Information Geometry of the Probability Simplex: A Short Course

Giovanni Pistone*
de Castro Statistics, Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino, Italy

This set of notes is intended for a short course aiming to provide an (almost) self-contained and (almost) elementary introduction to the topic of Information Geometry (IG) of the probability simplex. Such a course can be considered an introduction to the original monograph by Amari and Nagaoka [1], and to the recent monographs by Amari [2] and by Ay, Jost, Lê, and Schwachhöfer [3]. The focus is on a non-parametric approach, that is, I consider the geometry of the full probability simplex and compare the IG formalism with what is classically done in Statistical Physics.

I. INTRODUCTION

The key Amari's contribution to our subject has been convincingly showing that Fisher-Rao Riemannian structure on the probability simplex is just one among the geometric structures of interest. In fact, Amari describes the geometry of the probability simplex as an affine space with a natural system of dually flat connections. My aim here is to present Amari's ideas while avoiding the use of parametric differential geometry, my point being that such presentation better reveals the substantial connection with standard arguments in Boltzmann-Gibbs theory as introduced, for example, in Landau and Lifshits [4, Ch. I-III]. A second positive effect of the non-parametric approach is to provide a better preparation for interesting generalization, namely, infinite sample space, deformed logarithmic representation, Wasserstein geometry. In the text, I am freely using material from a number of papers that followed Pistone and Sempi [5].

I will not consider here any specific application. In fact, the presentation is limited to the consideration of the basic formalism. If Chance is to be accepted as a real object, as are Space, Time, Space-Time, ..., then something like IG should be the mathematics of Chance, in the same sense Cartesian Geometry is a mathematics of classical Space.

A statistical model is a parametrised set of probability functions. The point of view of Information geometry (IG) is that a statistical model must be viewed as a submanifold of a manifold on all probability functions. This statement requires a number of qualification to be technically feasible. Let us start by considering a few basic examples.

1) On the sample space of two binary trial, $\Omega = \{0,1\}^2$, the set of all possible probability functions is the probability simplex $\Delta(\Omega)$. It is a convex set whose dimension is 3, conveniently represented as 2-way table with elements $p(x,y) \geq 0$, $x,y = 0,1$, $\sum p(x,y) = 1$.

2) The model of two independent identically distributed binary trials is a 1-parameter model that can be seen as a curve in the probability simplex of *1)*. One possible parametrization is

$$[0,1] \ni \theta \mapsto p(x,y;\theta) = ((1-\theta)x + \theta x)((1-\theta)y + \theta y) \ .$$

Another one is $\theta \mapsto (1-\theta)^{1-x-y}\theta^{x+y}$.

3) The model of two independent binary trial has 2 independent parameters and can be seen as a surface in the probability simplex. The quadratic homogeneous equation $p(0,0)p(1,1) = p(0,1)p(1,0)$ defines the model as a semi-algebraic surface. It is a ruled surface that can be parametrized on the unit square by

$$[0,1] \ni (\theta_1,\theta_2) \mapsto ((1-\theta_1)^{1-x}\theta_1^x)((1-\theta_2)^{1-y}\theta_2^y) \ .$$

4) The set of all probability functions of the interior of the simplex of *1)* with a given entropy, $\mathcal{H}(p) = -\sum_{x,y} p(x,y)\log p(x,y) = \text{const}$, is a surface of dimension 2.

IG provides the tools for discussing in a geometric language a number of interesting problems about the examples above. For examples, I will define at each point of the open simplex an inner product such that the trajectories that are orthogonal to the surfaces of equal entropy are Gibbs models.

The language of IG is the language of differential geometry. All the IG monograph quoted above contain a short introduction to differential geometry. Non-parametric presentations of differential geometry can be found in Lang [6] and Klingenberg [7]. In these approaches, one the model space (coordinates space) can be any Banach space and different charts of the atlas are not required to have the same image space.

II. CALCULUS ON THE SIMPLEX

Convex analysis is a relevant topic in IG. Standard references are the monographs Rockafellar [8] and Barvinok [9]. Find below a short review of what is needed for IG.

A subset H of a vector space V is an *affine space* if $TH = \{x - y \,|\, x,y \in H\}$ is a sub-vector space of V. TH is called the vector subspace parallel to H (or tangent to H).

Our main example is $V = \mathbb{R}^n$ and $H = \{x \in \mathbb{R}^n \,|\, \mathbf{1}^t x = 1\}$. If $x,y \in H$, then $\mathbf{1}^t(x-y) = 0$.

* giovanni.pistone@carloalberto.org; https://www.giannidiorestino.it/; Lectures notes after the 6th Ph.D. School/Conference on Mathematical Modeling of Complex Systems Università "G. dAnnunzio", Pescara (IT), July 3–11, 2019. GP is supported by de Castro Statistics and Collegio Carlo Alberto and is a member of INdAM–GNAMPA

Conversely, if $\mathbf{1}^t z = 0$, then $z = (z + \boldsymbol{e}_1) - \boldsymbol{e}_1$, hence $TH = \{z \in \mathbb{R}^n \mid \mathbf{1}^t z = 0\}$.

The dimension of the affine space H is the dimension of its parallel vector subspace. Given $x_0, \dots, x_n \in V$, the set of all vectors of the form $x_0 + \sum_{j=1}^n \lambda_j x_j$, $\lambda_j \in \mathbb{R}$, is the affine space generated by the given vectors. An affine space of dimension $(n-1)$ in $\mathbb{R}^n$ is an *hyper-plane*.

A subset C of the vector space V is *convex* if for all $x, y \in C$ the segment $(1-\lambda)x + \lambda y$, $\lambda \in [0,1]$ is in C. The intersection of two convex sets is clearly convex. Given $x_0, \dots, x_n \in V$ the set of all $\lambda_0 x_0 + \cdots + \lambda_n x_n$ with $\lambda_0 + \cdots + \lambda_n = 1$ is the convex set generated by the given vectors. Such a set is called a *polytope* (or convex polytope). Notice that $\sum_{j=0}^n \lambda_j x_j = (1 - \sum_{j=1}^n \lambda_j) x_0 + \sum_{j=1}^n \lambda_j x_j = x_0 + \sum_{j=1}^n \lambda_j (x_j - x_0)$ that is, the polytope is a part of the affine space generated.

A notable example of convex set is the *half-space* of $v \in V$ such that $\langle c, v \rangle \leq b$ with $c \in V$ and $b \in \mathbb{R}$. A finite intersection of half-spaces is a convex set called a *polyhedron*.

The vectors $x_0, \dots, x_m$ are *affinely independent* if the vectors $x_1 - x_0, \dots, x_m - x_0$ are linearly independent. They form a vector basis of the sub-space parallel to the generated polytope which in this case is called a *simplex*. Two simplexes of the same dimension can be mapped one onto the other by an affine transformation that map their respective generators (the vertexes).

For example, the probability simplex $\Delta(\{1,2,3\})$ and its graphical representation as an equilateral triangle are well known in statistics.

Example Let us define more formally the example already used, the probability simplex on $\{0,1\}^2$. Let V be the vector space of real functions on $\{0,1\}^2$, $\mathbb{R}^{\{0,1\}^2} \simeq \mathbb{R}^4$. The four functions $\delta_{ij} = \delta_i \otimes \delta_j$, $i, j = 0, 1$, are linearly independent, in particular, affinely independent. The convex set generated is the probability simplex $\Delta\left(\{0,1\}^2\right) =$

$$\left\{ \sum_{i,j=0,1} p(i,j)\delta_{i,j} \,\middle|\, p(i,j) \geq 0, \sum_{i,j=0,1} p(i,j) = 1 \right\} .$$

Exercise Any other set of 4 affinely independent vectors can be used to represent the same probability simplex. For example, the 4 vertexes in $\mathbb{R}^3$ of the tetrahedron are affinely independent,

ij	θ	ϕ	x	y	z
00	0	0	$\sin(0)\cos(0)$	$\sin(0)\cos(0)$	$\cos(0)$
01	$\frac{2}{3}\pi$	0	$\sin\left(\frac{2}{3}\pi\right)\cos(0)$	$\sin\left(\frac{2}{3}\pi\right)\cos(0)$	$\cos\left(\frac{2}{3}\pi\right)$
10	$\frac{2}{3}\pi$	$\frac{2}{3}\pi$	$\sin\left(\frac{2}{3}\pi\right)\cos\left(\frac{2}{3}\pi\right)$	$\sin\left(\frac{2}{3}\pi\right)\cos\left(\frac{2}{3}\pi\right)$	$\cos\left(\frac{2}{3}\pi\right)$
11	$\frac{2}{3}\pi$	$\frac{4}{3}\pi$	$\sin\left(\frac{2}{3}\pi\right)\cos\left(\frac{4}{3}\pi\right)$	$\sin\left(\frac{2}{3}\pi\right)\cos\left(\frac{4}{3}\pi\right)$	$\cos\left(\frac{2}{3}\pi\right)$

The possibly most common representation uses the 0 vector together with the vectors of the standard basis, that is the set $\{p(0,1)\delta_{01} + p(1,0)\delta_{10} + p(11)\delta_{11}\}$ with conditions $p(i,j) \geq 0$, $p(0,1) + p(1,0) + p(1,1) \leq 1$.

Let us recall two basic results about convex sets.

Theorem 1. *Let K be a convex set of the finite dimensional vector space V. Assume K is closed, $K = \overline{K}$, and its interior is not empty, $K^\circ \neq \emptyset$. Let x be a point of the boundary, $x \in \partial K = \overline{K} - K^\circ$. The there exist $b \in \mathbb{R}$ and $A \in L(V)$, such that the affine function $h = A + b$* supports *K at x, namely $h(x) = 0$ and $h(y) \geq 0$ if $y \in K$.*

Proof. [9, §II.1-2] □

Theorem 2. *Every polytope is a polyhedron and every bounded polyhedron is a polytope.*

Proof. [9, §II.3] □

At this point, I recap the basic notations of the affine geometry of the probability simplex. Let λ be a probability function on Ω. As $\lambda \in \mathbb{R}^\Omega$, one can write $\lambda = \sum_{x \in \Omega} \lambda(x)\delta_x$, so that the set $\Delta(\Omega)$ is the convex set generated by the probability functions associated to the Dirac probability measures. Let us code Ω as $\{1, \dots, N\}$ and write $\lambda = \sum_{j=1}^n \lambda_j e_j$. The vectors $e_j - e_m$, $j = 1, \dots, N-1$ are linearly independent so that $\Delta(\Omega)$ is a special simplex which is called the *probability simplex*. The parallel vector space is the vector space of the vectors of the form $\sum_{j=1}^n \alpha_j (e_j - e_1)$ that is of the form $\sum_{j=1}^n \alpha_j e_j$ with $\sum_{j=1}^n \alpha_j = 0$. These are the vectors which are orthogonal to the constant vectors.

The set of probability functions with support $\Omega_1 \subset \Omega$ form a simplex of dimension $\#\Omega_1 - 1$. If $\#\Omega_1 = n - 1$ this sub-simplex is a *face* of $\Delta(\Omega)$.

There is another simplex that represents the probability simplex $\Delta(\Omega)$ namely, the *solid probability simplex*. In fact, one can represent a probability function by its $n-1$ values $\lambda_j, \dots, \lambda_{n-1}$ which form a vector in $\mathbb{R}^{n-1}$ satisfying the conditions $\lambda_j \geq 0$ and $\sum_{j=1}^{n-1} \lambda_j \leq 1$. The vectors $e_1, \dots, e_{n-1}, 0 \in \mathbb{R}^{n-1}$ are affinely independent and generate a simplex of dimension $n-1$ as $\sum_{J=1}^{n-1} \lambda_j e_j + \lambda_n 0$. The mapping between the two representations is given by $\mathbb{R}^n \ni e_j \mapsto e_j \in \mathbb{R}^{n-1}$ for $j = 1, \dots, n-1$ and $\mathbb{R}^n \ni e_n \mapsto 0 \in \mathbb{R}^{n-1}$.

Let us turn to the calculus on the simplex. Let $f \colon \mathcal{O} \to \mathbb{R}^n$, where $\mathcal{O}$ is an open sub-set of $\mathbb{R}^m$. The function is differentiable at $\bar{x} \in \mathcal{O}$ if there exists a linear mapping $df(\bar{x}) \in L(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$f(\bar{x} + h) - f(\bar{x}) - df(\bar{x})[h] = \mathrm{o}(h) .$$

The matrix representing the linear operator $df(\bar{x})$ is called the Jacobian matrix of f, $Jf(\bar{x})$, whose elements are the partial derivatives

$$Jf(\bar{x}) = \left[\tfrac{\partial}{\partial x_j} f_i(x_1, \dots, x_n) \right]_{i=1,\dots,n;j=1,\dots m} .$$

The derivative of the composite function $f \circ g$ at x is $df \circ g(x) = df(g(x)) \circ dg(x)$.

Example Here is a fundamental remark. *Let $I \ni \theta \mapsto \lambda(\theta)$ be a curve which stays in the probability simplex $\Delta(\Omega)$ and which is differentiable in $\mathbb{R}^\Omega$.* The derivative

$$\lambda'(\theta) = \lim_{h\to 0} h^{-1}(\lambda(\theta+h) - \lambda(\theta)$$

belongs to the subspace parallel to the simplex. If $\lambda(\bar\omega;\bar\theta) = 0$, then the real differentiable function $\theta \mapsto \lambda(\bar\omega,\theta)$ has a minimum at $\theta = \bar\theta$, so that $\lambda'(\bar\omega,\bar\theta) = 0$ and $\lambda'(\bar\theta)$ belong to the space parallel to the face of the simplex characterised by $\lambda(\bar\omega) = 0$. In the language of measure theory, *$\lambda'(t)$ is absolutely continuous with respect to $\lambda(t)$, that is, there exists a curve $t \mapsto s(t)$ such that $\lambda'(x;t) = s(x;t)\lambda(x;t)$ for all x and t.* Notice that, if $\lambda(x;t)$ stays positive in some time interval, then one can take $s(x;t) = \frac{d}{dt}\log\lambda(x;t)$ on that interval.

Exercise. The entropy $\mathcal{H}(\lambda) = -\sum_\omega \lambda(\omega)\log\lambda(\omega)$ is defined on the convex set $\Delta^\circ(\Omega)$, $\#\Omega = N$, of strictly positive probability functions. As $\phi(x) = -x\log x$, $x > 0$, is concave,

$$\frac{1}{N}\mathcal{H}(\lambda)) = \frac{1}{N}\sum_{\omega\in\Omega}\phi(\lambda(\omega)) \le \phi\left(\frac{1}{N}\sum_{\omega\in\Omega}\right) = \phi\left(\frac{1}{N}\right)\ ,$$

and the uniform probability function is a maximum of the entropy. Let us show that this maximum is unique. Assume there is a $\bar\lambda$ which is a maximum for the entropy and let $\theta \mapsto \lambda(\theta)$ be a differentiable curve in $\Delta^0(\Omega)$ such that $\lambda(0) = \bar\lambda$. Let us compute the derivative

$$\frac{d}{d\theta}H(\lambda(\theta))\Big|_{\lambda=0} = -\sum_{\omega\in\Omega}(\log\lambda(\omega;\theta)+1)\lambda'(\omega;\theta)\Big|_{\theta=0} = -\sum_{\omega\in\Omega}(\log\bar\lambda(\omega)+1)\lambda'(\omega;0) = 0\ .$$

As $\bar\lambda$ is in the $\Delta^0(\Omega)$, for each v in the space parallel to the simplex one can consider the curve $\theta \mapsto \bar\lambda + \theta v$ whose derivative at $\theta = 0$ is v. It follows that for each v one has

$$\sum_{\omega\in\Omega}(\log\bar\lambda(\omega)+1)v(\omega) = 0$$

hence, $\log\bar\lambda$ is constant that is, $\bar\lambda$ is constant $\bar\lambda(\omega) = 1/N$.

Let us move now to the discussion of convex functions. If a convex set $A \in \mathbb{R}^m$ is open, then every straight line intersects C in an open interval or an empty interval. For example, the subset of the solid probability simplex consisting of strictly positive probability functions is an open convex set. The closure $\overline{A}$ of an open convex set A is a convex set. The difference $\overline{A} \setminus A$ is the boundary of the convex set. Let x be a point of the boundary. A unit vector u applied at x enters A if there is a $y \in A$ such that $u = (y-x)/\|y-x\|$. The set of all entering vectors cannot contain two antipodal elements so that there is a unit vector w such that $\langle w,u\rangle < 0$ for all entering unit vector.

Theorem 3 (Isolation Theorem)**.** *Let A be an open convex set in $\mathbb{R}^m$ and let x be in the border of A. There exists a unit vector w such that $\langle w, y-x\rangle < 0$ for all $y \in A$ that is, the half-space contains the convex set*

Proof. See a full proof in Barvinok [9, p 45-46]. □

A function ϕ defined on $\mathbb{R}^n$ with values in $\overline{\mathbb{R}} = \mathbb{R}\cup\{+\infty\}$ is convex if the *epigraph* $\mathrm{epi}\,(\phi) = \{(x,t)\,|\,x\in\mathrm{dom}\,(\phi), t\in\mathbb{R}, \phi(x)\le t\}$ is a convex subset of $\mathbb{R}^{n+1}$. Define $\mathrm{dom}\,(\phi)$ to be the set where ϕ takes finite values. *If ϕ is convex, then $\mathrm{dom}\,(\phi)$ is a convex subset of $\mathbb{R}^n$.* If $x_1, x_2 \in \mathrm{dom}\,(\phi)$, then there exist $(x_1,t_1),(x_2,t_2) \in \mathrm{epi}\,(\phi)$ and for all $\lambda\in[0,1]$ it holds $((1-\lambda)x_1+\lambda x_2, (1-\lambda)t_1+\lambda t_2) \in \mathrm{epi}\,(\phi)$. In particular, $\phi((1-\lambda)x_1+\lambda x_2) < +\infty$. *If ϕ is convex, then $(1-\lambda)\phi(x_1)+\lambda\phi(x_2) \le \phi((1-\lambda)x_1+\lambda x_2)$ for all $x_1, x_2 \in \mathbb{R}^n$ and $\lambda\in[0,1]$.* If any of x_1, x_2 in nor in $\mathrm{dom}\,(\phi)$ the inequality is trivially satisfied. Otherwise, it is the same computation as above. Conversely, if $\phi\colon \mathrm{dom}\,(\phi) \to \mathbb{R}$ and $(1-\lambda)\phi(x_1)+\lambda\phi(x_2) \le \phi((1-\lambda)x_1+\lambda x_2)$ for all $x_1, x_2 \in \mathrm{dom}\,(\phi)$ and $\lambda\in[0,1]$, then the function extended with value $+\infty$ outside the domain is convex.

Let ϕ be convex, and define the strict epigraph be open convex set $\{(x,t)\,|\,x\in\mathrm{dom}\,(\phi), t\in\mathbb{R}, \phi(x) < t\}$. Assume that at a point $(x,\phi(x))$ the entering unit vectors are not all horizontal. Then the Isolation Theorem implies that there exist at least a *supporting hyper-plane.* In such a case, ϕ on all such points *ϕ is the point-wise maximum of the supporting affine functions.* In the differentiable case, the tangent plane is the unique supporting hyperplane. If $\phi \in C^2(\mathcal{O})$ then the Hessian matrix is non-negative definite.

The following result is important in the theory of exponential families. *Let ϕ be convex and let ϕ be differentiable on an open $\mathcal{O}$. Then $\nabla\phi\colon \mathcal{O}\to\mathbb{R}^n$ is* monotone *i.e., $\langle\nabla\phi(x) - \nabla\phi(y), x-y\rangle \ge 0$ for $x, y \in \mathcal{O}$.* The basic inequality can be rewritten as

$$\lambda^{-1}(\phi(x+\lambda(y-x)) - \phi(x)) \le \phi(y) - \phi(x)\ .$$

If $\lambda\to 0$, then $\langle\nabla\phi(x), y-x\rangle \le \phi(y)-\phi(x)$. By adding the same inequality with x and y exchanged, one obtains the monotonicity.

Conversely, *if ϕ is differentiable and monotone on an open set $\mathcal{O}$, then ϕ is convex on $\mathcal{O}$.* Write $z = (1-\lambda)x + \lambda y$ and assume $0 < \lambda < 1$ because otherwise there is nothing to prove. Observe that

$$\phi(z)-\phi(x)=\int_0^1\langle\nabla\phi(x+t(z-x)),z-x\rangle\ dt=$$
$$\int_0^1\langle\nabla\phi(x+t(z-x))-\nabla\phi(z),z-x\rangle\ dt+\langle\nabla\phi(z),z-x\rangle\leq\langle\nabla\phi(z),z-x\rangle=\lambda\langle\nabla\phi(z),y-x\rangle\ .$$

In fact, $z-x$ and $(x+t(x-z))-z$ are proportional with factor $-(1-t)\leq 0$. In a similar way,

$$\phi(y)-\phi(z)=\int_0^1\langle\nabla\phi(z+t(y-z)),y-z\rangle\ dt=$$
$$\int_0^1\langle\nabla\phi(z+t(y-z))-\nabla\phi(z),y-z\rangle\ dt+\langle\nabla\phi(z),y-z\rangle\geq\langle\nabla\phi(z),y-z\rangle=(1-\lambda)\langle\nabla\phi(z),y-x\rangle\ ,$$

as $y-z$ and $(z+t(y-z))-z$ are proportional with a factor $t\geq 0$. Rearrange the two inequalities as

$$\phi((1-\lambda)x+\lambda y)\leq\phi(x)+\lambda\langle\nabla\phi(z),y-x\rangle\ \text{and}\phi((1-\lambda)x+\lambda y)\leq\phi(y)+(1-\lambda)\langle\nabla\phi(z),y-x\rangle$$

and take the convex combination to conclude the proof. This proof is taken from Rockafellar [8, p. 26].

III. THE OPEN SIMPLEX

Let Ω be given a finite set with $N=\#\Omega$ points, the *sample space*. Denote by $\Delta(\Omega)$ the set of the *probability functions* $p\colon\Omega\to\mathbb{R}_{\geq 0}$, $\sum_{x\in\Omega}p(x)=1$. It is a $(N-1)$-simplex of $\mathbb{R}^\Omega$ that is, an $(N-1)$-dimensional polytope which is the convex hull of its N vertexes δ_x, $x\in\Omega$. It is a closed and convex subset of the affine space $\mathrm{Affine}(\Omega)=\left\{q\in\mathbb{R}^\Omega\,\middle|\,\sum_{x\in\Omega}q(x)=1\right\}$, the space of *signed probability functions*. It has a non empty relative topological interior $\Delta^\circ(\Omega)$, which is the set of the strictly positive probability functions,

$$\Delta^\circ(\Omega)=\left\{p\in\mathbb{R}^\Omega\,\middle|\,\sum_{x\in\Omega}p(x)=1,p(x)>0\right\}.$$

The border of the simplex $\Delta(\Omega)$ is the union of all its faces as a convex set. Recall that a face of maximal dimension $(n-1)$ is called *facet*. Each face is itself a simplex. An *edge* is a face of dimension 1. The focus will be now on the geometry of the open simplex $\Delta^\circ(\Omega)$.

Recall that our aim here is to provide a presentation of Information Geometry in the sense of the monographs by Amari and Nagaoka [1], Amari [2], and Ay, Jost, Lê, and Schwachhöfer [3]. Our presentation below does not use explicitly any specific parameterization of the sets $\Delta^\circ(\Omega)$, $\Delta(\Omega)$, $\mathrm{Affine}(\Omega)$, whose topological and geometrical structure is inherited from $\mathbb{R}^\Omega$. The basic arguments have a "kinetic" flavour, in contrast with the more frequently used "metric" approach. I.e., I consider curves $t\mapsto p(t)\in\Delta(\Omega)$ and look for a proper definition of velocity and acceleration.

The actual extension of this theory to non finite sample space requires a careful handling as most of the topological features of the finite case do not hold in the infinite case.

One possibility is given by the so called *exponential manifold*, which were first introduced in Pistone and Sempi [5], and which are Banach manifolds modeled on Orlicz spaces, see the review paper Pistone [10]. A different, more inclusive, option has been developed in the monograph by Ay, Jost, Lê, and Schwachhöfer [11]. They use as basic topological framework the Banach space of finite signed measures with the total variation norm. The two approaches coincide when the state space is finite.

Another option would be to consider differentiable densities as the image of a geometric measure under the action of a diffeomorphism and push-forward the geometry od the group of diffeomorphism to the densities. This approach is quite interesting because the distributions are identifies with their simulation. A further possibility is the use of Kantorovich and Wasserstein geometries. Montrucchio and Pistone [12] discusses the finite sample space case. There is an important literature about the general case, see in particular, the monograph by Ambrosio, Gigli, and Savaré [13].

III.1. The Fisher-Rao square root embedding

In 1945 C.R. Rao suggested the following construction of a Riemannian geometry on the open probability simplex $\Delta^\circ(\Omega)$. Nowadays, it is more commonly known under the joint name of Fisher-Rao.

Let us consider the strictly positive orthant of a sphere of radius 2 in $\mathbb{R}^\Omega$,

$$S_>=\left\{a\in\mathbb{R}^\Omega\,\middle|\,\|a\|=2,a(x)>0\right\}\ .$$

One has the 1-to-1 mapping of $S_>0$ to the open simplex

$$\sigma\colon a \mapsto \frac{1}{4}a^2 = \frac{1}{4}(a^2(x)\colon x \in \Omega) \ .$$

This mapping is a smooth mapping from the sub-manifold $S_>$ to the sub-manifold $\Delta^\circ(\Omega)$. In fact, the tangent at a is expressed as $T_a S_> = \left\{u \in \mathbb{R}^\Omega \,\middle|\, \langle u, a\rangle = 0\right\}$ and the tangent space at $p = \sigma(a)$ is expressed as $T_p\Delta^\circ(\Omega) = \left\{U \in \mathbb{R}^\Omega \,\middle|\, \sum_{x\in\Omega} U(x) = 0\right\}$. In such charts, the tangent application of σ at a is the ordinary differential, $d\sigma(a)[u] = \frac{1}{2}au$.

The same construction is frequently presented in the literature starting from the so-called embedding $\sigma^{-1}\colon p \mapsto 2\sqrt{p} = a$.

Now, I want to identify the push-forward g^{FR} with σ of the Riemannian metric defined by $g(u,v) = \langle u, v\rangle$ on $S_>$. For that, I require $g^{\mathrm{FR}}_p(U,V) = g_a(u,v)$ if $p = \sigma(a)$, $U = d\sigma(a)[u]$, $V = d\sigma(a)[v]$, that is,

$$g^{\mathrm{FR}}_p(U,V) = \sum_{x,y\in\Omega} \frac{U(x)V(x)}{p(x)} \ .$$

Here, I follow another approach, that leads to the same construction expressed in a different tangent bundle.

III.2. Statistical bundle

The main feature of this presentation of IG consists in the joint geometrical structure given to the probability simplex, that is the set of probability functions, together with the set of integrable functions. Precisely, the set of all couples (p, f) where p is a probability function and f is a random variable is the domain of the mapping $(p, f) \mapsto \sum_x f(x)p(x)$. The two, taken together, form a *vector bundle*. In the finite state case the bundle is trivial because all random variables f are p-integrable. This is not the case when the sample space is infinite.

This concept variously appears in the literature of IG with the name of Hilbert bundle. Cf. Amari [14], Lauritzen [15], Murray and Rice [16], Kass and Vos [17], Gibilisco and Pistone [18], Amari and Nagaoka [1], Lê [19].

More precisely, let us associate to each probability function $p \in \Delta(\Omega)$ a sub-space of the vector space of real random variables $L(p)$. In our finite setting, $L(p)$ is identified with the vector space $\mathbb{R}^\Omega$ if the support is full $\mathrm{Supp}(p) = \Omega$; otherwise, with $\Omega_0 = \mathrm{Supp}(p) \subset \Omega$, $L(p)$ is identified with $\mathbb{R}^{\Omega_0}$.

Definition 1.

1. For each $p \in \Delta^\circ(\Omega)$, $L^2(p)$ is the vector space of real functions of Ω endowed with the inner product $\langle U, V\rangle_p = \mathbb{E}_p[UV]$. It holds $L^2(p) = \mathbb{R} \oplus L^2_0(p)$.
2. The *statistical bundle* with base $\Delta^\circ(\Omega)$ is

$$S\Delta^\circ(\Omega) = \left\{(p, U) \,\middle|\, p \in \Delta^\circ(\Omega), U \in L^2_0(p)\right\} \ .$$

Remark. Notice that $S\Delta^\circ(\Omega)$ is a semi-algebraic subset of the polynomial ring

$$\mathbb{R}[p(x), U(x)\colon x \in \Omega] \ .$$

The geometry of statistical models on a finite sample space can be studied with the tool of real algebraic geometry i.e., as *Algebraic Statistics*. Cf. the monographs Pistone, Riccomagno, and Wynn [20], Pachter and Sturmfels [21], Drton, Sturmfels, and Sullivant [22], Watanabe [23], Aoki, Hara, and Takemura [24], Zwiernik [25], Sullivan [26]. The interplay between algebraic geometry and differential geometry is discussed in the conference proceedings Gibilisco *et al.* [27].

The geometry of the statistical bundle $S\Delta(\Omega)$ propts for a peculiar form of velocity vectors which are defined in terms of statistical *scores*, a name introduced by R. Fisher.

Let $t \mapsto p(t) \in \Delta(\Omega)$ be a curve which is differentiable as a curve in $\mathrm{Affine}(\Omega)$. Observe that $\langle \mathbf{1}, \dot p(t)\rangle = 0$ and call $t \mapsto (p(t), \dot p(t))$ the *velocity curve* which takes values in the trivial bundle $\Delta(\Omega) \times A_0(\Omega)$, $A_0(\Omega) = \{v \,|\, \mathbf{1}^t \cdot v = 0\}$.

The following is the finite state space version of a result in Ay *et al.* [11].

Proposition 1. *At each t the support of $\dot p(t)$ is contained in the support of $p(t)$, so that there exists a curve $t \mapsto (p(t), Sp(t))$ in $S\Delta(\Omega)$ such that $\dot p(t) = Sp(t) \cdot p(t)$. The expected value of $Sp(t)$ with respect to $p(t)$ is zero.*

Proof. For each t and $x \in \Omega$ the condition $p(x;t) = 0$ implies that t is a minimum, hence $\dot p(x;t) = 0$. It follows for all t that $\dot p(t) = Sp(t) \cdot p(t)$ where $Sp(t)$ is defined by

$$Sp(x;t) = \begin{cases} 0 & \text{if } p(x;t) = 0, \\ \frac{\dot p(x;t)}{p(x;t)} = \frac{d}{dt}\log p(x;t) & \text{if } p(x;t) > 0. \end{cases} \tag{1}$$

The expected value of $Sp(t)$ at $p(t)$ is $\sum_x Sp(x;t)\ p(x;t) = \sum_x \dot p(x;t) = 0$. □

Definition 2. The (differential) *score* of the differentiable curve $t \mapsto p(t) \in \Delta(\Omega)$ is the curve in the statistical bundle $t \mapsto (p(t), Sp(t)) \in S\Delta(\Omega)$.

I first discuss the statistical geometry on the open simplex by deriving it from a *vector bundle* with base $\Delta^\circ(\Omega)$. Later I will show that such a bundle can be identified with the tangent bundle of proper manifold structure. It is nevertheless interesting to observe that a number of geometrical properties do not require the actual definition of the statistical manifold, possibly opening the way to a new type of generalization outside the basic finite state space case.

Comment. For each $p \in \Delta^\circ(\Omega)$ consider the plane through the origin, orthogonal to the vector $\overrightarrow{Op}$. The set of positive probabilities each one associated to its plane forms a vector bundle which is the basic structure of our

presentation of Information Geometry. Note that, because of our orientation to Statistics, we call each element of $\mathbb{R}^\Omega = L(\Omega)$ a *random variable*.

In geometry, a mapping F defined on the probabilities $p \in \Delta(\Omega)$ to the bundle, compatible with the bundle structure, that is

$$F\colon p \mapsto (p, F(p)) \in \Delta(\Omega) \times \cup_{p\in\Delta(\Omega)} S_p\Delta(\Omega)) \ ,$$

such that $F(p) \in S_p\Delta(\Omega)$—that is, $F(p)$ is a random variable and $\mathbb{E}_p[F(p)] = 0$,— is called a *section* of the vector bundle. In Statistics, such a mapping is called an *estimating function* as the equation $F(\hat{p}, x) = 0$, $x \in \Omega$, provides an *estimator*, that is a distinguished mapping from the sample space Ω to the simplex of probabilities $\Delta(\Omega)$.

Comment. The previous definition is suggested by the classical set-up of statistics, as it is revealed by the Fisher-Rao computation that leads to the notion of score. However this set-up is too narrow in a number of situation.

1. The probability functions in the application of interest could have zero values at some $x \in \Omega$, that is the set of interest could be the full simplex $\Delta(\Omega)$.

2. There are simple examples where one wants to study a neighborhood of the border of the simplex, namely something in the full affine space $\mathrm{Affine}(\Omega)$. See below the discussion of optimization.

Exercise A formal extended definition is as follows. *A)* For each $\eta \in \mathrm{Affine}(\Omega)$ let B_η be the vector space of random variables U that are μ-centered,

$$B_\eta = \left\{ U\colon \Omega \to \mathbb{R} \,\middle|\, \mathbb{E}_\eta[U] = \sum_{x\in\Omega} U(x)\ \eta(x) = 0 \right\} \ .$$

B) Each B_η is endowed with the bi-linear form

$$\langle U, V\rangle_\eta = \mathbb{E}_\eta[UV] = \sum_{\{x\in\Omega\,|\,\eta(x)\neq 0\}} U(x)V(x)\ \eta(x) \ .$$

C) The *statistical bundle* of the affine space $\mathrm{Affine}(\Omega)$ is the linear bundle on $\mathrm{Affine}(\Omega)$

$$S\,\mathrm{Affine}(\Omega) = \{(\eta, U)\,|\,\eta \in \mathrm{Affine}(\Omega), U \in B_\eta\} \ .$$

D) It is a manifold isomorphic to the open subset of the Grassmanian manifold $\mathrm{Grass}(\mathbb{R}^\Omega, \#\Omega - 1)$ of sub-spaces B that do not contain constant vectors. In fact, each fiber B_η is a subspace of $\mathbb{R}^\Omega$ of co-dimension 1; Viceversa, for each subspace B of dimension $(n-1)$ and not containing the constant, there is a unique complement vector η such that $\sum_{x\in\Omega}\eta_x = 1$.

III.3. Natural gradient

Let us now discuss the notion of gradient in the statistical bundle of the open simplex.

Proposition 2. *Let $I \ni t \mapsto p(t)$ be a C^1 curve in $\Delta^\circ(\Omega)$. For each $f\colon \Omega \to \mathbb{R}$,*

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

where $Sp(t) = \frac{d}{dt}\log(p(t))$

Proof.

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}_{p(t)}[f] &= \frac{d}{dt}\sum_{x\in\Omega} f(x)p(x;t) \\
&= \sum_{x\in\Omega} f(x)\frac{d}{dt}p(x;t) \\
&= \sum_{x\in\Omega} f(x)\frac{d}{dt}\log p(x;t)\ p(x;t) \\
&= \mathbb{E}_{p(t)}[fSp(t)] \quad (\text{using } \mathbb{E}_{p(t)}[Sp(t)] = 0) \\
&= \mathbb{E}_{p(t)}\left[(f - \mathbb{E}_{p(t)}[f])Sp(t)\right] \\
&= \left\langle f - \mathbb{E}_{p(t)}[f], Sp(t)\right\rangle_{p(t)} \ .
\end{aligned}$$

□

Notice that $p \mapsto f - \mathbb{E}_p[f]$ is a *section* of $S\Delta^\circ(\Omega)$ and $t \mapsto Sp(\cdot)$ is a *lift* of $p(\cdot)$.

Example. I have chosen not discuss here the case of the closed simplex. The condition for the existence of the differential score means that the differential score exists if and only if the curve $t \mapsto \eta(t) \in \mathrm{Affine}(\Omega)$ hits the faces of $\Delta(\Omega)$ only *tangentially*. For example: $n = 3$, $p(0;t) = t$, $p(1;t) = \sqrt{\frac{1}{2} - t^2}$, $p(2;t) = 1 - t - \sqrt{\frac{1}{2} - t^2}$.

Definition 3 (Natural gradient)**.**

Given a function $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$, its *natural gradient* is a section

$$\Delta^\circ(\Omega) \ni p \mapsto (p, \mathrm{grad}\, F(p)) \in S\Delta^\circ(\Omega) \ .$$

such that for each regular curve $I \ni t \mapsto p(t)$ it holds

$$\frac{d}{dt}f(p(t)) = \langle \mathrm{grad}\, f(p(t)), Sp(t)\rangle_{p(t)}, \quad t \in I \ .$$

Proposition 3 (Computing grad)**.**

If f is a C^1 function on an open subset of $\mathbb{R}^\Omega$ containing $\Delta^\circ(\Omega)$, by writing $\nabla f(p)\colon \Omega \ni x \mapsto \frac{\partial}{\partial p(x)}f(p)$, the following relation between the statistical gradient and the ordinary gradient holds:

$$\mathrm{grad}\, f(p) = \nabla f(p) - \mathbb{E}_p[\nabla f(p)] \ .$$

Proof.

$$\begin{aligned}\frac{d}{dt}f(p(t)) &= \frac{d}{dt}f(p(x;t)\colon x\in\Omega)\\ &= \sum_{x\in\Omega}\frac{\partial}{\partial p(x)}f(p(x;t)\colon x\in\Omega)\frac{d}{dt}p(x;t)\\ &= \sum_{x\in\Omega}\frac{\partial}{\partial p(x)}f(p(x;t)\colon x\in\Omega)\frac{d}{dt}\log p(x;t)\ p(x;t)\\ &= \langle\nabla f(p(t)), Sp(t)\rangle_{p(t)}\\ &= \left\langle\nabla f(p(t)) - \mathbb{E}_{p(t)}\left[\nabla f(p(t))\right], Sp(t)\right\rangle_{p(t)}\\ &= \langle\operatorname{grad} f(\eta(t)), S\eta(t)\rangle_{p(t)}\ .\end{aligned}$$

□

Example: Natural gradient of the entropy Here is our basic example. The function

$$\mathcal{H}(p) = -\sum_{x\in\Omega}p(x)\log p(x)$$

satisfies the conditions of the proposition with

$$\nabla\mathcal{H}(p) = (x\mapsto -\log p(x) - 1).$$

Moreover,

$$\mathbb{E}_p\left[\nabla\mathcal{H}\right] = \sum_{x\in\Omega}(-\log p(x) - 1)p(x) = \mathcal{H}(p) - 1\ .$$

It follows that

$$\operatorname{grad}\mathcal{H}(p) = -\log p - 1 - \mathcal{H}(p) + 1 = -\log p - \mathcal{H}(p)\ .$$

The condition $\operatorname{grad}\mathcal{H}(q) = 0$ is satisfied by a constant $\log p$.

Remarks. The Information Geometry on the simplex does not coincide with the geometry of the embedding of the simplex $\Delta^\circ(\Omega)\to\mathbb{R}^\Omega$, in the sense the statistical bundle is not the tangent bundle of these embeddingIt will become the tangent bundle of the proper geometric structure which is given by special atlases.

The vector $Sp(t)\in S_{p(t)}\Delta^\circ(\Omega)$ is meant to represent the relative variation of the information in a one dimensional statistical model in the sense it is a relative derivative. Geometrically, the differential score is a representation of the velocity along a curve.

Consider the level surface of $f\colon \operatorname{Affine}(\Omega)\to\mathbb{R}$ at $\eta_0\in\operatorname{Affine}(\Omega)$, that is the surface $\{\eta\in\operatorname{Affine}(\Omega)\mid f(\eta) = f(\eta_0)\}$, and assume η_0 is not a critical point, $\operatorname{grad} f(\eta_0)\neq 0$. Then for each curve through η_0, $I\mapsto\eta(t)$ with $\eta(0) = \eta_0$, such that $f(\eta(t)) = f(\eta(0))$,

$$0 = \frac{d}{dt}f(\eta(t))\Big|_{t=0} = \langle\operatorname{grad} f(p(t)), p(t)\rangle_{\eta(t)}\Big|_{t=0} = \langle\nabla f(\eta_0), S\eta(t_0)\rangle_{\eta_0}\ ,$$

that is, *all velocities $Sp(t_0)$ tangential to the level set are orthogonal to the statistical gradient.* Note that I have not jet defined a manifold such that the statistical bundle is equal to the tangent bundle.

If the function $f\colon \operatorname{Affine}(\Omega)\to\mathbb{R}$ extends to a C^1 function on an open subset of $\mathbb{R}^\Omega$, then one can compute the statistical gradient via the ordinary gradient in the geometry of $\mathbb{R}^\Omega$, namely $\nabla f(\eta)\colon\Omega\ni x\mapsto\frac{\partial}{\partial\eta(x)}f(\eta)$. Note that the statistical gradient is zero if, and only if, the ordinary gradient is constant.

III.4. Flows

As already said, our emphasis is on a kinematic approach to IG. Let us start to consider differential equations.

Definition 4 (Flow)**.**

1. Given a section $F\colon\Delta^\circ(\Omega)$ the *trajectories along the section* are the solution of the (statistical) *differential equation*
$$Sp(t) = F(p(t))\ .$$
2. If F is defined on a open set of $\mathbb{R}^\Omega$ containg $\Delta^\circ(\Omega)$ with values in $\mathbb{R}^\Omega$, the statistical differential equation is equivalent to the system of ordinary differential equations
$$\frac{d}{dt}p(x;t) = p(x;t)F(x,\boldsymbol{p}(t))\qquad x\in\Omega\ .$$
3. The *gradient flow* is the flow of the section $F = \operatorname{grad} f$.

The right-end-side of differential equation is

$$G(x,\boldsymbol{p}) = p(x)F(x,\boldsymbol{p}(y))\ ,$$

so that $\sum_{x\in\Omega}G(x,\boldsymbol{p}) = 0$. And conversely. This class of differential equations is well studied in the literature under various names, e.g., replicator equation.

Example : Gradient flow of the expected value Given a random variable f, consider the section $F(p) = f - \mathbb{E}_p[f]$. The flow of F is the solution of

$$\dot{p}(x;t) = p(x;t)(f(x) - \sum_{y\in\Omega}f(y)p(y;t))\ .$$

The solution is an exponential family. Consider the 1-dimensional statistical model

$$p(x;t) = \exp\left(tf(x) - \psi(t)\right)p_0(x)\ , \tag{2}$$

with $\psi(t)$ normalising constant.

$$\exp(\psi(t)) = \sum_{x\in\Omega}\exp(tf(x))\ .$$

It is a curve in $\Delta^\circ(\Omega)$ with $p(x;0) = p_0(x)$. The differential score is

$$Sp(t) = \frac{d}{dt}\log p(t) = f(x) - \frac{d}{dt}\psi(t) \ .$$

As $\mathbb{E}_{p(t)}[Sp(t)] = 0$, $\frac{d}{dt}\psi(t) = \mathbb{E}_{p(t)}[f]$ and we have that Equation (2) is the solution of the natural flow starting at p_0.

Notice that the natural gradient of $f(p) = \mathbb{E}_p[f]$ is precisely

$$\operatorname{grad} f(p) = \nabla f(p) - \mathbb{E}_p[\nabla f(p)] = F(p) \ .$$

This is the solution of a gradient flow equation.

Example: Gradient flow of the entropy Consider the equation

$$Sp(t) = \operatorname{grad}\mathcal{H}(p(t)) = -\log p(t) - \mathcal{H}(p(t)) \ ,$$

or

$$\frac{d}{dt}\log p(t) = -\log p(t) - \mathcal{H}(p(t)) \ .$$

By setting $v(x,t) = \log p(x;t)$ the equation becomes

$$\dot v(x;t) = -v(x;t) + \sum_{x\in\Omega} v(x;t)\mathrm{e}^{v(x;t)} \ .$$

Let us look for a solution of the form

$$p(x;t) = \exp\left(a(t)\log p_0(x) - \psi(t)\right) \ .$$

with $a(0) = 1$, hence $p(0) = p_0$ and $\psi(0) = 0$.

In this case,

$$Sp(t) = \dot a(t)\log p_0 - \dot\psi(t) = \dot a(t)(\log p_0 - \mathbb{E}_{p(t)}[\log p_0])$$

and

$$\mathcal{H}(p(t)) = -\mathbb{E}_{p(t)}[a(t)\log p_0 - \psi(t)] = \\ - a(t)\,\mathbb{E}_{p(t)}[\log p_0] + \psi(t) \ .$$

Plugging the previous computations into the equation,

$$\dot a(t)(\log p_0 - \mathbb{E}_{p(t)}[\log p_0]) = \\ -(a(t)\log p_0 - \psi(t)) - (-\mathbb{E}_{p(t)}[\log p_0] + \psi(t))$$

which is satisfied if $\dot a(t) = -a(t)$. As $a(0) = 1$,

$$p(x;t) \propto \exp\left(\mathrm{e}^{-t}\log p_0(x)\right) = p_0(x)^{\mathrm{e}^{-t}} \ .$$

In conclusion, the natural gradient flow of the entropy is an exponential family with parameter $a(t) = \mathrm{e}^{-t}$, sufficient statistics $\log p_0$ and cumulant function ψ.

The same exponential family as before is, in the canonical parameter,

$$p(\theta) = \exp(\theta\log p_0 - \Psi(\theta)) \propto p_0^\theta \ , \quad \theta > 0 \ .$$

The differential score is

$$Sp(\theta) = \log p_0 - \dot\Psi(\theta) = \log p_0 - \mathbb{E}_{p(t)}[\log p_0] \ .$$

Example: KL-divergence Consider the Kulback-Leibler divergences $p \mapsto \mathrm{D}(p\,\|p_0)$ and $p \mapsto \mathrm{D}(p_0\,\|p)$ and compute the respective natural gradient.

In the first case,

$$\frac{\partial}{\partial p(x)}\mathrm{D}(p\,\|p_0) = \\ \frac{\partial}{\partial p(x)}\sum_{y\in\Omega} p(y)\log\frac{p(y)}{p_0(y)} = \log\frac{p(x)}{p_0(x)} + 1 \ ,$$

so that the natural gradient is

$$\operatorname{grad}(p \mapsto \mathrm{D}(p\,\|p_0)) = \left(p \mapsto \log\frac{p}{p_0} - \mathrm{D}(p\,\|p_0)\right) \ .$$

The solution on the gradient flow is similar to the solution for the entropy.

In the second case,

$$\frac{\partial}{\partial p(x)}\mathrm{D}(p_0\,\|p) = \\ \frac{\partial}{\partial p(x)}\sum_{y\in\Omega} p_0(y)\log\frac{p_0(y)}{p(y)} = 1 - \frac{p_0(x)}{p(x)},$$

so that the natural gradient is

$$\operatorname{grad}(p \mapsto \mathrm{D}(p_0\,\|p)) = \left(p \mapsto 1 - \frac{p_0(x)}{p(x)}\right) \ .$$

The gradient flow equation $Sp- = -\operatorname{grad}\mathrm{D}(p_0\,\|p)$ is

$$\frac{\dot p(t)}{p(t)} = \frac{p_0}{p(t)} - 1 \quad \text{that is} \quad \dot p(t) = p_0 - p(t) \ ,$$

and the solution is

$$p(t) = p_0 + (p(0) - p_0)\mathrm{e}^{-t} \ .$$

Comment. It is remarkable that the two variables p and q in $\mathrm{D}(p\,\|q)$ are clearly associated with two different affine geometries on the statistical bundle. In fact, it is possible to derive the structure of IG from the divergence. I do not discuss this approach here and refer to the general monographs for this development.

Proposition 4.

1. *Let $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$ be bounded and let $\mathbb{R}_+ \ni t \mapsto p(t)$ be a solution of the natural gradient flow*

$$Sp(t) = -\operatorname{grad} f(p(t)) \ , \quad t > 0 \ . \tag{3}$$

The value of f along the solution, $t \mapsto f(p(t))$, is decreasing and bounded below by $\min f$.

2. *Moreover, if $t \mapsto \|\operatorname{grad} f(p(t))\|^2_{p(t)} = \|Sp(t)\|^2_{p(t)}$ is uniformly continuous, then $\lim_{t\to\infty}\|Sp(t)\|_{p(t)} = 0$.*

3. *Assume in addition that* $p \mapsto \|\operatorname{grad} f(p)\|_p$ *continuously extend to a the full simplex as a function* $L\colon \Delta(\Omega) \to \mathbb{R}$ *and there exists a level set* $\{p \in \Delta^\circ(\Omega) \,|\, L(p) \le a\}$ *where* L *has an unique zero* $\bar{p} \in \Delta(\Omega)$. *In such a case,* $f(p(0)) \le \alpha$ *implies* $\lim_{t\to\infty} p(t) = \bar{p}$.

Proof. Item 1. From (3) and the definition of grad,

$$\frac{d}{dt} f(p(t)) = \langle \operatorname{grad} f(p(t)), Sp(t) \rangle_{p(t)} = - \|\operatorname{grad} f(p(t))\|^2_{p(t)} = - \|Sp(t)\|^2_{p(t)} \ ,$$

hence

$$f(p(t)) - f(p(0)) = - \int_0^t \|\operatorname{grad} f(p(t))\|^2_{p(t)} \ dt = - \int_0^t \|Sp(t)\|^2_{p(t)} \ dt \ ,$$

so that $t \mapsto f(p(t))$ is decreasing and converging to a limit $\alpha \ge \min f$.

Item 2. It follows from the boundedness below of f that α is finite and moreover

$$\int_0^\infty \|\operatorname{grad} f(p(t))\|^2_{p(t)} \ dt = \int_0^\infty \|Sp(t)\|^2_{p(t)} \ dt = f(p_0) - \alpha \le \max f < \infty \ .$$

If $t \mapsto \|\operatorname{grad} f(p(t))\|^2_{p(t)} = \|Sp(t)\|^2_{p(t)}$ is uniformly continuous, it follows from Barbalat's lemma that $\lim_{t\to\infty} \|\operatorname{grad} f(p(t))\|_{p(t)} = \lim_{t\to\infty} \|Sp(t)\|_{p(t)} = 0$.

Item 3. It holds $\lim_{t\to\infty} \|\operatorname{grad} f(p(t))\|_{p(t)} = \lim_{t\to\infty} L(p(t)) = 0$. Every solution that starts inside $\{f \le a\}$ stays in the level set. If $p \mapsto L(p)$ has a unique isolated zero at $\bar{p}$, then $\lim_{t\to\infty} p(t) = \bar{p}$. □

Example: Expected value. Let $f\colon \Omega \to \mathbb{R}$ have a unique maximum at $\bar{x}$ and relax $F(p) = \mathbb{E}_p[f]$, $p \in \Delta^\circ(\Omega)$. It holds $\operatorname{grad} F(p) = f - \mathbb{E}_p[f]$. The function $F\colon \Delta^\circ(\Omega)$ is bounded and $p \mapsto \|\operatorname{grad} f\|^2_p = \operatorname{Var}_p(f)$ is bounded and continuous on $\Delta(\Omega)$. The trajectory is the exponential family $p(t) = \mathrm{e}^{tf-\psi(t)} p_0$ and $Sp(t) = f - \psi'(t)$ is uniformly continuous because its derivative $\frac{d}{dt} Sp(t) = -\psi''(t) = \operatorname{Var}_{p(t)}(f)$ is bounded.

The natural gradient flow of the expected value has been intensively used as an optimization algorithm, see Malagò, Matteucci, and Pistone [28, 29, 30, 31].

IV. CONNECTIONS

I am now going to discuss now the notion of of differentiable function on the statistical bundle which provides a sort of second order calculus. Cf. Kass and Vos [17].

For each random variable $U \in S_p\Delta^\circ(\Omega) = B_p$, it holds

$$\mathbb{E}_q[U - \mathbb{E}_q[U]] = 0 \quad \text{and} \quad \mathbb{E}_q\left[\frac{p}{q} U\right] = 0 \ ,$$

so that both $U - \mathbb{E}_q[U]$ and $\frac{p}{q}U$ belong to $S_q\Delta^\circ(\Omega) = L^2_0(q)$. This prompts for the following definition.

Definition 5 (e- and m-transport)**.**

1. The *exponential transport*, or *e-transport*, is the family of linear mappings defined for each $p, q \in \Delta^\circ(\Omega)$ by
$$^{\mathrm{e}}\mathbb{U}^q_p\colon S_p\Delta^\circ(\Omega) \ni U \mapsto U - \mathbb{E}_q[U] \in S_q\Delta^\circ(\Omega) \ .$$
2. The *mixture transport*, or *m-transport*, is the family of linear mappings for each $p, q \in \Delta^\circ(\Omega)$ by
$$^{\mathrm{m}}\mathbb{U}^q_p\colon S_p\Delta^\circ(\Omega) \ni U \mapsto \frac{p}{q} U \in S_q\Delta^\circ(\Omega) \ .$$

Let us check now that the e-transport and the m-trasport are semi-groups of affine transformations which are compatible with the statistical bundle and are dual of each other with respect to the scalar product on each fiber.

Theorem 4. *The following properties hold for all* $p, q, r \in \Delta^\circ(\Omega)$.

1. *Exponential semi-group property:* $^{e}\mathbb{U}^r_q \, {}^{e}\mathbb{U}^q_p = {}^{e}\mathbb{U}^r_p$.
2. *Mixture semi-group property:* $^{m}\mathbb{U}^r_q \, {}^{m}\mathbb{U}^q_p = {}^{m}\mathbb{U}^r_p$.
3. *Duality:* $\left\langle {}^{e}\mathbb{U}^q_p U, V \right\rangle_q = \left\langle U, {}^{m}\mathbb{U}^p_q V \right\rangle_p$, $U \in S_p\Delta^\circ(\Omega)$ *and* $V \in S_q\Delta^\circ(\Omega)$.
4. *Conservation of the scalar product:* $\left\langle {}^{e}\mathbb{U}^q_p U, {}^{m}\mathbb{U}^q_p V \right\rangle_q = \langle U, V \rangle_p$, $U, V \in S_p\Delta^\circ(\Omega)$.

Proof. These are simple checks of the definitions. For example, the duality follows from

$$\left\langle {}^{\mathrm{e}}\mathbb{U}^q_p U, V \right\rangle_q = \mathbb{E}_q[(U - \mathbb{E}_q[U])V] = \mathbb{E}_q[UV] - \mathbb{E}_q[U]\,\mathbb{E}_q[V] = \mathbb{E}_q[UV] = \mathbb{E}_p\left[U\left(\frac{q}{p}V\right)\right] = \left\langle U, {}^{\mathrm{m}}\mathbb{U}^p_q V \right\rangle_p \ .$$

The conservation of the scalar product follows from the duality and the semi-group property:

$$\left\langle {}^{\mathrm{e}}\mathbb{U}^q_p U, {}^{\mathrm{m}}\mathbb{U}^q_p V \right\rangle_q = \left\langle {}^{\mathrm{e}}\mathbb{U}^p_q \, {}^{\mathrm{e}}\mathbb{U}^q_p U, V \right\rangle_q = \langle U, V \rangle_p \ .$$

□

Each transport defines a section of the statistical bundle: given $U \in S_q\Delta^\circ(\Omega)$, one has the sections

$$p \mapsto {}^{\mathrm{e}}\mathbb{U}^p_q U \quad \text{and} \quad p \mapsto {}^{\mathrm{m}}\mathbb{U}^p_q U \ ,$$

and can compute their respective flows as follows.

Proposition 5. *Let be given a random variable* $U \in S_q\Delta^\circ(\Omega)$.

1. *The flow of the section $p \mapsto {}^{e}\mathbb{U}_q^p U$, i.e., the solution of*

$$Sp(t) = {}^{e}\mathbb{U}_q^{p(t)} U \ , \quad p(0) = p \ ,$$

is

$$\Delta^\circ(\Omega) \times \mathbb{R} \ni (p,t) \mapsto \mathrm{e}^{t({}^{e}\mathbb{U}_q^p U) - \psi(t)} \cdot p \ ,$$

with $\psi(t) = \log\left(\mathbb{E}_p\left[\mathrm{e}^{{}^{e}\mathbb{U}_q^p U}\right]\right)$.

2. *The flow of the section $p \mapsto {}^{m}\mathbb{U}_q^p U$, $U \in S_q\Delta^\circ(\Omega)$ i.e., the solution of*

$$Sp(t) = {}^{m}\mathbb{U}_q^{p(t)} U, \quad p(0) = p,$$

is

$$\Delta^\circ(\Omega) \times I \ni (p,t) \mapsto (1 + t\,{}^{m}\mathbb{U}_q^p U)p \ ,$$

where $I =]-(\max {}^{m}\mathbb{U}_q^p U)^{-1}, -(\min {}^{m}\mathbb{U}_q^p U)^{-1}[$.

Proof. Item 1. This is a direct check:

$$\frac{d}{dt}(t\,{}^{e}\mathbb{U}_q^p U - \psi(t)) = {}^{e}\mathbb{U}_q^p U - \dot{\psi}(t) =$$
$${}^{e}\mathbb{U}_q^p U - \mathbb{E}_{p(t)}\left[{}^{e}\mathbb{U}_q^p U\right] = {}^{e}\mathbb{U}_p^{p(t)}\,{}^{e}\mathbb{U}_q^p U = {}^{e}\mathbb{U}_q^{p(t)} U \ .$$

Item 2. Assume $U \neq 0$ and let $V(x) = {}^{m}\mathbb{U}_q^p U(x)$. As $\mathbb{E}_p[V] = 0$, it holds both $V(x) < 0$ and $V(x) > 0$. In the first case, $1 + tV(x) > 0$ if $t \leq 0$ or $t > 0$ and $t < -(\min V)^{-1} \leq -V(x)^{-1}$. Similarly in the other case.

If $t \in I$, then $p(t) = (1+tV)p \in \Delta^\circ(\Omega)$ and

$$\frac{d}{dt}\log\left((1 + t\,{}^{m}\mathbb{U}_q^p)p\right) = \frac{{}^{m}\mathbb{U}_q^p U}{1 + t\,{}^{m}\mathbb{U}_q^p U} =$$
$$\frac{p}{p(t)}\,{}^{m}\mathbb{U}_q^p U = {}^{m}\mathbb{U}_p^{p(t)}\,{}^{m}\mathbb{U}_q^p U = {}^{m}\mathbb{U}_q^{p(t)} U \ .$$

□

The proposition, justifies the names given to the transports.

Other transports are of interest. In particular, look for an *isometry*

$${}^{0}\mathbb{U}_p^q \colon S_p\Delta^\circ(\Omega) \to S_q\Delta^\circ(\Omega) \ ,$$

so that $\left\langle {}^{0}\mathbb{U}_p^q U, {}^{0}\mathbb{U}_p^q V\right\rangle_q = \langle U, V\rangle_p$. Compare with item 4 of theorem 4. The remaining part of this section is essentially a long exercise.

Note that $\left\|\sqrt{\frac{p}{q}}U\right\|_q^2 = \|U\|_p^2$ for $U \in S_p\Delta^\circ(\Omega)$, but $\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] = \mathbb{E}_p\left[\sqrt{pq}U\right] = \mathrm{Cov}_p\left(\sqrt{pq}, U\right)$ would not be zero in general. Hence, there is a linear mapping of the form

$$S_p\Delta^\circ(\Omega) \ni U \mapsto \sqrt{\frac{p}{q}}U + A\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \ .$$

The expected value at q is

$$\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U + A\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right]\right] = (1 + \mathbb{E}_q[A])\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \ ,$$

which is zero if $\mathbb{E}_q[A] = -1$. Under this condition, it holds $\sqrt{\frac{p}{q}}U + A\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \in S_q\Delta^\circ(\Omega)$.

Let us compute the squared norm:

$$\left\|\sqrt{\frac{p}{q}}U + A\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right]\right\|_q^2 = \|U\|_p^2 + 2\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}UA\right]\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] + \mathbb{E}_q\left[A^2\right]\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right]^2 =$$
$$\|U\|_p^2 + \left(2\,\mathbb{E}_q\left[\sqrt{\frac{p}{q}}UA\right] + \mathbb{E}_q\left[A^2\right]\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right]\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] = \|U\|_p^2 + \mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\left(2A + \mathbb{E}_q\left[A^2\right]\right)\right]\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \ .$$

Taking $A = -(1 + \sqrt{\frac{p}{q}})(1 + \mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right])$ one has both $\mathbb{E}_q[A] = -1$ and $\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\left(2A + \mathbb{E}_q\left[A^2\right]\right)\right] =$

$$\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\left(-2\frac{1+\sqrt{\frac{p}{q}}}{1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]} + \frac{\mathbb{E}_q\left[\left(1+\sqrt{\frac{p}{q}}\right)^2\right]}{\left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^2}\right)\right] = \mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\left(-2\frac{1+\sqrt{\frac{p}{q}}}{1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]} + 2\frac{1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]}{\left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^2}\right)\right] =$$
$$\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\left(-2\frac{1+\sqrt{\frac{p}{q}}}{1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]} + 2\frac{1}{1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]}\right)\right] = -2\frac{\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\sqrt{\frac{p}{q}}\right]}{1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]} = 0 \ .$$

The previous computation justifies the following definition and proposition.

Definition 6. The *Hilbert transport*, or *h-transport*, is the family of linear mappings

$$ {}^0\mathbb{U}_p^q \colon S_p\Delta^\circ(\Omega) \ni U \mapsto \sqrt{\frac{p}{q}}U - \left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\sqrt{\frac{p}{q}}\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \in S_q\Delta^\circ(\Omega) \ , $$

for all $p, q \in \Delta^\circ(\Omega)$.

Proposition 6. *The following properties hold for all $p, q \in \Delta^\circ(\Omega)$.*

1. *Inverse:* ${}^0\mathbb{U}_q^p\,{}^0\mathbb{U}_p^q U = U$.
2. *Isometry:* $\left\langle {}^0\mathbb{U}_p^q U, {}^0\mathbb{U}_p^q V\right\rangle_q = \langle U, V\rangle_p$.

Proof. Item 1. It is a long computation. Let $V = {}^0\mathbb{U}_p^q U$, so that

$$ \sqrt{\frac{q}{p}}V = \sqrt{\frac{q}{p}}\left(\sqrt{\frac{p}{q}}U - \left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\sqrt{\frac{p}{q}}\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right]\right) = $$
$$ U - \left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\sqrt{\frac{q}{p}}\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \ , $$

$$ \mathbb{E}_p\left[\sqrt{\frac{q}{p}}V\right] = \mathbb{E}_p\left[U - \left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\sqrt{\frac{q}{p}}\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right]\right] = $$
$$ - \left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\mathbb{E}_p\left[1+\sqrt{\frac{q}{p}}\right]\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] = $$
$$ - \left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\mathbb{E}_p\left[\sqrt{\frac{q}{p}}\right]\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \ , $$

$$ \left(1+\mathbb{E}_p\left[\sqrt{\frac{q}{p}}\right]\right)^{-1}\left(1+\sqrt{\frac{q}{p}}\right)\mathbb{E}_p\left[\sqrt{\frac{q}{p}}V\right] = $$
$$ - \left(1+\mathbb{E}_p\left[\sqrt{\frac{q}{p}}\right]\right)^{-1}\left(1+\sqrt{\frac{q}{p}}\right)\left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\mathbb{E}_p\left[\sqrt{\frac{q}{p}}\right]\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] = $$
$$ - \left(1+\sqrt{\frac{q}{p}}\right)\left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\mathbb{E}_q\left[\sqrt{\frac{p}{q}}U\right] \ , $$

and the required equality follows. *Item 2.* The conservation of the norm has been already proved above. The conservation of the scalar product follows from that and from the linearity. □

Let us consider now the flow induced by the h-transport.

Proposition 7. *Given $p \in \Delta^\circ(\Omega)$ and $U \in S_p\Delta^\circ(\Omega)$ with $\mathbb{E}_p\left[U^2\right] = 1$, consider the mapping on an open interval I containing 0 defined by $I \ni t \mapsto \left(\cos\left(\frac{t}{2}\right) + \sin\left(\frac{t}{2}\right)U\right)^2 \cdot p$, with moreover $\cos\left(\frac{t}{2}\right) + \sin\left(\frac{t}{2}\right)U > 0$ for all $t \in I$. Such a mapping is a regular curve in $\Delta^\circ(\Omega)$ such that $p(0) = p$ and $Sp(t) = {}^0\mathbb{U}_p^{p(t)}U$.*

Proof. First, check that $p(t) \in \Delta^\circ(\Omega)$ for all $t \in I$:

$$ \mathbb{E}_p\left[\left(\cos\left(\frac{t}{2}\right) + \sin\left(\frac{t}{2}\right)U\right)^2\right] = \cos\left(\frac{t}{2}\right) + 2\cos\left(\frac{t}{2}\right)\sin\left(\frac{t}{2}\right)\mathbb{E}_p[U] + \sin\left(\frac{t}{2}\right)\mathbb{E}_p\left[U^2\right] = 1 \ . $$

Second, compute the differential score:

$$ Sp(t) = \frac{d}{dt}\left(2\log\left(\cos\left(\frac{t}{2}\right) + \sin\left(\frac{t}{2}\right)U\right) + \log p\right) = 2\frac{-\frac{1}{2}\sin\left(\frac{t}{2}\right) + \frac{1}{2}\cos\left(\frac{t}{2}\right)U}{\cos\left(\frac{t}{2}\right) + \sin\left(\frac{t}{2}\right)U} = \frac{-\sin\left(\frac{t}{2}\right) + \cos\left(\frac{t}{2}\right)U}{\cos\left(\frac{t}{2}\right) + \sin\left(\frac{t}{2}\right)U} \ . $$

Third, compute ${}^{0}\mathbb{U}_{p}^{p(t)}U$ in steps: $\sqrt{\frac{p}{p(t)}}=\frac{1}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}$;

$$\mathbb{E}_{p(t)}\left[\sqrt{\frac{p}{p(t)}}\right]=\mathbb{E}_{p}\left[\sqrt{\frac{p(t)}{p}}\right]=\mathbb{E}_{p}\left[\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U\right]=\cos\left(\frac{t}{2}\right)\ ;$$

$$\frac{1+\sqrt{\frac{p}{p(t)}}}{1+\mathbb{E}_{p(t)}\left[\sqrt{\frac{p}{p(t)}}\right]}=\left(1+\frac{1}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}\right)\frac{1}{1+\cos\left(\frac{t}{2}\right)}=\frac{1+\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}\frac{1}{1+\cos\left(\frac{t}{2}\right)}\ ;$$

$$\sqrt{\frac{p}{p(t)}}U=\frac{U}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}\ ;$$

$$\mathbb{E}_{p(t)}\left[\sqrt{\frac{p}{p(t)}}U\right]=\mathbb{E}_{p}\left[\sqrt{\frac{p(t)}{p}}U\right]=\mathbb{E}_{p}\left[\left(\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U\right)U\right]=\sin\left(\frac{t}{2}\right)\ ;$$

$${}^{0}\mathbb{U}_{p}^{p(t)}U=\frac{U}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}-\frac{1+\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}\frac{\sin\left(\frac{t}{2}\right)}{1+\cos\left(\frac{t}{2}\right)}=$$
$$\frac{\left(1+\cos\left(\frac{t}{2}\right)\right)U-\left(\sin\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U\right)}{\left(\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U\right)\left(1+\cos\left(\frac{t}{2}\right)\right)}=\frac{\left(1+\cos\left(\frac{t}{2}\right)-\sin\left(\frac{t}{2}\right)\right)U-\left(\sin\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)\cos\left(\frac{t}{2}\right)\right)}{\left(\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U\right)\left(1+\cos\left(\frac{t}{2}\right)\right)}=$$
$$\frac{\cos\left(\frac{t}{2}\right)U-\sin\left(\frac{t}{2}\right)}{\cos\left(\frac{t}{2}\right)+\sin\left(\frac{t}{2}\right)U}\ .$$

This concludes the proof. □

V. ACCELERATIONS

Second order geometry is usually derived from a notion of covariant derivative (connection), see e.g., Lang [6]. From the given connection one derives the relevant parallel transport. In our case, it is more natural to start from the transports already defined and derive the connections. This approach has been first applied to IG in Gibilisco and Pistone [18]. However, I do not construct directly the connections, but I introduce instead the accelerations associated to each transport.

Let us compute the acceleration of a curve $I\mapsto p(t)$. Let us start with the idea that the "velocity" is here the "log-velocity,"

$$t\mapsto\left(p(t),Sp(t)\right)=\left(p(t),\frac{d}{dt}\log\left(p(t)\right)\right)\in S\Delta^{\circ}\left(\Omega\right)$$

The vector $Sp(t)\in S_{p(t)}\Delta^{\circ}\left(\Omega\right)$ has to be checked against a curve in the statistical bundle, say $t\mapsto{}^{\mathrm{m}}\mathbb{U}_{p}^{p(t)}V$ for some $V\in S_{p}\Delta^{\circ}\left(\Omega\right)$. One can compute an acceleration as

$$\frac{d}{dt}\left\langle Sp(t),{}^{\mathrm{m}}\mathbb{U}_{p}^{p(t)}V\right\rangle_{p(t)}=$$
$$\frac{d}{dt}\left\langle{}^{\mathrm{e}}\mathbb{U}_{p(t)}^{p}Sp(t),V\right\rangle_{p}=\left\langle\frac{d}{dt}{}^{\mathrm{e}}\mathbb{U}_{p(t)}^{p}Sp(t),V\right\rangle_{p}=$$
$$\left\langle{}^{\mathrm{e}}\mathbb{U}_{p}^{p(t)}\frac{d}{dt}{}^{\mathrm{e}}\mathbb{U}_{p(t)}^{p}Sp(t),{}^{\mathrm{m}}\mathbb{U}_{p}^{p(t)}V\right\rangle_{p(t)}\ .$$

Definition 7 (e-acceleration)**.** The *exponential acceleration* ${}^{e}\mathrm{D}^{2}p(t)$ is

$${}^{\mathrm{e}}\mathbb{U}_{p}^{p(t)}\frac{d}{dt}{}^{\mathrm{e}}\mathbb{U}_{p(t)}^{p}Sp(t)=$$
$${}^{\mathrm{e}}\mathbb{U}_{p}^{p(t)}\frac{d}{dt}\left(\frac{\dot{p}(t)}{p(t)}-\mathbb{E}_{p}\left[\frac{\dot{p}(t)}{p(t)}\right]\right)=$$
$${}^{\mathrm{e}}\mathbb{U}_{p}^{p(t)}\left(\frac{\ddot{p}(t)p(t)-\dot{p}(t)^{2}}{p(t)^{2}}-\mathbb{E}_{p}\left[\frac{\ddot{p}(t)p(t)-\dot{p}(t)^{2}}{p(t)^{2}}\right]\right)=$$
$$\frac{\ddot{p}(t)p(t)-\dot{p}(t)^{2}}{p(t)^{2}}-\mathbb{E}_{p(t)}\left[\frac{\ddot{p}(t)p(t)-\dot{p}(t)^{2}}{p(t)^{2}}\right]=$$
$$\boxed{\frac{\ddot{p}(t)}{p(t)}-\left(Sp(t)\right)^{2}+\mathbb{E}_{p(t)}\left[\left(Sp(t)\right)^{2}\right]}\ .$$

Proposition 8. *Exponential families have null exponential acceleration.*

Proof. In fact, for $p(t) = \exp(tU - \psi(t)) \cdot p$, one has ${}^{\mathrm{e}}\mathbb{U}^{p}_{p(t)} Sp(t) = U - \mathbb{E}_p[U]$, so that $\frac{d}{dt}\,{}^{\mathrm{e}}\mathbb{U}^{p}_{p(t)} Sp(t) = 0$. □

Note that for the exponential family above,

$$(Sp(t))^2 = (u - \dot\psi(t))^2 \ ,$$
$$\ddot p(t) = \frac{d}{dt}[p(t)(u - \dot\psi(t))] = p(t)(u - \dot\psi(t))^2 - p(t)\ddot\psi(t) \ ,$$

so that

$$\frac{\ddot p(t)}{p(t)} - (Sp(t))^2 + \mathbb{E}_{p(t)}\left[(Sp(t))^2\right] = (u - \dot\psi(t))^2 - \ddot\psi(t) - (u - \dot\psi(t))^2 + \ddot\psi(t) = 0 \ .$$

A second option is to compute the acceleration as

$$\frac{d}{dt}\left\langle Sp(t), {}^{\mathrm{e}}\mathbb{U}^{p(t)}_{p} V\right\rangle_{p(t)} = \frac{d}{dt}\left\langle {}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} Sp(t), V\right\rangle_{p} = \left\langle \frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} Sp(t), V\right\rangle_{p} = \left\langle {}^{\mathrm{m}}\mathbb{U}^{p(t)}_{p}\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} Sp(t), {}^{\mathrm{e}}\mathbb{U}^{p(t)}_{p} V\right\rangle_{p(t)} \ .$$

Definition 8 (m-acceleration)**.** The *mixture acceleration* ${}^{m}\mathrm{D}^2 p(t)$ is

$${}^{\mathrm{m}}\mathbb{U}^{p(t)}_{p}\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} Sp(t) = \frac{p}{p(t)}\frac{d}{dt}\left(\frac{p(t)}{p}\frac{\dot p(t)}{p(t)}\right) = \boxed{\frac{\ddot p(t)}{p(t)}} \ .$$

Proposition 9. *Mixture models $t \mapsto (1+tU)p$ have null mixture acceleration.*

Proof. Obvious. □

Exercise One could define a Riemannian accelleration by ${}^{0}\mathbb{U}^{p(t)}_{p}\frac{d}{dt}\,{}^{0}\mathbb{U}^{p}_{p(t)} Sp(t)$. See some related computations in Pistone [10, 32], Lods and Pistone [33].

V.1. Taylor formula

Let us apply the definitions of acceleration to the compute the 2nd order Taylor formula. Given a regular function $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$ and a regular curve $t \mapsto p(t)$, the first derivative of $t \mapsto f(p(t))$ can be written in two ways:

$$\begin{aligned}\frac{d}{dt} f(p(t)) &= \langle \operatorname{grad} f(p(t)), Sp(t)\rangle_{p(t)} \\ &= \left\langle {}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t)), {}^{\mathrm{e}}\mathbb{U}^{p}_{p(t)} Sp(t)\right\rangle_{p} \\ &= \left\langle {}^{\mathrm{e}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t)), {}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} Sp(t)\right\rangle_{p} \ .\end{aligned}$$

Using the first one,

$$\frac{d^2}{dt^2} f(p(t)) = \left\langle \frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t)), {}^{\mathrm{e}}\mathbb{U}^{p}_{p(t} Sp(t)\right\rangle_{p} + \left\langle {}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t)), \frac{d}{dt}\,{}^{\mathrm{e}}\mathbb{U}^{p}_{p(t} Sp(t)\right\rangle_{p} = \left\langle {}^{\mathrm{m}}\mathbb{U}^{p(t)}_{p}\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t)), Sp(t)\right\rangle_{p(t)} + \left\langle \operatorname{grad} f(p(t)), {}^{e}\mathrm{D}^2 p(t)\right\rangle_{p(t)} \ .$$

Assume now that $p(t) = \mathrm{e}^{tU - \psi(t)} p$, $U \in S_p\Delta^\circ(\Omega)$, so that ${}^{e}\mathrm{D}^2 p(t) = 0$ and the second term above cancels, to give

$$\frac{d^2}{dt^2} f(p(t)) = \left\langle {}^{\mathrm{m}}\mathbb{U}^{p(t)}_{p}\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t)), Sp(t)\right\rangle_{p(t)} \ .$$

Definition 9 (m-Hessian)**.** Define the *mixture Hessian* ${}^{m}\mathrm{Hess}_U f(p)$ to be

$${}^{\mathrm{m}}\mathbb{U}^{p(t)}_{p}\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{p(t)} \operatorname{grad} f(p(t))\Big|_{t=0} \in S_p\Delta^\circ(\Omega)$$

when $p(0) = p$ and $Sp(0) = U$.

The second derivative above reduces to

$$\frac{d^2}{dt^2} f(p(t)) = \left\langle {}^{m}\mathrm{Hess}_{Sp(t)} f(p(t)), Sp(t)\right\rangle_{p(t)} = \left\langle {}^{m}\mathrm{Hess}_{U - \dot\psi(t)} f(p(t)), U - \dot\psi(t)\right\rangle_{p(t)}$$

and one can write for $q = \mathrm{e}^{U - \psi(1)} p$, the Taylor formula

$$f(q) - f(p) = \langle \operatorname{grad} f(p), U\rangle_p + \int_0^1 (1-t)\left\langle {}^{m}\mathrm{Hess}_{U - \dot\psi(t)} f(p(t)), U - \dot\psi(t)\right\rangle_{p(t)} \, dt = \boxed{\langle \operatorname{grad} f(p), U\rangle_p + \frac{1}{2}\left\langle {}^{m}\mathrm{Hess}_U f(p), U\right\rangle_p + R_2(p, U)} \ .$$

Exercise. In a similar way one could derive a Taylor formula for the e-Hessian,

$$f(q) - f(p) = \boxed{\langle \operatorname{grad} f(p), V\rangle_p + \frac{1}{2}\left\langle {}^{e}\mathrm{Hess}_V f(p), V\right\rangle_p + R_2(p, U)} \ .$$

Notice that the "increments" U and V in the equations above are quite different! The Riemannian Taylor formula could be derived along similar lines.

VI. ATLASES

I have shown in the previous sections how the calculus on the statistical bundle works. I now turn to the explicit introduction of special atlases of charts on the statistical bundle. Each of atlas will have the following special properties:

A. The tangent space computed in the atlas is the corresponding fiber of the statistical bundle;

B. The atlas induces induces one of the connections.

Notice that I define an atlas which is not the maximal atlas nor the atlas deduced from the embedding into $\mathbb{R}^{\Omega} \times \mathbb{R}^{\Omega}$. It is an *affine atlas*, that is all the transition maps are affine functions.

Definition 10 (Exponential atlas)**.** For each $p \in \Delta^{\circ}(\Omega)$, define

$$s_p \colon S\Delta^{\circ}(\Omega) \ni (q, w) \mapsto \left(\log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right], {}^{\mathrm{e}}\mathbb{U}_q^p w\right) \in S_p\Delta^{\circ}(\Omega) \times S_p\Delta^{\circ}(\Omega)$$

Recall the notation $S_p\Delta^{\circ}(\Omega) = B_p$. Notice there is a one chart for each $p \in \Delta^{\circ}(\Omega)$ and $s_p(p) = 0$. One can say that s_p is the chart centered at p..

Proposition 10 (Properties of the e-atlas)**.**

1. *If $u = s_p(q)$, then $q = \mathrm{e}^{u - K_p(u)} \cdot p$ with $K_p(u) = \log \mathbb{E}_p[\mathrm{e}^u]$.*
2. *The patches are*
$$s_p^{-1} \colon (u, v) \mapsto (\mathrm{e}^{u - K_p(u)} \cdot p, v - dK_p(u)[v])$$
3. *The transitions are given for $u, v \in B_{p_2}$ by*
$$s_{p_1} \circ s_{p_2}^{-1} \colon (u, v) \mapsto \left({}^{e}\mathbb{U}_{p_2}^{p_1} u + \log \frac{p_2}{p_1} - \mathbb{E}_{p_1}\left[\log \frac{p_2}{p_1}\right]\right) \in B_{p_1} \times B_{p_1}$$
4. *The tangent bundle identifies with the statistical bundle. If $p(0) = p$, then*
$$\frac{d}{dt} s_p(p(t))\Big|_{t=0} = {}^{e}\mathbb{U}_{p(t)}^{p} Sp(t)\Big|_{t=0} = Sp(0) \ .$$
5. *The velocity computed in the chart of the lift $t \mapsto (p(t), Sp(t))$ is $t \mapsto (Sp(t), {}^{e}\mathrm{D}^2 p(t))$.*

Proof. It is an exercise. □

The same ideas can be applied to the m-geometry. Notice that specific properties of the finite state space case are used. If such special properties are not available, one must carefully distinguish between B_p and its pre-dual ${}^*B_p$, see Pistone [10].

Definition 11 (Mixture atlas)**.** For each $p \in \Delta^{\circ}(\Omega)$, define

$$\eta_p \colon S\Delta^{\circ}(\Omega) \ni (q, w) \mapsto \left(\frac{q}{p} - 1, {}^{\mathrm{m}}\mathbb{U}_q^p w\right) \in S_p\Delta^{\circ}(\Omega) \times S_p\Delta^{\circ}(\Omega)$$

Proposition 11 (Properties of the m-atlas)**.**

1. *If $u = \eta_p(q)$, then $q = (1 + u)p$.*
2. *The patches are*
$$\eta_p^{-1} \colon (u, v) \mapsto ((1 + u)p, (1 + u)w)$$
3. *The transitions are*
$$\eta_{p_1} \circ \eta_{p_2}^{-1} \colon (u, v) \mapsto \left((1 + u)\frac{p_2}{p_1} - 1, {}^{m}\mathbb{U}_{p_1}^{p_2} v\right)$$
4. *The tangent bundle identifies with the statistical bundle. If $p(0) = p$, then*
$$\frac{d}{dt}\eta_p(p(t))\Big|_{t=0} = {}^{m}\mathbb{U}_{p(t)}^{p} Sp(t)\Big|_{t=0} = Sp(0) \ .$$
5. *The velocity computed in the chart of the lift $t \mapsto (p(t), Sp(t))$ is $t \mapsto (Sp(t), {}^{m}\mathrm{D}^2 p(t))$.*

Proof. It is an *Exercise.* □

VI.1. Using parameters

Even if one wants to study the geometry of the full simplex, it is possible to introduce parameters because the simplex is finite-dimensional. Parts of the presentation in this section are taken from Pistone and Rogantin [34].

Computations are frequently performed in a *parametrization*, even in applications such as Compositional Data Analysis, which is descriptive statistics of the full simplex, see Aitchison [35].

A parametrization of the open simplex is a 1-to-1 mapping

$$\pi \colon \Theta \ni \boldsymbol{\theta} \mapsto \pi(\boldsymbol{\theta}) \in \Delta^{\circ}(\Omega) \,,$$

Θ being an open set in $\mathbb{R}^n$, $n = \#\Omega - 1$. As the j-th coordinate curve is obtained by fixing the other $(n-1)$ components and moving θ_j only, the differential scores of each j-th coordinate curve are defined as the random variables

$$S_j\pi(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} \log \pi(\boldsymbol{\theta}), \quad j = 1, \dots, n.$$

The sequence $(S_j\pi(\boldsymbol{\theta}) \colon j = 1, \dots, n)$ is assumed to be a vector basis of the fiber $S_{\pi(\boldsymbol{\theta})}\Delta^{\circ}(\Omega)$. The expression of the inner product in such a basis is

$$\left\langle \sum_{i=1}^n \alpha_i S_i\pi(\boldsymbol{\theta}), \sum_{j=1}^n \beta_j S_j\pi(\boldsymbol{\theta}) \right\rangle_{\pi(\boldsymbol{\theta})} = \sum_{i,j=1}^n \alpha_i \beta_j \left\langle S_i\pi(\boldsymbol{\theta}), S_j\pi(\boldsymbol{\theta}) \right\rangle_{\pi(\boldsymbol{\theta})} \ .$$

Definition 12 (Fisher Information)**.** The matrix

$$I(\boldsymbol{\theta}) = \left[\langle S_i\pi(\boldsymbol{\theta}), S_j\pi(\boldsymbol{\theta})\rangle_{\pi(\boldsymbol{\theta})}\right]_{i,i=1}^{n} = \left[\langle \partial_i \log\pi(\boldsymbol{\theta}), \partial_j \log\pi(\boldsymbol{\theta})\rangle_{\pi(\boldsymbol{\theta})}\right]_{i,i=1}^{n} = \left[\sum_{x\in\Omega} \frac{\partial_i\pi(x;\boldsymbol{\theta})\partial_i\pi(x;\boldsymbol{\theta})}{\pi(x;\boldsymbol{\theta})}\right]_{i,i=1}^{n}$$

is the *Fisher information matrix* of the parametrization π. Notice that the Fisher Information matrix depends on the parametrization.

Consider a curve expressed in the parametrization,

$$t \mapsto p(t) = \pi(\boldsymbol{\theta}(t)) \ ,$$

and compute the differential score in the parametrization as

$$Sp(t) = \frac{d}{dt}\log\pi(\boldsymbol{\theta}(t)) = \sum_{i=1}^{n} S_i\pi(\boldsymbol{\theta}(t))\dot{\theta}_i(t) \ .$$

Let us now turn to consider the expression of the natural gradient in the parametrization. Let be given $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$. The random variable $\operatorname{grad} f(p)$ is defined by

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

that is, for $p(t) = \pi(\boldsymbol{\theta}(t))$,

$$\frac{d}{dt}f(\pi(\boldsymbol{\theta}(t))) = \langle \operatorname{grad} f(\pi(\boldsymbol{\theta}((t)))), S\pi(\boldsymbol{\theta}(t))\rangle_{\pi(\boldsymbol{\theta}(t))} \ .$$

If one writes $\tilde{f}(\boldsymbol{\theta}) = f\circ\pi(\boldsymbol{\theta})$ and expresses the differential score in the basis,

$$\sum_{i=1}^{n} \partial_i\tilde{f}(\boldsymbol{\theta}(t))\dot{\theta}_i(t) = \sum_{i=1}^{n} \langle \operatorname{grad} f(\pi(\boldsymbol{\theta}((t)))), S_j\pi(\boldsymbol{\theta}(t))\rangle_{\pi(\boldsymbol{\theta}(t))}\dot{\theta}_i(t) \ .$$

As $\boldsymbol{\theta}$ and $\dot{\boldsymbol{\theta}}$ in the equation above are generic, it follows the system of equations

$$\partial_i\tilde{f}(\boldsymbol{\theta}) = \langle \operatorname{grad} f(\pi(\boldsymbol{\theta})), S_j\pi(\boldsymbol{\theta})\rangle_{\pi(\boldsymbol{\theta})} \ , \quad i = 1,\dots,n \ .$$

This gives the form of the natural gradient that was originally proposed by Amari [36].

Proposition 12. *The expression of* $\operatorname{grad} f(\pi(\boldsymbol{\theta}))$ *has components in the basis* $(S_j\pi(\boldsymbol{\theta})\colon j = 1,\dots n)$ *given by*

$$I(\boldsymbol{\theta})^{-1}\nabla\tilde{f}(\boldsymbol{\theta}) \ .$$

VI.2. Special parametrizations

The *common parametrization* of the (flat) simplex $\Delta^\circ(\Omega)$ is the projection on the *solid simplex*

$$\Gamma_n = \left\{ \boldsymbol{\eta} \in \mathbb{R}^n \,\middle|\, 0 < \eta_j, \sum_{j=1}^{n}\eta_j < 1 \right\} \ ,$$

$$\pi\colon \Gamma_n \ni \boldsymbol{\eta} \mapsto \left(1 - \sum_{j=1}^{n}\eta_j, \eta_1, \dots, \eta_n\right) \in \Delta^\circ(\Omega) \ ,$$

in which case $\partial_j\pi(\boldsymbol{\eta})$, $j = 1,\dots,n$, is the random variable with values -1 at $x = 0$, 1 at $x = j$, 0 otherwise, hence $\partial_j\pi(\boldsymbol{\eta}) = ((X = j) - (X = 0))$ and

$$S_j\pi(\boldsymbol{\eta}) = ((X = j) - (X = 0))\,/\pi(\boldsymbol{\eta}).$$

The element $I_{jh}(\boldsymbol{\eta})$ of the Fisher information matrix is

$$\mathbb{E}_{\pi(\boldsymbol{\eta})}\left[\frac{(X = j) - (X = 0)}{\pi(X;\boldsymbol{\eta})}\frac{(X = h) - (X = 0)}{\pi(X;\boldsymbol{\eta})}\right] = \sum_x \pi(x,\boldsymbol{\eta})^{-1}\left((x = j)(j = h) + (x = 0)\right) = \eta_j^{-1}(j = h) + \left(1 - \sum_k \eta_k\right)^{-1} \ ,$$

hence,

$$I(\boldsymbol{\eta}) = \operatorname{diag}(\boldsymbol{\eta})^{-1} + \left(1 - \sum_{j=1}^{n}\eta_j\right)^{-1}[1]_{i,j=1}^{n} \ .$$

Example Consider $n = 3$. The Fisher information matrix, its inverse and the determinant of the inverse are, respectively,

$$I(\eta_1,\eta_2,\eta_3) = (1 - \eta_1 - \eta_2 - \eta_3)^{-1}\times \begin{bmatrix} \eta_1^{-1}(1 - \eta_2 - \eta_3) & 1 & 1 \\ 1 & \eta_2^{-1}(1 - \eta_1 - \eta_3) & 1 \\ 1 & 1 & \eta_3^{-1}(1 - \eta_1 - \eta_2) \end{bmatrix},$$

$$I(\eta_1,\eta_2,\eta_3)^{-1} = \begin{bmatrix} (1-\eta_1)\eta_1 & -\eta_1\eta_2 & -\eta_1\eta_3 \\ -\eta_1\eta_2 & (1-\eta_2)\eta_2 & -\eta_2\eta_3 \\ -\eta_1\eta_3 & -\eta_2\eta_3 & (1-\eta_3)\eta_3 \end{bmatrix} \ ,$$

$$\det\left(I(\eta_1,\eta_2,\eta_3)^{-1}\right) = (1 - \eta_1 - \eta_2 - \eta_3)\eta_1\eta_2\eta_3 \ .$$

Note that the computation of the inverse of $I(\boldsymbol{\eta})$ is an application of the Sherman-Morrison formula and the computation of the determinant of $I(\boldsymbol{\eta})^{-1}$ is an application of the matrix determinant lemma.

For general n,

Proposition 13.

1. *The inverse of the Fisher information matrix is*

$$I(\boldsymbol{\eta})^{-1} = \operatorname{diag}(\boldsymbol{\eta}) - \boldsymbol{\eta}\boldsymbol{\eta}^t \ .$$

2. *In particular, $I(\boldsymbol{\eta})^{-1}$ is zero on the vertexes of the simplex, only.*

3. *The determinant of the inverse Fisher information matrix is*

$$\det\left(I(\boldsymbol{\eta})^{-1}\right) = \left(1 - \sum_{i=1}^{n} \eta_i\right) \prod_{i=1}^{n} \eta_i \ .$$

4. *The determinant of $I(\boldsymbol{\eta})^{-1}$ is zero on the borders of the simplex, only.*

5. *On the interior of each facet, the rank of $I(\boldsymbol{\eta})^{-1}$ is $n-1$ and the $n-1$ liner independent column vectors generate the subspace parallel to the facet itself.*

Proof.

1. By direct computation, $I(\boldsymbol{\eta})I(\boldsymbol{\eta})^{-1}$ is the identity matrix.

2. The diagonal elements of $I(\boldsymbol{\eta})^{-1}$ are zero if $\eta_j = 1$ or $\eta_j = 0$, for $j = 1, \dots, n$. If, for a given j, $\eta_j = 1$, then the elements of $I(\boldsymbol{\eta})^{-1}$ are zero if $\eta_h = 0$, $h \neq j$. The remaining case corresponds to $\eta_j = 0$ for all j. Then $I(\boldsymbol{\eta})^{-1} = 0$ on all the vertexes of the simplex.

3. It follows from Matrix Determinant Lemma.

4. The determinant factors in terms corresponding to the equations of the facets.

5. Given i, the conditions $\eta_i = 0$ and $\eta_j \neq 0, 1$ for all $j \neq i$, define the interior of the facet orthogonal to standard base vector e_i. In this case the i-th row and the i-th column of $I(\boldsymbol{\eta})^{-1}$ are zero and the complement matrix corresponds to the inverse of a Fisher information matrix in dimension $n-1$ with non zero determinant. It follows that the subspace generated by the columns has dimension $n-1$ and coincides with the space orthogonal to η_i. Consider the facet defined by $(1 - \sum_{i=1}^{n} \eta_i) = 0$, $\eta_i \neq 0, 1$ for all i. For a given j, the matrix without the j-th row and the j-th column has determinant $\left(1 - \sum_{i=1, i\neq j}^{n} \eta_i\right) \prod_{i=1, i\neq j}^{n} \eta_i$. On the considered facet this determinant is different to zero and $I(\boldsymbol{\eta})^{-1}$ has rank $n-1$ and their columns are orthogonal to the constant vector. □

Exercise. Another parametrization is the *exponential parametrization* based on the exponential family with sufficient statistics $X_j = (X = j)$, $j = 1, \dots, n$,

$$\pi \colon \mathbb{R}^n \ni \boldsymbol{\theta} \mapsto \exp\left(\sum_{j=1}^{n} \theta_j X_j - \psi(\boldsymbol{\theta})\right) \frac{1}{n+1} \ , \quad \text{with}$$

$$\psi(\boldsymbol{\theta}) = \log\left(1 + \sum_j \mathrm{e}^{\theta_j}\right) - \log(n+1) \ .$$

Instead of using a parametrization on an open set of $\mathbb{R}^n$, $n = \#\Omega - 1$, one could use as parameter set a manifold with the correct dimension. This approach has been made systematic in the presentation of IG by Ay *et al.* [3]. The original example has been already discussed in section III.1.

VII. GENERALISED STATISTICAL BUNDLE

This section is devoted to a brief discuss of a generalisation of IG based on the idea of a "deformed" exponential functions. That is, every positive density is represented as a random variable transformed by a function with shape similar to that of the exponential function. Such models where first introduced as a replacement of Gibbs statistics by Tsallis [37]. The presentation below follows Naudts [38] and Montrucchio and Pistone [39].

The basic relation leading to the definition of differential score can be generalised as follows. Let be given a positive real function A with domain $]0, +\infty[$ and define the A-logarithm to be the strictly increasing function

$$\log_A(x) = \int_1^x \frac{du}{A(u)} \ .$$

Note that $\log_A(1) = 0$, that $\log_A$ is concave if A is increasing, and that $A(x) = x$ gives the usual logarithm. The A-exponential is $\exp_A = \log_A^{-1}$. It holds $\exp_S'(y) = A(\exp_A(y))$.

A notable example is the Tsallis logarithm, or q-logarithm, which is obtained when $A(x) = x^q$ for some given real q. The deformed cases, i.e., $q \neq 1$, can be computed explicitly as

$$\ln_q(x) = \int_1^x \frac{du}{u^q} = \frac{x^{1-q} - 1}{1-q}$$
$$\exp_q(y) = (q + (1-q)y)^{\frac{1}{1-q}}$$

It is possible to define a q-differential score,

$$S^{(q)}p(t) = \frac{d}{dt} \ln_q(p(t)) = \frac{\dot{p}(t)}{p(t)^q} \ ,$$

together with a q-statistical bundle

$$S^{(q)}\Delta^\circ(\Omega) = \left\{ (p, U) \,\middle|\, p \in \Delta^\circ(\Omega), \sum_{x\in\Omega} U(x) p(x)^q = 0 \right\}$$

It is possible to repeat in this setting the construction that led to a definition of entropy. One needs a section U of the q-statistical bundle such that $(p, U(p)) \in S^{(q)} \in \Delta^\circ(\Omega)$ and $p = \exp_q((U(q) - H_q(p))$. The conditions are satisfied with

$$p = \exp_q(\ln_q p) = \exp_q(\ln_q p - \mathbb{M}_{p^q}(\ln_q p) + \mathbb{M}_{p^q}(\ln_q p)) \ ,$$

where $\mathbb{M}_{p^q}$ is the sum with weight p^q. The generalised entropy is

$$H_q(p) = -\mathbb{M}_{p^q}(\ln_q p) = -\sum_{x\in\Omega} p(x)^q \ln_q p(x) = -\sum_{x\in\Omega} p(x)^q \frac{p(x)^{1-q}-1}{1-q} = \frac{-1+\sum_{x\in\Omega} p(x)^q}{1-q} ,$$

that is, the Tsallis entropy.

Other interesting examples are the deformed logarithms defined in Kaniadakis [40, 41], a special case being

$$\ln_1 = \frac{x-x^{-1}}{2} = \int_1^x \frac{2u^2}{1+u^2}\, du ,$$

and the deformed logarithm defined in Newton [42],

$$\ln_{\mathrm{N}}(x) = \log x + x - 1 = \int_1^x \frac{u}{1+u}\, du .$$

VIII. EXERCISES

1. Study the curve

$$t \mapsto \left(\frac{1}{2} + \frac{1}{2}(1-tU)^2 - \frac{t}{2}\,\mathbb{E}_p\left[U^2\right]\right)\cdot p, \quad U \in S_p\Delta^\circ(\Omega) .$$

See Eguchi [43].

2.. Study the curve

$$p(t) = \left(tU + \sqrt{1 - t^2\,\mathbb{E}_p\left[U^2\right]}\right)^2 \cdot p .$$

See Burdet *et al.* [44].

3. Check whether

$$t \mapsto p(t) = (1 + t^2\,\mathbb{E}_p\left[U^2\right])^{-1}(1+tU)^2 \cdot p$$

is the flow of the h-transport.

4. Chose a $U \in S_p\Delta^\circ(\Omega)$or each with unit p-norm, $\mathbb{E}_p\left[U^2\right] = 1$, and consider the model

$$t \mapsto p(t) = \frac{(1+(\sinh t)U)^2}{\cosh^2 t} \cdot p ,$$

where $1 + (\sinh t)U > 0$ if $t \in I$, I neighborhood of 0, so that $p(t) > 0$ and $\mathbb{E}_p\left[(1+(\sinh t)U)^2\right] = 1 + \sinh^2 t = \cosh^2 t$. $I \ni t \mapsto p(t)$ is a regular curve with differential score

$$Sp(t) = 2\left(\frac{(\cosh t)U}{1+(\sinh t)U} - \frac{\sinh t}{\cosh t}\right) .$$

Compute ${}^0\mathbb{U}_p^{p(t)}U$.

5 It is possible to define the Riemannian atlas through the embedding of $\Delta^\circ(\Omega)$ onto the the tangent space of the unit sphere, see a presentation in the style of the previous ones in Pistone [10].

[1] S. Amari and H. Nagaoka, *Methods of information geometry* (American Mathematical Society, 2000) pp. x+206, translated from the 1993 Japanese original by Daishi Harada.

[2] S.-i. Amari, *Information geometry and its applications*, Applied Mathematical Sciences, Vol. 194 (Springer, [Tokyo], 2016) pp. xiii+374.

[3] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer, *Information geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], Vol. 64 (Springer, Cham, 2017) pp. xi+407.

[4] L. D. Landau and E. M. Lifshits, *Course of Theoretical Physics. Statistical Physics.*, 3rd ed., Vol. V (Butterworth-Heinemann, 1980).

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

[6] S. Lang, *Differential and Riemannian manifolds*, 3rd ed., Graduate Texts in Mathematics, Vol. 160 (Springer-Verlag, 1995) pp. xiv+364.

[7] W. P. A. Klingenberg, *Riemannian geometry*, 2nd ed., De Gruyter Studies in Mathematics, Vol. 1 (Walter de Gruyter & Co., Berlin, 1995) pp. x+409.

[8] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28 (Princeton University Press, 1970) pp. xviii+451.

[9] A. Barvinok, *A course in convexity*, Graduate Studies in Mathematics, Vol. 54 (American Mathematical Society, Providence, RI, 2002) pp. x+366.

[10] G. Pistone, Nonparametric information geometry, in *Geometric science of information*, Lecture Notes in Comput. Sci., Vol. 8085, edited by F. Nielsen and F. Barbaresco (Springer, Heidelberg, 2013) pp. 5–36, first International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings.

[11] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer, *Information geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in

Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], Vol. 64 (Springer, Cham, 2017) pp. xi+407.
[12] L. Montrucchio and G. Pistone, Kantorovich distance on a weighted graph (2019), arXiv:1905.07547 [math.PR].
[13] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, 2nd ed., Lectures in Mathematics ETH Zrich (Birkhuser Verlag, Basel, 2008) pp. x+334.
[14] S. Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics, Vol. 28 (Springer-Verlag, 1985) pp. v+290.
[15] S. L. Lauritzen, Differential geometry in statistical inference (Institute of Mathematical Statistics, New York, NY, USA, 1987) Chap. Statistical Manifolds, pp. 163–216.
[16] M. K. Murray and J. W. Rice, *Differential Geometry and Statistics*, Monographs on Statistics and Applied Probability No. 48 (Chapman & Hall, 1993).
[17] R. E. Kass and P. W. Vos, *Geometrical foundations of asymptotic inference*, Wiley Series in Probability and Statistics: Probability and Statistics (John Wiley & Sons, Inc., New York, 1997) pp. xii+355, a Wiley-Interscience Publication.
[18] P. Gibilisco and G. Pistone, Connections on non-parametric statistical manifolds by Orlicz space geometry, IDAQP **1**, 325 (1998).
[19] H. V. Lê, The uniqueness of the Fisher metric as information metric, Ann. Inst. Statist. Math. **69**, 879 (2017).
[20] G. Pistone, E. Riccomagno, and H. P. Wynn, *Algebraic statistics: Computational commutative algebra in statistics*, Monographs on Statistics and Applied Probability, Vol. 89 (Chapman & Hall/CRC, Boca Raton, FL, 2001) pp. xvii+160.
[21] L. Pachter and B. Sturmfels, eds., *Algebraic Statistics for Computational Biology* (Cambridge University Press, 2005).
[22] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on algebraic statistics*, Oberwolfach Seminars, Vol. 39 (Birkhuser Verlag, 2009) pp. viii+171.
[23] S. Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge Monographs on Applied and Computational Mathematics, Vol. 25 (Cambridge University Press, Cambridge, 2009) pp. viii+286.
[24] S. Aoki, H. Hara, and A. Takemura, *Markov bases in algebraic statistics*, Springer Series in Statistics (Springer, New York, 2012) pp. xii+298.
[25] P. Zwiernik, *Semialgebraic statistics and latent tree models*, Monographs on Statistics and Applied Probability, Vol. 146 (Chapman & Hall/CRC, Boca Raton, FL, 2016) pp. xx+225.
[26] S. Sullivan, *Algebraic Statistics*, Graduate Studies in Mathematics No. 194 (AMS, 2018).
[27] P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, eds., *Algebraic and geometric methods in statistics* (Cambridge University Press, 2010) pp. xvi+430.
[28] L. Malagò, M. Matteucci, and G. Pistone, Stochastic relaxation as a unifying approach in 0/1 programming (2009), NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), December 11-12, 2009, Whistler Resort & Spa, Canada.
[29] L. Malagò, M. Matteucci, and G. Pistone, Stochastic natural gradient descent by estimation of empirical covariances., in *IEEE Congress on Evolutionary Computation* (IEEE, 2011) pp. 949–956.
[30] L. Malagò, M. Matteucci, and G. Pistone, Towards the geometry of estimation of distribution algorithms based on the exponential family, in *Proceedings of the 11th workshop on Foundations of genetic algorithms*, FOGA '11 (ACM, New York, NY, USA, 2011) pp. 230–242.
[31] L. Malagò, M. Matteucci, and G. Pistone, Natural gradient, fitness modelling and model selection: A unifying perspective, in *IEEE Congress on Evolutionary Computation* (IEEE, 2013) pp. 486–493.
[32] G. Pistone, Examples of the application of nonparametric information geometry to statistical physics, Entropy **15**, 4042 (2013).
[33] B. Lods and G. Pistone, Information geometry formalism for the spatially homogeneous Boltzmann equation, Entropy **17**, 4323 (2015).
[34] G. Pistone and M. P. Rogantin, The gradient flow of the polarization measure. with an appendix (2015), arXiv:1502.06718.
[35] J. Aitchison, *The statistical analysis of compositional data*, Monographs on Statistics and Applied Probability (Chapman & Hall, London, 1986) pp. xvi+416.
[36] S.-I. Amari, Natural gradient works efficiently in learning, Neural Computation **10**, 251 (1998).
[37] C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics, J. Statist. Phys. **52**, 479 (1988).
[38] J. Naudts, *Generalised thermostatistics* (Springer-Verlag London Ltd., 2011) pp. x+201.
[39] L. Montrucchio and G. Pistone, Deformed exponential bundle: the linear growth case, in *Geometric Science of Information*, LNCS No. 10589, edited by F. Nielsen and F. Barbaresco (Springer, 2017) pp. 239–246, third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings.
[40] G. Kaniadakis, Non-linear kinetics underlying generalized statistics, Physica A **296**, 405 (2001).
[41] G. Kaniadakis, H-theorem and generalized entropies within the framework of nonlinear kinetics, Physics Letters A **288**, 283 (2001).
[42] N. J. Newton, An infinite-dimensional statistical manifold modelled on Hilbert space, J. Funct. Anal. **263**, 1661 (2012).
[43] S. Eguchi, Tubular modelling approach to statistical method for observational studies (2005), 2nd International Symposium on Information Geometry and its Applications Tokyo.
[44] G. Burdet, P. Combe, and H. Nencka, On real Hilbertian info-manifolds, in *Disordered and complex systems (London, 2000)*, AIP Conf. Proc., Vol. 553 (Amer. Inst. Phys., 2001) pp. 153–158.