NON-PARAMETRIC INFORMATION GEOMETRY WITH DERIVATIVES
HANDOUT NOT INTENDED FOR PUBLICATION OR DISTRIBUTION

GIOVANNI PISTONE

Abstract. Non-parametric Information Geometry according to a series of papers starting with [16] consists of a manifold on the set of positive densities of a measure space. The manifold is modeled on the Banach space of exponentially integrable random variables. In a more recent presentation [14] the relevant structure is described a Banach bundle of couples (p, u) where p is a positive density and u is a random variable such that $E_p(u) = 0$. Each connected component of the base manifold, consisting of densities which are connected by an open exponential family, is fully described in [17]. Other methods for dealing with the infinite-dimensional geometry of probabilities are available, in particular [3]. The main limitation of this approach is the inability to deal with properties of the statistical models depending on the structure of the sample space where the densities are defines e.g., the smoothness. In the framework of Gaussian spaces [6] it is actually possible to study such properties while retaining the same bundle structure. Preliminary results have been published in [7, 15] and further research is in progress. An example of application is the study of Hyvärinen divergence [6].

Contents

1. Gradient of a density

We read from [6, 13]. Given a statistical model $\mathcal{M}$ of positive densities on a measure space $(X, \mathcal{X}, \mu)$, a *local scoring rule* is a mapping $S\colon \mathcal{M}$ with values in random variables $x \mapsto S(x, q)$. The "local" means that the scoring rule depends on the sample point x only. The risk under a positive density $p \in \mathcal{P}$ is $d(p, q) = E_p(S(q))$. We assume that the expected value is defined for each couple p, q $\mathcal{M}$. The scoring rule is *proper* is $q \mapsto d(p, q)$ is minimized at $q = p$ only that is, $d(p, q) \geq d(p, p)$ and $d(p, q) = d(p, p)$ implies $q = p$. Notice that there is a sampling version of the objective function namely, $\hat{d}(q) = \sum_{j=1}^{N} S(X_j, Q)$ with (X_j) IID p and $\hat{q} = \operatorname{argmin} \hat{d}(q)$ is an estimator of p. The divergence associate to S is $D(p, q) = d(p, p) - d(p, q)$ and minimization of $q \mapsto D(p, q)$ is equivalent to the minimization of $d(p, q)$. However, $D(p, q)$ has no sampling version.

1.1. **Example: log-score.** Take $S(x, q) = -\log q(x)$. As $-\log q \geq 1 - q$, the expectation $E_p(-\log q)$ is well defined, possibly $+\infty$, if $\int q(x)p(x)\ \mu(dx) < +\infty$ for all $p, q \in \mathcal{M}$. We have

$$d(p, q) = -\int p(x) \log q(x)\ \mu(dx) = \int \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} q(x)\ \mu(dx) - \int p(x) \log p(x)\ \mu(dx) \geq$$
$$\int \left(1 - \frac{p(x)}{q(x)}\right) q(x)\ \mu(dx) + d(p, p) = d(p, p)\ .$$

Date: DISMA Polito, June 13 2018; revised June 24, 2018.

The divergence can be translated to the minimum value to get a non-negative divergence,

$$d(p,q) - d(p,p) = \int p(x) \log \frac{p(x)}{q(x)} \; \mu(dx) = \int \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} q(x) \; \mu(dx) = D_{\mathrm{KL}}(p\|q) \; ,$$

the Kullback-Leibler divergence. The KL-divergence is always well defined and faithful because, if we write $f(t) = t \log t$, then f is strictly convex and bounded below, so

$$D_{\mathrm{KL}}(p,q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x) \; \mu(dx) \geq f\left(\int \frac{p(x)}{q(x)} q(x) \; \mu(dx)\right) = f(1) = 0 \; .$$

Notice that the application of the LLN in the sample case requires a further assumption i.e., $\log q$ must be p-integrable for all $p, q \in \mathcal{M}$.

1.2. **Hyvärinen divergence.** Here we use unpublished notes by M.P. Rogantin (2018). Let us assume now that the sample space is the n-dimensional real space and each density q in $\mathcal{M}$ is strictly positive and such that $\partial_j \log q = \partial_j q/q$ is square integrable for each $p \in \mathcal{M}$. The *Hyvärinen divergence* is

$$D_{\mathrm{H}}(p,q) = \frac{1}{2} \int |\nabla \log p(x) - \nabla \log q(x)|^2 \, p(x) \; dx \; .$$

By expanding the squared norm of the difference, we obtain

$$D_{\mathrm{H}}(p,q) = \frac{1}{2} \int |\nabla \log p(x)|^2 \, p(x) \; dx + \frac{1}{2} \int |\nabla \log q(x)|^2 \, p(x) \; dx - \int \nabla \log p(x) \cdot \nabla \log q(x) \; p(x) \; dx \; .$$

The first term does not depend on q. Integration by parts in the last term gives

$$- \int \nabla \log p(x) \cdot \nabla \log q(x) \; p(x) \; dx = - \int \nabla p(x) \cdot \nabla \log q(x) \; dx = \int \Delta \log q(x) \; p(x) \; dx \; ,$$

if the second derivatives exist and the border terms vanish. In such a case, we define the *Hyvärinen score* to be

$$S_{\mathrm{H}}(q) = \Delta \log q(x) + \frac{1}{2} \left|\nabla \log q(x)\right|^2 \; .$$

Minimization of the expected Hyvärinen score is the same as minimization of the Hyvärinen divergence.

All assumptions are satisfied if $\mathcal{M}$ is the multivariate Gaussian model. This provides us with an example where a statistical problem requires a ditailed discussin of the properties of the spatial derivatives. This methodology was originally motivated by the need of a divergence that does not require the computation of the normalizing constant. That is, if $p(x) = f(x)/Z$, then $\log p(x) = \log f(x) - \log Z$ and $\nabla \log p(x) = \nabla \log f(x)$.

Variations on the theme are possible. On the Gaussian space $(\mathbb{R}^n, \gamma)$, $\gamma(x) = (2\pi)^{-n/2} \mathrm{e}^{-|x|^2/2}$, we can define

$$D_{\mathrm{GH}}(p,q) = \frac{1}{2} \int |\nabla \log p(x) - \nabla \log q(x)|^2 \, p(x) \gamma(x) \; dx \; .$$

In this case, the derivation operator is defined in the sense of the analysis of the Gaussian space, see the next section.

Another option is to substitute the log function with the Nigel Newton deformed logarithm $\log_A(t) = \int ds/A(s)$, $A(t) = s/(1+s)$. See the references to this formalism in [10]. A possible definition in this case is

$$D_{A\mathrm{H}(p,q)} = \frac{1}{2} \int |\nabla \log_A p(x) - \nabla \log_A q(x)|^2 \, A(p(x)) \; dx \; .$$

2. Gaussian space and derivation

Let us first review a few facts about the Gaussian space as it is defined in P. Malliavin textbook [8, Ch. V]. We restrict ourselves to the finite-dimensional sample space. References for the infinite-dimensional case are P. Malliavin monograph [9] and I. Nourdin and G. Peccati monograph [12].

We denote by τ_h the translation operator $\tau_h u(x) = u(x-h)$.

Proposition 1 (Translation)**.** *If $u \in L^2(\gamma)$ then $\tau_h u \in L^1(\gamma)$ and the mapping $u \mapsto \tau_h u$ is continuous.*

Proof. We have

$$\|u\|_{L^1(\gamma)} = \int |u(x-h)|\gamma(x)\ dx\ = \int |u(y)|\,\gamma(y+h)\ dy = \int |u(y)|\,\gamma(x+h)\gamma^{-1}(y)\gamma(y)\ dy\ \le \left(\int \gamma^2(y+h)\gamma^{-1}(y)\ dy\right)^{1/2} \|u\|_{L^2(\gamma)} = \mathrm{e}^{|h|^2}\ \|u\|_{L^2(\gamma)}\ ,$$

where the last equality follows from the computation

$$\gamma^2(y+h)\gamma^{-1}(y) = (2\pi)^{n/2}\mathrm{e}^{\frac{1}{2}|y|^2-|y+h|^2} = (2\pi)^{n/2}\mathrm{e}^{|h|^2}\mathrm{e}^{-\frac{1}{2}|y-2h|^2}\ .$$

□

Compare with the case of Lebesgue spaces where the translation is an isometry from each space into itself.

2.1. **The space $\mathbb{D}$.** If $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable define $\delta_j f(x) = x_j f(x) - \frac{\partial}{\partial x_j} f(x)$ and $\delta^\alpha = \prod_{j=1}^n \delta_j^{\alpha_j}$, $\alpha \in A = \mathbb{Z}_{\ge}^n$, and the *Hermite polynomials* are defined by $H_\alpha(x) = \delta^\alpha 1$. It is an orthogonal total system in $L^2(\gamma) = L^2(\mathbb{R}^n, \mathcal{B}, \gamma)$ with $\|H_\alpha\|_\gamma^2 = \int H_\alpha(x)^2\gamma(x)\ dx\ = \alpha!$.

Each $u \in L^2(\gamma)$ has the Fourier expansion

$$u = \sum_{\alpha\in A} c_\alpha(u)\frac{1}{\alpha!}H_\alpha\ , \quad c_\alpha(u) = \langle u, H_\alpha\rangle_\gamma = \int u(x)H_\alpha(x)\gamma(x)\ dx\ \ ,$$

with $\|u\|_\gamma^2 = \sum_{\alpha\in A} c_\alpha^2 \frac{1}{\alpha!}$. Let π be the finite measure on A defined by $\pi(\alpha) = 1/\alpha!$. The mapping $u \leftrightarrow c.$ is an isometry between $L^2(\gamma)$ and $L^2(\pi)$. In [9] the space (A, π) is called the *numerical model* of the Gaussian space.

As $\partial_j H_\alpha = \alpha_j H_{\alpha-e_j}$ if $\alpha_j \ge 1$, zero otherwise, and $\delta_j H_\alpha = H_{\alpha+e_j}$, we can define the operators on $L^2(\gamma)$

$$\partial_j\left(\sum_{\alpha\in A} c_\alpha\frac{1}{\alpha!}H_\alpha\right) = \sum_{\alpha\in A} c_{\alpha+e_j}\frac{1}{\alpha!}H_\alpha\ ,$$

$$\delta_j\left(\sum_{\alpha\in A} c_\alpha\frac{1}{\alpha!}H_\alpha\right) = \sum_{\alpha\in A\colon\,\alpha_j\ge 1} \alpha_j c_{\alpha-e_j}\frac{1}{\alpha!}H_\alpha\ ,$$

$$\delta_j\partial_j\left(\sum_{\alpha\in A} c_\alpha\frac{1}{\alpha!}H_\alpha\right) = \sum_{\alpha\in A} \alpha_j c_\alpha\frac{1}{\alpha!}H_\alpha\ ,$$

whose domains are, respectively,

$$\mathrm{Dom}\,(\partial_j) = \left\{\sum_{\alpha\in A}\frac{c_{\alpha+e_j}^2}{\alpha!} < \infty\right\}\ ,$$

$$\mathrm{Dom}\,(\delta_j) = \left\{\sum_{\alpha,\neq 0}\frac{\alpha_j^2 c_{\alpha-e_j}^2}{\alpha!} < \infty\right\}\ ,$$

$$\mathrm{Dom}\,(\delta_j\partial_j) = \left\{\sum_{\alpha\in A}\frac{\alpha_j^2 c_\alpha^2}{\alpha!} < \infty\right\}\ .$$

Proposition 2. *The operators ∂_j, δ_j, $\delta_j\partial_j$ are closed.*

Proposition 3. *If $u \in \mathrm{Dom}\,(\partial_j)$ and $v \in \mathrm{Dom}\,(\delta_j)$. then $\langle \partial_j u, v\rangle_\gamma = \langle u, \delta_j v\rangle_\gamma$.*

In particular, if $u \in \mathrm{Dom}\,(\partial_j)$ and $\phi \in C_0^\infty(\mathbb{R}^n)$ (compact support) then $\phi, \delta_j\phi \in L^2(\gamma)$ and $\phi \in \mathrm{Dom}\,(\delta_j)$ so that $\langle \partial_j u, \phi\rangle_\gamma = \langle u, \delta_j\phi\rangle_\gamma$.

Under the same assumptions, let us consider the ordinary integral and the distributional definition of partial derivative. The space of infinitely differentiable functions with compact support is denoted

$C_0^\infty(\mathbb{R}^n)$. Notice that, if B is a ball, the restriction to B is continuous mapping from $L^2(\gamma)$ into $L^2(B)$.

$$\int \partial_j u(x)\phi(x)\ dx = \sum_{\alpha\in A} c_{\alpha+e_j}\frac{1}{\alpha!}\int H_\alpha(x)\phi(x)\ dx = \sum_{\alpha\in A} c_{\alpha+e_j}\frac{1}{(\alpha+e_j)!}\int \partial_j H_{\alpha+e_j}(x)\phi(x)\ dx = -\int u(x)\partial_j\phi(x)\ dx\ ,$$

so the operator ∂_j coincides with the derivative in the sense of distributions. The following proposition is a converse statement.

Proposition 4. *If $u \in L^2(\gamma)$ and u' is the j-partial derivative in the sense of distributions,*

$$\int u'(x)\phi(x)\ dx = -\int u(x)\partial_j\phi(x)\ dx\ , \quad \phi \in C_0^\infty(\mathbb{R}^n)\ ,$$

and $u' \in L^2(\gamma)$, then $u \in \mathrm{Dom}\,(\partial_j)$ and $\partial_j = u'$.

If $u \in \mathrm{Dom}\,(\partial_j)$ for all j, the gradient operator ∇ is defined as the vector field whose components are the $\partial_j u(x)$. Its domain is the intersection of the domains $\mathrm{Dom}\,(\nabla) = \cap_{j=1}^n \mathrm{Dom}\,(\partial_j)$.

Proposition 5 (Poincaré inequality)**.** *If $u \in \mathrm{Dom}\,(\nabla)$ then*

$$\int \left| u(x) - \int u(y)\gamma(y)\ dy \right|^2 \gamma(x)\ dx \ \le \int \|\nabla u(x)\|^2 \gamma(x)\ dx\ .$$

Proof. The following proof is given in the numerical model. Other proof are given in the quoted literature. We have $\partial_j u = \sum_\alpha c_{\alpha+e_j}\frac{1}{\alpha!}H_\alpha$ for each j, hence

$$\|\partial_j u\|_\gamma^2 = \sum_\alpha \frac{c_{\alpha+e_j}^2}{\alpha!} = \sum_\alpha (\alpha_j+1)\frac{c_{\alpha+e_j}^2}{(\alpha+e_j)!} = \sum_{\alpha_j\ge 1}\alpha_j\frac{c_\alpha^2}{\alpha!} \ge \sum_{\alpha_j\ge 1}\frac{c_\alpha^2}{\alpha!}\ .$$

It follows

$$\|\nabla u\|_\gamma^2 = \sum_{j=1}^n \|\partial_j u\|_\gamma^2 = \sum_{j=1}^n\sum_{\alpha_j\ge 1}\frac{c_\alpha^2}{\alpha!} \ge \sum_{\alpha\ne 0}\frac{c_\alpha^2}{\alpha!}\ .$$

As $c_0 = \int u(x)\gamma(x)\ dx$, we have proved the Poincaré inequality, □

Proposition 6 (Gauss-Taylor expansion)**.** *If $f \in C^\infty(\mathbb{R}^n)$ and $\partial^\alpha f \in L^2(\gamma)$ for all $\alpha \in A$, then*

$$f = \sum_\alpha \langle f, H_\alpha\rangle_\gamma \frac{1}{\alpha!}H_\alpha = \sum_{\alpha\in A}\left(\int \partial^\alpha f(x)\gamma(x)\ dx\ \right)\frac{1}{\alpha!}H_\alpha$$

and

$$\|f\|_\gamma^2 = \sum_{\alpha\in A}\left(\int \partial^\alpha f(x)\gamma(x)\ dx\ \right)^2\frac{1}{\alpha!}\ .$$

Definition 7 (The space $\mathbb{D}$)**.** We denote by $\mathbb{D}$ the domain of ∇ endowed with the Hilbert norm

$$\|u\|_{\mathbb{D}}^2 = \|u\|_\gamma^2 + \sum_{j=1}^2 \|\partial_j u\|_\gamma^2\ .$$

Proposition 8. *If $u \in \mathrm{Dom}\,(\nabla)$ and $h \in \mathbb{R}^n$, then $h \mapsto \tau_h u$ is differentiable as a mapping in $L^1(\gamma)$ with derivative $\nabla u \cdot h \in L^2(\gamma)$.*

3. Maximal exponential model on the Gaussian space

Here we read from [8, Ch. V], [14], [17].

If γ is the standard n-dimensional Gaussian density, consider a 1-dimensional Gibbs model $t \mapsto \mathrm{e}^{tv}/Z(t)\cdot\gamma$, with $t \in I$, I open and $0 \in I$. The partition function $Z(t) = \int \mathrm{e}^{tv(x)}\ \gamma(x)\ dx < +\infty$, the "energy"random variable v is subject to a restrictive condition.

More generally, given any positive density $p \in \mathcal{P}_\ge$ of the n-dimensional real space endowed with the standard Gaussian, the class of possible "energy" random variables is

$$L^{(\cosh -1)}(p) = \left\{ v \in L^0(p) \middle| \mathbb{E}_p\left[\cosh(\alpha v)\right] < +\infty \text{ for some } \alpha > 0 \right\}\ .$$

It is the Orlicz space we call *exponential space* [11]. The closed unit ball is

$$\left\{v \in L^{(\cosh -1)}(p) \middle| \mathbb{E}_\gamma\left[\mathrm{e}^v\right] \leq 1\right\} \ .$$

It is easy to check that

$$L^\infty(p) \subset L^{(\cosh -1)}(p) \subset L^{\infty-0} = \cap_{\alpha \geq 1} L^\alpha(p)$$

with continuous injections. We define $B_p = \left\{v \in L^{(\cosh -1)}(p) \middle| \mathbb{E}_p[v] = 0\right\}$. The linear bundle

$$\{(p, v) | p \in \mathcal{P}_\geq, v \in B_p\}$$

is the natural non-parametric set-up for Information geometry in the sense of [1, 2, 16].

The function

$$K_p \colon B_p \ni u \mapsto \log \mathbb{E}_p\left[\mathrm{e}^u\right] \in [0, +\infty]$$

is convex and lower semi-continuous. The proper domain $\mathrm{Dom}(K_p)$ is a convex set and the interior of the proper domain $\mathcal{S}_p$ is an open convex set containing the open unit ball of B_p. For each $u \in \mathcal{S}_p$ we define the density

$$e_p(u) = \mathrm{e}^{u - K_p(u)} \cdot p \in \mathcal{P}_\geq \ .$$

The set of all such densities in the *maximal exponential model* at p, $\mathcal{E}(p)$. If $q = e_p(u)$, then $u = s_p(q) = \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right]$. That is, $e_p \colon \mathcal{S}_p \to \mathcal{E}(p)$ with inverse $s_p \colon \mathcal{E}(p) \to \mathcal{S}_p$. We define the binary relation $\smile$ on $\mathcal{P}_\geq$ by saying that $p \smile q$ if p and q are connected by an open exponential arc. It is an equivalence relation [5].

The global structure as p varies is clarified by the following "Portmanteau theorem," cf. [17, Th. 4.7]. The following propositions are equivalent:

(1) $q \in \mathcal{E}(p)$;
(2) $p \smile q$;
(3) $\mathcal{E}(p) = \mathcal{E}(q)$;
(4) $L^{(\cosh -1)}(p) = L^{(\cosh -1)}(q)$;
(5) $\log \frac{q}{p} \in L^{(\cosh -1)}(p)$ and $\log \frac{q}{p} \in L^{(\cosh -1)}(q)$;
(6) $\frac{q}{p} \in L^\alpha(p)$ and $\frac{p}{q} \in L^\alpha(q)$ for some $\alpha > 1$.

As a consequence, given a $\smile$-class of densities $\mathcal{E}$, the atlas of charts

$$s_p(q) = \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \in L^{(\cosh -1)}(p) \ , \ q \in \mathcal{E} \ ,$$

$p \in \mathcal{E}$, defines the *exponential affine manifold* and the *statistical bundle*

$$S\mathcal{E} = \{(p, u) | p \in \mathcal{E}, u \in B_p\}$$

is the expression of the tangent bundle in the atlas [14].

In the rest of the talk we focus on the Gaussian case that is $\mathcal{E} = \mathcal{E}(1)$.

Let $(\cosh -1)_*$ the convex conjugate of $(\cosh -1)$,

$$(cosh - 1)_*(y) = \sup_x \left(xy - (\cosh -1)(x)\right) \ .$$

This convex function defines the Orlicz space $L^{(\cosh -1)_*}(p)$ whose dual is $L^{(\cosh -1)}(p)$ in the bilinear form

$$L^{(\cosh -1)}(p) \times L^{(\cosh -1)_*}(p) \ni (u, f) \mapsto \int u(x) f(x) \gamma(x) \, dx \ .$$

We have, for each $p \in \mathcal{E}$ and $a > 1$, that

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

with continuous injections.

Remark 1. The Gaussian version of the Hyvärinen divergence can be discussed with the assumption $\log p \in \mathbb{D}$ to get similar expression for the Hyvärinen score with some partial derivatives replaced by the operator δ. However, extra assumptions are still necessary to ensure finite values on the integrals and smoothness of the relevant quantities.

4. Maximal exponential model modeled on Orlicz-Sobolev spaces with Gaussian weight

It is clear from the preceding discussion that we need to introduce a class of random variables that ensures both the existence of the exponential manifold and the existence of derivatives. This is accomplished by the following definitions taken from [15].

Definition 9. The exponential and the mixture Orlicz-Sobolev-Gauss (OSG) spaces are, respectively,

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

$$W^{1,(\cosh -1)_*}(M) = \left\{ f \in L^{(\cosh -1)_*}(M) \middle| \partial_j f \in L^{(\cosh -1)_*}(M) \right\} , \tag{2}$$

where ∂_j, $j = 1, \ldots, n$, is the partial derivative in the sense of distributions.

As $\phi \in C_0^\infty(\mathbb{R}^n)$ implies $\phi M \in C_0^\infty(\mathbb{R}^n)$, for each $f \in W^{1,(\cosh -1)_*}(M)$ we have, in the sense of distributions, that

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

with $\delta_j \phi = (X_j - \partial_j)\phi$. The *Stein operator* δ_i acts on $C_0^\infty(\mathbb{R}^n)$.

The meaning of both operators ∂_j and $\delta_j = (X_j - \partial_j)$ when acting on square-integrable random variables of the Gaussian space is well known, but here we are interested in the action on OSG-spaces. Let us denote by $C_{\mathrm{p}}^\infty(\mathbb{R}^n)$ the space of infinitely differentiable functions with polynomial growth. Polynomial growth implies the existence of all M-moments of all derivatives, hence $C_{\mathrm{p}}^\infty(\mathbb{R}^n) \subset W^{1,(\cosh -1)_*}(M)$. If $f \in C_{\mathrm{p}}^\infty(\mathbb{R}^n)$, then the distributional derivative and the ordinary derivative are equal and moreover $\delta_j f \in C_{\mathrm{p}}^\infty(\mathbb{R}^n)$. For each $\phi \in C_0^\infty(\mathbb{R}^n)$ we have $\langle \phi, \delta_j f \rangle_M = \langle \partial_j \phi, f \rangle_M$.

The OSG spaces $W^1_{\cosh -1}(M)$ and $W^1_{(\cosh -1)_*}(M)$ are both Banach spaces. In fact, both the product functions $(u, x) \mapsto (\cosh -1)(u)M(x)$ and $(u, x) \mapsto (\cosh -1)_*(u)M(x)$ are ϕ-functions according the Musielak's definition. The norm on the OSG-spaces are the graph norms,

$$\|f\|_{W^1_{(\cosh -1)}(M)} = \|f\|_{L^{(\cosh -1)}(M)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(M)} \tag{3}$$

and

$$\|f\|_{W^1_{(\cosh -1)_*}(M)} = \|f\|_{L^{(\cosh -1)}(M)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(M)} . \tag{4}$$

We review some relations between OSG-spaces and ordinary Sobolev spaces. For all $R > 0$

$$(2\pi)^{-\frac{n}{2}} \geq M(x) \geq M(x)(|x| < R) \geq (2\pi)^{-\frac{n}{2}} \mathrm{e}^{-\frac{R^2}{2}} (|x| < R), \quad x \in \mathbb{R}^n.$$

Proposition 10. *Let $R > 0$ and let Ω_R denote the open sphere of radius R.*

1. *We have the continuous mappings*
$$W^{1,(\cosh -1)}(\mathbb{R}^n) \subset W^{1,(\cosh -1)}(M) \to W^{1,p}(\Omega_R), \quad p \geq 1.$$
2. *We have the continuous mappings*
$$W^{1,p}(\mathbb{R}^n) \subset W^{1,(\cosh -1)_*}(\mathbb{R}^n) \subset W^{1,(\cosh -1)_*}(M) \to W^{1,1}(\Omega_R), \quad p > 1.$$
3. *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.*

Proof of Item 3. See [4]. □

4.1. Hyvärinen divergence in the Gaussian space.

The Hyvärinen divergence between q and p in $\mathcal{E}$ is

$$D_{\mathrm{H}}(p, q) = \frac{1}{2} \int |\nabla \log q(x) - \nabla \log p(x)|^2 \, p(x)\gamma(x) \, dx .$$

As $\log q = v - K_1(v)$ and $\log p = u - K_1(v)$ we assume $u, v \in B_1$ to be differentiable in the sense of distributions with derivatives in $L^{(\cosh -1)}(1)$. It follows that the expression of the GH-divergence in the chart at 1 is

$$D_{\mathrm{H}}(u, v) = \frac{1}{2} \int |\nabla v - \nabla u|^2 \, \mathrm{e}^{u(x) - K_1(u)} \gamma(x) \, dx .$$

We proceed as in Hyvärinen computation by parts. First, decompose the squred norm of the difference to get

$$D_{\mathrm{H}}(u,v) = \frac{1}{2}\int |\nabla v(x)|^2 \, \mathrm{e}^{u(x)-K_1(u)}\gamma(x)\, dx \; - \int \nabla v(x)\cdot\nabla u(x)\mathrm{e}^{u(x)-K_1(u)}\gamma(x)\, dx \; + \frac{1}{2}\int |\nabla u(x)|^2 \, \mathrm{e}^{u(x)-K_1(u)}\gamma(x)\, dx \;\; .$$

The last term does not depend on v. If we write $\nabla u \mathrm{e}^{y-K_1(u)} = \nabla \mathrm{e}^{u-K_1(u)}$ and assume the equality $\partial_j^* = \delta_j$ is correct, the middle term is

$$-\int \nabla v(x)\cdot\nabla u(x)\mathrm{e}^{u(x)-K_1(u)}\gamma(x)\, dx \; = -\int \nabla v(x)\cdot\nabla \mathrm{e}^{u(x)-K_1(u)}\gamma(x)\, dx \; = -\int \delta\cdot\nabla v(x)\mathrm{e}^{u(x)-K_1(u)}\gamma(x)\, dx \; = -\,\mathbb{E}_p\left[\delta\nabla v\right] \;\; ,$$

where

$$\delta\cdot\nabla v(x) = \sum_{j=1}^n \delta_j \partial_j v(x) = -x\cdot\nabla v(x) - \Delta v(x) \;\; .$$

The formal derivative of $v \mapsto J(v) = \mathbb{E}_p\left[\frac{1}{2}|\nabla v|^2 - \delta\cdot\nabla v\right]$ in the direction h is

$$d_h J(v) = \mathbb{E}_p\left[\nabla h\cdot\nabla v - \delta\cdot\nabla h\right] \;\; .$$

References

1. Shun-ichi Amari, *Differential geometry of curved exponential families—curvatures and information loss*, Ann. Statist. **10** (1982), no. 2, 357–385. MR MR653513 (84g:62027)
2. Shun-ichi Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics, vol. 28, Springer-Verlag, 1985. MR 86m:62053
3. Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer, *Information geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], vol. 64, Springer, Cham, 2017. MR 3701408
4. Haim Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011. MR 2759829 (2012a:35002)
5. Alberto Cena and Giovanni Pistone, *Exponential statistical manifold*, Ann. Inst. Statist. Math. **59** (2007), no. 1, 27–56. MR MR2396032 (2009b:62011)
6. Aapo Hyvärinen, *Estimation of non-normalized statistical models by score matching*, J. Mach. Learn. Res. **6** (2005), 695–709. MR 2249836
7. Betrand Lods and Giovanni Pistone, *Information geometry formalism for the spatially homogeneous Boltzmann equation*, Entropy **17** (2015), no. 6, 4323–4363.
8. Paul Malliavin, *Integration and probability*, Graduate Texts in Mathematics, vol. 157, Springer-Verlag, 1995, With the collaboration of Hlne Airault, Leslie Kay and Grard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky. MR MR1335234 (97f:28001a)
9. ———, *Stochastic analysis*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 313, Springer-Verlag, 1997. MR MR1450093 (99b:60073)
10. Luigi Montrucchio and Giovanni Pistone, *Deformed exponential bundle: the linear growth case*, arXiv:1709.01430, 2017.
11. Julian Musielak, *Orlicz spaces and modular spaces*, Lecture Notes in Mathematics, vol. 1034, Springer-Verlag, 1983. MR 724434 (85m:46028)
12. Ivan Nourdin and Giovanni Peccati, *Normal approximations with Malliavin calculus*, Cambridge Tracts in Mathematics, vol. 192, Cambridge University Press, Cambridge, 2012, From Stein's method to universality. MR 2962301
13. Matthew Parry, A. Philip Dawid, and Steffen Lauritzen, *Proper local scoring rules*, Ann. Statist. **40** (2012), no. 1, 561–592. MR 3014317
14. Giovanni Pistone, *Nonparametric information geometry*, Geometric science of information (Frank Nielsen and Frédéric Barbaresco, eds.), Lecture Notes in Comput. Sci., vol. 8085, Springer, Heidelberg, 2013, First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings, pp. 5–36. MR 3126029
15. ———, *Information geometry of the Gaussian space*, arXiv:1803.08135, 2018.
16. Giovanni Pistone and Carlo Sempi, *An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one*, Ann. Statist. **23** (1995), no. 5, 1543–1561. MR 97j:62006
17. Marina Santacroce, Paola Siri, and Barbara Trivellato, *New results on mixture and exponential models by Orlicz spaces*, Bernoulli **22** (2016), no. 3, 1431–1447. MR 3474821

de Castro Statistics, Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino