Imperial College London

Stochastic Analysis seminar

Information Geometry with Differentiable Densities

Giovanni Pistone
www.giannidiorestino.it

DE CASTRO STATISTICS
Collegio Carlo Alberto

Feb 2, 2016

Abstract

- We focus on a specific way to put a global differentiable structure on positive densities of a measure space, namely the Banach manifold modelled on Orlicz spaces introduced by Pistone and Sempi (1995).
- This framework is able to overcome the limitations of other approaches (such as the classical embedding a density in the Hilbert sphere by its square root) which are due to the fact the relative interior of the positive cone of square integrable functions is empty unless the sample space is finite. Recent research has improved our structure so that the current version allows to construct with a minimum of technicalities a differential structure which is able to support first and second order calculus and reduces to Amari's Information Geometry on parametric sub-models. While classical statistical applications can use parameters, other applications, such as Stochastic Analysis, are intrinsically nonparametric, hence the advantage of a way to avoid parameters. On the other side, this result is obtained at the cost of reducing the set of densities available to essentially those which have a finite relative divergence from a given one.
- When [the] reference density is the Gaussian density we obtain a special set-up that allows for space differentiability through the introduction of Orlicz-Sobolev model spaces. . . .
- I plan to hint to a number of potential applications of such a Calculus, e.g. Continuity Equation, Kolmogorov Forward Equation, Hyvrinen Divergence, Gradient Flow, Wasserstein distance, Continuous Martingale.
- I refer to recent joint work: with L. Malagò (2015); with B. Lods (2015); with D. Brigo (2016).

Thanks to:

- Carlo Sempi (Università di Lecce)
- Damiano Brigo (Imperial College, London)
- Paolo Gibilisco (II Università di Roma Tor Vergata)
- Maria Piera Rogantin (DIMA Università di Genova)
- Alberto Cena (ITCS E. Bona, Biella)
- Daniele Imparato (dom Placido, San Miniato Firenze)
- Paola Siri (DISMA Politecnico di Torino)
- Barbara Trivellato (DISMA Politecnico di Torino)
- Marina Santacroce (DISMA Politecnico di Torino)
- Luigi Malagò (DEEE Shinshu University, Japan)
- Bertrand Lods (Università di Torino & Collegio Carlo Alberto)

Cette conversation est dédiée à Michel Metivier, mon maitre à Rennes (1973-75)

Why a nonparametric IG?

- Applications without natural parameters:
  - Flows e.g. gradient flows on the probability simplex.
  - Homogeneous Boltzmann equation

$$\partial_t f = Q(f, f);$$

  - Divergences in Machine Learning e.g., Hyvärinen divergence

$$\mathrm{DH}\,(p\|q) = \int \left|\nabla \log p(x) - \nabla \log q(x)\right|^2 q(x)\ dx$$

  - Evolution equation for densities e.g., heat equation

$$\partial_t f = \Delta_{\boldsymbol{x}} f;$$

- G. Pistone. Examples of the application of nonparametric information geometry to statistical physics. *Entropy*, 15(10):4042–4065, 2013
- G. Pistone. Nonparametric information geometry. In F. Nielsen and F. Barbaresco, editors, *Geometric science of information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings
- B. Lods and G. Pistone. Information geometry formalism for the spatially homogeneous Boltzmann equation. *Entropy*, 17(6):4323–4363, 2015
- D. Brigo and G. Pistone. Projection based dimensionality reduction for measure valued evolution equations in statistical manifolds. arXiv:1601.04189 [math.PR], 2016

Why "exponential family"?

- The cone of strictly positive unnormalized densities is an affine space for the multiplication. The additive representation of this affine geometry is the exponential family.
    - H. Gzyl and L. Recht. A geometry on the space of probabilities. I. The finite dimensional case. *Rev. Mat. Iberoam.*, 22(2):545–558, 2006
    - H. Gzyl and L. Recht. A geometry on the space of probabilities. II. Projective spaces and exponential families. *Rev. Mat. Iberoam.*, 22(3):833–849, 2006
- Previous work (and current work) on generalising exponential families was focused on the generalisation of parameters to infinite dimension. Our idea is to avoid parameters at all.
- Non-parametric = coordinate-free differential geometry exits, and it is simpler than its version based on coordinates
    - R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, tensor analysis, and applications*, volume 75 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1988
    - S. Lang. *Differential and Riemannian manifolds*, volume 160 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 1995

MY THESIS: IG is the geometry of the statistical bundle

In Information Geometry we want to perform computations such as

$$\frac{d}{d\theta} \int u(x) p(x;\theta) \ \mu(dx) = \int u(x) \frac{d}{d\theta} p(x;\theta) \ \mu(dx) =$$

$$\int u(x) \frac{d}{d\theta} \log p(x;\theta) \ p(x;\theta) \ \mu(dx) = \mathrm{E}_\theta \left[(u - \mathrm{E}_\theta [u]) \left(\frac{d}{d\theta} \log p_\theta \right) \right]$$

- Δ is the probability simplex on a given sample space $(\Omega, \mathcal{F})$.
- The statistical bundle of Δ is

$$T\Delta = \left\{ (\pi, u) \middle| \pi \in \Delta, u \in L^2(\pi), \mathrm{E}_\pi [u] = 0 \right\}$$

- We want the fibers $L^2_0(\mu)$ to be isomorphic and express the tangent space.
  - P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *IDAQP*, 1(2):325–347, 1998
- This program is easily feasible if the sample space Ω is finite. If Ω is not finite, we have a problem.

Model space

Orlicz Φ-space

If $\phi(y) = \cosh y - 1$, the Orlicz Φ-space $L^\Phi(p)$ is the vector space of all random variables U such that $\mathrm{E}_p[\Phi(\alpha U)]$ is finite for some $\alpha > 0$.

Properties of the Φ-space

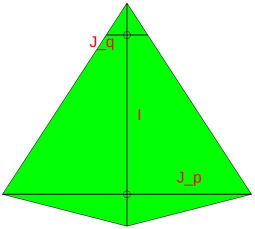
1. $U \in L^\Phi(p)$ if, and only if, the moment generating function $\alpha \mapsto \mathrm{E}_p[\mathrm{e}^{\alpha u}]$ is finite in a neighbourhood of 0. $L^\Phi(p)$ is the space of sufficient statistics in an exponential family.
2. The set
$$\left\{u \in L^\Phi(p) \middle| \mathrm{E}_p[\Phi(u)] \leq 1\right\}$$
is the closed unit ball of a Banach space, hence
$$\|u\|_p = \inf\left\{\rho > 0 \middle| \mathrm{E}_p\left[\Phi\left(\frac{u}{\rho}\right)\right] \leq 1\right\}.$$
3. $\lim_{n\to\infty} u_n = 0$ in $L^\Phi(p)$ if and only if for all $\epsilon > 0$
$$\limsup_{n\to\infty} \mathrm{E}_p\left[\Phi(\epsilon^{-1} u_n)\right] \leq 1$$

Isomorphism of L^Φ spaces

$L^\Phi(p) = L^\Phi(q)$ as Banach spaces if $\theta \mapsto \int p^{1-\theta} q^\theta \, d\mu$ is finite on an open neighbourhood I of $[0,1]$. It is an equivalence relation $p \smile q$ and we denote by $\mathcal{E}(p)$ the class containing p.

Proof.

Assume $u \in L^\Phi(p)$ and consider the convex function $C \colon (s,\theta) \mapsto \int \mathrm{e}^{su} p^{1-\theta} q^\theta \, d\mu$. The restriction $s \mapsto C(s,0) = \int \mathrm{e}^{su} p \, d\mu$ is finite on an open neighbourhood J_p of 0; the restriction $\theta \mapsto C(0,\theta) = \int p^{1-\theta} q^\theta \, d\mu$ is finite on the open set $I \supset [0,1]$. hence, there exists an open interval $J_q \ni 0$ where $s \mapsto C(s,1) = \int \mathrm{e}^{su} q \, d\mu$ is finite. □



- G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, October 1995

Portmanteau theorem

The following statements are equivalent:

- $q \in \mathcal{E}(p)$;
- $p \smile q$;
- $\mathcal{E}(p) = \mathcal{E}(q)$;
- $L^{\Phi}(p) = L^{\Phi}(q)$;
- $\log\left(\frac{q}{p}\right) \in L^{\Phi}(p) \cap L^{\Phi}(q)$.
- $\frac{q}{p} \in L^{1+\epsilon}(p)$ and $\frac{p}{q} \in L^{1+\epsilon}(q)$ for some $\epsilon > 0$.

- A. Cena. *Geometric structures on the non-parametric statistical manifold*. PhD thesis, Dottorato in Matematica, Università di Milano, 2002
- A. Cena and G. Pistone. Exponential statistical manifold. *Ann. Inst. Statist. Math.*, 59(1):27–56, 2007
- M. Santacroce, P. Siri, and B. Trivellato. New results on mixture and exponential models by Orlicz spaces. *Bernoulli*, 2015. online first

Maximal exponential family

- For each $p \in \mathcal{P}_>$, the moment generating functional is the positive lower-semi-continuous convex function $G_p \colon B_p \ni U \mapsto \mathrm{E}_p\left[\mathrm{e}^U\right]$ and
- the cumulant generating functional is the non-negative lower semicontinuous convex function $K_p = \log G_p$.
- The interior of the common proper domain $\{U|G_p(U) < +\infty\}^\circ = \{U|K_p(U) < \infty\}^\circ$ is an open convex set $\mathcal{S}_p$ containing the open unit ball (for the norm of the Orlicz space).
- For each $p \in \mathcal{P}_>$, the maximal exponential family at p is

$$\mathcal{E}(p) = \left\{ \mathrm{e}^{u - K_p(u)} \cdot p \middle| u \in \mathcal{S}_p \right\}.$$

e-charts

- For each $p \in \mathcal{P}_>$, $p \in \mathcal{E}$, consider the chart $s_p \colon \mathcal{E} \to L_0^\Phi(p) = B_p$

$$\boxed{s_p \colon \mathcal{E} \ni q \mapsto \log\left(\frac{q}{p}\right) + D(p\|q) = \log\left(\frac{q}{p}\right) - \mathrm{E}_p\left[\log\left(\frac{q}{p}\right)\right]}$$

- For $U \in B_p$ let $K_p(U) = \log \mathrm{E}_p\left[\mathrm{e}^U\right]$ the cumulant generating function of U and let $\mathcal{S}_p$ the interior of the proper domain. Define

$$\boxed{e_p = s_p^{-1} \colon \mathcal{S}_p \ni U \mapsto \mathrm{e}^{U - K_p(U)} \cdot p}$$

- $\{s_p \colon \mathcal{E}(p) | p \in \mathcal{P}_>\}$ is an affine atlas on $\mathcal{P}_>$ that defines the exponential manifold.
- Each $\mathcal{E}(p)$ is a connected component.
- The information closure of any $\mathcal{E}(p)$ is $\mathcal{P}_\geq$. The reverse information closure of any $\mathcal{E}(p)$ is $\mathcal{P}_>$.

- D. Imparato and B. Trivellato. Geometry of extended exponential models. In *Algebraic and geometric methods in statistics*, pages 307–326. Cambridge Univ. Press, Cambridge, 2010

Cumulant functional

- The r-divergence $q \mapsto \mathrm{D}\left(p \| q\right)$ is represented in the chart centered at p by $\mathrm{D}\left(p \| e_p(U)\right) = K_p(U) = \log \mathrm{E}_p\left[\mathrm{e}^U\right]$, where $q = e_p(U) = \mathrm{e}^{U - K_p(U)} \cdot p$, $u \in B_p$.
- $K_p : B_p \to \mathbb{R}_{\geq} \cup \{+\infty\}$ is convex and its proper domain $\mathrm{Dom}\left(K_p\right)$ contains the open unit ball of B_p.
- K_p is infinitely Gâteaux-differentiable on the interior $\mathcal{S}p$ of its proper domain and analytic on the unit ball of B_p.
- For all $V, V_1, V_2, V_3 \in B_p$ the first derivatives are:

$$\begin{aligned} \mathrm{d}\, K_p(U)[V] &= \mathrm{E}_q\left[V\right] \\ \mathrm{d}^2\, K_p(U)[V_1, V_2] &= \mathrm{Cov}_q\left(V_1, V_2\right) \\ \mathrm{d}^3\, K_p(U)[V_1, V_2, V_3] &= \mathrm{Cov}_q(V_1, V_2, V_3) \end{aligned}$$

Summary

$$\boxed{p \smile q} \Longrightarrow \boxed{\begin{array}{ccccccc} \mathcal{E}(p) & \xrightarrow{s_p} & \mathcal{S}p & \xrightarrow{I} & B_p & \xrightarrow{I} & L^{\Phi}(p) \\ \Big\| & & \Big\downarrow s_q \circ s_p^{-1} & & \Big\downarrow d(s_q \circ s_p^{-1}) & & \Big\| \\ \mathcal{E}(q) & \xrightarrow[s_q]{} & \mathcal{S}q & \xrightarrow[I]{} & B_q & \xrightarrow[I]{} & L^{\Phi}(q) \end{array}}$$

- If $p \smile q$, then $\mathcal{E}(p) = \mathcal{E}(q)$ and $L^{\Phi}(p) = L^{\Phi}(q)$.
- $B_p = L_0^{\Phi}(p)$, $B_q = L_0^{\Phi}(q)$
- $\mathcal{S}p \neq \mathcal{S}q$ and $s_q \circ s_p^{-1} \colon \mathcal{S}p \to \mathcal{S}q$ is affine

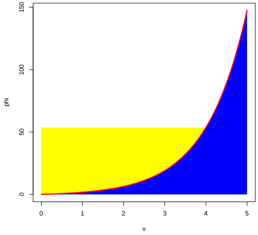
$$s_q \circ s_p^{-1}(U) = U - \mathrm{E}_q[U] + \log\left(\frac{p}{q}\right) - \mathrm{E}_q\left[\log\left(\frac{p}{q}\right)\right]$$

- The tangent application is $d(s_q \circ s_p^{-1})(U)[V] = V - \mathrm{E}_q[V]$ (does not depend on p)

Duality

Young pair (N–function)

- $\phi^{-1} = \phi_*$,
- $\Phi(x) = \int_0^{|x|} \phi(u)\ du$
- $\Phi_*(y) = \int_0^{|y|} \phi_*(v)\ dv$
- $|xy| \leq \Phi(x) + \Phi_*(y)$



$\phi_*(u)$	$\phi(v)$	$\Phi_*(x)$	$\Phi(y)$
$\log(1+u)$	$\mathrm{e}^v - 1$	$(1+\|x\|)\log(1+\|x\|) - \|x\|$	$\mathrm{e}^{\|y\|} - 1 - \|y\|$
$\sinh^{-1} u$	$\sinh v$	$\|x\| \sinh^{-1} \|x\| - \sqrt{1+x^2} + 1$	$\cosh y - 1$

- $L^{\Phi_*}(p) \times L^{\Phi}(p) \ni (v,u) \mapsto \langle u, v \rangle_p = \mathrm{E}_p[uv]$
- $\left| \langle u, v \rangle_p \right| \leq 2 \|u\|_{\Phi_*,p} \|v\|_{\Phi,p}$
- $(L^{\Phi_*}(p))' = L^{\Phi}(p)$ because $\Phi_*(ax) \leq a^2 \Phi_*(x)$ if $a > 1$ (Δ_2).

PART II

Second order geometry

Parallel transport

- e-transport:

$$ {}^{\mathrm{e}}\mathbb{U}_p^q \colon B_p \ni U \mapsto U - \mathrm{E}_q\left[U\right] \in B_q \ . $$

- m-transport: for each $V \in {}^*B_q$

$$ {}^{\mathrm{m}}\mathbb{U}_q^p \, {}^*B_q \ni V \mapsto \frac{q}{p} V \in {}^*B_p $$

Properties

- $\left\langle U, {}^{\mathrm{m}}\mathbb{U}_q^p V \right\rangle_p = \left\langle {}^{\mathrm{e}}\mathbb{U}_p^q U, V \right\rangle_q$
- ${}^{\mathrm{e}}\mathbb{U}_q^r \, {}^{\mathrm{e}}\mathbb{U}_p^q = {}^{\mathrm{e}}\mathbb{U}_p^r$
- ${}^{\mathrm{m}}\mathbb{U}_q^r \, {}^{\mathrm{m}}\mathbb{U}_p^q = {}^{\mathrm{m}}\mathbb{U}_p^r$
- $\left\langle {}^{\mathrm{e}}\mathbb{U}_p^q U, {}^{\mathrm{m}}\mathbb{U}_p^q V \right\rangle_q = \langle U, V \rangle_p$
- $d^2 K_p(q)[U, V] = \left\langle {}^{\mathrm{e}}\mathbb{U}_p^q U, {}^{\mathrm{e}}\mathbb{U}_p^q V \right\rangle_q = \left\langle {}^{\mathrm{m}}\mathbb{U}_q^p \, {}^{\mathrm{e}}\mathbb{U}_p^q U, V \right\rangle_p$.

Statistical exponential manifold and bundles

- The exponential manifold is the maximal exponential family $\mathcal{E}$ with the affine atlas of global charts $(s_p \colon p \in \mathcal{E})$,

$$s_p(q) = \log \frac{q}{p} - \mathrm{E}_p\left[\log \frac{q}{p}\right].$$

- The statistical exponential bundle $S\,\mathcal{E}$ is the manifold defined on the set

$$\{(p, V) | p \in \mathcal{E}, V \in B_p\}$$

by the affine atlas of global charts

$$\sigma_p \colon (q, V) \mapsto \left(s_p(q), {}^{\mathrm{e}}\mathbb{U}_q^p V\right) \in B_p \times B_p, \quad p \in \mathcal{E}$$

- The statistical predual bundle ${}^*S\,\mathcal{E}$ is the manifold defined on the set

$$\{(p, W) | p \in \mathcal{E}, W \in {}^*B_p\}$$

by the affine atlas of global charts

$${}^*\sigma_p \colon (q, W) \mapsto \left(s_p(q), {}^{\mathrm{m}}\mathbb{U}_q^p W\right) \in B_p \times {}^*B_p, \quad p \in \mathcal{E}$$

Score and statistical gradient

Definition

$t \mapsto p(t)$ is a curve in $\mathcal{E}(p)$ and $f \colon \mathcal{E} \to \mathbb{R}$.

- The score of the curve $t \mapsto p(t)$ is a curve in the statistical bundle $t \mapsto (p(t), Dp(t)) \in S\,\mathcal{E}(p)$ such that for all $X \in L^{\Phi}(p)$ it holds

$$\frac{d}{dt}\mathrm{E}_{p(t)}[X] = \left\langle X - \mathrm{E}_{p(t)}[X], Dp(t) \right\rangle_{p(t)}$$

- Usually,

$$Dp(t) = \frac{\dot{p}(t)}{p(t)} = \frac{d}{dt}\log p(t)$$

- The statistical gradient of f is a section of the statistical bundle, $p \mapsto (p, \operatorname{grad} f(p)) \in S\,\mathcal{E}(p)$ such that for each regular curve $t \mapsto p(t)$, it holds

$$\frac{d}{dt} f(p(t)) = \langle \operatorname{grad} f(p(t)), Dp(t) \rangle_{p(t)}$$

Taylor formula in the Statistical Bundle

- For a curve $t \mapsto p(t) \in \mathcal{E}$ connecting $p = p(0)$ to $q = p(1)$ and a function $f \colon \mathcal{E} \to \mathbb{R}$ the Taylor formula is

$$f(q) = f(p) + \left. \frac{d}{dt} f(p(t)) \right|_{t=o} + \frac{1}{2} \left. \frac{d^2}{dt^2} f(p(t)) \right|_{t=o} + R_2(f, p, q)$$

- The first derivative is computed with the statistical gradient and the score

$$f(q) = f(p) + \langle \operatorname{grad} f(p(0)), Dp(0) \rangle_p + \frac{1}{2} \left. \frac{d}{dt} \langle \operatorname{grad} f(p(t)), Dp(t) \rangle_{p(t)} \right|_{t=o} + R_2(f, p, q)$$

Accellerations

- Let us define the acceleration at t of a curve $t \mapsto p(t) \in \mathcal{E}$. The velocity is defined to be
  $t \mapsto (p(t), Dp(t)) = \left(p(t), \frac{d}{dt} \log\left(p(t)\right)\right) \in S\,\mathcal{E}$
- The exponential acceleration is

$$^{e}\mathrm{D}^2 p(t) = \left. \frac{d}{ds} \, {}^{\mathrm{e}}\mathbb{U}_{p(s)}^{p(t)} Dp(s) \right|_{s=t}$$

- The mixture acceleration is

$$^{m}\mathrm{D}^2 p(t) = \left. \frac{d}{ds} \, {}^{\mathrm{m}}\mathbb{U}_{p(s)}^{p(t)} Dp(s) \right|_{s=t}$$

- The e-accelleration of a 1d-exponential family is zero
- The m-accelleration of a 1d-mixture family is zero

Taylor's formulæ I

1. $t \mapsto p(t)$ is the mixture geodesic connecting $p = p(0)$ to $q = p(1)$.

$$f(q) = f(p) + \langle \operatorname{grad} f(p), Dp(0) \rangle_p + \frac{1}{2} \left\langle {}^e\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p + R_2^+(p,q)$$

$$R_2^+(p,q) = \int_0^1 dt \left((1-t) \left\langle {}^e\mathrm{Hess}_{Dp(t)} f(p(t)), Dp(t) \right\rangle_{p(t)} \right) - \frac{1}{2} \left\langle {}^e\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p$$

Taylor's formulæ II

2. $t \mapsto p(t)$ is the exponential geodesic connecting $p = p(0)$ to $q = p(1)$.

$$f(q) = f(p) + \langle \operatorname{grad} f(p), Dp(0) \rangle_p + \frac{1}{2} \left\langle {}^m\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p + R_2^-(p,q)$$

$$R_2^-(p,q) = \int_0^1 dt \left((1-t) \left\langle {}^m\mathrm{Hess}_{Dp(t)} f(p(t)), Dp(t) \right\rangle_{p(t)} \right) - \frac{1}{2} \left\langle {}^m\mathrm{Hess}_{Dp(0)} f(p), Dp(0) \right\rangle_p$$

PART III

Information Geometry of the Gaussian space

- B. Lods and G. Pistone. Information geometry formalism for the spatially homogeneous Boltzmann equation. *Entropy*, 17(6):4323–4363, 2015
- Unpublished paper in progress (2016)

Gaussian space

- We consider $\mathcal{E}(M)$ with

$$M(x) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{|x|^2}{2}\right), \qquad x \in \mathbb{R}^n$$

- The spaces are denoted $L^{(\cosh -1)}(M)$ and $L^{(\cosh -1)_*}(M)$ with conjugate Young functions

$$(\cosh -1)(x) = \cosh x - 1,$$
$$(\cosh -1)_*(y) = y \log\left(y + \sqrt{1+y^2}\right) - \sqrt{1+y^2} - 1,$$

- The space $L^{(\cosh -1)_*}(M)$ is separable with dual space $L^{(\cosh -1)}(M)$ because

$$(\cosh -1)_*(ay) = \int_0^{ay} \frac{ay - t}{\sqrt{1+t^2}}\, dt \leq \max(1, a^2)(\cosh -1)_*(y).$$

Notable functions in $L^{(\cosh -1)}(M)$ and $L^{(\cosh -1)_*}(M)$

- The general inclusions hold, if $1 < a < \infty$,

$$L^\infty(M) \subset L^{(\cosh -1)}(M) \subset L^a(M) \subset L^{(\cosh -1)_*}(M) \subset L^1(M)$$

- Local inclusion holds, if $1 \leq a < \infty$, $\Omega_R = \{x \in \mathbb{R}^n | |x| < R\}$,

$$L^{(\cosh -1)}(M) \subset L^a(\Omega_R)$$

- The Orlicz space $L^{(\cosh -1)}(M)$ contains all polynomials with degree up to 2 and all functions which are bounded by such a polynomial.
- The Orlicz space $L^{(\cosh -1)_*}(M)$ contains all random variables $f \colon \mathbb{R}^d \to \mathbb{R}$ which are bounded by a polynomial, in particular, all polynomials.
- As $\log M \in L^{(\cosh -1)}(M)$ and $p \in \mathcal{E}(M)$ then $\log p \in L^{(\cosh -1)}(M)$, hence $\log p \in L^{(\cosh -1)}(q)$ for all $q \in \mathcal{E}(M)$. The entropy function $H \colon \mathcal{E}(M) \ni p \mapsto -\mathrm{E}_p[\log p]$ is diffentiable with statistical gradient $\operatorname{grad} H(p) = -\log p + H(p)$. The gradien flow trajectories are Gibbs models.

$C_c^\infty(\mathbb{R}^n)$ is boudedly a.s. dense

- For each $f \in L^{(\cosh -1)_*}(M)$ there exist a nonnegative function $h \in L^{(\cosh -1)_*}(M)$ and a sequence $f_n \in C_c(\mathbb{R}^n)$ (respectively $C_c^\infty(\mathbb{R}^n)$) with $|f_n| \leq h$, $n = 1, 2, \dots$, such that $\lim_{n\to\infty} f_n = f$ a.s.
- For each $f \in L^{(\cosh -1)}(M)$ there exist a nonnegative function $h \in L^{(\cosh -1)}(M)$ and a sequence $f_n \in C_c(\mathbb{R}^n)$ (respectively $C_c^\infty(\mathbb{R}^n)$) with $|f_n| \leq h$, $n = 1, 2, \dots$, such that $\lim_{n\to\infty} f_n = f$ a.s.
- $C_c^\infty(\mathbb{R})$ is dense in $L^{(\cosh -1)_*}(M)$ and it is weakly*-dense in $L^{(\cosh -1)}(M)$.

Proof

Let $\mathcal{L}$ be a maximal subset of $L^{(\cosh -1)_*}(M)$, respectively $L^{(\cosh -1)}(M)$, such that the property is true. $\mathcal{L}$ is a vector space, contains the constant functions, is closed for $\wedge$, contains $C_c(\mathbb{R}^n)$. By the monotone class theorem, $\mathcal{L}$ contains all measurable functions that are bounded by an element of $L^{(\cosh -1)_*}(M)$, respectively $L^{(\cosh -1)}(M)$.

Remarks

- If $f \in L^{(\cosh -1)_*}(M)$ and $g \in L^{(\cosh -1)}(M)$ there exists sequences $f_n, g_n \in \mathbb{C}_c^\infty(\mathbb{R}^n)$, $n = 1, 2, \dots,$ such that $f_n g_n \to uv$ in $L^1(M)$
- If $f, h \in L^{(\cosh -1)}(M)$ and $C_c^\infty(\mathbb{R}^n)$ with $|f_n| \le h$, $n = 1, 2, \dots,$ and $\lim_{n\to\infty} f_n = f$ a.s., then $\mathrm{e}^{f_n} \in \mathbb{C}^\infty(\mathbb{R}^n) \cap L^{(\cosh -1)_*}(M)$ and $\lim_{n\to\infty} \mathrm{e}^{f_n} = f$ a.s.
- Let $1 \le a < \infty$. The mapping $g \mapsto gM^{\frac{1}{a}}$ is an isometry of $L^a(M)$ onto $L^a(\mathbb{R}^n)$. As a consequence, for each $f \in L^1(\mathbb{R}^n)$ and each $g \in L^a(M)$ we have $\left\| \left[f * (gM^{\frac{1}{a}}) \right] M^{-\frac{1}{a}} \right\|_{L^a(M)} \le \|f\|_{L^1(\mathbb{R}^n)} \|g\|_{L^a(M)}$.
- The mapping $g \mapsto \operatorname{sign}(g) (\cosh -1)_*^{-1}(M(\cosh -1)_*(g))$ is a surjection of $L^{(\cosh -1)_*}(\mathbb{R}^n)$ onto $L^{(\cosh -1)_*}(M)$ with inverse $h \mapsto \operatorname{sign}(h) (\cosh -1)_*^{-1}(M^{-1}(\cosh -1)_*(f))$. It is surjective from unit vectors (for the Luxenbourg norm) onto unit vectors.

Orlicz-Sobolev with weight

- The O-S spaces with weight M are the vector spaces

$$W^1_{\cosh -1}(M) = \left\{ f \in L^{(\cosh -1)}(M) \middle| \partial_j f \in L^{(\cosh -1)}(M), j = 1, \dots, n \right\}$$

$$W^1_{(\cosh -1)_*}(M) = \left\{ f \in L^{(\cosh -1)_*}(M) \middle| \partial_j f \in L^{(\cosh -1)_*}(M), j = 1, \dots, n \right\}$$

where ∂_j is the derivative in the sense of distributions.

- Both are Banach spaces with the norm of the graph

$$\|f\|_{W^1_{(\cosh -1)}(M)} = \|f\|_{L^{(\cosh -1)}(M)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(M)}$$

$$\|f\|_{W^1_{(\cosh -1)_*}(M)} = \|f\|_{L^{(\cosh -1)}(M)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(M)}$$

- J. Musielak. *Orlicz spaces and modular spaces*, volume 1034 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1983, §10

Remarks

- As $\phi \in C_c^\infty(\mathbb{R}^n)$ implies $\phi M \in C_c^\infty(\mathbb{R}^n)$, for each $f \in W^1_{(\cosh -1)*}(M)$ we have

$$\langle \partial_j f, \phi \rangle_M = \langle \partial_j f, \phi M \rangle = - \langle f, M \partial_j \phi - X_j M \phi \rangle = \langle f, M(X_j - \partial_j)\phi \rangle = \langle f, (X_j - \partial_j)\phi \rangle_M .$$

  We want the operator $\delta_j = X_j - \partial_j$ exteded to $L^{(\cosh -1)_*}(M)$.

- $W^1_{(\cosh -1)}(M) \subset W^1_{(\cosh -1)}(\Omega_R) \subset W^{1,p}(\Omega_R)$, $p \geq 1$
- $W^1_{(\cosh -1)_*}(M) \subset W^1_{(\cosh -1)_*}(\Omega_R) \subset W^{1,1}(\Omega_R)$.
- Each $u \in W^1_{(\cosh -1)}(M)$ is a.s. continuous and Hölder of all orders on each $\overline{\Omega}_R$

- H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

Compositions and operators

Differentiable densities

1. If $u \in \mathcal{S}_M$ and $f_1, \dots, f_m \in L^{(\cosh -1)}(M)$, then $f_1 \cdots f_m \mathrm{e}^{u - K_M(u)} \in L^{(\cosh -1)_*}(M)$.
2. If $u \in \mathcal{S}_M \cap W^1_{(\cosh -1)}(M)$ and $f \in W^1_{\cosh -1}(M)$, then $f\mathrm{e}^{u-K_M(u)} \in W^1_{(\cosh -1)_*}(M) \cap C^1(\mathbb{R}^n)$.

Proof.

The equality $\mathrm{e}^{u-K_M(u)} = \partial_i(f\mathrm{e}^{u-K_M(u)})$ in the sense of distributions is checked in pointwise approximation by $C_c^\infty(\mathbb{R}^n)$ functions. □

The δ_j operator

- The injection $W^1_{(\cosh -1)_*}(M) \ni f \mapsto X_j f \in L^{(\cosh -1)_*}(M)$, where X_j is the multiplication operator by the j-th coordinate x_j, is defined and continuous.
- If $f \in W^1_{(\cosh -1)_*}(M)$ and $g \in W^1_{\cosh -1}(M)$, then

$$\langle f, \partial_j g \rangle_M = \langle X_j f - \partial_j f, g \rangle_M = \langle \delta_j f, g \rangle_M$$

Exponential family modeled on $W^1_{(\cosh -1)}(M)$

- If we restrict the exponential family $\mathcal{E}(M)$ to $W^1_{\cosh -1}(M)$,
$$W_M = W^1_{\cosh -1}(M) \cap B_M = \left\{ U \in W^1_{\cosh -1}(M) \middle| \mathrm{E}_M[U] = 0 \right\}$$
we obtain the following non parametric exponential family
$$\mathcal{E}_1(M) = \left\{ \mathrm{e}^{U - K_M(U)} \cdot M \middle| U \in W^1_{\cosh -1}(M) \cap \mathcal{S}_M \right\}$$
- Because of $W^1_{\cosh -1}(M) \hookrightarrow L^{\cosh -1}(M)$ the set $W^1_{\cosh -1}(M) \cap \mathcal{S}M$ is open in W_M and the cumulant functional $K_M : W^1_{\cosh -1}(M) \cap \mathcal{S}M \to \mathbb{R}$ is convex and differentiable.
- Every feature of the exponential manifold carries over to this case. In particular, we can define the spaces
$$W_f = W^1_{\cosh -1}(M) \cap B_M = \left\{ U \in W^1_{\cosh -1}(M) \middle| \mathrm{E}_f[U] = 0 \right\}, \quad f \in \mathcal{E}_1(M)$$
to be models for the tangent spaces of $\mathcal{E}_1(M)$. The e-transport acts on these spaces
$$\mathbb{U}^g_f \colon W_f \ni U \mapsto U - \mathrm{E}_g[U] \in W_g \ ,$$
so that we can define the statistical bundle to be
$$S\,\mathcal{E}_1(M) = \{(g, V) | g \in \mathcal{E}_1(M), V \in W_f\}$$

Application: Hyvärinen divergence

- For each $f, g \in \mathcal{E}_1(M)$ the Hyvärinen divergence is

$$\mathsf{DH}\left(g \| f\right) = \mathsf{E}_g\left[\left|\nabla \log f - \nabla \log g\right|^2\right].$$

- The expression in the chart centered at M is

$$\mathsf{DH}_M(v \| u) := \mathsf{DH}\left(\mathrm{e}_M(v) \| \mathrm{e}_M(u)\right) = \mathsf{E}_M\left[\left|\nabla u - \nabla v\right|^2 \mathrm{e}^{v - K_M(v)}\right],$$

where $f = \mathrm{e}_M(u)$, $g = \mathrm{e}_M(v)$.

- $\operatorname{grad}(f \mapsto \mathsf{DH}\left(g \| f\right)) = -2\nabla \log g \cdot \nabla \log \frac{f}{g} - 2\Delta \log \frac{f}{g}$
- $\operatorname{grad}(g \mapsto \mathsf{DH}\left(f \| g\right)) = 2\nabla \log g \cdot \nabla \log \frac{f}{g} + 2\Delta \log \frac{f}{g} + \mathsf{DH}\left(f \| g\right)$

Elliptic operator (Brigo & Pistone 2016)

- Elliptic operator as section of the tangent bundle is

$$\mathcal{A}p(x) = p(x)^{-1} \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial}{\partial x_j} p(x) \right), \quad x \in \mathbb{R}^d \ .$$

- The expression in the statistical bundle is

$$U \mapsto \widehat{\mathcal{A}}_M(U) = \mathrm{e}^{U - K_M(U)} \mathcal{A}(\mathrm{e}^{U-K_M(U)} \cdot M) = \\ \frac{\mathrm{e}^{U-K_M(U)}}{\mathrm{e}^{U-K_M(U)} \cdot M} \mathcal{A}(\mathrm{e}^{U-K_M(U)} \cdot M) = M^{-1} \mathcal{L}^*(\mathrm{e}^{U-K_M(U)} \cdot M)$$

- Computation gives

$$M^{-1}\mathcal{L}^*(\mathrm{e}^{U-K_M(U)} \cdot M) = \\ \mathrm{e}^{U-K_M(U)} \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \left[a_{ij}(x) \left(\frac{\partial}{\partial x_j} U(x) - x_j \right) \right] p(x) + \\ \mathrm{e}^{U-K_M(U)} \sum_{i,j=1}^{d} a_{ij}(x) \left(\frac{\partial}{\partial x_i} U(x) - x_i \right) \left(\frac{\partial}{\partial x_j} U(x) - x_j \right) p(x).$$