Information geometry:

L^p-Fisher-Rao metrics and the α-connections

Alice Le Brigant
SAMM, Université Paris 1

Joint work with Martin Bauer, Yuxiu Lu and Cy Maor

Infinite-dimensional Geometry: Theory and Applications

ESI, 4 February 2025

Table of contents

Table of contents

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \ldots, x_n$

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \ldots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \ldots, x_n) \mapsto T(x_1, \ldots, x_n)$, how much information on θ do I loose ?

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \ldots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \ldots, x_n) \mapsto T(x_1, \ldots, x_n)$, how much information on θ do I loose ?

Answers :

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta ; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \ldots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \ldots, x_n) \mapsto T(x_1, \ldots, x_n)$, how much information on θ do I loose ?

Answers :

1. Fisher information :
$$I_\theta(X) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right) = -\mathbb{E}_\theta \left(\mathrm{Hess}_\theta \log f_\theta(X) \right).$$

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \dots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \dots, x_n) \mapsto T(x_1, \dots, x_n)$, how much information on θ do I loose ?

Answers :

1. Fisher information :
$$I_\theta(X) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right) = -\mathbb{E}_\theta \left(\mathrm{Hess}_\theta \log f_\theta(X) \right).$$

$\rightarrow$ if f_θ does not depend on θ, $I_\theta(X) = 0$

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \dots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \dots, x_n) \mapsto T(x_1, \dots, x_n)$, how much information on θ do I loose ?

Answers :

1. Fisher information :

$$I_\theta(X) = \mathbb{E}_\theta\left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top\right) = -\mathbb{E}_\theta\left(\mathrm{Hess}_\theta \log f_\theta(X)\right).$$

$\rightarrow$ if f_θ does not depend on θ, $I_\theta(X) = 0$
$\rightarrow$ if $X, Y \sim f_\theta$ are independant, then $I_\theta((X,Y)) = I_\theta(X) + I_\theta(Y)$.

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \dots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \dots, x_n) \mapsto T(x_1, \dots, x_n)$, how much information on θ do I loose ?

Answers :

1. Fisher information :
$$I_\theta(X) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right) = -\mathbb{E}_\theta \left(\mathrm{Hess}_\theta \log f_\theta(X) \right).$$

   $\rightarrow$ if f_θ does not depend on θ, $I_\theta(X) = 0$
   $\rightarrow$ if $X, Y \sim f_\theta$ are independant, then $I_\theta((X,Y)) = I_\theta(X) + I_\theta(Y)$.

2. We have $I_\theta(T(X)) \leq I_\theta(X) \rightarrow$ Information lost : $I_\theta(X) - I_\theta(T(X))$

Parametric statistics

Consider a parametric family of probability densities

$$P_\Theta = \{f_\theta; \theta \in \Theta\}, \quad \text{with} \quad \Theta \subset \mathbb{R}^d \text{ open.}$$

Typical problem (parameter estimation) :

find $\theta \in \Theta$ such that f_θ best "fits" observations $x_1, \dots, x_n$

Natural questions :

1. how much *information* on θ is contained in a sample $x \sim f_\theta$?
2. if I transform my observations $(x_1, \dots, x_n) \mapsto T(x_1, \dots, x_n)$, how much information on θ do I loose ?

Answers :

1. Fisher information :
$$I_\theta(X) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right) = -\mathbb{E}_\theta \left(\mathrm{Hess}_\theta \log f_\theta(X) \right).$$

   $\rightarrow$ if f_θ does not depend on θ, $I_\theta(X) = 0$
   $\rightarrow$ if $X, Y \sim f_\theta$ are independant, then $I_\theta((X,Y)) = I_\theta(X) + I_\theta(Y)$.

2. We have $I_\theta(T(X)) \leq I_\theta(X) \rightarrow$ Information lost : $I_\theta(X) - I_\theta(T(X))$
   equality iff T is a sufficient statistic : $P_\theta(X|T(X))$ is independant of θ.

The parametric Fisher-Rao metric

The Fisher information is a $d \times d$ symmetric positive semi-definite matrix

$$I(\theta) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right)$$

The parametric Fisher-Rao metric

The Fisher information is a $d \times d$ symmetric positive semi-definite matrix

$$I(\theta) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right)$$

When definite, it defines a Riemannian metric on the parameter space $\Theta \subset \Theta$ (Rao 1945, Jeffreys 1946), called the Fisher-Rao metric

$$\langle u, v \rangle_\theta = u^\top I(\theta) v, \quad u, v \in T_\theta \Theta \simeq \mathbb{R}^d$$

The parametric Fisher-Rao metric

The Fisher information is a $d \times d$ symmetric positive semi-definite matrix

$$I(\theta) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right)$$

When definite, it defines a Riemannian metric on the parameter space $\Theta \subset \Theta$ (Rao 1945, Jeffreys 1946), called the Fisher-Rao metric

$$\langle u, v \rangle_\theta = u^\top I(\theta) v, \quad u, v \in T_\theta \Theta \simeq \mathbb{R}^d$$

- The Fisher-Rao metric is invariant w.r.t. transformation of the statistical model by a sufficient statistic : if T is a sufficient statistic,

$$T_* P_\Theta = \{T_* f_\theta; \theta \in \Theta\} \quad \text{is isometric to} \quad P_\Theta = \{f_\theta; \theta \in \Theta\}$$

for the Fisher-Rao metric.

The parametric Fisher-Rao metric

The Fisher information is a $d \times d$ symmetric positive semi-definite matrix

$$I(\theta) = \mathbb{E}_\theta \left(\nabla_\theta \log f_\theta(X) \cdot \nabla_\theta \log f_\theta(X)^\top \right)$$

When definite, it defines a Riemannian metric on the parameter space $\Theta \subset \Theta$ (Rao 1945, Jeffreys 1946), called the Fisher-Rao metric

$$\langle u, v \rangle_\theta = u^\top I(\theta) v, \quad u, v \in T_\theta \Theta \simeq \mathbb{R}^d$$

- The Fisher-Rao metric is invariant w.r.t. transformation of the statistical model by a sufficient statistic : if T is a sufficient statistic,

  $$T_* P_\Theta = \{T_* f_\theta ; \theta \in \Theta\} \quad \text{is isometric to} \quad P_\Theta = \{f_\theta ; \theta \in \Theta\}$$

  for the Fisher-Rao metric.

- Cencov 1972 : it is the *only* Riemannian metric with this property (up to scalar multiplication).

The parametric Fisher-Rao metric

- The Fisher-Rao metric is invariant to diffeomorphic parametrization change $\varphi : \theta \mapsto \tilde{\theta} \rightarrow$ The geometric structure does not depend on the parametrization.

The parametric Fisher-Rao metric

- The Fisher-Rao metric is invariant to diffeomorphic parametrization change $\varphi : \theta \mapsto \tilde{\theta} \rightarrow$ The geometric structure does not depend on the parametrization.

- It gives a metric approximation of the Kullback-Leibler divergence $KL(p|q) = E_p \log(p/q)$

$$KL(f_\theta | f_{\theta + d\theta}) = \frac{1}{2} d\theta^\top I(\theta) d\theta + O(|d\theta|^3) \approx \frac{1}{2} \|d\theta\|_\theta^2$$

The parametric Fisher-Rao metric

- The Fisher-Rao metric is invariant to diffeomorphic parametrization change $\varphi : \theta \mapsto \tilde{\theta} \rightarrow$ The geometric structure does not depend on the parametrization.

- It gives a metric approximation of the Kullback-Leibler divergence $KL(p|q) = E_p \log(p/q)$

$$KL(f_\theta | f_{\theta + d\theta}) = \frac{1}{2} d\theta^\top I(\theta) d\theta + O(|d\theta|^3) \approx \frac{1}{2} \|d\theta\|_\theta^2$$

- "Displacement is cheaper in zones of lower information."

The parametric Fisher-Rao metric

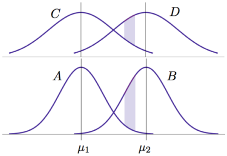
- The Fisher-Rao metric is invariant to diffeomorphic parametrization change $\varphi : \theta \mapsto \tilde{\theta}$ $\rightarrow$ The geometric structure does not depend on the parametrization.

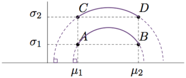
- It gives a metric approximation of the Kullback-Leibler divergence $KL(p|q) = E_p \log(p/q)$

$$KL(f_\theta | f_{\theta+d\theta}) = \frac{1}{2} d\theta^\top I(\theta) d\theta + O(|d\theta|^3) \approx \frac{1}{2} \|d\theta\|_\theta^2$$

- "Displacement is cheaper in zones of lower information."

  Example : normal distributions (Atkinson & Mitchell 81, Skovgaard 84)





$$I((\mu,\sigma)) = \frac{1}{\sigma^2} I_2$$

Figure from
Costa, Santos, Strapasson 2015

Fisher-Rao geometry amounts to hyperbolic geometry.

The parametric Fisher-Rao metric

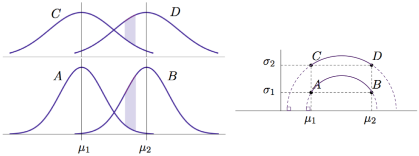
- The Fisher-Rao metric is invariant to diffeomorphic parametrization change $\varphi : \theta \mapsto \tilde{\theta}$ → The geometric structure does not depend on the parametrization.

- It gives a metric approximation of the Kullback-Leibler divergence $KL(p|q) = E_p \log(p/q)$

$$KL(f_\theta | f_{\theta+d\theta}) = \frac{1}{2} d\theta^\top I(\theta) d\theta + O(|d\theta|^3) \approx \frac{1}{2} \|d\theta\|_\theta^2$$

- "Displacement is cheaper in zones of lower information."

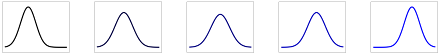
  Example : normal distributions (Atkinson & Mitchell 81, Skovgaard 84)



$$I((\mu, \sigma)) = \frac{1}{\sigma^2} I_2$$

Figure from Costa, Santos, Strapasson 2015

Fisher-Rao geometry amounts to hyperbolic geometry.



Non parametric Fisher-Rao metric

Let M be a closed manifold.

The spaces of all smooth (probability) densities on M w.r.t. the volume measure $\mathrm{d}x$

$$\mathrm{Dens}_+(M) = \{\rho \mathrm{d}x;\ \rho > 0\}, \qquad \mathrm{Prob}(M) = \{\rho \mathrm{d}x;\ \rho > 0, \int_M \rho\, \mathrm{d}x = 1\}$$

are Fréchet manifolds with tangent spaces

$$T_\rho \mathrm{Dens}_+(M) = \{a\, \mathrm{d}x\}, \qquad T_\rho \mathrm{Prob}(M) = \{a\, \mathrm{d}x;\ \int_M a\, \mathrm{d}x = 0\}.$$

Friedrich, 1991 : the Fisher-Rao metric on $\mathrm{Dens}_+(M)$ or $\mathrm{Prob}(M)$ is

$$\langle a, b\rangle_\rho = \int_M \frac{a}{\rho} \frac{b}{\rho} \rho \mathrm{d}x.$$

Link to the parametric setting

A parametric statistical model $P_\Theta = \{f_\theta \mathrm{d}x;\ \theta \in \Theta\}$ defines a submanifold of Prob.

The non parametric Fisher-Rao metric restricted to such a submanifold P_Θ is the parametric Fisher-Rao metric defined by the Fisher information.

The tangent space to P_Θ at $\rho = f_\theta$ is spanned by $\{e_i = \partial f_\theta / \partial \theta_i\}_{i=1}^d$ and

$$\langle e_i, e_j \rangle_\rho = \int \frac{\frac{\partial f_\theta}{\partial \theta_i}}{f_\theta} \frac{\frac{\partial f_\theta}{\partial \theta_j}}{f_\theta} f_\theta \mathrm{d}x = E_\theta \left(\frac{\partial}{\partial \theta_i} \log f_\theta(X) \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right) = I(\theta)_{ij}$$

Link to the parametric setting

A parametric statistical model $P_\Theta = \{f_\theta \mathrm{d}x;\ \theta \in \Theta\}$ defines a submanifold of Prob.

The non parametric Fisher-Rao metric restricted to such a submanifold P_Θ is the parametric Fisher-Rao metric defined by the Fisher information.

The tangent space to P_Θ at $\rho = f_\theta$ is spanned by $\{e_i = \partial f_\theta / \partial \theta_i\}_{i=1}^d$ and

$$\langle e_i, e_j \rangle_\rho = \int \frac{\frac{\partial f_\theta}{\partial \theta_i}}{f_\theta} \frac{\frac{\partial f_\theta}{\partial \theta_j}}{f_\theta} f_\theta \mathrm{d}x = E_\theta \left(\frac{\partial}{\partial \theta_i} \log f_\theta(X) \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right) = I(\theta)_{ij}$$

If M is a compact manifold without boundary of dimension greater than 1, it is the only Riemannian metric, up to a multiplicative factor, invariant under the action of diffeomorphisms of M (Bauer, Bruveris, Michor 16, Ay, Jost, Lê, Schwachhöfer 15).

Link to the parametric setting

A parametric statistical model $P_\Theta = \{f_\theta \mathrm{d}x;\ \theta \in \Theta\}$ defines a submanifold of Prob.

The non parametric Fisher-Rao metric restricted to such a submanifold P_Θ is the parametric Fisher-Rao metric defined by the Fisher information.

The tangent space to P_Θ at $\rho = f_\theta$ is spanned by $\{e_i = \partial f_\theta / \partial \theta_i\}_{i=1}^{d}$ and

$$\langle e_i, e_j \rangle_\rho = \int \frac{\frac{\partial f_\theta}{\partial \theta_i}}{f_\theta} \frac{\frac{\partial f_\theta}{\partial \theta_j}}{f_\theta} f_\theta \mathrm{d}x = E_\theta \left(\frac{\partial}{\partial \theta_i} \log f_\theta(X) \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right) = I(\theta)_{ij}$$

If M is a compact manifold without boundary of dimension greater than 1, it is the only Riemannian metric, up to a multiplicative factor, invariant under the action of diffeomorphisms of M (Bauer, Bruveris, Michor 16, Ay, Jost, Lê, Schwachhöfer 15).

Remark : Parametric and non parametric Fisher-Rao define very different geometries !

Square-root transform

Let $\phi : \mathrm{Dens}_+(M) \to C^\infty(M)$,

$$\phi(\rho) = \sqrt{\rho}$$

Square-root transform

Let $\phi : \mathrm{Dens}_+(M) \to C^\infty(M)$,

$$\phi(\rho) = \sqrt{\rho}$$

Then

$$\|T_\rho \phi(\delta\rho)\|_{L^2}^2$$

Square-root transform

Let $\phi : \mathrm{Dens}_+(M) \to C^\infty(M)$,

$$\phi(\rho) = \sqrt{\rho}$$

Then

$$\|T_\rho \phi(\delta\rho)\|_{L^2}^2 = \frac{1}{4} \int_M \frac{\delta\rho}{\rho^{1/2}} \frac{\delta\rho}{\rho^{1/2}} dx$$

Square-root transform

Let $\phi : \mathrm{Dens}_+(M) \to C^\infty(M)$,

$$\phi(\rho) = \sqrt{\rho}$$

Then

$$\|T_\rho \phi(\delta\rho)\|_{L^2}^2 = \frac{1}{4}\int_M \frac{\delta\rho}{\rho^{1/2}}\frac{\delta\rho}{\rho^{1/2}}dx = \frac{1}{4}\int_M \frac{\delta\rho}{\rho}\frac{\delta\rho}{\rho}\rho dx$$

Square-root transform

Let $\phi : \mathrm{Dens}_+(M) \to C^\infty(M)$,

$$\phi(\rho) = \sqrt{\rho}$$

Then

$$\|T_\rho \phi(\delta\rho)\|_{L^2}^2 = \frac{1}{4}\int_M \frac{\delta\rho}{\rho^{1/2}} \frac{\delta\rho}{\rho^{1/2}} dx = \frac{1}{4}\int_M \frac{\delta\rho}{\rho} \frac{\delta\rho}{\rho} \rho dx = \frac{1}{4}\langle \delta\rho, \delta\rho \rangle_\rho .$$

Square-root transform

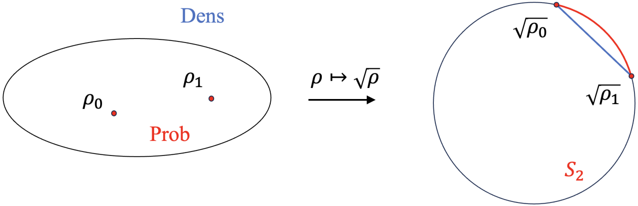
Let $\phi : \mathrm{Dens}_+(M) \to C^\infty(M)$,

$$\phi(\rho) = \sqrt{\rho}$$

Then

$$\|T_\rho \phi(\delta\rho)\|_{L^2}^2 = \frac{1}{4}\int_M \frac{\delta\rho}{\rho^{1/2}}\frac{\delta\rho}{\rho^{1/2}}dx = \frac{1}{4}\int_M \frac{\delta\rho}{\rho}\frac{\delta\rho}{\rho}\rho dx = \frac{1}{4}\langle\delta\rho,\delta\rho\rangle_\rho.$$

The Fisher-Rao metric is the pullback of the L^2 metric by the square-root transform.



The Fisher-Rao geometry is thus flat on $\mathrm{Dens}_+(M)$, and its geodesics are pullbacks of straight lines

$$\rho(t) = ((1-t)\sqrt{\rho_0} + t\sqrt{\rho_1})^2.$$

On $\mathrm{Prob}(M)$, it is spherical, and its geodesics are pullbacks of L^2-sphere geodesics.

Remarks

- Linked to known PDEs in mathematical physics : using

$$\mathrm{Prob}(M) \equiv \mathrm{Diff}(M)/\mathrm{Diff}_{\mathrm{d}x}(M)$$

Khesin, Lenells, Misiolek & Preston (2013) show that a right-invariant $\dot{H}^1$-metric on $\mathrm{Diff}(M)$ induces the Fisher-Rao metric on $\mathrm{Prob}(M)$ and complete integrability of a generalization of the Hunter-Saxton equation. cf Modin (2015).

Remarks

- Linked to known PDEs in mathematical physics : using

$$\mathrm{Prob}(M) \equiv \mathrm{Diff}(M)/\mathrm{Diff}_{\mathrm{d}x}(M)$$

  Khesin, Lenells, Misiolek & Preston (2013) show that a right-invariant $\dot{H}^1$-metric on $\mathrm{Diff}(M)$ induces the Fisher-Rao metric on $\mathrm{Prob}(M)$ and complete integrability of a generalization of the Hunter-Saxton equation. cf Modin (2015).

- Used for Riemannian geometric learning on probability distributions, in
  - image processing (Schwander & Nielsen 2012, Angulo et al. 2024)
  - diffusion tensor imaging (Pennec, Sommer, Fletcher 2019)
  - econometrics (Marriott & Salmon 2000)
  - functional shape and data analysis (Srivastava & Klassen 2016)

  Many parametric Fisher-Rao geometries are implemented in the Python package Geomstats (Miolane et al. 2020, Le Brigant et al. 2023).

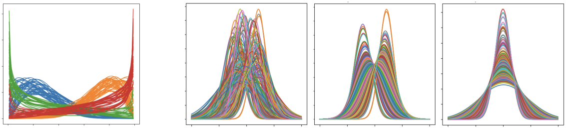


Table of contents

Statistical curvature

In statistics, exponential families are parametric families of probability distributions

$$\mathscr{E} = \{f_\theta;\ \theta \in \Theta\}$$

that enjoy nice properties :

- for hypothesis testing (locally most powerful test is uniformly most powerful)
- for estimation (MLE for θ is a sufficient statistic, achieves the Cramer-Rao bound)

Statistical curvature

In statistics, exponential families are parametric families of probability distributions

$$\mathscr{E} = \{f_\theta;\ \theta \in \Theta\}$$

that enjoy nice properties :

- for hypothesis testing (locally most powerful test is uniformly most powerful)
- for estimation (MLE for θ is a sufficient statistic, achieves the Cramer-Rao bound)

In 1975, Efron introduces a "statistical curvature" that quantifies how much a parametric family deviates from exponentiality. Exponential families have zero curvature.

Statistical curvature

In statistics, exponential families are parametric families of probability distributions

$$\mathscr{E} = \{f_\theta;\ \theta \in \Theta\}$$

that enjoy nice properties :

- for hypothesis testing (locally most powerful test is uniformly most powerful)
- for estimation (MLE for θ is a sufficient statistic, achieves the Cramer-Rao bound)

In 1975, Efron introduces a "statistical curvature" that quantifies how much a parametric family deviates from exponentiality. Exponential families have zero curvature.

Efron's "statistical curvature" corresponds to the curvature induced by a non-metric affine connection ∇^e (Dawid, 1975).

Statistical curvature

In statistics, exponential families are parametric families of probability distributions

$$\mathscr{E} = \{f_\theta;\ \theta \in \Theta\}$$

that enjoy nice properties :

- for hypothesis testing (locally most powerful test is uniformly most powerful)
- for estimation (MLE for θ is a sufficient statistic, achieves the Cramer-Rao bound)

In 1975, Efron introduces a "statistical curvature" that quantifies how much a parametric family deviates from exponentiality. Exponential families have zero curvature.

Efron's "statistical curvature" corresponds to the curvature induced by a non-metric affine connection ∇^e (Dawid, 1975).

(Amari) Inside a exponential family, a (smoothly embedded) sub-family is again an exponential family if and only if its extrinsic curvature w.r.t. ∇^e is zero.

Statistical curvature

In statistics, exponential families are parametric families of probability distributions

$$\mathscr{E} = \{f_\theta;\ \theta \in \Theta\}$$

that enjoy nice properties :

- for hypothesis testing (locally most powerful test is uniformly most powerful)
- for estimation (MLE for θ is a sufficient statistic, achieves the Cramer-Rao bound)

In 1975, Efron introduces a "statistical curvature" that quantifies how much a parametric family deviates from exponentiality. Exponential families have zero curvature.

Efron's "statistical curvature" corresponds to the curvature induced by a non-metric affine connection ∇^e (Dawid, 1975).

(Amari) Inside a exponential family, a (smoothly embedded) sub-family is again an exponential family if and only if its extrinsic curvature w.r.t. ∇^e is zero.

A similar notion of curvature can be defined for mixture families, corresponding to an affine mixture connection ∇^m (Kass, 1989).

α-connections

Let $P_\Theta = \{f_\theta;\ \theta \in \Theta\}$ be parametric family of probability distributions.

The α-connections on P_Θ are a family of affine connections

$$\nabla^{(\alpha)}, \quad -1 \le \alpha \le 1$$

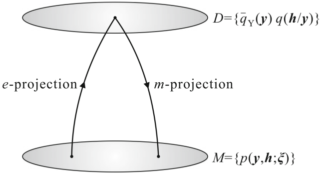
such that

- $\nabla^{(-1)} = \nabla^m, \quad \nabla^{(1)} = \nabla^e, \quad \nabla^{(0)} = \nabla^{FR}$
- $\nabla^{(-\alpha)}$ and $\nabla^{(\alpha)}$ are dual w.r.t. the Fisher-Rao metric.

  i.e. $X\langle Y,Z\rangle = \langle \nabla_X^{(\alpha)} Y, Z\rangle + \langle X, \nabla_X^{(-\alpha)} Z\rangle, \quad \forall X,Y,Z$ vector fields

Introduced by Amari (1982), discussed by Centsov (1972) in the discrete case.
Allowed Amari to express statistical estimation results in geometric terms.

e.g. Expectation-Maximization algorithm (Amari 2016)



α-connections

Let $P_\Theta = \{f_\theta;\ \theta \in \Theta\}$ be parametric family of probability distributions.

The α-connections on P_Θ are a family of affine connections

$$\nabla^{(\alpha)}, \quad -1 \leq \alpha \leq 1$$

such that

- $\nabla^{(-1)} = \nabla^m, \quad \nabla^{(1)} = \nabla^e, \quad \nabla^{(0)} = \nabla^{FR}$
- $\nabla^{(-\alpha)}$ and $\nabla^{(\alpha)}$ are dual w.r.t. the Fisher-Rao metric.

  i.e. $X\langle Y,Z\rangle = \langle \nabla_X^{(\alpha)} Y, Z\rangle + \langle X, \nabla_X^{(-\alpha)} Z\rangle, \quad \forall X,Y,Z$ vector fields

More explicitly, $\nabla^{(\alpha)}$ is defined by

$$\langle \nabla_{e_i}^{(\alpha)} e_j, e_k\rangle_\theta = E_\theta\left[\left(\partial_i\partial_j\ell_\theta + \tfrac{1-\alpha}{2}\,\partial_i\ell_\theta\,\partial_j\ell_\theta\right)\partial_k\ell_\theta\right]$$

where $\langle\cdot,\cdot\rangle$ is the Fisher-Rao metric, $\ell_\theta := \log f_\theta$, and $e_i := \frac{\partial}{\partial\theta_i}$.

Link to α-divergences

The α-divergences are a family of divergences that include the Kullback-Leibler divergence :

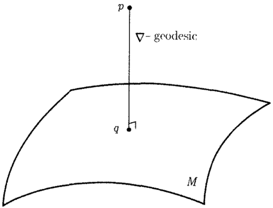
$$D^{(\alpha)}(f_\theta | f_{\theta'}) = \frac{4}{1-\alpha^2}\left(1 - \int f_\theta^{\frac{1-\alpha}{2}} f_{\theta'}^{\frac{1+\alpha}{2}}\,\mathrm{d}x\right) \quad \text{if} \quad \alpha \neq \pm 1$$

$$D^{(-1)}(f_\theta | f_{\theta'}) = D^{(1)}(f_{\theta'} | f_\theta) = \int f_\theta \log \frac{f_\theta}{f_{\theta'}}\,\mathrm{d}x = KL(f_\theta | f_{\theta'})$$

The α-connection can be obtained as

$$\langle \nabla^{(\alpha)}_{e_i} e_j, e_k \rangle_\theta = - \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta'_k} D^{(\alpha)}(f_\theta | f_{\theta'}) \Big|_{\theta = \theta'}$$

where $\langle \cdot, \cdot \rangle$ is the Fisher-Rao metric and $e_i := \frac{\partial}{\partial \theta_i}$.



In an exponential family, the orthogonal projection onto a sub-family w.r.t $\nabla^{(\alpha)}$ gives the best approximation in terms of the α-divergence (Amari).

Non parametric setting : densities

For a, b, c vector fields on $\mathrm{Dens}_+(M)$, define the α-divergence

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{2}{1-\alpha}\int_M \bar{\rho}(x)\mathrm{d}x + \frac{2}{1+\alpha}\int_M \rho(x)\mathrm{d}x - \frac{4}{1-\alpha^2}\int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x,$$

On Prob :

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{4}{1-\alpha^2}\left(1 - \int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x\right).$$

Non parametric setting : densities

For a, b, c vector fields on $\mathrm{Dens}_+(M)$, define the α-divergence

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{2}{1-\alpha}\int_M \bar{\rho}(x)\mathrm{d}x + \frac{2}{1+\alpha}\int_M \rho(x)\mathrm{d}x - \frac{4}{1-\alpha^2}\int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x,$$

On Prob :

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{4}{1-\alpha^2}\left(1 - \int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x\right).$$

The α-connection $\overline{\nabla}^{(\alpha)}$ on $\mathrm{Dens}_+(M)$ is defined by

$$\langle(\overline{\nabla}^{(\alpha)}_a b)_\rho, c_\rho\rangle_\rho := -\partial_\rho\left(\partial_\rho\partial_{\bar{\rho}}\, D^{(\alpha)}(\rho|\bar{\rho})[b,c]\right)[a]\Big|_{\bar{\rho}=\rho} \quad \text{with} \quad \langle a_\rho, b_\rho\rangle_\rho := \int_M \frac{a}{\rho}\frac{b}{\rho}\rho\mathrm{d}x.$$

Non parametric setting : densities

For a,b,c vector fields on $\mathrm{Dens}_+(M)$, define the α-divergence

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{2}{1-\alpha}\int_M \bar{\rho}(x)\mathrm{d}x + \frac{2}{1+\alpha}\int_M \rho(x)\mathrm{d}x - \frac{4}{1-\alpha^2}\int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x,$$

On Prob :

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{4}{1-\alpha^2}\left(1 - \int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x\right).$$

The α-connection $\overline{\nabla}^{(\alpha)}$ on $\mathrm{Dens}_+(M)$ is defined by

$$\langle(\overline{\nabla}^{(\alpha)}_a b)_\rho, c_\rho\rangle_\rho := -\partial_\rho\left(\partial_\rho\partial_{\bar{\rho}}\, D^{(\alpha)}(\rho|\bar{\rho})[b,c]\right)[a]\Big|_{\bar{\rho}=\rho} \quad \text{with} \quad \langle a_\rho, b_\rho\rangle_\rho := \int_M \frac{a}{\rho}\frac{b}{\rho}\rho\mathrm{d}x.$$

This yields $\langle(\overline{\nabla}^{(\alpha)}_a b)_\rho, c_\rho\rangle_\rho = \int_M \frac{Db(a)}{\rho}\frac{c}{\rho}\rho\mathrm{d}x - \frac{1+\alpha}{2}\int_M \frac{a}{\rho}\frac{b}{\rho}\frac{c}{\rho}\rho\mathrm{d}x,$

Non parametric setting : densities

For a,b,c vector fields on $\mathrm{Dens}_+(M)$, define the α-divergence

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{2}{1-\alpha}\int_M \bar{\rho}(x)\mathrm{d}x + \frac{2}{1+\alpha}\int_M \rho(x)\mathrm{d}x - \frac{4}{1-\alpha^2}\int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x,$$

On Prob :

$$D^{(\alpha)}(\rho|\bar{\rho}) := \frac{4}{1-\alpha^2}\left(1-\int_M \rho(x)^{\frac{1-\alpha}{2}}\bar{\rho}(x)^{\frac{1+\alpha}{2}}\mathrm{d}x\right).$$

The α-connection $\overline{\nabla}^{(\alpha)}$ on $\mathrm{Dens}_+(M)$ is defined by

$$\langle(\overline{\nabla}_a^{(\alpha)}b)_\rho, c_\rho\rangle_\rho := -\partial_\rho\left(\partial_\rho\partial_{\bar{\rho}}\, D^{(\alpha)}(\rho|\bar{\rho})[b,c]\right)[a]\Big|_{\bar{\rho}=\rho} \quad \text{with} \quad \langle a_\rho, b_\rho\rangle_\rho := \int_M \frac{a}{\rho}\frac{b}{\rho}\rho\mathrm{d}x.$$

This yields $\langle(\overline{\nabla}_a^{(\alpha)}b)_\rho, c_\rho\rangle_\rho = \int_M \frac{Db(a)}{\rho}\frac{c}{\rho}\rho\mathrm{d}x - \frac{1+\alpha}{2}\int_M \frac{a}{\rho}\frac{b}{\rho}\frac{c}{\rho}\rho\mathrm{d}x,$ and so

$$(\overline{\nabla}_a^{(\alpha)}b)_\rho = Db(a) - \frac{1+\alpha}{2}\frac{a}{\rho}\frac{b}{\rho}\rho.$$

Non parametric setting : probability densities

On $\mathrm{Prob}(M)$, tangent vectors are zero-mean functions $\rightarrow$ for any $a, b \in T\mathrm{Prob}(M)$,

$$\langle (\nabla_a^{(\alpha)} b)_\rho, c_\rho \rangle_\rho := \int_M \frac{Db(a)}{\rho} \frac{c}{\rho} \rho \mathrm{d}x - \frac{1+\alpha}{2} \int_M \frac{a}{\rho} \frac{b}{\rho} \frac{c}{\rho} \rho \mathrm{d}x \qquad \forall c \text{ s.t. } \int_M c \mathrm{d}x = 0,$$

with $\displaystyle\int_M \nabla_a^{(\alpha)} b \, \mathrm{d}x = 0.$

Non parametric setting : probability densities

On $\mathrm{Prob}(M)$, tangent vectors are zero-mean functions $\to$ for any $a, b \in T\mathrm{Prob}(M)$,

$$\langle (\nabla_a^{(\alpha)} b)_\rho, c_\rho \rangle_\rho := \int_M \frac{Db(a)}{\rho} \frac{c}{\rho} \rho \mathrm{d}x - \frac{1+\alpha}{2} \int_M \frac{a}{\rho} \frac{b}{\rho} \frac{c}{\rho} \rho \mathrm{d}x \qquad \forall c \text{ s.t. } \int_M c \mathrm{d}x = 0,$$

with $\displaystyle\int_M \nabla_a^{(\alpha)} b \, \mathrm{d}x = 0.$

This gives $\displaystyle (\nabla_a^{(\alpha)} b)_\rho = \underbrace{Db(a) - \frac{1+\alpha}{2} \frac{a}{\rho} \frac{b}{\rho} \rho}_{\overline{\nabla}_a^{(\alpha)} b \text{ on } \mathrm{Dens}_+} \quad + \quad \text{element of } (T_\rho \mathrm{Prob})^\perp.$

Since $(T_\rho \mathrm{Prob})^\perp = \mathrm{span}(\rho)$,

$$\text{element of } (T_\rho \mathrm{Prob})^\perp = k\rho \quad \text{with} \quad k = \frac{1+\alpha}{2} \int_M \frac{a}{\rho} \frac{b}{\rho} \rho \mathrm{d}x.$$

Non parametric setting : probability densities

On $\mathrm{Prob}(M)$, tangent vectors are zero-mean functions $\to$ for any $a, b \in T\mathrm{Prob}(M)$,

$$\langle(\nabla_a^{(\alpha)} b)_\rho, c_\rho\rangle_\rho := \int_M \frac{Db(a)}{\rho}\frac{c}{\rho}\rho \mathrm{d}x - \frac{1+\alpha}{2}\int_M \frac{a}{\rho}\frac{b}{\rho}\frac{c}{\rho}\rho \mathrm{d}x \qquad \forall c \text{ s.t. } \int_M c\mathrm{d}x = 0,$$

with $\displaystyle\int_M \nabla_a^{(\alpha)} b \,\mathrm{d}x = 0.$

This gives $\displaystyle(\nabla_a^{(\alpha)} b)_\rho = \underbrace{Db(a) - \frac{1+\alpha}{2}\frac{a}{\rho}\frac{b}{\rho}\rho}_{\overline{\nabla}_a^{(\alpha)} b \text{ on } \mathrm{Dens}_+} \quad + \quad$ element of $(T_\rho \mathrm{Prob})^\perp$.

Since $(T_\rho \mathrm{Prob})^\perp = \mathrm{span}(\rho)$,

$$\text{element of } (T_\rho \mathrm{Prob})^\perp = k\rho \quad \text{with} \quad k = \frac{1+\alpha}{2}\int_M \frac{a}{\rho}\frac{b}{\rho}\rho \mathrm{d}x.$$

and we get

$$(\nabla_a^{(\alpha)} b)_\rho = Db(a) - \frac{1+\alpha}{2}\left(\frac{a}{\rho}\frac{b}{\rho} - \int_M \frac{a}{\rho}\frac{b}{\rho}\rho \mathrm{d}x\right)\rho.$$

Remarks

- The geodesic equation of $\overline{\nabla}^{(\alpha)}$ on Dens_+ is locally well-posed

$$\overline{\nabla}^{(\alpha)}_{\dot\rho}\dot\rho = 0 \quad \Leftrightarrow \quad \ddot\rho = \frac{1+\alpha}{2}\frac{\dot\rho^2}{\rho}.$$

Remarks

- The geodesic equation of $\overline{\nabla}^{(\alpha)}$ on Dens_+ is locally well-posed
$$\overline{\nabla}^{(\alpha)}_{\dot\rho}\dot\rho = 0 \quad \Leftrightarrow \quad \ddot\rho = \frac{1+\alpha}{2}\frac{\dot\rho^2}{\rho}.$$

- The geodesic equation of $\nabla^{(\alpha)}$ on $\mathrm{Prob}(M)$ is locally well-posed
$$\nabla^{(\alpha)}_{\dot\rho}\dot\rho = 0 \quad \Leftrightarrow \quad \ddot\rho - \frac{2}{1+\alpha}\rho^{-1}\dot\rho^2 = -\frac{2}{1+\alpha}\left(\int\left(\frac{\dot\rho}{\rho}\right)^2\rho\right)\rho.$$

Remarks

- The geodesic equation of $\overline{\nabla}^{(\alpha)}$ on Dens_+ is locally well-posed

$$\overline{\nabla}^{(\alpha)}_{\dot\rho}\dot\rho = 0 \quad \Leftrightarrow \quad \ddot\rho = \frac{1+\alpha}{2}\frac{\dot\rho^2}{\rho}.$$

- The geodesic equation of $\nabla^{(\alpha)}$ on $\mathrm{Prob}(M)$ is locally well-posed

$$\nabla^{(\alpha)}_{\dot\rho}\dot\rho = 0 \quad \Leftrightarrow \quad \ddot\rho - \frac{2}{1+\alpha}\rho^{-1}\dot\rho^2 = -\frac{2}{1+\alpha}\left(\int \left(\frac{\dot\rho}{\rho}\right)^2 \rho\right)\rho.$$

- The α-connection projects orthogonally, with respect to the Fisher-Rao metric, from Dens_+, to Prob, to any parametric statistical model P_Θ.

$$\begin{array}{ccccc} \mathrm{Dens}_+ & \supset & \mathrm{Prob} & \supset & P_\Theta \\ \overline{\nabla}^{(\alpha)} & \xrightarrow[\text{FR}\perp\text{proj}]{} & \nabla^{(\alpha)} & \xrightarrow[\text{FR}\perp\text{proj}]{} & \nabla^{(\alpha)} \end{array}$$

Table of contents

The L^p-Fisher-Rao metrics

Let $p \in (1, +\infty)$. Given a (probability) density ρ and a tangent vector a at ρ, we define the L^p-Fisher-Rao metric

$$F_p(\rho, a) = \left(\int_M \left| \frac{a}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p}$$

The L^p-Fisher-Rao metrics

Let $p \in (1, +\infty)$. Given a (probability) density ρ and a tangent vector a at ρ, we define the L^p-Fisher-Rao metric

$$F_p(\rho, a) = \left(\int_M \left| \frac{a}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p}$$

- F_2 coincides with the Fisher-Rao norm $F_2(\rho, a) = \sqrt{\langle a, a \rangle_\rho}$.

The L^p-Fisher-Rao metrics

Let $p \in (1, +\infty)$. Given a (probability) density ρ and a tangent vector a at ρ, we define the L^p-Fisher-Rao metric

$$F_p(\rho, a) = \left(\int_M \left| \frac{a}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p}$$

- F_2 coincides with the Fisher-Rao norm $F_2(\rho, a) = \sqrt{\langle a, a \rangle_\rho}$.
- F_p defines a Finsler metric : collection of norms on the tangent spaces.

The L^p-Fisher-Rao metrics

Let $p \in (1, +\infty)$. Given a (probability) density ρ and a tangent vector a at ρ, we define the L^p-Fisher-Rao metric

$$F_p(\rho, a) = \left(\int_M \left| \frac{a}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p}$$

- F_2 coincides with the Fisher-Rao norm $F_2(\rho, a) = \sqrt{\langle a, a \rangle_\rho}$.
- F_p defines a Finsler metric : collection of norms on the tangent spaces.
- F_p defines a notion of geodesics as minimizers of the p-length

$$L(\rho) = \int_0^1 \left(\int_M \left| \frac{\dot{\rho}}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p} dt,$$

where $\rho : [0,1] \to \mathrm{Dens}_+$ such that $\rho(0) = \rho_0$, $\rho(1) = \rho_1$, or equivalently, local minimizers of the p-energy

$$E_p(\rho) = \frac{1}{p} \int_0^1 \int_M \left| \frac{\dot{\rho}}{\rho} \right|^p \rho \mathrm{d}x \, dt.$$

The L^p-Fisher-Rao metrics

Let $p \in (1, +\infty)$. Given a (probability) density ρ and a tangent vector a at ρ, we define the L^p-Fisher-Rao metric

$$F_p(\rho, a) = \left(\int_M \left| \frac{a}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p}$$

- F_2 coincides with the Fisher-Rao norm $F_2(\rho, a) = \sqrt{\langle a, a \rangle_\rho}$.
- F_p defines a Finsler metric : collection of norms on the tangent spaces.
- F_p defines a notion of geodesics as minimizers of the p-length

$$L(\rho) = \int_0^1 \left(\int_M \left| \frac{\dot{\rho}}{\rho} \right|^p \rho \mathrm{d}x \right)^{1/p} dt,$$

  where $\rho : [0,1] \to \mathrm{Dens}_+$ such that $\rho(0) = \rho_0$, $\rho(1) = \rho_1$, or equivalently, local minimizers of the p-energy

$$E_p(\rho) = \frac{1}{p} \int_0^1 \int_M \left| \frac{\dot{\rho}}{\rho} \right|^p \rho \mathrm{d}x \, dt.$$

- The F_p are an information-geometric counterpart of the L^p-Wasserstein metrics.

The p-root transform

For $p \in (1, \infty)$, define the p-root transform

$$\Phi_p(\rho \mathrm{d}x) = \rho^{1/p}.$$

Theorem (Bauer, L., Lu, Maor)

- Φ_p is an isometric embedding

$$(\mathrm{Dens}_+(M), \tfrac{1}{p} F_p) \to (C^\infty(M), \|\cdot\|_{L^p}).$$

  Thus the F_p-geodesic and geodesic distance between ρ_0 and ρ_1 are

$$\rho(t) = ((1-t)\sqrt[p]{\rho_0} + t\sqrt[p]{\rho_1})^p, \quad d_p(\rho_0, \rho_1) = \left(\int_M |\sqrt[p]{\rho_1} - \sqrt[p]{\rho_0}|^p \, \mathrm{d}x\right)^{1/p}.$$

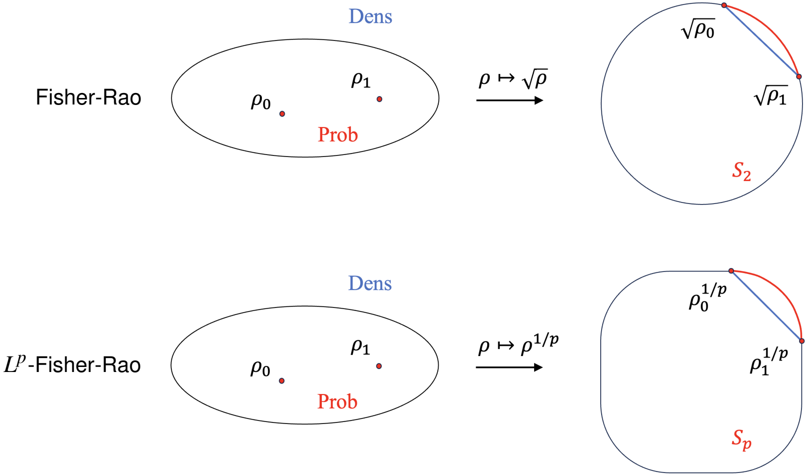
- Φ_p is an isometric embedding

$$(\mathrm{Prob}(M), \tfrac{1}{p} F_p) \to (S_p, \|\cdot\|_{L^p})$$

  where $S_p := \{f \in C^\infty(M) : \|f\|_{L^p} = 1\}$ is the L^p-sphere equipped with the restriction of the standard L^p-norm.

The p-root transform

This generalizes the square root transform and its link to the Fisher-Rao metric.



Link between F_p and $\nabla^{(\alpha)}$ on Dens_+

Theorem (Bauer, L., Lu, Maor) Let M be a closed manifold, $p > 1$, and $\alpha = 1 - \frac{2}{p}$.

- On $\mathrm{Dens}_+(M)$, the L^p-Fisher-Rao metric and the α-connection define the same geodesics.
- Equivalently : the Chern connection associated to the Finsler L^p-Fisher-Rao metric coincides with the α connection :

$$\nabla^{\nu}_{a}\nu = \nabla^{(\alpha)}_{a}\nu$$

  for every nowhere vanishing vector field ν and any $a \in T\mathrm{Dens}_+(M)$.
- Equivalently : the $\nabla^{(\alpha)}$ geodesics are energy-minimizing curves for

$$E_{\frac{2}{1-\alpha}}(\rho) = \frac{1-\alpha}{2}\int_0^1 \int_M \left|\frac{\dot{\rho}}{\rho}\right|^{\frac{2}{1-\alpha}} \rho \mathrm{d}x\, dt.$$

Link between F_p and $\nabla^{(\alpha)}$ on Dens_+

Theorem (Bauer, L., Lu, Maor) Let M be a closed manifold, $p > 1$, and $\alpha = 1 - \frac{2}{p}$.

- On $\mathrm{Dens}_+(M)$, the L^p-Fisher-Rao metric and the α-connection define the same geodesics.
- Equivalently : the Chern connection associated to the Finsler L^p-Fisher-Rao metric coincides with the α connection :

$$\nabla^{\nu}_{a}\nu = \nabla^{(\alpha)}_{a}\nu$$

  for every nowhere vanishing vector field ν and any $a \in T\mathrm{Dens}_+(M)$.
- Equivalently : the $\nabla^{(\alpha)}$ geodesics are energy-minimizing curves for

$$E_{\frac{2}{1-\alpha}}(\rho) = \frac{1-\alpha}{2}\int_0^1 \int_M \left|\frac{\dot{\rho}}{\rho}\right|^{\frac{2}{1-\alpha}} \rho \mathrm{d}x\, dt.$$

The F_p-geodesics on $\mathrm{Dens}_+(M)$ previously described are also the $\nabla^{(\alpha)}$-geodesics ! This relates to the work of Giblisco and Pistone (1998), who have studied the α-connections in the non-parametric setting using a similar p-root transform.

Link between F_p and $\nabla^{(\alpha)}$ on Prob

On $\mathrm{Prob}(M)$, F_p and $\nabla^{(\alpha)}$ no longer define the same geodesics !

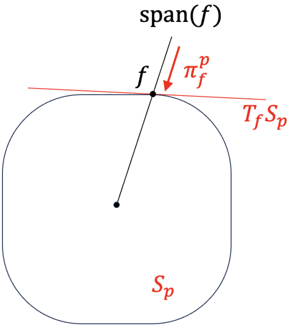
$$
\begin{tikzcd}
\mathrm{Dens}_+(M), F_p, \overline{\nabla}^{(\alpha)} & & C^\infty(M), \|\cdot\|_{L^p}, \nabla^{tr} \arrow[d, "\pi^p"] \\
\mathrm{Prob}(M), F_p, \nabla^{(\alpha)} \arrow[u, hook] \arrow[rr, leftrightarrow, "\rho \mapsto \rho^{1/p}", "f^p \leftarrow f"'] & & S_p(M), \|\cdot\|_{L^p|S_p}, \pi^p(\nabla^{tr}) \arrow[u, hook]
\end{tikzcd}
$$

$\nabla^{(\alpha)}$-geodesics on Prob

Consider the connection on $S_p := \{f \in C^\infty(M) : \|f\|_{L^p} = 1\}$ defined as

$$\nabla^p_U V = \pi^p \left(\nabla^{\mathrm{tr}}_U V\right)$$

the projection of the trivial connection w.r.t. the splitting $T_f C^\infty = T_f S_p \oplus \mathrm{span}(f)$.



$\nabla^{(\alpha)}$-geodesics on Prob

Theorem (Bauer, L., Lu, Maor)

- The pullback of $\Phi_p^*\nabla^p$ coincides with $\nabla^{(\alpha)}$ up to a constant depending only on the footpoint. In particular, the geodesics of $\Phi_p^*\nabla^p$ and $\nabla^{(\alpha)}$ coincide.
- Geodesics on S_p for ∇^p with initial conditions $\gamma(0) = f$, $\dot\gamma(0) = U$ are given by

$$\gamma(t) = \frac{f + \tau(t)U}{\|f + \tau(t)U\|_{L^p}}, \quad t \in I,$$

where $\tau : I \to \mathbb{R}$ verifies

$$\ddot\tau(t) = 2\frac{\int_M |f + \tau(t)U|^{p-2}(f + \tau(t)U)U\,\mathrm{d}x}{\int_M |f + \tau(t)U|^p\,\mathrm{d}x}\dot\tau(t)^2, \quad \tau(0) = 0, \quad \dot\tau(0) = 1.$$

Geodesics of $\nabla^{(\alpha)}$ are obtained by pulling back these geodesics using Φ_p.

Summary

Let $\alpha \in (-1,1)$ and $p = \frac{2}{1-\alpha}$.

$$\begin{tikzcd}
\mathrm{Dens}_+(M), F_p, \overline{\nabla}^{(\alpha)} \arrow[rr, leftrightarrow, "\rho \mapsto \rho^{1/p}", "f^p \leftarrow f"'] & & C^\infty(M), ||\cdot||_{L^p}, \nabla^{tr} \arrow[d, shift left=2, "\pi^p"] \\
\mathrm{Prob}(M), F_p, \nabla^{(\alpha)} \arrow[u, hook] & & S_p(M), ||\cdot||_{L^p|_{S_p}}, \pi^p(\nabla^{tr}) \arrow[u, hook, shift left=2]
\end{tikzcd}$$

3 notions of geodesics :

(1) the $\nabla^{(\alpha)}$-geodesic = L^p-Fisher-Rao geodesic on $\mathrm{Dens}_+(M)$

(2) the $\nabla^{(\alpha)}$-geodesic on $\mathrm{Prob}(M)$

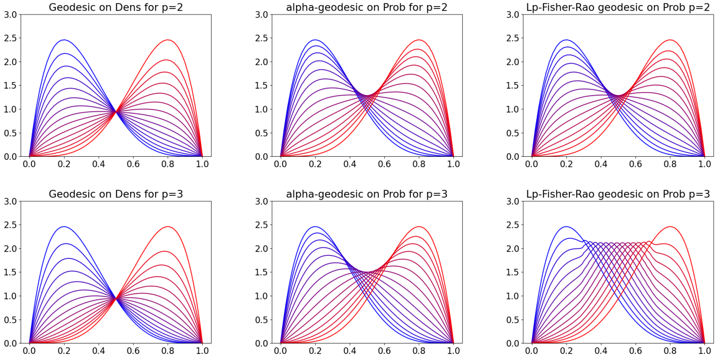
(3) the L^p-Fisher-Rao geodesic on $\mathrm{Prob}(M)$

3 notions of geodesics

$\mathrm{Dens}_+,\ F_p/\nabla^{(\alpha)}$ | $\mathrm{Prob},\ \nabla^{(\alpha)}$ | $\mathrm{Prob},\ F_p$

$p = 2$

$p = 3$



Different notions of geodesics between distributions on $[0, 1]$.

...and sometimes, a fourth !

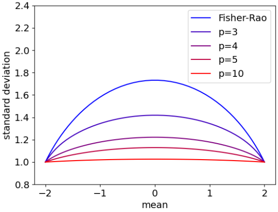
L^p-Fisher-Rao metric on a parametric statistical model $P_\Theta = \{f_\theta \mathrm{d}x;\ \theta \in \Theta\}$:

$$F_p(\theta, v) = \mathbb{E}\left(|\langle \nabla_\theta \ell(X,\theta), v\rangle|^p\right)^{1/p}$$

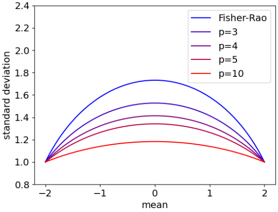
- F_p defines a Finsler metric on the parameter space Θ
- For $p = 2$ we retrieve the Fisher information metric

$$F_2(\theta,v)^2 = \mathbb{E}\left(\langle \nabla_\theta \ell(X,\theta), v\rangle^2\right) = v^\top \mathbb{E}\left(\nabla_\theta \ell(X,\theta) \nabla_\theta \ell(X,\theta)^\top\right) v = v^\top I(\theta) v = \langle v, v\rangle_\theta.$$

Example : normal distributions



F_p geodesics



$\nabla^{(\alpha)}$-geodesics

Concluding remarks

- The L^p-Fisher-Rao metric F_p already appeared in the definition of the generalized unbalanced optimal transport metric in Chizat, Schmitzer, Peyré, Vialard (2018)

- F_p relates to known PDEs on $\mathrm{Prob}(M) \equiv \mathrm{Diff}(M)/\mathrm{Diff}_{\mathrm{d}x}(M)$:

  - On any closed M, a family of right-invariant $\dot{W}^{1,p}$-Finsler metrics on $\mathrm{Diff}(M)$ induce F_p on $\mathrm{Prob}(M)$. This generalizes the work of Khesin et al. (2013) on Hunter-Saxton.

  - F_p - geodesic on $\mathrm{Prob}(S^1) \to$ periodic r-Hunter–Saxton equation ($r = 1/p$)
    $\nabla^{(\alpha)}$-geodesic on $\mathrm{Prob}(S^1) \to$ generalized periodic inviscid Proudman-Johnson eq. (Lenells Misiolek 2014).

  - $\nabla^{(\alpha)}$ / F_p-geodesics on $\mathrm{Dens}_+(\mathbb{R}) \to$ generalized non-periodic inviscid Proudman-Johnson equation $\equiv$ non-periodic r-Hunter-Saxton equation ($r = 1/p$) (Bauer, Lu, Maor 2022).

- More details in our paper :

  Bauer, Lu, Le Brigant, Maor 2024 : The L^p-Fisher-Rao metric and Amari-Cencov alpha-connections. *Calculus of Variations and Partial Differential Equations*.

Thank you for your attention !

References

S. Amari and H. Nagaoka. Methods of information geometry, volume 191. American Mathematical Soc., 2000.

C. Atkinson and A. F. Mitchell. Rao's distance measure. Sankhya : The Indian Journal of Statistics, Series A, 45–365, 1981.

N. Ay, J. Jost, H. V. Le, and L. Schwachhofer. Information geometry and sufficient statistics. Probability Theory and Related Fields, 162 :327–364, 2015.

M. Bauer, M. Bruveris, and P. W. Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. Bulletin of the London Mathematical Society, 48(3) :499–506, 2016.

M. Bauer, A. Le Brigant, Y. Lu and C. Maor. The L^p-Fisher–Rao metric and Amari–Cencov α-Connections. Calculus of Variations and Partial Differential Equations, 63(2), 1-33, 2024.

N.N. Cencov. Statistical Decision Rules and Optimal Inference. American Mathematical Society, Rhode Island, U.S.A., 1982. (Originally published in Russian, Nauka, Moscow, 1972).

R. Chakraborty, D. Seo, B. C. Vemuri. An efficient exact-pga algorithm for constant curvature manifolds. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3976-3984, 2016.

L. Chizat, G. Peyré, B. Schmitzer, and F. X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. Foundations of Computational Mathematics, 18, 1-44, 2018.

Costa, Santos, Strapasson. Fisher information distance : A geometrical reading. Discrete Applied Mathematics 197, 59-69, 2015.

A. P. Dawid. Discussion of "Defining the curvature of a statistical problem (with applications to second-order efficiency)," by B. Efron. Annals of Statistics 3, 1231-1234, 1975.

B. Efron. Defining the curvature of a statistical problem. Annals of Statistics 3, 1189-1242, 1975.

T. Friedrich. Die Fisher-information und symplektische strukturen. Mathematische Nachrichten, 153(1) :273–296, 1991.

References

P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. Infinite Dimensional Analysis, Quantum Probability and Related Topics, 1(02) :325–347, 1998.

H. Jeffreys. An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. London Ser. A, 196, 453-461, 1946.

B. Khesin, J. Lenells, G. Misiolek, and S. C. Preston. Geometry of diffeomorphism groups, complete integrability and geometric statistics. Geometric and Functional Analysis, 23(1) :334–366, 2013.

A. Le Brigant, J. Deschamps, A. Collas and N. Miolane. Parametric information geometry with the package Geomstats. ACM Transactions on Mathematical Software, 49(4), 1-26, 2023.

J. Lenells and G. Misiolek. Amari–Chentsov connections and their geodesics on homogeneous spaces of diffeomorphism groups. Journal of Mathematical Sciences, 196(2) :144–151, 2014.

P. Marriott and M. Salmon. Applications of differential geometry to econometrics. Cambridge University Press, 2000.

N. Miolane et al. Geomstats : a Python package for Riemannian geometry in machine learning. Journal of Machine Learning Research 21.223 (2020) : 1-9.

X. Pennec, S. Sommer, and T. Fletcher, eds. Riemannian geometric statistics in medical image analysis. Academic Press, 2019.

C.R. Rao, Information and the accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc. 37, 1945.

O. Schwander and F. Nielsen. Model centroids for the simplification of kernel density estimators. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 737–740. IEEE, 2012.

L. T. Skovgaard, A Riemannian geometry of the multivariate normal model, Scand. J. Stat. (1984) 211–223.

A. Srivastava, and E. Klassen. Functional and shape data analysis. Vol. 1. New York : Springer, 2016.