ORLICZ AND ORLICZ-SOBOLEV EXPONENTIAL STATISTICAL BUNDLE

GIOVANNI PISTONE

Abstract. These are draft lecture notes about the key-words in the title and about how they interact to provide a convenient set-up for Information Geometry.

Contents

Date: Compiled 2019-10-11 16:39:18Z.

G.P. acknowledges the support of de Castro Statistics, Collegio Carlo Alberto, and GNAMPA-INdAM.

Part 1. General probability space

1. Nonparametric Information Geometry: A point of view

Most philosophical accounts of Chance, Randomness, and Probability also discuss Entropy (or Information) as an essential element of the whole picture. For example, see Hájek (2019). Chance and Randomness are metaphysical concepts, while probability is the mathematical model of Chance as it appears in the theory games. See, for example, Hacking (2006). Randomness belongs to the metaphysics of Statistical Physics and requires from the beginning a treatment of Entropy. See, for example, Sklar (1993).

Information Geometry (IG) has not jet found its way into philosophical literature but should have done. In fact, it is a refinement of the basic mathematical model of chance and randomness that uses a formalism that comes from Statistics. In particular, it explains why information and entropy appear necessarily in the theory. Moreover, it shows the intrinsic geometrical flavour of a mathematical model of Statistics.

I start with a simple presentation that any (analytic) philosopher could use if she is willing to. Then I move to more technical material but the aim is always a generic presentation of the formalism without touching to any specific application.

My basic references for IG are the proceedings Amari, Barndorff-Nielsen, Kass, Lauritzen, and Rao (1987) and the monographs Amari and Nagaoka (2000); Amari (2016); Ay, Jost, Lê, and Schwachhöfer (2017). The last one is probably the most comprehensive, and most of the topics in my presentation are actually to be found there, but a difference in emphasis. I try to avoid as much as possible the use of parametric notation and to follow the style of Lang (1995). In the last part on these notes, I focus on the case of smooth densities.

My general reference for standard Mathematical Statistics is Schervish (1995).

1.1. Probability simplex.

The first ingredient of IG is the the base set of the formalism, that is, the "probability simplex". Namely, the convex set $\mathcal{P}$ of all probability measures on a measurable space $(X, \mathcal{X})$. The probability simplex is a base set of the cone of finite (non-negative) measures $\mathcal{M}$. In turn, $\mathcal{M}$ is a subset of the Banach space of signed finite measures $\mathcal{S}$ with the norm of total variation. In this picture, the probability simplex is a closed convex subset of $\mathcal{S}$ whose affine space is the Banach sub-space $\mathcal{S}_0$ of signed finite measures with 0 total value.[1]

The conceptual point here is that the probability simplex is not a vector space hence the notions of affine geometry is not really adapted: Space and Chance are quite different methaphisical notions. In practice, statisticians know very well that one cannot look directly to frequences f; much better is to look at some "score", for example, $\log f$ or $\sqrt{f}$.

Let us look to the geometry of $\mathcal{P}$ from the point of view of kinematics. Let

$$I \ni\mapsto \mu(t) \in \mathcal{P} \subset \mathcal{S}$$

be a differentiable curve with derivative (affine velocity) $\dot{\mu}$. The affine velocity is a curve $t \mapsto \dot{\mu}(t) \in \mathcal{S}_0$, more precisely,

$$t \mapsto (\mu(t), \dot{\mu}(t)) \in \mathcal{P} \times \mathcal{S}_0 \quad .$$

The obvious property that at any t such that $\mu(t)$ is on the border of the probability simplex the affine velocity $\dot{\mu}(t)$ cannot point outside is best expressed in the following form, see Ay et al. (2017).

[1] Aside: Even if the actual object on interest is the probability, there are many reasons to chose to work in the larger set $\mathcal{M}$ of (non-negative) finite measures. In particular, from the applied point of view, it is frequently useful to deal with the projective structure of $\mathcal{M}$ as "unnormalized probabilities."

Proposition 1. *Let $I \ni\mapsto \mu(t) \in \mathcal{P}_1$ be a differentiable curve in $\mathcal{M}$ with derivative (velocity) $t \mapsto \dot\mu(t) \in \mathcal{S}_0$. For all t, the velocity $\dot\mu(t)$ is absolutely continuous with respect to the position $\mu(t)$.*

Proof. Let $A \in \mathcal{X}$ be an event such that $\mu(A)\colon s \mapsto \mu(A;s)$ is zero at $s = t$. Then t is a minimum point for $\mu(A)$, hence, the derivative is 0, $\dot\mu(A;t) = 0$. □

Using a term from the Fisherian Statistics, we call the Radon-Nikodim derivative the *score* of the curve $\mu(\cdot)$,

$$S\mu(t) = \frac{\dot\mu}{\mu} \quad \text{that is,} \quad \dot\mu(A;t) = \int S\mu(x;t)\ \mu(dx;t)\ , \quad t \in I, A \in \mathcal{X}\ . \tag{1}$$

Cf. the classical definition in § 2.3 of Schervish (1995).

The score is an element of $L^1_0(\mu(t))$. Its relevance is explained by the following computation. Let f be a bounded random variable. Then

$$\frac{d}{dt}\int f(x)\ \mu(dx;t) = \int f(x)\ \dot\mu(dx;t) = \int f(x) S\mu(x;t)\ \mu(dx;t) = \\ \int \left(f(x) - \int f(x)\ \mu(dx;t)\right) S\mu(x;t)\ \mu(dx;t)\ .$$

Or,

$$\frac{d}{dt}\int f\ d\mu(t) = \left\langle f - \int f\ d\mu(t), S\mu(t) \right\rangle_{\mu(t)}\ , \tag{2}$$

where we have introduced a separating duality between $L^\infty_0(\mu(t))$ and $L^1_0(\mu(t))$. Fisher's idea was to evaluate (to score) the variation of expected value of a statistics f along a statistical model with a scalar product at the "true distribution" $\mu(t)$ of the variation of f with a score function. The variance of the score at the "true distribution" is the Fisher Information,

$$t \mapsto \int |S\mu(x;t)|^2\ \mu(dx;t)\ .$$

Moreover, this set-up leads, for example, to the Cramer-Rao inequality.

The mathematical back-ground of applied mathematicians at Fisher's time was Mathematical Physics. Hence, we can be pressy sure that Fisher himself was quite conscious of geometrical meaning of his ouw contruction. Precisely, we can see the score term $S\mu(t)$ in eq. (2) as an alternative way to compute the velocity, due to the existence of constraints to the motion. From

$$0 = \dot\mu(X;t) = \int_X S\mu(x;t)\ \mu(dx;t)\ ,$$

we see that $\dot\mu(t) \in L^1_0(\mu(t))$, hence the score is a section of a vector bundle,

$$(\mu(t), \dot\mu(t)) \in \mathcal{P} \times L^1_)(\mu(t))\ .$$

In this geometrical perspective, $f \mapsto f - \int f\ d\mu(t)$ of eq. (2) is an alternative way to compute the gradient of the function $\mu \mapsto \int f\ d\mu$. This argument prompts for the following definition of *natural gradined.* This is the name introduced by Amari (1998) in a different but equivalent way.

Consider the following informal definition. We say that $F\colon \mathcal{P}_1 \to \mathbb{R}$ has natural gradient $\operatorname{grad} F\colon X \times \mathbb{R} \to \mathbb{R}$, if for all smooth curve $t \mapsto \mu(t) \in \mathcal{P}_1$ it holds

$$\frac{d}{dt}F(\mu(t)) = \int \operatorname{grad} F(x;t) S\mu(x;t)\ \mu(dx;t)\ .$$

Clearly, there are many loose end to take care of. This theory is classically presented decorated with sufficient technical conditions to make everything work smoothly. Our ambitions is to find some set of "natural" assumptions. Inspiration comes from the notion of exponential family, see, for example, the short version in § 2.2 of Schervish (1995), and the long version in Brown (1986).

1.2. **Exponential models.** Given the probability space $(X, \mathcal{X}, \mu)$, consider the exponential model

$$I \ni \theta \mapsto p_\theta = \mathrm{e}^{\theta u - \psi(\theta)} p \ , \tag{3}$$

where I is an open interval containing 0, u is a random variable, p is a probability density, and the cumulant function $\psi(\theta) = \log \int \mathrm{e}^{tu} \ p \cdot d\mu$ is assumed to be finite for all $t \in I$. This model is the venerable Botzmann-Gibbs factor of Statistical Physics if $\theta = -\frac{1}{kT}$, $T > 0$, see, for example, Landau and Lifshits (1980). This is the real historical origin of IG.

For each fixed θ, the model in eq. (3) identifies the sufficient statistics u up to a constant. in fact,

$$\mathrm{e}^{\theta u_1 - \psi_1(\theta)} = \mathrm{e}^{\theta u_2 - \psi_2(\theta)}$$

implies that $u_1 - u_2$ is a constant random variable.

Let us drop the paramenter for a moment. There are tree interesting way to identify a unique u-statistics for a density $q = \mathrm{e}^{u-\psi} p$.

1. One option is to assume $\int u \ p d\mu = 0$, so that

$$\int \log \frac{q}{p} \ p d\mu = \int (u - \psi) \ p d\mu = -\psi$$

and

$$u = \log \frac{q}{p} - \psi = \log \frac{q}{p} - \int \log p \frac{p}{q} \ d\mu \ .$$

Notice that in this case the normalizing constant is a KL divergence, $\psi = \int p \log \frac{p}{q} \ d\mu = \mathrm{D}\left(p \,\|q\right)$.

2. A second option is to assume $\int u \ q d\mu = 0$, so that

$$\int q \log \frac{q}{p} \ d\mu = \int (u - \psi) \ q d\mu = -\psi$$

and

$$u = \log \frac{q}{p} + \int q \log \frac{q}{p} \ d\mu$$

In this case, $-\psi = \int q \log \frac{p}{q} \ d\mu = \mathrm{D}\left(q \,\|p\right)$.

3. A third option is available in some cases. It is based on the assumption that $u \geq 0$ and $\min u = 0$. In this form, the computation of the limits $\theta \to \pm\infty$ is expecially simple.

From 1. and 2. it follows that the KL-divergences appear in the theory from the beginning as a direct consequence of the exponential representation. In particular, if $q = \mathrm{e}^{u-\psi} \cdot p$, then it is easy to see that

$$\mathrm{D}\left(p \,\|q\right) + \mathrm{D}\left(q \,\|p\right) = \mathbb{E}_\mu\left[(q-p)u\right] \ .$$

An important issue is the actual μ-integrability of $\mathrm{e}^{\theta u} \cdot p$ in eq. (3). This remark leads to the following definition. The MGF $M_p(u)$ is a convex analytic function in the interior of its proper domain. Moreover, the Cramer Class of $p \cdot \mu$ is a vector space. In conclusion, the sufficient statistics of *eq.* (3) form a vector space of random variables. We first discuss the properties of this space.

Definition 2. Consider the set of random variables u such that the moment generating function $\theta \mapsto M_p(u;\theta) = \int \mathrm{e}^{\theta u} \ d\mu$ is defined in a neigborhood of zero. Equivalently, there exists a positive $\alpha > 0$ such that $\int (\cosh(\alpha u) - 1) d\mu < +\infty$. This set is a Banach space whose closed unit ball is $\left\{u \,\middle|\, \int (\cosh u - 1) d\mu \leq 1\right\}$ denoted $L^{(\cosh - 1)}\left(\mu\right)$.

This actually defines a class of classical Banach spaces called Orlicz spaces, see for example the monograph by J. Musielak (Musielak, 1983, Ch. II). The same class has been used in Information Geometry to provide a model for statistical manifolds, see G. Pistone and C. Sempi Pistone and Sempi (1995), A. Cena and G. Pistone Cena and Pistone (2007), and M. Santacroce, P. Siri, and B. Trivellato Santacroce et al. (2016b).

Precisely, given a probability space $(\Omega, \mathcal{F}, \mu)$, the moment generating function of the random variable u is finite in a neighborhood of 0 if, and only if, $\mathbb{E}_\mu [\cosh(\lambda u) - 1] < \infty$ for some $\lambda > 0$. The class of such random variables is a vector space that we denote by $L^{(\cosh -1)}(\mu)$ and

$$L^{(\cosh -1)}(\mu) \ni u \mapsto \|u\|_{L^{(\cosh -1)}(\mu)} = \inf\left\{\alpha > 0 \,\middle|\, \mathbb{E}_\mu \left[\cosh(\alpha^{-1}u) - 1\right] \leq 1\right\}$$

is a complete norm. In particular, the closed unit ball is

$$\left\{u \in L^{(\cosh -1)}(\mu) \,\middle|\, \mathbb{E}\left(\cosh u\right) \leq 2\right\} \ .$$

From now on, we write briefly $\|\cdot\| = \|\cdot\|_{L^{(\cosh -1)}(\mu)}$. Let us consider some special cases.

In the case of a constant random variable $u = a \in \mathbb{R}$, $\cosh(a/\|a\|) = 2$, that is $\|a\| = |a| / \cosh^{-1} 2$. It would be possible to use an equivalent norm which reduced to the absolute value on constants, but our choice is more convenient in the computations to follow.

In the case of an indicator random variable, $u = a\mathbf{1}_A$,

$$\mathbb{E}_\mu \left[\cosh(\alpha^{-1}a\mathbf{1}_A)\right] = 1 - \mu(A) + \cosh(\alpha^{-1}a)\mu(A) \ ,$$

hence, $\|a\mathbf{1}_A\| = |a| / \cosh^{-1}\left((1 + \mu(A))/\mu(A)\right)$.

Both cases above belong to the general class of random variable whose moment generating function is finite everywhere. In such a case, one has to solve for $\alpha = \|u\|$ the equation

$$\mathbb{E}_\mu \left[\mathrm{e}^{u/\alpha}\right] + \mathbb{E}_\mu \left[\mathrm{e}^{-u/\alpha}\right] = 4 \ .$$

We will discuss below the important issue of the equivalence of the Orlicz norms for different μ's.

1.3. **Cramer class, sub-exponential random variable, Orlicz space.** Two very recent monograph (Vershynin, 2018, Ch. 2) and (Wainwright, 2019, Ch. 2) discuss concentration inequalities for randon variables in the exponential Orlicz space $L^{(\cosh -1)}(\mu)$.[2] A basic statement is reproduced below.

Proposition 3. *Let u be a random variable of the probability space $(X, \mathcal{X}, \mu)$. The following conditions are equivalent.*

(1) *The moment generating function of u is finite in a neighborhood of 0.*
(2) *The random variable u is* sub-exponential*, namely, there exist constants $c_1 \geq 1$ and $c_2 > 0$ such that $\mathbb{P}(|u| \geq t) \leq c_1 \mathrm{e}^{-c_2 t}$.*
(3) *It holds $\sup_{k \geq 1}(\mathbb{E}\left(|u|^k\right)/k!)^{1/k} = c < \infty$.*

Proof: (1) *implies* (2). For all $t > 0$ and each $\lambda > 0$ such that both $\mathbb{E}\left(\mathrm{e}^{\lambda u}\right), \mathbb{E}\left(\mathrm{e}^{-\lambda u}\right) < \infty$,

$$\mathbb{P}(u \geq t) = \mathbb{P}\left(\mathrm{e}^{\lambda u} \geq \mathrm{e}^{\lambda t}\right) \leq \mathrm{e}^{-\lambda t}\, \mathbb{E}\left(\mathrm{e}^{\lambda u}\right) \ ,$$

and

$$\mathbb{P}(u \leq -t) = \mathbb{P}(-\lambda u \geq \lambda t) = \mathbb{P}\left(\mathrm{e}^{-\lambda u} \geq \mathrm{e}^{\lambda t}\right) \leq \mathrm{e}^{-\lambda t}\, \mathbb{E}\left(\mathrm{e}^{-\lambda u}\right) \ .$$

We have

$$\mathbb{P}(|u| \geq t) = \mathbb{P}(u \geq t) + \mathbb{P}(u \leq -t) \leq 2\mathrm{e}^{-\lambda t}\, \mathbb{E}\left(\cosh(\lambda u)\right) \ ,$$

and we can take $c_1 = 2\,\mathbb{E}\left(\cosh(\lambda u)\right)$ and $c_2 = \lambda$. □

[2] I have introduced this space in IG in the ninties propted by the argument reproduced above about exponential families. It took to me about 25 years to realize that this assumption is not only imposted by the problem itself, but has some interesting implication from the point of view of applications

Proof: (2) *implies* (3). We have

$$\mathbb{E}\left(|u|^k\right) = \int_0^\infty \mathbb{P}\left(|u|^k > t\right)\, dt = \int_0^\infty \mathbb{P}\left(|u| > s\right) ks^{k-1}\, ds \le kc_1 \int_0^\infty s^{k-1}\mathrm{e}^{-c_2 s}\, ds =$$

$$kc_1 \int_0^\infty \left(\frac{u}{c_2}\right)^{k-1} \mathrm{e}^{-u}\frac{1}{c_2}\, du = k\frac{c_1}{c_2^k}\int_0^\infty u^{k-1}\mathrm{e}^{-u}\, du = k\frac{c_1}{c_2^k}(k-1)! = k!\frac{c_1}{c_2^k} \ .$$

Notice that $c_1 \le c_1^k$ because $c_1 \ge 1$, so that $(\mathbb{E}\left(|u|^k\right)/k!)^{1/k} \le c_1/c_2$ for all k. □

Proof: (3) *implies* (1). If $\mathbb{E}\left(|u|^k\right)/k! \le c^k$ and $0 < \lambda < 1/c$, then

$$\mathbb{E}\left(\mathrm{e}^{\lambda|u|}\right) = 1 + \sum_{k=1}^\infty \frac{\mathbb{E}\left(|u|^k\right)}{k!}\lambda^k \le \sum_{k=0}^\infty (c\lambda)^k = \frac{1}{1-c\lambda} \ .$$

□

The sub-exponential inequality of Prop. 3(2) takes the following form when one express the constants in terms of the Orlicz norm.

For all $u \in L^{(\cosh -1)}(\mu)$ and all $t > 0$, we have,

$$\mathbb{P}\left(|u| \ge t\right) = \mathbb{P}\left(\frac{|u|}{\|u\|} \ge \frac{t}{\|u\|}\right) =$$

$$\mathbb{P}\left(\cosh\left(\frac{u}{\|u\|}\right) \ge \cosh\left(\frac{t}{\|u\|}\right)\right) \le$$

$$2/\cosh\left(t/\|u\|\right) \le 4\exp\left(-\|u\|^{-1} t\right) \ .$$

The bound on the absolute moment of Prop. 3(3) becomes

$$\mathbb{E}\left(|u|^k\right) = \int_0^\infty \mathbb{P}\left(|u|^k > t\right)\, dt = \int_0^\infty \mathbb{P}\left(|u| > s\right) ks^{k-1}\, ds \le 4k\int_0^\infty s^{k-1}\mathrm{e}^{-s/\|u\|}\, ds =$$

$$4k\|u\|^k \int_0^\infty u^{k-1}\mathrm{e}^{-u}\, du = 4k!\,\|u\|^k \ ,$$

so that the bound is

$$\left(\mathbb{E}\left(\frac{|u|^k}{k!}\right)\right)^{1/k} \le 4\|u\| \ . \tag{4}$$

An important application of sub-exponentiality i.e., Orlicz norm, is to provide warranties in the law of large numbers for small samples.

Proposition 4 (Centered sub-exponential u)**.** *If the sub-exponential random variable u is centered, $\mathbb{E}(u) = 0$, then for all $\lambda \le 1/(8\|u\|)$ it holds*

$$\mathbb{E}\left(\mathrm{e}^{\lambda|u|}\right) \le \mathrm{e}^{32\|u\|^2\lambda^2} \ .$$

Proof. We use the bound in Eq. (4). For each $\lambda < \frac{1}{2\|u\|}$ we have

$$\mathbb{E}\left(\mathrm{e}^{\lambda u}\right) \le 1 + \sum_{k=2}^\infty \frac{\mathbb{E}\left(|u|^k\right)}{k!}\lambda^k \le 1 + \sum_{k=2}^\infty (4\|u\|\lambda)^k = 1 + \frac{(4\|u\|\lambda)^2}{1-4\|u\|\lambda} \le 1 + 32(\|u\|\lambda)^2 \le \mathrm{e}^{32\|u\|^2\lambda^2} \ .$$

□

Proposition 5 (Bernstein inequality)**.** *Let be the given sub-exponential centered independent random variables $u_1, \dots, u_n$ and let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Then, for all $t \geq 0$, it holds*

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i u_i\right| \geq t\right) \leq 2 \exp\left(-\inf\left(\frac{t^2}{128 \sum_{i=1}^n |a_i|^2 \|u_i\|^2}, \frac{t}{16 \max_i |a_i| \|u_i\|}\right)\right) .$$

Proof. From Prop. 4, we have, for all $t > 0$, and $0 < \lambda \leq \dfrac{1}{8|a_i|\|u_i\|}$, $i = 1, \dots, n$, that is, $0 < \lambda \leq \dfrac{1}{8 \max_i |a_i| \|u_i\|}$. that

$$\mathbb{P}\left(\sum_{i=1}^n a_i u_i \geq t\right) \leq \mathrm{e}^{-\lambda t} \mathbb{E}\left(\exp\left(\sum_{i=1}^n \lambda a_i u_i\right)\right) = \mathrm{e}^{-\lambda t} \prod_{i=1}^n \mathbb{E}\left(\mathrm{e}^{\lambda a_i u_i}\right) \leq$$

$$\mathrm{e}^{-\lambda t} \prod_{i=1}^n \mathrm{e}^{32|a_i|^2\|u_i\|^2\lambda^2} = \mathrm{e}^{-\lambda t} \exp\left(32 \sum_{i=1}^n |a_i|^2 \|u_i\|^2 \lambda^2\right) = \exp\left(-\lambda t + 32 \sum_{i=1}^n |a_i|^2 \|u_i\|^2 \lambda^2\right) .$$

Let us find the minimum of the parabola $\lambda \mapsto -\lambda t + 32\left(\sum_{i=1}^n |a_i|^2 \|u_i\|^2\right)\lambda^2$ under the constraint $0 \leq \lambda \leq \dfrac{1}{8 \max_i |a_i| \|u_i\|}$. The minimum is at $\dfrac{t}{64 \sum_{i=1}^n |a_i|^2 \|u_i\|^2} \wedge \dfrac{1}{8 \max_i |a_i| \|u_i\|}$. The value in the first point is $-\frac{t^2}{128 \sum_{i=1}^n |a_i|^2\|u_i\|^2}$. The line through the vertex of the parabola is $\lambda \mapsto -\frac{t}{2}\lambda$, hence it is above the parabola in the positive interval up to the point of minimum and it is below the minimum otherwise. It follows that the minimum value is bounded by

$$-\frac{t}{2} \inf\left(\frac{t}{64 \sum_{i=1}^n |a_i|^2 \|u_i\|^2}, \frac{1}{8 \max_i |a_i| \|u_i\|}\right) = -\inf\left(\frac{t^2}{128 \sum_{i=1}^n |a_i|^2 \|u_i\|^2}, \frac{t}{16 \max_i |a_i| \|u_i\|}\right)$$

The same bound applies to the other half of the inequality. □

Proposition 6 (Law of large numbers)**.** *Let be the given sub-exponential centered independent random variables $u_1, \dots, u_n$. Then, for all $t \geq 0$, it holds*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n u_i\right| \geq t\right) \leq 2 \exp\left(-n \inf\left(\frac{t^2}{128 \max_{i=1} \|u_i\|^2}, \frac{t}{16 \max_i \|u_i\|}\right)\right) .$$

Proof. Let us apply Prop. 5 with $a_i = 1/n$. The bound becomes

$$\frac{t^2}{128 \sum_{i=1}^n \frac{1}{n^2} \|u_i\|^2} \wedge \frac{t}{16 \max_i \frac{1}{n} \|u_i\|} = \frac{n^2 t^2}{128 \sum_{i=1}^n \|u_i\|^2} \wedge \frac{nt}{16 \max_i \|u_i\|} \geq$$

$$n\left(\frac{t^2}{128 \max_{i=1} \|u_i\|^2} \wedge \frac{t}{16 \max_i \|u_i\|}\right) .$$

□

1.4. **Orlicz spaces.** If $\Phi(x) = \cosh x - 1$, a real random variable u belongs to the vector space $L^\Phi(p)$ if $\mathbb{E}_p[\Phi(\alpha v)] < +\infty$ for some $\alpha > 0$. A norm is obtained by defining the set $\{v \mid \mathbb{E}_p[\Phi(v)] \leq 1\}$ to be the closed unit ball. It follows that the open unit ball consists of those u's such that αu is in the closed unit ball for some $\alpha > 1$. The corresponding norm $\|\cdot\|_{\Phi,p}$ is called Luxemburg norm and defines a Banach space, see e.g. (Musielak, 1983, Th 7.7). The function $\cosh -1$ has been chosen here because the condition $\mathbb{E}_p[\Phi(\alpha v)] < +\infty$ is clearly equivalent to $\mathbb{E}_p\left[\mathrm{e}^{tv}\right] < +\infty$ for $t \in [-\alpha, \alpha]$, but other choices will define the same Banach space e.g., $\Phi(x) = \mathrm{e}^{|x|} - |x| - 1$. By abuse of notation, we will denote all these equivalent functions by Φ.

The main technical issue in working with Orlicz spaces such as $L^{(\cosh -1)}(p)$ is the regularity of its unit sphere $S = \left\{u \,\middle|\, \|u\|_{(\cosh -1),p} = 1\right\}$. In fact, while $\mathbb{E}_p[\cosh u - 1] = 1$ implies $u \in S$, the latter implies $\mathbb{E}_p[\cosh u - 1] \leq 1$. Subspaces of L^Φ where this cannot happen are called *steep*, see examples in Pistone (2013b). If the state space is finite, the full space is steep.

The relevance of steep families in exponential families is discussed in Barndorff-Nielsen (1978). Steepness is important when related with the idea of embedding. Consider the mapping $\Phi_+^{-1} \colon \mathcal{P}_> \ni p \mapsto v = \Phi_+^{-1}(p)$, $\Phi_+ = \Phi_{|\mathbb{R}_>}$. Then $\int \Phi(v)\ d\mu = \int p\ d\mu = 1$ hence $\|u\|_\Phi = 1$ and we have an embedding of $\mathcal{P}_>$ into the sphere of a Banach space.

If the functions Φ and Φ_* are Young pair, for each $u \in L^\Phi(p)$ and $v \in L^{\Phi_*}(p)$, such that $\|u\|_{\Phi,p}, \|v\|_{\Phi_*,p} \leq 1$, we have $\mathbb{E}_p[uv] \leq 2$, hence

$$L^{\Phi_*}(p) \times L^\Phi(p) \ni (v,u) \mapsto \mathbb{E}_p[uv]$$

is a duality mapping, $\left|\langle u, v\rangle_p\right| \leq 2\|u\|_{\Phi_*,p}\|v\|_{\Phi,p}$.

A sequence u_n, $n = 1, 2, \dots$ is convergent to 0 for such a norm if and only if for all $\epsilon > 0$ there exists a $n(\epsilon)$ such that $n > n(\epsilon)$ implies $\mathbb{E}_p\left[\Phi_1(\frac{u_n}{\epsilon})\right] \leq 1$. Note that $|u| \leq |v|$ implies

$$\mathbb{E}_p\left[\Phi_1\left(\frac{u}{\|v\|_{\Phi_1,p}}\right)\right] \leq \mathbb{E}_p\left[\Phi_1\left(\frac{v}{\|v\|_{\Phi_1,p}}\right)\right] \leq 1$$

so that $\|u\|_{\Phi_1,p} \leq \|v\|_{\Phi_1,p}$.

In defining our manifold, we need to show that Orlicz spaces defined at different points of statistical models are isomorphic, we will use frequently the fact that following lemma, see (Cena and Pistone, 2007, Lemma 1).

Lemma 7. *Let $p \in \mathcal{M}$ and let Φ_0 be a Young function. If the Orlicz spaces $L^{\Phi_0}(p)$ and $L^{\Phi_0}(q)$ are equal as sets, then their norms are equivalent.*

The condition $u \in L^{\cosh -1}(p)$ is equivalent to the existence of the moment generating function $g(t) = \mathbb{E}_p\left[\mathrm{e}^{tu}\right]$ on a neighbourhoods of 0. The case when such a moment generating function is defined on all of the real line is special and defines a notable subspace of the Orlicz space see e.g., Rao and Ren (2002). Such spaces could be the model of an alternative definition of as in Grasselli (2001).

In fact, the Banach space $L^\Phi(p)$, $\phi = \cosh -1$ is not separable, unless the basic space has a finite number of atoms. In this sense it is an unusual choice from the point of view of functional analysis and manifold's theory. However, $L^\Phi(p)$ is natural for statistics because for each $u \in L^{\Phi_1}(p)$ the Laplace transform of u is well defined at 0, then the one-dimensional exponential model $p(\theta) \propto \mathrm{e}^{\theta u}$ is well defined.

However, the space $L^{\Phi_*}(p)$ is separable and its dual space is $L^\Phi(p)$, the duality pairing being $(u, v) \mapsto \mathbb{E}_p[uv]$. This duality extends to a continuous chain of spaces:

$$L^{\Phi_1}(p) \to L^a(p) \to L^b(p) \to L^{\Psi_1}(p), \quad 1 < b \leq 2, \quad \frac{1}{a} + \frac{1}{b} = 1$$

where $\to$ denotes continuous injection.

From the duality pairing of conjugate Orlicz spaces and the characterization of the closed unit ball it follows a definition of dual norm on $L^{\Phi_*}(p)$:

$$N_p(v) = \sup\left\{\mathbb{E}_p[uv] \mid \mathbb{E}_p[\Phi(u)] \leq 1\right\}.$$

2. Non parametric Information Geometry: exponential bundle

2.1. **Moment generating functional and cumulant generating functional.** In this section we review a number of key technical results. Most of the results are related with the smoothness of the superposition operator $L^\Phi(p) \colon v \mapsto \exp \circ v$. Superposition operators on Orlicz spaces are discussed e.g. in Krasnosel'skii and Rutickii (1961) and (Appell and Zabrejko, 1990, Ch 4). Banach analytic functions are discussed in Bourbaki (1971), Upmeier (1985) and Ambrosetti and Prodi (1993).

Let $p \in \mathcal{P}_>$ be given. The following theorem has been proved in (Cena, 2002, Ch 2), see also Cena and Pistone (2007).

Proposition 8.

(1) *For $a \geq 1$, $n = 0, 1, \dots$ and $u \in L^{\Phi}(p)$,*

$$\lambda_{a,n}(u)\colon\ (w_1, \dots, w_n) \mapsto \frac{w_1}{a} \cdots \frac{w_n}{a}\ \mathrm{e}^{\frac{u}{a}}$$

is a continuous, symmetric, n-multi-linear map from $L^{\Phi}(p)$ to $L^{a}\left(p\right)$.

(2) $v \mapsto \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{v}{a}\right)^n$ *is a power series from $L^{\Phi}(p)$ to $L^a(p)$ with radius of convergence ≥ 1.*

(3) *The superposition mapping $v \mapsto \mathrm{e}^{v/a}$ is an analytic function from the open unit ball of $L^{\Phi}(p)$ to $L^a(p)$.*

Definition 9. Let $\Phi = \cosh - 1$ and $B_p = L^{\Phi}_0(p)$, $p \in \mathcal{P}_>$. The *moment generating functional* is $M_p\colon L^{\Phi}(p) \ni u \mapsto \mathbb{E}_p\left[\mathrm{e}^u\right] \in \mathbb{R}_> \cup \{+\infty\}$. The *cumulant generating functional* is $K_p\colon B_p \ni u \mapsto \log M_p(u) \in \mathbb{R}_> \cup \{+\infty\}$.

Proposition 10.

(1) $M_p(0) = 1$*; otherwise, for each centered random variable $u \neq 0$, $M_p(u) > 1$.*

(2) M_p *is convex and lower semi-continuous, and its proper domain is a convex set which contains the open unit ball of $L^{\Phi}(p)$; in particular the interior of such a domain is a non empty convex set.*

(3) M_p *is infinitely Gâteaux-differentiable in the interior of its proper domain, the nth-derivative at u in the direction $v \in L^{\Phi}(p)$ being*

$$\left.\frac{d^n}{dt^n} M_p(u + tv)\right|_{t=0} = \mathbb{E}_p\left[v^n \mathrm{e}^u\right];$$

(4) M_p *is bounded, infinitely Fréchet-differentiable and analytic on the open unit ball of $L^{\Phi}(p)$, the nth-derivative at u evaluated in $(v_1, \dots, v_n) \in L^{\Phi}(p) \times \cdots \times L^{\Phi}(p)$ is*

$$D^n M_p(u)(v_1, \dots, v_n) = \mathbb{E}_p\left[v_1 \cdots v_n \mathrm{e}^u\right].$$

Proposition 11.

(1) $K_p(0) = 0$*; otherwise, for each $u \neq 0$, $K_p(u) > 0$.*

(2) K_p *is convex and lower semi-continuous, and its proper domain is a convex set which contains the open unit ball of B_p; in particular the interior of such a domain is a non empty convex set.*

(3) K_p *is infinitely Gâteaux-differentiable in the interior of its proper domain.*

(4) K_p *is bounded, infinitely Fréchet-differentiable and analytic on the open unit ball of $\mathcal{V}_p$.*

Other properties of the key functional K_p are described below as they relate directly to the exponential manifold.

2.2. **Families of Orlicz spaces.** Let $\mathcal{P}_+(\mu)$ be the set of positive probability densities of the measure space $(X, \mathcal{X}, \mu)$. As explained above, we associate to each density p a space of p-centered random variables to represent scores of one-dimensional statistical models. That is, if the one-parameter statistical model $I \ni t \mapsto p(t) \in \mathcal{P}_+(\mu)$, I open interval, is regular enough, then $u(t) = \frac{d}{dt} \log p(t)$ satisfies $\mathbb{E}_{p(t)}\left[u(t)\right] = 0$ for all $t \in I$. In particular, if $p(t) = \mathrm{e}^{tu - \psi(t)} \cdot p$, with $u \in B_p = \left\{u \in L^{(\cosh - 1)}\left(p\right) \middle| \mathbb{E}_p\left[u\right] = 0\right\}$.

It is crucial to discuss how the relevant spaces of p-centered random variables depend on the variation of the density p, that is it is crucial to understand the variation of the spaces $B_p = L^{\Phi}_0(p)$ and ${}^*B_p = L^{\Phi_*}_0(p)$ along a one-dimensional statistical model $p(t)$, $t \in I$. In Information Geometry, those spaces contain models for the tangent and pre-tangent spaces of the statistical models. We require that they must be isomophic at two different points of a regular model, they must be isomorphic. In particular, they must be equal with equivalent norms.

We use a peculiar notion of connection by arcs, which is different from what is usually meant with this name. Given $p, q \in \mathcal{P}_>$, the exponential model $p\left(\theta\right) \propto p^{1-\theta} q^{\theta}$, $0 \leq \theta \leq 1$ connects the two given densities as end points of a curve, $p(\theta) \propto \exp\left(\theta \log \frac{q}{p}\right) \cdot p$, where $\log \frac{q}{p}$ is not in the exponential Orlicz space at p unless θ can be extended to assume negative values.

Definition 12. We say that $p, q \in \mathcal{P}_>$ are connected by an open exponential arc if there exist $r \in \mathcal{P}_>$ and an open interval I, such that $p(t) \propto \mathrm{e}^{tu} r$, $t \in I$, is an exponential model containing both p and q at t_0, t_1 respectively. By the change of parameter $s = t - t_0$, we can always reduce to the case where $r = p$ and $u \in L^{\Phi}(p)$.

The open connection of Def. 12 is an equivalence relation.

Definition 13. Let us denote by $\mathcal{S}_p$ the interior of the proper domain of the cumulant generating functional K_p. For every density $p \in \mathcal{P}_>$, the *maximal exponential model at* p is defined to be the family of densities

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

Proposition 14. *The following statements are equivalent:*

(1) $q \in \mathcal{M}$ *is connected to* p *by an open exponential arc;*
(2) $q \in \mathcal{E}(p)$*;*
(3) $\mathcal{E}(p) = \mathcal{E}(q)$*;*
(4) $\log \frac{q}{p}$ *belongs to both* $L^{\Phi_1}(p)$ *and* $L^{\Phi_1}(q)$*.*
(5) $L^{\Phi_1}(p)$ *and* $L^{\Phi_1}(q)$ *are equal as vector spaces and their norms are equivalent.*

In the following proposition we have collected a number of properties of the maximal exponential model $\mathcal{E}(p)$ which are relevant for its manifold structure.

Proposition 15. *Assume* $q = \mathrm{e}^{u - K_p(u)} \cdot p \in \mathcal{E}(p)$*.*

(1) *The first two derivatives of* K_p *on* $\mathcal{S}_p$ *are*

$$DK_p(u)v = \mathbb{E}_q[v],$$
$$D^2 K_p(u)(v_1, v_2) = \mathrm{Cov}_q(v_1, v_2)$$

(2) *The random variable* $\frac{q}{p} - 1$ *belongs to* ${}^*B_p$ *and*

$$DK_p(u)v = \mathbb{E}_p\left[\left(\frac{q}{p} - 1\right) v\right].$$

In other words the gradient of K_p *at* u *is identified with an element of* ${}^*B_p$*, denoted by* $\nabla K_p(u) = e^{u - K_p(u)} - 1 = \frac{q}{p} - 1$*.*

(3) *The mapping* $B_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$ *is monotonic, in particular one-to-one.*
(4) *The weak derivative of the map* $\mathcal{S}_p \ni u \mapsto \nabla K_p(u) \in {}^*B_p$ *at* u *applied to* $w \in B_p$ *is given by*

$$D(\nabla K_p(u))w = \frac{q}{p}(w - \mathbb{E}_q[w]),$$

and it is one-to-one at each point.

(5) *The mapping* ${}^m\mathbb{U}_p^q : v \mapsto \frac{p}{q} v$ *is an isomorphism of* ${}^*B_p$ *onto* ${}^*B_q$*.*
(6) $q/p \in L^{\Phi_*}(p)$*.*
(7) $D(q\|p) = DK_p(u)u - K_p(u)$ *with* $q = \mathrm{e}^{u - K_p(u)} p$*, in particular* $-D(q \,\|\, p) < +\infty$*.*
(8)

$$B_q = L_0^{\Phi_1}(q) = \left\{ u \in L^{\Phi_1}(p) \,\middle|\, \mathbb{E}_p\left[u \frac{q}{p}\right] = 0 \right\}.$$

(9) ${}^e\mathbb{U}_p^q$: $u \mapsto u - \mathbb{E}_q[u]$ *is an isomorphism of* B_p *onto* B_q*.*

2.3. Hilbert bundle, exponential bundle.

3. Nonparametric Information Geometry: Smooth densities

Any statistical method that requires the computation of a density function p at a given sample point x requires the continuity of the density to ensure the existence of $p(x) = \int p \ d\delta_x$, that is, the basic sampling operation. Some applications require the computation of derivatives at a given sample point. In this section, I will consider some examples of this type in order to motivate a further specification of the formalism of the exponential statistical bundle.

The basic measure space is $(\mathbb{R}^n, \mathcal{B}, \mu)$. The reference measure is either the Lebesgue measure or, the Gaussian probability measure. Here $\mathcal{E}\left(\mu\right)$ denotes the exponential manifold at μ and $S\mathcal{E}\left(\mu\right)$ is the exponential bundle.

(1) A inner product on the Hilbert bundle of probability simplex that involves the space derivatives has been introduced by Otto (2001) in the context of evolution equations. The construction is not fully formalised in that paper. One full development appears in Ambrosio, Gigli, and Savaré (2008). Another possible setup is developed Lott (2008), who, in turn, refers to Kriegl and Michor (1997). I try to retell the story in the language of exponential manifolds in section 3.1 below.

(2) An interesting similar case was independently developed in Hyvärinen (2005) and Parry, Dawid, and Lauritzen (2012). They define statistical divergences that depend on derivatives and prompt for a non-Fisherian Information Geometry, that is, a construction based on a different inner product on the fibers of the relevant bundle. The classical geometry of divergences is discussed in § 2.7 of Ay et al. (2017). See sections 3.2 and 3.3. below. Compare § 13.6 of Amari (2016)

(3) Parabolic equations

3.1. F. Otto's metric.

Here is Otto's definition in the context of the the statistical bundle $S\mathcal{E}\left(\mu\right)$ with $\mu(dx) = dx$. For each $\rho \in \mathcal{E}\left(\mu\right)$ and

$$s \in S_1\mathcal{E}\left(\mu\right) = \left\{ s \in L^{(\cosh -1)}\left(\mu\right) \middle| \mathbb{E}_1\left[s\right] = \int s(x)dx = 0 \right\} \ ,$$

assume there is a solution to the elliptic equation

$$-\nabla \cdot (\rho \nabla u) = s \ , \quad u \in S_\rho\mathcal{E}\left(\mu\right) \ . \tag{5}$$

Notice that the more usual border condition is replaced by an integral condition. This implies

$$\int s(x)\phi(x) \ dx = \int \nabla u(x) \cdot \nabla\phi(x) \ \rho(x)dx$$

for all test functions ϕ and prompts for the definition of the following family of inner product on the statistical bundle:

$$S_\rho\mathcal{E}\left(p\right) \times S_\rho\mathcal{E}\left(p\right) \ni (u,v) \mapsto \left\langle\left\langle u, v\right\rangle\right\rangle_\rho = \mathbb{E}_\rho\left[\nabla u \cdot \nabla v\right] \ . \tag{6}$$

If the velocity of the curve $t \mapsto \rho(t) \in \mathcal{E}\left(\mu\right)$ is computed with the score $S\rho(t) = \frac{d}{dt}\log\rho(t)$, we want the natural gradient of the mapping $F(\rho) = \mathbb{E}_\rho\left[f\right]$ with respect to the inner product of eq. (6). We know how to compute the natural gradient in the statistical bundle:

$$\frac{d}{dt}F(\rho(t)) = \int (f - \mathbb{E}_{\rho(t)}\left[f\right])S\rho(t) \ \rho(t)d\mu \ .$$

Now consider that $\int (f - \mathbb{E}_\rho\left[f\right])\rho \ d\mu = 0$, so that it is possible to assume that $(f - \mathbb{E}_\rho\left[\right)]\rho \in S_1\mathcal{E}\left(\mu\right)$ and hence solve eq. (5). In conclusion, if V is the section of $S\mathcal{E}\left(\mu\right)$ defined by

$$-\nabla \cdot (\rho\nabla V(\rho)) = (f - \mathbb{E}_\rho\left[f\right])\rho \ , \quad V(\rho) \in S_\rho\mathcal{E}\left(\mu\right) \ ,$$

then

$$\left\langle\left\langle V(\rho(t)), S\rho(t)\right\rangle\right\rangle_{\rho(t)} = -\int \nabla\cdot(\rho(t)\nabla V(\rho(t)))S\rho(t) \ d\mu = \int (f - \mathbb{E}_{\rho(t)}\left[f\right])\rho(t)S\rho(t) = \frac{d}{dt}F(\rho(t)) \ .$$

A similar computation provides the natural gradient for other funcional of interest, for example, the entropy $\mathcal{H}\left(\rho\right)$.

3.2. **Local scoring rules.** Consider a statistical model $\mathcal{C}$ on a real measure space $(\mathbb{R}^n, \mu)$ and assume each density in the model is positive, continuous and differentiable in some (possibly weak) sense up to an order k.

Definition 16. A *local scoring rule* is a mapping $S\colon \mathcal{C}$ with values in Borel function $x \mapsto S(x; q)$ which is k-local, that is, such that $S(x; q)$ depends on the value at x of q and the derivatives up to order k. Moreover, assume that the *risk* under a positive $p \in \mathcal{C}$ is well defined as $d(p, q) = E_p(S(q))$. The local scoring rule $q \mapsto S(q)$ is *proper* if $q \mapsto d(p, q)$ is minimized at $q = p$ only, that is, $d(p, q) \geq d(p, p)$ and $d(p, q) = d(p, p)$ implies $q = p$. The *divergence* associated to S is $D(p, q) = d(p, q) - d(p, p)$.

Remark 1. The minimization of $q \mapsto D(p, q)$ is equivalent to the minimization of $q \mapsto d(p, q)$. But notice that d and D are not equivalent from the statistician's point of view. In fact, there is a sampling version of the risk namely, $\hat{d}(q) = \sum_{j=1}^N S(X_j, q)$ with (X_j) IID p. Moreover, $\hat{q} = \operatorname{argmin} \hat{d}(q)$ is an estimator of p. On the other side, $D(p, q)$ has no sampling version.

3.2.1. *Example: log-score.* It is possible to formalize in this language the construction of the Kullback-Leibler divergence. Take $\mathcal{C}$ be the set of all possible continuous and bounded densities on the Borel space $(X, \mathcal{X}, \mu)$ and define the 0-local scoring rule $S(x, q) = -\log q(x)$. The expectation $E_p(-\log q)$ is finite because all densities are bounded.[3] Clearly,

$$
\begin{aligned}
d(p,q) = -\int p(x) \log q(x)\ \mu(dx) = \\
\int \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} q(x)\ \mu(dx) - \int p(x) \log p(x)\ \mu(dx) \geq \\
\int (p(x) - q(x))\ \mu(dx) + d(p,p) = d(p,p)\ .
\end{aligned}
$$

The divergence can be translated to the minimum value to get a non-negative divergence, which is precisely the Kullback-Leibler divergence,

$$
d(p,q) - d(p,p) = \int p(x) \log \frac{p(x)}{q(x)}\ \mu(dx) = \int \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} q(x)\ \mu(dx) = \mathrm{D}\left(p \,\|q\right)\ .
$$

Conversely, we could procede the other way round. 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

$$
\mathrm{D}\left(p \,\|q\right) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)\ \mu(dx) \geq f\left(\int \frac{p(x)}{q(x)} q(x)\ \mu(dx)\right) = f(1) = 0
$$

and $\mathrm{D}\left(p \,\|q\right) = 0$ implies $p/q = 1$.

From the KL-divergence one could recover the score, which is that part of the divergence that has a sample version. 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 $p, q \in \mathcal{M}$. A warranty on convergence requires further assumptions on $\log q$, for example, sub-exponentiality. This is the case when the model $\mathcal{C}$ is a sub-model of an exponential model.

3.3. **Hyvärinen divergence.** Here is our main case.

Definition 17 (Hyvärinen divergence)**.** Let us assume now that the sample space is the n-dimensional real space and each density $q \in \mathcal{C}$ 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

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

[3] Weaker assumptions are enough. As $-\log q \geq 1 - q$, $D(p, q)$ is well defined, possibly $+\infty$, if $\int q(x)p(x)\ \mu(dx) < +\infty$ for all $p, q \in \mathcal{C}$ which is the case if $\mathcal{C} \subset L^2(\mu)$.

Let us see that this divergence comes from a local scoring rule. By expanding the squared norm of the difference, we obtain

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

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

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

Recall that minimization of the expected Hyvärinen score is the same as minimization of the Hyvärinen divergence.

The computations above were done first in Hyvärinen (2005). The paper Parry et al. (2012) discus the possible forms of a local scoring rule. The Hyvärinen divergence provides us with an example where a statistical problem requires a detailed discussion 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 $\nabla \log p(x) = \nabla \log f(x)$.

3.3.1. *Example: Gaussian case.* All assumptions are satisfied if $\mathcal{C}$ is the multivariate Gaussian model.

Let us discuss now the case when $\mathcal{C}$ is a subset of the maximal exponential model $\mathcal{E}(\mu)$, that is $q \in \mathcal{C}$ implies $q = \mathrm{e}^{u - K_p(u)} \cdot p$, $u \in B_p$. The Kyvärinen score is computable if p and u are differentiable as

$$S_{\mathrm{H}}(q) = \Delta(u - K_p(u)) + \Delta \log p + \frac{1}{2}\left|\nabla(u - K_p(u)) + \log p\right|^2 = \Delta u + \Delta \log p + \frac{1}{2}\left|\nabla u + \nabla \log p\right|^2 \ .$$

The risk at p (the same as in the chart) is

$$\begin{aligned} \mathbb{E}_p\left[S_{\mathrm{H}}(q)\right] &= \mathbb{E}_p\left[\Delta u\right] + \mathbb{E}_p\left[\Delta \log p\right] + \frac{1}{2}\mathbb{E}_p\left[\left|\nabla u + \nabla \log p\right|^2\right] = \\ &\frac{1}{2}\mathbb{E}_p\left[\left|\nabla u\right|^2\right] + \mathbb{E}_p\left[\nabla u \cdot \nabla \log p\right] + \frac{1}{2}\mathbb{E}_p\left[\left|\nabla \log p\right|^2\right] = \frac{1}{2}\mathbb{E}_p\left[\left|\nabla u\right|^2\right] + \frac{1}{2}\mathbb{E}_p\left[\left|\nabla \log p\right|^2\right] \ . \end{aligned}$$

In conclusion, in the exponential parametrization, the Hyvärinen divergence is associated with a special inner product on the exponential bundle,

$$\langle\langle u, v\rangle\rangle_p = \mathbb{E}_p\left[\nabla u \cdot \nabla v\right] \ , \quad u, v \in B_p \ .$$

3.3.2. *Variations on the theme: deformed exponentials.* 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 Montrucchio and Pistone (2017). A possible definition in this case is

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

Part 2. Gaussian probability space

In this second part, the idea is to review the construction of the exponential manifold using tools from the analysis of Gaussian spaces. A short recap of facts about the Gaussian space as it is defined in P. Malliavin textbook (Malliavin, 1995, Ch. V) is offered in section 4.1. Here, I consider only the finite-dimensional sample space. References for the infinite-dimensional case are monographs Malliavin (1997) and Nourdin and Peccati (2012). These notes are partly taken from Pistone (2017, 2018a).

Fisherian Information Geometry considers properties of statistical models that are invariant under measurable transformation of the sample space. This fact is well expressed by Chentsov characterization of the Fisher information. See Ch. 5 of Ay et al. (2017).

Any assumption on a specific structure of the sample space actually changes the picture. For example, on the real space is natural to consider statistical models induced by the action of a flow and the properties of such models could ardly be described with the tools of IG alone.

4. Information Geometry on the Gaussian space

When the sample space is $\mathbb{R}^n$, there is a particular class of statistical models of interest, namely, translation models. If f is a probability density of the Lebesgue measure, for each $h \in \mathbb{R}^n$, the curve $\theta \mapsto f(x - \theta h)$ defines a curve whose Fisher's score is

$$\frac{d}{d\theta} \log f(x - \theta h) = \frac{\nabla f(x - \theta h) \cdot h}{f(x - \theta h)}$$

and whose Fisher information is

$$\int \frac{(\nabla f(x - \theta h) \cdot h)^2}{f(x - \theta h)^2} f(x - \theta h) \; dx = \int \frac{(\nabla f(y) \cdot h)^2}{f(y)} \; dy \; .$$

It is interesting to note that the quantity above appears in Statistical Physics literature with the name *Fisher Information* but without any reference to the Fisher-Rao-Cernov theory. Let us see an example of its use, taken from § 5.5.2-3 of McKean (2014).

Example. In dimension 1, take a density q with mean value 0 and variance 1. Observe that

$$F(q) = \int \frac{(q' + xq)^2}{q} = \int \frac{(q')^2}{q} + 2 \int \frac{xqq'}{q} + \int \frac{x^2 q^2}{q} = \int \frac{(q')^2}{q} - 1$$

is zero only if $q' + xq = 0$, that is, $q(x) = (2\pi)^{-1/2} \mathrm{e}^{-x^2/2}$.

Let $q(t)$ be the density of $Y = \mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} Z$ with $X \sim q$ independent of $Z \sim \mathrm{N}(0, 1)$. One has $\frac{\partial}{\partial t} q(y; t) = \frac{\partial^2}{\partial y^2} q(y; t) + \frac{\partial}{\partial y}(y q(y; t))$. In fact,

$$\begin{gathered}
\int \phi(y) \frac{d}{dt} q(y; t) \; dy = \frac{d}{dt} \mathbb{E}\left(\phi\left(\mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} Z\right)\right) = \\
\mathbb{E}\left(\phi'\left(\mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} Z\right)\left(-\mathrm{e}^{-t} X + \frac{\mathrm{e}^{-2t}}{\sqrt{1 - \mathrm{e}^{-2t}}} Z\right)\right) = \\
\mathbb{E}\left(\phi'\left(\mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} Z\right)\left(-\mathrm{e}^{-t} X - \sqrt{1 - \mathrm{e}^{-2t}} Z + \frac{1}{\sqrt{1 - \mathrm{e}^{-2t}}} Z\right)\right) = \\
\int \phi'(y) y q(y; t) \; dy + \frac{1}{\sqrt{1 - \mathrm{e}^{-2t}}} \int \mathbb{E}\left(\phi'\left(\mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} z\right)\right) \; z \gamma(z) dz = \\
\int \phi'(y) y q; t(y) \; dy - \frac{1}{\sqrt{1 - \mathrm{e}^{-2t}}} \int \mathbb{E}\left(\phi'\left(\mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} z\right)\right) \; \gamma'(z) dz = \\
\int \phi'(y) y q(y; t) \; dy + \int \mathbb{E}\left(\phi''\left(\mathrm{e}^{-t} X + \sqrt{1 - \mathrm{e}^{-2t}} z\right)\right) \; \gamma(z) dz = \int (\phi'(y) y + \phi''(y)) q(y; t)
\end{gathered}$$

Let us compute the *rate of entropy production* of $t \mapsto q(t)$:

$$\frac{d}{dt}\left(-\int q(t)\log q(t)\right) = -\int(\log q(t)+1)\dot{q}(t) =$$
$$-\int(\log q(y;t)+1)\left(\frac{\partial^2}{\partial y^2}q(y;t)+\frac{\partial}{\partial y}(yq(y;t))\right)\,dy =$$
$$\int\frac{\left(\frac{\partial}{\partial y}q(y;t)\right)^2}{q(y;t)}+\int y\frac{\partial}{\partial y}q(y;t)\,dy+\int\frac{\left(\frac{\partial}{\partial y}q(y;t)\right)^2}{q(y;t)}-1 = F(q(t))\ .$$

This has to be compared with the generic equation:

$$\frac{d}{dt}H(q(t)) = \langle \operatorname{grad} H(q(t)), Sq(t)\rangle_{q(t)} = \langle -\log(q(t)) - H(q(t)), Sq(t)\rangle_{q(t)}\ .$$

Here we have used the specific evolution equation for $t \mapsto q(t)$.

McKean proceeds with the proof of a “logaritmic Sobolev inequality.”

4.1. **Analysis of the Gaussian space.** On the Gaussian space, computations like those illustrated in the previous section must be adapted. Let us show, in particular, what happens with the translations.

We denote by τ_h the translation operator $\tau_h u(x) = u(x-h)$. The following result shows the kind of checking we need in the Gaussian space.

Proposition 18 (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 with norm bounded by* $\mathrm{e}^{|h|^2}$.

Proof. We have

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

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. If $u \in L^2(\gamma)$, we want to consider the differentiability of $h \mapsto \tau_h u \in L^1(\gamma)$.

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 $\delta^\alpha = \prod_{j=1}^n \delta_j^{\alpha_j}$, $\alpha \in A = \mathbb{Z}^n_\geq$, 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 $\|u\|_\gamma^2 = \sum_{\alpha\in A} c_\alpha^2\dfrac{1}{\alpha!}$. 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 Malliavin (1997) 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 $\alpha_j \geq 1$, zero otherwise, and $\delta_j H_\alpha = H_{\alpha+e_j}$, we can define the operators on $L^2(\gamma)$

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

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

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

whose domains are, respectively,

$$\operatorname{dom}(\partial_j) = \left\{ \sum_{\alpha\in A} \frac{c^2_{\alpha+e_j}}{\alpha!} < \infty \right\} \ ,$$

$$\operatorname{dom}(\delta_j) = \left\{ \sum_{\alpha,\neq 0} \frac{\alpha_j^2 c^2_{\alpha-e_j}}{\alpha!} < \infty \right\} \ ,$$

$$\operatorname{dom}(\delta_j\partial_j) = \left\{ \sum_{\alpha\in A} \frac{\alpha_j^2 c^2_\alpha}{\alpha!} < \infty \right\} \ .$$

Proposition 19. *The operators ∂_j, δ_j, $\delta_j\partial_j$ are closed.*

Proposition 20. *If $u \in \operatorname{dom}(\partial_j)$ and $v \in \operatorname{dom}(\delta_j)$. then $\langle \partial_j u, v\rangle_\gamma = \langle u, \delta_j v\rangle_\gamma$.*

In particular, if $u \in \operatorname{dom}(\partial_j)$ and $\phi \in C_0^\infty(\mathbb{R}^n)$ (compact support) then $\phi, \delta_j\phi \in L^2(\gamma)$ and $\phi \in \operatorname{dom}(\delta_j)$ 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 21. *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 $u \in \operatorname{dom}(\partial_j)$ and $\partial_j = u'$.

If $u \in \operatorname{dom}(\partial_j)$ 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 $\operatorname{dom}(\nabla) = \cap_{j=1}^n \operatorname{dom}(\partial_j)$.

Proposition 22 (Poincaré inequality)**.** *If $u \in \operatorname{dom}(\nabla)$ then*

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

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

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

It follows

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

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

Proposition 23 (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 24 (The space $\mathbb{D}$)**.** We denote by $\mathbb{D}$ the domain of ∇ endowed with the Hilbert norm

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

Proposition 25. *If $u \in \operatorname{dom}(\nabla)$ and $h \in \mathbb{R}^n$, then $h \mapsto \tau_{-h}u$ is differentiable as a mapping in $L^1(\gamma)$ with derivative in $L^2(\gamma)$ given by*

$$\tau_{-h} - u = \int_0^1 \tau_{-sh}\nabla u \cdot h\ ds = \nabla u\cdot h + \int_0^1 (\tau_{-sh}\nabla u - u)\cdot h\ ds$$

Proof. Consider the measurable mapping

$$[0,1]\times\mathbb{R}^n \ni (t,x) \mapsto \nabla u(x+th)\cdot h$$

and observe that $\nabla u(x+th)\cdot h = \tau_{-th}\nabla u(x+th)\cdot h$. By proposition 18

$$\int_0^1 dt \int |\nabla u(x+th)\cdot h|\gamma(x)\ dx\ \ \leq \int_0^1 \mathrm{e}^{t^2|h|^2}\ dt < +\infty \ .$$

By Fubini theorem, the partial integration

$$v(x) = \int_0^1 \nabla u(x+th)\cdot h\ dt$$

is a γ-integrable random variable, in particular it is locally integrable. For each $\phi \in C_0^\infty(\mathbb{R}^n)$,

$$\int \phi(x)v(x)\ dx = \int \phi(x)\left(\int_0^1 \nabla u(x+th)\cdot h\ dt\right)\ dx =$$

$$\int_0^1\left(\int \phi(x)\sum_{j=1}^n h_j\partial_j u(x+sh)\ dx\right)\ dt = \int_0^1\left(\int \phi(y-sh)\sum_{j=1}^n h_j\partial_j u(y)\ dy\right)\ dt =$$

$$\int_0^1\left(\int -\nabla\phi(y-sh)\cdot h\ u(y)\ dy\right)\ dt = \int\left(\int_0^1 \frac{d}{ds}\phi(y-sh)\right)u(y)\ dy =$$

$$\int (\phi(y-h)-\phi(y))\,u(y)dy = \int \phi(x)\,(u(x+h)-u(x))\ dx \ ,$$

hence $v(x) = u(x+h) - u(x)$. □

4.2. **Gaussian Hyvärinen.** On the Gaussian space $(\mathbb{R}^n,\gamma)$, $\gamma(x) = (2\pi)^{-n/2}\mathrm{e}^{-|x|^2/2}$, consider two densities of the exponential manifold $p, q \in \mathcal{E}$. We have $\log p, \log q \in L^{(\cosh-1)}(\gamma) \in L^2(\gamma)$. If we assume moreover $\log p, \log q \in \mathbb{D}$, the quantity

$$\left|\nabla \log p - \nabla \log q\right|^2$$

is well defined in $L^1(\gamma)$, but this is not enough to ensure the finitness of

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

unless $p = 1$. Here, the issue is the dependence on a variable p, that appears both in the gradient and in the integration.

5. Gaussian Orlicz-Sobolev exponential manifold

This part of the lectures follows closely Pistone (2018a) and is more detailed that the previous parts. We give essentially full proofs or detailed references and offer the discussion of a number of critical examples, precisely, the same examples that have been informally discussed in section 3.

In all this section the sample space is the real Gaussian space of dimension n, $(\mathbb{R}^n, \mathcal{B}, \gamma)$. The standard Gaussian density is denote by γ.

5.1. **Orlicz spaces.** First, we review basic facts about Orlicz spaces, see, for example, Ch. II of Musielak (1983).

The couple of Young functions $(\cosh-1)$ and its conjugate $(\cosh-1)_*$ are associated with the Orlicz space $L^{(\cosh-1)}(\gamma)$ and $L^{(\cosh-1)_*}(\gamma)$, respectively. The choice of this specific couple is inessential. For example, as $\frac{1}{2}\mathrm{e}^x \leq \cosh x \leq \mathrm{e}^x$, the couple $\mathrm{e}^x - 1$ and $y\log y - y$ would give the same results.

The space $L^{(\cosh-1)}(\gamma)$ is called *exponential space* and is the vector space of all functions such that $\int (\cosh-1)(\alpha f(x))\gamma(x)\ dx\ < \infty$ for some $\alpha > 0$. This is the same as saying that the moment generating function $G_f(t) = \int \mathrm{e}^{tf(x)}\gamma(x)\ dx$ is finite on a open interval containing 0.

Computations. If $x, y \geq 0$, we have $(\cosh-1)'(x) = \sinh(x)$, $(\cosh-1)'_*(y) = \sinh^{-1}(y) = \log\left(y + \sqrt{1+y^2}\right)$, $(\cosh-1)_*(y) = \int_0^y \sinh^{-1}(t)\ dt$. The Fenchel-Young inequality is

$$xy \leq (\cosh-1)(x) + (\cosh-1)_*(y) = \int_0^x \sinh(s)\ ds + \int_0^y \sinh^{-1}(t)\ dt$$

and

$$\begin{aligned}(\cosh-1)(x) &= x\sinh(x) - (\cosh-1)_*(\sinh(x))\ ;\\ (\cosh-1)_*(y) &= y\sinh^{-1}(y) - (\cosh-1)(\sinh^{-1}(y))\\ &= y\log\left(y + (1+y^2)^{1/2}\right) - (1+y^2)^{-1/2}\ .\end{aligned}$$

The conjugate Young function $(\cosh-1)_*$ is associated with the *mixture space* $L^{(\cosh-1)_*}(\gamma)$. In this case, we have the inequality

$$(\cosh-1)_*(ay) \leq C(a)(\cosh-1)_*(y), \quad C(a) = \max(|a|, a^2)\ . \tag{7}$$

In fact

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

The inequality (7) follows easily by considering the two cases $a > 1$ and $a < 1$. As a consequence, $g \in L^{(\cosh-1)_*}(\gamma)$ if, and only if, $\int (\cosh-1)_*(g(y))\gamma(y)\ dy\ < \infty$.

In the theory of Orlicz spaces, the existence of a bound of the type (7) is called Δ_2-property, and it is quite relevant. In our case, it implies the following. See the proof in the reference given above.

Proposition 26. *The mixture space $L^{(\cosh -1)_*}(\gamma)$ is the dual space of its conjugate, the exponential space $L^{(\cosh -1)}(\gamma)$. Moreover, a separating sub-vector space e.g., $C_0^\infty(\mathbb{R}^n)$, is norm-dense.*

Other Young couple. In the definition of the associated spaces, the couple $(\cosh -1)$ and $(\cosh -1)_*$ is equivalent to the couple defined for $x, y > 0$ by $\Phi(x) = \mathrm{e}^x - 1 - x$ and $\Psi(y) = (1+y)\log(1+y) - y$. In fact, for $t > 0$ we have $\log(1+t) \le \log\left(y + \sqrt{1+t^2}\right)$ and

$$\log\left(t + \sqrt{1+t^2}\right) \le \log\left(t + \sqrt{1+2t+t^2}\right) = \log(1+2t) \ ,$$

so that we derive by integration the inequality

$$\Psi(y) \le (\cosh -1)_*(y) \le \frac{1}{2}\Psi(2y) \ .$$

In turn, conjugation gives

$$\frac{1}{2}\Phi(x) \le (\cosh -1)(x) \le \Phi(x) \ .$$

The *exponential space* $L^{(\cosh -1)}(\gamma)$ and the *mixture space* $L^{(\cosh -1)_*}(\gamma)$ are the spaces of real functions on $\mathbb{R}^n$ respectively defined using the conjugate Young functions $\cosh -1$ and $(\cosh -1)_*$. The exponential space and the mixture space are given norms by defining the closed unit balls of $L^{(\cosh -1)}(\gamma)$ and $L^{(\cosh -1)_*}(\gamma)$, respectively, by

$$\left\{f \,\middle|\, \int (\cosh -1)(f(x))\ M(x)dx \le 1\right\}, \quad \left\{g \,\middle|\, \int (\cosh -1)_*(g(x))\ M(x)dx \le 1\right\} \ .$$

Such a norm is called Luxemburg norm.

The Fenchel-Young inequality

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

implies that $(f, g) \mapsto \mathbb{E}_M[fg]$ is a separating duality, precisely

$$\left|\int f(x)g(x)\gamma(x)\ dx\ \right| \le 2\,\|f\|_{L^{(\cosh -1)}(\gamma)}\,\|g\|_{L^{(\cosh -1)_*}(\gamma)} \ .$$

A random variable g has norm $\|g\|_{L^{(\cosh -1)_*}(\gamma)}$ bounded by ρ if, and only if, $\|g/\rho\|_{L^{(\cosh -1)_*}(\gamma)} \le 1$, that is $\mathbb{E}_M[(\cosh -1)_*(g/\rho)] \le 1$, which in turn implies

$$\mathbb{E}_M[(\cosh -1)_*(\alpha g)] = \mathbb{E}_M[(\cosh -1)_*(\alpha\rho(g/\rho))] \le \rho\alpha$$

for all $\alpha \ge 0$. This is not true for the exponential space $L^{(\cosh -1)}(\gamma)$.

It is possible to define a dual norm, called Orlicz norm, on the exponential space, as follows. We have $\|f\|_{(L^{(\cosh -1)_*}(\gamma))^*} \le 1$ if, and only if $\left|\int f(x)g(x)\gamma(x)\ dx\ \right| \le 1$ for all g such that $\int (\cosh -1)_*(g(x))\gamma(x)\ dx\ \le 1$. With this norm, we have

$$\left|\int f(x)g(x)\gamma(x)\ dx\ \right| \le \|f\|_{(L^{(\cosh -1)_*}(\gamma))^*}\,\|g\|_{L^{(\cosh -1)_*}(\gamma)} \tag{8}$$

The Orlicz norm and the Luxemburg norm are equivalent, precisely,

$$\|f\|_{L^{(\cosh -1)}(\gamma)} \le \|f\|_{L^{(\cosh -1)_*}(\gamma)^*} \le 2\,\|f\|_{L^{(\cosh -1)}(\gamma)} \ .$$

We define the Banach space of centered random variables in the exponential space,

$$S_\gamma\mathcal{E} = \left\{u \in L^{(\cosh -1)}(\gamma) \,\middle|\, \int u(x)\gamma(x)\ dx\ = 0\right\}$$

and the moment function $Z_\gamma \colon S_\gamma\mathcal{E} \to \mathbb{R}_> \cup \{\infty\}$, $Z_\gamma(u) = \int \mathrm{e}^{u(x)}\gamma(x)\ dx$.

Proposition 27. *Z_γ is strictly convex ad its proper domain contains the open unit ball of $S_\gamma\mathcal{E}$. The interior $\mathcal{S}_1$ of the proper domain of the moment generating function is a convex open nonempty subset of $S_\gamma\mathcal{E}$.*

5.2. **Exponential arcs.** The use of the exponential space $L^{(\cosh -1)}(\gamma)$ is justified by the fact that for every 1-dimensional exponential family of the Gaussian space

$$J \ni \theta \mapsto p(\theta) \propto \mathrm{e}^{\theta v} \ , \quad J \text{ neighborhood of } 0 \ ,$$

the sufficient statistics v belongs to the exponential space, $v \in L^{(\cosh -1)}(\gamma)$.

Actually, we are going to see now the simple, but crucial, result that extends the construction to a manifold of positive densities containing γ. This topic has been incrementally developed in a series of paper, Cena and Pistone (2007); Santacroce, Siri, and Trivellato (2016b, 2018).

Definition 28. Two positive densities p and q of the Gaussian space, $p, q \in \mathcal{E}(\gamma)$, are said to connected by an open exponential arc, $p \smile q$, if there exists an interval J containing $[0,1]$ such that

$$\int p(x)^{1-\theta} q(x)^{\theta} \gamma(x) \, dx \ < +\infty \ , \quad \theta \in J \ . \tag{9}$$

This is the same as the existence of an exponential family $p(t) = \mathrm{e}^{tu - \psi(t)} \cdot p$ with $t \in J$, $I = \overset{\circ}{J}$ and $p(0) = p$, $p(1) = q$.

The following theorem is the key result for our construction.

Theorem 29. *Consider positive densities $p, q \in \mathcal{P}_+(\gamma)$. If the two densities are connected by an open exponential arc, $p \smile q$, then $L^{(\cosh -1)}(p \cdot \gamma) = L^{(\cosh -1)}(q \cdot \gamma)$.*

Proof. The relation $p \smile q$ is symmetric, then it is enough to show one inclusion. Let be given $w \in L^{(\cosh -1)}(p \cdot \mu)$ and let $p(t) \propto \mathrm{e}^{tu} p$, $t \in J$, be the open exponential curve that connects p and q. The function

$$\mathbb{R} \times J \ni (\alpha, t) \mapsto \int \frac{1}{2} (\exp(\alpha w + tu) + \exp(-\alpha w + tu)) \ \gamma d\mu \ , \tag{10}$$

is clearly convex. At $t = 0$ the value is proportional to $\mathbb{E}_p [(\cosh -1)(\alpha w)]$. Because of the assumption $w \in L^{(\cosh -1)}(p \cdot \gamma)$, that value is finite for all α in an open interval I around 0. If $\alpha = 0$, the value is equal to the normalizing constant of the model so that it is finite for all $t \in J$. It follows that the value of **??** is finite on the convex set generated by $I \times \{0\}$ and $\{0\} \times J$, in particular on the vertical section at $t = 1$, where it is equal to $\mathbb{E}_q [\cosh(\alpha w)]$. □

Remark 2. Note that the equality of the two Orlicz spaces implies the equivalence of the norms, that is the spaces are homeomorphic as Banach spaces. This is a general result for Orlicz spaces, see Lemma 1 of Cena and Pistone (2007). This is an essential ingredient of the full picture.

Let us show that the open exponential connection is an equivalent relation.

Proposition 30. *If $p, q \in \mathcal{P}_+(\gamma)$ are connected by an open exponential arc, $p \smile q$, then $\log \frac{p}{q}$ and $\log \frac{q}{p}$ belong to both $L^{(\cosh -1)}(p \cdot \gamma)$ and $L^{(\cosh -1)}(q \cdot \gamma)$. Moreover, $\cdot \smile \cdot$ is an equivalence relation.*

Proof. The integral in eq. (9) can be rewritten in exponential form in two ways,

$$\int \left(\frac{q}{p}\right)^{\theta} p \, \gamma = \int \exp\left(\theta \log \frac{q}{p}\right) p \, \gamma < +\infty \ , \quad \theta \in J \ ,$$

and

$$\int \left(\frac{p}{q}\right)^{\theta} q \, \gamma = \int \exp\left((1-\theta) \log \frac{p}{q}\right) q \, \gamma < +\infty \ , \quad \theta \in J \ .$$

This shows that the condition is necessary.

Conversely, consider that the exponential family $p(\theta) \propto \exp\left(\theta \log \frac{q}{p}\right) p$ is well defined in a neighborhood of 0 because of $\log \frac{q}{p} \in L^{(\cosh -1)}(p \cdot \gamma)$. Moreover,

$$\exp\left(\theta \log \frac{q}{p}\right) p = \left(\frac{q}{p}\right)^{\theta} p = \left(\frac{p}{q}\right)^{1-\theta} q = \exp\left((1-\theta) \log \frac{p}{q}\right) q$$

so that the exponential model is defined in a neighborhood of 1 because of $\log \frac{q}{p} \in L^{(\cosh -1)}(q \cdot \gamma)$.

The relation $\smile$ is reflexive and symmetric. Assume now $p \smile q$ and $q \smile r$. In follows from proposition 29 that $L^{(\cosh -1)}(p \cdot \gamma) = L^{(\cosh -1)}(q \cdot \gamma) = L^{(\cosh -1)}(r \cdot \gamma)$. We want to show that $\log \frac{p}{r} \in L^{(\cosh -1)}(r \cdot \gamma)$. But $\log \frac{p}{r} = \log \frac{p}{q} + \log \frac{q}{r}$ with $\log \frac{p}{q} \in L^{(\cosh -1)}(q \cdot \gamma)$ and $\log \frac{q}{r} \in L^{(\cosh -1)}(r \cdot \gamma)$. The equality of all spaces yelds the conclusion. The same when p and r are exchanged. □

Definition 31. The equivalence class for the relation $\smile$ that contains p is the maximal exponential model at p, denoted $\mathcal{E}(p)$.

We focus here on the equivalence class containing 1. It is the maximal exponential model $\mathcal{E} = \mathcal{E}(1)$ of the Gaussian space.

Let us write eq. (9) with $p = 1$ and $q(x) = \mathrm{e}^{v(x)}$,

$$\int \mathrm{e}^{\theta v(x)} \gamma(x)\, dx \ < +\infty \ , \quad \theta \in J \ . \tag{11}$$

It is clear that $q \in \mathcal{E}$ implies $v = \log q \in L^{(\cosh -1)}(\gamma)$. Conversely, if $q = \mathrm{e}^v$ and $v \in L^{(\cosh -1)}(\gamma)$ i.e.,

$$\int \cosh(\alpha v(x)) \gamma(x)\, dx \ \leq 2 \quad \text{for some } \alpha > 0 \ ,$$

then $\int \mathrm{e}^{\theta v} \gamma \leq 4$ for $\theta \in]-\alpha, 0]$. For $\theta \in [0,1]$ the convexity of the exponential implies

$$\int \mathrm{e}^{\theta v(x)} \gamma(x)\, dx \ \leq (1-\theta) + \theta \int \mathrm{e}^{v(x)} \gamma(x)\, dx \ = 1 \ .$$

The condition $\theta \in]-\alpha, 1]$ define an exponential arc which connects γ and q, but it is not open. In conclusion, the condition $q = \mathrm{e}^v$, $v \in L^{(\cosh -1)}(\gamma)$, does not imply $q\mathcal{E}$.

Here is a possible description of $\mathcal{E}$.

Proposition 32.
(1) *The cumulant mapping $K \colon L^{(\cosh -1)}(\gamma) \to [0, \infty]$ is convex and L.S.C. The proper domain $\{K < \infty\} = \operatorname{dom} K$ contains unit ball. The interior of the proper domain $(\operatorname{dom} K)^\circ$ is a convex set that contains the open unit ball.*
(2) *The positive density $q \in \mathcal{P}_+(\gamma)$ belongs to the class $\mathcal{E}$ if, and only if, has the exponential form $q = \exp(u - K(u))$ with $u \in (\operatorname{dom} K)^\circ$.*

Proof.
(1) The closed unit ball is $\left\{u \in L^{(\cosh -1)}(\gamma) \,\middle|\, \mathbb{E}(\cosh(u) - 1) \leq 2\right\}$. For such an u, $\mathbb{E}(\mathrm{e}^u) \leq 2\,\mathbb{E}(\cosh(u)) \leq 4 < +\infty$. From this, the second statement is follows. See theorem III.(2.5) in Barvinok (2002).
(2) Assume $u \in (\operatorname{dom} K)^\circ$. It follows that the set of θ's such that $\theta u \in (\operatorname{dom} K)^\circ$ is an open interval that contains $[0,1]$, so that $q = \mathrm{e}^{u - K(u)} \smile 1$. Conversely, assume $q \smile 1$, and consider the exponential curve $\theta \mapsto q(\theta) \propto \mathrm{e}^{\theta u}$, $\theta \in J$, J open super-interval of $[0,1]$. For each $\theta_1 \in J$ such that $\theta_1 > 1$, from the existence of the exponential curve we get $\theta_1 u \in \operatorname{dom} K$. As 0 is contained in the interior of $\operatorname{dom} K$, and u is a strict convex combination of 0 and θu, it follows $u \in (\operatorname{dom} K)^\circ$ (see lemma III.(2.4) of Barvinok (2002)).

□

5.3. **Entropy.** The statistical interest of the mixture space $L^{(\cosh -1)_*}(\gamma)$ resides in its relation with entropy.

If f is a positive density of the Gaussian space, $\int f(x)\gamma(x)\, dx \ = 1$, we define its entropy to be $\mathcal{H}(f) = -\int f(x) \log f(x) \gamma(x)\, dx$. As $x \log x \geq x - 1$ and it is strictly convex, the integral is well defined and $\mathcal{H}(f) > 0$ unless $f = 1$. It holds

$$-\int f(x) \log^+ f(x) \gamma(x)\, dx \ \leq \mathcal{H}(f) \leq \mathrm{e}^{-1} - \int f(x) \log^+ f(x) \gamma(x)\, dx \ \ , \tag{12}$$

where $\log^+$ is the positive part of log.

Proposition 33. *A positive density f of the Gaussian space has finite entropy if, and only if, f belongs to the mixture space $L^{(\cosh -1)_*}(\gamma)$.*

Proof. We use Eq. (12) in order to show the equivalence. For $x \geq 1$ it holds

$$2x \leq x + \sqrt{1+x^2} \leq (1+\sqrt{2})x \ .$$

It follows

$$\log 2 + \log x \leq \log\left(x + \sqrt{1+x^2}\right) = \sinh^{-1}(x) \leq \log\left(1+\sqrt{2}\right) + \log x \ ,$$

and, taking the integral $\int_1^y$ with $y \geq 1$, we get

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

then, substituting $y > 1$ with $\max(1, f(x))$, $f(x) > 0$,

$$(\log 2 - 1)(f(x)-1)^+ + f(x)\log^+ f(x) \leq \\ (\cosh -1)_*(\max(1,f(x))) - (\cosh -1)_*(1) \leq \\ (\log\left(1+\sqrt{2}\right) - 1)(f(x)-1)^+ + f(x)\log^+ f(x) \ .$$

By taking the Gaussian integral, we have

$$(\log 2 - 1)\int (f(x)-1)^+\gamma(x)\ dx\ + \int f(x)\log^+ f(x)\gamma(x)\ dx\ \leq \\ \int (\cosh -1)_*(\max(1,f(x)))\gamma(x)\ dx\ - (\cosh -1)_*(1) \leq \\ (\log\left(1+\sqrt{2}\right) - 1)\int (f(x)-1)^+\gamma(x)\ dx\ + \int f(x)\log^+ f(x)\gamma(x)\ dx \ \ ,$$

which in turn implies the statement because $f \in L^1(M)$ and

$$\int (\cosh -1)_*(f(x))\gamma(x)\ dx\ + (\cosh -1)_*(1) = \\ \int (\cosh -1)_*(\max(1,f(x)))\gamma(x)\ dx\ + \int (\cosh -1)_*(\min(1,f(x)))\gamma(x)\ dx \ \ .$$

□

Of course, this proof does not depend on the Gaussian assumption.

5.4. **Orlicz and Lebesgue spaces.** We discuss now the relations between the exponential space, the mixture space, and the Lebesgue spaces. This provides a first list of classes of functions that belong to the exponential space or to the mixture space. The first item in the proposition holds for a general base probability measure, while the other is proved in the Gaussian case.

Proposition 34. *Let $1 < a < \infty$.*

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

(2) *If $\Omega_R = \{x \in \mathbb{R}^n \mid |x| < R\}$, the restriction operator is defined and continuous in the cases*
$$L^{(\cosh -1)}(\gamma) \to L^a(\Omega_R), \quad L^{(\cosh -1)_*}(\gamma) \to L^1(\Omega_R)$$

Proof. (1) See (Musielak, 1983, Ch. II).

(2) For all integers $n \geq 1$,

$$1 \geq \int (\cosh -1) \left(\frac{f(x)}{\|f\|_{L^{(\cosh -1)}(\gamma)}} \right) M(x)\ dx \geq$$

$$\int_{\Omega_R} \frac{1}{(2n)!} \left(\frac{f(x)}{\|f\|_{L^{(\cosh -1)}(\gamma)}} \right)^{2n} M(x)\ dx \geq$$

$$\frac{(2\pi)^{-n/2} \mathrm{e}^{-R^2/2}}{(2n)!\, \|f\|_{L^{(\cosh -1)}(\gamma)}} \int_{\Omega_R} (f(x))^{2n}\ dx.$$

□

5.5. **Maximal exponential model on the Gaussian space.** Here, we read from Ch. V of Malliavin (1995), Pistone (2013b), Santacroce, Siri, and Trivellato (2016b).

Let us repeat again our construction. 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 $p \in \mathcal{P}_{\geq}$ 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 Orlicz space*, see Musielak (1983). 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 *statistical bundle*

$$S\mathcal{E}(\gamma) = \left\{ (p, v) \,|\, p \in \mathcal{E}(\gamma), v \in B_p \right\}$$

is the natural non-parametric set-up for Information geometry in the sense of Amari (1982, 1985); Pistone and Sempi (1995).

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 $\operatorname{dom}(K_p)$ 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

$$e_p(u) = \mathrm{e}^{u - K_p(u)} \cdot p \in \mathcal{E}(\gamma) .$$

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 Cena and Pistone (2007).

The global structure as p varies is clarified by the following "Portmanteau theorem," cf. (Santacroce et al., 2016b, 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 Pistone (2013b).

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 $(\cosh -1)$,

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

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.

5.5.1. *Gaussian Hyvärinen divergence: special case.* The Gaussian version of the Hyvärinen divergence can be discussed with assumptions the type $\log p \in \mathbb{D}$ 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.

The GH divergence between the densities $p = \mathrm{e}^{u - K(u)}$ and $q = \mathrm{e}^{v - K(v)}$, $q \in \mathcal{E}(\gamma)$, is

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

which is well defined if $u, v \in L^{(\cosh -1)}(\gamma) \cap \mathbb{D}$.

Let us compute the risk. We have

$$\frac{1}{2} \mathbb{E}_p \left[|\nabla u - \nabla v|^2 \right] = \frac{1}{2} \mathbb{E}_p \left[|\nabla u|^2 \right] + \frac{1}{2} \mathbb{E}_p \left[|\nabla v|^2 \right] + \mathbb{E}_p \left[\nabla u \cdot \nabla v \right] = \\ \frac{1}{2} \mathbb{E}_p \left[|\nabla u|^2 \right] + \frac{1}{2} \mathbb{E}_p \left[|\nabla v|^2 \right] + \mathbb{E}_p \left[\nabla u \cdot \nabla v \right] =$$

5.6. **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 Pistone (2018b).

Definition 35. The exponential and the mixture Orlicz-Sobolev-Gauss (OSG) spaces are, respectively,

$$W^{1,(\cosh -1)}(M) = \left\{ f \in L^{(\cosh -1)}(M) \,\middle|\, \partial_j f \in L^{(\cosh -1)}(M) \right\} \ , \tag{13}$$

$$W^{1,(\cosh -1)_*}(M) = \left\{ f \in L^{(\cosh -1)_*}(M) \,\middle|\, \partial_j f \in L^{(\cosh -1)_*}(M) \right\} \ , \tag{14}$$

where ∂_j, $j = 1, \dots, n$, 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,

$$\|f\|_{W^1_{(\cosh -1)}(M)} = \|f\|_{L^{(\cosh -1)}(M)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(M)} \tag{15}$$

and

$$\|f\|_{W^1_{(\cosh -1)_*}(M)} = \|f\|_{L^{(\cosh -1)}(M)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(M)} \ . \tag{16}$$

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 36. *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 Brezis (2011). □

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

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

6. Formal results: Gaussian space and derivation

All along this paper, the sample space is the real Borel space $(\mathbb{R}^n, \mathcal{B})$ and γ denotes the standard n-dimensional Gaussian density, $\gamma(z) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \left|z\right|^2\right)$, $z \in \mathbb{R}^n$. The probability space $(\mathbb{R}^n, \mathcal{B}, \gamma)$ is the *Gaussian space.*

We list below some useful computations about the Gaussian standard density γ.

1. $\gamma(z+h)/\gamma(z) = \exp\left(-\frac{1}{2} \left|z+h\right|^2 + \frac{1}{2} \left|z\right|^2\right) = \exp\left(- \langle h, z \rangle - \frac{1}{2} \left|h\right|^2\right)$, $z, h \in \mathbb{R}^n$.
2. $\int \mathrm{e}^{\langle \theta, z \rangle} \gamma(z) \ dz \ = \mathrm{e}^{|\theta|/2}$.
3. $\int (\gamma(z+h)/\gamma(z))^2 \gamma(z) \ dz \ \leq \mathrm{e}^{|h|^2}$.

7. Notable bounds and examples

There is a large body of literature about the analysis of the Gaussian space $L^2(M)$. In order to motivate our own construction and to connect it up, in this section we have collected some results about notable classes of functions that belongs to the exponential space $L^{(\cosh -1)}(\gamma)$ or to the mixture space $L^{(\cosh -1)_*}(\gamma)$. Some of the examples will be used in the applications of Orlicz-Sobolev spaces in the Information Geometry of the Gaussian space. Basic references on the analysis of the Gaussian space are (Malliavin, 1995, V.1.5), (Stroock, 2008, 4.2.1), and (Nourdin and Peccati, 2012, Ch. 1).

7.1. Polynomial bounds.

The exponential space $L^{(\cosh -1)}(\gamma)$ contains all functions $f \in C^2(\mathbb{R}^n; \mathbb{R})$ whose Hessian is uniformly dominated by a constant symmetric matrix. In such a case, $f(x) = f(0) + \nabla f(0)x + \frac{1}{2} x^* \operatorname{Hess} f(\bar{x}) x$, with $x^* \operatorname{Hess} f(y) x \leq \lambda \left|x\right|^2$, $y \in \mathbb{R}^n$, and $\lambda \geq 0$ being the largest non-negative eigen-value of the dominating matrix. Then for all real α,

$$\int_{\mathbb{R}^n} \mathrm{e}^{\alpha f(x)} M(x) \ dx < \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\alpha f(0) + \nabla f(0) x + \frac{1}{2}(\alpha\lambda - 1)|x|^2} \ dx$$

and the RHS is finite for $\alpha < \lambda^{-1}$. In particular, $L^{(\cosh -1)}(\gamma)$ contains all polynomials with degree up to 2.

An interesting simple application of the same argument is the following. Assume $p = \mathrm{e}^v$ is a positive density on the Gaussian space such that

$$\mathrm{e}^{A_1(x)} \leq \mathrm{e}^{v(x)} \leq \mathrm{e}^{A_2(x)}, \quad x \in \mathbb{R}^n \ ,$$

for suitable second order polynomials A_1, A_2. Then $v \in L^{(\cosh -1)}(\gamma)$. Inequalities of this type appear in the theory of parabolic equations e.g., see (Brezis, 2011, Ch. 4).

The mixture space $L^{(\cosh -1)_*}(\gamma)$ contains all random variables $f \colon \mathbb{R}^d \to \mathbb{R}$ which are bounded by a polynomial, in particular, all polynomials. In fact, all polynomials belong to $L^2(M) \subset L^{(\cosh -1)_*}(\gamma)$.

7.2. **Densities of exponential form.** In this paper, we are specially interested in densities of the Gaussian space of the form $f = \mathrm{e}^v$, that is $\int \mathrm{e}^{v(x)}\gamma(x)\ dx\ = 1$. Let us now consider simple properties of the mappings $f \mapsto v = \log f$ and $v \mapsto f = \mathrm{e}^v$.

We have seen in Prop. 33 that $f = \mathrm{e}^v \in L^{(\cosh -1)_*}(\gamma)$ if, and only if,

$$- \mathcal{H}(\mathrm{e}^v) = \int \mathrm{e}^{v(x)} v(x)\gamma(x)\ dx\ < \infty\ .$$

As $\lim_{x\to+\infty} \frac{\cosh(x)}{x\mathrm{e}^x} = 0$, we do not expect $v \in L^{(\cosh -1)}(\gamma)$ to imply $f = \mathrm{e}^v \in L^{(\cosh -1)_*}(\gamma)$.

As $(\cosh -1)(\alpha \log y) = (y^\alpha + y^{-\alpha})/2 - 1$, $\alpha > 0$, then $v = \log f \in L^{(\cosh -1)}(\gamma)$ if, and only if, both f^α and $f^{-\alpha}$ both belong to $L^1(M)$ for some $\alpha > 0$. In the case $\|v\|_{L^{(\cosh -1)}(\gamma)} < 1$, then we can take $\alpha > 1$ and $f \in L^\alpha(M) \subset L^{(\cosh -1)_*}(\gamma)$. In conclusion, $\exp\colon v \mapsto \mathrm{e}^v$ maps the open unit ball of $L^{(\cosh -1)}(\gamma)$ into $\cup_{\alpha>1} L^\alpha(M) \subset L^{(\cosh -1)_*}(\gamma)$.

This issue is discussed in the next Sec. 8.

7.3. **Poincaré-type inequalities.** Let us denote by $C_b^k(\mathbb{R}^n)$ the space of functions with derivatives up to order k, each bounded by a constant. We write $C_p^k(\mathbb{R}^n)$ if all the derivative are bounded by a polynomial. We discuss below inequalities related to the classical Gaussian Poincaré inequality, which reads, in the 1-dimensional case,

$$\int \left(f(x) - \int f(y)\gamma(y)\ dy\right)^2 \gamma(x)\ dx\ \leq \int \left|f'(x)\right|^2 \gamma(x)\ dx\ \ , \tag{17}$$

for all $f \in C_p^1(\mathbb{R}^n)$. We are going to use the same techniques used in the classical proof of (17) e.g., see Nourdin and Peccati (2012).

If u, Y are independent standard Gaussian variables, then

$$u' = \mathrm{e}^{-t} + \sqrt{1-\mathrm{e}^{-2t}}Y, \quad Y' = \sqrt{1-\mathrm{e}^{-2t}}u - \mathrm{e}^{-t}Y$$

are independent standard Gaussian random variables for all $t \geq 0$. Because of that, it is useful to define Ornstein-Uhlenbeck semi-group by the Mehler formula

$$P_t f(x) = \int f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y)\gamma(y)\ dy\ , \quad t \geq 0, \quad f \in C_p(\mathbb{R}^n)\ . \tag{18}$$

For any convex function Φ, Jensen's inequality gives

$$\int \Phi(P_t f(x))\gamma(x)\ dx\ \leq \\ \int\int \Phi(f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y))\gamma(y)\ dy\ \gamma(x)\ dx\ = \\ \int \Phi(f(x))\gamma(x)\ dx\ \ .$$

In particular, this shows that, for all $t \geq 0$, $f \mapsto P_t f$ is a contraction for the norm of both the mixture space $L^{(\cosh -1)_*}(\gamma)$ and the exponential space $L^{(\cosh -1)}(\gamma)$.

Moreover, if $f \in C_p^1(\mathbb{R}^n)$, we have

$$\begin{aligned}
& f(x) - \int f(y)\gamma(y)\ dy \\
& \quad = P_0(x) - P_\infty f(x) \\
& \quad = - \int_0^\infty \frac{d}{dt} P_t f(x)\ dt \\
& \quad = \int_0^\infty \int \nabla f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y) \cdot \left(\mathrm{e}^{-t}x - \frac{\mathrm{e}^{-2t}}{\sqrt{1-\mathrm{e}^{-2t}}}y \right) \gamma(y)\ dy\ \ dt \qquad (19) \\
& \quad \leq \int_0^\infty \frac{\mathrm{e}^{-t}}{\sqrt{1-\mathrm{e}^{-2t}}}\ dt \quad \times \\
& \qquad \int \left| \nabla f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y) \right| \left| \sqrt{1-\mathrm{e}^{-2t}}x - \mathrm{e}^{-t}y \right| \gamma(y)\ dy \ \ . \qquad (20)
\end{aligned}$$

Note that

$$\int_0^\infty \frac{\mathrm{e}^{-t}}{\sqrt{1-\mathrm{e}^{-2t}}}\ dt = \int_0^1 \frac{ds}{\sqrt{1-s^2}} = \frac{\pi}{2}\ .$$

We use this remark and (20) to prove our first inequality.

Proposition 37. *If $f \in C_p^1(\mathbb{R}^n)$ and $\lambda > 0$ is such that*

$$C\left(\lambda \frac{\pi}{2}\right) \int C(|y|)\gamma(y)\ dy\ = 1\ , \quad C(a) = \max(|a|\,, a^2)\ , \qquad (21)$$

then

$$\int (\cosh -1)_* \left(\lambda \left(f(x) - \int f(y)M(y)\ dy \right) \right) M(x)\ dx \leq \int (\cosh -1)_*(|\nabla f(x)|) M(x)\ dx\ ,$$

that is

$$\left\| f - \int f(y)M(y)\ dy \right\|_{L^{(\cosh -1)_*}(\gamma)} \leq \lambda^{-1} \left\| |\nabla f| \right\|_{L^{(\cosh -1)_*}(\gamma)}\ .$$

Proof. Jensen's inequality applied to Eq. (20) gives

$$\begin{aligned}
(\cosh -1)_* \left(\lambda \left(f(x) - \int f(y)\gamma(y)\ dy \right) \right) \leq \int_0^\infty \frac{2}{\pi} \frac{\mathrm{e}^{-t}}{\sqrt{1-\mathrm{e}^{-2t}}}\ dt\ \times \qquad (22) \\
\int (\cosh -1)_* \left(\lambda \frac{\pi}{2} \left| \nabla f(\sqrt{1-\mathrm{e}^{-2t}}x + \mathrm{e}^{-t}y) \right| \left| \sqrt{1-\mathrm{e}^{-2t}}x - \mathrm{e}^{-t}y \right| \right) \gamma(y)\ dy
\end{aligned}$$

Now we use of the bound in Eq. (7), namely $(\cosh -1)_*(ay) \leq C(a)(\cosh -1)_*(y)$ if $a > 0$, where $C(a) = \max(|a|\,, a^2)$, and further bound for $a, k > 0$

$$C(ka) = ka \vee k^2a^2 \leq kC(a) \vee k^2C(a) = C(k)C(a)\ ,$$

to get

$$\begin{aligned}
(\cosh -1)_* \left(\lambda \frac{\pi}{2} \left| \nabla f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y) \right| \left| \sqrt{1-\mathrm{e}^{-2t}}x - \mathrm{e}^{-t}y \right| \right) \leq \qquad (23) \\
C\left(\lambda \frac{\pi}{2}\right) C\left(\left| \sqrt{1-\mathrm{e}^{-2t}}x - \mathrm{e}^{-t}y \right| \right) (\cosh -1)_* \left(\left| \nabla f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y) \right| \right)\ .
\end{aligned}$$

Taking the expected value of both sides of the inequality resulting from (22) and (23), we get

$$\int (\cosh -1)_* \left(\lambda \left(f(y) - \int f(x)M(x)\ dx \right) \right) M(y)\ dy \leq C\left(\lambda \frac{\pi}{2}\right) \int C(|y|)\gamma(y)\ dy\ \int (\cosh -1)_*(|\nabla f(x)|)\gamma(x)\ dx\ \ ,$$

We conclude by choosing a proper value of λ. □

The same argument does not work in the exponential space. We have assume the boundedness of derivatives i.e., a Lipschitz assumption.

Proposition 38. *If* $f \in C_b^1(\mathbb{R}^n)$ *with* $\sup\{|\nabla f(x)| \,|\, x \in \mathbb{R}^n\} = m$ *then*

$$\left\| f - \int f(y)\gamma(y)\ dy \ \right\|_{L^{(\cosh-1)}(\gamma)} \leq \frac{\pi}{2\sqrt{2\log 2}} m\ .$$

Proof. Jensen's inequality applied to Eq. (20) and the assumption give

$$(\cosh -1)\left(\lambda\left(f(x) - \int f(y)\gamma(y)\ dy\ \right)\right) \leq \\ \int (\cosh -1)\left(\lambda \frac{\pi}{2} m x\right)\gamma(x)\ dx\ = \exp\left(\frac{\lambda^2}{2}\frac{\pi^2}{4}m^2\right) - 1\ .$$

To conclude, choose λ such that the the RHS equals 1. □

Remark 3. Both Prop. 37 and Prop. 38 are related with interesting results on the Gaussian space other then bounds on norms. For example, if f is a density of the Gaussian space, then the first one is a bound on the lack of uniformity $f-1$, which, in turn, is related with the entropy of f. As a further example, consider a case where $\int f(x)\gamma(x)\ dx\ = 0$ and $\|\nabla f\|_\infty < \infty$. In such a case, we have a bound on the Laplace transform of f, which in turn implies a bound on large deviations of the random variable f.

To prepare the proof of an inequality for the exponential space, we start from Eq. (19) and observe that for $f \in C_p^2(\mathbb{R}^n)$ we can write

$$f(x) - \int f(y)\gamma(y)\ dy\ = \\ \int_0^\infty \mathrm{e}^{-t}\left(\int \nabla f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y)\gamma(y)\ dy\ \right)\cdot x\ dt \\ - \int_0^\infty \mathrm{e}^{-2t}\int \nabla\cdot\nabla f(\mathrm{e}^{-t}x + \sqrt{1-\mathrm{e}^{-2t}}y)\gamma(y)\ dy\ \ dt\ ,$$

where integration by parts and $(\partial/\partial y_i)M(y) = -y_i M(y)$ have been used to get the last term.

If we write $f_i(z) = \frac{\partial}{\partial z_i}$ and $f_{ii}(z) = \frac{\partial^2}{\partial {z_i}^2} f(z)$ then

$$\frac{\partial}{\partial x_i} P_t f(x) = \mathrm{e}^{-t} P_t f_i(x)$$

and

$$\frac{\partial^2}{\partial {x_i}^2} P_t f(x) = \mathrm{e}^{-2t} P_t f_{ii}(x)\ ,$$

so that

$$f(x) - \int f(y)\gamma(y)\ dy\ = \int_0^\infty \left(x\cdot\nabla P_t f(x) - \nabla\cdot\nabla P_t f(x)\right)\ dt\ .$$

If $g \in C_b^2(\mathbb{R}^n)$ we have

$$\begin{aligned}(24)\quad & \int g(x) \left(f(x) - \int f(y)\gamma(y)\ dy \right)\gamma(x)\ dx\ = \\ & \int_0^\infty \left(\int g(x) x \cdot \nabla P_t f(x) \gamma(x)\ dx\ - \int g(x) \nabla \cdot \nabla P_t f(x) \gamma(x)\ dx \right)\ dt = \\ & \int_0^\infty \left(\int g(x) x \cdot \nabla P_t f(x) \gamma(x)\ dx\ + \int \nabla (g(x) M(x)) \cdot \nabla P_t f(x)\ dx \right)\ dt = \\ & \int_0^\infty \int \nabla g(x) \cdot \nabla P_t f(x) \gamma(x)\ dx\ \ dt = \\ & \int_0^\infty \mathrm{e}^{-t} \int \nabla g(x) \cdot P_t \nabla f(x) \gamma(x)\ dx\ \ dt\ .\end{aligned}$$

Let $|\cdot|_1$ and $|\cdot|_2$ be two norms on $\mathbb{R}^n$ such that $|x \cdot y| \leq |x|_1 |y|_2$. Define the covariance of $f, g \in C_p^2(\mathbb{R}^n)$ to be

$$\begin{aligned}\operatorname{Cov}_M(f,g) = & \\ & \int \left(f(x) - \int f(y)\gamma(y)\ dy \right) g(x)\gamma(x)\ dx\ = \\ & \int \left(f(x) - \int f(y)\gamma(y)\ dy \right) \left(g(x) - \int g(y)\gamma(y)\ dy \right)\gamma(x)\ dx\ \ .\end{aligned}$$

Proposition 39. *If $f, g \in C_p^2(\mathbb{R}^n)$, then*

$$|\operatorname{Cov}_M(f,g)| \leq \left| \|\nabla f\|_{L^{(\cosh -1)_*}(\gamma)} \right|_1 \left| \|\nabla g\|_{(L^{(\cosh -1)_*}(\gamma))^*} \right|_2\ .$$

Proof. We use Eq. (24) and the inequality (8).

$$\begin{aligned}& \left| \int \nabla g(x) \cdot P_t \nabla f(x) \gamma(x)\ dx\ \right| \leq \\ & \sum_{i=1}^n \left| \int g_i(x) P_t f_i(x) \gamma(x)\ dx\ \right| \leq \\ & \sum_{i=1}^n \|g_i\|_{L^{(\cosh -1)_*}(\gamma)^*} \|P_t f_i\|_{L^{(\cosh -1)_*}(\gamma)} \leq \\ & \sum_{i=1}^n \|g_i\|_{L^{(\cosh -1)_*}(\gamma)^*} \|f_i\|_{L^{(\cosh -1)_*}(\gamma)} \leq \\ & \left| \|\nabla g\|_{L^{(\cosh -1)_*}(\gamma)^*} \right|_1 \left| \|\nabla f\|_{L^{(\cosh -1)_*}(\gamma)} \right|_2\ .\end{aligned}$$

□

If g_n is a sequence such that $\nabla g_n \to 0$ in $L^{(\cosh -1)}(\gamma)$, then the inequality above shows that $g_n - \int g_n(x)\gamma(x)\ dx\ \to 0$.

8

8.1. **Maximal exponential manifold as an affine manifold.** The maximal exponential model $\mathcal{E}(M) = \left\{ \mathrm{e}^{U - K_M(U)} \,\middle|\, U \in B_M \right\}$ is an elementary manifold embedded into $L^{(\cosh -1)_*}(M)$ by the smooth mapping $e_M \colon \mathcal{S}_M \to L^{(\cosh -1)_*}(M)$. There is actually an atlas of charts that makes it into an affine manifold, see Pistone (2013a). We discuss here some preliminary results about this important topic.

An elementary computation shows that

$$(\cosh -1)^2(u) = \frac{1}{2}(\cosh -1)(2u) - 2(\cosh -1)(u) \leq \frac{1}{2} .(\cosh -1)(2u)$$

Iterating,

$$(\cosh -1)^{2k}(u) \leq \left(\frac{1}{2}\right)^{a(k)} (\cosh -1)(2^k u) \ , \tag{25}$$

with $a(1) = 1$ and $a(k+1) = 2a(k) + 1$ i.e., $a(k) = 2^{k+1} - 1$.

If $\Phi = \cosh -1$ and $2^k = b$, the inequality becomes

$$\Phi(u)^b \leq \frac{1}{2^{2b-1}} \Phi(bu) \ .$$

Proposition 40. *If $f, g \in \mathcal{E}(M)$, then $L^{(\cosh -1)}(f \cdot M) = L^{(\cosh -1)}(g \cdot M)$.*

Proof. Given any $f \in \mathcal{E}(M)$, with $f = \mathrm{e}^{U - K_M(U)}$ and $U \in \mathcal{S}_M$, and any $V \in L^{(\cosh -1)}(M)$, we have from Fenchel-Young inequality and Eq. (25) that

$$\int (\cosh -1)(\alpha V(x)) f(x) \gamma(x) \ dx \ \leq \\ \frac{1}{2^{k+1}k} \int (\cosh -1)(2k\alpha V(x)) \gamma(x) \ dx \ + \\ \frac{2k-1}{2k} Z_M(U)^{\frac{2k}{2k-1}} \int \exp\left(\frac{2k}{2k-1} U\right) \gamma(x) \ dx \ \ .$$

If k is such that $\frac{2k}{2k-1} U \in \mathcal{S}_M$, one sees that $V \in L^{(\cosh -1)}(f \cdot M)$. We have proved that $L^{(\cosh -1)}(M) \subset L^{(\cosh -1)}(f \cdot M)$.

Conversely,

$$\int (\cosh -1)(\alpha V(x)) \gamma(x) \ dx \ = \int (\cosh -1)(\alpha V(x)) f^{-1}(x) f(x) \gamma(x) \ dx \ \leq \\ \frac{1}{2^{k+1}k} \int (\cosh -1)(2k\alpha V(x)) f(x) \gamma(x) \ dx \ + \\ \frac{2k-1}{2k} Z_M(U)^{\frac{1}{2k-1}} \int \exp\left(\frac{1}{2k-1} U\right) \gamma(x) \ dx \ \ .$$

If $\frac{1}{2k-1} U \in \mathcal{S}_M$, one sees that $V \in L^{(\cosh -1)}(f \cdot M)$ implies $V \in L^{(\cosh -1)}(M)$. □ □

The affine manifold is defined as follows. For each $f \in \mathcal{E}(M)$, we define the Banach space

$$B_f = \left\{ U \in L^{(\cosh -1)}(f \cdot M) \middle| \mathbb{E}_{f \cdot M}[U] = 0 \right\} = \left\{ U \in L^{(\cosh -1)}(M) \middle| \mathbb{E}_M[Uf] = 0 \right\} \ ,$$

and the chart

$$s_f \colon \mathcal{E}(M) \ni g \mapsto \log \frac{g}{f} - \mathbb{E}_{f \cdot M}\left[\log \frac{g}{f}\right] \ .$$

It is easy to verify the following statement, which defines the *exponential affine manifold.* Specific properties related with the Gaussian space are discussed in the next Sec. 8.2 and space derivatives in Sec. **??**.

Proposition 41. *The set of charts $s_f \colon \mathcal{E}(M) \to B_f$ is an affine atlas of global charts on $\mathcal{E}(M)$.*

On each fiber $S_p\mathcal{E}(M) = B_p$ of the statistical bundle the covariance $(U, V) \mapsto \mathbb{E}_M[UV] = \langle U, V \rangle_p$ provides a natural metric. In that metric the natural gradient of a smooth function $F \colon \mathcal{E}(M) \to \mathbb{R}$ is defined by

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

where $t \mapsto p(t)$ is a smooth curve in $\mathcal{E}(M)$ and $Dp(t) = \frac{d}{dt} \log p(t)$ is the expression of the velocity.

8.2. **Translations and mollifiers.** In this section, we begin to discuss properties of the exponential affine manifold of Prop. 41 which depend on the choice of the Gaussian space as base probability space.

Because of the lack of norm density of the space of infinitely differentiable functions with compact support $C_0^\infty\left(\mathbb{R}^n\right)$ in the exponential space $L^{(\cosh-1)}\left(M\right)$, we introduce the following classical the definition of Orlicz class.

Definition 42. The *exponential class*, $C_0^{(\cosh-1)}\left(\gamma\right)$, is the closure of $C_0\left(\mathbb{R}^n\right)$ in the exponential space $L^{(\cosh-1)}\left(\gamma\right)$. The space $C_0^\infty\left(\mathbb{R}^n\right)$ is dense in $C_0^{(\cosh-1)}\left(\gamma\right)$.

A characteristic property of the exponential class is the convergence of restrictions to a bounded domain.

Proposition 43. *Assume $f \in L^{(\cosh-1)}\left(\gamma\right)$ and write $f_R(x) = f(x)(|x| > R)$. The following conditions are equivalent:*

(1) *The real function $\rho \mapsto \int (\cosh-1)(\rho f(x))\gamma(x)\ dx$ is finite for all $\rho > 0$;*
(2) *$f \in C_0^{(\cosh-1)}\left(\gamma\right)$;*
(3) $\lim_{R\to\infty} \left\|f_R\right\|_{L^{(\cosh-1)}(\gamma)} = 0$.

Proof. This is well known e.g., see (Musielak, 1983, Ch. II). A short proof is given in our note (Pistone, 2017, Prop. 3). □ □

Here we study of the action of translation operator on the exponential space $L^{(\cosh-1)}\left(\gamma\right)$ and on the exponential class $C_0^{(\cosh-1)}\left(M\right)$. We consider both translation by a vector, $\tau_h f(x) = f(x-h)$, $h \in \mathbb{R}^n$, and translation by a probability measure, or convolution, μ, $\tau_\mu f(x) = \int f(x-y)\ \mu(dy) = f * \mu(x)$. A small part of this material was published in the conference paper (Pistone, 2017, Prop. 4–5).

Proposition 44 (Translation by a vector)**.**

(1) *For each $h \in \mathbb{R}^n$, the translation mapping $L^{(\cosh-1)}\left(\gamma\right) \ni f \mapsto \tau_h f$ is linear and bounded from $L^{(\cosh-1)}\left(\gamma\right)$ to itself. In particular,*

$$\left\|\tau_h f\right\|_{L^{(\cosh-1)}(\gamma)} \leq 2\left\|f\right\|_{L^{(\cosh-1)}(\gamma)} \quad \text{if} \quad |h| \leq \sqrt{\log 2}\ .$$

(2) *For all $g \in L^{(\cosh-1)_*}\left(M\right)$ we have*

$$\left\langle \tau_h f, g\right\rangle_M = \left\langle f, \tau_h^* g\right\rangle_M\ , \quad \tau_h^* g(x) = \mathrm{e}^{-h\cdot x - \frac{1}{2}|h|^2}\tau_{-h}g(x)\ ,$$

and $|h| \leq \sqrt{\log 2}$ implies $\left\|\tau_h^* g\right\|_{L^{(\cosh-1)}(\gamma))^*} \leq 2\left\|g\right\|_{L^{(\cosh-1)}(\gamma))^*}$. The translation mapping $h \mapsto \tau_h^* g$ is continuous in $L^{(\cosh-1)_*}\left(M\right)$.

(3) *If $f \in C_0^{(\cosh-1)}\left(M\right)$ then $\tau_h f \in C_0^{(\cosh-1)}\left(M\right)$, $h \in \mathbb{R}^n$, and the mapping $\mathbb{R}^n \colon h \mapsto \tau_h f$ is continuous in $L^{(\cosh-1)}\left(\gamma\right)$.*

Proof. (1) Assume $\left\|f\right\|_{L^{(\cosh-1)}(\gamma)} \leq 1$. For each $\rho > 0$, writing $\Phi = \cosh-1$,

$$\int \Phi(\rho\tau_h f(x))\gamma(x)\ dx\ = \int \Phi(\rho f(x-h))\gamma(x)\ dx\ = \int \Phi(\rho f(y))\ \gamma(y+h)\ dy = \mathrm{e}^{-\frac{1}{2}|h|^2}\int \mathrm{e}^{-h\cdot y}\Phi(\rho f(y))\gamma(y)\ dy\ \ ,$$

hence, using Hölder inequality and the inequality in Eq. (25),

$$(26)\quad \int \Phi(\rho\tau_h f(x))\gamma(x)\,dx \;\le\; \mathrm{e}^{-\frac{1}{2}|h|^2}\left(\int \mathrm{e}^{-2h\cdot y}\gamma(y)\,dy\right)^{\frac{1}{2}}\left(\int \Phi^2(\rho f(y))\gamma(y)\,dy\right)^{\frac{1}{2}} \le \frac{1}{\sqrt{2}}\mathrm{e}^{\frac{|h|^2}{2}}\left(\int \Phi(2\rho f(y))\gamma(y)\,dy\right)^{\frac{1}{2}} .$$

Take $\rho = 1/2$, so that $\mathbb{E}_\gamma\left[\Phi\left(\frac{1}{2}\tau_h f(x)\right)\right] \le \frac{1}{\sqrt{2}}\mathrm{e}^{\frac{|h|^2}{2}}$, which in turn implies $\tau_h f \in L^{(\cosh -1)}(\gamma)$. Moreover, $\|\tau_h f\|_{L^{(\cosh -1)}(\gamma)} \le 2$ if $\frac{1}{\sqrt{2}}\mathrm{e}^{\frac{|h|^2}{2}} \le 1$.

The semi-group property $\tau_{h_1+h_2}f = \tau_{h_1}\tau_{h_2}f$ implies the boundedness for all h.

(2) The computation of τ_h^* is

$$\begin{aligned}\langle \tau_h f, g\rangle_M &= \int f(x-h)g(x)\;M(x)dx\\ &= \int f(x)g(x+h)M(x+h)\;dx\\ &= \int f(x)\mathrm{e}^{-h\cdot x-\frac{1}{2}|h|^2}\tau_{-h}g(x)\;M(x)dx\\ &= \langle f, \tau_h^* g\rangle_M\;.\end{aligned}$$

Computing Orlicz norm of the mixture space, we find

$$\|\tau_h^* g\|_{(L^{(\cosh -1)}(\gamma))^*} = \sup\left\{\langle f, \tau_h^* g\rangle_M \,\middle|\, \|f\|_{L^{(\cosh -1)}(\gamma)} \le 1\right\} = \sup\left\{\langle \tau_h f, g\rangle_M \,\middle|\, \|f\|_{L^{(\cosh -1)}(\gamma)} \le 1\right\} .$$

From the previous item we know that $|h| \le \sqrt{\log 2}$ implies

$$\langle \tau_h f, g\rangle_M \le \|\tau_h f\|_{L^{(\cosh -1)}(\gamma)}\,\|g\|_{(L^{(\cosh -1)}(\gamma))^*} \le 2\,\|f\|_{L^{(\cosh -1)}(\gamma)}\,\|g\|_{(L^{(\cosh -1)}(\gamma))^*}\;,$$

hence $\|\tau_h^* g\|_{(L^{(\cosh -1)}(\gamma))^*} \le 2\,\|g\|_{L^{(\cosh -1)}(\gamma))^*}$.

Consider first the continuity a 0. We have for $|h| \le \sqrt{\log 2}$ and any $\phi \in C_0^\infty(\mathbb{R}^n)$ that

$$\|\tau_h g - g\|_{(L^{(\cosh -1)}(\gamma))^*} \le \|\tau_h(g-\phi)\|_{(L^{(\cosh -1)}(\gamma))^*} + \|\tau_h\phi - \phi\|_{(L^{(\cosh -1)}(\gamma))^*} + \|\phi - g\|_{(L^{(\cosh -1)}(\gamma))^*} \le 3\,\|g-\phi\|_{(L^{(\cosh -1)}(\gamma))^*} + \sqrt{2}\,\|\tau_h\phi - \phi\|_\infty\;.$$

The first term in the RHS is arbitrary small because of the density of $C_0^\infty(\mathbb{R}^n)$ in $L^{(\cosh -1)_*}(M)$, while the second term goes to zero as $h \to 0$ for each ϕ.

The general case follows from the boundedness and the semi-group property.

(3) If $f \in C_0^{(\cosh -1)}(M)$, then , by Prop. 43, the RHS of Eq. (26) is finite for all ρ, which in turn implies that $\tau_h f \in C_0^{(\cosh -1)}(M)$ because of Prop. 43.1. Other values of h are obtained by the semi-group property.

The continuity follows from the approximation argument, as in the previous item.

□ □

We denote by $\mathcal{P}$ the convex set of probability measures on $\mathbb{R}^n$ and call *weak convergence* the convergence of sequences in the duality with $C_b(\mathbb{R}^n)$. In the following proposition we denote by $\mathcal{P}_\mathrm{e}$ the set of probability measures μ such that $h \mapsto \mathrm{e}^{\frac{1}{2}|h|^2}$ is integrable. For example, this is the case when μ is Gaussian with variance $\sigma^2 I$, $\sigma^2 < 1$, or when μ has a bounded support.

Weak convergence in $\mathcal{P}_e$ means $\mu_n \to \mu$ weakly and $\int \mathrm{e}^{\frac{1}{2}|h|^2} \ \mu_n(dh) \to \int \mathrm{e}^{\frac{1}{2}|h|^2} \ \mu(dh)$. Note that we study here convolutions for the limited purpose of deriving the existence of smooth approximations obtained by *mollifiers*, see 108–109 of Brezis (2011).

Proposition 45 (Translation by a probability)**.** *Let $\mu \in \mathcal{P}_\mathrm{e}$.*

(1) *The mapping $f \mapsto \tau_\mu f$ is linear and bounded from $L^{(\cosh -1)}(\gamma)$ to itself. If, moreover, $\int \mathrm{e}^{\frac{1}{2}|h^2|} \ \mu(dh) \le \sqrt{2}$, then $\|\tau_\mu f\|_{L^{(\cosh -1)}(\gamma)} \le 2 \|f\|_{L^{(\cosh -1)}(\gamma)}$.*

(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 continuous at δ_0 from the weak convergence to the $L^{(\cosh -1)}(\gamma)$ norm.*

Proof.

(1) Let us write $\Phi = \cosh -1$ and note the Jensen's inequality

$$\Phi\left(\rho \tau_\mu f(x)\right) = \Phi\left(\rho \int f(x-h) \ \mu(dh)\right) \le \int \Phi\left(\rho f(x-h)\right) \ \mu(dh) = \int \Phi\left(\rho \tau_h f(x)\right) \ \mu(dh) \ .$$

By taking the Gaussian expectation of the previous inequality we have, as in the previous item,

$$\mathbb{E}_M\left[\Phi\left(\rho \tau_\mu f\right)\right] \le \int \int \Phi\left(\rho f(x-h)\right) \gamma(x) \ dx \ \ \mu(dh) = \int \mathrm{e}^{-\frac{1}{2}|h|^2} \int \mathrm{e}^{-h \cdot z} \Phi\left(\rho f(z)\right) \gamma(z) \ dz \ \ \mu(dh) \le \frac{1}{\sqrt{2}} \int \mathrm{e}^{\frac{1}{2}|h|^2} \ \mu(dh) \ \mathbb{E}_M\left[\Phi(2\rho f)\right] \ . \tag{27}$$

If $\|f\|_{L^{(\cosh -1)}(\gamma)} \le 1$ and $\rho = 1/2$, the RHS is bounded, hence $\tau_\mu f \in L^{(\cosh -1)}(\gamma)$. If, moreover, $\int \mathrm{e}^{\frac{1}{2}|h|^2} \le \sqrt{2}$, then the RHS is bounded by 1, hence $\|\tau_\mu f\|_{L^{(\cosh -1)}(\gamma)} \le 2$.

(2) We have found above that for each $\rho > 0$ it holds (27), where the right-end-side if finite for all ρ under the current assumption. It follows from Prop. 43 that $\tau_h f \in C_0^{(\cosh -1)}(M)$.

To prove the continuity at δ_0, assume $\int \mathrm{e}^{\frac{1}{2}|h|^2} \ \mu(dh) \le \sqrt{2}$, which is always feasible if $\mu \to \delta_0$ in $\mathcal{P}_\mathrm{e}$ weakly. Given $\epsilon > 0$, choose $\phi \in C_0^\infty(\mathbb{R}^n)$ so that $\|f - \phi\|_{L^{(\cosh -1)}(\gamma)} < \epsilon$. We have

$$\|\tau_\mu f - f\|_{L^{(\cosh -1)}(\gamma)} \le \|\tau_\mu (f - \phi)\|_{L^{(\cosh -1)}(\gamma)} + \|\tau_\mu \phi - \phi\|_{L^{(\cosh -1)}(\gamma)} + \|\phi - f\|_{L^{(\cosh -1)}(\gamma)} \le 3\epsilon + A^{-1} \|\tau_\mu \phi - \phi\|_\infty \ ,$$

where $A = \|1\|_{L^{(\cosh -1)}(\gamma)}$. As $\lim_{\mu \to \delta_0} \|\tau_\mu \phi - \phi\|_\infty = 0$, see e.g. § III-1.9 ofMalliavin (1995), the conclusion follows.

□ □

We use the previous propositions to show the existence of smooth approximations through sequences of mollifiers. A *bump* function is a non-negative function ω in $C_0^\infty(\mathbb{R}^n)$ such that $\int \omega(x) \ dx = 1$. It follows that $\int \lambda^{-n} \omega(\lambda^{-1} x) \ dx = 1$, $\lambda > 0$ and the family of mollifiers $\omega_\lambda(dx) = \lambda^{-n} \omega(\lambda^{-1} x) dx$, $\lambda > 0$, converges weakly to the Dirac mass at 0 as $\lambda \downarrow 0$ in $\mathcal{P}_\mathrm{e}$. Without restriction of generality, we shall assume that the support of ω is contained in $[-1, +1]^n$.

For each $f \in L^{(\cosh -1)}(\gamma)$ we have

$$\tau_{\omega_\lambda}(x) = f * \omega_\lambda(x) = \int f(x-y) \lambda^{-n} \omega(\lambda^{-1} y) \ dy = \int_{[-1,+1]^n} f(x - \lambda z) \omega(z) \ dz \ .$$

For each Φ convex we have by Jensen's inequality that

$$\Phi\left(f * \omega_\lambda(x)\right) \leq \left(\Phi \circ f\right) * \omega_\lambda(x)$$

and also

$$\int \Phi\left(f * \omega_\lambda(x)\right) M(x)\ dx \leq \int \int_{[-1,+1]^n} \Phi \circ f(x - \lambda z) \omega(z)\ dz M(x)\ dx =$$

$$\int \Phi \circ f(y) \left(\int_{[-1,+1]^n} \exp\left(-\lambda \left\langle z, y \right\rangle - \frac{\lambda^2}{2} \left|z\right|^2 \right) \omega(z)\ dz \right) M(y)\ dy \leq$$

$$\int \Phi \circ f(y) M(y)\ dy\ .$$

Proposition 46 (Mollifiers)**.** *Let be given a family of mollifiers ω_λ, $\lambda > 0$. For each $f \in C_0^{(\cosh -1)}(M)$ and 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 $C^\infty(\mathbb{R}^n)$. Moreover,

$$\lim_{\lambda \to 0} \left\| f * \omega_\lambda - f \right\|_{L^{(\cosh -1)}(\gamma)} = 0\ .$$

Proof. Any function in $L^{(\cosh -1)}(\gamma)$ belongs to $L^1_{\mathrm{loc}}(\mathbb{R}^n)$, hence

$$x \mapsto \int f(x-y) \omega_\lambda(y) dy = \int f(z) \omega_\lambda(z - x) dz$$

belongs to $C^\infty(\mathbb{R}^n)$, see e.g. Ch. 4 of . Note that $\int \mathrm{e}^{|h|^2/2} \omega_\lambda(dh) < +\infty$ and then apply Prop. 45(2).

□

Remark 4. Properties of weighted Orlicz spaces with the Δ_2-property can be sometimes deduced from the properties on the un-weighted spaces by suitable embeddings, but this is not the case for the exponential space. Here are two examples.

(1) Let $1 \leq a < \infty$. The mapping $g \mapsto g M^{\frac{1}{a}}$ 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 $\left\| \left[f * (g M^{\frac{1}{a}}) \right] M^{-\frac{1}{a}} \right\|_{L^a(M)} \leq \left\|f\right\|_{L^1(\mathbb{R}^n)} \left\|g\right\|_{L^a(M)}$.

(2) The mapping

$$g \mapsto \operatorname{sign}(g) \left(\cosh -1\right)_*^{-1} \left(M (\cosh -1)_*(g)\right)$$

is a surjection of $L^{(\cosh -1)_*}(\mathbb{R}^n)$ onto $L^{(\cosh -1)_*}(M)$ with inverse

$$h \mapsto \operatorname{sign}(h) \left(\cosh -1\right)_*^{-1} \left(M^{-1} (\cosh -1)_*(f)\right)\ .$$

It is surjective from unit vectors (for the Luxemburg norm) onto unit vectors.

We conclude this section by recalling the following tensor property of the exponential space and of the mixture space, see Lods and Pistone (2015).

Proposition 47. *Let us split the components $\mathbb{R}^n x \mapsto (x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and denote by M_1, M_2, respectively, the Maxwell densities on the factor spaces.*

(1) *A function f belongs to $L^{(\cosh -1)}(\gamma)$ if and only if for one $\alpha > 0$ the partial integral $x_1 \to \int (\cosh -1)(\alpha f(x_1, x_2)) M(x_2)\ dx_2$ is M_1-integrable.*

(2) *A function f belongs to $L^{(\cosh -1)_*}(M)$ if and only if the partial integral $x_1 \to \int (\cosh -1)_*(f(x_1, x_2)) M(x$ is M_1-integrable.*

8.3. **Gaussian statistical bundle.** It is an essential feature of the exponential affine manifold on $\mathcal{E}(M)$ discussed in Sec. 8.1 that the exponential *statistical bundle*

$$S\mathcal{E}(M) = \{(p,U) \mid p \in \mathcal{E}(M), U \in B_p\} \ ,$$

with $B_p = \left\{U \in L^{(\cosh -1)}(p \cdot M) \,\middle|\, \mathbb{E}_{p \cdot M}[U] = 0\right\}$ is an expression of the tangent bundle in the atlas $\{s_p \mid p \in \mathcal{E}(M)\}$. This depends on the fact that all fibers B_p are actually a closed subspace of the exponential space $L^{(\cosh -1)}(\gamma)$. This has been proved in Prop. 40. The equality of the spaces $L^{(\cosh -1)}(p \cdot M)$ and $L^{(\cosh -1)}(\gamma)$ is equivalent to $p \in \mathcal{E}(M)$, see the set of equivalent conditions called Portmanteau Theorem in Santacroce et al. (2016a).

We now investigate whether translation statistical models are sub-set of the maximal exponential model $\mathcal{E}(M)$ and whether they are sub-manifolds. Proper sub-manifolds of the exponential affine manifold should have a tangent bundle that splits the statistical bundle.

Let $p \in \mathcal{E}(M)$ and write $f = p \cdot M$. Then f is a positive probability density of the Lebesgue space and so are all its translations

$$\tau_h f(x) = p(x-h)M(x-h) = \mathrm{e}^{h\cdot x - \frac{1}{2}|h|^2} \tau_h p(x) \cdot M(x) = \tau_{-h}^* p(x) \cdot M(x) \ .$$

From Prop. **??** and Prop. 44.2 we know that the translated densities $\tau_{-h}^* p$, are in $L^{(\cosh -1)_*}(M)$ for all $h \in \mathbb{R}^n$ and the dependence on h is continuous.

Let us consider now the action of the translation on the values of the chart s_M. If $s_M(p) = U$, that is $p = \mathrm{e}^{U - K_M(U)}$ with $U \in \mathcal{S}_M$, then

$$\begin{aligned}\tau_{-h}^* p(x) = &\\ &\mathrm{e}^{h\cdot u - \frac{1}{2}|h|^2} \mathrm{e}^{U(x-h) - K_M(U)} = \exp\left(h \cdot u - \frac{1}{2}|h|^2 + \tau_h U - K_M(U)\right) = \\ &\exp\left((h\cdot u + \tau_h U - \mathbb{E}_M[\tau_h U]) - \left(K_M(U) + \frac{1}{2}|h|^2 - \mathbb{E}_M[\tau_h U]\right)\right) \ .\end{aligned}$$

Here $\tau_h U \in L^{(\cosh -1)}(\gamma)$ because of Prop. 44.1. If $\tau_{-h}^* p \in \mathcal{E}(M)$, then

$$s_M(\tau_{-h}^* p) = h \cdot u + \tau_h U - \mathbb{E}_M[\tau_h U] \ .$$

The expected value of the translated $\tau_h U$ is

$$\mathbb{E}_M[\tau_h U] = \int U(x-h)\gamma(x)\ dx \ = \mathrm{e}^{-\frac{1}{2}|h|^2} \int \mathrm{e}^{-h\cdot x} U(x) \gamma(x)\ dx \ .$$

We have found that the action of the translation on the affine coordinate $U = s_M(p)$ of a density $p \in \mathcal{E}(M)$ is

$$U \mapsto h \cdot u + \tau_h U - \mathrm{e}^{-\frac{1}{2}|h|^2} \mathbb{E}_M\left[\mathrm{e}^{-h\cdot u} U\right] \ , \tag{28}$$

and we want the resulting value belong to $\mathcal{S}_M$, i.e. we want to show that

$$\begin{aligned}\mathbb{E}_M[\exp(\gamma(h\cdot u + \tau_h U - \mathbb{E}_M[\tau_h U]))] = &\\ &\mathrm{e}^{\gamma\,\mathbb{E}_M[\tau_h U]}\, \mathbb{E}_M\left[\mathrm{e}^{\gamma h \cdot u}\right] \mathbb{E}_M\left[\mathrm{e}^{\gamma \tau_h U}\right] = \\ &\mathrm{e}^{\frac{\gamma^2}{2}|h|^2 + \gamma\,\mathbb{E}_M[\tau_h U]}\, \mathbb{E}_M\left[\mathrm{e}^{\gamma\tau_h U}\right] \ .\end{aligned}$$

is finite for γ in a neighborhood of 0.

We have the following result.

Proposition 48. (1) *If $p \in \mathcal{E}(M)$, for all $h \in \mathbb{R}^n$ the translated density $\tau_{-h}^* p$ is in $\mathcal{E}(M)$.*

(2) *If $s_M(p) \in C_0^{(\cosh -1)}(M)$, then $s_M(\tau_{-h}^* p) \in C_0^{(\cosh -1)}(M) \cap \mathcal{S}_M$ for all $h \in \mathbb{R}^n$ and dependence in h is continuous.*

Proof. (1) For each γ and conjugate exponents α, β, we have

$$\mathbb{E}_M\left[\mathrm{e}^{\gamma\tau_h U}\right] = \mathrm{e}^{-\frac{1}{2}|h|^2}\int \mathrm{e}^{-h\cdot x}\mathrm{e}^{\gamma U(x)}\gamma(x)\ dx\ \le$$

$$\mathrm{e}^{-\frac{1}{2}|h|^2}\left(\frac{1}{\alpha}\mathrm{e}^{\frac{\alpha^2}{2}|h|^2} + \frac{1}{\beta}\,\mathbb{E}_M\left[\mathrm{e}^{\beta\gamma U}\right]\right)\ .$$

As $U \in \mathcal{S}_M$, then $\mathbb{E}_M\left[\mathrm{e}^{\pm aU}\right] < \infty$ for some $a > 1$, and we can take $\beta = \sqrt{a}$ and $\gamma = \pm\sqrt{a}$.

(2) Under the assumed conditions on U the mapping $h \mapsto \tau_h U$ is continuous in $C_0^{(\cosh-1)}(M)$ because of Prop. 44.3. So is $h \mapsto \mathbb{E}_M\left[\tau_h U\right]$. As $u_i \in C_0^{(\cosh-1)}(M)$, $i = 1, \dots, n$, the same is true for $h \mapsto h \cdot u$. In conclusion, the translated U of (28) belongs to $C_0^{(\cosh-1)}(M)$.

□ □

The proposition above shows that the translation statistical model $\tau_{-h}^{*}p$, $h \in \mathbb{R}^m$ is well defined as a subset of $\mathcal{E}(M)$. To check if it is a differentiable sub-manifold, we want to compute the velocity of a curve $t \mapsto \tau_{h(t)}^{*}p$, that is

$$\frac{d}{dt}\left(h(t)\cdot u + \tau_{h(t)}U - \mathbb{E}_M\left[\tau_{h(t)}\right]U\right)\ .$$

That will require first of all the continuity in h, hence $U \in C_0^{(\cosh-1)}(M)$, and moreover we want to compute $\partial/\partial h_i U(x-h)$, that is the gradient of U. This task shall be the object of the next section.

Cases other than translations are of interest. Here are two sufficient conditions for a density to be in $\mathcal{E}(M)$.

Proposition 49.

(1) *Assume* $p > 0$ *M-a.s.,* $\mathbb{E}_M[p] = 1$, *and*

$$\mathbb{E}_M\left[p^{n_1/(n_1-1)}\right] \le 2^{n_1/(n_1-1)}, \quad \mathbb{E}_M\left[p^{-1/(n_2-1)}\right] \le 2^{n_2/(n_2-1)} \tag{29}$$

for some natural $n_1, n_2 > 2$. *Then* $p \in \mathcal{E}(M)$, *the exponential spaces are equal,* $L^{(\cosh-1)}(\gamma) = L^{(\cosh-1)}(p\cdot M)$, *and for all random variable* U

$$\|U\|_{L^{(\cosh-1)}(p\cdot M)} \le 2^{n_1}\|U\|_{L^{(\cosh-1)}(\gamma)}\ , \tag{30}$$

$$\|U\|_{L^{(\cosh-1)}(\gamma)} \le 2^{n_2}\|U\|_{L^{(\cosh-1)}(p\cdot M)}\ . \tag{31}$$

(2) *Condition* (29) *holds for* $p = \sqrt{\pi/2}\,|u_i|$ *and for* $p = u_i^2$, $i = 1, \dots, n$.

(3) *Let* χ *be a diffeomorphism of* $\mathbb{R}^n$ *and such that both the derivatives are uniformly bounded in norm. Then the density of the image under* χ *of the standard Gaussian measure belongs to* $\mathcal{E}(M)$.

Proof. (1) The bound on the moments in Eq.s (29) is equivalent to the inclusion in $\mathcal{E}(M)$ because of the definition of $\mathcal{S}_M$, or see (Santacroce et al., 2016a, Th. 4.7(vi)). Assume $\|U\|_{L^{(\cosh-1)}(\gamma)} \le 1$, that is $\mathbb{E}_M\left[(\cosh-1)(U)\right] \le 1$. From Hölder inequality and the elementary inequality in Eq. (25), we have

$$\mathbb{E}_{f\cdot M}\left[(\cosh-1)\left(\frac{U}{2^{n_1}}\right)\right] = \mathbb{E}_M\left[(\cosh-1)\left(\frac{U}{2^{n_1}}\right)f\right] \le$$

$$\mathbb{E}_M\left[(\cosh-1)\left(\frac{U}{2^{n_1}}\right)^{n_1}\right]^{1/n_1}\mathbb{E}_M\left[f^{n_1/(n_1-1)}\right]^{(n_1-1)/n_1} \le \frac{1}{2}\cdot 2 = 1$$

For the other direction, assume $\|U\|_{L^{(\cosh-1)}(f\cdot M)} \le 1$, that is $\mathbb{E}_M\left[\Phi(U)f\right] \le 1$, so that

$$\mathbb{E}_M\left[(\cosh -1)\left(\frac{U}{2^{n_2}}\right)\right] = \mathbb{E}_M\left[(\cosh -1)\left(\frac{U}{2^{n_2}}\right) f^{1/n_2} f^{-1/n_2}\right] \leq \\ \mathbb{E}_M\left[(\cosh -1)\left(\frac{U}{2^{n_2}}\right)^{n_2} f\right]^{1/n_2} \mathbb{E}_M\left[f^{-1/(n_2-1)}\right]^{(n_2-1)/n_2} \leq \frac{1}{2}\cdot 2 = 1 \ .$$

(2) Simple computations of moments.

(3) We consider first the case where $\chi(0) = 0$, in which case we have the following inequalities. If we define $\alpha^{-1} = \sup\left\{\left\|d\chi(x)\right\|^2 \,\middle|\, x \in \mathbb{R}^n\right\}$, then $\alpha\left|\chi(x)\right| \leq |x|$ for all $x \in \mathbb{R}^n$ and equivalently, $\alpha\left|x\right| \leq \left|\chi^{-1}(x)\right|$. In a similar way, if we define $\beta^{-1} = \sup\left\{\left\|d\chi^{-1}(y)\right\|^2 \,\middle|\, y \in \mathbb{R}^n\right\}$, then $\beta\left|\chi^{-1}(y)\right| \leq |y|$ and $\beta\left|x\right| \leq \left|\chi(x)\right|$.

The density of the image probability is $M\circ\chi^{-1}\left|\det d\chi^{-1}\right|$ and we want to show that for some $\epsilon > 0$ the following inequalities both hold,

$$\mathbb{E}_M\left[\left(\frac{M\circ\chi^{-1}\left|\det d\chi^{-1}\right|}{M}\right)^{1+\epsilon}\right] < \infty$$

and

$$\mathbb{E}_{M\circ\chi^{-1}\left|\det d\chi^{-1}\right|}\left[\left(\frac{M}{M\circ\chi^{-1}\left|\det d\chi^{-1}\right|}\right)^{1+\epsilon}\right] < \infty \ .$$

The first condition is satisfied as

$$\int \left|\det d\chi^{-1}(x)\right|^{1+\epsilon}\left(\frac{M\left(\chi^{-1}(x)\right)}{M(x)}\right)^{1+\epsilon} M(x)\ dx = \\ \int \left|\det d\chi^{-1}(x)\right|^{1+\epsilon} M\left(\chi^{-1}(x)\right)^{1+\epsilon} M(x)^{-\epsilon}\ dx \leq \\ (2\pi)^{-n/2}\beta^{-\frac{(1+\epsilon)n}{2}} \int \exp\left(-\frac{1}{2}\left((1+\epsilon)\left|\chi^{-1}(x)\right|^2 - \epsilon\left|x\right|^2\right)\right)\ dx = \\ (2\pi)^{-n/2}\beta^{-\frac{(1+\epsilon)n}{2}} \int \exp\left(-\frac{|x|}{2}\left((1+\epsilon)\frac{\left|\chi^{-1}(x)\right|}{|x|} - \epsilon\right)\right)\ dx \leq \\ (2\pi)^{-n/2}\beta^{-\frac{(1+\epsilon)n}{2}} \int \exp\left(-\frac{|x|}{2}\left((1+\epsilon)\alpha - \epsilon\right)\right)\ dx \ ,$$

where we have used the Hadamard's determinant inequality

$$\left|\det d\chi^{-1}(x)\right| \leq \left\|d\chi^{-1}(x)\right\|^n \leq \beta^{-n/2}$$

and the lower bound $\alpha \leq \frac{\left|\chi^{-1}(x)\right|}{|x|}$, $x \in \mathbb{R}^n_*$. If $\alpha \geq 1$ then $(1+\epsilon)\alpha - \epsilon = \alpha + \epsilon(\alpha - 1) \geq \alpha > 0$ for all ϵ. If $\alpha < 1$, then $(1+\epsilon)\alpha - \epsilon > 0$ if $\epsilon < \alpha/(1-\alpha)$ e.g., $\epsilon = \alpha/2(1-\alpha)$, which in turn gives $(1+\epsilon)\alpha - \epsilon = \alpha/2$. In conclusion, there exist an $\epsilon > 0$ such that

$$\int \left|\det d\chi^{-1}(x)\right|^{1+\epsilon}\left(\frac{M\left(\chi^{-1}(x)\right)}{M(x)}\right)^{1+\epsilon} M(x)\ dx \leq \\ (2\pi)^{-n/2}\left|\det d\chi^{-1}(x)\right|^{1+\epsilon} \int \exp\left(-\frac{\alpha\left|x\right|}{4}\right)\ dx = \left(\frac{\alpha}{2}\right)^{n/2} \ .$$

For the second inequality,

$$\int \left(\frac{M(y)}{M\left(\chi^{-1}(y)\right) \left|\det d\chi^{-1}(y)\right|} \right)^{1+\epsilon} M\left(\chi^{-1}(y)\right) \left|\det d\chi^{-1}(y)\right| \, dy =$$
$$\int M(y)^{1+\epsilon} M\left(\chi^{-1}(y)\right)^{-\epsilon} \left|\det d\chi^{-1}(y)\right|^{-\epsilon} \, dy =$$
$$\int M(\chi(x))^{1+\epsilon} M(x)^{-\epsilon} \left|\det d\chi^{-1}(\chi(x))\right|^{-\epsilon} \left|\det d\chi(x)\right| \, dx =$$
$$\int \left|\det d\chi(x)\right|^{1+\epsilon} M(\chi(x))^{1+\epsilon} M(x)^{-\epsilon} \, dx \ .$$

As the last term is equal to the expression in the previous case with χ^{-1} replaced by χ, the same proof applies with the bounds α and β exchanged.

□ □

Remark 5. While the moment condition for proving $p \in \mathcal{E}(M)$ has been repeatedly used, nonetheless the results above have some interest. The first one is an example where an explicit bound for the different norms on the fibers of the statistical bundle is derived. The second case is the starting point for the study of transformation models where a group of transformation χ_θ is given.

8.4. **Orlicz-Sobolev spaces with Gaussian weight.** The section offers an improvement upon the results presented in previous works i.e., Lods and Pistone (2015) ,Pistone (2018a). The aim is to discuss in detail the calculus of Orlicz-Sobolev spaces with Gaussian weight. There is an obvious relation with Stochastic Analysis in the sense of Malliavin Malliavin (1997).

Definition 50. The exponential and the mixture Orlicz-Sobolev-Gauss (OSG) spaces are, respectively,

$$W^{1,(\cosh -1)}(\gamma) = \left\{ f \in L^{(\cosh -1)}(\gamma) \,\middle|\, \partial_j f \in L^{(\cosh -1)}(\gamma) \right\} \ , \tag{32}$$

$$W^{1,(\cosh -1)_*}(\gamma) = \left\{ f \in L^{(\cosh -1)_*}(\gamma) \,\middle|\, \partial_j f \in L^{(\cosh -1)_*}(\gamma) \right\} \ , \tag{33}$$

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

$$\langle \partial_j f, \phi \rangle_\gamma = \langle \partial_j f, \phi\gamma \rangle = -\langle f, \partial_j(\phi\gamma) \rangle = \langle f, \gamma(u_j - \partial_j)\phi \rangle = \langle f, \delta_j \phi \rangle_\gamma \ ,$$

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

The meaning of both operators ∂_j and $\delta_j = (u_j - \partial_j)$ 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 γ-moments of all derivatives, hence $C_{\mathrm{p}}^\infty(\mathbb{R}^n) \subset W^{1,(\cosh -1)_*}(\gamma)$. 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 $\langle \phi, \delta_j f \rangle_\gamma = \langle \partial_j \phi, f \rangle_\gamma$.

The OSG spaces $W^1_{\cosh -1}(M)$ and $W^1_{(\cosh -1)_*}(M)$ are both Banach spaces, see (Musielak, 1983, Sec. 10). 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 is the graph norms,

$$\|f\|_{W^1_{(\cosh -1)}(\gamma)} = \|f\|_{L^{(\cosh -1)}(\gamma)} + \sum_{j=1}^n \|\partial_j f\|_{L^{(\cosh -1)}(\gamma)} \tag{34}$$

and

$$\|f\|_{W^1_{(\cosh-1)_*}(\gamma)} = \|f\|_{L^{(\cosh-1)}(\gamma)} + \sum_{j=1}^{n} \|\partial_j f\|_{L^{(\cosh-1)}(\gamma)} \ . \tag{35}$$

Because of Prop. 40, see also Th. 4.7 of Santacroce, Siri, and Trivellato (2016b), for each $p \in \mathcal{E}(M)$, we have both equalities and isomorphisms on $L^{(\cosh-1)}(p) = L^{(\cosh-1)}(\gamma)$ and $L^{(\cosh-1)_*}(p) = L^{(\cosh-1)_*}(\gamma)$. It follows

$$\begin{aligned} W^{1,(\cosh-1)}(\gamma) &= W^{1,(\cosh-1)}(p\cdot\gamma) \\ &= \left\{ f \in L^{(\cosh-1)}(p) \,\middle|\, \partial_j f \in L^{(\cosh-1)}(p) \right\} \ , \qquad (36) \\ W^{1,(\cosh-1)_*}(\gamma) &= W^{1,(\cosh-1)_*}(p\cdot\gamma) \\ &= \left\{ f \in L^{(\cosh-1)_*}(p) \,\middle|\, \partial_j f \in L^{(\cosh-1)_*}(p) \right\} \ , \qquad (37) \end{aligned}$$

and the graph norms are equivalent for any density $p \in \mathcal{E}(\gamma)$. The OSG spaces are compatible with the structure of the maximal exponential family $\mathcal{E}(\gamma)$. In particular, as all Gaussian densities of a given dimension belong into the same exponential manifold, one could have defined the OSG spaces with respect to any of such densities.

Example 6. Assume $q = \mathrm{e}^{v-K_\gamma(v)}\cdot\gamma$ and $p = \mathrm{e}^{u-K_\gamma(u)}\cdot\gamma$ with $q,p \in \mathcal{E}(\gamma)$ and $v,u \in \mathcal{S}_\gamma \cap W^{1,(\cosh-1)}(\gamma)$. The the Hyvärinen divergence is

$$\mathrm{DH}(p|q) = \frac{1}{2}\int \|\nabla(u-v)\|^2 p(x)\gamma(x)\ dx \ < +\infty$$

because $\nabla(u-v) \in L^{(\cosh-1)}(\gamma) = L^{(\cosh-1)}(p) \subset L^2(p\cdot\gamma)$.

Moreover, if we write $\partial_j u(x) p = \partial_j u(x)\mathrm{e}^{u-K_1(u)} = \partial_j \mathrm{e}^{u-K_1(u)}$, it holds

$$\int \nabla u(u)\cdot\nabla v(x) p(x)\gamma(x)\ dx \ = \sum_{j=1}^{n}\int \partial_j u(x)\partial_j v(x) p(x)\gamma(x)\ dx \ = \\ \sum_{j=1}^{n}\int \partial_j p(x)\partial_j v(x)\gamma(x)\ dx \ \ ,$$

and if we assume also $\partial_j v \in \mathbb{D}$ in the Gaussian space to get

$$\int \nabla p(x)\cdot\nabla v(x)\gamma(x)\ dx \ = \int p(x)\left(\sum_{i=1}^{n}\delta_j\partial_j\right) v(x)\gamma(x)\ dx \ = \mathbb{E}_p\left[\left(\sum_{i=1}^{n}\delta_j\partial_j\right) v\right] \ .$$

We review some relations between OSG-spaces and ordinary Sobolev spaces. We underline that much more precise results are available in the specialised literature on OSG-spaces.

For all $R > 0$

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

Proposition 51. *Let Ω_R denote the open sphere of radius $R > 0$ and consider the restriction $u \mapsto u_R$ of u to Ω_R.*

(1) *We have the continuous mappings*

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

(2) *We have the continuous mappings*

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

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

(4)

Proof. (1) From the inequality on M and from $(\cosh -1)(y) \geq y^{2n}/(2n)!$.

(2) From the inequality on M and from $y^2/2 \geq (\cosh -1)_*(y)$ and $\cosh(1)-1+(\cosh -1)_*(y) \geq |y|$.

(3) It is one of the Sobolev's embedding theorem, see Ch. 9 of Brezis (2011).

(4) By definition, $u \in W^{1,(\cosh -1)}(\gamma)$ if, and only if, $u, \partial_j u \in L^{(\cosh -1)}(\gamma)$, $j = 1, \dots, n$. Clearly, for all $a, b > 1$

$$
\begin{aligned}
u\mathrm{e}^{-\frac{1}{2a}|x|^2} &\in L^a(\mathbb{R}^n) \\
\partial_j u\mathrm{e}^{-\frac{1}{2b}|x|^2} &\in L^b(\mathbb{R}^n)
\end{aligned}
$$

For all $\phi \in C_0^\infty(\mathbb{R}^n)$,

$$
\begin{aligned}
\left\langle \partial_j \left(u\mathrm{e}^{-\frac{1}{2b}|u|^2}\right), \phi \right\rangle = -\left\langle u\mathrm{e}^{-\frac{1}{2b}|u|^2}, \partial_j\phi \right\rangle = -\left\langle u, \mathrm{e}^{-\frac{1}{2b}|u|^2}\partial_j\phi \right\rangle = \\
-\left\langle u, \partial_j\left(\mathrm{e}^{-\frac{1}{2b}|u|^2}\phi\right) - \frac{x_j}{a}\mathrm{e}^{-\frac{1}{2b}|u|^2}\phi \right\rangle = \\
\left\langle \partial_j u, \mathrm{e}^{-\frac{1}{2b}|u|^2}\phi \right\rangle + \left\langle u, \frac{u_j}{b}\mathrm{e}^{-\frac{1}{2b}|u|^2}\phi \right\rangle = \\
\left\langle \partial_j u\mathrm{e}^{-\frac{1}{2b}|u|^2} + u\frac{u_j}{b}\mathrm{e}^{-\frac{1}{2b}|u|^2}, \phi \right\rangle \ .
\end{aligned}
$$

We have shown that

$$
\partial_j \left(u\mathrm{e}^{-\frac{1}{2b}|u|^2}\right) = \partial_j u\mathrm{e}^{-\frac{1}{2b}|u|^2} + u\frac{u_j}{b}\mathrm{e}^{-\frac{1}{2b}|u|^2} \ .
$$

The first term in the RHS belongs to $L^b(\mathbb{R}^n)$. For the second term,

$$
\int \left| u(x)\frac{x_j}{b}\mathrm{e}^{-\frac{1}{2b}|x|^2} \right|^b \ dx \propto \int |x_j u(x)|^b \gamma(x) \ dx \quad ,
$$

which is bounded.

We have shown that $u \in W^{1,b}(\mathbb{R}^n)$ for all b. It follows from Morrey's theorem (Brezis, 2011, Th. 9.12) that $x \mapsto u(x)\mathrm{e}{-}\frac{1}{2b}|x|^2$ with $b > n$ is a.s. continuous (hence u is continuous) and for the continuous version

$$
\left| u(x)\mathrm{e}{-}\frac{1}{2b}|x|^2 - u(y)\mathrm{e}{-}\frac{1}{2b}|x|^2 \right| \leq C(b,n)
$$

□

Let us consider now the extension of the ∂_j operator to the OSG-spaces and its relation with the translation operator.

The operator given by the ordinary partial derivative $\partial_j \colon C_\mathrm{p}^\infty(\mathbb{R}^n) \to C_\mathrm{p}^\infty(\mathbb{R}^n) \subset L^{(\cosh -1)_*}(\gamma)$ is closable. In fact, if both $f_n \to 0$ and $\partial_j f_n \to \eta$ in $L^{(\cosh -1)_*}(\gamma)$, then for all $\phi \in C_0^\infty(\mathbb{R}^n)$,

$$
\langle \phi, \eta \rangle_\gamma = \lim_{n\to\infty} \langle \phi, \partial_j f_n \rangle_\gamma = \lim_{n\to\infty} \langle \delta\phi, f_n \rangle_\gamma = 0 \ ,
$$

hence $\eta = 0$. The same argument shows that $\partial_j \colon C_0^\infty(\mathbb{R}^n) \to C_0^\infty(\mathbb{R}^n) \subset L^{(\cosh -1)}(\gamma)$ is closable.

For $f \in L^{(\cosh -1)}(\gamma)$ we define $\tau_h f(x) = f(x-h)$ and it holds $\tau_h f \in L^{(\cosh -1)}(\gamma)$ because of Prop. 44(1). For each given $f \in W^{1,(\cosh -1)}(\gamma)$ we denote by $\partial_j f \in W^{1,(\cosh -1)}(\gamma)$, $j = 1, \dots, n$ its distributional partial derivatives and write $\nabla f = (\partial_j f \colon j = 1, \dots, n)$.

Proposition 52 (Continuity and directional derivative)**.**

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

$$
v(x+th) - v(x) = t\int_0^1 \nabla v(x+sth) \cdot h \ ds \ .
$$

Moreover, $|t| \leq \sqrt{2}$ implies

$$\left\| v(x+th) - v(x) \right\|_{L^{(\cosh -1)}(\gamma)} \leq 2t \left\| \nabla v \right\|_{L^{(\cosh -1)}(\gamma)} \ ,$$

especially, $\lim_{t \to 0} \left\| v(x+th) - v(x) \right\|_{L^{(\cosh -1)}(\gamma)} = 0$ uniformly in h.

(2) *For each $v \in W^{1,(\cosh -1)}(\gamma)$ the mapping $h \mapsto \tau_h v$ is differentiable from $\mathbb{R}^n$ to $L^{\infty - 0}(M)$ with gradient ∇v at $h = 0$.*

(3) *For each $v \in W^{1,(\cosh -1)}(\gamma)$ and each $f \in L^{(\cosh -1)_*}(\gamma)$, the mapping $h \mapsto \left\langle \tau_h v, f \right\rangle_\gamma$ is differentiable with derivative $\left\langle \tau_h \nabla v \cdot h, f \right\rangle_\gamma$. Conversely, if $v \in L^{(\cosh -1)}(\gamma)$ and $h \mapsto \left\langle \tau_h v, v \right\rangle_\gamma$ is differentiable for all $f \in L^{(\cosh -1)_*}(\gamma)$, with derivative $\left\langle d(f,h), f \right\rangle_\gamma$ then $f \in W^{1,(\cosh -1)}(\gamma)$ and .*

(4) *If $\partial_j v \in C_0^{(\cosh -1)}(\gamma)$, $j = 1, \dots, n$, then strong differentiability in $L^{(\cosh -1)}(\gamma)$ holds.*

Proof. (1) The measurable mapping $(s,x) \mapsto t \nabla v(x+sth) \cdot h$ is integrable as

$$\int_0^1 \int \left| t \nabla v(x+sth) \cdot h \right| \gamma(x) \ dx \ \ ds = \int_0^1 \int \left| t \nabla v(y) \cdot h \right| \gamma(y - sth) \ dy \ ds = \\ t \int \left| \nabla v(y) \cdot h \right| \int_0^1 \mathrm{e}^{sth \cdot y - \frac{s^2 t^2 |h|^2}{2}} \ ds \ \gamma(y) \ dy \ \leq \\ t \left| h \right| \left\| \nabla v \right\|_{L^2(\gamma)} \left(\int_0^1 \int \mathrm{e}^{2sth \cdot y - s^2 t^2 |h|^2} \gamma(y) \ dy \ \ ds \right)^{1/2}$$

and the value of the last integral is bounded,

$$\int_0^1 \int \mathrm{e}^{2sth \cdot y - s^2 t^2 |h|^2} \gamma(y) \ dy \ \ ds = \int_0^1 \mathrm{e}^{-s^2 t^2 |h|^2} \int \mathrm{e}^{2sth \cdot y} \gamma(y) \ dy \ \ ds = \\ \int_0^1 \mathrm{e}^{-s^2 t^2 |h|^2 + 4 s^2 t^2 |h|^2} \ ds = \int_0^1 \mathrm{e}^{3 s^2 t^2 |h|^2} \ ds \leq \mathrm{e}^{3 t^2 |h|^2} \ .$$

The partial integral $x \mapsto t \int_0^1 \nabla v(x+sth) \cdot h \ ds$ is a.s. defined and it is integrable with respect to $\gamma(x) dx$. Recall that for each test function $\phi \in C_0^\infty(\mathbb{R}^n)$ we have

$$\left\langle \partial_j v, \phi \right\rangle_\gamma = - \left\langle v, \partial_j (\phi \gamma) \right\rangle = \left\langle v, \delta_j \phi \right\rangle_\gamma \ .$$

We check the equality $\tau_{-th}v - v = t\int_0^1 \tau_{-sth}(\nabla v)\cdot h\ ds$ in the scalar product with a generic $\phi \in C_0^\infty(\mathbb{R}^n)$:

$$\begin{aligned}
\langle \tau_{-th}v - v, \phi\rangle_\gamma &= \int v(x+th)\phi(x)\gamma(x)\ dx - \int v(x)\phi(x)\gamma(x)\ dx \\
&= \int v(x)\phi(x-th)\gamma(x-th)\ dx - \int v(x)\phi(x)\gamma(x)\ dx \\
&= \int v(x)\left(\phi(x-th)\gamma(x-th) - \phi(x)\gamma(x)\right)\ dx \\
&= -t\int v(x)\int_0^1 \sum_{j=1}^n \partial_j(\phi\gamma)(x-sth)h_j\ ds\ dx \\
&= -t\int_0^1 \int v(x)\sum_{j=1}^n \partial_j(\phi\gamma)(x-sth)h_j\ dx\ ds \\
&= t\int_0^1 \int \sum_{j=1}^n \partial_j v(x)h_j\ \phi(x-sth)\gamma(x-sth)\ dx\ ds \\
&= t\int_0^1 \int \sum_{j=1}^n \partial_j v(x+sth)h_j\ \phi(x)\gamma(x)\ dx\ ds \\
&= \left\langle t\int_0^1 \tau_{-sth}(\nabla v)\cdot h\ ds, \phi\right\rangle_\gamma \ .
\end{aligned}$$

If $|t| \le \sqrt{\log 2}$ then the translation sth is small, $|sth| \le \sqrt{\log 2}$ so that, according to Prop. 44(1), we have$\|\tau_{-sth}(\nabla v\cdot h)\|_{L^{(\cosh -1)}(\gamma)} \le 2\,\|\nabla v\cdot h\|_{L^{(\cosh -1)}(\gamma)}$ and the thesis follows.

(2) We want to show that the following limit holds in all $L^\alpha(M)$-norms, $\alpha > 1$:

$$\lim_{t\to 0}\frac{\tau_{-th}v - v}{t} = h\cdot\nabla v\ .$$

Because of the identity in the previous Item, we need to check the limit

$$\lim_{t\to 0}\int\left|\int_0^1 (\tau_{-sth}(\nabla v(x)\cdot h) - \nabla v(x)\cdot h)\ ds\right|^\alpha \gamma(x)\ dx\ = 0\ .$$

Jensen's inequality gives

$$\int\left|\int_0^1 (\tau_{-sth}(\nabla v(x)\cdot h) - \nabla v(x)\cdot h)\ ds\right|^\alpha \gamma(x)\ dx\ \le \int_0^1\int \left|\tau_{-sth}(\nabla v(x)\cdot h) - \nabla v(x)\cdot h)\right|^\alpha \gamma(x)\ dx\ \ ds$$

and the result follows because translations are bounded and continuous in $L^\alpha(M)$.

(3) We have

$$\left\langle \int_0^1 (\tau_{(-sth)}v - v)\ ds, f\right\rangle_\gamma = \int_0^1 \left\langle \tau_{(-sth)}v - v, f\right\rangle_\gamma\ ds = \int_0^1 \left\langle v, \tau^*_{(-sth)}f - f\right\rangle_\gamma\ ds\ .$$

Conclusion follows because $t \mapsto \tau^*_{-sth}f$ is continuous in $L^{(\cosh -1)_*}(\gamma)$.

Assume now $v \in L^{(\cosh -1)}(\gamma)$ and $h \mapsto \tau_h v$ is weakly differentiable. There exist $v_1, \dots, v_n \in L^{(\cosh -1)}(\gamma)$ such that for all $\phi \in C_0(\mathbb{R}^n)$ and j

$$\langle v_j, \phi\gamma \rangle = \langle v_j, \phi \rangle_\gamma = \left. \frac{d}{dt} \left\langle \tau_{-te_j} v, \phi \right\rangle_\gamma \right|_{t=0} = \left. \frac{d}{dt} \left\langle \tau_{-te_j} v, \phi\gamma \right\rangle \right|_{t=0} = \frac{d}{dt} \left\langle v, \tau_{te_j}(\phi\gamma) \right\rangle = - \left\langle v, \partial_j(\phi\gamma) \right\rangle \ .$$

The distributional derivative holds because $\phi\gamma$ is the generic element of $C_0^\infty(\mathbb{R}^n)$.

(4) For each $\rho > 0$ Jensen's inequality implies

$$\left\| \int_0^1 (\tau_{-sth}(\nabla f \cdot h) - \nabla f \cdot h) \ ds \ \gamma(x)dx \right\|_{L^{(\cosh -1)}(\gamma)} \leq \int_0^1 \left\| (\tau_{-sth}(\nabla f \cdot h) - \nabla f \cdot h) \ \gamma(x)dx \right\|_{L^{(\cosh -1)}(\gamma)} \ ds \ .$$

As in Prop. 44(1) we choose $|t| \leq \sqrt{\log 2}$ to get $|ste_j| \leq \sqrt{\log 2}$, $0 \leq s \leq 1$, so that $\|\tau_{-sth} \nabla f \cdot h\|_{L^{(\cosh -1)}(\gamma)} \leq 2 \|\nabla f \cdot h\|_{L^{(\cosh -1)}(\gamma)}$, hence the integrand is bounded by $\|\nabla f \cdot h\|_{L^{(\cosh -1)}(\gamma)}$. The convergence for each s follows from the continuity of the translation on $C_0^{(\cosh -1)}(\gamma)$.

□

Notice that in Item 2. of the proposition we could have derived a stronger differentiability if the mapping $h \mapsto \tau_h \nabla f$ were continuous in $L^{(\cosh -1)}(\gamma)$. That, and other similar observations, lead to the following definition.

Definition 53. The *Orlicz-Sobolev-Gauss exponential class* is

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

The following density results will be used frequently in approximation arguments. We denote by $(\omega_n)_{n \in \mathbb{N}}$ a sequence of mollifiers.

Proposition 54 (Calculus in $C_0^{1,(\cosh -1)}(\gamma)$)**.** (1) *For each $f \in C_0^{1,(\cosh -1)}(\gamma)$ the sequence $f * \omega_n$, $n \in \mathbb{N}$, belongs to $C^\infty(\mathbb{R}^n) \cap W^{1,(\cosh -1)}(\gamma)$. Precisely, for each n and $j = 1, \dots, n$, we have the equality $\partial_j(f * \omega_n) = (\partial_j f) * \omega_n$; the sequences $f * \omega_n$, respectively $\partial_j f * \omega_n$, $j = 1, \dots, n$, converge to f, respectively $\partial_j f$, $j = 1, \dots, n$, strongly in $L^{(\cosh -1)}(\gamma)$.*

(2) *Same statement is true if $f \in W^{1,(\cosh -1)_*}(\gamma)$.*

(3) *Let be given $f \in C_0^{1,(\cosh -1)}(\gamma)$ and $g \in W^{1,(\cosh -1)_*}(\gamma)$. Then $fg \in W^{1,1}(M)$ and $\partial_j(fg) = \partial_j fg + f\partial_j g$.*

(4) *Let be given $F \in C^1(\mathbb{R})$ with $\|F'\|_\infty < \infty$. For each $U \in C_0^{1,(\cosh -1)}(\gamma)$, we have $F \circ U, F' \circ U \partial_j U \in C_0^{(\cosh -1)}(\gamma)$ and $\partial_j F \circ U = F' \circ U \partial_j U$, in particular $F(U) \in C_0^{1,(\cosh -1)}(\gamma)$.*

Proof. (1) We need only to note that the equality $\partial_j(f * \omega_n) = (\partial_j f) * \omega_n$ is true for $f \in W^{1,(\cosh -1)}(\gamma)$. Indeed, the sequence $f * \omega_n$ belongs to $C^\infty(\mathbb{R}^n) \cap L^{(\cosh -1)}(\gamma)$ and converges to f in $L^{(\cosh -1)}(\gamma)$-norm according from Prop. 46. The sequence $\partial_j f * \omega_n = (\partial_j f) * \omega_n$ converges to $\partial_j f$ in $L^{(\cosh -1)}(\gamma)$-norm because of the same theorem.

(2) Same proof.

(3) Note that $fg, \partial_j fg + f\partial_j g \in L^1(M)$. The following converge in $L^1(M)$ holds

$$\partial_j fg + f\partial_j g = \lim_{n \to \infty} \partial_j f * \omega_n g * \omega_n + f * \omega_n \partial_j * \omega_n = \lim_{n \to \infty} \partial_j f * \omega_n g * \omega_n \ ,$$

so that for all $\phi \in C_0^\infty(\mathbb{R}^n)$

$$\langle \partial_j f g + f \partial_j g, \phi \rangle = \lim_{n\to\infty} \langle \partial_j f * \omega_n g * \omega_n, \phi \rangle = \lim_{n\to\infty} - \langle f * \omega_n g * \omega_n, \partial_j \phi \rangle = - \langle fg, \partial_j \phi \rangle \ .$$

It follows that the distributional partial derivative of the product is $\partial_j fg = \partial_j f g + f \partial_j g$, in particular belongs to $L^1(M)$, hence $fg \in W^{1,1}(M)$.

(4) From the assumption on F we have $|F(U)| \leq |F(0)| + \|F'\|_\infty |U|$. It follows $F \circ U \in L^{(\cosh -1)}(\gamma)$ because

$$\int (\cosh -1)\left(\rho F(U(x))\right) \ \gamma(x) dx \leq \frac{1}{2}(\cosh -1)(2\rho F(0)) + \frac{1}{2} \int (\cosh -1)\left(2\rho \left\|F'\right\|_\infty U(x)\right) \ \gamma(x) dx \ ,$$

and $\rho \left\|F(U)\right\|_{L^{(\cosh -1)}(\gamma)} \leq 1$ if both

$$(\cosh -1)(2\rho F(0)) \leq 1, \quad 2\rho \left\|F'\right\|_\infty \left\|U\right\|_{L^{(\cosh -1)}(\gamma)} \leq 1 \ .$$

In the same way we show that $F' \circ U \partial_j U \in L^{(\cosh -1)}(\gamma)$. Indeed,

$$\int (\cosh -1)\left(\rho F'(U(x)) \partial_j U(x)\right) \quad \gamma(x) dx \quad \leq \quad \int (\cosh -1)\left(\rho \left\|F'\right\|_\infty \partial_j U(x)\right) \quad \gamma(x) dx \quad ,$$

so that $\rho \left\|F' \circ U \partial_j U\right\|_{L^{(\cosh -1)}(\gamma)} \leq 1$ if $\rho \left\|F'\right\|_\infty \left\|\partial_j U(x)\right\|_{L^{(\cosh -1)}(\gamma)} = 1$. Because of the Item (1) the sequence $U * \omega_n$ belongs to C^∞ and converges strongly in $L^{(\cosh -1)}(\gamma)$ to U, so that from

$$\left\|F \circ (U * \omega_n) - F \circ U\right\|_{L^{(\cosh -1)}(\gamma)} \leq \left\|F'\right\|_\infty \left\|U * \omega_n - U\right\|_{L^{(\cosh -1)}(\gamma)}$$

we see that $F \circ (U * \omega_n) \to F \circ U$ in $L^{(\cosh -1)}(\gamma)$. In the same way,

$$\begin{aligned} \left\|F' \circ (U * \omega_n) \partial_j (U * \omega_n) - F' \circ U \partial_j U\right\|_{L^{(\cosh -1)}(\gamma)} \leq \\ \left\|F' \circ (U * \omega_n)(\partial_j (U * \omega_n) - \partial_j U)\right\|_{L^{(\cosh -1)}(\gamma)} \\ + \left\|(F' \circ (U \circ \omega_n) - F' \circ U) \partial_j U\right\|_{L^{(\cosh -1)}(\gamma)} \leq \\ \left\|F'\right\|_\infty \left\|\partial_j (U * \omega_n) - \partial_j U\right\|_{L^{(\cosh -1)}(\gamma)} + \\ \left\|(F' \circ (U \circ \omega_n) - F' \circ U) \partial_j U\right\|_{L^{(\cosh -1)}(\gamma)} \ . \end{aligned}$$

The first term goes clearly to 0, while the second term requires consideration. Note the bound

$$\left|(F' \circ (U \circ \omega_n) - F' \circ U) \partial_j U\right| \leq 2 \left\|F'\right\|_\infty \left|\partial_j U\right| \ ,$$

so that the sequence $(F' \circ (U \circ \omega_n) - F' \circ U) \partial_j U$ goes to zero in probability and is bounded by a function in $C_0^{(\cosh -1)}(\gamma)$. This in turn implies the convergence in $L^{(\cosh -1)}(\gamma)$.

Finally we check that the distributional derivative of $F \circ U$ is $F' \circ U \partial_j U$: for each $\phi \in C_0^\infty(\mathbb{R}^n)$

$$\begin{aligned}\langle \partial_j F \circ U, \phi\gamma \rangle &= - \langle F \circ U, \partial_j(\phi M) \rangle \\ &= - \langle F \circ U, \delta_j \phi \rangle_\gamma \\ &= \lim_{n\to\infty} \langle F \circ (U * \omega_n), \delta_j \phi \rangle_\gamma \\ &= \lim_{n\to\infty} \langle \partial_j F \circ (U * \omega_n), \phi \rangle_\gamma \\ &= \lim_{n\to\infty} \left\langle F' \circ (U * \omega_n) \partial_j (U * \omega_n), \phi \right\rangle_\gamma \\ &= \left\langle F' \circ U \partial_j U, \phi \right\rangle_\gamma \\ &= \left\langle F' \circ U \partial_j U, \phi\gamma \right\rangle \ .\end{aligned}$$

□

We conclude our presentation by re-stating a technical result from Prop. 15 of Lods and Pistone (2015), where the assumptions where not sufficient for the stated result.

Proposition 55.

(1) *If $U \in \mathcal{S}_\gamma$ and $f_1, \dots, f_m \in L^{(\cosh -1)}(\gamma)$, then $f_1 \cdots f_m \mathrm{e}^{M - K_M(M)} \in L^\gamma(M)$ for some $\gamma > 1$, hence it is in $L^{(\cosh -1)_*}(\gamma)$.*

(2) *If $U \in \mathcal{S}_\gamma \cap C_0^{1,(\cosh -1)}(\gamma)$ and $f \in C_0^{1,(\cosh -1)}(\gamma)$, then*

$$f \mathrm{e}^{u - K_M(u)} \in W^{1,(\cosh -1)_*}(\gamma) \cap C(\mathbb{R}^n) \ ,$$

and its distributional partial derivatives are $(\partial_j f + f \partial_j u) \mathrm{e}^{u - K_M(u)}$

Proof. (1) From We know that $\mathrm{e}^{U - K_M(U)} \cdot \gamma \in \mathcal{E}(\gamma)$ and $\mathrm{e}^{U - K_M(U)} \in L^{1+\varepsilon}(M)$ for some $\varepsilon > 0$. From that, let us prove that $f_1 \cdots f_m \mathrm{e}^{U - K_M(U)} \in L^\gamma(M)$ for some $\gamma > 1$. According to classical (m+1)-term Fenchel-Young inequality,

$$|f_1(x) \cdots f_n(x)| \, \mathrm{e}^{U(x) - K_M(U)} \leq \sum_{i=1}^m \frac{1}{\alpha_i} |f_i(x)|^{\alpha_i} + \frac{1}{\beta} \left| \mathrm{e}^{U(x) - K_M(U)} \right|^\beta ,$$

$$\alpha_1, \dots, \alpha_m, \beta > 1, \sum_{i=1}^m \frac{1}{\alpha_i} + \frac{1}{\beta} = 1, x \in \mathbb{R}^n.$$

Since $(\cosh -1)_*$ is convex, we have

$$\mathbb{E}_M \left[(\cosh -1)_* (|f_1 \cdots f_m| \, \mathrm{e}^{U - K_M(U)}) \right] \leq \sum_{i=1}^m \frac{1}{\alpha_i} \mathbb{E}_\gamma \left[(\cosh -1)_* (|f_i|^{\alpha_i}) \right] + \frac{1}{\beta} \mathbb{E}_M \left[(\cosh -1)_* \left(\mathrm{e}^{\beta(U - K_M(U))} \right) \right] .$$

Since $f_1, \dots, f_m \in L^{(\cosh -1)}(\gamma) \subset \cap_{\alpha > 1} L^\alpha(M)$, one has $|f_i|^{\alpha_i} \in L^{(\cosh -1)_*}(\gamma)$, for $i = 1, \dots, m$ and all $\alpha_i > 1$, hence $\mathbb{E}_M \left[(\cosh -1)_* (|f_i|^{\alpha_i}) \right] < \infty$ for $i = 1, \dots, m$ and all $\alpha_i > 1$. By choosing $1 < \beta < 1 + \varepsilon$ one has $\mathrm{e}^{\beta(U(x) - K_M(U))} \in L^\gamma(M) \subset L^{(\cosh -1)_*}(\gamma)$, $\gamma = \frac{1+\varepsilon}{\beta}$, so that $\mathbb{E}_M \left[(\cosh -1)_* \left(\mathrm{e}^{\beta(U - K_M(U))} \right) \right] < \infty$. This proves that $(\cosh -1)_*(f_1 \cdots f_m \mathrm{e}^{U - K_M(U)}) \in L^1(M)$, which implies $f_1 \cdots f_m \mathrm{e}^{U(x) - K_M(U)} \in L^{(\cosh -1)_*}(\gamma)$.

(2) From the previous item we know $f \mathrm{e}^{U - K_M(U)} \in L^{(\cosh -1)_*}(\gamma)$. For each $j = 1, \dots, n$ from prop. 54(3) we have the distributional derivative $\partial_j (f \mathrm{e}^U) = \partial f \mathrm{e}^U + f \partial_j \mathrm{e}^{U - K_M(U)}$ we we need to show a composite function derivation, namely $\partial_j \mathrm{e}^{U - K_M(U)} = \partial_j u \mathrm{e}^{U - K_M(U)}$. Let $\chi \in C_0^\infty(\mathbb{R}^n)$ be a cut-off equal to 1 on the ball of radius 1, zero outside the ball of radius

2, derivative bounded by 2, and for $n \in \mathbb{N}$ consider the function $x \mapsto F_n(x) = \chi(x/n)\mathrm{e}^x$ which is $C^\infty(\mathbb{R}^n)$ and whose derivative is bounded:

$$F_n'(x) = \left(\frac{1}{n}\chi'(x/n) + \chi(x/n)\right)\mathrm{e}^x \leq \left(\frac{2}{n} + 1\right)\mathrm{e}^{2n} \ .$$

As Prop. 54(4) applies, we have $\partial_j F_n(U) = F_n'(U)\partial_j U \in C_0^{(\cosh -1)}(\gamma)$. Finally, for each $\phi \in C_0^\infty(\mathbb{R}^n)$,

$$\begin{aligned}
\left\langle \partial_j \mathrm{e}^U, \phi \right\rangle &= -\left\langle \mathrm{e}^U, \partial_j \phi \right\rangle \\
&= -\lim_{n\to\infty} \left\langle F_n(U), \partial_j \phi \right\rangle \\
&= \lim_{n\to\infty} \left\langle \partial F_n(U), \phi \right\rangle \\
&= \lim_{n\to\infty} \left\langle (\frac{1}{n}\chi'(U/n) + \chi(U/n))\partial_j U \mathrm{e}^U, \phi \right\rangle \\
&= \left\langle \partial_j U \mathrm{e}^U, \phi \right\rangle \ .
\end{aligned}$$

□

Remark 7. As a particular case of the above proposition, we see that $U \in \mathcal{S}_\gamma \cap C_0^{1,(\cosh -1)}(\gamma)$ implies

$$\mathrm{e}^{U - K_M(U)} \in W^{1,(\cosh -1)*}(\gamma) \qquad \text{with} \quad \vec{\nabla}\mathrm{e}^{U-K_M(U)} = \vec{\nabla}u\,\mathrm{e}^{U-K_M(U)} \ .$$

References

S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao. *Differential geometry in statistical inference*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 10. Institute of Mathematical Statistics, 1987. ISBN 0-940600-12-9.

Shun-ichi Amari. Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.*, 10(2):357–385, 1982. ISSN 0090-5364.

Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985. ISBN 3-540-96056-2.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, feb 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746. URL http://dx.doi.org/10.1162/089976698300017746.

Shun-ichi Amari. *Information geometry and its applications*, volume 194 of *Applied Mathematical Sciences*. Springer, [Tokyo], 2016. ISBN 978-4-431-55977-1; 978-4-431-55978-8. URL https://doi.org/10.1007/978-4-431-55978-8.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. American Mathematical Society, 2000. Translated from the 1993 Japanese original by Daishi Harada.

Antonio Ambrosetti and Giovanni Prodi. *A primer of nonlinear analysis*, volume 34 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1993. ISBN 0-521-37390-5.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zrich. Birkhuser Verlag, Basel, second edition, 2008. ISBN 978-3-7643-8721-1.

Jrgen Appell and Petr P. Zabrejko. *Nonlinear superposition operators*, volume 95 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1990. ISBN 0-521-36102-8. doi: 10.1017/CBO9780511897450. URL http://dx.doi.org/10.1017/CBO9780511897450.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64 of *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]*. Springer, Cham, 2017. ISBN 978-3-319-56477-7; 978-3-319-56478-4.

Ole E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* John Wiley & Sons, 1978.

Alexander Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2002. ISBN 0-8218-2968-8.

Nicolas Bourbaki. *Varits differentielles et analytiques. Fascicule de rsultats / Paragraphes 1 7.* Number XXXIII in lments de mathmatiques. Hermann, 1971.

Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Universitext. Springer, New York, 2011. ISBN 978-0-387-70913-0.

Lawrence D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory.* Number 9 in IMS Lecture Notes. Monograph Series. Institute of Mathematical Statistics, 1986.

Alberto Cena. *Geometric structures on the non-parametric statistical manifold.* PhD thesis, Università degli Studi di Milano, 2002.

Alberto Cena and Giovanni Pistone. Exponential statistical manifold. *Ann. Inst. Statist. Math.*, 59(1):27–56, 2007. ISSN 0020-3157.

Matheus R. Grasselli. Dual connections in nonparametric classical information geometry. arXiv:math-ph/0104031v1, 2001.

Ian Hacking. *The Emergence of Probability.* Cambridge University Press, New York, 2nd edition, 2006. ISBN ISBN 978-0-521-86655-2.

Alan Hájek. Interpretations of probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. ISSN 1532-4435.

M. A. Krasnosel'skii and Ya. B. Rutickii. *Convex Functions and Orlicz Spaces.* Noordhoff, 1961. Russian original: (1958) Fizmatgiz, Moskva.

Andreas Kriegl and Peter W. Michor. *The convenient setting of global analysis*, volume 53 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 1997. ISBN 0-8218-0780-3. doi: 10.1090/surv/053. URL `https://doi.org/10.1090/surv/053`.

Lev D. Landau and Eugenij M. Lifshits. *Course of Theoretical Physics. Statistical Physics.*, volume V. Butterworth-Heinemann, 3rd edition, 1980.

Serge Lang. *Differential and Riemannian manifolds*, volume 160 of *Graduate Texts in Mathematics.* Springer-Verlag, third edition, 1995. ISBN 0-387-94338-2.

Betrand Lods and Giovanni Pistone. Information geometry formalism for the spatially homogeneous Boltzmann equation. *Entropy*, 17(6):4323–4363, 2015.

John Lott. Some geometric calculations on Wasserstein space. *Comm. Math. Phys.*, 277(2):423–437, 2008. ISSN 0010-3616. doi: 10.1007/s00220-007-0367-3. URL `https://doi.org/10.1007/s00220-007-0367-3`.

Paul Malliavin. *Integration and probability*, volume 157 of *Graduate Texts in Mathematics.* Springer-Verlag, 1995. ISBN 0-387-94409-5. With the collaboration of Héléne Airault, Leslie Kay and Gérard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky.

Paul Malliavin. *Stochastic analysis*, volume 313 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, 1997. ISBN 3-540-57024-1.

Henry McKean. *Probability: the classical limit theorems.* Cambridge University Press, Cambridge, 2014. ISBN 978-1-107-62827-4; 978-1-107-05321-2. doi: 10.1017/CBO9781107282032. URL `https://doi.org/10.1017/CBO9781107282032`.

Luigi Montrucchio and Giovanni Pistone. Deformed exponential bundle: the linear growth case. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information*, number 10589 in LNCS, pages 239–246. Springer, 2017. Third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings.

Julian Musielak. *Orlicz spaces and modular spaces*, volume 1034 of *Lecture Notes in Mathematics*. Springer-Verlag, 1983. ISBN 3-540-12706-2.

Ivan Nourdin and Giovanni Peccati. *Normal approximations with Malliavin calculus*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2012. ISBN 978-1-107-01777-1. doi: 10.1017/CBO9781139084659. URL `http://dx.doi.org/10.1017/CBO9781139084659`. From Stein's method to universality.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001. ISSN 0360-5302. URL `../publications/Riemann.ps`.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *Ann. Statist.*, 40(1):561–592, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS971. URL `http://dx.doi.org/10.1214/12-AOS971`.

Giovanni Pistone. Nonparametric information geometry. In *Geometric Science of Information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013a.

Giovanni Pistone. Nonparametric information geometry. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric science of information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013b. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings.

Giovanni Pistone. Translations in the exponential Orlicz space with Gaussian weight. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information*, number 10589 in LNCS, pages 569–576. Springer, 2017. Third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings.

Giovanni Pistone. Information geometry of the Gaussian space. In *Information geometry and its applications*, volume 252 of *Springer Proc. Math. Stat.*, pages 119–155. Springer, Cham, 2018a.

Giovanni Pistone. Information geometry of the Gaussian space. arXiv:1803.08135, 2018b.

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(5):1543–1561, October 1995. ISSN 0090-5364.

M. M. Rao and Z. D. Ren. *Applications of Orlicz spaces*, volume 250 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., 2002. ISBN 0-8247-0730-3.

Marina Santacroce, Paola Siri, and Barbara Trivellato. New results on mixture and exponential models by Orlicz spaces. *Bernoulli*, 22(3):1431–1447, 2016a. ISSN 1350-7265. doi: 10.3150/15-BEJ698. URL `https://doi.org/10.3150/15-BEJ698`.

Marina Santacroce, Paola Siri, and Barbara Trivellato. New results on mixture and exponential models by Orlicz spaces. *Bernoulli*, 22(3):1431–1447, 2016b. ISSN 1350-7265. doi: 10.3150/15-BEJ698. URL `https://doi.org/10.3150/15-BEJ698`.

Marina Santacroce, Paola Siri, and Barbara Trivellato. Exponential models by Orlicz spaces and applications. *J. Appl. Probab.*, 55(3):682–700, 2018. ISSN 0021-9002. doi: 10.1017/jpr.2018.45. URL `https://doi.org/10.1017/jpr.2018.45`.

Mark J. Schervish. *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995. ISBN 0-387-94546-6. doi: 10.1007/978-1-4612-4250-5. URL `https://doi.org/10.1007/978-1-4612-4250-5`.

Lawrence Sklar. *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge University Press, 1993. doi: 10.1017/CBO9780511624933.

Daniel W. Stroock. *Partial differential equations for probabilists*, volume 112 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2008. ISBN 978-0-521-88651-2. doi: 10.1017/CBO9780511755255. URL `http://dx.doi.org/10.1017/CBO9780511755255`.

Harald Upmeier. *Symmetric Banach manifolds and Jordan C^*-algebras*, volume 104 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., 1985. ISBN 0-444-87651-0. Notas de Matemtica [Mathematical Notes], 96.

Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL `https://doi.org/10.1017/9781108231596`. With a foreword by Sara van de Geer.

Martin J. Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019. ISBN 9781108498029. doi: 10.1017/9781108627771.

de Castro Statistics, Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino IT
E-mail address: `giovanni.pistone@carloalberto.org`
URL: `www.giannidiorestino.it`