The Ninth International Conference on Guided Self-Organization
GSO-2018: Information Geometry and Statistical Physics

Information Geometry: finite vs infinite sample space

Giovanni Pistone
www.giannidiorestino.it

DE CASTRO STATISTICS

Collegio Carlo Alberto

Leipzig March 28, 2018

My case

- The basic structure of Information Geometry (IG) is a (dually) affine Hessian manifold. The Riemannian structure is not enough.
- A generic parametric presentation could fit the need of Statistics and Machine Learning, but it is not natural in other applications e.g., Statistical Physics, Evolution Equation.
- It is useful exercise, even in studying statistical models on a finite state space or Gaussian statistical models, to avoid the parameterization.
- The affine and Hessian structure is feasible in the infinite dimensional case. There are many options of that generalization, the exponential representation of positive densities being one. The use of smooth densities is another possible choice.

PART I: Non-parametric Information Geometry: finite dimension

Even in the case of a finite state space there is an advantage in avoiding the explicit parameterization of probabilities. Probabilities are represented by random variables as centered log-likelihoods, tangent vectors are represented by the score random variable.

- G. Pistone and M. P. Rogantin. The algebra of reversible Markov chains. *Ann. Inst. Statist. Math.*, 65(2):269–293, 2013
- G. Pistone. Lagrangian function on the finite state space statistical bundle. *Entropy*, 20(2):139, Feb 2018
- L. Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian geometry of positive definite matrices. arXiv:1801.09269, 2018

Maximal exponential family

- We consider a finite sample space Ω, with $\#\Omega = N$. The probability simplex is $\Delta(\Omega)$ and $\Delta^\circ(\Omega)$ is its interior. The uniform probability on Ω is denoted μ, $\mu(x) = \frac{1}{N}$, $x \in \Omega$. The expected value of $f : \Omega \to \mathbb{R}$ with respect to the density $P \in \mathcal{E}(\mu)$ is denoted $\mathbb{E}_P[f] = \mathbb{E}_\mu[fP] = \frac{1}{N}\sum_{x\in\Omega} f(x)P(x)$.
- The *maximal exponential family* $\mathcal{E}(\mu)$ is the set of all strictly positive probability densities of (Ω, μ) i.e., $\Delta^\circ(\Omega)$. The name is suggested by the fact every $Q \in \mathcal{E}(\mu)$ can be expressed as

$$Q = \exp\left(\sum_{j=1}^{n} \theta_j T_j - \psi(\theta)\right)$$

where T_j is a basis of the vector space of the random variables on Ω.

Statistical bundle

- The (exponential) *statistical bundle* is

$$S\mathcal{E}(\mu) = \{(Q, V) | Q \in \mathcal{E}(\mu), \mathbb{E}_Q[V] = 0\} \ ,$$

and the linear fibers are

$$S_Q\mathcal{E}(\mu) = \{V | \mathbb{E}_Q[V] = 0\} \ .$$

The base of the bundle is $\mathcal{E}(\mu)$, the *exponential manifold*.

- A *section* of the statistical bundle is a mapping

$$X : \ \mathcal{E}(\mu) \to \cup_P S_P\mathcal{E}(\mu) = L^2(\mu) \ .$$

such that $X(P) \in S_P\mathcal{E}(\mu)$, $P \in \mathcal{E}(\mu)$.

- In statistical terms, a section is a mapping from the probability densities to square-integrable random variable such that $\mathbb{E}_P[X(P)] = 0$. If $\widehat{X}$ is a sample of X, then $\widehat{P}$ such that $\widehat{X}(\widehat{P}) = 0$ is an estimation of the density P.

Algebra

- The statistical bundle is a semi-algebraic subset of $\mathbb{R}^{2N}$ i.e., it is defined by algebraic equations and strict inequalities namely,

$$\begin{cases} \sum_j q_j = N \ , \\ \sum_j q_j v_j = 0 \ , \qquad j = 1, \dots, N \ . \\ q_j > 0 \ , \end{cases} \tag{1}$$

  It a real sub-manifold of $\mathbb{R}^{2N}$. A section is a mapping $\Delta^\circ \ni q \mapsto X(q) \in \mathbb{R}^N$ such that $\sum_j q_j X_j(q) = 0$. If X is smooth, $\{(q, X(q)\}$ is a sub-manifold.

- If we relax the positivity condition we get the algebraic variety whose ideal is generated in the polynomial ring $\mathbb{R}[(p_j, q_j) : j = 1, \dots, n]$ by the polynomials $\sum_j q_j - N$ and $\sum_j q_j v_j$. The statistical interpretation is a model for the Information Geometry of densities with total mass 1 but unrestricted in the sign.

Metric

- Each fiber $S_Q\,\mathcal{E}\,(\mu)$ is endowed with the inner product

$$(V_1, V_2) \mapsto \langle V_1, V_2 \rangle_Q = \mathbb{E}_Q\,[V_1\,V_2] = \mathrm{Cov}_Q\,(V_1, V_2)\ ,$$

  which gives to each fiber the structure of the Hilbert space of $L^2_0(Q)$.

- While the statistical bundle inherits its manifold structure from $\mathbb{R}^{2N}$, we want to add a special atlas of charts in order to show a structure of affine manifold which is of interest. In particular, the atlas produces the exponential form of the positive densities that comes from Statistical Physics.

- The exponential atlas of the statistical bundle $S\,\mathcal{E}\,(\mu)$ is the collection of charts given for each $P \in \mathcal{E}\,(\mu)$ by

$$\sigma_P \colon S\,\mathcal{E}\,(\mu) \ni (Q, V) \mapsto (s_P(Q), {}^{\mathrm{e}}\mathbb{U}^P_Q V) \in S_P\,\mathcal{E}\,(\mu) \times S_P\,\mathcal{E}\,(\mu)\ ,$$

  where

$$s_P(Q) = \log\frac{Q}{P} - \mathbb{E}_P\left[\log\frac{Q}{P}\right], \quad {}^{\mathrm{e}}\mathbb{U}^P_Q V = V - \mathbb{E}_P\,[V]\ .$$

Exponential form, divergence

- If $s_P(Q) = U$, the exponential form of Q as a density with respect to P is
$$Q = \mathrm{e}^{U - \mathbb{E}_P\left[\log \frac{Q}{P}\right]} \cdot P \ .$$
- In the statistical language, $\log \frac{Q}{P}$ is the sample information of Q with respect to P, $\mathbb{E}_P\left[\log \frac{Q}{P}\right]$ is the P-mean value of the information, $s_P(Q)$ is the available relative information.
- In the language of Physics, $-\log \frac{Q}{P} = \log \frac{P}{Q}$ is proportional to the Boltzmann entropy of Q with respect to P, and $\mathbb{E}_P\left[-\log \frac{Q}{P}\right] = \mathbb{E}_P\left[\log \frac{P}{Q}\right] = \mathrm{D}\left(P \,\|\, Q\right)$. is proportional to the P-mean value of the Boltzmann entropy, and $-s_P(Q) = \mathrm{D}\left(P \,\|\, Q\right) + \log \frac{P}{Q}$.
- We could represent each Q as $\mathrm{e}^{V + \psi_P(Q)} \dot{P}$ with $(Q, V) \in S\mathcal{E}(\mu)$. In such a case $\psi_P(Q) = \mathbb{E}_Q\left[\log \frac{Q}{P}\right] = \mathrm{D}\left(Q \,\|\, P\right)$. It follows the conjugation relation
$$\mathrm{D}\left(P \,\|\, Q\right) + \mathrm{D}\left(Q \,\|\, P\right) = \mathbb{E}_P\left[V\right] = \mathbb{E}_Q\left[U\right] \ .$$

Patches

- Let us compute the inverse of $Q \mapsto s_P(Q) = U$. We have $\mathbb{E}_P\left[\mathrm{e}^U\right] = \exp\left(-\mathbb{E}_P\left[\log \frac{Q}{P}\right]\right)$, and

$$s_P^{-1}(U) = \mathrm{e}^{U - K_P(U)} \cdot P \ , \quad K_P(U) = \log \mathbb{E}_P\left[\mathrm{e}^U\right] \ .$$

- The patch centered at P is

$$\sigma_P^{-1} = \epsilon_P \colon (S_P\,\mathcal{E}\,(\mu))^2 \ni (U, W) \mapsto (e_P(U), {}^{\mathrm{e}}\mathbb{U}_P^{e_P(U)} W) \in S\,\mathcal{E}\,(\mu) \ ,$$

with $e_P = s_P^{-1}$.

- The transition maps of the exponential atlas are

$$\sigma_{P_2} \circ \epsilon_{P_1}(U, W) =$$

$$\sigma_{P_2}\left(\mathrm{e}^{U - K_{P_1}(U)} \cdot P_1, W - \mathbb{E}_{\mathrm{e}_{P_1}(U)}[W]\right) =$$

$$\left(U - K_{P_1}(U) + \log \frac{P_1}{P_2} - \mathbb{E}_{P_2}\left[U - K_{P_1}(U) + \log \frac{P_1}{P_2}\right],\right.$$

$$\left. W - \mathbb{E}_{\mathrm{e}_{P_1}(U)}[W] - \mathbb{E}_{P_2}\left[W - \mathbb{E}_{\mathrm{e}_{P_1}(U)}[W]\right]\right) =$$

$$\left({}^{\mathrm{e}}\mathbb{U}_{P_1}^{P_2} U + s_{P_2}(P_1), {}^{\mathrm{e}}\mathbb{U}_{P_1}^{P_2} W\right) \ ,$$

so that the exponential atlas is indeed affine.

Velocity and score

- Given a curve $t \mapsto Q(t) \in \mathcal{E}(\mu)$, the expression in the chart s_P has derivative

$$\frac{d}{dt} s_P(Q(t)) = \frac{d}{dt}\left(\log \frac{Q(t)}{P} - \mathbb{E}_P\left[\log \frac{Q(t)}{P}\right]\right) = \frac{\dot{Q}(t)}{Q(t)} - \mathbb{E}_P\left[\frac{\dot{Q}(t)}{Q(t)}\right] .$$

- We define the Fisher score $\overset{\star}{Q}(t) = \frac{\dot{Q}(t)}{Q(t)}$. It is the velocity expressed in the moving frame, that is

$$\overset{\star}{Q}(t) = {}^{\mathrm{e}}\mathbb{U}_P^{Q(t)} \frac{d}{dt} s_P(Q(t)) .$$

- Every scalar of the exponential manifold $\phi\colon \mathcal{E}(\mu) \to \mathbb{R}$ is expressed in the chart at P by $\phi_P = \phi \circ e_p$.
- The natural gradient is the section $\operatorname{grad}\phi$ such that

$$\frac{d}{dt}\phi(Q(t)) = \left\langle \operatorname{grad}\phi(Q(t)), \overset{\star}{Q}(t) \right\rangle_{Q(t)} .$$

- In fact, if $\phi\colon \mathbb{R}^N$, then $\operatorname{grad}\phi(Q) = \nabla\phi(Q) - \mathbb{E}_Q[\nabla\phi]$.

Hessian manifold

- The base manifold $\mathcal{E}(\mu)$ is actually an Hessian manifold with respect to any of the convex functions $K_p(U) = \log \mathbb{E}_p\left[\mathrm{e}^U\right]$, $U \in S_p\,\mathcal{E}(\mu)$.
- Many computations are actually performed using the Hessian structure e.g.,

$$
\begin{aligned}
\mathbb{E}_{\mathrm{e}_P(U)}[H] &= dK_P(U)[H]\ ; \\
{}^{\mathrm{e}}\mathbb{U}_P^{\mathrm{e}_P(U)} H &= H - dK_P(U)[H]\ ; \\
d^2K_P(U)[H_1, H_2] &= \left\langle {}^{\mathrm{e}}\mathbb{U}_P^{\mathrm{e}_P(U)} H_1, {}^{\mathrm{e}}\mathbb{U}_P^{\mathrm{e}_P(U)} H_2 \right\rangle_{\mathrm{e}_P(U)}\ ; \\
d^3K_P(U)[H_1, H_2, H_3] &= \mathbb{E}_{\mathrm{e}_P(U)}\left[\left({}^{\mathrm{e}}\mathbb{U}_P^{\mathrm{e}_P(U)} H_1\right)\left({}^{\mathrm{e}}\mathbb{U}_P^{\mathrm{e}_P(U)} H_2\right)\left({}^{\mathrm{e}}\mathbb{U}_P^{\mathrm{e}_P(U)} H_3\right)\right]\ .
\end{aligned}
$$

Transports

- The mapping

$$ {}^{\mathrm{e}}\mathbb{U}_P^Q \colon S_P\,\mathcal{E}\,(\mu) \ni V \mapsto V - \mathbb{E}_Q\,[V] \in S_Q\,\mathcal{E}\,(\mu) $$

  is the exponential transport between the fibers.

- Given $V \in S_P\,\mathcal{E}\,(\mu)$ and $W \in S_Q\,\mathcal{E}\,(\mu)$,

$$ \left\langle {}^{\mathrm{e}}\mathbb{U}_P^Q V, W \right\rangle_Q = \mathbb{E}_Q\,[(V - \mathbb{E}_Q\,[V])W] = $$

$$ \mathbb{E}_Q\,[VW] = \mathbb{E}_P\left[V\left(\frac{Q}{P}W\right)\right] = \left\langle V, {}^{\mathrm{m}}\mathbb{U}_Q^P W \right\rangle_P \ , $$

  where

$$ {}^{\mathrm{m}}\mathbb{U}_Q^P \colon S_Q\,\mathcal{E}\,(\mu) \ni W \mapsto \frac{Q}{P}W \in S_P\,\mathcal{E}\,(\mu) $$

  is the mixture transport.

- If $V, W \in S_Q\,\mathcal{E}\,(\mu)$, with expression at P given by V_P and W_P,

$$ \langle V, W \rangle_Q = \left\langle {}^{\mathrm{e}}\mathbb{U}_P^Q V_P, {}^{\mathrm{e}}\mathbb{U}_P^Q W_P \right\rangle_Q = \left\langle {}^{\mathrm{m}}\mathbb{U}_Q^P\, {}^{\mathrm{e}}\mathbb{U}_P^Q V_P, W_P \right\rangle_P $$

  i.e., the self-adjoint operator $\left({}^{\mathrm{m}}\mathbb{U}_Q^P\, {}^{\mathrm{e}}\mathbb{U}_P^Q\right)$ on $S\,\mathcal{E}\,(\mu)$ is the expression of the Riemannian metric in the chart at P.

Gradient flow of the entropy I

- Consider the scalar field

$$Q \mapsto \mathcal{H}(Q) = -\mathbb{E}_Q[\log Q]$$

along the curve

$$t \mapsto Q(t) = e_P(V(t)) = \mathrm{e}^{V(t) - K_P(V(t))} \cdot P \ .$$

- We have

$$\begin{aligned}\mathcal{H}(Q(t)) &= -\mathbb{E}_{Q(t)}[V(t) - K_P(V(t)) + \log P] = \\ &K_P(V(t)) - \mathbb{E}_{Q(t)}[V(t) + \log P + \mathcal{H}(P)] + \mathcal{H}(P) = \\ &K_P(V(t)) - dK_P(V(t))[V(t) + \log P + \mathcal{H}(P)] + \mathcal{H}(P)\end{aligned}$$

Gradient flow of the entropy II

- The derivative of the entropy along the given curve is

$$\frac{d}{dt}\mathcal{H}(Q(t)) =$$
$$\frac{d}{dt}K_P(V(t)) - \frac{d}{dt}\, dK_P(V(t))[V(t) + \log P + \mathcal{H}(P)] =$$
$$- d^2K_P(V(t))[V(t) + \log P + \mathcal{H}(P), \dot{V}(t)] =$$
$$- \mathbb{E}_{Q(t)}\left[{}^{\mathrm{e}}\mathbb{U}_P^{Q(t)}(V(t) + \log P + \mathcal{H}(P))\, {}^{\mathrm{e}}\mathbb{U}_P^{Q(t)}\dot{V}(t)\right]$$

- We rewrite as

$$\frac{d}{dt}\mathcal{H}(Q(t)) = -\left\langle \log Q(t) + \mathcal{H}(Q(t)), \overset{\star}{Q}(t)\right\rangle_{Q(t)}$$

  so that natural gradient of the entropy is

- $$\operatorname{grad}\mathcal{H}(Q) = -(\log Q + \mathcal{H}(Q))$$

Gradient flow of the entropy III

- The natural gradient flow is

$$\overset{\star}{Q}(t) = -(\log Q(t) + \mathcal{H}(Q(t))) \ .$$

- The solution from Q is $Q(t) \propto Q^{\mathrm{e}^{-t}}$:

$$Q(t) = \frac{Q^{\mathrm{e}^{-t}}}{\mathbb{E}_\mu\left[Q^{\mathrm{e}^{-t}}\right]} \ ;$$

$$\log Q(t) = \mathrm{e}^{-t} \log Q - \log \mathbb{E}_\mu\left[Q^{\mathrm{e}^{-t}}\right] \ ;$$

$$\mathcal{H}(Q(t)) = -\mathrm{e}^{-t}\, \mathbb{E}_{Q(t)}[\log Q] + \log \mathbb{E}_\mu\left[Q^{\mathrm{e}^{-t}}\right] \ ;$$

$$\overset{\star}{Q}(t) = \frac{d}{dt} \log Q(t) = -\mathrm{e}^{-t} \log Q + \mathrm{e}^{-t}\, \mathbb{E}_{Q(t)}[\log Q] \ .$$

e-acceleration

- For each curve $t \mapsto \gamma(t) = (Q(t), W(t))$ in the statistical bundle, define its velocity at t to be

$$\overset{\star}{\gamma}(t) = \left(Q(t), W(t), \overset{\star}{Q}(t), \dot{W}(t) - \mathbb{E}_{Q(t)} \left[\dot{W}(t) \right] \right) \ ,$$

- In particular, the velocity of $t \mapsto \chi(t) = (Q(t), \overset{\star}{Q}(t))$ is

$$\overset{\star}{\chi}(t) = \left(Q(t), \overset{\star}{Q}(t), \overset{\star}{Q}(t), \overset{\star\star}{Q}(t) \right) \ ,$$

  where the acceleration $\overset{\star\star}{Q}(t)$ is

$$\overset{\star\star}{Q}(t) = \frac{d}{dt} \frac{\dot{Q}(t)}{Q(t)} - \mathbb{E}_{Q(t)} \left[\frac{d}{dt} \frac{\dot{Q}(t)}{Q(t)} \right] = \frac{\ddot{Q}(t)}{Q(t)} - \left(\overset{\star}{Q}(t)^2 - \mathbb{E}_{Q(t)} \left[\overset{\star}{Q}(t)^2 \right] \right)$$

- The acceleration of $t \mapsto \mathrm{e}^{tU - \psi(t)}$ is zero.

m-acceleration, 0-acceleration

- We could consider an *exponential acceleration* ${}^{e}\mathrm{D}^2 Q(t) = \overset{**}{Q}(t)$, a mixture acceleration ${}^{m}\mathrm{D}^2 Q(t) = \ddot{Q}(t)/Q(t)$, and a 0-acceleration

$$ {}^{0}\mathrm{D}^2 Q(t) = \frac{1}{2}\left({}^{e}\mathrm{D}^2 Q(t) + {}^{m}\mathrm{D}^2 Q(t)\right) $$

- The mixture acceleration of $t \mapsto (1-t)P + tQ$ is zero.
- The Boltzmann-Gibbs density

$$ Q(\theta) = N\mathrm{e}^{-(1/\theta)H}/Z(\theta),. \quad \theta > 0 $$

has velocity $\overset{*}{Q}(\theta) = \theta^{-2}(H - \mathbb{E}_\theta[H])$ and acceleration is

$$ \overset{**}{Q}(\theta) = -2\theta^{-3}(H - \mathbb{E}_\theta[H]) = -2\theta^{-1}\overset{*}{Q}(\theta) $$

Levi-Civita covariant derivative

- Let U, Y, W be section of the statistical bundle and $\overset{\star}{Q}(t) = Y(Q(t))$. Then

$$\frac{d}{dt}\left\langle U(Q(t)), W(Q(t))\right\rangle_{Q(t)}\Big|_{t=0} = d^3K_Q(0)[U(Q), Y(Q), Z(Q)]+$$
$$d^2K_Q(0)\left[({}^{\mathrm{e}}\mathrm{D}_Y U)(Q), W(Q)\right] + d^2K_Q(0)\left[U(Q), ({}^{\mathrm{e}}\mathrm{D}_Y W)(Q)\right] =$$
$$\mathbb{E}_Q\left[U(Q)Y(Q)W(Q)\right]+\left\langle({}^{\mathrm{e}}\mathrm{D}_Y U)(Q), W(Q)\right\rangle_Q+\left\langle U(Q), ({}^{\mathrm{e}}\mathrm{D}_Y W)(Q)\right\rangle_Q$$

- It follows that the covariant derivative

$$D_Y U(Q) = {}^{\mathrm{e}}\mathrm{D}_Y U(Q) + \frac{1}{2}\left(U(Q)Y(Q) - \mathbb{E}_Q\left[U(Q)Y(Q)\right]\right)$$

  is compatible with the metric

- Moreover

$$D_V U - D_U V = {}^{\mathrm{e}}\mathrm{D}_V U - {}^{\mathrm{e}}\mathrm{D}_V U = [U, V]$$

- $D_V U$ is the Levi-Civita covariant derivative and, with $V(Q(t)) = U(Q(t)) = \overset{\star}{Q}(t)$, we find

$$\overset{\star\star}{Q}(t) + \frac{1}{2}(\overset{\star}{Q}(t)^2 - \mathbb{E}_{Q(t)}\left[\overset{\star}{Q}(t)^2\right] = {}^{0}\mathrm{D}^2 Q(t)\ .$$

Lagrangian function

- A Lagrangian function is a smooth scalar field on the statistical bundle

$$L\colon S\mathcal{E}(\mu) \ni (Q, W) \mapsto L(Q, W) \in \mathbb{R}\ . \tag{2}$$

- If $t \mapsto (Q(t), W(t))$ is a smooth curve in $S\mathcal{E}(\mu)$, then

$$\frac{d}{dt} L(Q(t), W(t)) = d_1 L(Q(t), W(t))[\overset{\star}{Q}(t)] + d_2 L(Q(t), W(t))[\overset{\star}{W}(t)]$$

where d_2 is the fiber derivative and

$$d_1(Q, W)[H_1] = d_1 L_P(U, V)[{}^{\mathrm{e}}\mathbb{U}^P_{\mathrm{e}_P(U)} H_1]\ , \quad H_1 \in S_Q\mathcal{E}(\mu)$$

- In particular

$$\frac{d}{dt} L(Q(t), \overset{\star}{Q}(t)) = d_1 L(Q(t), \overset{\star}{Q}(t))[\overset{\star}{Q}(t)] + d_2 L(Q(t), \overset{\star}{Q}(t))[\overset{\star\star}{Q}(t)]$$

Euler-Lagrange equation

- At a critical point of the action integral, we can derive the Euler-Lagrange equations

$$d_1 L(Q(t), \overset{\star}{Q}(t))[H] - \frac{d}{dt} d_2 L(Q(t), \overset{\star}{Q}(t))[H] = 0, \quad H \in S_{Q(t)}\,\mathcal{E}\,(\mu)$$

- For example,

$$L(Q, W) = \frac{1}{2}\,\langle W, W\rangle_Q + \kappa\,\mathcal{H}\,(Q)\ , \quad \kappa \geq o$$

has Euler-Lagrange equations

$$\overset{\star\star}{Q}(t) + \left(\overset{\star}{Q}(t)^2 - \mathbb{E}_{Q(t)}\left[\overset{\star}{Q}(t)^2\right]\right) =$$
$$\frac{1}{2}\left(\overset{\star}{Q}(t)^2 - \mathbb{E}_{Q(t)}\left[\overset{\star}{Q}(t)^2\right]\right) - \kappa(\log\,(Q(t)) + \mathcal{H}\,(Q(t)))$$

that is

$${}^0\mathrm{D}^2 Q(t) = \kappa\,\mathrm{grad}\,\mathcal{H}\,(Q(t))$$

PART II

Information Geometry of the Gaussian space

- G. Pistone. Examples of the application of nonparametric information geometry to statistical physics. *Entropy*, 15(10):4042–4065, 2013
- 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. Optimal approximations of the Fokker-Planck-Kolmogorov equation: projection, maximum likelihood eigenfunctions and Galerkin methods. arXiv:1603.04348, 2016
- D. Brigo and G. Pistone. Projection based dimensionality reduction for measure valued evolution equations in statistical manifolds. In F. Nielsen, F. Critchley, and C. Dodson, editors, *Computational Information Geometry. For Image and Signal Processing*, Signals and Communication Technology, pages 217–265. Springer, 2017
- G. Pistone. Information geometry of the Gaussian space. arXiv:1803.08135, 2018

Gaussian space

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

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

- For a convex Φ, the Orlicz space $L^{\Phi}(M)$ is the vector space of all random variables U such that $\mathbb{E}_M[\Phi(\alpha U)]$ is finite for some $\alpha > 0$.
- The Orlicz spaces of interest 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 set

$$\left\{u \in L^{\Phi}(p) \middle| \mathbb{E}_p[\Phi(u)] \leq 1\right\}$$

is the *closed unit ball of a Banach space*, hence

$$\|u\|_p = \inf\left\{\rho > 0 \middle| \mathbb{E}_p\left[\Phi\left(\frac{u}{\rho}\right)\right] \leq 1\right\}.$$

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

Model space

- $U \in L^{(\cosh -1)}(p)$ if, and only if, the moment generating function $\alpha \mapsto \mathbb{E}_p\left[\mathrm{e}^{\alpha u}\right]$ is finite in a neighborhood of 0.
- $L^{\Phi}(p)$ is the space of sufficient statistics in an exponential family.
- The space $L^{(\cosh -1)_*}(M)$ is separable with dual space $L^{(\cosh -1)}(M)$ because

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

- A positive density f has finite entropy if, and only if, $f \in L^{(\cosh -1)_*}(M)$.

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 semi-continuous 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

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

Isomorphism of L^Φ spaces

- For positive densities p, q, the condition $\int p^{1-\theta}(x)q^{\theta}(x)M(x)dx < \infty$ on an open neighborhood of $[0,1]$ is an equivalence relation $p \smile q$.
- 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 $\epsilon > 0$.

- 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, 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*, 22(3):1431–1447, 2016

Transports

- As all the spaces $L^{(\cosh -1)}(p)$ are isomorphic to $L^{(\cosh -1)}(M)$, for all $p \in \mathcal{E}(M)$, the fibers

$$B_p = \left\{ u \in L^{(\cosh -1)}(p) \middle| \mathbb{E}_p[V] = 0 \right\}$$

  are sub-spaces of $L^{(\cosh -1)}(M)$.

- The exponential transports are defined by

$${}^{\mathrm{e}}\mathbb{U}_p^q \colon B_p \to B_q \qquad u \mapsto u - \mathbb{E}_q[u]$$

- Each of the spaces

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

  is the pre-dual of B_p.

- The mixture transports are defined by

$${}^{\mathrm{m}}\mathbb{U}_q^p \colon {}^*B_q \to {}^*B_p \qquad v \mapsto \frac{q}{p} v$$

- For all $v \in {}^*B_q$ and $u \in B_p$ it holds

$$\left\langle v, {}^{\mathrm{e}}\mathbb{U}_p^q u \right\rangle_q = \left\langle {}^{\mathrm{m}}\mathbb{U}_q^p v, u \right\rangle_p$$

Random variables 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 $f : \mathbb{R}^d \to \mathbb{R}$ which are bounded by a polynomial.

Statistical exponential manifold and bundles

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

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

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

$$\left\{ (p, V) \middle| p \in \mathcal{E}(M), V \in L^{(\cosh -1)}(M), \mathbb{E}_p[V] = 0 \right\}$$

by the affine atlas of global charts

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

- The statistical pre-dual 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

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

Space derivatives

In the continuous case we want to discuss the action of translations and diffeomorphisms on densities, and differentiable densities.

- The 1-d Poincaré inequality is

$$\int \left(f(x) - \int f(y) M(y)\, dy \right)^2 M(x)\, dx \;\leq \int |f'(x)|^2 M(x)\, dx \quad .$$

- In our case we have

$$|\mathrm{Cov}_M(f,g)| \leq \text{const} \times \|\nabla f\|_{L^{(\cosh -1)_*}(M)} \, \|\nabla g\|_{L^{(\cosh -1)}(M)} \quad .$$

- For example, f g_n, g is a sequence on differentiable densities in $\mathcal{E}(M)$, $\lim_{n\to\infty} \nabla g_n = \nabla g$ in $L^{(\cosh -1)_*}(M)$, and $u \in S_1 \mathcal{E}(M)$, with $\nabla u \in L^{(\cosh -1)}(M)$, then

$$\lim_{n\to\infty} \int g_n(x) u(x) M(x)\, dx \;= \int g(x) u(x) M(x)\, dx$$

Orlicz-Sobolev space with Gaussian weight

- The Orlicz-Sobolev spaces with Gaussian weight M are the vector spaces $W^1_{\cosh -1}(M)$, respectively $W^1_{(\cosh -1)_*}(M)$, defined by

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

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

- Both are Banach spaces with the norm of the graph

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

Embeddings

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

- We have the continuous mappings

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

- We have the continuous mappings for $p > 1$

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

- Each $u \in W^{1,(\cosh -1)}(M)$ is a.s. Hölder of all orders on each $\overline{\Omega}_R$ and hence a.s. continuous.
- The restriction $W^{1,(\cosh -1)}(M) \to C(\overline{\Omega}_R)$ is compact.

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

- If we restrict the exponential family $\mathcal{E}(M)$ to $W^1_{\cosh -1}(M)$,

$$W_M = W^1_{\cosh -1}(M) \cap B_M = \left\{ U \in W^1_{\cosh -1}(M) \middle| \mathbb{E}_M[U] = 0 \right\}$$

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

$$W_f = W^1_{\cosh -1}(M) \cap B_M = \left\{ U \in W^1_{\cosh -1}(M) \middle| \mathbb{E}_f[U] = 0 \right\}, \quad f \in \mathcal{E}_1(M)$$

to be models for the tangent spaces of $\mathcal{E}_1(M)$. The e-transport acts on these spaces

$$^0\mathbb{U}^g_f : W_f \ni U \mapsto U - \mathbb{E}_g[U] \in W_g \ ,$$

so that we can define the statistical bundle to be

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

Example: Hyvärinen divergence

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

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

- The expression in the chart centered at M is

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

where $f = e_M(u)$, $g = e_M(v)$.

Elliptic operator

- Elliptic operator as a densely defined section of the statistical 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 affine atlas 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

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

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

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