FISHER-RAO GEOMETRY OF DIRICHLET DISTRIBUTIONS

ALICE LE BRIGANT, STEPHEN C. PRESTON, AND STÉPHANE PUECHMOREL

Abstract. In this paper, we study the geometry induced by the Fisher-Rao metric on the parameter space of Dirichlet distributions. We show that this space is a Hadamard manifold, i.e. that it is geodesically complete and has everywhere negative sectional curvature. An important consequence for applications is that the Fréchet mean of a set of Dirichlet distributions is uniquely defined in this geometry.

1. Introduction

The differential geometric approach to probability theory and statistics has met increasing interest in the past years, from the theoretical point of view as well as in applications. In this approach, probability distributions are seen as elements of a differentiable manifold, on which a metric structure is defined through the choice of a Riemannian metric. Two very important ones are the Wasserstein metric, central in optimal transport, and the Fisher-Rao metric (also called Fisher information metric), essential in information geometry. Unlike optimal transport, information geometry is foremost concerned with parametric families of probability distributions, and defines a Riemannian structure on the parameter space using the Fisher information matrix [14]. It was Rao who showed in 1945 [26] that the Fisher information could be used to locally define a scalar product on the space of parameters, and interpreted as a Riemannian metric. Later on, Čencov [13] proved that it was the only metric invariant with respect to sufficient statistics, for families with finite sample spaces. This result has been extended more recently to non parametric distributions with infinite support [8, 9].

Information geometry has been used to obtain new results in statistical inference as well as gain insight on existing ones. In parameter estimation for example, Amari [3] shows that conditions for consistency and efficiency of estimators can be expressed in terms of geometric conditions; in the presence of hidden variables, the famous Expectation-Maximisation (EM) algorithm can be described in an entirely geometric manner; and in order to insure invariance to diffeomorphic change of parametrization, the so-called natural gradient [2] can be used to define accurate parameter estimation algorithms [21].

Another important use of information geometry is for the effective comparison and analysis of families of probability distributions. The geometric tools provided by the Riemannian framework, such as the geodesics, geodesic distance and intrinsic mean, have proved useful to interpolate, compare, average or perform segmentation between objects modeled by probability densities, in applications such as signal processing [5], image [29, 4] or shape analysis [24, 31], to name a few. These applications rely on the specific study of the geometries of usual parametric families of distributions, which has started in the early work of Atkinson and Mitchell. In [7], the authors study the trivial geometries of one-parameter families of distributions, the hyperbolic geometry of the univariate normal model as well as special cases of the multivariate normal model, a work that is continued by Skovgaard in [30]. The family of gamma distributions has been studied by Lauritzen in [19], and more recently by

Arwini and Dodson in [6], who also focus on the log-normal, log-gamma, and families of bivariate distributions. Power inverse Gaussian distributions [35], location-scale models and in particular the von Mises distribution [28], and the generalized gamma distributions [27] have also received attention.

In this work, we are interested in Dirichlet distributions, a family of probability densities defined on the $(n-1)$-dimensional probability simplex, that is the set of vectors of $\mathbb{R}^n$ with non-negative components that sum up to one. The Dirichlet distribution models a random probability distribution on a finite set of size n. It generalizes the beta distribution, a two-parameter probability measure on $[0,1]$ used to model random variables defined on a compact interval. Beta and Dirichlet distributions are often used in Bayesian inference as conjugate priors for several discrete probability laws [23, 16, 11], but also come up in a wide variety of other applications, e.g. to model percentages and proportions in genomic studies [33], distribution of words in text documents [20], or for mixture models [10]. Up to our knowledge, the information geometry of Dirichlet distributions has not yet received much attention. In [12], the authors give the expression of the Fisher-Rao metric for the family of beta distributions, but nothing is said about the geodesics or the curvature.

In this paper, we give new results and properties for the geometry of Dirichlet distributions, and its sectional curvature in particular. The derived expressions depend on the trigamma function, the second derivative of the logarithm of the gamma function, however we will avoid using its properties when possible to obtain our results. Instead, we consider a more general metric written using a function f, for which we only make the strictly necessary assumptions. Section 2 gives the setup for our problem by considering the Fisher-Rao metric on the space of parameters of Dirichlet distributions. In Section 3, we consider the more general metric where f replaces the trigamma function, and show that it induces the geometry of a submanifold in a flat Lorentzian space. This allows us to show geodesic completeness, and that the sectional curvature is everywhere negative. Section 4 focuses on the two-dimensional case, i.e. beta distributions.

2. Fisher-Rao metric on the manifold of Dirichlet distributions

Let Δ_n denote the $(n-1)$-dimensional probability simplex, i.e. the set of vectors in $\mathbb{R}^n$ with non-negative components that sum up to one

$$\Delta_n = \{q = (q_1, \dots, q_n) \in \mathbb{R}^n, \sum_{i=1}^{n} q_i = 1,\ q_i \geq 0, i = 1, \dots, n\}.$$

The family of Dirichlet distributions is a family of probability distributions on Δ_n parametrized by n positive scalars $x_1, \dots, x_n > 0$ (Figure 1), that admits the following probability density function with respect to the Lebesgue measure

$$f_n(q|x_1, \dots, x_n) = \frac{\Gamma(x_1 + \dots + x_n)}{\Gamma(x_1) \dots \Gamma(x_n)} q_1^{x_1 - 1} \dots q_n^{x_n - 1}.$$

As an open subset of $\mathbb{R}^n$, the space of parameters $M = (\mathbb{R}_+^*)^n$ is a differentiable manifold and can be equipped with a Riemannian metric defined in its matrix form by the Fisher information matrix

$$g_{ij}(x_1, \dots, x_n) = -\mathbb{E}\left[\frac{\partial^2}{\partial x_i \partial x_j} \log f_n(Q|x_1, \dots, x_n)\right], \quad i, j = 1, \dots, n,$$

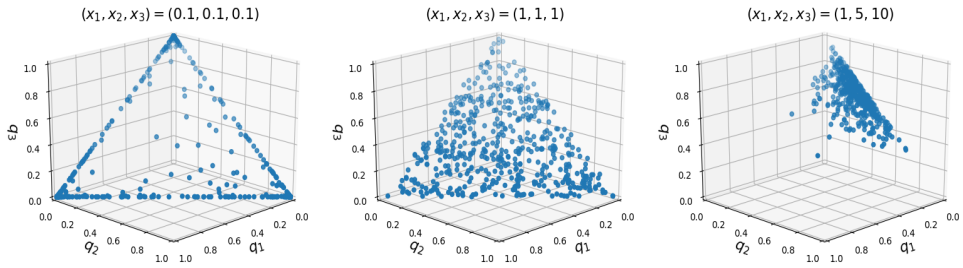

FIGURE 1. Random samples drawn from Dirichlet distributions on the 2-dimensional simplex Δ_3 for different values of the parameters (x_1, x_2, x_3).

where $\mathbb{E}$ denotes the expectation taken with respect to Q, a random variable with density $f_n(\cdot|x_1, \ldots, x_n)$. The Dirichlet distributions form an exponential family and so the Fisher-Rao metric is the hessian of the log-partition function [3], namely

$$g_{ij}(x_1, \ldots, x_n) = \frac{\partial^2}{\partial x_i \partial x_j} \varphi(x_1, \ldots, x_n), \quad i, j = 1, \ldots, n,$$

where φ is the logarithm of the normalizing factor

$$\varphi(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log \Gamma(x_i) - \log \Gamma(x_1 + \ldots + x_n).$$

We obtain the following metric tensor.

$$g_{ij}(x_1, \ldots, x_n) = \psi'(x_i)\delta_{ij} - \psi'(x_1 + \ldots + x_n), \tag{1}$$

where δ_{ij} is the Kronecker delta function, and ψ denotes the digamma function, that is the first derivative of the logarithm of the gamma function, i.e.

$$\psi(x) = \frac{d}{dx} \log \Gamma(x).$$

Its derivative ψ' is called the trigamma function. As noted below, the trigamma function is a function whose reciprocal is increasing, convex, and sublinear on $\mathbb{R}^+$. For slightly greater generality, and to emphasize what properties of this function are needed for our results, we will work in the sequel with a more general function f on which we make only the necessary assumptions; in our special case we have $f = 1/\psi'$.

3. The general framework

3.1. **The metric.** In this section we consider a more general geometry, that admits the Fisher-Rao geometry of Dirichlet distributions as a special case. The goal is to avoid using the properties of the trigamma function when possible. For this, we consider the quadrant $M = (\mathbb{R}_+^*)^n$ equipped with a metric of the form

$$ds^2 = \frac{dx_1^2}{f(x_1)} + \cdots + \frac{dx_n^2}{f(x_n)} - \frac{(dx_1 + \cdots + dx_n)^2}{f(x_1 + \cdots + x_n)}, \tag{2}$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a function on which we make the following assumptions:

$$f(x) \underset{x\to 0}{=} O(x^2), \quad f'(x) \underset{x\to 0}{=} O(x), \quad f(x) \underset{x\to \infty}{=} O(x^2), \quad f'' > 0 \text{ and } \frac{d^2}{dx^2}\left(\frac{f}{f'}\right) > 0. \tag{3}$$

We retrieve the Fisher-Rao metric (1) when

$$f(x) = \frac{1}{\psi'(x)}.$$

Notice that this choice for f satisfies the conditions of (3). Indeed, that $f(0) = f'(0) = 0$ comes from the asymptotic formula $\psi'(x) \approx x^{-2}$ valid near $x = 0$, since

$$f'(0) = \lim_{x\to 0} \frac{-\psi''(x)}{\psi'(x)^2} = \lim_{x\to 0} \frac{2x^{-3}}{x^{-4}} = 0.$$

The fact that the reciprocal of the trigamma function $f(x) = \frac{1}{\psi'(x)}$ is convex comes from an argument of Trimble-Wells-Wright [32], based on an inequality later proved in Alzer-Wells [1]. The fact that f/f' is convex comes from Yang [34].

Another example of a function satisfying the conditions (3) is

$$\tilde{f}(x) = \frac{(2x+1)x^2}{2x^2+2x+1},$$

a simple rational function which approximates the reciprocal of the trigamma function well, in both the small-x and large-x regions.

Some useful consequences of our assumptions (3) are given in the following lemma. These results are well-known, but we include the simple proofs for completeness.

Lemma 1. *If f satisfies (3), then we have $f'(x) > 0$ and $f(x) > 0$ for all $x > 0$. In addition f and f/f' are superadditive:*

$$f(x_1 + \cdots + x_n) > f(x_1) + \cdots + f(x_n), \tag{4}$$

$$\frac{f(x_1 + \cdots + x_n)}{f'(x_1 + \cdots x_n)} > \frac{f(x_1)}{f'(x_1)} + \cdots + \frac{f(x_n)}{f'(x_n)}, \tag{5}$$

for all $x_1, \ldots, x_n > 0$.

Proof. That $f'' > 0$ implies $f' > 0$ and thus $f > 0$ for all $x > 0$ is obvious. It has been known since Petrovich [25] that a convex function f with $f(0) = 0$ is superadditive: an easy argument in the differentiable case is that

$$f(x+y) - f(x) - f(y) = \int_0^x \int_0^y f''(s+t)\, dt\, ds \geq 0.$$

By induction the general case (4) follows. Since $\lim_{x\to 0} f(x)/f'(x) = 0$, the same argument applies to f/f' to give (5). □

3.2. **Lorentzian submanifold geometry.** We now show that after a change of coordinates, M can be seen as a codimension 1 submanifold of the $(n+1)$-dimensional flat Minkowski space $L^{n+1} = (\mathbb{R}^{n+1}, ds_L^2)$, where

$$ds_L^2 = dy_1^2 + \ldots + dy_n^2 - dy_{n+1}^2. \tag{6}$$

In the sequel, we will denote by $\langle \cdot, \cdot \rangle$ the scalar product induced by this metric.

Proposition 2. *The mapping*

$$\Phi : M \to L^{n+1},$$
$$(x_1, \ldots, x_n) \mapsto (\eta(x_1), \ldots, \eta(x_n), \eta(x_1 + \ldots + x_n))$$

where $\eta\colon \mathbb{R}_+ \to \mathbb{R}$ is defined by

$$\eta(x) = \int_1^x \frac{dr}{\sqrt{f(r)}},$$

is an isometric embedding.

Proof. Since $f(x) \approx \frac{1}{2}f''(0)x^2$ for $x \approx 0$, we see that

$$\int_0^1 \frac{dr}{\sqrt{f(r)}} = \infty,$$

so that the image of η must include all negative reals. Therefore η maps $\mathbb{R}_+$ bijectively to $(-\infty, N)$ for some $N \in (0, \infty]$. The behavior of f at infinity assumed in (3) implies that $f(x) \leq Cx^2$ for all $x \geq K$, for some $K > 0$, $C > 0$, which in turns leads to

$$\int_K^\infty \frac{dx}{\sqrt{f(x)}} = \infty.$$

Therefore η maps bijectively $\mathbb{R}_+$ to $\mathbb{R}$, and Φ is a homeomorphism onto its image. Since $\eta'(x) > 0$ for all x, it is also an immersion. Finally, if $(y_1, \ldots, y_{n+1}) = \Phi(x_1, \ldots, x_n)$,

$$dy_i^2 = \eta'(x_i)^2 dx_i^2 = \frac{dx_i^2}{f(x_i)}, i = 1, \ldots, n, \quad dy_{n+1}^2 = \frac{(dx_1 + \ldots + dx_n)^2}{f(x_1 + \ldots + x_n)},$$

and Φ is isometric. □

Proposition 3. *$S = \Phi(M)$ is a codimension 1 submanifold of L^{n+1} given by the graph of*

$$y_{n+1} = \eta(\xi(y_1) + \ldots + \xi(y_n)), \quad y_i > 0, \tag{7}$$

where $\xi = \eta^{-1}$. On this submanifold the metric is positive-definite and thus Riemannian. A basis of tangent vectors of T_yS is defined by

$$e_i = \frac{\partial}{\partial y_i} + \sqrt{\frac{f \circ \xi(y_i)}{f \circ \xi(y_{n+1})}} \frac{\partial}{\partial y_{n+1}}, \quad i = 1, \ldots, n, \tag{8}$$

Proof. Let $\gamma(u) = (y_1(u), \ldots, y_{n+1}(u))$ be a parametrized curve in S. Then its coordinates verify the following relations

$$\begin{aligned} y_{n+1} &= \eta(\xi(y_1) + \ldots + \xi(y_n)), \\ y'_{n+1} &= \eta'(\xi(y_1) + \ldots + \xi(y_n))(\xi'(y_1)y'_1 + \ldots + \xi'(y_n)y'_n), \end{aligned}$$

and so, since $\xi'(x) = \sqrt{f(\xi(x))}$,

$$\begin{aligned} \gamma'(u) &= \sum_{i=1}^n y'_i(u) \frac{\partial}{\partial y_i} + \eta'(\xi(y_{n+1}(u)))(\xi'(y_1(u))y'_1(u) + \ldots + \xi'(y_n(u))y'_n(u)) \frac{\partial}{\partial y_{n+1}} \\ &= \sum_{i=1}^n y'_i(u) \left(\frac{\partial}{\partial y_i} + \frac{\sqrt{f \circ \xi(y_i(u))}}{\sqrt{f \circ \xi(y_{n+1}(u))}} \frac{\partial}{\partial y_{n+1}} \right), \end{aligned}$$

yielding (8) as basis tangent vectors. The metric components on S take the form

$$g_{ij} = \langle e_i, e_j \rangle = \delta_{ij} - W_i W_j, \qquad \text{or} \quad g = I - WW^T,$$

where $\langle \cdot, \cdot \rangle$ denotes the flat Minkowskian metric (6) and $W_i = \sqrt{f(\xi(y_i))/f(\xi(y_{n+1}))}$ for $i = 1, \ldots, n$. Applying Lemma 16 of the appendix gives the result upon computing

$$W^T W = \frac{\sum_{i=1}^n f\big(\xi(y_i)\big)}{f\big(\sum_{i=1}^n \xi(y_i)\big)} < 1,$$

by superadditivity of f, as in (4). □

In other words, the metric (2) is the restriction of the flat Lorentzian metric

$$ds^2 = \frac{dx_1^2}{f(x_1)} + \cdots + \frac{dx_n^2}{f(x_n)} - \frac{dt^2}{f(t)}$$

to the hyperplane $t = x_1 + \cdots + x_n$. In the sequel, we will use this Lorentzian submanifold geometry to study the sectional curvature and geodesic completeness of M. We state the results in the original coordinate system of M when possible, using the following notations for any $y = (y_1, \ldots, y_{n+1}) \in S$:

$$x_i = \xi(y_i),\ i = 1, \ldots, n, \quad t = x_1 + \ldots + x_n = \xi(y_{n+1}). \tag{9}$$

3.3. **Negative sectional curvature.** The goal of this section is to prove that the sectional curvature of M is everywhere negative. We start by computing the shape operator.

Proposition 4. *The shape operator of $S = \Phi(M)$ has the following components in the basis (8) of tangent vectors*

$$\langle \Sigma(e_i), e_j \rangle = -\frac{1}{2\sqrt{f(t) - \sum_{\ell=1}^n f(x_\ell)}} \left(f'(x_i)\delta_{ij} - \frac{f'(t)}{f(t)} \sqrt{f(x_i) f(x_j)} \right).$$

Proof. We first observe that the basis vectors (8) can be expressed in coordinates (9) as

$$e_i = \frac{\partial}{\partial y_i} + \sqrt{\frac{f(x_i)}{f(t)}} \frac{\partial}{\partial y_{n+1}}, \quad i = 1, \ldots, n. \tag{10}$$

Since S can be obtained as the graph of $F(y_1, \ldots, y_n) = \eta(\xi(y_1), \ldots, \xi(y_n))$, a normal vector field to S at y is given by

$$N = \sum_{i=1}^n \frac{\partial F}{\partial y_i} \frac{\partial}{\partial y_i} + \frac{\partial}{\partial y_{n+1}} = \sum_{i=1}^n \sqrt{\frac{f(x_i)}{f(t)}} \frac{\partial}{\partial y_i} + \frac{\partial}{\partial y_{n+1}}, \tag{11}$$

which yields a timelike vector since

$$\langle N, N \rangle = \frac{1}{f(t)} \left(f(x_1) + \ldots + f(x_n) - f(t) \right) < 0, \tag{12}$$

by superadditivity of f. Since $\langle N, e_i \rangle = 0$, the shape operator is then given by

$$\langle \Sigma(e_i), e_j \rangle = -\langle \nabla_{e_i} \left(\frac{N}{\sqrt{-\langle N, N \rangle}} \right), e_j \rangle = -\frac{\langle \nabla_{e_i} N, e_j \rangle}{\sqrt{-\langle N, N \rangle}}, \tag{13}$$

∇ is the flat connection of the Minkowski space. Denoting $\partial_i = \partial / \partial y_i$, we get from (10), (11) and the flatness of ∇,

$$\nabla_{e_i} N = \nabla_{\partial_i} N + \frac{f(x_i)}{f(t)} \nabla_{\partial_{n+1}} N = \sum_{j=1}^n \partial_i \partial_j F \partial_j.$$

Inserting this last equation along with (12) into (13) yields

$$\langle \Sigma(e_i), e_j \rangle = -\sqrt{\frac{f(t)}{f(t) - \sum_{\ell=1}^n f(x_\ell)}} \partial_i \partial_j F.$$

Straightforward computations give

$$\partial_i F = \eta'(t)\xi'(y_i) = \sqrt{f(x_i)/f(t)},$$
$$\partial_i\partial_j F = \eta''(t)\xi'(y_i)\xi'(y_j) + \eta'(t)\xi''(y_i)\delta_{ij} = \frac{1}{2\sqrt{f(t)}}\left(\frac{-f'(t)}{f(t)}\sqrt{f(x_i)f(x_j)} + f'(x_i)\delta_{ij}\right),$$

and the result follows after simplification. □

Corollary 5. *The second fundamental form given by Proposition 4 is positive-definite.*

Proof. This follows from Lemma 16 and the decomposition of the matrix Σ with components $\Sigma_{ij} = \langle \Sigma(e_i), e_j\rangle$ as

$$\Sigma = -\frac{1}{2}k(D - cVV^T),$$

where $D = \mathrm{diag}(d_1, \ldots, d_n)$ is a diagonal matrix, $V = (v_i)_{1\leq i\leq n}$ is a column vector and c and k are constants, defined for $i = 1, \ldots, n$ by

$$d_i = f'(x_i), \quad v_i = \sqrt{f(x_i)}, \quad k = \frac{1}{\sqrt{f(t) - \sum_{\ell=1}^n f(x_\ell)}}, \quad c = \frac{f'(t)}{f(t)}. \tag{14}$$

Recalling that $f > 0$ and $f' > 0$ by Lemma 1, we see that the matrix D and constant c are positive. There remains to verify that

$$cV^T D^{-1} V = \frac{f'(t)}{f(t)}\sum_{i=1}^n \frac{f(x_i)}{f'(x_i)} < 1,$$

by the superadditivity property (5). □

We can now show our main result.

Theorem 6. *The sectional curvature of the Riemannian metric* (2) *is negative on* M.

Proof. We use a result from O'Neill [22, Chapter 4, Corollary 20], which states that if the normal vector field N of a hypersurface M in a flat Lorentzian manifold L is timelike, then the sectional curvature of the submanifold is given by

$$K(U,V) = -\frac{\langle \Sigma(U), U\rangle\langle \Sigma(V), V\rangle - \langle \Sigma(U), V\rangle^2}{\langle U, U\rangle\langle V, V\rangle - \langle U, V\rangle^2}, \tag{15}$$

where U and V are tangent to the submanifold and Σ is the shape operator. The result now follows by the Cauchy-Schwarz inequality: since Σ is a positive-definite symmetric matrix, we know that $\langle \Sigma(U), U\rangle\langle \Sigma(V), V\rangle \geq \langle \Sigma(U), V\rangle^2$ with equality iff V is a multiple of U, but in that case the denominator vanishes as well. So the sectional curvature must be strictly negative. □

We now give more specifically the formula of the sectional curvature of the planes generated by the basis tangent vectors (8).

Proposition 7. *The sectional curvature along the axes defined by* (8) *is given by*

$$K(e_i, e_j) = \frac{f(x_i)f'(x_j)f'(t) + f'(x_i)f(x_j)f'(t) - f'(x_i)f'(x_j)f(t)}{4(f(t) - \sum_{\ell=1}^n f(x_\ell))(f(t) - f(x_i) - f(x_j))}.$$

Proof. This follows from applying formula (15) for the sectional curvature of a hypersurface in a flat Lorentzian manifold, with

$$\langle e_i, e_i\rangle = 1 - \frac{f(x_i)}{f(t)}, \quad \langle e_i, e_j\rangle = \sqrt{\frac{f(x_i)f(x_j)}{f(t)^2}}, \quad i \neq j,$$

and $\langle \Sigma(e_i), e_j\rangle$ given by Proposition 4. □

Finally, we state a result about the eigenvalues of the shape operator, which will be useful to show geodesic completeness in the next section.

Proposition 8. *The principal curvatures at any point in $S = \Phi(M)$ are bounded.*

Proof. The principal curvatures at a given point $(y_1, \dots, y_{n+1}) = \Phi(x_1, \dots, x_n) \in S$ are the eigenvalues of the shape operator

$$\Sigma = -\frac{1}{2}k(D - cVV^T),$$

where D, V, c and k are defined by (14). Without loss of generality, we assume that the n-tuple $(x_1, \dots, x_n)$ is ordered. Let us first show that when at least $n-1$ variables go to zero, i.e. $x_i \to 0$ for $i = 1, \dots, n-1$ with the previous assumption, the principal curvatures go to zero. Let $\tau = x_1 + \ldots + x_{n-1}$, then $t = x_n + \tau$ and

$$f(t) - f(x_1) - \ldots - f(x_n) \underset{\tau\to 0}{\sim} \tau f'(x_n)$$

since f has limit zero in zero, and so

$$k \underset{\tau\to 0}{\sim} \frac{1}{\sqrt{\tau f'(x_n)}}.$$

Using the fact that when $\tau \to 0$, recalling assumptions (3),

$$f(x_i) = O(x_i^2), \quad f'(x_i) = O(x_i), \quad x_i = O(\tau), \quad i = 1, \dots, n-1,$$

we see that the diagonal terms of $D - cVV^T$ behave as

$$f'(x_i) - \frac{f'(t)}{f(t)}f(x_i) = O(\tau), \quad i = 1, \dots, n-1,$$

$$f'(x_n) - \frac{f'(t)}{f(t)}f(x_n) = \frac{f'(x_n)f(x_n+\tau) - f'(x_n+\tau)f(x_n)}{f(x_n+\tau)} \underset{\tau\to 0}{\sim} \tau\left(f''(x_n) + \frac{f'(x_n)^2}{f(x_n)}\right),$$

while the antidiagonal terms verify

$$-\frac{f'(t)}{f(t)}\sqrt{f(x_i)f(x_j)} = O(\tau^2), \quad 1 \leq i, j \leq n-1,$$

$$-\frac{f'(t)}{f(t)}\sqrt{f(x_i)f(x_n)} = O(\tau).$$

Finally, we obtain that

$$\Sigma_{ij} = \langle \Sigma(e_i), e_j\rangle \underset{\tau\to 0}{=} O(\sqrt{\tau}), \quad 1 \leq i, j \leq n,$$

and so the principal curvatures go to zero when $\tau \to 0$. Therefore there exists $\delta > 0$ such that, at any point $(x_1, \dots, x_n)$ belonging to the set

$$\mathcal{D}_\delta = \{(x_1, \dots, x_n) \in (\mathbb{R}_+^*)^n, x_i < \delta \text{ for at least } n-1 \text{ indices } i \in \{1, \dots, n\}\},$$

the principal curvatures are upper bounded by, say, 1. Now let us consider an n-tuple $(x_1, \ldots, x_n) \notin \mathcal{D}_\delta$, ordered as before. Then the diagonal elements of D are ordered as well since f' is increasing, and the ordered eigenvalues of $k(D - cVV^T)$ verify

$$0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq k d_n,$$

where the lower bound comes from the positive-definiteness shown in Corollary 5, and the upper bound comes from [15]. Since $d_n = f'(x_n)$ and f' is increasing and upper bounded by $\lim_{x\to\infty} f'(x) = 1$, we have that $d_n \leq 1$. Since the function

$$(x_1, \ldots, x_n) \mapsto f(x_1 + \ldots + x_n) - f(x_1) - \ldots - f(x_n)$$

is increasing in all of its variables, it is larger than its limit as the first $n-2$ variables go to zero, and since at least $x_{n-1} > \delta$ and $x_n > \delta$, we obtain

$$k = \frac{1}{\sqrt{f(t) - f(x_1) - \ldots - f(x_n)}} \leq \frac{1}{\sqrt{f(2\delta) - 2f(\delta)}},$$

and the principal curvatures are again bounded. □

3.4. **Geodesics and geodesic completeness.** The geodesics of M for the metric (2) are parametrized curves $u \mapsto (x_1(u), \ldots, x_n(u))$ solution of the standard second-order ODEs

$$\ddot{x}_k + \sum_{1\leq i,j\leq n} \Gamma^k_{ij} \dot{x}_i \dot{x}_j = 0, \quad k = 1, \ldots, n,$$

whose coefficients can be computed using the following result.

Proposition 9. *The Christoffel symbols for metric* (2) *are given by*

$$\Gamma^k_{ij} = \frac{1}{2}\left[\frac{f(x_k)}{f(t) - \sum_{\ell=1}^n f(x_\ell)}\left(g(t) - g(x_j)\delta_{ij}\right) - g(x_k)\delta_{ij}\delta_{jk}\right],$$

where $t = x_1 + \ldots + x_n$ *and* $g(x) = f'(x)/f(x)$, *while* δ *denotes the Kronecker delta function.*

Proof. The Christoffel symbols of the second kind Γ^k_{ij} can be obtained from the Christoffel symbols of the first kind Γ_{ijk} and the coefficients g^{ij} of the inverse of the metric matrix using the formula

$$\Gamma^k_{ij} = \Gamma_{ijl} g^{kl},$$

where we have used the Einstein summation convention. It is easy to see that the Christoffel symbols of the first kind are given by

$$\Gamma_{ijk} = \frac{1}{2}\left(\frac{f'(t)}{f(t)^2} - \frac{f'(x_k)}{f(x_k)^2}\delta_{ik}\delta_{jk}\right).$$

Applying the Sherman-Morrison formula, we obtain that the inverse of the metric matrix

$$g(x_1, \ldots, x_n) = \operatorname{diag}\left(\frac{1}{f(x_1)}, \ldots, \frac{1}{f(x_n)}\right) - \frac{1}{f(t)} J,$$

where J denotes the n-by-n matrix with all entries equal to one, is given by

$$g(x_1, \ldots, x_n)^{-1} = \operatorname{diag}(f(x_1), \ldots, f(x_n)^{-1}) + \frac{1}{f(t) - \sum_{\ell=1}^n f(x_\ell)} [f(x_i) f(x_j)]_{1\leq i,j\leq n}. \tag{16}$$

Noticing that the sum of all the elements of the kth line (or column) of the inverse of the metric matrix is given by

$$\sum_{\ell=1}^n g^{k\ell} = f(x_\ell) + \frac{\sum_{\ell=1}^n f(x_k) f(x_\ell)}{f(t) - \sum_{\ell=1}^n f(x_\ell)} = \frac{f(x_k) f(t)}{f(t) - \sum_{\ell=1}^n f(x_\ell)}, \tag{17}$$

we obtain

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{\ell=1}^n \left(\frac{f'(t)}{f(t)^2} - \frac{f'(x_\ell)}{f(x_\ell)^2}\delta_{ij}\delta_{j\ell}\right) g^{k\ell} = \frac{1}{2}\frac{f'(t)}{f(t)^2}\sum_{\ell=1}^n g^{k\ell} - \frac{1}{2}\delta_{ij}\frac{f'(x_j)}{f(x_j)^2}g^{kj}.$$

Inserting (17) and the general term of the inverse matrix (16) in the above yields

$$\Gamma_{ij}^k = \frac{1}{2}\frac{f'(t)}{f(t)^2}\frac{f(x_k)f(t)}{f(t)-\sum_{\ell=1}^n f(x_\ell)} - \frac{1}{2}\delta_{ij}\frac{f'(x_j)}{f(x_j)^2}\left(f(x_k)\delta_{kj} + \frac{f(x_k)f(x_j)}{f(t)-\sum_{\ell=1}^n f(x_\ell)}\right)$$

and the result follows. □

Now, using the result of Proposition 8 and a theorem from [17], we can show that M is geodesically complete.

Theorem 10. *M equipped with the Riemannian metric (2) is geodesically complete.*

Proof. The image of M by Φ is a hypersurface of the $(n+1)$-Minkowski space L^{n+1}. Moreover, Φ is an embedding and it is closed since $\Phi(M)$ is a closed subset of L^{n+1} as preimage of the singleton $\{0\}$ by the continuous map $(y_1,\dots,y_{n+1}) \mapsto \xi(y_1)+\ldots+\xi(y_n)-\xi(y_{n+1})$. Therefore Φ is proper [17, Theorem 1]. Then, [17, Theorem 6] allows us to conclude that since Φ has bounded principal curvatures by Proposition 8, M equipped with the pullback (2) of the Minkowski metric by Φ is complete. □

3.5. **Uniqueness of the Fréchet mean.** Since M is simply connected, we deduce from Theorems 6 and 10 the following.

Corollary 11. *M equipped with the Riemannian metric (2) is a Hadamard manifold.*

This has important implications in information geometry, as it guarantees the uniqueness of the Fréchet mean of a set of points in this geometry. The Fréchet mean, also called intrinsic mean, is a popular choice to extend the notion of barycenter to a Riemannian manifold. It is defined for a set of points $p_1,\dots,p_N \in M$ as the minimizer of the sum of the squared geodesic distances to the points of the set

$$\bar{p} = \operatorname*{argmin}_{p\in M} \sum_{i=1}^N d(p,p_i)^2.$$

It exists as long as M is complete, however it is in general not unique and refers to a set. Uniqueness holds however for Hadamard manifolds [18]. This implies that the notion of barycenter of Dirichlet distributions is well defined in the Fisher-Rao geometry.

4. The two-dimensional case of beta distributions

The simplest case is obviously when $n = 2$, and even in this case the formulas are nontrivial. When $f = \frac{1}{\psi'}$, the metric comes from the well-known two-parametric family of beta distributions defined on the compact interval $[0,1]$, which is important in statistics and useful in many applications.

Proposition 12. *The geodesic equations are given by*

$$\begin{aligned} a(x,y)\ddot{x} + b(x,y)\dot{x}^2 + c(x,y)\dot{x}\dot{y} + d(x,y)\dot{y}^2 = 0,\\ a(y,x)\ddot{y} + b(y,x)\dot{y}^2 + c(y,x)\dot{x}\dot{y} + d(y,x)\dot{x}^2 = 0, \end{aligned} \tag{18}$$

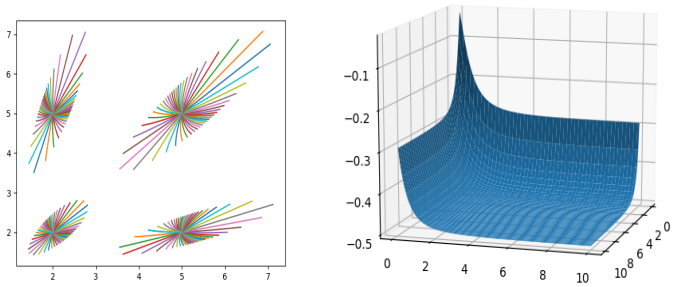

FIGURE 2. On the left, geodesic balls and on the right, sectional curvature of the manifold of beta distributions ($n = 2$).

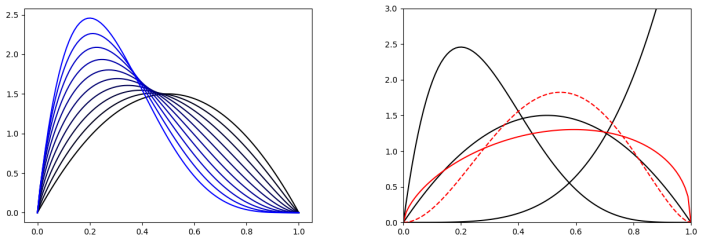

FIGURE 3. On the left, geodesic between the beta distributions of parameters $(2, 5)$ and $(2, 2)$ and on the right, Fréchet mean (full red line) compared to the Euclidean mean (dashed red line) of the beta distributions of parameters $(2, 5)$, $(2, 2)$ and $(5, 1)$, shown in terms of probability density function.

where

$$\begin{aligned} a(x,y) &= 2\big[f(x+y) - f(x) - f(y)\big] \\ b(x,y) &= f(y)g(x) + f(x)g(x+y) - f(x+y)g(x) \\ c(x,y) &= 2f(x)g(x+y) \\ d(x,y) &= f(x)g(x+y) - g(y)f(x), \end{aligned}$$

with the shorthand $g(x) = f'(x)/f(x)$.

Proof. The geodesic equations can be expressed in terms of the Christoffel symbols as

$$\begin{aligned} \ddot{x} + \Gamma^1_{11}\dot{x}^2 + 2\Gamma^1_{12}\dot{x}\dot{y} + \Gamma^1_{22}\dot{y}^2 = 0, \\ \ddot{y} + \Gamma^2_{22}\dot{y}^2 + 2\Gamma^2_{12}\dot{x}\dot{y} + \Gamma^2_{11}\dot{x}^2 = 0, \end{aligned}$$

and the coefficients can be computed using Proposition 9. □

No closed form is known for the geodesics, but they can be computed numerically by solving (18), see the left-hand side of Figure 2. Nonetheless we can notice that, due to the symmetry of the metric with respect to parameters x and y, both equations in (18) yield a unique ordinary differential equation when $x = y$.

Corollary 13. *Solutions of the geodesic equation* (18) *with* $x(0) = y(0)$ *and* $\dot{x}(0) = \dot{y}(0)$ *satisfy*

$$\sqrt{q(x(t))}\dot{x}(t) = \text{constant}, \tag{19}$$

where $q(x) = \frac{1}{f(x)} - \frac{2}{f(2x)}$*, and thus can be found by quadratures.*

Proof. If at some time t_0 we have $x = y$ and $\dot{x} = \dot{y}$, then the equations (18) imply that $\ddot{x} = \ddot{y}$ at t_0. Differentiating repeatedly in time shows that all higher derivatives must also be equal at t_0, and we conclude by analyticity of the solutions that $x(t) = y(t)$ on some interval. The usual extension arguments for ODEs then imply that $x(t) = y(t)$ on the entire domain of the solution, which by Theorem 10 is $\mathbb{R}$.

When $x = y$ equation (18) reduces to

$$2\big[f(2x) - 2f(x)\big]\ddot{x} + \left(\frac{4f(x)f'(2x)}{f(2x)} - \frac{f(2x)f'(x)}{f(x)}\right)\dot{x}^2 = 0,$$

which is equivalent to

$$2q(x)\ddot{x} + q'(x)\dot{x}^2 = 0.$$

This clearly implies the conservation law (19). The differential equation (19) can then be solved by writing

$$t = \frac{1}{\dot{x}_0\sqrt{q(x_0)}}\int_{x_0}^{x}\sqrt{q(s)}\,ds$$

and inverting the resulting function. □

For example, if $f(x) = 1/\psi'(x)$, then the duplication formula for the trigamma function implies

$$q(x) = \psi'(x) - 2\psi'(2x) = \tfrac{1}{2}[\psi'(x) - \psi'(x + \tfrac{1}{2})].$$

Asymptotically this looks like $q(x) \approx \frac{1}{2x^2}$ for $x \approx 0$ and $q(x) \approx \frac{1}{4x^2}$ as $x \to \infty$. We conclude that it takes infinite time for a geodesic along the diagonal to either reach "diagonal infinity" or the origin, as Theorem 10 of course implies.

From an applications point of view, the geodesics for the Fisher-Rao geometry allow us to define a notion of optimal interpolation between beta and more generally Dirichlet distributions. An example of such an optimal interpolation is shown on the left-hand side of Figure 3, in terms of probability density function.

Now we give the formula for the sectional curvature in two dimensions.

Proposition 14. *If* $n = 2$*, the sectional curvature is given by*

$$K(x,y) = -\frac{1}{4}\,\frac{f(t)f'(x)f'(y) - f(x)f'(t)f'(y) - f(y)f'(t)f'(x)}{\big[f(t) - f(x) - f(y)\big]^2}, \tag{20}$$

where $t = x + y$*.*

Proof. This is just a particular case of Proposition 7. □

Notice that in two dimensions, the negativity of the sectional curvature is straightforward, as there is only one Gaussian curvature to consider, which is given by (20), in which one can easily see that the numerator is positive by factorizing by $f'(x)f'(y)f'(t) > 0$ and using the superadditivity property (5) of f/f'.

As previously mentioned, the negative curvature of the Fisher-Rao geometry also has interesting implications for applications: it entails that the Fréchet mean of a set of beta, or more generally Dirichlet distributions is well defined. An example of Fréchet mean of

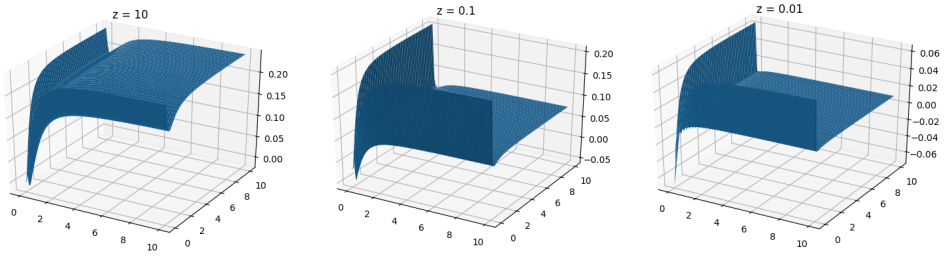

FIGURE 4. The difference (23) between the sectional curvatures of the plane generated by e_1 and e_2 in two and three dimensions changes sign for $z = 0.01$.

beta distributions is shown in terms of probability density function on the right-hand side of Figure 3.

Numerically we observe that when $f = 1/\psi'$, the function $K(x, y)$ given by (20) is decreasing in both the x and y variables – see the right-hand side of Figure 2 – but we do not yet have a proof of this fact. However we may analyze the asymptotics of the function relatively easily.

Proposition 15. *If $f = 1/\psi'$, then the asymptotic behavior of the sectional curvature given by* (20) *approaching the boundary square is given by*

$$\lim_{y\to 0} K(x, y) = \lim_{y\to 0} K(y, x) = \frac{3}{4} - \frac{\psi'(x)\psi'''(x)}{2\,\psi''(x)^2}, \tag{21}$$

$$\lim_{y\to \infty} K(x, y) = \lim_{y\to \infty} K(y, x) = \frac{x\,\psi''(x) + \psi'(x)}{4(x\,\psi'(x) - 1)^2}. \tag{22}$$

Moreover, we have the following limits at the asymptotic corners:

$$\lim_{x,y\to 0} K(x, y) = 0, \quad \lim_{x,y\to \infty} K(x, y) = -\frac{1}{2}, \quad \lim_{x\to 0, y\to \infty} K(x, y) = \lim_{x\to \infty, y\to 0} K(x, y) = -\frac{1}{4}.$$

Proof. Writing $K(x, y) = -\frac{A(x,y)}{4B(x,y)^2}$, with

$$A(x, y) = f(x + y)f'(x)f'(y) - f(x)f'(x + y)f'(y) - f(y)f'(x + y)f'(x),$$
$$B(x, y) = f(x + y) - f(x) - f(y),$$

we note that $A(x, 0) = M_y(x, 0) = 0$ and $N(x, 0) = 0$, so that

$$\lim_{y\to 0} K(x, y) = \frac{A_{yy}(x, 0)}{8B_y(x, 0)^2},$$

which gives (21) after rewriting in terms of ψ'.

For the infinite limits, we use the facts that $\lim_{y\to\infty} f(y) - y = -\frac{1}{2}$ and $\lim_{y\to\infty} f'(y) = 1$, and that $\lim_{y\to\infty} y(f'(y) - 1) = 0$, to obtain limits of $A(x, y)$ and $B(x, y)$ separately with elementary computations. □

These limits and strong numerical evidence allow us to conjecture that the sectional curvature in two dimensions is lower bounded by $-1/2$. Comparing the two-dimensional sectional curvature $K_2(x, y) = K(x, y)$ with the sectional curvature of the plane generated

by e_1 and e_2 in three dimensions, that we denote by $K_3(x, y, z)$, we observe numerically that for a given $z > 0$, the function

$$(x, y) \mapsto K_3(x, y, z) - K_2(x, y) \tag{23}$$

does not have a fixed sign in general, as can be observed on Figure 4 for small values of x, y and z.

Acknowledgments

S. C. Preston was partially supported by Simons Foundation, Collaboration Grant for Mathematicians, no. 318969. A. Le Brigant and S. Puechmorel would like to thank Fabrice Gamboa and Thierry Klein for bringing this problem to their attention and for fruitful discussions.

Appendix

Here we give a well-known principle to establish positivity of matrices.

Lemma 16. *Suppose A is a positive-definite symmetric matrix, V is a vector, and c is a positive real number. Then $B = A - cVV^T$ is positive-definite if and only if*

$$cV^TA^{-1}V < 1.$$

Proof. Since A is positive-definite and symmetric, we may write $A = P^2$ for some positive-definite symmetric matrix P. Let $X = P^{-1}V$; then we may write

$$B = P^2 - cVV^T = P\big(I - c(P^{-1}V)(P^{-1}V)^T\big)P = P(I - cXX^T)P.$$

Denoting by $\langle U|U\rangle = U^TU$ the usual scalar product on $\mathbb{R}^n$, we have for any vector U,

$$\begin{aligned}\langle U|BU\rangle &= \langle PU|PU\rangle - c\langle PU|X\rangle^2 = |Y|^2 - c\langle Y|X\rangle^2\\ &\geq |Y|^2 - c|X|^2|Y|^2 = |Y|^2(1 - c|X|^2),\end{aligned}$$

where $Y = PU$, using the Cauchy-Schwarz inequality. This is positive for all U if and only if the right side is positive for all Y, which translates into $c|X|^2 < 1$. Since $|X|^2 = \langle P^{-1}V|P^{-1}V\rangle = \langle V|A^{-1}V\rangle$, we obtain the claimed result. □

References

[1] Horst Alzer and Jim Wells. Inequalities for the polygamma functions. *SIAM Journal on Mathematical Analysis*, 29(6):1459–1466, 1998.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[4] Jesus Angulo and Santiago Velasco-Forero. Morphological processing of univariate gaussian distribution-valued images based on poincaré upper-half plane representation. In *Geometric Theory of Information*, pages 331–366. Springer, 2014.

[5] Marc Arnaudon, Frédéric Barbaresco, and Le Yang. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604, 2013.

[6] Khadiga Arwini and Christopher TJ Dodson. *Information geometry: near randomness and near independence*. Springer Science & Business Media, 2008.

[7] Colin Atkinson and Ann FS Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 345–365, 1981.

[8] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. Information geometry and sufficient statistics. *Probability Theory and Related Fields*, 162(1-2):327–364, 2015.

[9] Martin Bauer, Martins Bruveris, and Peter W Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

[10] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, 2004.
[11] Andrew H Briggs, AE Ades, and Martin J Price. Probabilistic sensitivity analysis for decision trees with multiple branches: use of the dirichlet distribution in a bayesian framework. *Medical Decision Making*, 23(4):341–350, 2003.
[12] Ovidiu Calin and Constantin Udrişte. *Geometric modeling in probability and statistics*. Springer, 2014.
[13] Nikolai Nikolaevich Cencov. Statistical decision rules and optimal inference. transl. math. *Monographs, American Mathematical Society, Providence, RI*, 1982.
[14] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.
[15] Gene H Golub. Some modified matrix eigenvalue problems. *Siam Review*, 15(2):318–334, 1973.
[16] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. 2002.
[17] Stephen G Harris. Closed and complete spacelike hypersurfaces in minkowski space. *Classical and Quantum Gravity*, 5(1):111, 1988.
[18] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.
[19] Stefan L Lauritzen. Statistical manifolds. *Differential geometry in statistical inference*, 10:163–216, 1987.
[20] Rasmus E Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552, 2005.
[21] Yann Ollivier. True asymptotic natural gradient optimization. *arXiv preprint arXiv:1712.08449*, 2017.
[22] Barrett O'neill. *Semi-Riemannian geometry with applications to relativity*. Academic press, 1983.
[23] Philip D O'Neill and Gareth O Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.
[24] Adrian Peter and Anand Rangarajan. Shape analysis using the fisher-rao riemannian metric: Unifying shape representation and deformation. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, pages 1164–1167. IEEE, 2006.
[25] M Petrovich. Sur une fonctionnelle. *Publ. Math. Beograd, TL*, 1932.
[26] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37, 01 1945.
[27] Sana Rebbah, Florence Nicol, and Stéphane Puechmorel. The geometry of the generalized gamma manifold and an application to medical imaging. *Mathematics*, 7(8):674, 2019.
[28] Salem Said, Lionel Bombrun, and Yannick Berthoumieu. Warped riemannian metrics for location-scale models. In *Geometric Structures of Information*, pages 251–296. Springer, 2019.
[29] Olivier Schwander and Frank Nielsen. Model centroids for the simplification of kernel density estimators. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 737–740. IEEE, 2012.
[30] Lene Theil Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, pages 211–223, 1984.
[31] Anuj Srivastava, Ian Jermyn, and Shantanu Joshi. Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
[32] SY Trimble, Jim Wells, and FT Wright. Superadditive functions and a statistical application. *SIAM journal on mathematical analysis*, 20(5):1255–1259, 1989.
[33] Shengping Yang and Zhide Fang. Beta approximation of ratio distribution and its application to next generation sequencing read counts. *Journal of applied statistics*, 44(1):57–70, 2017.
[34] Zhen-Hang Yang. Some properties of the divided difference of psi and polygamma functions. *Journal of Mathematical Analysis and Applications*, 455(1):761 – 777, 2017.
[35] Zhenning Zhang, Huafei Sun, and Fengwei Zhong. Information geometry of the power inverse gaussian distribution. *Applied Sciences*, 9, 2007.

SAMM 4543, Université Paris 1 Panthéon Sorbonne, Centre PMF, Paris, France.
Email address: alice.le-brigant@univ-paris1.fr

Department of Mathematics, Brooklyn College and CUNY Graduate Center, New York, USA.
Email address: stephen.preston@brooklyn.cuny.edu

Ecole Nationale de l'Aviation Civile, Université de Toulouse, Toulouse, France.
Email address: stephane.puechmorel@enac.fr