_i F^i(x)$$

  $\eta = (\eta_i)$: an m-affine chart

  $$\theta^i(p) = \sum_x F^i(x) \log p(x)$$

  $\theta = (\theta^i)$: an e-affine chart

Chains of autoparallel submanifolds

Given a dually flat space (S, g, ∇, ∇^*) with canonical divergence D,

- If $K \overset{\nabla}{\subset} S$ and $M \overset{\nabla|_K}{\subset} S$, then

  - $M \overset{\nabla}{\subset} S$,

  - M is dually flat w.r.t. $(g|_M, \nabla|_M, \pi_M(\nabla^*))$
  with canonical divergence $D|_{M\times M}$.

- For simplicity, we write this implication as

$$M \overset{\nabla}{\subset} K \overset{\nabla}{\subset} S \Longrightarrow M \overset{\nabla}{\subset} S \quad \text{and} \quad M \text{ is dually flat with canonical divergence } D.$$

  Note: the canonical divergence determines the dually flat structure.

- Similarly,

$$M \overset{\nabla^*}{\subset} K \overset{\nabla^*}{\subset} S \Longrightarrow M \overset{\nabla^*}{\subset} S \quad \text{and} \quad M \text{ is dually flat with canonical divergence } D.$$

- On the other hand,

  either $M \overset{\nabla^*}{\subset} K \overset{\nabla}{\subset} S$ or $M \overset{\nabla}{\subset} K \overset{\nabla^*}{\subset} S$

  $\Longrightarrow$ not $M \overset{\nabla}{\subset} S$ nor $M \overset{\nabla^*}{\subset} S$ in general,

  but M is still dually flat with canonical divergence D.

$\boxed{\text{For instance}}$

- $M \overset{\mathrm{e}}{\subset} K \overset{\mathrm{e}}{\subset} \mathscr{P}(\mathscr{X}) \implies M \overset{\mathrm{e}}{\subset} \mathscr{P}(\mathscr{X})$ (M is an e-family) and M is dually flat with canonical divergence D_{KL}.
- $M \overset{\mathrm{m}}{\subset} K \overset{\mathrm{m}}{\subset} \mathscr{P}(\mathscr{X}) \implies M \overset{\mathrm{m}}{\subset} \mathscr{P}(\mathscr{X})$ (M is an m-family) and M is dually flat with canonical divergence D_{KL}.
- either $M \overset{\mathrm{m}}{\subset} K \overset{\mathrm{e}}{\subset} \mathscr{P}(\mathscr{X})$ or $M \overset{\mathrm{e}}{\subset} K \overset{\mathrm{m}}{\subset} \mathscr{P}(\mathscr{X})$

  $\implies$ not $M \overset{\mathrm{e}}{\subset} \mathscr{P}(\mathscr{X})$ nor $M \overset{\mathrm{m}}{\subset} \mathscr{P}(\mathscr{X})$ in general,

  but M is dually flat with canonical divergence D_{KL}.

$\boxed{\text{Main theorem}}$

Given a dually flat space (S, g, ∇, ∇^*) with canonical divergence D and a submanifold $M \subset S$, the following conditions are equivalent.

(1) M is dually flat with canonical divergence D.

(2) $M \overset{\nabla^*}{\subset} \exists K \overset{\nabla}{\subset} S$.

(3) $M \overset{\nabla}{\subset} \exists K \overset{\nabla^*}{\subset} S$.

(4) $\exists K_1 \overset{\nabla}{\subset} S$ and $\exists K_2 \overset{\nabla^*}{\subset} S$ such that

$$M = K_1 \cap K_2 \quad \text{and} \quad \forall p \in M,\ T_p(K_1)^\perp \perp T_p(K_2)^\perp$$

(5) $\exists K_1 \overset{\nabla}{\subset} S$ and $\exists K_2 \overset{\nabla^*}{\subset} S$ such that

$$M = K_1 \cap K_2 \quad \text{and} \quad \exists p \in M,\ T_p(K_1)^\perp \perp T_p(K_2)^\perp$$

$\boxed{\text{Proof of (1)} \implies \text{(2): 1/5 slides}}$

We show that

(1) M is dually flat with canonical divergence D.

implies

(2) $M \overset{\nabla^*}{\subset} \exists K \overset{\nabla}{\subset} S$.

- Let $m := \dim M \leq n := \dim S$.
- Since (S, g, ∇, ∇^*) is dually flat, there exist a ∇- affine chart $\sigma : S \to \mathbb{R}^{n\times 1}$ (column vectors) and a ∇^*-affine chart $\zeta : S \to \mathbb{R}^{1\times n}$ (row vectors) such that

  $\forall p, q, r \in S$

  $$D(p\|q) + D(q\|r) - D(p\|r) = (\zeta(p) - \zeta(q))\,(\sigma(r) - \sigma(q))\,.$$
- Assume (1), so that there exists a pair of affine charts $\theta : M \to \mathbb{R}^{m\times 1}$ and $\eta : M \to \mathbb{R}^{1\times m}$ satisfying

  $\forall p, q, r \in M$

  $$D(p\|q) + D(q\|r) - D(p\|r) = (\eta(p) - \eta(q))\,(\theta(r) - \theta(q))\,.$$

$\boxed{\text{Proof of (1)} \implies \text{(2), 2/5 slides}}$

- Fix a point $p_0 \in M$ arbitrarily, and let

$$V := \operatorname{span}\{\sigma(p) - \sigma(p_0) \,|\, p \in M\} \subset \mathbb{R}^{n\times 1}$$

and

$$K := \{p \in S \,|\, \sigma(p) - \sigma(p_0) \in V\} \subset S.$$

Note: the definitions of V and K do not depend on p_0.

- It is obvious that

$$M \subset K \overset{\nabla}{\subset} S.$$

- Let $k := \dim V = \dim K$. $\quad (m \leq k \leq n)$

Then there exists a matrix $F : n \times k$ such that $V = \mathrm{Im}F$.

- Define a ∇-affine chart $\rho : K \to \mathbb{R}^{k\times 1}$ (column vectors) of K by

$$\forall p \in K, \quad \sigma(p) - \sigma(p_0) = F\rho(p).$$

- Let $\xi : K \to \mathbb{R}^{1\times k}$ (row vectors) be defined by

$$\forall p \in K, \quad \xi(p) = \zeta(p)F.$$

Proof of (1) $\Longrightarrow$ (2), 3/5 slides

- For $\forall p, q, r \in K$, we have

$$
\begin{aligned}
(\xi(p) - \xi(q))\,(\rho(r) - \rho(q)) &= (\zeta(p) - \zeta(q))\,F\,(\rho(r) - \rho(q)) \\
&= (\zeta(p) - \zeta(q))\,(\sigma(r) - \sigma(q)) \\
&= D(p\|q) + D(q\|r) - D(p\|r),
\end{aligned}
$$

which implies that ξ is ∇^*-affine chart of K.

manifold	dimension	∇-chart (column vectors)	∇^*-chart (row vectors)
S	n	σ	ζ
K	k	ρ	ξ
M	m	θ	η

Proof of (1) $\Longrightarrow$ (2), 4/5 slides

- Lemma:

$$\forall f \in V\ (\subset \mathbb{R}^{n\times 1}),\ \ \exists a \in \mathbb{R}^{m\times 1},\ \ \forall p \in M,$$

$$(\zeta(p) - \zeta(p_0))\, f = (\eta(p) - \eta(p_0))\, a. \qquad (\sharp)$$

$\because$) It suffices to show ($\sharp$) in the case when $f = \sigma(q) - \sigma(p_0)$ for an arbitrary $q \in M$, since V is the linear span of such f's. In this case, for $\forall p \in M$ we have

$$\begin{aligned}(\zeta(p) - \zeta(p_0))\, f &= (\zeta(p) - \zeta(p_0))\,(\sigma(q) - \sigma(p_0)) \\ &= D(p\|p_0) + D(p_0\|q) - D(p\|q) \\ &= (\eta(p) - \eta(p_0))\,(\theta(q) - \theta(p_0))\,,\end{aligned}$$

which means that ($\sharp$) holds by setting $a := \theta(q) - \theta(p_0)$.

- Since $V = \mathrm{Im}F$, the previous lemma implies the existence of a matrix $A : m \times k$ such that

$$\forall p \in M,\ \ (\zeta(p) - \zeta(p_0))\, F = (\eta(p) - \eta(p_0))\, A. \qquad (\flat)$$

Proof of (1) $\Longrightarrow$ (2), 5/5 slides

- Since the LHS of ($\flat$) is $(\zeta(p) - \zeta(p_0))\,F = \xi(p) - \xi(p_0)$, we have

$$\forall p \in M, \;\; \xi(p) = \eta(p)A + b \qquad (\natural)$$

where $b := \xi(p_0) - \eta(p_0)A$.

- ($\natural$) means that $\xi(M)$ forms an affine subspace of the ξ-coordinate space $\mathbb{R}^k$ and hence $M \overset{\nabla^*}{\subset} K$. (QED)

manifold	dimension	∇-chart (column vectors)	∇^*-chart (row vectors)
S	n	σ	ζ
K	k	ρ	ξ
M	m	θ	η

Proof of (2) $\Longrightarrow$ (4), an outline

(2) $M \overset{\nabla^*}{\subset} \exists K \overset{\nabla}{\subset} S$.

$\Longrightarrow$ (4) $\exists K_1 \overset{\nabla}{\subset} S$ and $\exists K_2 \overset{\nabla^*}{\subset} S$ such that

$$M = K_1 \cap K_2 \quad \text{and} \quad \forall p \in M,\ T_p(K_1)^\perp \perp T_p(K_2)^\perp$$

- Assume that $M \overset{\nabla^*}{\subset} K \overset{\nabla}{\subset} S$ and fix a point $p_0 \in M$ arbitrarily. Let $\sigma : S \to \mathbb{R}^{n\times 1}$ and $\zeta : S \to \mathbb{R}^{1\times n}$ be a pair of dual affine charts.

- Since $K \overset{\nabla}{\subset} S$, there exits a matrix $F : n \times k$ such that

$$K = \{p \in S \,|\, \sigma(p) - \sigma(p_0) \in \mathrm{Im} F\}.$$

  Let $\xi : K \to \mathbb{R}^{1\times k}$ be defined by $\xi(p) = \zeta(p)F$ for $\forall p \in K$. Then ξ becomes a ∇^* chart of K.

- Since $M \overset{\nabla^*}{\subset} K$, there exists a linear subspace W of $\mathbb{R}^{1\times k}$ such that

$$\begin{aligned} M &= \{p \in K \,|\, \xi(p) - \xi(p_0) \in W\} \\ &= \{p \in K \,|\, \zeta(p)F - \xi(p_0) \in W\} = K_1 \cap K_2 \end{aligned}$$

  where $K_1 := K$ and $K_2 := \{p \in S \,|\, \zeta(p)F - \xi(p_0) \in W\}$.

- $K_1 \overset{\nabla}{\subset} S$ and $K_2 \overset{\nabla^*}{\subset} S$ are obvious, and

$$\forall p \in M,\ T_p(K_1)^\perp \perp T_p(K_2)^\perp$$

  can be verified by some linear algebraic argument.

Let us observe that the set of stationary markov joint distributions forms a dually flat space which can be regarded as an example of our theorem.

Markov joint distributions

- Let $\mathscr{X}$ be an arbitrary finite set and let $\mathscr{X}^n = \underbrace{\mathscr{X} \times \cdots \times \mathscr{X}}_{n}$.
- $S := \mathscr{P}(\mathscr{X}^n)$, the set of positive n-joint distributions.
- An n-joint distribution $p \in S$ is markov

$$\overset{\text{def}}{\iff} \; X_1 - X_2 - \cdots - X_n \text{ is a Markov process}$$

for $X^n = (X_1, \cdots, X_n) \sim p$.

$$\iff \; \exists u_1, \ldots, u_{n-1} : \mathscr{X}^2 \to \mathbb{R}^+, \quad \forall x^n = (x_1, \ldots, x_n) \in \mathscr{X}^n,$$

$$p(x^n) = \prod_{t=1}^{n-1} u_t(x_t, x_{t+1}).$$

$$\iff \; \exists \pi \in \mathscr{P}(\mathscr{X}), \;\; \exists w_1, \ldots, w_{n-1} \in \mathscr{P}(\mathscr{X} \,|\, \mathscr{X}),$$

$$\forall x^n = (x_1, \ldots, x_n) \in \mathscr{X}^n,$$

$$p(x^n) = \pi(x_1) \prod_{t=1}^{n-1} w_t(x_{t+1}|x_t).$$

Markov joint distributions, cont.

- $K_{\mathrm{mar}} := \{p \in S \mid p \text{ is markov}\}$.
- K_{mar} is an exponential family.
  ($\because$ the Hammersley-Clifford theorem)
- For $\mathscr{X} = \{0, 1, \ldots, m-1\}$,

$$N := \dim K_{\mathrm{mar}} = n(m-1) + (n-1)(m-1)^2,$$

and $K_{\mathrm{mar}} = \{p_\rho \mid \rho = (\rho^{it}; \rho^{ijt}) \in \mathbb{R}^N\}$ where

$$\log p_\rho(x^n) = \sum_{t=1}^{n} \sum_{i=1}^{m-1} \rho^{it} \delta_i(x_t) + \sum_{t=1}^{n-1} \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \rho^{ijt} \delta_{ij}(x_t, x_{t+1}) - \psi(\rho).$$

Stationary joint distributions

- For an n-joint distribution $p \in S = \mathscr{P}(\mathscr{X}^n)$, its marginal distributions $p_t^{(k)} \in \mathscr{P}(\mathscr{X}^k)$ for $k \in \{1, \ldots, n-1\}$ and $t \in \{1, \ldots, n-k+1\}$ are defined by

$$p_1^{(n-1)}(x^{n-1}) = \sum_{x'} p(x^{n-1}, x'),$$

$$p_2^{(n-1)}(x^{n-1}) = \sum_{x'} p(x', x^{n-1}),$$

$$p_1^{(n-2)}(x^{n-2}) = \sum_{x'} p_1^{(n-1)}(x^{n-2}, x'),$$

$$p_2^{(n-2)}(x^{n-2}) = \sum_{x'} p_1^{(n-1)}(x', x^{n-2}) = \sum_{x'} p_2^{(n-1)}(x^{n-2}, x'),$$

$$p_3^{(n-2)}(x^{n-2}) = \sum_{x'} p_2^{(n-1)}(x', x^{n-2}),$$

$$\cdots\cdots\cdots\cdots$$

- An n-joint distribution $p \in S$ is stationary

$$\stackrel{\text{def}}{\Longleftrightarrow} \quad p_1^{(n-1)} = p_2^{(n-1)}$$

$$\Longleftrightarrow \quad \forall k \in \{1, \ldots, n-1\},\ \ p_t^{(k)} \text{ does not depend on } t.$$

Stationary joint distributions, cont.

- $K_{\mathrm{sta}} := \{p \in S \mid p \text{ is stationary}\}$.
- K_{sta} is a mixture family.

Stationary markov joint distributions

- $M := K_{\mathrm{mar}} \cap K_{\mathrm{sta}}$.
- $p \in M \iff \exists w \in \mathscr{P}(\mathscr{X}|\mathscr{X})$ s.t. $\forall x^n \in \mathscr{X}^n$,

$$p(x^n) = p_w^{(n)}(x^n) := \pi_w(x_1) \prod_{t=1}^{n-1} w(x_{t+1}|x_t)$$

where π_w is the stationary distribution of w:

$$\pi_w(x) = \sum_{x'} w(x|x')\pi_w(x').$$

- It does not hold that $\forall p \in M$, $T_p(K_{\mathrm{mar}})^{\perp} \perp T_p(K_{\mathrm{sta}})^{\perp}$.

Stationary markov joint distributions, cont.

- Let

$$K_{\mathrm{sta},2} := \left\{ p \in S \mid p_t^{(2)} \text{ does not depend on } t \right\} \quad (\supset K_{\mathrm{sta}}).$$

  Then

  - $K_{\mathrm{sta},2}$ is a mixture family.
  - $M = K_{\mathrm{mar}} \cap K_{\mathrm{sta},2}$.
  - $\forall p \in M,\ T_p(K_{\mathrm{mar}})^{\perp} \perp T_p(K_{\mathrm{sta},2})^{\perp}$.

- Therefore, we have

  - M is dually flat with canonical divergence D_{KL}.
  - $M \overset{\mathrm{m}}{\subset} K_{\mathrm{mar}}$ and $M \overset{\mathrm{e}}{\subset} K_{\mathrm{sta},2}$.

- $D = D_{\mathrm{KL}}|_{M \times M}$ is represented as

$$D(p_{w_1}^{(n)} \| p_{w_2}^{(n)}) = D(\pi_{w_1} \| \pi_{w_2}) + (n-1) \sum_x \pi_{w_1}(x) D(w_1(\cdot|x) \| w_2(\cdot|x))$$

- M is not an e-family nor m-family.

Stationary markov joint distributions, cont. cont.

- Recall that $K_{\mathrm{mar}} = \{p_\rho \,|\, \rho \in \mathbb{R}^N\}$ with

$$\log p_\rho(x^n) = \sum_{t=1}^{n} \sum_{i=1}^{m-1} \rho^{it} \delta_i(x_t) + \sum_{t=1}^{n-1} \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \rho^{ijt} \delta_{ij}(x_t, x_{t+1}) - \psi(\rho).$$

- There exists a subset $U \subset \mathbb{R}^N$ such that $M = \{p_\rho \,|\, \rho \in U\}$.

- A pair of dual affine charts of M is given by

  - e-affine chart $\theta = (\theta^i, \theta^{ij})$:

$$\theta^i = \sum_{t=1}^{n} \rho^{it}, \quad \theta^{ij} = \sum_{t=1}^{n-1} \rho^{ijt} \qquad (\forall \rho \in U)$$

  - m-affine chart: $\eta = (\eta_i, \eta_{ij})$:

$$\eta_i(p) = E_p\left[\delta_i(X_t)\right], \quad \eta_{ij}(p) = E_p\left[\delta_{ij}(X_t, X_{t+1})\right]$$

(do not depend on t)

Thank you for listening!