IGAIA IV Information Geometry and its Applications IV

The Orlicz-Sobolev-Gauss Exponential Manifold

Giovanni Pistone
www.giannidiorestino.it

DE CASTRO STATISTICS Collegio Carlo Alberto

Liblice June 13 2016

My four parts

1. Amari's Information Geometry when the state space is not finite and the model is not parametric
2. An example: computing the Wasserstein's distance
3. Gauss-Orlicz-Sobolev model spaces
4. Second order geometry

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

Part I

Amari's Information Geometry when the state space is not finite and the model in not parametric

In IG the velocity is the score

- $\theta \mapsto p_\theta$ is a *curve*
- The score $\theta \mapsto \frac{d}{d\theta} \log p_\theta$ is an estimating function because $\mathbb{E}_\theta \left[\frac{d}{d\theta} \log p_\theta \right] = 0$
- Fisher-Rao computation:

$$\begin{aligned} \frac{d}{d\theta} \mathbb{E}_\theta [U] &= \int U(x) \frac{d}{d\theta} p(x;\theta) \; \mu(dx) \qquad p(x;\theta) > 0 \\ &= \int U(x) \frac{d}{d\theta} \log p(x;\theta) \; p(x;\theta) \; \mu(dx) = \\ &= \left\langle U - \mathbb{E}_\theta [U] \, , \frac{d}{d\theta} \log p_\theta \right\rangle_\theta \end{aligned}$$

- $U - \mathbb{E}_\theta [U]$ is the statistical gradient of $\theta \mapsto \mathbb{E}_\theta [U]$.

- Cf recent work by Ay, Jost, Lê, Schwachhöfer on *measure models*

IG is the geometry of the statistical bundle

- $\mathcal{P}$ is a set of probabilities on a given sample space $(\Omega, \mathcal{F})$
- For each $p \in \mathcal{P}$, $B_p \hookrightarrow L_0^1(p)$
- A *statistical bundle* is

$$T\mathcal{P} = \{(p, U) | p \in \mathcal{P}, U \in B_p\}$$

- We expect the fibers B_p to be *isomorphic* and express a tangent space at $p \in \mathcal{P}$
- A chart at p $\sigma_p : (q, V) \mapsto (s_p(q), \dot{s}_p(V)) \in B_p \times B_p$

- S.-i. Amari and M. Kumon. Estimation in the presence of infinitely many nuisance parameters—geometry of estimating functions.

  Ann. Statist., 16(3):1044–1068, 1988

- P. Gibilisco and G. Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry.

  IDAQP, 1(2):325–347, 1998

- Cf Otto, cf Lê

Fibers: $B_p = L^{(\cosh -1)}(p)$

- The exponential space $L^{(\cosh -1)}(p)$ and the mixture space $L^{(\cosh -1)_*}(p)$ are the Orlicz spaces respectively defined by the conjugate Young functions

$$(\cosh -1)(x) = \cosh x - 1$$

- with

$$xy \leqslant (\cosh -1)(x) + (\cosh -1)_*(y)$$

- The closed unit balls of the exponential and mixture space are, respectively,

$$\left\{ f \middle| \|f\|_{L^{(\cosh -1)}(p)} \leqslant 1 \right\} = \left\{ f \middle| \int (\cosh -1)(f(x))\ p(x)dx \leqslant 1 \right\}$$

$$\left\{ g \middle| \|\leqslant\|_{L^{(\cosh -1)_*}(p)} 1 \right\} = \left\{ g \middle| \int (\cosh -1)_*(g(x))\ p(x)dx \leqslant 1 \right\}.$$

- $B_p = \left\{ U \in L^{(\cosh -1)}(p) \middle| \mathbb{E}_p[U] = 0 \right\}$ is the dual of ${}^*B_p = \left\{ V \in L^{(\cosh -1)_*}(p) \middle| \mathbb{E}_p[V] = 0 \right\}$

B_p is the space of scores

B_p is *exactly* the space of scores of Gibbs model through p

Theorem

1. $U \in B_p$ *iff* $\mathbb{E}_p[U] = 0$ *and* $\mathbb{E}_p[(\cosh -1)(\rho U)] < \infty$ *for some* $\rho > 0$
2. $U \in B_p$ *iff* $\mathbb{E}_p[U] = 0$ *and the moment generating function* $\alpha \mapsto \mathbb{E}_p\left[\mathrm{e}^{\theta U}\right]$ *is finite in a neighbourhood of 0*
3. *The Gibbs model* $\theta \mapsto \frac{\mathrm{e}^{\theta U}}{\mathbb{E}_p[\mathrm{e}^{\theta U}]}$ *is defined in a neighborhood of 0 and* $\frac{d}{d\theta}\mathbb{E}_p\left[\mathrm{e}^{\theta U}\right]\big|_{\theta=0} = 0.$
4. *The score of the Gibbs model at 0 is* $\frac{d}{d\theta}\log p_\theta\big|_{\theta=0} = U$

This set-up applies to the set $\mathcal{P}_>$ of strictly positive densities.

- 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

Isomorphism of the $L^{(\cosh -1)}(p)$ spaces

Theorem

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

Proof.

Assume $U \in L^{(\cosh -1)}(p)$ and consider the restrictions to the axes of the convex function

$$(s,\theta) \mapsto \int \mathrm{e}^{sU} p^{1-\theta} q^{\theta}\, d\mu = \int \exp\left(sU + \theta \log \frac{q}{p}\right) p\, d\mu$$

□

- 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

- A. Cena. *Geometric structures on the non-parametric statistical manifold.*

  PhD thesis, Dottorato in Matematica, Università di Milano, 2002

Portmanteau theorem

Theorem

The following statements are equivalent for $p, q \in \mathcal{P}_>$*:*

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

- 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, 22(3):1431–1447, 2016

Maximal exponential family

- For each $p \in \mathcal{P}_>$, the *moment generating functional* is the positive lower-semi-continuous convex function $G_p \colon B_p \ni U \mapsto \mathbb{E}_p\left[\mathrm{e}^U\right]$ and
- the *cumulant generating functional* is the non-negative lower semicontinuous convex function $K_p = \log G_p$.
- The interior of the proper domain

$$\mathcal{S}_p = \left\{ U \in L^{(\cosh -1)}(p) \middle| G_p(U) < +\infty \right\}^\circ$$

  is an open convex set containing the open unit ball of $L^{(\cosh -1)}(p)$.
- For each $p \in \mathcal{P}_>$, the *maximal exponential family* at p is

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

From now on the maximal exponential family of interest is the family of the Maxwell density on $\mathbb{R}^n$, $\mathcal{E}(M)$

e-chart at $p \in \mathcal{E}(M)$

- For each $p \in \mathcal{E}(M)$ we define a chart $s_p \colon \mathcal{E}(M) \to \mathcal{S}_p \subset B_p$.
- The chart is defined by

$$s_p(q) \mapsto \log\left(\frac{q}{p}\right) + D(p\|q) = \log\left(\frac{q}{p}\right) - \mathbb{E}_p\left[\log\left(\frac{q}{p}\right)\right]$$

- The inverse of the chart $e_p^{-1} = s_p \colon \mathcal{S}_p \to \mathcal{E}(M)$ is

$$e_p(U) = \exp\left(U - K_p(U)\right) \cdot p$$

- $\{s_p | p \in \mathcal{E}(M)\}$ is an affine atlas on $\mathcal{E}(M)$ that defines the *exponential manifold*
- The information closure of any $\mathcal{E}(M)$ is $\mathcal{P}_{\geqslant}$. The reverse information closure of any $\mathcal{E}(M)$ is $\mathcal{P}_{>}$.

- I. Csiszár and F. Matúš. Information projections revisited.
  IEEE Trans. Inform. Theory, 49(6):1474–1490, 2003
- 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

e-chart at $(p, U) \in T\mathcal{E}(M)$

- A curve $t \mapsto p(t)$, $p(0) = p$ in the exponential manifold $\mathcal{E}(M)$ is expressed in the chart s_p as $p(t) = \mathrm{e}^{U(t) - K_p(U(t))} \cdot p$.
- The expression of the velocity at $t = 0$ is $\dot{U}(0) = \frac{d}{dt} \log p(t)\big|_{t=0}$
- It follows that the exponential bundle

$$T\mathcal{E}(M) = \{(p, U) | p \in \mathcal{E}(M), U \in B_p\}$$

  is the expression of the tangent bundle of the exponential manifold
- The transition map $s_{p_2} \circ e_{p_1} : \mathcal{S}_{p_1} \to \mathcal{S}_{p_2}$ is affine with derivative ${}^{\mathrm{e}}\mathbb{U}_{p_1}^{p_2} : B_{p_1} \to B_{p_2}$ given by ${}^{\mathrm{e}}\mathbb{U}_{p_1}^{p_2} U = U - \mathbb{E}_{p_2}[U]$
- We define an atlas of charts on $T\mathcal{E}(M)$ by

$$\sigma_p(q, V) = \left(s_p(q), {}^{\mathrm{e}}\mathbb{U}_q^p V\right)$$

- 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

Cumulant functional

- The r-divergence $q \mapsto \mathbf{D}(p \,\|\, q)$ is represented in the chart centered at p by $\mathbf{D}(p \,\|\, e_p(U)) = K_p(U) = \log \mathbb{E}_p\left[\mathrm{e}^U\right]$.
- $K_p : B_p \to \mathbb{R}_{\geqslant} \cup \{+\infty\}$ is convex and its proper domain contains the open unit ball of B_p. It 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:

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

- G. Pistone and M. Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations.

  Bernoulli, 5(4):721–760, August 1999

- A. Cena and G. Pistone. Exponential statistical manifold.

  Ann. Inst. Statist. Math., 59(1):27–56, 2007

Pre-dual statistical bundle

- Recall $L^{(\cosh -1)_*}(M)$ is the pre-dual of $L^{(\cosh -1)}(M)$
- Define the pre-dual statistical bundle with fibers ${}^*B_p = \left\{ V \in L^{(\cosh -1)_*}(M) \middle| \mathbb{E}_p[V] = 0 \right\}$.
- Compute the adjoint of the transport ${}^{\mathrm{e}}\mathbb{U}_p^q$. For $U \in B_p$ and $V \in {}^*B_q$,

$$\begin{aligned} \left\langle {}^{\mathrm{e}}\mathbb{U}_p^q U, V \right\rangle_q &= \langle U - \mathbb{E}_q[U], V \rangle_q = \mathbb{E}_q[UV] \\ &= \mathbb{E}_p\left[\frac{q}{p}UV\right] = \left\langle U, \frac{q}{p}V \right\rangle_p = \langle U, {}^{\mathrm{m}}\mathbb{U}_q^p V \rangle_p \end{aligned}$$

- Define the charts on ${}^*T\mathcal{E}(M)$ by

$$\sigma_p^*(q, W) = \left(s_p(q), {}^{\mathrm{m}}\mathbb{U}_q^p W\right)$$

Statistical gradient

- The score of the curve $t \mapsto p(t)$ is a curve in the statistical bundle $t \mapsto (p(t), Dp(t)) \in T\mathcal{E}(M)$ such that for all $X \in L^{(\cosh -1)_*}(M)$ it holds

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

- $Dp(t)$ is the expression in the exponential atlas of the velocity

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

- The statistical gradient of $F : \mathcal{E}(M) \to \mathbb{R}$ is a section of the pre-dual statistical bundle ${}^*T\mathcal{E}(M)$, $p \mapsto (p, \operatorname{grad} F(p)) \in {}^*T\mathcal{E}(p)$ such that for each regular curve

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

- L. Malagò, M. Matteucci, and G. Pistone. Towards the geometry of estimation of distribution algorithms based on the exponential family.

  In *Proceedings of the 11th workshop on Foundations of genetic algorithms*, FOGA '11, pages 230–242, New York, NY, USA, 2011. ACM

- G. Pistone. Examples of the application of nonparametric information geometry to statistical physics.

  Entropy, 15(10):4042–4065, 2013

Part II
An example: computing the Wasserstein distance

- This is an example of the use of the formalism. There is considerable literature, e.g. F. Otto
- Unpublished talk at the workshop *Computational information geometry for image and signal processing* Sep 21st25th, 2015 at ICMS, Edinburgh. Finite state space.
- Unpublished work in progress with Luigi Malagò.

Transport plan

$(\mathbb{R}^{2n}, M_{2n}) = (\mathbb{R}^n, M_n) \otimes (\mathbb{R}^n, M_n)$ with projection X and Y

- The marginalization mapping

$$\mathcal{E}(M_{2n}) \ni \gamma \mapsto (\gamma_1, \gamma_2) \in \mathcal{E}(M_n) \times \mathcal{E}(M)_n$$

has fibers

$$\Gamma(\mu_1, \mu_2) = \{\gamma \in \mathcal{E}(M_{2n}) | X_{\#}\gamma = \mu_1, Y_{\#}\gamma = \mu_2\}$$

which are convex subsets

- If $t \mapsto \gamma(t) \in \Gamma(\mu_1, \mu_2)$, then

$$0 = \frac{d}{dt}\mathbb{E}_{\gamma(t)}[f \circ X] = \langle f \circ X - \mathbb{E}_{\mu_1}[f], D\gamma(t)\rangle_{\gamma(t)} = \left\langle f \circ X - \mathbb{E}_{\mu_1}[f], \mathbb{E}_{\gamma(t)}(D\gamma(t)|X)\right\rangle_{\gamma(t)}$$

-

$$\mathbb{E}_{\gamma(t)}(D\gamma(t)|X) = 0, \qquad \mathbb{E}_{\gamma(t)}(D\gamma(t)|Y) = 0, \qquad D\gamma(t) \in B_{\gamma(t)}$$

Splitting of $T\Gamma(\mu_1, \mu_2)$

- Consider subspaces of the ANOVA

$$B_{\gamma,1} = \{f \circ X | f \in B_{\gamma_1}\},$$
$$B_{\gamma,2} = \{f \circ Y | f \in B_{\gamma_2}\},$$
$${}^{*}B_{\gamma,12} = (B_{\gamma,0} + B_{\gamma,1} + B_{\gamma,2})^{\perp}$$

- For $U \in B_\gamma$, let $U = U_1 + U_2 + U_{12}$ be a splitting. Then

$$\mathbb{E}_\gamma (U - (U_1 + U_2)|X) = 0$$
$$\mathbb{E}_\gamma (U - (U_1 + U_2)|Y) = 0$$

- or in terms of transport

$$\mathbb{E}_{M_{2n}} \left({}^{\mathrm{m}}\mathbb{U}_\gamma^{M_{2n}} U \middle| X\right) - \mathbb{E}_{M_{2n}} \left(\frac{\gamma}{M_{2n}} \middle| X\right) U_1 + \mathbb{E}_{M_{2n}} \left({}^{\mathrm{m}}\mathbb{U}_\gamma^{M_{2n}} U_2 \middle| X\right) = 0$$
$$\mathbb{E}_{M_{2n}} \left({}^{\mathrm{m}}\mathbb{U}_\gamma^{M_{2n}} U \middle| X\right) - \mathbb{E}_{M_{2n}} \left({}^{\mathrm{m}}\mathbb{U}_\gamma^{M_{2n}} U_1 \middle| X\right) + \mathbb{E}_{M_{2n}} \left(\frac{\gamma}{M_{2n}} \middle| X\right) U_2 = 0$$

Gradient flow

- Given a cost function $w\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, the *expected cost function* is $W\colon \mathcal{E}(M_{2n}) \ni \gamma \mapsto \mathbb{E}_\gamma[w]$. Then the function W restricted to the open plan $\Gamma(\mu_1, \mu_2) \cap \mathcal{E}(M_{2n})$ has statistical gradient obtained by the projection of the unconstrained gradient $w - \mathbb{E}_\gamma[w]$ onto the the space of interactions $B_{\gamma,12}$

$$\operatorname{grad} W(\gamma)\colon \gamma \mapsto w - \mathbb{E}_\gamma[w] - w_{1,\gamma} - w_{2,\gamma}$$

- The gradient flow equation is

$$D\gamma(t) = -\left(w - \mathbb{E}_\gamma[w] - w_{1,\gamma} - w_{2,\gamma}\right) \ .$$

- Any solution $t \mapsto \gamma(t)$ of the gradient flow converges to a measure $\gamma* = \lim_{t\to\infty} \gamma(t)$, in $\Gamma(\mu_1, \mu_2)$ (but not in $\mathcal{E}(M_{2n})$) such that

$$\mathbb{E}_{\gamma*}[w] = \min\{\mathbb{E}_\gamma[w] | \gamma \in \Gamma(\mu_1, \mu_2)\}$$

Part III
Gauss-Orlicz-Sobolev model spaces

- M. R. Grasselli. Dual connections in nonparametric classical information geometry.
  Ann. Inst. Statist. Math., 62(5):873–896, 2010
- 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, 2016
- D. Brigo and G. Pistone. Eigenfunctions based maximum likelihood estimation of the fokker planck equation and hellinger projection.
  submitted, 2016
- Luigi Montrucchio and GP. Unpublished working paper (2016) based on N. Newton deformed logarithm

Inclusions

1. If $1 < a < \infty$,

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

2. Local inclusions hold, if $1 \leqslant a < \infty$, $\Omega_R = \{x \in \mathbb{R}^n | |x| < R\}$,

$$L^{(\cosh -1)}(M) \hookrightarrow L^{a}(\Omega_R), \quad L^{(\cosh -1)_*}(M) \hookrightarrow L^{1}(\Omega_R)$$

3. The Orlicz space $L^{(\cosh -1)}(M)$ contains all functions $f \in C^2(\mathbb{R}^n; \mathbb{R})$ whose Hessian is uniformely bounded in operator's norm. In particular, it contains all polynomials with degree up to 2 and, moreover, all functions which are bounded by such a polynomial.

4. The Orlicz space $L^{(\cosh -1)_*}(M)$ contains all random variables $f : \mathbb{R}^d \to \mathbb{R}$ which are bounded by a polynomial, in particular, all polynomials.

Pointwise density

The space $L^{(\cosh -1)}(M)$ is *not* separable nor reflexive. However, we have the following monotone class theorem

- For each $f \in L^{(\cosh -1)}(M)$ there exist a nonnegative function $h \in L^{(\cosh -1)}(M)$ and a sequence $f_n \in C_0(\mathbb{R}^n)$ with $|f_n| \leqslant h$, $n = 1, 2, \ldots$, such that $\lim_{n\to\infty} f_n = f$ a.s..
- The space $C_0(\mathbb{R}^n)$ is strongly dense in $L^{(\cosh -1)_*}(M)$ and it is weakly*-dense in $L^{(\cosh -1)}(M)$.
- For each $f \in L^{(\cosh -1)}(M)$ there exist a nonnegative function $h \in L^{(\cosh -1)}(M)$ and a sequence $\phi_n \in C_0^\infty(\mathbb{R}^n)$ with $|f_n| \leqslant h$, $n = 1, 2, \ldots$, such that $\lim_{n\to\infty} \phi_n = f$ a.s..
- The space $C_0^\infty(\mathbb{R}^n)$ is strongly dense in $L^{(\cosh -1)_*}(M)$ and it is weakly*-dense in $L^{(\cosh -1)}(M)$.

Orlicz class

Definition

We define the *exponential (Orlicz) class*, $C_0^{(\cosh -1)}(M)$, to be the closure of $C_0(\mathbb{R}^n)$ in the exponential (Orlicz) space $L^{(\cosh -1)}(M)$.

Theorem

Assume $f \in L^{(\cosh -1)}(M)$ *and write* $f_R(x) = f(x)(|x| > R)$. *The following conditions are equivalent:*

1. *The real function* $\rho \mapsto \int (\cosh -1)(\rho f(x))\, M(x)dx$ *is finite for all* $\rho > 0$.
2. $f \in C^{\cosh -1}(M)$.
3. $\lim_{R\to\infty} \|f_R\|_{L^{(\cosh -1)}(M)} = 0$.

For example $(x \mapsto \|x\|^2) \in L^{(\cosh -1)}(M) \setminus C_0^{(\cosh -1)}(M)$

Translation models

- We look for statistical models induced by the gromrtry of the state space. E.g. the n-dimensional model defined by the translation of $p = \mathrm{e}^{U - K_M(U)} \cdot M \in \mathcal{E}(M)$

$$p(x;h) = p(x-h) = \mathrm{e}^{U(x-h) - K_M(U)} \mathrm{e}^{h \cdot x - \frac{|h|^2}{2}} \cdot M$$

- $$\mathbb{E}_M [U(X-h) + h \cdot X] = \int U(x-h) M(x)\, dx = \int U(x) \mathrm{e}^{-h \cdot x - \frac{|h|^2}{2}}\, M(x) dx = \mathbb{E}_M \left[U \mathrm{e}^{-h \cdot X \frac{|h|^2}{2}} \right]$$

- $$p(x;h) = p(x-h) = \exp\left(U_h - K_M(U_h)\right) \cdot M$$

  with $U_h = \tau_h U + h \cdot X - \mathbb{E}_M [\tau_h U] \in B_M$ and $K_M(\tau_h U) = K_M(U) - \frac{1}{2} |h|^2$

Translations in $L^{(\cosh -1)}(M)$

- For each $h \in \mathbb{R}^n$, the mapping $f \mapsto \tau_h f$ is linear and bounded from $L^{(\cosh -1)}(M)$ to itself and $\|\tau_h f\|_{L^{(\cosh -1)}(M)} \leqslant 2 \|f\|_{L^{(\cosh -1)}(M)}$ if $|h| \leqslant \sqrt{\log 2}$.
- For each $f \in L^{(\cosh -1)}(M)$ and $h \in \mathbb{R}^n$ we have $\tau_h f \in L^{(\cosh -1)}(M)$. For all $g \in L^{(\cosh -1)_*}(M)$ we have

$$\langle \tau_h f, g \rangle_M = \langle f, \tau_h^* g \rangle_M , \quad \tau_h^* g(x) = \mathrm{e}^{-h\cdot x - \frac{|h|^2}{2}} \tau_{-h} g(x) ,$$

  and $|h| \leqslant \sqrt{2}$ implies $\|\tau_h^* g\|_{L^{(\cosh -1)_*}(M)} \leqslant 4 \|g\|_{L^{(\cosh -1)_*}(M)}$. Moreover, $h \mapsto \tau_h^* g$ is continuous in $L^{(\cosh -1)_*}(M)$.
- If $f \in C_0^{(\cosh -1)}(M)$ then $\tau_h f \in C_0^{(\cosh -1)}(M)$, $h \in \mathbb{R}^n$ and the mapping $\mathbb{R}^n \colon h \mapsto \tau_h f$ is continous in $L^{(\cosh -1)}(M)$.

Translation by a probability

- Let
$$\tau_\mu f(x) = \int f(x-y)\ \mu(dy) = f * \mu(x)$$
for $\mu \in \mathcal{P}_{\mathrm{e}}$, namely $\mathbb{E}_M\left[\mathrm{e}^{\frac{1}{2}|h|^2}\right] < \infty$.
- The mapping $f \mapsto \tau_\mu f$ is linear and bounded from $L^{(\cosh -1)}(M)$ to itself. If, moreover, $\int \mathrm{e}^{|h|^2/2}\ \mu(dh) \leqslant \sqrt{2}$, then its norm is bounded by 2.
- If $f \in C_0^{(\cosh -1)}(M)$ then $\tau_\mu f \in C_0^{(\cosh -1)}(M)$. The mapping $\mathcal{P}\colon \mu \mapsto \tau_\mu f$ is continous at δ_0 from the weak convergence to the $L^{(\cosh -1)}(M)$ norm.

Mollifiers

- Let be given a family of mollifiers $\omega_\lambda \in C_0^\infty(\mathbb{R}^n)$. Let $f \in C_0^{(\cosh -1)}(M)$. For each $\lambda > 0$ the function

$$\tau_{\omega_\lambda} f(x) = \int f(x-y)\lambda^{-n}\omega(\lambda^{-1}y)\, dy = f * \omega_\lambda(x)$$

  belongs to $C^\infty(\mathbb{R}^n)$ and $\lim_{\lambda\to 0} f * \omega_\lambda = f$ in $L^{(\cosh -1)}(M)$

- For each $f \in L^{(\cosh -1)}(M)$ there exists a sequence f_n, in $C_0^\infty(\mathbb{R}^n)$, $n \in \mathbb{N}$, and a bound $h \in L^{(\cosh -1)}(M)$ such that $|f_n(x)| \leqslant h(x)$ and $\lim_{n\to\infty} f_n(x) = f(x)$ for all x.

- For each $f \in C_0^{(\cosh -1)}(M)$ there exists a sequence f_n, in $C_0^\infty(\mathbb{R}^n)$, $n \in \mathbb{N}$, with $\lim_{n\to\infty} \|f_n - f\|_{L^{(\cosh -1)}(M)} = 0$.

Differentiable densities

Definition

The Orlicz-Sobolev (O-S) spaces with weight M are

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

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

where ∂_j is the derivative in the sense of distributions. The spaces $W^{1,(\cosh -1)}(M)$ and $W^{1,(\cosh -1)_*}(M)$ are both Banach spaces for the graph norms.

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

- Cf Malliavin Calculus

Inclusions

Theorem

Let $R > 0$ and let Ω_R denote the open sphere of radius R.

1. *We have the embeddings*

$$W^{1,(\cosh -1)}(\mathbb{R}^n) \subset W^{1,(\cosh -1)}(M) \subset W^{1,(\cosh -1)}(\Omega_R) \subset W^{1,p}(\Omega_R), \quad p \geqslant 1.$$

2. *We have the embeddings*

$$W^{1,p}(\mathbb{R}^n) \subset W^{1,(\cosh -1)_*}(\mathbb{R}^n) \subset W^{1,(\cosh -1)_*}(M) \subset W^{1,(\cosh -1)_*}(\Omega_R) \subset W^{1,1}(\Omega_R), \quad p > 1.$$

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

Directional derivative

Theorem

- *For each* $f \in W^{1,(\cosh -1)}(M)$, *each unit vector* $h \in S^n$, *and all* $t \in \mathbb{R}$, *it holds*

$$f(x+th) - f(x) = t \int_0^1 \sum_{j=1}^{n} \partial_j f(x+sth) h_j \, ds \ .$$

  Moreover, $|t| \leqslant \sqrt{2}$ *implies*

$$\|f(x+th) - f(x)\|_{L^{(\cosh -1)}(M)} \leqslant 2t \, \|\nabla f\|_{L^{(\cosh -1)}(M)} \ ,$$

  especially, $\lim_{t\to 0} \|f(x+th) - f(x)\|_{L^{(\cosh -1)}(M)} = 0$ *uniformely in* h.

- *For each* $f \in W^{1,(\cosh -1)}(M)$ *and each* $g \in L^{(\cosh -1)_*}(M)$, *the mapping* $h \mapsto \langle \tau_h f, g \rangle_M$ *is differentiable. Viceversa, if* $f \in L^{(\cosh -1)}(M)$ *and* $h \mapsto \tau_h f$ *is weakly differentiable, then* $f \in W^{1,(\cosh -1)}(M)$

- *If* $\partial_j f \in C_0^{(\cosh -1)}(M)$, $j = 1, \ldots, n$, *then strong differentiability in* $L^{(\cosh -1)}(M)$ *holds.*

Orlicz-Sobolev class

Definition

The *Orlicz-Sobolev-Gauss exponential class* is

$$C_0^{1,(\cosh -1)}(M) = \left\{ f \in W^{1,(\cosh -1)}(M) \middle| f, \partial_j f \in C_0^{(\cosh -1)}(M), j = 1, \ldots, n \right\}$$

- The translation model is qualified as

$$p_h = \tau_h p, \quad p = \mathrm{e}^{U - K_M(U)} \cdot M, \quad U \in \mathcal{S}_M \cap C_0^{1,(\cosh -1)}(M)$$

- The score in the direction j is $(x_j - te_j) - \partial_j U(x - te_j)$:

$$\frac{\frac{d}{dt} p(x - te_j)}{p(x - te_j)} = \frac{\frac{d}{dt} \mathrm{e}^{U(x - te_j) - K_M(U)} M(x - te_j)}{p(x - te_j)} =$$

$$\frac{-\partial_j U(x - te_j) \mathrm{e}^{U(x - te_j) - K_M(U)} M(x - te_j) + (x_j - te_j) \mathrm{e}^{U(x - te_j) - K_M(U)} M(x - te_j)}{p(x - te_j)} =$$

$$(x_j - te_j) - \partial_j U(x - te_j)$$

Calculus in $C_0^{1,(\cosh -1)}(M)$

Theorem

- *For each* $f \in C_0^{1,(\cosh -1)}(M)$ *the sequence* $f * \omega_n$, $n \in \mathbb{N}$, *belongs to* $C^\infty(\mathbb{R}^n) \cap W^{1,(\cosh -1)}(M)$. *Precisely, for each* n *and* $j = 1, \ldots, n$, *we have the equality* $\partial_j(f * \omega_n) = (\partial_j f) * \omega_n$; *the sequences* $f * \omega_n$, *respectively* $\partial_j f * \omega_n$, $j = 1, \ldots, n$, *converge to* f, *respectively* $\partial_j f$, $j = 1, \ldots, n$, *strongly in* $W^{1,(\cosh -1)_*}(M)$.
- *Same statement is true if* $f \in W^{1,(\cosh -1)_*}(M)$.
- *Let be given* $f \in C_0^{(\cosh -1)}(M)$ *and* $g \in W^{1,(\cosh -1)_*}(M)$. *Then* $fg \in W^{1,1}(M)$ *and* $\partial_j(fg) = \partial_j fg + f\partial_j g$.
- *Let be given* $F \in C^1(\mathbb{R})$ *with* $\|F'\|_\infty < \infty$. *For each* $U \in C_0^{(\cosh -1)}(M)$, *we have* $F \circ U, F' \circ U\partial_j U \in C_0^{(\cosh -1)}(M)$ *and* $\partial_j F \circ U = F' \circ U\partial_j U$, *in particular* $F(U) \in C_0^{1,(\cosh -1)}(M)$.

Exponential family modeled on $C_0^{1,(\cosh -1)}(M)$

- Restrict the exponential family $\mathcal{E}(M)$ to $C_0^{1,(\cosh -1)}(M)$,

$$\mathcal{E}_1(M) = \left\{ \mathrm{e}^{U - K_M(U)} \cdot M \middle| U \in C_0^{1,(\cosh -1)}(M) \cap \mathcal{S}_M \right\}$$

- Because of $C_0^{1,(\cosh -1)}(M) \hookrightarrow L^{\cosh -1}(M)$ the domain $C_0^{1,(\cosh -1)}(M) \cap \mathcal{S}_M$ is open and the cumulant functional $K_M : C_0^{1,(\cosh -1)}(M) \cap \mathcal{S}_M \to \mathbb{R}$ remains convex and differentiable.
- Every feature of the exponential manifold carries over to this case. Define $B_1(p) = B_p \cap C_0^{1,(\cosh -1)}(M)$ to be models for the tangent spaces of $\mathcal{E}_1(M)$. The e-transport acts on these spaces

$$^{\mathrm{e}}\mathbb{U}_f^g : B_1(f) \ni U \mapsto U - \mathbb{E}_g[U] \in B_1(g) \ ,$$

so that we can define the statistical bundle to be

$$T\mathcal{E}_1(M) = \{(g, V) | g \in \mathcal{E}_1(M), V \in B_1(g)\}$$

and take as charts the restrictions of the charts defined on $T\mathcal{E}(M)$.

Application: Hyvärinen divergence

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

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

- The expression in the chart centered at M is

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

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

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

Example: Elliptic operator

- 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

$$M^{-1} \mathcal{L}^*(\mathrm{e}^{U - K_M(U)} \cdot M) =$$

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

$$\mathrm{e}^{U - K_M(U)} \sum_{i,j=1}^{d} a_{ij}(x) \left(\frac{\partial}{\partial x_i} U(x) - x_i \right) \left(\frac{\partial}{\partial x_j} U(x) - x_j \right) p(x).$$

Part IV
Second order geometry

- L. Malagò and G. Pistone. Combinatorial optimization with information geometry: Newton method. *Entropy*, 16:4260–4289, 2014
- L. Malagò and G. Pistone. Second-order optimization over the multivariate Gaussian distribution. In F. Barbaresco and F. Nielsen, editors, *Geometric Science of Information*, number 9389 in LNCS, pages 349–358. Springer, 2015

Parallel transport

- *e-transport*:
$$ {}^{\mathrm{e}}\mathbb{U}_p^q \colon B_p \ni U \mapsto U - \mathbb{E}_q[U] \in B_q \ . $$
- *m-transport*: for each $V \in {}^*B_q$
$$ {}^{\mathrm{m}}\mathbb{U}_q^p \colon {}^*B_q \ni V \mapsto \frac{q}{p} V \in {}^*B_p $$

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 $(s_p \colon p \in \mathcal{E})$,

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

- The *statistical exponential bundle* $T\mathcal{E}(M)$ 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* ${}^*T\mathcal{E}(M)$ 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 $F : \mathcal{E}(p) \to \mathbb{R}$.

- 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 $X \in L^{(\cosh -1)_*}(p)$ it holds
$$\frac{d}{dt}\mathbb{E}_{p(t)}[X] = \left\langle X - \mathbb{E}_{p(t)}[X], Dp(t) \right\rangle_{p(t)}$$
- The *statistical gradient* of F is a *section* of the statistical bundle, $p \mapsto (p, \operatorname{grad} f(p)) \in eTofp$ such that for each regular curve $t \mapsto p(t)$, it holds
$$\frac{d}{dt} f(p(t)) = \langle \operatorname{grad} f(p(t)), Dp(t) \rangle_{p(t)}$$

Everithing applies if the tangent space is in $C_0^{1,(\cosh -1)}(p)$, but technical detail have to be checked, e.d. ${}^{\mathrm{m}}\mathbb{U}_p^q$

Taylor formula in the Statistical Bundle

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

$$F(q) = F(p) + \frac{d}{dt} F(p(t))\Big|_{t=0} + \frac{1}{2} \frac{d^2}{dt^2} F(p(t))\Big|_{t=0} + R_2(f, p(\cdot))$$

with $R_2(f, p(\cdot)) = \int_0^1 (1-t) \left(\frac{d^2}{dt^2} F(p(t)) - \frac{d^2}{dt^2} F(p(t))\Big|_{t=o} \right) dt$

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

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

- where $\frac{d}{dt} \langle \operatorname{grad} F(p(t)), Dp(t) \rangle_{p(t)}\Big|_{t=o}$ depends on $p(\cdot)$

Accellerations

- $p(t) = \mathrm{e}^{U(t) - K_p(U(t))} \cdot p,\ U \in B_p$.
- Let us define the *acceleration* at t of a curve $t \mapsto p(t) \in \mathcal{E}(M)$. The velocity is defined to be $t \mapsto (p(t), Dp(t)) = \left(p(t), \frac{d}{dt} \log(p(t))\right) \in T\mathcal{E}(M)$
- The *exponential acceleration* is $t \mapsto (p(t), {}^{e}\mathrm{D}^2 p(t)) \in T\mathcal{E}(M)$ with

$$ {}^{e}\mathrm{D}^2 p(t) = \frac{d}{ds} {}^{\mathrm{e}}\mathbb{U}_{p(s)}^{p(t)} Dp(s) \Big|_{s=t} = \ddot{U}(t) - \mathbb{E}_{p(t)}\left[\ddot{U}(t)\right] $$

- The *mixture acceleration* is

$$ {}^{m}\mathrm{D}^2 p(t) = \frac{d}{ds} {}^{\mathrm{m}}\mathbb{U}_{p(s)}^{p(t)} Dp(s) \Big|_{s=t} = \frac{\ddot{p}(t)}{p(t)} $$

Computation

$$
\begin{aligned}
{}^{m}\mathrm{D}^2 p(t) &= \frac{d}{ds}\,{}^{\mathrm{m}}\mathbb{U}_{p(s)}^{p(t)}\, Dp(s)\Big|_{s=t} \\
&= \frac{d}{ds}\,{}^{\mathrm{m}}\mathbb{U}_{p(s)}^{p(t)}\,\frac{d}{ds}(U(s) - K_p(U(s)))\Big|_{s=t} \\
&= \frac{d}{ds}\,{}^{\mathrm{m}}\mathbb{U}_{p(s)}^{p(t)}(\dot{U}(s) - dK_p(U(s))\dot{U}(s))\Big|_{s=t} \\
&= \frac{d}{ds}\frac{p(s)}{p(t)}(\dot{U}(s) - \mathbb{E}_{p(t)}\left[\dot{U}(s)\right])\Big|_{s=t} \\
&= \frac{\dot{p}(s)}{p(t)}(\dot{U}(s) - \mathbb{E}_{p(t)}\left[\dot{U}(s)\right]) + \frac{p(s)}{p(t)}(\ddot{U}(s) - \mathbb{E}_{p(t)}\left[\ddot{U}(s)\right])\Big|_{s=t} \\
&= \frac{\dot{p}(t)}{p(t)}(\dot{U}(t) - \mathbb{E}_{p(t)}\left[\dot{U}(t)\right]) + \frac{p(t)}{p(t)}(\ddot{U}(t) - \mathbb{E}_{p(t)}\left[\ddot{U}(t)\right]) \\
&= (\dot{U}(t) - \mathbb{E}_{p(t)}\left[\dot{U}(t)\right])^2 + (\ddot{U}(t) - \mathbb{E}_{p(t)}\left[\ddot{U}(t)\right]) \\
&= \frac{\ddot{p}(t)}{p(t)}
\end{aligned}
$$

Autoparallel curves

- The *exponential accelleration* is

$$ {}^{e}\mathrm{D}^2 p(t) = \ddot{U}(t) - \mathbb{E}_{p(t)}\left[\ddot{U}(t)\right] $$

- The e-accelleration of a 1d-exponential family $t \mapsto \mathrm{e}^{tU - K_M(tU)} \cdot M$ is zero, $t\ddot{U} = 0$.

- The *mixture acceleration* is

$$ {}^{m}\mathrm{D}^2 p(t) = \frac{\ddot{p}(t)}{p(t)} $$

- The m-accelleration of a 1d-mixture family $t \mapsto p(0) + t(p(1) - p(0))$ is zero,

$$ \frac{p(0) + t(\ddot{p(1)} - p(0))}{p(0) + t(p(1) - p(0))} = 0 $$

Taylor's formulæ

The computation of $\frac{d}{dt}\left\langle \operatorname{grad} F(p(t)), Dp(t)\right\rangle_{p(t)}\Big|_{t=o}$ reduces to one term by chosing an autoparallel segment connecting p and q

- If $t \mapsto p(t)$ is the *mixture geodesic* connecting $p = p(0)$ to $q = p(1)$,

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

- If $t \mapsto p(t)$ is the *exponential geodesic* connecting $p = p(0)$ to $q = p(1)$,

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