Self-study: Information Geometry Basis

Tianpei Xie

Nov. 6th., 2022

Contents

1 Geometry of $\mathcal{P}(\mathcal{X})$

1.1 Definitions

- Let $\mathcal{P}(\mathcal{X})$ be the set of ***probability density functions*** on $\mathcal{X}$ with respect to base measure μ

$$\mathcal{P}(\mathcal{X}) := \left\{ p : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} p(x) d\mu(x) = 1,\ p(x) > 0 \, (\forall x \in \mathcal{X}). \right\}$$

  In general, $p = \frac{dP}{d\mu}$ is ***the Radon-Nikodym derivative*** where μ is σ-finite measure on a measurable set $(\mathcal{X}, \mathcal{B})$ with $\mathcal{B}$ being *the Borel field* consisting of $\mathcal{X}$ and its subsets. P is ***the probability measure*** that is *absolutely continuous* with respect to μ. We also assume that ***the support of*** p ***covers*** $\mathcal{X}$ so that $p(x) > 0$ for all $x \in \mathcal{X}$.

- Define $S \subseteq \mathcal{P}(\mathcal{X})$ as a family of probability densities on $\mathcal{X}$. Suppose for each probability function can be parameterized as $p_{\xi} = p(x; \xi) \in S$, where $\xi = (\xi^1, \ldots, \xi^n) \in \Xi \subseteq \mathbb{R}^n$. Thus

$$S := \{ p_{\xi} = p(x; \xi) : \xi \in \Xi \subseteq \mathbb{R}^n \}$$

  and $\xi \mapsto p_{\xi}$ is injective. We call S as an ***n-dimensional statistical model***, a ***parametric model***, simply a ***model*** on $\mathcal{X}$.

- Define the space of all *real-valued* ***measurable functions*** on $\mathcal{X}$ as $\mathbb{R}^{\mathcal{X}} := \{ f : \mathcal{X} \to \mathbb{R} \}$. $\mathbb{R}^{\mathcal{X}}$ is an ***infinite-dimensional vector space*** under function addition and scalar multiplication. We see that $\mathcal{P}(\mathcal{X}) \subseteq \mathbb{R}^{\mathcal{X}}$, is an ***affine subspace*** of $\mathbb{R}^{\mathcal{X}}$. Moreover, since $\mathbb{R}^{\mathcal{X}}$ is a metric space, with metric topology, we assume that $\mathcal{P}(\mathcal{X})$ has ***subspace topology***.

- Assume that the statistical model $S = \{ p(x; \xi) : \xi \in \Xi \}$ is ***a topological manifold*** equipped with ***smooth structure*** $\{ (U_{\alpha}, \varphi_{\alpha}) \}$ where each smooth chart (U, φ) is defined and $\varphi : U \to \widehat{U} \subseteq \mathbb{R}^n$ is defined by $\varphi(p_{\xi}) = \xi := (\xi^1, \ldots, \xi^n)$. For any $(U_{\alpha}, \varphi_{\alpha})$ and $(U_{\beta}, \varphi_{\beta})$ such that $U_{\alpha} \cap U_{\beta} \neq \emptyset$, we have $\varphi_{\beta} \circ \varphi_{\alpha}^{-1}$ being a diffeomorphism. That is, S ***is a n-dimenisional smooth manifold***. We may call S ***a statistical manifold***.

- Define $\ell : \mathcal{P} \to \mathbb{R}^{\mathcal{X}}$ as $\ell(p) = \log(p)$. ℓ is the ***log-likelihood function***. Under the subspace topology in $\mathcal{P}$, ℓ is ***continous*** mapping, and is ***injective***. It is a ***homemorphism*** onto its image $\ell : \mathcal{P} \to \ell(\mathcal{P}) \subseteq \mathbb{R}^{\mathcal{X}}$ with its inverse being $(\ell)^{-1}(f) = \exp(f)$ for $f \in \ell(\mathcal{P})$. The ***restriction*** of ℓ on statistical manifold S is a ***smooth injection*** since the *differential* of ℓ at p as $d\ell_p = p^{-1} dp = p_{\xi}^{-1} (\partial_i \, p_{\xi}) d\xi^i \neq 0$ for all $\xi \in \Xi$. Moreover, $d\ell_p$ is also ***injective***, thus ℓ is ***an injective immersion***. Since ℓ is also a homemorphism onto its image, the log-likelihood ℓ is ***a smooth embedding***.

- The ***Fisher Information matrix*** for $p_{\xi} \in S$ is defined as

$$\begin{aligned} g_{i,j}(\xi) &= \mathbb{E}_p \left[\frac{\partial}{\partial \xi^i} \ell_{\xi} \, \frac{\partial}{\partial \xi^j} \ell_{\xi} \right] := \int_{\mathcal{X}} \frac{\partial}{\partial \xi^i} \log p(x; \xi) \frac{\partial}{\partial \xi^j} \log p(x; \xi) d\mu \qquad (1) \\ &= -\mathbb{E}_p \left[\frac{\partial^2}{\partial \xi^i \, \xi^j} \ell_{\xi} \right], \quad \forall i, j = 1, \ldots, n \\ G(\xi) &= [g_{i,j}(\xi)] \succeq 0 \end{aligned}$$

  since $\partial_i \int_{\mathcal{X}} p_{\xi} d\mu = \int_{\mathcal{X}} \partial_i p_{\xi} d\mu = 0$, thus $\mathbb{E}_p \left[\partial_i \ell_{\xi} \right] = \int \partial_i \ell_{\xi} = \int p_{\xi}^{-1} \partial_i p_{\xi} = 0$.

Let us ***assume*** that ***the Fisher Information matrix is positive definite*** for all $\xi \in \Xi$. This is ***equivalent*** to say that the n-tuple

$$\left(\frac{\partial}{\partial \xi^1}\ell_\xi, \ldots, \frac{\partial}{\partial \xi^n}\ell_\xi\right) \subset \mathbb{R}^{\mathcal{X}} \textit{ are } \textbf{\textit{linearly independent}}.$$

1.2 $\mathcal{P}(\mathcal{X})$ as Embedded Submanifold

- As discussed above, $\mathcal{P}(\mathcal{X}) \subseteq \mathbb{R}^{\mathcal{X}}$ is a subspace in $\mathbb{R}^{\mathcal{X}}$. In fact, it is *an open subset* of ***the affine subspace*** $\mathcal{A}_0 := \left\{A : \int_{\mathcal{X}} A(x) = 1\right\}$.
- Given $|\mathcal{X}| < \infty$, $\mathcal{P}(\mathcal{X})$ is ***an embedded submanifold of*** $\mathbb{R}^{\mathcal{X}}$ under two different embeddings:
  1. The ***natural inclusion map*** $\iota : \mathcal{P} \hookrightarrow \mathbb{R}^{\mathcal{X}}$ is an ***embedding***. If we assume that the probability density function is smooth, then ι is a *smooth embedding* as well. We call it ***the mixture embedding***.

     The ***tangent space*** $T_p^{(m)}\mathcal{P}$ ***under this embedding*** is the *subspace* of $T_p\mathbb{R}^{\mathcal{X}} \simeq \mathbb{R}^{\mathcal{X}}$. In particular,

     $$T_p^{(m)}\mathcal{P} = \mathcal{A}_0 = \left\{A \in \mathbb{R}^{\mathcal{X}} : \int_{\mathcal{X}} A(x) d\mu = 0\right\}$$

     Denote the tangent vector under this embedding as $X^{(m)} = d\iota_p(X)$. That is, $X^{(m)}$ is a representation of the tangent vector $X \in T_p\mathcal{P}$ when considered as an element of $\mathcal{A}_0$, It is called ***the mixture representation*** of the tangent vector $X \in T_p\mathcal{P}$ [Amari and Nagaoka, 2007]. Thus the tangent space under the mixture embedding is

     $$T_p^{(m)}\mathcal{P} := \left\{X^{(m)} : X \in T_p\mathcal{P}\right\} = \mathcal{A}_0 = \left\{A \in \mathbb{R}^{\mathcal{X}} : \int_{\mathcal{X}} A(x) d\mu = 0\right\}. \tag{2}$$

     Note that ***the basis tangent vector*** under this embedding is still

     $$\left(\frac{\partial}{\partial \xi^i}\Big|_p\right)^{(m)} = \frac{\partial}{\partial \xi^i}\Big|_{\iota(p)} = \frac{\partial}{\partial \xi^i}\Big|_p. \tag{3}$$

  2. The ***log-likelihood function*** $\ell : \mathcal{P} \to \ell(\mathcal{P})\mathbb{R}^{\mathcal{X}}$ is also a ***smooth embedding*** as shown above. It is called ***the exponential embedding***. Note that $\ell(\mathcal{P}) = \{\log(p) : p \in \mathcal{P}\}$. A tangent vector $X \in T_p\mathcal{P}$ under this embedding is then represented by the result of mapping $p \mapsto \log(p)$, which is denoted as $X^{(e)}$ and call ***the exponential representation*** [Amari and Nagaoka, 2007]. Note that

     $$X^{(e)} = d\ell_p(X) = X\ell = p(x;\xi)^{-1} X^{(m)}(x).$$

     Thus ***the basis tangent vector*** under *the exponential embedding*

     $$\left(\frac{\partial}{\partial \xi^i}\Big|_p\right)^{(e)} = \frac{\partial}{\partial \xi^i}\Big|_{\ell(p)} = \frac{\partial \ell}{\partial \xi^i}\Big|_p. \tag{4}$$

     Denote *the* ***tangent space*** *under this embedding* as $T_p^{(e)}\mathcal{P}$. We can verify that

     $$T_p^{(e)}\mathcal{P} = \left\{X^{(e)} : X \in T_p\mathcal{P}\right\} = \left\{A \in \mathbb{R}^{\mathcal{X}} : \int_{\mathcal{X}} A(x)p(x) d\mu = \mathbb{E}_p[A] = 0\right\}. \tag{5}$$

- **Remark** $\mathcal{P}(\mathcal{X})$ is $|\mathcal{X}|$-*dimensional submanifold* if the domain $\mathcal{X}$ is finite. Otherwise, $\mathcal{P}(\mathcal{X})$ is ***not seen as a manifold itself***. However, the above discussion is still valid if we restrict our attention to the n-***dimensional statistical manifold*** $S \subseteq \mathcal{P}(\mathcal{X})$. We just need to replace $\mathcal{P}$ with S above. Without noticing, we will focus on S instead of $\mathcal{P}$ for our discussion.

1.3 Fisher Information Metrics

- **Remark** For probabilty models, the ambient space $\mathbb{R}^{\mathcal{X}}$ denotes *the set of all* ***random variables*** on $\mathcal{X}$. Moreover, it has a natural definition of ***inner product*** as

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

The inner product induced by the embedding map ι in $T_p^{(m)}S$ is formulated as

$$\langle d\iota_p(X)\,,\,d\iota_p(Y)\rangle := \left\langle X^{(m)}\,,\,Y^{(m)}\right\rangle := \int_{\mathcal{X}} X^{(m)}(s)\,Y^{(m)}(s)d\mu(s) \tag{6}$$

Similarly, *the* ***inner product*** induced by the embedding map ℓ in $T_p^{(e)}S$ becomes

$$\langle d\ell_p(X)\,,\,d\ell_p(Y)\rangle := \left\langle X^{(e)}\,,\,Y^{(e)}\right\rangle_p := \mathbb{E}_p\left[X^{(e)}\,Y^{(e)}\right] = \int \left[X^{(e)}(s)\,Y^{(e)}(s)\right]p(s)d\mu(s) \tag{7}$$

where the additional $p(s)$ comes from *the* ***Jacobian*** *for the* ***inverse*** *of the log-likelihood.*

- By definition, ***the Riemannian metric*** on S under ***the exponential representation*** is defined as

$$\begin{aligned}\hat{g}_{i,j} &:= \left\langle \left(\frac{\partial}{\partial \xi^i}\Big|_p\right)^{(e)}\,,\,\left(\frac{\partial}{\partial \xi^j}\Big|_p\right)^{(e)}\right\rangle_p \\ &= \mathbb{E}_p\left[\frac{\partial}{\partial \xi^i}\ell(p)\,\frac{\partial}{\partial \xi^i}\ell(p)\right] := \text{Fisher information } g_{i,j}.\end{aligned}$$

$g_{i,j}$ is called ***the Fisher metric*** *or* ***the Information metric*** [Amari and Nagaoka, 2007]. It is seen that ***the Fisher metric is a Riemannian metric on*** S.

Thus, S is a n***-dimensional Riemannian submanifold***.

1.4 α-Connections

- [Amari and Nagaoka, 2007] proposed ***the*** α***-connections*** $\nabla^{(\alpha)}$ as ***a family of affine connections*** on the tangent bundle TS, for $\alpha \in [-1, 1]$. The ***coefficient of the*** α***-connection*** under ***the Fisher metric*** is formulated as

$$\Gamma_{i,j;k}^{(\alpha)} = \mathbb{E}_\xi\left[\left(\frac{\partial}{\partial \xi^i}\frac{\partial}{\partial \xi^j}\ell_\xi + \frac{1-\alpha}{2}\frac{\partial}{\partial \xi^i}\ell_\xi\,\frac{\partial}{\partial \xi^j}\ell_\xi\right)\left(\frac{\partial}{\partial \xi^k}\ell_\xi\right)\right] \tag{8}$$

where

$$\Gamma_{i,j;k}^{(\alpha)} := \left\langle \nabla_{\partial_i}^{(\alpha)}\partial_j\,,\,\partial_k\right\rangle,$$

where $g = \langle \cdot\,, \cdot \rangle_p$ is ***the Fisher metric***.

We see that for $\alpha = 0$, the coefficient for 0-connection

$$\Gamma^{(0)}_{i,j;k} = \mathbb{E}_{\xi}\left[\left(\frac{\partial}{\partial \xi^i}\frac{\partial}{\partial \xi^j}\ell_{\xi}\right)\left(\frac{\partial}{\partial \xi^k}\ell_{\xi}\right)\right] + \frac{1}{2}\mathbb{E}_{\xi}\left[\left(\frac{\partial}{\partial \xi^i}\ell_{\xi}\,\frac{\partial}{\partial \xi^j}\ell_{\xi}\right)\left(\frac{\partial}{\partial \xi^k}\ell_{\xi}\right)\right]$$

Thus

$$\partial_k\, g_{i,j} = \partial_k \mathbb{E}_p\left[(\partial_i \ell)(\partial_j \ell)\right] = \mathbb{E}_p\left[(\partial_k\partial_i \ell)(\partial_j \ell)\right] + \mathbb{E}_p\left[(\partial_i \ell)(\partial_k\partial_j \ell)\right] + \mathbb{E}_p\left[(\partial_i \ell)(\partial_j \ell)(\partial_k \ell)\right]$$

The last terms from ∂_k acting on the expectation function $\mathbb{E}_p\left[\cdot\right]$. Thus

$$\begin{aligned}\partial_k\, g_{i,j} &= \mathbb{E}_p\left[(\partial_k\partial_i \ell)(\partial_j \ell)\right] + \mathbb{E}_p\left[(\partial_i \ell)(\partial_k\partial_j \ell)\right] + \mathbb{E}_p\left[(\partial_i \ell)(\partial_j \ell)(\partial_k \ell)\right] \\ &= \Gamma^{(0)}_{k,i;j} + \Gamma^{(0)}_{k,j;i}\end{aligned}$$

- Note that for Levi-Civita connection (i.e. connection that is both metric and symmetric), the relationship between the Riemannian metric and the coefficients of connection under the metric is

$$\frac{\partial}{\partial \xi^k} g_{i,j} = \Gamma_{k,i;j} + \Gamma_{k,j;i}$$
$$\text{where } \Gamma_{i,j;k} := \left\langle \nabla_{\partial_i}\partial_j\,,\, \partial_k \right\rangle,$$

  Thus ***the*** α***-connection is the Levi-Civita connection with respect to the Fisher metric if and only if*** $\alpha = 0$.

- ***The family of*** α***-connections*** *forms* ***an affine space*** *itself*, i.e.

$$\begin{aligned}\nabla^{(\alpha)} &= \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)} \\ &= (1-\alpha)\nabla^{(0)} + \alpha\,\nabla^{(1)}\end{aligned}$$

  Also since $\nabla^{(0)}$ is the Levi-Civita connection (Riemannian connections) on S and also that this connection is unique, we see that $\nabla^{(\alpha)}$ ***is not the Levi-Civita connection for all*** $\alpha \neq 0$. In fact, $\nabla^{(\alpha)}$ ***is not a metric connection for all*** $\alpha \neq 0$

- There are two special α-connections:

  1. When $\alpha = -1$, the $\nabla^{(-1)}$ is called <u>***the mixture connection***</u> and is denoted as $\nabla^{(m)}$.

     <u>***The mixture family***</u> of distributions is seen as a <u>m***-affine subspaces***</u> since it is considered ***flat*** (i.e. $\Gamma^{(-1)}_{i,j;k} = 0$) under ***the mixture connections*** $\nabla^{(m)}$.

$$p(x;\xi) = \sum_{i=1}^{n} \xi^i\,\phi_i(x) + C(x) \tag{9}$$

  2. When $\alpha = 1$, the $\nabla^{(1)}$ is called <u>***the exponential connection***</u> and is denoted as $\nabla^{(e)}$.

     <u>***The exponential family***</u> of distributions is seen as an <u>e***-affine subspaces***</u> since it is considered ***flat*** (i.e. $\Gamma^{(1)}_{i,j;k} = 0$) under ***the exponential connections*** $\nabla^{(e)}$.

$$p(x;\xi) = \exp\left\{\sum_{i=1}^{n} \xi^i\,\phi_i(x) - A(\xi)\right\} C(x) \tag{10}$$

1.5 Dual Connections

- **Definition** Let (S, g) be a Riemannian manifold and ∇ and ∇^* are two connections on TS. If for all vector fields $X, Y, Z \in \mathfrak{X}(S)$,
$$Z \langle X\,,\,Y \rangle = \langle \nabla_Z X\,,\,Y \rangle + \langle X\,,\,\nabla^*_Z(Y) \rangle \tag{11}$$
holds, then we say that ∇ and ∇^* are ***duals*** to each other with respect to the Riemannian metric g. We call one either ***the dual connection*** or ***the conjugate connection***.

  We call the triple (g, ∇, ∇^*) ***a dualistic structure*** on S.

- We see that the coefficients $\Gamma_{i,j;k}$ and $\Gamma^*_{i,j;k}$ for ∇ and ∇^* have the relationship:
$$\partial_k\, g_{i,j} = \Gamma_{k,i;j} + \Gamma^*_{k,j;i}$$

- Similarly, define ***the covariant derivative*** of vector field *along curve* with respect to ∇ and its dual connection ∇^* as D_t and D^*_t, then
$$\frac{d}{dt} \langle X(t)\,,\,Y(t) \rangle = \langle D_t X(t)\,,\,Y(t) \rangle + \langle X(t)\,,\,D^*_t Y(t) \rangle$$

- For ***the parallel transport map*** Π_γ and Π^*_γ along the curve γ (from t_0 to t_1) with respect to ∇ and its dual ∇^*, we have
$$\left\langle \Pi_\gamma(X)\,,\,\Pi^*_\gamma(Y) \right\rangle_q = \langle X\,,\,Y \rangle_p\,.$$
where $p = \gamma(t_0)$ and $q = \gamma(t_1)$. This is a generalization of "***the invariance of the inner product under parallel translation with respect to metric connections***."

- Also ***the Riemannian curvature tensor*** with respect to ∇ and its dual ∇^* has the relationship
$$\langle R(X,Y)Z\,,\,W \rangle = -\, \langle R^*(X,Y)Z\,,\,W \rangle\,.$$
Thus $Rm = -Rm^*$, so $R = 0 \Leftrightarrow R^* = 0$.

  In other word, *a Riemannian manifold S with dualistic structure (g, ∇, ∇^*) is **flat** in ∇ **if and only** if it is **flat** in its **dual connection** ∇^*.*

- It is clear that if ∇ is ***a metric connection***, then $\nabla = \nabla^*$. The concept of dual connections (∇, ∇^*) is a generalization of the metric connection. Moreover, $\frac{1}{2}(\nabla + \nabla^*)$ becomes *a metric connection.*

- Within α-connections, $(\nabla^{(-\alpha)}, \nabla^{(\alpha)})$ are ***duals*** to each other with respect to *the Fisher metric.* Specifically, $(\nabla^{(m)}, \nabla^{(e)})$, i.e. ***the mixture connection** and **the exponential connection** are **duals** to each other.*

  From above statement, we see that
$$S \text{ is } (\alpha)\text{-flat } \Leftrightarrow S \text{ is } (-\alpha)\text{-flat} \tag{12}$$
That (S, g, ∇, ∇^*) is called ***a dually flat space***

- **Remark** *The exponential family* is *a dually flat space* since it is both 1-***flat*** and (-1)-**flat**. The former corresponds to ***the natural parameterization*** (ξ^i) which is $\nabla^{(e)}$***-affine*** and the latter corresponds to ***the mean parameterization*** (μ_i) which is $\nabla^{(m)}$***-affine***. It has ***two mutually dual coordinate systems***.

1.6 Embedding Associated with α-Connections

- We have seen the mixture embeddings and the exponential embeddings and their associated definition of inner product. In this section, we see the embedding associated with α-connections, which includes both embeddings above as its special cases.

- Consider the extension of $\mathcal{P}(\mathcal{X})$ by dropping the normalization constraint:

$$\widetilde{\mathcal{P}} := \left\{ p : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} p(x) d\mu(x) < \infty,\ p(x) > 0\, (\forall x \in \mathcal{X}). \right\}$$

- **Definition** For each $\alpha \in \mathbb{R}$, define the following ***α-likelihood function***:

$$L^{(\alpha)}(x) := \begin{cases} \frac{2}{(1-\alpha)} x^{\frac{(1-\alpha)}{2}} & \text{if } \alpha \neq 1, \\ \log(x), & \text{if } \alpha = 1. \end{cases} \quad . \tag{13}$$

$$\ell^{(\alpha)}(x;\xi) := L^{(\alpha)}(p(x;\xi)) \tag{14}$$

  Note in particular that $\ell^{(1)}(x;\xi) = \ell(x;\xi)$ and that $\ell^{(-1)}(x;\xi) = p(x;\xi)$.

- **Definition** For a tangent vector $X \in T_p(S)$, we call

$$X^{(\alpha)}(x) := X\, \ell^{(\alpha)}(x;\xi) \tag{15}$$

  as a function of x ***the α-representation of X***. The *e-representation* and *m-representation* correspond to $\alpha = 1$ and $\alpha = -1$.

- **Definition** With the α-representation, we have ***the induced inner product*** by *the α-likelihood function* $\ell^{(\alpha)}$:

$$\langle X\,,\, Y\rangle_g^{(\alpha)} := \left\langle X^{(\alpha)}\,,\, Y^{(-\alpha)} \right\rangle = \int_{\mathcal{X}} \left(X\ell^{(\alpha)}(x;\xi) \right) \left(Y\ell^{(-\alpha)}(x;\xi) \right) d\mu(x) \tag{16}$$

- We can compute the first and second order partial derivatives of the α-likelihood as

$$\frac{\partial}{\partial \xi^i} \ell^{(\alpha)} = p^{(1-\alpha)/2} \frac{\partial}{\partial \xi^i} \ell \tag{17}$$

$$\frac{\partial}{\partial \xi^i} \frac{\partial}{\partial \xi^j} \ell^{(\alpha)} = p^{(1-\alpha)/2} \left(\frac{\partial}{\partial \xi^i} \frac{\partial}{\partial \xi^j} \ell + \frac{1-\alpha}{2} \frac{\partial}{\partial \xi^i} \ell\, \frac{\partial}{\partial \xi^j} \ell \right) \tag{18}$$

- We may rewrite ***the Fisher metric*** and ***the Christoffel symbol*** of α-connection as

$$g_{i,j}(\xi) = \int_{\mathcal{X}} \frac{\partial}{\partial \xi^i} \ell^{(\alpha)}(x;\xi) \frac{\partial}{\partial \xi^j} \ell^{(-\alpha)}(x;\xi) d\mu(x) \tag{19}$$

$$\Gamma_{i,j;k}^{(\alpha)} = \int_{\mathcal{X}} \frac{\partial}{\partial \xi^i} \frac{\partial}{\partial \xi^j} \ell^{(\alpha)}(x;\xi) \frac{\partial}{\partial \xi^k} \ell^{(-\alpha)}(x;\xi) d\mu(x) \tag{20}$$

- **Remark** From (20), we see that the α-likelihood defines an ***embedding*** $\ell^{(\alpha)} : \widetilde{\mathcal{P}} \to \mathbb{R}^{\mathcal{X}}$. And the α-connection on $S \subset \widetilde{\mathcal{P}}$ is ***the induced connection*** from ***the affine structure*** of the space $\mathbb{R}^{\mathcal{X}}$ of functions on $\mathcal{X}$ through the embedding $\ell^{(\alpha)}$.

- **Remark** For probability distribution, since $\int \partial_i p = 0$, we have

$$\int p(x;\xi)^{\frac{1+\alpha}{2}} \partial_i \ell^{(\alpha)}(x;\xi)dx = 0$$
$$\frac{1+\alpha}{2} g_{i,j}(\xi) = -\int_{\mathcal{X}} p(x;\xi)^{\frac{1+\alpha}{2}} \partial_i \partial_j \ell^{\alpha}(x;\xi)dx$$

- **Definition** For given α, if under some coordinate system (ξ^i) of S,

$$\frac{\partial}{\partial \xi^i}\frac{\partial}{\partial \xi^j}\ell^{(\alpha)}(x;\xi) = 0, \tag{21}$$

then it is seen from (20) that $\Gamma^{(\alpha)}_{i,j;k} = 0$. Thus S is ***α-flat***.

We call (ξ^i) an ***α-affine coordinate system***, and such an S an ***α-affine manifold***.

- **Remark** Thus we can say that:
  1. A *mixture family* is a (-1)***-affine manifold***,
  2. An *exponential family* is ***not*** a 1***-affine manifold***.
  3. For ***finite*** $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ is an α***-affine manifold*** for ***every*** $\alpha \in \mathbb{R}$
- **Definition** We can also extend $S \subset \mathcal{P}$ by varying the sum of mass:

$$\widetilde{S} := \{\tau p_{\xi} : \xi \in \Xi, \tau > 0\} \subset \widetilde{\mathcal{P}}$$

We see that $\widetilde{S}$ is *a manifold of dimension dim* $S+1$ *which contains* S. We call $\widetilde{S}$ a ***denormalization*** of S. The adopted coordinate system of $\widetilde{S}$ is $(\xi^1, \ldots, \xi^n, \tau)$. We can extend our definition of ℓ^{α} as $\widetilde{\ell}^{(\alpha)} := \ell^{(\alpha)}(x;\xi,\tau) := L^{(\alpha)}(\tau\, p(x;\xi))$. We then extend compuatation of derivatives with τ added.

- The following is the relation between $\widetilde{S}$ and S:

Proposition 1.1 *S is (-1)-autoparallel in $\widetilde{S}$.*

- **Proposition 1.2** *Let M be a submanifold of S and $\widetilde{M}$ be its denormalization. For every $\alpha \in \mathbb{R}$, the following conditions (1) and (2) are* ***equivalent****.*
  1. *M is α-autoparallel in S.*
  2. *$\widetilde{M}$ is α-autoparallel in $\widetilde{S}$.*
- **Definition** We call a statistical model $S = \{p(x;\xi)\}$ whose **denormalization** $\widetilde{S}$ is *an α-affine manifold* ***an α-family***.
- **Remark** We have the following results
  1. ***An exponential family*** is a 1***-family***; and conversely, ***every*** 1***-family is exponential family***.
  2. ***A mixture family*** is a (-1)***-family***; and conversely, ***every*** (-1)***-family is mixture family***.
  3. For ***finite*** $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ is an α**-family** for ***every*** $\alpha \in \mathbb{R}$

2 Differential Geometry vs. Information Geometry

Table 1: Comparison between differential geometry and information geometry

base	***smooth manifold*** M	***statistical manifold*** $S \subseteq \mathcal{P}$.
embeddings	$M \subseteq \mathcal{R}$ with smooth embedding $\iota : M \hookrightarrow \mathcal{R}$	$\mathcal{P} \subset \mathbb{R}^{\mathcal{X}}$ with a ***smooth embedding*** as ***the log-likelihood*** $\ell : \mathcal{P} \to \mathbb{R}^{\mathcal{X}} : \ell(p) = \log(p)$.
element	a point $p \in M$	a ***parametric model*** $p(x;\xi) \in S,\ \xi \in \Xi$
coordinate map	$\varphi(p) = (x^1, \ldots, x^n)$	$\varphi(p_\xi) = (\xi^1, \ldots, \xi^n)$
smooth map	$f : M \to \mathbb{R}$	e.g. $\kappa : \mathcal{P} \to \mathbb{R}$, $\kappa(p) := \mathbb{E}_p\left[f\right]$ for some ***random variable*** $f \in \mathbb{R}^{\mathcal{X}}$.
space of smooth maps	$\mathcal{C}^\infty(M)$	$\mathcal{C}^\infty(\mathcal{S}) \subseteq \mathcal{C}^\infty(\mathcal{P})$
tangent vector at p	a ***derivation operator*** at p: $v : \mathcal{C}^\infty(M) \to \mathbb{R}$	a ***derivation operator*** at p: $X : \mathcal{C}^\infty(S) \to \mathbb{R}$
tangent space at p	**tangent space** T_pM	**tangent space** $T_pS \subseteq T_p\mathcal{P}$
embedding representation of $T_p\mathcal{P}$	$\{\widetilde{v} := d\iota_p(v) : v \in T_pM\} \subseteq T_p\mathcal{R}$	***exponential-representation*** $T_p^{(e)}\mathcal{P} = \left\{X^{(e)} := X\ell :\ X \in T_p\mathcal{P}\right\}$ $= \left\{f \in \mathbb{R}^{\mathcal{X}} : \mathbb{E}_p\left[f\right] = 0\right\} \subseteq T_p\mathbb{R}^{\mathcal{X}} \simeq \mathbb{R}^{\mathcal{X}}$
dim T_pM	n	$n = \dim\, T_pS < \dim\, T_p\mathcal{P} = +\infty$
basis of tangent space	$\left(\frac{\partial}{\partial x^1}\Big\vert_p, \ldots, \frac{\partial}{\partial x^n}\Big\vert_p\right)$	$\left(\frac{\partial}{\partial \xi^1}\Big\vert_p, \ldots, \frac{\partial}{\partial \xi^n}\Big\vert_p\right)$
basis of embedding tangent space	$\left(\frac{\partial}{\partial x^1}\Big\vert_{\iota(p)}, \ldots, \frac{\partial}{\partial x^n}\Big\vert_{\iota(p)}\right)$	$\left(\frac{\partial}{\partial \xi^1}\Big\vert_{\ell(p)}, \ldots, \frac{\partial}{\partial \xi^n}\Big\vert_{\ell(p)}\right)$
inner product on tangent space	$\left\langle v\,,\, w\right\rangle_g := g(v,w)$	The ***cross correlation*** $\left\langle X\,,\, Y\right\rangle_p := \mathbb{E}_p\left[(X\ell)\,(Y\ell)\right]$
Riemanian metric	The ***Riemanian metric*** $g = g_{i,j}\, dx^i dx^j$ where $g_{i,j} = \left\langle \frac{\partial}{\partial x^i}\,,\, \frac{\partial}{\partial x^j}\right\rangle_g$	The ***Fisher information metric*** $g = g_{i,j}\, d\xi^i\, d\xi^j$ where $g_{i,j} = \mathbb{E}_p\left[\partial_i \ell\, \partial_j \ell\right] := \left\langle \partial_i\,,\, \partial_j\right\rangle_p$, and $\partial_i \equiv \frac{\partial}{\partial \xi^i}$
Riemanian matrix	$(g_{i,j}) \in \mathcal{S}_+^n$	The ***Fisher information matrix*** I where $(g_{i,j}(\xi)) \in \mathcal{S}_+^n$
connections / Christoffel symbols	***Riemannian connection*** $\Gamma_{i,j;k} := \left\langle \nabla_{\partial_i}\partial_j\,,\, \partial_k\right\rangle_g$ $= \frac{1}{2}\left(\partial_i g_{j,k} + \partial_j g_{k,i} - \partial_k g_{i,j}\right)$ $\Rightarrow \partial_k g_{i,j} = \Gamma_{k,i;j} + \Gamma_{k,j;i}$	***α-connection*** $\Gamma_{i,j;k}^{(\alpha)} := \langle \nabla^{(\alpha)}_{(\partial_i)^{(e)}}(\partial_j)^{(e)}, (\partial_k)^{(e)}\rangle_p$ $= \mathbb{E}_\xi\left[\left(\partial_i\partial_j\ell_\xi + \frac{1-\alpha}{2}\partial_i\ell_\xi\partial_j\ell_\xi\right)\partial_k\ell_\xi\right]$ $\Rightarrow \partial_k g_{i,j} = \Gamma_{k,i;j}^{(0)} + \Gamma_{k,j;i}^{(0)}$

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.