TUTORIAL HANDOUT GENOVA JUNE 11

INFORMATION GEOMETRY AND ALGEBRAIC STATISTICS ON A FINITE STATE SPACE AND ON GAUSSIAN MODELS

LUIGI MALAGÒ AND GIOVANNI PISTONE

Contents

Date: June 11, 2015.

The authors wish to thank Henry Wynn for his comments on a draft of this tutorial. Some matherial reproduced here is part of work in progress by the authors and collaborators. G. Pistone is supported by the de Castro Statistics Initiative, Collegio Carlo Alberto, Moncalieri; and he is a member of GNAMPA–INdAM.

1. Introduction

It was shown by C. R. Rao in a paper published 1945 [23] that the set of positive probabilities $\Delta^\circ(\Omega)$ on a finite state space Ω is a *Riemannian manifold*, as it is defined in classical treatises such as [7] and [11], but in a way which is of interest for Statistics. It was later pointed out by Sun-Ichi Amari that it is actually possible to define two affine geometries of Hessian type [29] on top of the Rao's Riemannian geometry, but see also the original contribution by Steffen Laurizen [12]. This development was somehow stimulated by two papers by Efron [8, 9]. The original work of Amari was published in the '80s, see Shun'ichi Amari [1], see monograph presentations in [10] and [3]. Amari gave to this new topic the name of *Information Geometry* and provided many applications, in particular in Machine Learning [2].

Information Geometry and Algebraic statistics are deeply connected because of the central place occupied by statistical exponential families [4] in both fields. There is possibly a simpler connection, which is the object of the first part of this presentation. The present tutorial is focused mainly on Differential Geometry.

The present tutorial treats only cases where the statistical model is *parametric*. However, there is an effort to use methods that scale well to the case where the statistical model is essentially infinite dimensional, e.g. [5, 6], *parry—dawid—lauritzen:2012*, [14], and, in general, all applications in Statistical Physics.

Where to start from? Here is my choice, but read the comments by C. R. Rao to Scholarpedia.

1.1. C. R. Rao.

In [23] we find the following computation:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}_{p(t)}\left[U\right] &= \frac{d}{dt}\sum_x U(x)p(x;t) \\
&= \sum_x U(x)\frac{d}{dt}p(x;t) \\
&= \sum_x U(x)\frac{d}{dt}\log\left(p(x;t)\right)p(x;t) \\
&= \sum_x \left(U(x)-\mathbb{E}_{p(t)}\left[U\right]\right)\frac{d}{dt}\log\left(p(x;t)\right)p(x;t) \\
&= \mathbb{E}_{p(t)}\left[\left(U-\mathbb{E}_{p(t)}\left[U\right]\right)\frac{d}{dt}\log\left(p(t)\right)\right] \\
&= \left\langle \left(U-\mathbb{E}_{p(t)}\left[U\right]\right), \frac{d}{dt}\log\left(p(t)\right)\right\rangle_{p(t)} .
\end{aligned}$$

Here the relevant point is fact the scalar product is computed at the running $p(t)$ and the Fisher's score $\frac{d}{dt}\log\left(p(t)\right)$ appears as a measure of velocity.

1.2. S.-H. Amari.

In [2] there are applications of computations of the following type. Give a function $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$,

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

where $\operatorname{grad} f(p(t)) - \mathbb{E}_{p(t)}\left[\operatorname{grad} f(p(t))\right]$ appears as the gradient of f in the scalar product $\langle\cdot,\cdot\rangle_{\cdot}$.

2. Finite State Space: Full Simplex

Let Ω be a finite set with $n+1 = \#\Omega$ points. We denote by $\Delta(\Omega)$ the simplex of the probability functions $p\colon \Omega \to \mathbb{R}_{\geq 0}$, $\sum_{x\in\Omega} p(x) = 1$. It is a n-simplex, i.e. an n-dimensional polytope which is the convex hull of its n + 1 vertices. It is a closed and convex convex subset of the affine space $A_1(\Omega) = \left\{q \in \mathbb{R}^\Omega \middle| \sum_{x\in\Omega} q(x) = 1\right\}$ and it has non empty relative topological interior.

The interior of the simplex, Δ_n°, 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 is the union of all the faces of $\Delta(\Omega)$ as a convex set. We 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.

Remark 1. The presentation below does not use explicitely any specific parameterization of the sets $\Delta^\circ(\Omega)$, $\Delta(\Omega)$, $A_1(\Omega)$. However, the actual extension of this theory to non finite sample space requires a carefull handling as most of the topological features do not hold in such a case. One possibility is given by the so called exponential manifolds, see [21].

2.1. **Statistical bundle.** We first discuss the statistical geometry on the open simplex as deriving from a *vector bundle* with base $\Delta^\circ(\Omega)$. *The notion of vector bundle has been introduced in non-parametric Information Geometry by* [13]. Later we 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 generalization.

For each $p \in \Delta^\circ(\Omega)$ we 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, see Fig. 1. Note that, because of our orientation to Statistics, we call each element of $\mathbb{R}^\Omega$ a *random variable*. A section mapping S from probabilities $p \in \Delta^\circ(\Omega)$ to the bundle, $\mathbb{E}_p[S(p)] = 0$ is 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 $\Delta(\Omega)$.

We can give a formal definition as follows.

Definition 1.

(1) *For each $p \in \Delta^\circ(\Omega)$ let B_p be the vector space of random variables U that are p-centered,*

$$B_p = \left\{ U\colon \Omega \to \mathbb{R} \middle| \mathbb{E}_p[U] = \sum_{x\in\Omega} U(x)p(x) = 0 \right\} .$$

Each B_p is an Hilbert space for the scalar product $\langle U, V\rangle_p = \mathbb{E}_p[UV]$.

(2) *The* statistical bundle *of the open simplex $\Delta^\circ(\Omega)$ is the linear bundle on $\Delta^\circ(\Omega)$*

$$T\Delta^\circ(\Omega) = \{(p, U) | p \in \Delta^\circ(\Omega), U \in B_p\} .$$

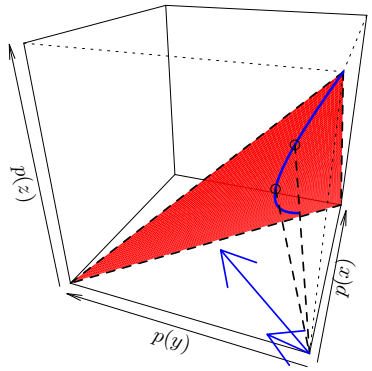

FIGURE 1. The red triangle is the simplex on the sample space with 3 points $\Omega = \{x, y, z\}$ viewed from below. The blue curve on the simplex is a one-dimensional statistical model. The probabilities p are represented by vectors from O to the point whose coordinates are $p = (p(x), p(y), p(z))$. The velocity vectors $Dp(t)$ of a curve $I \mapsto p(t)$ are represented by arrows; they are orthogonal to the vector from O to p.

It is an open subset of the variety of $\mathbb{R}^\Omega \times \mathbb{R}^\Omega$ *defined by the polynomial equations*

$$\begin{cases} \sum_{x\in\Omega} p(x) = 1 \ , \\ \sum_{x\in\Omega} U(x)p(x) = 0 \ . \end{cases}$$

(3) *A* vector field F *of the statistical bundle is a* section *of the bundle i.e.,*

$$F\colon \Delta^\circ(\Omega) \ni p \mapsto (p, F(p)) \in T\Delta^\circ(\Omega)$$

The term estimating function *is also used in the statistical literature.*

(4) *If* $I \ni t \mapsto p(t) \in \Delta^\circ(\Omega)$ *is a* C^1 *curve, its* score *is defined by*

$$Dp(t) = \frac{\dot p(t)}{p(t)} = \frac{d}{dt}\log p(t), \quad t \in I \ .$$

As the score $Dp(t)$ *is a* $p(t)$*-centered random variable which belongs to* $B_{p(t)}$ *for all* $t \in I$*, the mapping* $I \ni t \mapsto (p(t), Dp(t))$ *is a curve in the statistical bundle. Note that the notion of score extends to any curve* $p(\cdot)$ *in the affine space* $A_1(\Omega)$ *by its relation to the statistical gradient, i.e.* $Dp(t)$ *is implicitely defined by*

$$\left\langle \operatorname{grad} f(p(t)) - \mathbb{E}_{p(t)}\left[\operatorname{grad} f(p(t))\right], , \right\rangle_{p(t)} \quad f \in C^1(\mathbb{R}^\Omega) \ .$$

(5) *Given a function* $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$*, its* statistical gradient *is a vector field* $\nabla f\colon \Delta^\circ(\Omega) \ni p \mapsto (p, \nabla F(p)) \in T\Delta^\circ(\Omega)$ *such that for each differentiable curve* $t \mapsto p(t)$ *it holds*

$$\frac{d}{dt} f(p(t)) = \langle \nabla f(p(t)), Dp(t)\rangle_{p(t)}$$

Remark 2. 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 embedding, see Fig. 1. It will become the tangent bundle of the proper geometric structure which is given by special atlases.

Remark 3. We could extend the statistical bundle by taking the linear fiberts on $\Delta(\Omega)$ or $A_1(\Omega)$. In such cases the bilinear form is not always a scalar product. In fact $\langle\cdot,\cdot\rangle_p$ is not faithful where at least a component of the probability vector is zero, while it is not positive definite ourside the simplex $\Delta(\Omega)$.

Remark 4. The vector $Dp(t) \in B_{p(t)}$ is meant to represent the relative variation of the information in a one dimensional statistical model. The score is a representation of the velocity along a curve, because of the geometric interpretation of C. R. Rao's already mentioned classical computation of [23],

$$\frac{d}{dt}\mathbb{E}_{p(t)}[U] = \frac{d}{dt}\sum_x U(x)p(x;t) = \sum_x U(x)\frac{d}{dt}p(x;t) =$$
$$\sum_x U(x)\frac{d}{dt}\log(p(x;t))\,p(x;t) = \sum_x \left(U(x) - \mathbb{E}_{p(t)}[U]\right)\frac{d}{dt}\log(p(x;t))\,p(x;t) =$$
$$E_t\left[\left(U - \mathbb{E}_{p(t)}[U]\right)\frac{d}{dt}\log(p(t))\right] = \left\langle U - \mathbb{E}_{p(t)}[U], Dp(t)\right\rangle_{p(t)}$$

We observe that the scalar product above is the scalar product on $B_{p(t)}$ because $U \mapsto U - \mathbb{E}_p[U]$ takes values in B_p.

Remark 5. Consider the level surface of f at $p_0 \in \Delta^\circ(\Omega)$, that is $\{p \in \Delta^\circ(\Omega) | f(p) = f(p_0)\}$, and assume $\nabla f(p_0) \neq 0$. Then for each curve through p_0, $I \mapsto p(t)$ with $p(0) = p_0$, such that $f(p(t)) = f(p(0))$, we have

$$0 = \frac{d}{dt}f(p(t))\Big|_{t=0} = \langle \nabla f(p(t)), Dp(t)\rangle_{p(t)}\Big|_{t=0} = \langle \nabla f(p_0), Dp(t_0)\rangle_{p_0}$$

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

Remark 6. If the function $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$ extends to a C^1 funtion on an open subset of $\mathbb{R}^\Omega$, then we can compute the statistical gradient via the ordinary gradient in the geometry of $\mathbb{R}^\Omega$, namely $\operatorname{grad} f(p) = \left(\frac{\partial}{\partial p(x)}f(p)\colon x \in \Omega\right)$. In fact,

$$\frac{d}{dt}f(p(t)) = \sum_{x\in\Omega}\frac{\partial}{\partial p(x)}f(p(x;t)\colon x \in \Omega)\frac{d}{dt}p(x;t) =$$
$$\langle \operatorname{grad} f(p(t)), Dp(x;t)\rangle_{p(t)} = \left\langle \operatorname{grad} f(p(t)) - \mathbb{E}_{p(t)}[\operatorname{grad} f(p(t))], Dp(t)\right\rangle_{p(t)} =$$
$$\langle \nabla f(p(t)), Dp(t)\rangle_{p(t)} \ ,$$

where $\operatorname{grad} f(p(t)) - \mathbb{E}_{p(t)}[\operatorname{grad} f(p(t))]$ is the projection of $\operatorname{grad} f(p(t))$ onto B_p with respect to the scalar product $\langle ., . \rangle_{p(t)}$. Note that the statistical gradient is zero if, and only if, the ordinary gradient is constant.

Definition 2 (Flow)**.**

(1) *Given a vector field F, the trajectories along the vector field are the solution of the differential equation*

$$\frac{D}{dt}p(t) = F(p(t)) \quad \textit{or} \quad \frac{d}{dt}p(t) = p(t)F(p(t)) \ .$$

(2) *The* flow *is a mapping $S\colon \Delta^\circ(\Omega) \times \mathbb{R}_{>0} \ni (p,t) \mapsto S(p,t) \in \Delta^\circ(\Omega)$ such that $S(p,0) = p$ and $t \mapsto S(p,t)$ is a trajectory along F.*

(3) *Given $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$ with statistical gradient $p \mapsto (p, \nabla f(p)) \in T\Delta^\circ(\Omega)$ a solution of the s-*gradient flow equation*, $s = \pm 1$, starting at $p_0 \in \Delta^\circ(\Omega)$ at time t_0 is a curve $I \mapsto (p(t), Dp(t)) \in T\Delta^\circ(\Omega)$ such that $Dp(t) = s\nabla f(p(t))$, $t \in I$, and such that $p(t_0) = p_0$.*

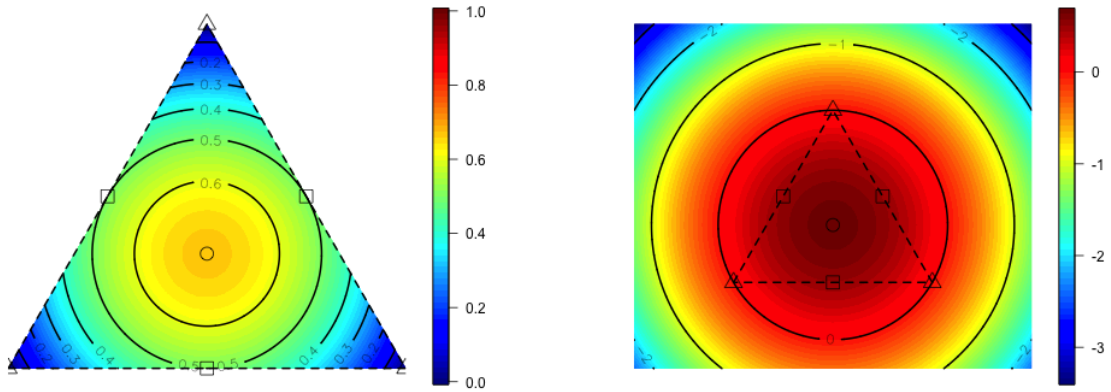

FIGURE 2. The FRAC function on the 3-point simplex (left) and extended to the affine space (right). Note difference between the flow of the ordinary gradient and the flow of the statistical gradient.

(4) *The* s-gradient flow *is a mapping*

$$\Delta^\circ(\Omega) \times I \ni (p,t) \mapsto S(p;t) \in \Delta^\circ(\Omega) \ ,$$

$0 \in I$, *such that* $t \mapsto S(p,t)$ *is the solution of the gradient flow equation starting at* p *at time 0,*

$$\frac{D}{dt} S(p,t) = \nabla f(S(p,t)), \quad t \in I \text{ and } S(p,0) = p \ .$$

Remark 7. If we have a solution of the $+1$ gradient flow $Dp(t) = \nabla f(p(t))$, $t \in I$, then

$$\frac{d}{dt} f(p(t)) = \langle \nabla f(p(t)), Dp(t) \rangle_{p(t)} = \|\nabla f(p(t))\|^2_{p(t)} = \|Dp(t)\|^2_{p(t)} \ .$$

It follows that

$$f(p(t)) - f(p_0) = \int_0^t \|\nabla f(p(t))\|^2_{p(t)} \ dt = \int_0^t \|Dp(t)\|^2_{p(t)} \ dt \ .$$

Assume that $p \mapsto f(p)$ is bounded and $t \mapsto \|\nabla f(p(t))\|^2_{p(t)} = \|Dp(t)\|^2_{p(t)}$ is uniformely continuous. Then $\lim_{t\to\infty} f(p(t)) - f(p_0) = \int_0^\infty \|\nabla f(p(t))\|^2_{p(t)} \ dt = \int_0^\infty \|Dp(t)\|^2_{p(t)} \ dt$ is finite, hence $\lim_{t\to\infty} \|\nabla f(p(t))\|_{p(t)} = \lim_{t\to\infty} \|Dp(t)\|_{p(t)} = 0$ (Barbalat lemma). If $p \mapsto \|\nabla f(p)\|^2_p$ has an isolated zero $\bar{p}$, then $\lim_{t\to\infty} p(t) = \bar{p}$

2.2. **Example: fractionalization.** Consider the function

$$\mathrm{FRAC}(p) = \sum_{x\in\Omega} p(x)(1-p(x)) = 1 - \sum_{x\in\Omega} p(x)^2, \quad p \in \Delta^\circ(\Omega) \ ,$$

see Fig. 2.

Its value is 0 if, and only if, p is a vertex of the simplex, otherwise it is positive. It is an index used in the Social Sciences to measure the fractionalization of the society in

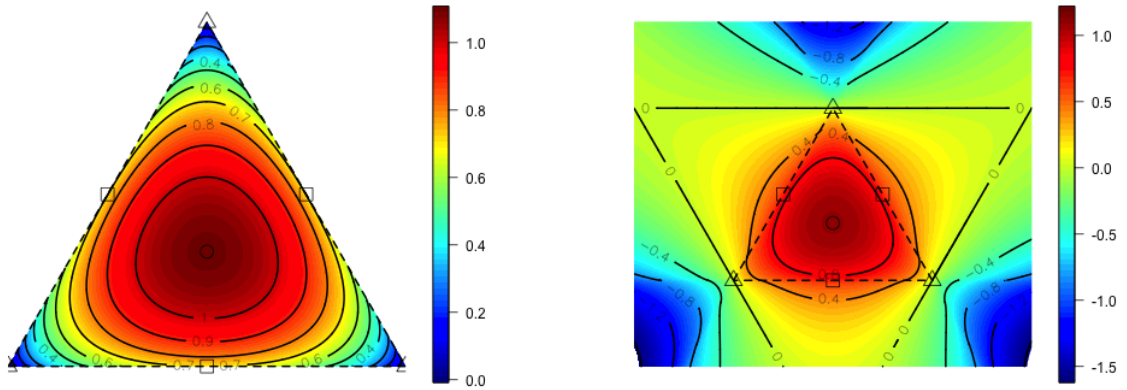

FIGURE 3. The entropy function on the 3-point simplex (left) and extended as $-\sum_{x\in\Omega} p(x)\log|p(x)|$ to the affine space (right). Note the large inverted triangle corresponding to the cases of the type $p(x)=0$, $p(y)+p(z)=0$.

various groups. In fact, the gradient is

$$\nabla \operatorname{FRAC}(p) = \left(-2p(x) + 2\sum_{y\in\Omega} p(y)^2 \middle| x\in\Omega \right)$$

which is 0 on the uniform distribution. The same Eq. extends to $A_1(\Omega)$, where the statistical gradient is 0 at the vertexes of the simplex. On the uniform distribution FRAC takes the maximal value $(n-1)/n$. The $+1$ gradient flow equation is

$$Dp(t) = -2p(x) + 2\sum_{y\in\Omega} p(y)^2 \quad \text{or} \quad \dot{p}(t) = 2p(t)(\sum_{y\in\Omega} p(y)^2 - p(x)) \ .$$

I do not know a closed form solution. The function $V(p)=\sum_{x\in\Omega} p(x)^2$ has statistical gradient $\nabla V(p) = \left(2(p(x) - \sum_{y\in\Omega} p(y)^2) \middle| x\in\Omega\right) = -\nabla \operatorname{FRAC}(p)$, so that

$$\langle \nabla V(p), \nabla \operatorname{FRAC}(p)\rangle_p < 0 \ ,$$

that is V is a Lyapunov function for FRAC.

2.3. **Example: entropy.** The combination of Dp as a definition of the derivative and of the metric $\langle\cdot,\cdot\rangle_{p(t)}$ on the vector fiber provides a computable rule for the gradient. We compute the gradient of the entropy $H(p) = -\mathbb{E}_p[\log p]$, $p\in\Delta^\circ(\Omega)$, see Fig. 3.

Let us compute the gradient:

$$\frac{d}{dt}H(p(t)) =$$

$$-\frac{d}{dt}\sum_{x\in\Omega} p(x;t)\log(p(x;t)) = -\sum_{x\in\Omega}(\log(p(x;t))+1)\dot{p}(x;t) =$$

$$\langle -\log(p(t) - H(p(t)), Dp(t)\rangle_{p(t)} \ ,$$

hence $\nabla H(p) = -\log p - H(p)$. Note that the gradient is zero if, and only if, the distribution is uniform.

Let us solve the flow of the gradient equation

$$Dp(t) = \nabla H(p(t)), \quad p(0) = p_0 \ ,$$

that is

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

It follows

$$\frac{d}{dt} \left(\mathrm{e}^t \log\left(p(t)\right)\right) = -\mathrm{e}^t H(p(t)) \ ,$$

hence

$$\mathrm{e}^t \log\left(p(t)\right) = \log\left(p_0\right) - \int_0^t \mathrm{e}^s H(p(s)) \ ds \ ,$$

so that

$$p(t) = \frac{p_0^{\mathrm{e}^{-t}}}{\exp\left(\int_0^t \mathrm{e}^{-(t-s)} H(p(s)) \ ds\right)}$$

and

$$\exp\left(\int_0^t \mathrm{e}^{-(t-s)} H(p(s))\right) \ ds = \sum_{x \in \Omega} p_0(x)^{\mathrm{e}^{-t}} \ .$$

In conclusion, we have proved that the solution of the gradient flow tends to the probability with maximal entropy, that is to the uniform distribution.

Let us consider the negative gradient flow,

$$Dp(t) = -\nabla H(p(t)), \quad p(0) = p_0 \ ,$$

that is

$$\frac{d}{dt} \log\left(p(t)\right) = \log\left(p(t)\right) + H(p(t)) \ .$$

It follows

$$\frac{d}{dt} \left(\mathrm{e}^{-t} \log\left(p(t)\right)\right) = \mathrm{e}^{-t} H(p(t)) \ ,$$

hence

$$\mathrm{e}^{-t} \log\left(p(t)\right) = \log\left(p_0\right) + \int_0^t \mathrm{e}^{-s} H(p(s)) \ ds \ ,$$

so that

$$p(t) = \frac{p_0^{\mathrm{e}^{t}}}{\exp\left(\int_0^t \mathrm{e}^{(t-s)} H(p(s)) \ ds\right)}$$

and

$$\exp\left(\int_0^t \mathrm{e}^{(t-s)} H(p(s)) \ ds\right) = \sum_{x \in \Omega} p_0(x)^{\mathrm{e}^{t}} \ .$$

In conclusion, it is easy to check that the solution of the negative gradient flow tend to the probability in $\Delta(\Omega)$ which is uniform on the set $\{x \in \Omega | p_0(x) = \max_y p_0(y)\}$.

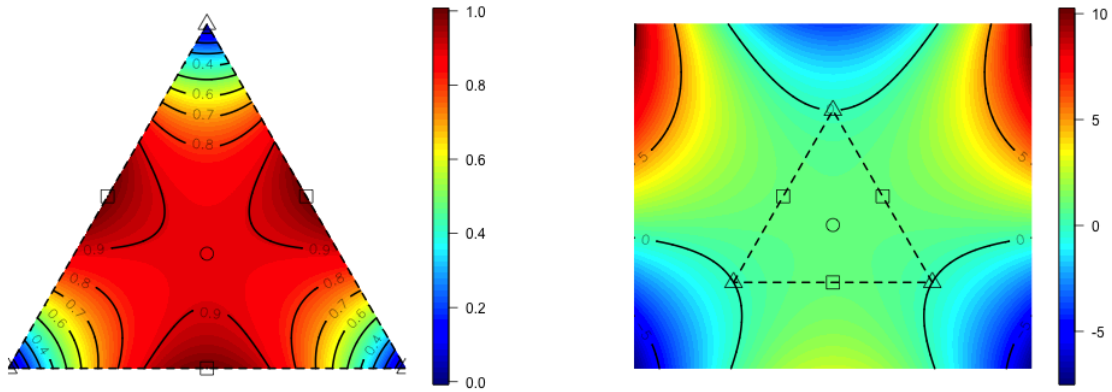

FIGURE 4. Polarization

2.4. **Example: the polarization measure.** M. Roynal-Quenol has defined the index of polarization in her PhD thesis, discussed at LSE on 2001, see [25] to be

$$\mathrm{POL}\colon \Delta_n \ni p \mapsto 1 - 4\sum_{x=0}^{n}\left(\frac{1}{2} - p(x)\right)^2 p(x) = 4\sum_{x=0}^{n} p(x)^2(1-p(x))$$

From the first form it follows that the value is 1 if, and only if, $p(x) = 1/2$ or $p(x) = 0$ for all x, that is p is the uniform distribution on a couple of points, $\binom{n+1}{2} = \frac{n(n+1)}{2}$ cases. Otherwise it is < 1. From the second form, we see it is ≥ 0 and 0 on the vertexes of the simplex only.

Note that we can extend the funtion POL on the affine space

$$A_1(\Omega) = \left\{ q \in \mathbb{R}^{\Omega} \middle| \sum_{x\in\Omega} q(x) = 1 \right\} .$$

See in Fig. 4 the two cases. On such extension it is still true that the function is zero on the vertexes of the simplex $\Delta(\Omega)$, but it is not anymore true that the maximum is reached on the middle points of the edges.

The statistical gradient is the vector

$$\nabla \mathrm{POL}\,(p) = \left(8p(x) - 12p(x)^2 - 8\sum_{y\in\Omega} p(y)^2 + 12\sum_{y\in\Omega} p(y)^3 \middle| x \in \Omega \right)$$

This example shall be discussed again in the context of parameterization.

2.5. **Transports.** For each random variable $U \in B_p$, note that $\mathbb{E}_q\left[U - \mathbb{E}_q\left[U\right]\right] = 0$ and $\mathbb{E}_q\left[\frac{p}{q}U\right] = 0$, so that we can define the following transport mappings.

Definition 3.

(1) *The* e-transport *is the family of linear mappings*

$${}^{+}\mathbb{U}_p^q\colon B_p \ni U \mapsto U - \mathbb{E}_q\left[U\right] \in B_q \ .$$

(2) *The* m-transport *is the family of linear mappings*

$$ {}^{-}\mathbb{U}_p^q \colon B_p \ni U \mapsto \frac{q}{p} U \in B_q \ . $$

(3) *The* h-transport *is the family of linear mappings*

$$ {}^{0}\mathbb{U}_p^q \colon B_p \ni U \mapsto \sqrt{\frac{p}{q}} u - \left(1 + \mathbb{E}_q \left[\sqrt{\frac{p}{q}} \right] \right)^{-1} \left(1 + \sqrt{\frac{p}{q}} \right) \mathbb{E}_q \left[\sqrt{\frac{p}{q}} u \right] \in B_q $$

Proposition 4.

(1) ${}^{+}\mathbb{U}_q^r \, {}^{+}\mathbb{U}_p^q = {}^{+}\mathbb{U}_p^r$, ${}^{-}\mathbb{U}_q^r \, {}^{-}\mathbb{U}_p^q = {}^{-}\mathbb{U}_p^r$, ${}^{0}\mathbb{U}_q^p \, {}^{0}\mathbb{U}_p^q u = u$.
(2) $\left\langle {}^{+}\mathbb{U}_p^q U, V \right\rangle_q = \left\langle U, {}^{-}\mathbb{U}_q^p V \right\rangle_p$.
(3) $\left\langle {}^{+}\mathbb{U}_p^q U, {}^{-}\mathbb{U}_p^q V \right\rangle_q = \langle U, V \rangle_p$.
(4) $\left\langle {}^{0}\mathbb{U}_p^q U, {}^{0}\mathbb{U}_p^q V \right\rangle_q = \langle U, V \rangle_p$.

2.6. **Connections.** Second order geometry is usually based on a notion of covariant derivative or a notion of connection. It leads to interesting applications in optimiziation, such as the use of a Newton method based on the computation of the proper Heffian. Here we restrict to a definition of accelleration based on the use of transports.

Let us compute the accelleration of a curve $I \mapsto p(t)$. The velocity is $t \mapsto (p(t), Dp(t)) = \left(p(t), \frac{d}{dt} \log (p(t)) \right) \in T\Delta^\circ(\Omega)$. The vector $Dp(t)$ has to be checked against an element of $B_{p(t)}$, say ${}^{-}\mathbb{U}_p^{p(t)} v$. We can compute an accelleration as

$$ \begin{aligned} \frac{d}{dt} \left\langle Dp(t), {}^{-}\mathbb{U}_p^{p(t)} v \right\rangle_{p(t)} &= \frac{d}{dt} \left\langle {}^{+}\mathbb{U}_{p(t)}^{p} Dp(t), v \right\rangle_p \\ &= \left\langle \frac{d}{dt} {}^{+}\mathbb{U}_{p(t)}^{p} Dp(t), v \right\rangle_p \\ &= \left\langle {}^{+}\mathbb{U}_p^{p(t)} \frac{d}{dt} {}^{+}\mathbb{U}_{p(t)}^{p} Dp(t), {}^{-}\mathbb{U}_p^{p(t)} v \right\rangle_{p(t)} \end{aligned} $$

The *exponential accelleration* is

$$ \begin{aligned} {}^{+}\mathbb{U}_p^{p(t)} \frac{d}{dt} {}^{+}\mathbb{U}_{p(t)}^{p} Dp(t) &= {}^{+}\mathbb{U}_p^{p(t)} \frac{d}{dt} \left(\frac{\dot{p}(t)}{p(t)} - \mathbb{E}_p \left[\frac{\dot{p}(t)}{p(t)} \right] \right) \\ &= {}^{+}\mathbb{U}_p^{p(t)} \left(\frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} - \mathbb{E}_p \left[\frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} \right] \right) \\ &= \frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} - \mathbb{E}_{p(t)} \left[\frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} \right] \\ &= \frac{\ddot{p}(t)}{p(t)} - (Dp(t))^2 + \mathbb{E}_{p(t)} \left[(Dp(t))^2 \right] \ . \end{aligned} $$

Exponential families have null exponential accelleration: for $p(t) = \exp(tu - \psi(t))\, p$, we have

$$ \begin{aligned} Dp(t) &= u - \dot{\psi}(t) \\ \frac{d}{dt} Dp(t) &= -\ddot{\psi}(t) \\ \mathbb{E}_{p(t)} \left[(Dp(t))^2 \right] &= \ddot{\psi}(t) \end{aligned} $$

We could have computed the accelleration as

$$\begin{aligned}\frac{d}{dt}\left\langle Dp(t), {}^{+}\mathbb{U}_{p}^{p(t)}v\right\rangle_{p(t)} &= \frac{d}{dt}\left\langle {}^{-}\mathbb{U}_{p(t)}^{p}Dp(t), v\right\rangle_{p} \\ &= \left\langle \frac{d}{dt}\,{}^{-}\mathbb{U}_{p(t)}^{p}Dp(t), v\right\rangle_{p} \\ &= \left\langle {}^{-}\mathbb{U}_{p}^{p(t)}\frac{d}{dt}\,{}^{-}\mathbb{U}_{p(t)}^{p}Dp(t), {}^{+}\mathbb{U}_{p}^{p(t)}v\right\rangle_{p(t)}\end{aligned}$$

hence we can compute the *mixture accelleration* as follows

$$\begin{aligned}{}^{-}\mathbb{U}_{p}^{p(t)}\frac{d}{dt}\,{}^{-}\mathbb{U}_{p(t)}^{p}Dp(t) &= \frac{p}{p(t)}\frac{d}{dt}\left(\frac{p(t)}{p}\frac{\dot{p}(t)}{p(t)}\right) \\ &= \frac{\ddot{p}(t)}{p(t)}\ .\end{aligned}$$

In follows that *mixture models have null mixture accelleration.*

We could compute the *Riemannian accelleration* as ${}^{0}\mathbb{U}_{p}^{p(t)}\frac{d}{dt}\,{}^{0}\mathbb{U}_{p(t)}^{p}Dp(t)$. See some developments in [20, 21, 14]

2.7. **Atlases.** We now turn to the explicit introduction of atlases of charts. Each of these chart correspond to a different geometry, in particular to different notion of geodesics.

Definition 5 (Exponential atlas)**.** *For each $p \in \Delta^{\circ}(\Omega)$ we define*

$$s_p\colon T\Delta^{\circ}(\Omega) \ni (q, w) \mapsto \left(\log\frac{q}{p} - \mathbb{E}_p\left[\log\frac{q}{p}\right], {}^{+}\mathbb{U}_{q}^{p}w\right) \in B_p \times B_p$$

Proposition 6 (Exponential atlas)**.**

(1) *If $u = s_p(q)$, then $q = \mathrm{e}^{u - K_p(u)} \cdot p$ with $K_p(u) = \log \mathbb{E}_p\left[\mathrm{e}^{u}\right]$.*
(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*

$$s_{p_1} \circ s_{p_2}^{-1}\colon (u, v) \mapsto \left({}^{+}\mathbb{U}_{p_2}^{p_1}u + \log\frac{p_2}{p_1} - \mathbb{E}_{p_1}\left[\log\frac{p_2}{p_1}\right]\right) \in B_{p_1} \times B_{p_1}$$

(4) *The tangent bundle identifies with the statistical bundle,*

$$\frac{d}{dt}s_p(p(t)) = {}^{+}\mathbb{U}_{p(t)}^{p}Dp(t)\ .$$

Definition 7 (Mixture atlas)**.** *For each $p \in \Delta^{\circ}(\Omega)$ we define*

$$\eta_p\colon T\Delta^{\circ}(\Omega) \ni (q, w) \mapsto \left(\frac{q}{p} - 1, {}^{-}\mathbb{U}_{q}^{p}w\right) \in B_p \times B_p$$

Proposition 8 (Mixture 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*

$$\eta_{p_1} \circ \eta_{p_2}^{-1}\colon (u, v) \mapsto \left((1 + u)\frac{p_2}{p_1} - 1, {}^{-}\mathbb{U}_{p_1}^{p_2}v\right)$$

(4) *The tangent bundle identifies with the statistical bundle,*

$$\frac{d}{dt}\eta_p(p(t)) = {}^{-}\mathbb{U}^p_{p(t)} Dp(t) \ .$$

It is possible to define the Riemannian atlas, see [21].

2.8. **Using parameters.** We still study the geometry of the full simplex, but now we introduce parameters. This section is taken from [22].

Computations are usually performed in a *parametrization*

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

Θ being an open set in $\mathbb{R}^n$. The j-th coordinate curve is obtained by fixing the other $n-1$ components and moving θ_j only. The scores of the j-th coordinate curves are the random variables

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

The sequence $(D_j\pi(\boldsymbol{\theta})\colon j = 1,\dots,n)$ is a vector basis of the tangent space $B_{\pi(\boldsymbol{\theta})}$. The representation of the scalar product in such a basis is

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

Definition 9 (Fisher Information)**.** *The matrix* $I(\boldsymbol{\theta}) = \Big[\langle D_i\pi(\boldsymbol{\theta}), D_j\pi(\boldsymbol{\theta})\rangle_{\pi(\boldsymbol{\theta})}\Big]_{i,i=1}^n$ *is the* Fisher information *matrix.*

If $\boldsymbol{\theta} \mapsto \widetilde{\phi}(\boldsymbol{\theta})$ is the expression in the parameters of a function $\phi\colon \Delta^\circ(\Omega) \to \mathbb{R}$, that is $\widetilde{\phi}(\boldsymbol{\theta}) = \phi(\pi(\boldsymbol{\theta}))$, and $t \mapsto \boldsymbol{\theta}(t)$ is the expression in the parameters of a generic curve $p\colon I \to \Delta^\circ(\Omega)$, then the components of the gradient in (11) are expressed in terms of the ordinary gradient by observing that

$$\frac{d}{dt}\phi(p(t)) = \frac{d}{dt}\widetilde{\phi}(\boldsymbol{\theta}(t)) = \sum_{j=1}^n \frac{\partial}{\partial\theta_j}\widetilde{\phi}(\boldsymbol{\theta}(t))\dot{\theta}_j(t).$$

As $Dp(t) = \sum_j D_j\pi(\boldsymbol{\theta}(t))\dot{\theta}_j(t)$, we obtain from (11)

$$\frac{d}{dt}\phi(p(t)) = \langle\nabla\phi(p(t)), Dp(t)\rangle_{p(t)} = \sum_j \left\langle \nabla\phi(p(t)), D_j\pi(\boldsymbol{\theta}(t))\dot{\theta}_j(t)\right\rangle_{p(t)} . \tag{1}$$

Definition 10 ([2])**.** *The* natural gradient *is a vector* $\widetilde{\nabla}\widetilde{\phi}(\boldsymbol{\theta})$ *whose components are the coordinates of the gradient* $\nabla\widetilde{\phi}(\pi(\boldsymbol{\theta})) \in B_{\pi(\boldsymbol{\theta})}$ *in its* π*-basis, that is*

$$\nabla\phi(\pi(\boldsymbol{\theta})) = \sum_{j=1}^n (\widetilde{\nabla}\widetilde{\phi}(\boldsymbol{\theta}))_j D_j\pi(\boldsymbol{\theta}).$$

By substitution of the expression in (1) *we obtain*

$$\widetilde{\nabla}\widetilde{\phi}(\boldsymbol{\theta}) = \nabla\widetilde{\phi}(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta}). \tag{2}$$

The *common parametrization* of the (flat) simplex $\Delta^\circ(\Omega)$ is the projection on the solid simplex $\Gamma_n = \left\{\boldsymbol{\eta} \in \mathbb{R}^n \middle| 0 < \eta_j, \sum_{j=1}^n \eta_j < 1\right\}$, that is

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

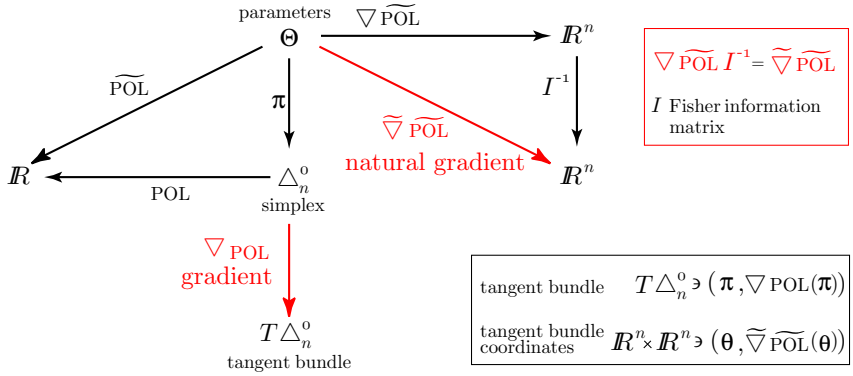

FIGURE 5. Diagram of the action of the natural gradient in a given parametrization $\pi\colon \Theta \to \Delta^\circ(\Omega)$.

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

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

The element (j, h) of the Fisher information matrix is

$$I_{jh}(\boldsymbol{\eta}) = \mathbb{E}_{\pi(\boldsymbol{\eta})}\left[\frac{(X = j) - (X = 0)}{\pi(X;\boldsymbol{\eta})}\frac{(X = h) - (X = 0)}{\pi(X;\boldsymbol{\eta})}\right] =$$

$$\sum_x \pi(x,\boldsymbol{\eta})^{-1}\left((x = j)(j = h) + (x = 0)\right) = \eta_j^{-1}(j = h) + \left(1 - \sum_k \eta_k\right)^{-1}$$

hence

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

As an example we consider $n = 3$. The Fisher information matrix, its inverse and the determinant of the inverse are, respectively,

$$I(\eta_1,\eta_2,\eta_3) =$$

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

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

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

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, we have the following Proposition, whose interest stems from the definition of natural gradient, see Eq. (2).

Proposition 11.

(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 $j = 1, \ldots, n$. 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. □

□

An other parametrization is the *exponential parametrization* based on the exponential family with sufficient statistics $X_j = (X = j)$, $j = 1, \ldots, n$,

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

where

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

See [22] for an illustration of the gradient flow in case of th function POL.

3. Finite State Space: Exponential Families

This section is taken from our paper "Natural Gradient Flow in the Mixture Geometry of a Discrete Exponential Family" to appear in *Entropy*.

Let Ω be a finite set of points $\boldsymbol{x} = (x_1, \dots, x_n)$ and μ the counting measure on Ω. In this case a density $p \in \mathcal{P}_{\geq}$ is a probability function i.e., $p\colon \Omega \to \mathbb{R}_+$ such that $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$.

Given a set $\mathcal{B} = \{T_1, \dots, T_d\}$ of random variables such that, if $\sum_{j=1}^d c_j T_j$ is constant, then $c_1 = \cdots = c_d = 0$. E.g., if $\sum_{\boldsymbol{x}\in\Omega} T_j(\boldsymbol{x}) = 0$, $j = 0, \dots, d$, and the $\mathcal{B}$ is a linear basis. We say that $\mathcal{B}$ is a set of *affinely independent* random variables. If $\mathcal{B}$ is a linear basis, it is affinely independent if and only if $\{1, T_1, \dots, T_d\}$ is a linear basis.

We consider the statistical model $\mathcal{E}$ whose elements are uniquely identified by the natural parameters $\boldsymbol{\theta}$ in the exponential family with sufficient statistics $\mathcal{B}$, namely

$$p_{\boldsymbol{\theta}} \in \mathcal{E} \quad \Leftrightarrow \quad \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{d} \theta_i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^d \ ,$$

see [4].

The proper convex function $\psi\colon \mathbb{R}^d$,

$$\boldsymbol{\theta} \mapsto \psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x}\in\Omega} \mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{T}(\boldsymbol{x})} = \boldsymbol{\theta} \cdot \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\boldsymbol{T}\right] - \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\log\left(p_{\boldsymbol{\theta}}\right)\right]$$

is the *cumulant generating function* of the sufficient statistics $\boldsymbol{T}$, in particular,

$$\nabla\psi(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{T}\right], \quad \operatorname{Hess}\psi(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}\left(\boldsymbol{T}, \boldsymbol{T}\right) \ .$$

Moreover, the entropy of $p_{\boldsymbol{\theta}}$ is

$$H(p_{\boldsymbol{\theta}}) = -\mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\log\left(p_{\boldsymbol{\theta}}\right)\right] = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \nabla\psi(\boldsymbol{\theta}) \ .$$

The mapping $\nabla\psi$ is 1-to-1 onto the interior M° of the *marginal polytope*, that is the convex span of the values of the sufficient statistics $M = \{\boldsymbol{T}(\boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$. Note that no extra condition is required because on a finite state space all random variables are bounded. Nonetheless, even in this case the proof is not trivial, see [4].

Convex conjugation applies, see [27, §25], with the definition

$$\psi_*(\boldsymbol{\eta}) = \sup\left\{\boldsymbol{\theta} \in \mathbb{R}^d \middle| \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\right\}, \quad \boldsymbol{\eta} \in \mathbb{R}^d \ .$$

The concave function $\boldsymbol{\theta} \mapsto \boldsymbol{\eta} \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta})$ has divergence mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\eta} - \nabla\psi(\boldsymbol{\theta})$ and the equation $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$ has solution if and only if $\boldsymbol{\eta}$ belongs to the interior M° of the marginal polytope. The restriction $\phi = \psi_*|_{M^\circ}$ is the *Legendre conjugate* of ψ, and it is computed by

$$\phi\colon M^\circ \ni \boldsymbol{\eta} \mapsto \in (\nabla\psi)^{-1}(\boldsymbol{\eta}) \cdot \boldsymbol{\eta} - \psi \circ (\nabla\psi)^{-1}(\boldsymbol{\eta}) \in \mathbb{R} \ .$$

The Legendre conjugate ϕ is such that $\nabla\phi = (\nabla\psi)^{-1}$ and it provides an alternative parameterization of $\mathcal{E}$ with the so called *extectation* or *mixture* parameter $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$,

$$p_{\boldsymbol{\eta}} = \exp\left((\boldsymbol{T} - \boldsymbol{\eta}) \cdot \nabla\phi(\boldsymbol{\eta}) + \phi(\boldsymbol{\eta})\right). \tag{3}$$

While in the $\boldsymbol{\theta}$-parameters the entropy is $H(p_{\boldsymbol{\theta}}) = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta}\cdot\nabla\psi(\boldsymbol{\theta})$, in the $\boldsymbol{\eta}$-parameters the ϕ function gives the the negative of the entropy: $-H(p_{\boldsymbol{\eta}}) = \mathbb{E}_{p_{\boldsymbol{\eta}}}\left[\log p_{\boldsymbol{\eta}}\right] = \phi(\boldsymbol{\eta})$.

Proposition 12.

(1) $\operatorname{Hess}\phi(\boldsymbol{\eta}) = \left(\operatorname{Hess}\psi(\boldsymbol{\theta})\right)^{-1}$ *when* $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$.

(2) *The Fisher information matrix of the statistical model given by the exponential family in the $\boldsymbol{\theta}$-parameters is* $I_e(\boldsymbol{\theta}) = \mathrm{Cov}_{p_{\boldsymbol{\theta}}}(\nabla \log p_{\boldsymbol{\theta}}, \nabla \log p_{\boldsymbol{\theta}}) = \mathrm{Hess}\,\psi(\boldsymbol{\theta})$.
(3) *The Fisher information matrix of the statistical model given by the exponential family in the $\boldsymbol{\eta}$-parameters is* $I_m(\boldsymbol{\theta}) = \mathrm{Cov}_{p_{\boldsymbol{\eta}}}(\nabla \log p_{\boldsymbol{\eta}}, \nabla \log p_{\boldsymbol{\eta}}) = \mathrm{Hess}\,\phi(\boldsymbol{\eta})$.

Proof. Derivation of the equality $\nabla\phi = (\nabla\psi)^{-1}$ gives the first item. The second item is a property of the cumulant generating function ψ. The third item follows from Eq. (3). □

3.1. **Statistical manifold.** The exponential family $\mathcal{E}$ is an elementary manifold in either the $\boldsymbol{\theta}$- or the $\boldsymbol{\eta}$-parameterization, named respectively *exponential* of *mixture* parameterization. We discuss now the proper definition of the tangent bundle $T\mathcal{E}$.

Definition 13 (Velocity)**.** *If $I \ni t \mapsto p_t$, I open interval, is a differentiable curve in $\mathcal{E}$, then its* velocity vector *is identified with its Fisher score:*

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log(p_t) \ .$$

The capital-D notation is taken from differential geometry, see the classical monograph [7].

Definition 14 (Tangent space)**.** *In the expression of the curve by the exponential parameters, the velocity is*

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log(p_t) = \frac{d}{dt}\left(\boldsymbol{\theta}(t)\cdot\boldsymbol{T} - \psi(\boldsymbol{\theta}(t))\right) = \dot{\boldsymbol{\theta}}(t)\cdot\left(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}(t)}[\boldsymbol{T}]\right), \tag{4}$$

that is it equals the statistics whose coordinates are $\dot{\boldsymbol{\theta}}(t)$ in the basis of the sufficient statistics centered at p_t. As a consequence, we identify the tangent space *at each $p \in \mathcal{E}$ with the vector space of centered sufficient statistics, that is*

$$T_p\mathcal{E} = \mathrm{Span}\left(T_j - \mathbb{E}_p[T_j] | j = 1, \ldots, d\right) \ .$$

In the mixture parameterization of Eq. (3) the computation of the velocity is

$$\begin{aligned}\frac{D}{dt}p(t) = \frac{d}{dt}\log(p_t) = \frac{d}{dt}\left(\nabla\phi(\boldsymbol{\eta}(t))\cdot(\boldsymbol{T} - \boldsymbol{\eta}(t)) + \phi(\boldsymbol{\eta}(t))\right) = \\ \left(\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\dot{\boldsymbol{\eta}}(t)\right)\cdot(\boldsymbol{T} - \boldsymbol{\eta}(t)) = \dot{\boldsymbol{\eta}}(t)\cdot\left[\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\,(\boldsymbol{T} - \boldsymbol{\eta}(t))\right].\end{aligned} \tag{5}$$

The last equality provides the interpretation of $\dot{\boldsymbol{\eta}}(t)$ as the coordinate of the velocity in the *conjugate* vector basis $\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\,(\boldsymbol{T} - \boldsymbol{\eta}(t))$, that is the basis of velocities along the $\boldsymbol{\eta}$ coordinates.

In conclusion, the first order geometry is characterized as follows.

Definition 15 (Tangent bundle $T\mathcal{E}$)**.** *The* tangent space *at each $p \in \mathcal{E}$ is a vector space of random variables $T_p\mathcal{E} = \mathrm{Span}\left(T_j - \mathbb{E}_p[T_j] | j = 1, \ldots, d\right)$ and the* tangent bundle *$T\mathcal{E} = \{(p, V) | p \in \mathcal{E}, V \in T_p\mathcal{E}\}$, as a manifold, is defined by the chart*

$$T\mathcal{E} \ni \left(\mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{T} - \psi(\boldsymbol{\theta})}, \boldsymbol{v}\cdot(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])\right) \mapsto (\boldsymbol{\theta}, \boldsymbol{v}). \tag{6}$$

Proposition 16.

(1) *If* $V = \boldsymbol{v} \cdot (\boldsymbol{T} - \boldsymbol{\eta}) \in T_{p_{\boldsymbol{\eta}}}\mathcal{E}$*, then* V *is represented in the conjugate basis as*

$$(7)\quad V = \boldsymbol{v} \cdot (\boldsymbol{T} - \boldsymbol{\eta}) = \boldsymbol{v} \cdot (\operatorname{Hess}\phi(\boldsymbol{\eta}))^{-1} \operatorname{Hess}\phi(\boldsymbol{\eta})\,(\boldsymbol{T} - \boldsymbol{\eta}) = \left((\operatorname{Hess}\phi(\boldsymbol{\eta}))^{-1}\,\boldsymbol{v}\right) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta})\,(\boldsymbol{T} - \boldsymbol{\eta})\,.$$

(2) *The mapping* $(\operatorname{Hess}\phi(\boldsymbol{\eta}))^{-1}$ *maps the coordinates* $\boldsymbol{v}$ *of a tangent vector* $V \in T_{p_{\boldsymbol{\eta}}}\mathcal{E}$ *with respect to the basis of centered sufficient statistics to the coordinates* $\boldsymbol{v}^*$ *with respect to the conjugate basis.*

(3) *In the* $\boldsymbol{\theta}$*-parameters the transformation is* $\boldsymbol{v} \mapsto \boldsymbol{v}^* = \operatorname{Hess}\psi(\boldsymbol{\theta})\boldsymbol{v}$*.*

Remark 8. In the finite state space case it is not necessary to go on to the formal construction of a dual tangent bundle because all finite dimensional vector spaces are isomorphic. However, this step is compulsolry in the infinite state space case, as it was done in [21]. Moreover, the explicit construction of natural connections and natural parallel transports of the tangent and dual tangent bundle is unavoidable when considering the second order calculus as it was done in [21, 17] in order to compute Hessians and implement Newton methods of optimization. However, the scope of the present paper is restricted to a basic study of gradient flows, hence from now on we focus on the Riemannian structure and disregard all second order topics.

Proposition 17 (Riemannian metric)**.** *The tangent bundle has a* Riemannian structure *with the natural scalar product of each* $T_p\mathcal{E}$*,* $\langle V, W\rangle_p = \mathbb{E}_p[VW]$*. In the basis of sufficient statistics the metric is expressed by the* Fisher information matrix $I(p) = \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$*, while in the conjugate basis it is expressed by the inverse Fisher matrix* $I^{-1}(p)$*.*

Proof. In the basis of the sufficient statistics, $V = \boldsymbol{v} \cdot (\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])$, $W = \boldsymbol{w} \cdot (\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])$, so that

$$(8)\qquad \langle V, W\rangle_p = \boldsymbol{v}'\mathbb{E}_p\left[(\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])\,(\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])'\right]\boldsymbol{w} = \boldsymbol{v}'\operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})\,\boldsymbol{w} = \boldsymbol{v}' I(p)\boldsymbol{w}\ ,$$

where $I(p) = \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$ is the *Fisher information matrix.*

If $p = p_{\boldsymbol{\theta}} = p_{\boldsymbol{\eta}}$, the conjugate basis at p is

$$(9)\qquad \operatorname{Hess}\phi(\boldsymbol{\eta})(\boldsymbol{T} - \boldsymbol{\eta}) = \operatorname{Hess}\psi(\boldsymbol{\theta})^{-1}(\boldsymbol{T} - \nabla\phi(\boldsymbol{\theta})) = I^{-1}(p)(\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}]),$$

so that for elements of the tangent space expressed in the conjugate basis we have $V = \boldsymbol{v}^* \cdot I^{-1}(p)\,(\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])$, $W = \boldsymbol{w}^* \cdot I^{-1}(p)\,(\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])$, thus

$$(10)\qquad \langle V, W\rangle_p = \boldsymbol{v}^{*\prime}\mathbb{E}_p\left[I^{-1}(p) \cdot (\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])\,(\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}])'\,I^{-1}(p)\right]\boldsymbol{w}^* = \boldsymbol{v}^{*\prime} I^{-1}(p)\boldsymbol{w}^*.$$

□

3.2. **Gradient.** For each C^1 real function $F\colon \mathcal{E} \to \mathbb{R}$, its gradient is defined by taking the derivative along a C^1 curve $I \mapsto p(t)$, $p = p(0)$, and writing it in the Riemannian metrics,

$$(11)\qquad \frac{d}{dt}\hat{F}(\boldsymbol{\theta}(t))\bigg|_{t=0} = \left\langle \nabla F(p), \frac{D}{dt}p(t)\bigg|_{t=0}\right\rangle_p, \quad \nabla F(p) \in T_p\mathcal{E}\ .$$

If $\boldsymbol{\theta} \mapsto \hat{F}(\boldsymbol{\theta})$ is the expression of F in the parameter $\boldsymbol{\theta}$, and $t \mapsto \boldsymbol{\theta}(t)$ is the expression of the curve, then $\frac{d}{dt}\hat{F}(\boldsymbol{\theta}(t)) = \nabla\hat{F}(\boldsymbol{\theta}(t)) \cdot \dot{\boldsymbol{\theta}}(t)$ so that at $p = p_{\boldsymbol{\theta}(0)}$, with velocity $V =$

$\frac{D}{dt}p(t)\big|_{t=0} = \dot{\boldsymbol{\theta}}(0) \cdot (\boldsymbol{T} - \nabla\psi(\boldsymbol{\theta}(0))$, so that we obtain the celebrated Amari's *natural gradient* of [2]:

$$\langle \nabla F(p), V\rangle_p = \left(\operatorname{Hess}\psi(\boldsymbol{\theta}(0))^{-1}\nabla\hat{F}(\boldsymbol{\theta}(0)\right)' \operatorname{Hess}\psi(\boldsymbol{\theta}(0))\dot{\boldsymbol{\theta}}(0) \ . \tag{12}$$

If $\boldsymbol{\eta} \mapsto \check{F}(\boldsymbol{\eta})$ is the expression of F in the parameter $\boldsymbol{\eta}$, and $t \mapsto \boldsymbol{\eta}(t)$ is the expression of the curve, then $\frac{d}{dt}\check{F}(\boldsymbol{\eta}(t)) = \nabla\check{F}(\boldsymbol{\eta}(t)) \cdot \dot{\boldsymbol{\eta}}(t)$ so that at $p = p_{\boldsymbol{\eta}(0)}$, with velocity $V = \frac{d}{dt}\log(p(t))\big|_{t=0} = \dot{\boldsymbol{\eta}}(0) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta}(0))(\boldsymbol{T} - \boldsymbol{\eta}(0))$,

$$\langle \nabla F(p), V\rangle_p = (\operatorname{Hess}\phi(\boldsymbol{\eta}(0))^{-1}\nabla\hat{F}(\boldsymbol{\eta}(0))' \operatorname{Hess}\phi(\boldsymbol{\eta}(0))\dot{\boldsymbol{\eta}}(0). \tag{13}$$

We sumarize all notions od gradient in the following definition.

Definition 18 (Gradients)**.**

(1) *The random variable* $\nabla F(p)$ *uniquely defined by Eq.* (11) *is called* statistical gradient *of* F *at* p*. The mapping* $\nabla F\colon \mathcal{E} \ni p \mapsto \nabla F(p)$ *is a vector field of* $T\mathcal{E}$*.*

(2) *The vector* $\widetilde{\nabla}\,\hat{F}(\boldsymbol{\theta}) = \operatorname{Hess}\psi(\boldsymbol{\theta})^{-1}\nabla\hat{F}(\boldsymbol{\theta})$ *of* (12) *is the expression of the statistical gradient in the* $\boldsymbol{\theta}$ *in the basis of sufficient statistics, and it is called* Amari's natural gradient*, while* $\nabla\hat{F}(\boldsymbol{\theta})$*, which is the expression in the conjugate basis of the sufficient statistics, is called* Amari's vanilla gradient*.*

(3) *The vector* $\widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})^{-1}\nabla\check{F}(\boldsymbol{\eta})$ *of* (12) *is the expression of the statistical gradient in the* $\boldsymbol{\eta}$ *parameter and in the conjugate basis of sufficient statistics, and it is called* Amari's natural gradient*, while* $\nabla\check{F}(\boldsymbol{\eta})$*, which is the expression in the basis of sufficient statistics, is called* Amari's vanilla gradient*.*

Given a vector field of $\mathcal{E}$ i.e., a mapping G defined on $\mathcal{E}$ such that $G(p) \in T_p\mathcal{E}$—which is called a *section* of the tangent bundle in the standard differential geometric language—an integral curve from p is a curve $I \ni t \mapsto p(t)$ such that $p(0) = p$ and $\frac{D}{dt}p(t) = G(p(t))$. In the $\boldsymbol{\theta}$ parameters $G(p_{\boldsymbol{\theta}}) = \hat{\boldsymbol{G}}(\boldsymbol{\theta}) \cdot (\boldsymbol{T} - \nabla\psi(\boldsymbol{\theta}))$, so that the differential equation is expressed by $\dot{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{G}}(\boldsymbol{\theta}(t))$. In the $\boldsymbol{\eta}$ parameters, $G(p_{\boldsymbol{\eta}}) = \check{\boldsymbol{G}}(\boldsymbol{\eta}) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta})(\boldsymbol{T} - \boldsymbol{\eta})$ and the differential equation is $\dot{\boldsymbol{\eta}}(t) = \check{\boldsymbol{G}}(\boldsymbol{\eta}(t))$.

Definition 19 (Gradient flow)**.**

- *The* gradient flow *of the real function* $F\colon \mathcal{E}$ *is the flow of the differential equation* $\frac{D}{dt}p(t) = \nabla F(p(t))$ *i.e.* $\frac{d}{dt}p(t) = p(t)\nabla F(p(t))$*.*
- *The expression in the* $\boldsymbol{\theta}$ *parameters is* $\dot{\boldsymbol{\theta}}(t) = \widetilde{\nabla}\,\hat{F}(\boldsymbol{\theta}(t))$*.*
- *The expression in the* $\boldsymbol{\eta}$ *parameters is* $\dot{\boldsymbol{\eta}}(t) = \widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}(t))$*.*

The cases of gradient computation we have discussed above are just a special case of a generic argument. Let us brefly study the gradient flow in a general chart $f\colon \boldsymbol{\zeta} \mapsto p_{\boldsymbol{\zeta}}$. Consider the change of of parametrization from $\boldsymbol{\zeta}$ to $\boldsymbol{\theta}$,

$$\boldsymbol{\zeta} \mapsto p_{\boldsymbol{\zeta}} \mapsto \boldsymbol{\theta}(p_{\boldsymbol{\zeta}}) = I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}}) \ ,$$

and denote the Jacobian matrix of the parameters' change by $J(\boldsymbol{\zeta})$. We have

$$\begin{aligned}\log p_{\boldsymbol{\zeta}} &= \boldsymbol{T}\cdot\boldsymbol{\theta}(\boldsymbol{\zeta}) - \psi(\boldsymbol{\theta}(\boldsymbol{\zeta})) \\ &= \boldsymbol{T}\cdot I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}}) - \psi\left(I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}})\right) \ ,\end{aligned}$$

and the $\boldsymbol{\zeta}$-coordinate basis of the tangent space $T_{p_{\boldsymbol{\zeta}}}\mathcal{E}$ consists of the components of the gradient with respect to $\boldsymbol{\zeta}$,

$$\nabla(\boldsymbol{\zeta} \mapsto \log p_{\boldsymbol{\zeta}}) = J^{-1}(\boldsymbol{\zeta}) \left(\boldsymbol{T} - \mathbb{E}_{p_{\boldsymbol{\zeta}}}[\boldsymbol{T}]\right)$$

It should be noted that in this case the expression Fisher information matrix does not have the form od an Hessian of a potential function. In fact, the case of the exponential and the mixture parameters point to a special structure which is called *Hessian manifold*, see [29].

3.3. **Gradient flow in the mixture geometry.** From now on, we are going to focus on the expression of the gradient flow in the $\boldsymbol{\eta}$ parameters. From Def. 18 we have

$$\widetilde{\nabla}\, \check{F}(\boldsymbol{\eta}) = \operatorname{Hess} \phi(\boldsymbol{\eta})^{-1} \nabla \check{F}(\boldsymbol{\eta}) = \operatorname{Hess} \psi(\nabla \phi(\boldsymbol{\eta})) \nabla \check{F}(\boldsymbol{\eta}) = I(p_{\boldsymbol{\eta}}) \nabla \check{F}(\boldsymbol{\eta}) \ ,$$

where $I(p) = \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$. As $p \mapsto \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$ is the restriction to the simplex of a quadratic function, while $p \mapsto \boldsymbol{\eta}$ is the restriction to the exponential family $\mathcal{E}$ of a linear function, in some cases we can naturally consider the extension of the gradient flow equation outside M°. One notable case is when the function F is a relaxation of a non constant state space function $f \colon \Omega \to \mathbb{R}$, as it is defined in e.g. [15].

Proposition 20. *Let $f \colon \Omega \to \mathbb{R}$ and let $F(p) = \mathbb{E}_p[f]$ be its* relaxation *on $p \in \mathcal{E}$, It follows:*

(1) *$\nabla F(p)$ is the least square projection of f onto $T_p\mathcal{E}$, that is*

$$\nabla F(p) = I(p)^{-1} \operatorname{Cov}_p(f, \boldsymbol{T}) \cdot (\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}]) \ .$$

(2) *The expressions in the exponential parameters $\boldsymbol{\theta}$ are $\widetilde{\nabla}\, \hat{F}(\boldsymbol{\theta}) = (\operatorname{Hess} \psi(\boldsymbol{\theta}))^{-1} \operatorname{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T})$, $\nabla \hat{F}(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T})$, respectively.*

(3) *The expressions in the mixture parameters $\boldsymbol{\eta}$ are $\widetilde{\nabla}\, \check{F}(\boldsymbol{\eta}) = \operatorname{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$ and $\nabla \check{F}(\boldsymbol{\eta}) = \operatorname{Hess} \phi(\boldsymbol{\eta}) \operatorname{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$, respectively.*

Proof. On a generic curve thought p with velocity V, we have

$$\frac{d}{dt} \mathbb{E}_{p(t)}[f] \bigg|_{t=0} = \operatorname{Cov}_p(f, V) = \langle f, V \rangle_p \ .$$

If $V \in T_p\mathcal{E}$ we can orthogonally project f to get $\langle \nabla F, V \rangle_p = \langle (I^{-1}(p) \operatorname{Cov}_p(f, \boldsymbol{T})) \cdot (\boldsymbol{T} - \mathbb{E}_p[\boldsymbol{T}]), V \rangle_p$. □

3.4. **Gradient flow of the expected value function.** Let us breafly recall the behaviour of the gradient flow in the relaxation case. Let $\boldsymbol{\theta}_n$, $n = 1, 2, \dots$, be a minimizing sequence for $\hat{F}$ and let $\bar{p}$ be a limit point of the sequence $(p_{\boldsymbol{\theta}_n})_n$. It follows that $\bar{p}$ has a defective support, in particular $\bar{p} \notin \mathcal{E}$, see [26] and [24]. For a proof along lines coherent with the present paper, see [16, Th. 1]. It is found that the support $\underline{F} \subset \Omega$ is *exposed*, that is $\boldsymbol{T}(\underline{F})$ is a face of the marginal polytope $M = \operatorname{con}\{\boldsymbol{T}(\boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$. In particular, $\mathbb{E}_{\bar{p}}[\boldsymbol{T}] = \bar{\boldsymbol{\eta}}$ belongs to a face of the marginal polytope M. If $\boldsymbol{a}$ is the (interior) orthogonal of the face, that is $\boldsymbol{a} \cdot \boldsymbol{T}(\boldsymbol{x}) + b \geq 0$ for all $\boldsymbol{x} \in \Omega$ and $\boldsymbol{a} \cdot \boldsymbol{T}(\boldsymbol{x}) + b = 0$ on the exposed set, then $\boldsymbol{a} \cdot (\boldsymbol{T}(\boldsymbol{x}) - \bar{\boldsymbol{\eta}}) = 0$ on the face, so that $\boldsymbol{a} \cdot \operatorname{Cov}_{\bar{p}}(f, \boldsymbol{T}) = 0$. If we extend the mapping $\boldsymbol{\eta} \mapsto \operatorname{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$ on the closed marginal polytope M to be the limit of the vector field of the gradient on the faces of the marginal polytope, we expect to see that such a vector field is tangent to the faces.

4. Gaussian Models

This matherial is taken from a conference poster presented by G.P. at the 2014 SIS conerence.

Gaussian models form an exponential family with special analytical features. The following notes are intended to underline some of these special feature namely the connection between differentiation with respect to parameters and differentiation with respect to the space variables.

The geometry of the multivariate Gaussian model $\mathrm{N}(\vec{\mu}, \Sigma)$ has been studied in detail by Skovgaard in [30], where normal densities are parameterised by the mean parameter $\vec{\mu}$ and the covariance matrix Σ, and the relevant Riemannian geometry is based on an explicit form of the Fisher information.

We review Hermite polynomials from [18, ch V]. If $Z \sim \nu_1 = \mathrm{N}(0,1)$ and f, g real smooth functions, then $\mathbb{E}\,[df(Z)g(Z)] = \mathbb{E}\,[f(Z)\delta g(Z)]$ where $df(x) = f'(x)$ and $\delta g = xg(x) - g'(x)$ is the Stein operator. The Hermite polynomial of order $n = 0, 1, \dots$ is $H_n(x) = \delta^n 1$ e.g., $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, ... It follows that each H_n is a monic polynomial of degree n, $dH_n = nH_{n-1}$, $\mathbb{E}\,[H_n(Z)H_m(Z)] = 0$ for $n \neq m$, $\mathbb{E}\,[H_n(Z)^2] = n!$. The sequence $H_n/\sqrt{n!}$, $n = 0, 1, \dots$ is a complete orthonormal basis of $L^2(\nu_1)$. In dimension d, for each multi-index α, we define $H_\alpha(\vec{(x)}) = \prod_{i=1}^d H_{\alpha_i}(x_i)$ to get an orthogonal basis of $L^2(\nu_d)$, $\nu_d = \mathrm{N}_d(\vec{0}, I_d)$. If we define $d^\alpha = \prod_{i=1}^d d_{x_i}^{\alpha_i}$, $\delta^\alpha = \prod_{i=1}^d \delta_{x_i}^{\alpha_i}$, we have for functions $f, g \colon \mathbb{R}^d \to \mathbb{R}$ and $\vec{Z} \sim \nu_d$ that $\mathbb{E}\left[d^\alpha f(\vec{Z})g(\vec{Z})\right] = \mathbb{E}\left[f(\vec{Z})\delta^\alpha g(\vec{Z})\right]$. If f is infinitely differentiable, then its Fourier-Hermite expansion is $\sum_\alpha \mathbb{E}\left[d^\alpha f(\vec{Z})\right] H_\alpha(\vec{Z})/\alpha!$. Sometimes it is convenient to use $\widetilde{H}_\alpha = H_\alpha/\alpha!$.

4.1. **Gaussian model in the Hermite basis.** Given a vector of means $\boldsymbol{\mu} \in \mathbb{R}^m$ and and a full-rank covariance matrix $\Sigma \in S_m^+$, with $\Sigma = [\sigma_{ij}]$ and $\Sigma^{-1} = [\sigma^{ij}]$, the exponent $-(1/2)(\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ in the Gaussian density $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ can be written

$$-\frac{1}{2}\left(\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} + \mathrm{Tr}\left(\Sigma^{-1}\right) + 2\sum_i (\boldsymbol{\mu}^t\Sigma^{-1})_i H_i(\boldsymbol{x}) + 2\sum_{i<j}\sigma^{ij}H_{ij}(\boldsymbol{x}) + \sum_i \sigma^{ii}H_{ii}(\boldsymbol{x})\right),$$

where $H_i(\boldsymbol{x}) = H_1(x_i) = x_i$ and $H_{ii}(\boldsymbol{x}) = H_2(x_i) = x_i^2 - 1$ for $i = 1, \dots, m$, and $H_{ij} = H_1(x_i)H_1(x_j) = x_i x_j$ for $1 \leq i < j \leq m$. The likelihood of $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ with respect to the standard Gaussian with density $w(\boldsymbol{x}) = (2\pi)^{-1/2}\exp\left(-\boldsymbol{x}^t\boldsymbol{x}/2\right)$ has exponent

$$-\frac{1}{2}\boldsymbol{\mu}^t\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\right) - \frac{m}{2} + \sum_i (\boldsymbol{\mu}^t\Sigma^{-1})_i H_i(\boldsymbol{x}) - \sum_{i<j}\sigma^{ij}H_{ij}(\boldsymbol{x}) - \sum_i (\sigma^{ii} - 1)\frac{H_{ii}(\boldsymbol{x})}{2}$$

Vice-versa, given $I - \Theta \in S_m^+$ and $\boldsymbol{\theta} \in \mathbb{R}^n$, then

$$p(\boldsymbol{x}; \theta_i, \theta_{ij} \colon i \leq j) = \exp\left(\sum_i \theta_i H_i(\boldsymbol{x}) + \sum_{i<j}\theta_{ij}H_{ij}(\boldsymbol{x}) + \sum_i \theta_{ii}\frac{H_{ii}(\boldsymbol{x})}{2} - \psi(\theta_i, \theta_{ij} \colon i \leq j)\right) w(\boldsymbol{x}) \tag{14}$$

is the multivariate Gaussian density with $\Sigma^{-1}\boldsymbol{\mu} = \boldsymbol{\theta} = (\theta_i \colon i = 1, \dots, n)$, $I - \Sigma^{-1} = \Theta$ with upper entries $(\theta_{ij} \colon i < j)$, and cumulant generating function

$$\psi(\theta_i, \theta_{ij} \colon i \le j) = \frac{1}{2}\boldsymbol{\theta}^t (I - \Theta)^{-1}\boldsymbol{\theta} - \frac{1}{2}\operatorname{Tr}(\Theta) - \frac{1}{2}\log\det(I - \Theta). \tag{15}$$

In Eq. (14) the Gaussian model is presented as an exponential family with natural parameters $(\theta_i \colon i = 1, \dots, m; \theta_{ij} \colon 1 \le i \le j \le m)$ in the open convex set $\mathbb{R}^n \times (I + S_m^-)$ and w-orthogonal sufficient statistics. From $(\partial/\partial\theta_i)\boldsymbol{\theta} = \boldsymbol{e}_i$, $(\partial/\partial\theta_{ij})\Theta = E^{ij}$ and Eq. (15) we can compute the first derivatives of the cumulant generating function ψ, that is the expected values of the sufficient statistics,

$$\begin{aligned} \frac{\partial}{\partial\theta_i}\psi &= \boldsymbol{\theta}^t (I - \Theta)^{-1}\boldsymbol{e}_i = \mu_i, && (16)\\ \frac{\partial}{\partial\theta_{ij}}\psi &= \frac{1}{2}\operatorname{Tr}\left((I - \Theta)^{-1}E^{ij}\right) + \frac{1}{2}\boldsymbol{\theta}^t (I - \Theta)^{-1}E^{ij}(I - \Theta)^{-1}\boldsymbol{\theta} \\ &= \sigma_{ij} + \mu_i\mu_j, \quad i < j, && (17)\\ \frac{\partial}{\partial\theta_{ii}}\psi &= \frac{1}{2}\operatorname{Tr}\left((I - \Theta)^{-1}E^{ii}\right) + \frac{1}{2}\boldsymbol{\theta}^t (I - \Theta)^{-1}E^{ii}(I - \Theta)^{-1}\boldsymbol{\theta} - \frac{1}{2} \\ &= \frac{1}{2}\left(\sigma_{ii} + \mu_i^2 - 1\right). && (18) \end{aligned}$$

The second derivatives, that is the covariances of the the sufficient statistics, are

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\psi = \boldsymbol{e}_j^t (I - \Theta)^{-1}\boldsymbol{e}_i, \quad \frac{\partial^2}{\partial\theta_i\partial\theta_{jh}}\psi = \boldsymbol{\theta}^t (I - \Theta)^{-1}E^{jk}(I - \Theta)^{-1}\boldsymbol{e}_i,$$

and

$$\begin{aligned} \frac{\partial^2}{\partial\theta_{ij}\partial\theta_{hk}}\psi = \frac{1}{2}\operatorname{Tr}&\left((I - \Theta)^{-1}E^{hk}(I - \Theta)^{-1}E^{ij}\right) \\ &+ \frac{1}{2}\boldsymbol{\theta}^t (I - \Theta)^{-1}E^{hk}(I - \Theta)^{-1}E^{ij}(I - \Theta)^{-1}\boldsymbol{\theta} \\ &+ \frac{1}{2}\boldsymbol{\theta}^t (I - \Theta)^{-1}E^{ij}(I - \Theta)^{-1}E^{hk}(I - \Theta)^{-1}\boldsymbol{\theta}. \end{aligned}$$

This formulæ are to be compared with the expression of the Riemannian metric in [30]. Yo Sheena [28] has a different parameterisation in which the Fisher matrix is diagonal.

We have used up to now standard change-of-parameter computations. We turn now to exploit specific properties of the Hermite system. Let us write $U(\boldsymbol{x}; \boldsymbol{\theta}, \Theta) = \sum_i \theta_i H_i(\boldsymbol{x}) + \sum_{i<j} \theta_{ij} H_{ij}(\boldsymbol{x}) + \sum_i \theta_{ii}\frac{H_{ii}(\boldsymbol{x})}{2}$. The vector space generated by the sufficient statistics $\operatorname{Span}(U_{\boldsymbol{\theta},\Theta}|\boldsymbol{\theta}, \Theta)$ is the space of polynomials up to degree 2 in the variables $X_1, \dots, X_n$ that are centered with respect to w. In the geometrical picture, it is the tangent space at w of the Gaussian model, while the tangent space at $p_{\boldsymbol{\theta},\Theta}$ is generated by the Fisher's scores, i.e. the partial derivatives of the log-density, see the discussion in [21]. We have

$$\begin{aligned} &\partial U(\boldsymbol{x}; \boldsymbol{\theta}, \Theta)/\partial x_i = \\ &\quad \frac{\partial}{\partial x_i}\left(\theta_i H_1(x_i) + H_1(x_i)\sum_{j<i}\theta_{ji}H_1(x_j) + \frac{1}{2}\theta_{ii}H_2(x_i) + H_1(x_i)\sum_{i<j}\theta_{ij}H_1(x_j)\right) \\ &\qquad = \theta_i + \sum_{j<i}\theta_{ji}H_1(x_j) + \theta_{ii}H_1(x_i) + \sum_{i<j}\theta_{ij}H_1(x_j) \end{aligned}$$

and $\partial^2 U(\boldsymbol{x};\boldsymbol{\theta},\Theta)/\partial x_i \partial x_j = \theta_{ij}$. In matrix form, the basic relation between parameters of the Gaussian model and Hermite polynomials is

$$\nabla_{\boldsymbol{x}} U(\boldsymbol{x};\boldsymbol{\theta},\Theta) = \boldsymbol{\theta} + \Theta\boldsymbol{x}, \quad \mathrm{Hess}_{\boldsymbol{x}}\, U(\boldsymbol{x};\boldsymbol{\theta},\Theta) = \Theta. \tag{19}$$

Let us write the expectation parameters as $\eta_i = \mathbb{E}_{\boldsymbol{\theta},\Theta}[H_i]$, $\eta_{ij} = \mathbb{E}_{\boldsymbol{\theta},\Theta}[H_{ij}]$, $i < j$, $\eta_{ii} = \mathbb{E}_{\boldsymbol{\theta},\Theta}[H_{ij}]/2$, and $\mathbb{E}_{\mathbf{0},I} = \mathbb{E}$. We can compute the η's as moments, instead of derivatives of the cumulant generating function. From $H_i = \delta_i 1$,

$$\begin{aligned}\eta_i = \mathbb{E}\left[H_i \mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right] &= \mathbb{E}\left[\partial_i \mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right] \\ &= \mathbb{E}\left[(\theta_i + \sum_j \theta_{ij} H_j)\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right] = \theta_i + \sum_j \theta_{ij}\eta_j,\end{aligned}$$

or $\boldsymbol{\eta} = \boldsymbol{\theta} + \Theta\boldsymbol{\eta}$, $\boldsymbol{\eta} = (I-\Theta)^{-1}\boldsymbol{\theta}$, cfr. Eq. (16). For η_{ij} we need

$$\begin{aligned}\partial_i\partial_j \mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)} &= \left(\theta_{ij} + (\theta_i + \sum_h \theta_{ih}H_h)(\theta_j + \sum_k \theta_{jk}H_k)\right)\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)} \\ &= \left(\theta_{ij} + \theta_i\theta_j + \sum_h(\theta_i\theta_{jh} + \theta_j\theta_{ih})H_h + \sum_{h,k}\theta_{ik}\theta_{jh}H_hH_k\right)\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}.\end{aligned}$$

From $H_{ij} = \delta^i\delta_j 1$, $\mathbb{E}_{\boldsymbol{\theta},\Theta}[H_hH_k] = \eta_{hk}$ if $h \neq k$, $\mathbb{E}_{\boldsymbol{\theta},\Theta}[H_h^2] = 2\eta_{hh} + 1$, we obtain

$$\eta_{ij} = \theta_{ij} + \theta_i\theta_j + \sum_h(\theta_i\theta_{jh} + \theta_j\theta_{ih})\eta_h + \sum_{h\neq k}\theta_{ik}\theta_{jh}\eta_{hk} + \sum_h \theta_{ih}\theta_{jh}(2\eta_{hh} + 1),$$

to be compared with Eqs. (17) and (18).

4.2. **Optimisation.** Let $f\colon \mathbb{R}^m \to R$ be a continuous bounded function, with maximum at a point $\boldsymbol{m} \in \mathbb{R}^m$. We define the *relaxed function* $F(\boldsymbol{\theta},\Theta) = \mathbb{E}_{\boldsymbol{\theta},\Theta}[f] = \mathbb{E}\left[f\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right]$. Then $F(\boldsymbol{\theta},\Theta) \leq f(\boldsymbol{m})$ and for each sequence $(\boldsymbol{\theta}_n,\Theta_n)$, $n = 1,2,\ldots$, such that $\lim_{n\to\infty}(I-\Theta_n)^{-1} = \lim_{n\to\infty}\Sigma_n = 0$ and $\lim_{n\to\infty}(I-\Theta_n)^{-1}\boldsymbol{\theta}_n = \lim_{n\to\infty}\boldsymbol{\mu}_n = \boldsymbol{m}$, we have $\lim_{n\to\infty} F(\boldsymbol{\theta}_n,\Theta_n) = f(\boldsymbol{m})$. This remark has been used in Optimisation when the function f is a *black box* that is when no analytic expression is known, but the function can be computed at each point $\boldsymbol{x}$, see for example [19]. In fact, the gradient of the relaxed function has components

$$\begin{aligned}\frac{\partial}{\partial\theta_i}F &= \mathrm{Cov}_{\boldsymbol{\theta},\Theta}(f,H_i), \\ \frac{\partial}{\partial\theta_{ij}}F &= \mathrm{Cov}_{\boldsymbol{\theta},\Theta}(f,H_{ij}), \quad i<j, \\ \frac{\partial}{\partial\theta_{ii}}F &= \frac{1}{2}\mathrm{Cov}_{\boldsymbol{\theta},\Theta}(f,H_{ii}),\end{aligned}$$

so that the direction of steepest ascent at $(\boldsymbol{\theta},\Theta)$ can be learned from a sample of $\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\phi(\boldsymbol{\theta},\Theta)}v$ for example from sample covariances. This method does not require any smoothness in the original function and it is expected to have a better robustness vs local maxima than the ordinary gradient search because a mean values of the function f are used. A reduction of dimensionality is obtained by considering sub-models, for example Θ diagonal.

We note that the gradient of the relaxed function is related with the f, ∇f, Hess f as follows. We have $\mathrm{Cov}_{\boldsymbol{\theta},\Theta}(f, H_i) = \mathbb{E}_{\boldsymbol{\theta},\Theta}[fH_i] - \mathbb{E}_{\boldsymbol{\theta},\Theta}[f]\,\mathbb{E}_{\boldsymbol{\theta},\Theta}[H_i]$ and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta},\Theta}[fH_i] &= \mathbb{E}\left[H_i f \mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right] = \mathbb{E}\left[\partial_i\left(f\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right)\right]\\ &= \mathbb{E}\left[(\partial_i f + f\partial_i U)\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right] = \mathbb{E}_{\boldsymbol{\theta},\Theta}[\partial_i f] + \theta_i \mathbb{E}_{\boldsymbol{\theta},\Theta}[f] + \sum_j \theta_{ij}\mathbb{E}_{\boldsymbol{\theta},\Theta}[fH_j].\end{aligned}$$

If $\boldsymbol{H}_1$ is the vector with components $H_1, \ldots, H_m$,

$$\mathbb{E}_{\boldsymbol{\theta},\Theta}[f\boldsymbol{H}_1] = \mathbb{E}_{\boldsymbol{\theta},\Theta}\left[(I-\Theta)^{-1}\nabla f\right] + \mathbb{E}_{\boldsymbol{\theta},\Theta}[f]\,(I-\Theta)^{-1}\boldsymbol{\theta},$$

so that $\nabla_{\boldsymbol{\theta}} F = \mathbb{E}_{\boldsymbol{\theta},\Theta}[(I-\Theta)^{-1}\nabla f] = \mathbb{E}_{\boldsymbol{\mu},\Sigma}[\Sigma\nabla f]$.

In a similar way, $\mathrm{Cov}_{\boldsymbol{\theta},\Theta}(f, H_{ij}) = \mathbb{E}_{\boldsymbol{\theta},\Theta}[fH_{ij}] - \mathbb{E}_{\boldsymbol{\theta},\Theta}[f]\,\mathbb{E}_{\boldsymbol{\theta},\Theta}[H_{ij}]$ and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta},\Theta}[fH_{ij}] &= \mathbb{E}\left[H_{ij} f \mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right] = \mathbb{E}\left[\partial_i\partial_j\left(f\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right)\right]\\ &= \mathbb{E}\left[\partial_i\left[(\partial_j f + f\partial_j U_{\boldsymbol{\theta},\Theta})\mathrm{e}^{U_{\boldsymbol{\theta},\Theta}-\psi(\boldsymbol{\theta},\Theta)}\right]\right]\\ &= \mathbb{E}_{\boldsymbol{\theta},\Theta}[\partial_i\partial_j f + \partial_i f\partial_j U_{\boldsymbol{\theta},\Theta} + \partial_j f\partial_i U_{\boldsymbol{\theta},\Theta} + f\partial_i\partial_j U_{\boldsymbol{\theta},\Theta} + \partial_i U_{\boldsymbol{\theta},\Theta}\partial_j U_{\boldsymbol{\theta},\Theta}].\end{aligned}$$

Now we can substitute in the equation above $\partial_i U_{\boldsymbol{\theta},\Theta} = \theta_i + \sum_h \theta_{ih}H_h$, $\partial_j U_{\boldsymbol{\theta},\Theta} = \theta_j + \sum_h \theta_{jh}H_h$, $\partial_i\partial_j U_{\boldsymbol{\theta},\Theta} = \theta_{ij}$ and

$$\begin{aligned}\partial_i U_{\boldsymbol{\theta},\Theta}\partial_j U_{\boldsymbol{\theta},\Theta} &= (\theta_i + \sum_h \theta_i h H_h)(\theta_j + \sum_k \theta_{ik}H_k)\\ &= \theta_i\theta_j + \sum_h(\theta_i\theta_{jh} + \theta_j\theta_{ih})H_h + \sum_{h,k}\theta_{ih}\theta_{jk}H_hH_k\\ &= \theta_i\theta_j + \sum_h(\theta_i\theta_{jh} + \theta_j\theta_{ih})H_h + 2\sum_{h<k}\theta_{ih}\theta_{jk}H_{hk} + \sum_h \theta_{ih}\theta_{jh}(H_{hh}+1),\end{aligned}$$

to obtain the required relation. We leave the rest of the computation to the reader.

References

1. Shun-ichi Amari, *Differential-geometrical methods in statistics*, Lecture Notes in Statistics, vol. 28, Springer-Verlag, New York, 1985. MR 86m:62053
2. Shun-Ichi Amari, *Natural gradient works efficiently in learning*, Neural Computation **10** (1998), no. 2, 251–276.
3. Shun-ichi Amari and Hiroshi Nagaoka, *Methods of information geometry*, American Mathematical Society, Providence, RI, 2000, Translated from the 1993 Japanese original by Daishi Harada. MR 1 800 071
4. Lawrence D. Brown, *Fundamentals of statistical exponential families with applications in statistical decision theory*, IMS Lecture Notes. Monograph Series, no. 9, Institute of Mathematical Statistics, 1986. MR MR882001 (88h:62018)
5. A. P. Dawid, *Discussion of a paper by Bradley Efron*, Ann. Statist. **3** (1975), no. 6, 1231–1234.
6. ———, *Further comments on: "Some comments on a paper by Bradley Efron" (Ann. Statist.* **3** *(1975), 1189–1242)*, Ann. Statist. **5** (1977), no. 6, 1249. MR MR0471125 (57 #10863)
7. Manfredo Perdigão do Carmo, *Riemannian geometry*, Mathematics: Theory & Applications, Birkhäuser Boston Inc., Boston, MA, 1992, Translated from the second Portuguese edition by Francis Flaherty. MR 1138207 (92i:53001)
8. Bradley Efron, *Defining the curvature of a statistical problem (with applications to second order efficiency)*, Ann. Statist. **3** (1975), no. 6, 1189–1242, With a discussion by C. R. Rao, Don A. Pierce, D. R. Cox, D. V. Lindley, Lucien LeCam, J. K. Ghosh, J. Pfanzagl, Niels Keiding, A. P. Dawid, Jim Reeds and with a reply by the author. MR MR0428531 (55 #1552)
9. ———, *The geometry of exponential families*, Ann. Statist. **6** (1978), no. 2, 362–376. MR 57 #10890
10. Robert E. Kass and Paul W. Vos, *Geometrical foundations of asymptotic inference*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, 1997, A Wiley-Interscience Publication. MR 1461540 (99b:62032)
11. Serge Lang, *Differential and Riemannian manifolds*, third ed., Graduate Texts in Mathematics, vol. 160, Springer-Verlag, New York, 1995. MR 96d:53001
12. Steffen L. Lauritzen, *Statistical manifolds*, Differential geometry in statistical inference, Institute of Mathematical Statistics Lecture Notes—Monograph Series, 10, Institute of Mathematical Statistics, Hayward, CA, 1987.
13. Hông Vân Lê, *The uniqueness of the Fisher metric as information metric*, ArXiV.1306.1465, 2014.
14. Betrand Lods and Giovanni Pistone, *Information geometry formalism for the spatially homogeneous Boltzmann equation*, arXiv:1502.06774, 2015.
15. Luigi Malagò, Matteo Matteucci, and Giovanni Pistone, *Towards the geometry of estimation of distribution algorithms based on the exponential family*, Proceedings of the 11th workshop on Foundations of genetic algorithms (New York, NY, USA), FOGA '11, ACM, 2011, pp. 230–242.
16. Luigi Malagò and Giovanni Pistone, *A note on the border of an exponential family*, arXiv:1012.0637v1, 2010.
17. ———, *Combinatorial optimization with information geometry: Newton method*, Entropy **16** (2014), 4260–4289.
18. Paul Malliavin, *Integration and probability*, Graduate Texts in Mathematics, vol. 157, Springer-Verlag, New York, 1995, With the collaboration of Hélène Airault, Leslie Kay and Gérard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky. MR MR1335234 (97f:28001a)
19. Y. Ollivier, L. Arnold, A. Auger, and N. Hansen, *Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles*, arXiv:1106.3708, 2011v1; 2013v2.
20. Giovanni Pistone, *Examples of the application of nonparametric information geometry to statistical physics*, Entropy **15** (2013), no. 10, 4042–4065. MR 3130268
21. ———, *Nonparametric information geometry*, Geometric science of information (Frank Nielsen and Frèdèric Barbaresco, eds.), Lecture Notes in Comput. Sci., vol. 8085, Springer, Heidelberg, 2013, First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings, pp. 5–36. MR 3126029
22. Giovanni Pistone and Maria Piera Rogantin, *The gradient flow of the polarization measure. with an appendix*, arXiv:1502.06718, 2015.
23. C. Radhakrishna Rao, *Information and the accuracy attainable in the estimation of statistical parameters*, Bull. Calcutta Math. Soc. **37** (1945), 81–91. MR MR0015748 (7,464a)

24. Johannes Rauh, Thomas Kahle, and Nihat Ay, *Support sets in exponential families and oriented matroid theory*, International Journal of Approximate Reasonoing **52** (2011), no. 2, 613–626, Special Issue Workshop on Uncertainty Processing WUPES09.
25. Marta Reynal-Querol, *Ethnicity, political systems and civil war*, Journal of Conflict Resolution **46** (2002), no. 1, 29–54.
26. Alessandro Rinaldo, Stephen E. Fienberg, and Yi Zhou, *On the geometry of discrete exponential families with application to exponential random graph models*, Electron. J. Stat. **3** (2009), 446–484. MR 2507456 (2010f:62077)
27. R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970. MR MR0274683 (43 #445)
28. Yo Sheena, *Inference on the eigenvalues of the covariance matrix of a multivariate normal distribution–geometrical view–*, ArXiv1211.5733, November 2012.
29. Hirohiko Shima, *The geometry of Hessian structures*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007. MR 2293045 (2008f:53011)
30. Lene Theil Skovgaard, *A Riemannian geometry of the multivariate normal model*, Scand. J. Statist. **11** (1984), no. 4, 211–223. MR 793171 (86m:62102)

L. Malagò: Department of Electrical and Electronic Engineering, Shinshu University, 4-17-1 Wakasato, Nagano 380-8553 Japan
E-mail address: malago@shinshu-u.ac.jp
URL: http://homes.di.unimi.it/~malago/

G. Pistone: de Castro Statistics, Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy
E-mail address: giovanni.pistone@carloalberto.org
URL: www.giannidiorestino.it