Signals and Communication Technology

Frank Nielsen *Editor*

Geometric Structures of Information

Springer

Signals and Communication Technology

The series "Signals and Communications Technology" is devoted to fundamentals and applications of modern methods of signal processing and cutting-edge communication technologies. The main topics are information and signal theory, acoustical signal processing, image processing and multimedia systems, mobile and wireless communications, and computer and communication networks. Volumes in the series address researchers in academia and industrial R&D departments. The series is application-oriented. The level of presentation of each individual volume, however, depends on the subject and can range from practical to scientific.

More information about this series at http://www.springer.com/series/4748

Frank Nielsen
Editor

Geometric Structures of Information

Springer

Editor
Frank Nielsen
Sony Computer Science Laboratories, Inc.
Tokyo, Japan

and

Ecole Polytechnique
Palaiseau, France

ISSN 1860-4862 ISSN 1860-4870 (electronic)
Signals and Communication Technology
ISBN 978-3-030-02519-9 ISBN 978-3-030-02520-5 (eBook)
https://doi.org/10.1007/978-3-030-02520-5

Library of Congress Control Number: 2018960356

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This collective book is devoted to the recent advancements of the *Geometric Structures of Information* (GSI) and *Information Geometry* (IG). The book consists of twelve contributed chapters that have been carefully peer reviewed and revised. Each chapter has been assigned to two to five reviewers. I am very thankful for the reviewers' expertise, feedback, and insight and would like to acknowledge them in alphabetical order:

Ayan Basu, Charles Casimiro Cavalcante, Shinto Eguchi, Gaëtan Hadjeres, Tomoyuki Ichiba, Peter Jupp, Amor Keziou, Takashi Kurose, Bertrand Lods, Hiroshi Matsuzoe, Subrahamanian Moosath K. S., Cyrus Mostajeran, Jan Naudts, Nigel Newton, Frank Nielsen, Richard Nock, Atsumi Ohara, Xavier Pennec, Giovanni Pistone, Hector Roman Quiceno-Echavarría, Johannes Ruf, George Ruppeiner, Salem Said, Aida Toma, Barbara Trivellato, Pavan Turaga, Paul Vos, and Konstantinos Zografos. I apologize for any potential omission.

I list below the collection of chapters in order of appearance in this book:

1. Naudts and Zhang present in their paper entitled "Rho-Tau Embedding of Statistical Models" the (ρ, τ) monotone embeddings of statistical models (for any two increasing functions ρ and τ) and discuss the gauge freedom and its induced geometry.
2. Montrucchio and Pistone studied a class of nonparametric deformed statistical models, its associated statistical divergence, and the underlying geometry in their paper "A Class of Non-parametric Deformed Exponential Statistical Models."
3. Henmi and Matsuzoe survey and report recent results on statistical manifolds admitting torsion (SMAT) and precontrast functions and show how to get this SMAT structure from an estimating function in their paper entitled "Statistical Manifolds Admitting Torsion and Partially Flat Spaces."
4. Ohara studies some transformation on the probability simplex from the viewpoint of affine differential geometry and provides some applications in his paper entitled "Conformal Flattening on the Probability Simplex and Its Applications to Voronoi Partitions and Centroids."

5. Nielsen and Hadjeres introduce a series of computationally friendly information-geometric dualistic manifolds approximating a computationally untractable statistical manifold in their paper "Monte Carlo Information-Geometric Structures."
6. Wong in his paper "Information Geometry in Portfolio Theory" applies the principle of information geometry to financial problems.
7. Maroufy and Marriott describe the use of information geometry for Cox regression in survival analysis in their paper entitled "Generalising Frailty Assumptions in Survival Analysis: A Geometric Approach."
8. Broniatowski and Stummer present a unifying view of dissimilarities in their detailed paper "Some Universal Insights on Divergences for Statistics, Machine Learning and Artificial Intelligence."
9. Chirikjian studies information theory in interaction with Lie groups in his paper called "Information-Theoretic Matrix Inequalities and Diffusion Processes on Unimodular Lie Groups."
10. Said, Bombrun, and Berthoumieu proved that the Fisher-Rao information metric of any location-scale model is a warped Riemannian metric provided that the model is invariant under the action of some Lie group in their paper "Warped Riemannian Metrics for Location-Scale Models."
11. Nielsen and Sun propose to use Hilbert's projective geometry for modeling the probability simplex and the elliptope of correlation matrices in their paper entitled "Clustering in Hilbert's Projective Geometry: The Case Studies of the Probability Simplex and the Elliptope of Correlation Matrices."
12. Finally, Barbaresco presents an introduction on Koszul's pioneering work on homogeneous bounded domains that has revealed itself as the elementary structures of information geometry in his paper "Jean-Louis Koszul and the Elementary Structures of Information Geometry."

I would also like to thank Christoph Baumann and his support staff at Springer Nature that made the publishing process very smooth. Finally, I heartily thank Prof. Hiroaki Kitano (President of Sony Computer Science Laboratories, Inc.) for his kind support and encouragements over the years.

Tokyo, Japan
August 2018

Frank Nielsen
Sony Computer Science Laboratories, Inc.

Contents

Rho-Tau Embedding of Statistical Models

Jan Naudts and Jun Zhang

Abstract Two strictly increasing functions ρ and τ determine the rho-tau embedding of a statistical model. The Riemannian metric tensor is derived from the rho-tau divergence. It depends only on the product $\rho'\tau'$ of the derivatives of ρ and τ. Hence, once the metric tensor is fixed still some freedom is left to manipulate the geometry. We call this the *gauge freedom*. A sufficient condition for the existence of a dually flat geometry is established. It is shown that, if the coordinates of a parametric model are affine then the rho-tau metric tensor is Hessian and the dual coordinates are affine as well. We illustrate our approach using models belonging to deformed exponential families, and give a simple and precise characterization for the rho-tau metric to become Hessian.

1 Introduction

A *statistical manifold* [1, 2, 7] is an abstract manifold $\mathbb{M}$ equipped with a Riemannian metric g and an Amari—Chentsov tensor T. If the manifold is a smooth differentiable manifold then it can be realized [8] as a *statistical model*.

Most studies of statistical models are based on the widely used logarithmic embedding of probability density functions. Here, more generally embeddings are considered. Recent work [11, 12, 23] unifies the formalism of rho-tau embeddings [19] with statistical models belonging to deformed exponential families [10]. The present exposition continues this investigation.

The notion of a statistical manifold has been generalized in the non-parametric setting [14, 15, 20, 21] to include Banach manifolds. The corresponding terminology is used here, although up to now only a few papers have combined non-parametric manifolds with deformed exponential families [9, 13, 16, 18].

J. Naudts (✉)
Universiteit Antwerpen, Antwerpen, Belgium
e-mail: jan.naudts@uantwerpen.be

J. Zhang
University of Michigan, Ann Arbor, MI, USA
e-mail: junz@umich.edu

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication
Technology, https://doi.org/10.1007/978-3-030-02520-5_1

The rho-tau divergence is discussed in the next section. Eguchi [4, 5] proved under rather general conditions that, given a differentiable manifold, a divergence function defines a metric tensor and a pair of connections. These are derived in Sect. 4, respectively Sect. 6. Parametric statistical models are discussed in Sect. 7, which discusses Hessian geometry, and Sect. 8, which deals with deformed exponential families.

2 The Statistical Manifold

The points of a given statistical manifold $\mathbb{M}$ are assumed to be random variables over some measure space $(\mathscr{X}, \mu)$. A random variable X is defined as any measurable real function. The expectation, if it exists, is denoted $\mathbb{E}_\mu X$. Throughout the text it is assumed that the manifold is differentiable and that for each X in $\mathbb{M}$ the tangent plane $T_X\mathbb{M}$ is well-defined.

The derivative of a random variable is again a random variable. Therefore one can expect that the tangent vectors at a point X of $\mathbb{M}$ are random variables with vanishing expectation value. Let us assume that these tangent vectors can be used as a local chart in the vicinity of the point X and that they belong to some Banach space $\mathscr{B}$. Then $\mathbb{M}$ is a Banach manifold, provided a number of technical conditions are satisfied.

In the simplest case the manifold $\mathbb{M}$ consists of all strictly positive probability distributions on a discrete set $\mathscr{X}$. These probability distributions can be considered as positive-valued random variables with expectation equal to 1. The space $\mathscr{B}$ of all random variables is a Banach space for instance for the L^1 norm. The manifold $\mathbb{M}$ is a Banach manifold. Our approach here is the same as that adopted in [21], where random variables are called χ-functions, and functions of random variables are called χ-functionals.

In the more general situation the choice of an appropriate norm for the tangent vectors is not so simple. See the work of Pistone et al. [14–16].

3 Rho-Tau Divergence

Given a strictly convex differentiable function h and a pair of real-valued random variables P and Q the Bregman divergence [3] is given by

$$\mathscr{D}(P, Q) = \mathbb{E}_\mu \left[h(P) - h(Q) - (P - Q)h'(Q)\right], \tag{1}$$

where h' denotes the derivative of h. A generalization involving two strictly increasing real functions $\rho(u)$ and $\tau(u)$ is proposed in [19]. For the sake of completeness the definition is repeated here. Throughout the text these functions ρ and τ are assumed to be at least once, sometimes twice differentiable.

There exists a strictly convex function f with the property that $f' \circ \rho = \tau$. It is given by

$$f(u) = \int^{\rho^{-1}(u)} \tau(v)\mathrm{d}\rho(v). \tag{2}$$

The convex conjugate function f^* is therefore given by

$$f^*(u) = \int^{\tau^{-1}(u)} \rho(v)\mathrm{d}\tau(v), \tag{3}$$

provided the lower boundary of the integrals is chosen appropriately.

The original definition [19] of the rho-tau divergence can be written as

$$\mathscr{D}_{\rho,\tau}(P, Q) = \mathbb{E}_\mu \left[f(\rho(P)) + f^*(\tau(Q)) - \rho(P)\tau(Q)\right] \tag{4}$$

which is assumed to be $\leq +\infty$. The reformulation given below simplifies the proof of some of its properties.

Definition 1 Let be given two strictly increasing differentiable functions ρ and τ, defined on a common open interval D in $\mathbb{R}$. The rho-tau divergence of two random variables P and Q with values in D is given by

$$\mathscr{D}_{\rho,\tau}(P, Q) = \mathbb{E}_\mu \left(\int_Q^P [\tau(v) - \tau(Q)]\, \mathrm{d}\rho(v)\right). \tag{5}$$

This definition is equivalent to (4). To see this, split (5) into two parts. Use (2) to write the former contribution as $\mathbb{E}_\mu f \circ \rho(P) - \mathbb{E}_\mu f \circ \rho(Q)$ and the latter as $-\mathbb{E}_\mu \tau(Q)[\rho(P) - \rho(Q)]$. Use partial integration to prove that $f \circ \rho + f^* \circ \tau = \rho\tau$. This definition also generalizes (1). To see this take $I = f$, $\rho = \mathrm{id}$, and $\tau = I'$.

Note that the integral in (5) is a Stieltjes integral, which is well-defined because ρ and τ are strictly increasing functions. The result is non-negative. Hence, the μ-expectation is either convergent or it diverges to $+\infty$.

Let P and Q be two random variables with joint probability distribution $p(\zeta, \eta)$. Then (5) can be written as

$$\begin{aligned}\mathscr{D}_{\rho,\tau}(P, Q) &= \int p(\zeta, \eta)\mathrm{d}\zeta\mathrm{d}\eta \left(\int_\eta^\zeta [\tau(v) - \tau(\eta)]\, \mathrm{d}\rho(v)\right)\\ &\leq \int p(\zeta, \eta)\mathrm{d}\zeta\mathrm{d}\eta\, |\tau(\zeta) - \tau(\eta)|\, |\rho(\zeta) - \rho(\eta)|\\ &\leq \left\{\mathbb{E}_\mu|\tau(P) - \tau(Q)|^2\mathbb{E}_\mu|\rho(P) - \rho(Q)|^2\right\}^{1/2}. \end{aligned} \tag{6}$$

To obtain the latter the Cauchy–Schwarz inequality is used.

Theorem 1 $\mathscr{D}_{\rho,\tau}(P, Q) \geq 0$ *with equality if* $P = Q$. *If* μ *is faithful, i.e.* $\mathbb{E}_\mu P = 0$ *implies* $P = 0$ *for any non-negative* P, *then* $\mathscr{D}_{\rho,\tau}(P, Q) = 0$ *implies* $P = Q$.

Proof From (5) it is immediately clear that $\mathscr{D}_{\rho,\tau}(P, Q) \geq 0$ and $\mathscr{D}_{\rho,\tau}(P, P) = 0$. Assume now that $\mathscr{D}_{\rho,\tau}(P, Q) = 0$. By assumption this implies that

$$\int_Q^P [\tau(v) - \tau(Q)]\, \mathrm{d}\rho(v) = 0 \quad \mu\text{-almost everywhere.}$$

However, because τ and ρ are strictly increasing the integral is strictly positive unless $P = Q$, μ-almost everywhere. □

It can be easily verified that the rho-tau divergence satisfies the following generalized Pythagorean equality for any three points P, Q, R

$$\mathscr{D}_{\rho,\tau}(P, Q) + \mathscr{D}_{\rho,\tau}(Q, R) - \mathscr{D}_{\rho,\tau}(P, R) = \mathbb{E}_\mu \left\{[\rho(P) - \rho(Q)][\tau(R) - \tau(Q)]\right\}.$$

The general expression for the rho-tau entropy is

$$S_{\rho,\tau}(P) = -\mathbb{E}_\mu f(\rho(P)) + \text{ constant} = -\mathbb{E}_\mu \int^P \tau(u)\mathrm{d}\rho(u). \tag{7}$$

See for instance Section 2.6 of [23]. The function f is a strictly convex function which, given ρ, can still be chosen arbitrarily and then determines τ. The following identity holds

$$\mathscr{D}_{\rho,\tau}(P, Q) = -S_{\rho,\tau}(P) + S_{\rho,\tau}(Q) - \mathbb{E}_\mu [\rho(P) - \rho(Q)]\, \tau(Q). \tag{8}$$

In [12, 23], we also discuss rho-tau cross-entropy, as well as the notion of "dual entropy" arising out of rho-tau embedding.

Rho-tau divergence $\mathscr{D}_{\rho,\tau}(P, Q)$ is a special form of the more general divergence function $\mathscr{D}^{(\alpha)}_{f,\rho}(P, Q)$ arising out of convex analysis, see [19, 20]:

$$\begin{aligned}\mathscr{D}^{(\alpha)}_{f,\rho}(P, Q) &= \frac{4}{1-\alpha^2}\\ &\times \mathbb{E}_\mu \left\{\frac{1-\alpha}{2} f(\rho(P)) + \frac{1+\alpha}{2} f(\rho(Q)) - f\left(\frac{1-\alpha}{2}\rho(P) + \frac{1+\alpha}{2}\rho(Q)\right)\right\}.\end{aligned} \tag{9}$$

Clearly

$$\lim_{\alpha\to 1} \mathscr{D}^{(\alpha)}_{f,\rho}(P, Q) = \mathscr{D}_{\rho,\tau}(P, Q) = \mathscr{D}_{\tau,\rho}(Q, P);$$
$$\lim_{\alpha\to -1} \mathscr{D}^{(\alpha)}_{f,\rho}(P, Q) = \mathscr{D}_{\rho,\tau}(Q, P) = \mathscr{D}_{\tau,\rho}(P, Q);$$

with $f' \circ \rho = \tau$ (and equivalent $(f^*)' \circ \tau = \rho$, with f^* denoting convex conjugate of f). Though in $\mathscr{D}^{(\alpha)}_{f,\rho}(P, Q)$ the two free functions are f (a strictly convex function) and ρ (a strictly monotone increasing function), as reflected in its subscripts, there is only notational difference from the ρ, τ specification of two function's choice. This is because for f, f^*, ρ, τ, a choice of any two functions (one of which would have to be either ρ or τ) would specify the remaining two. See [19, 22].

4 Tangent Vectors

The rho-tau divergence introduced above can be used to fix a Riemannian metric on the tangent planes of the statistical manifold $\mathbb{M}$.

In the standard situation of the Fisher-Rao metric the point P is a probability density function p^θ, parametric with $\theta \in \mathbb{R}^n$. A short calculation gives

$$\partial_j \mathbb{E}_\mu p^\theta Y = \left\langle \partial_j \log p^\theta, Y \right\rangle_\theta, \tag{10}$$

with $\langle X, Y\rangle_\theta = \mathbb{E}_\mu p^\theta XY$, and where ∂_j is an abbreviation for $\partial/\partial\theta^j$. The metric tensor is then given by

$$g_{ij}(\theta) = \langle \partial_i \log p^\theta, \partial_j \log p^\theta \rangle_\theta.$$

The score variables $\partial_j \log p^\theta$ have vanishing expectation and span the tangent plane at the point p^θ.

These expressions are now generalized. Fix P in $\mathbb{M}$. Make the assumption that there exists some open neighborhood U of P in $\mathbb{M}$ and a one-to-one correspondence χ_P between elements Q of U and tangent vectors $X = \chi_P(Q)$ of $T_P\mathbb{M}$, satisfying $\chi_P(P) = 0$. This map χ_P is used as a local chart centered at the point P. The directional derivative d_X is then defined as

$$d_X P := \lim_{\varepsilon \to 0} \frac{\chi_P^{-1}(\varepsilon X) - \chi^{-1}(0)}{\varepsilon},$$

and is assumed to exist for all $X \in T_P\mathbb{M}$. Here, we leave the topology unspecified.

Now we take one of the two increasing functions ρ and τ, say ρ, to define a two-point correlation function $\mathbb{E}_\mu \rho(P)Y$, and the other function, τ, to act as a deformed logarithmic function replacing the logarithmic function which appears in the definition of the standard scores. The expression analogue to (10) now involves derivatives of $\mathbb{E}_\mu \rho(P)Y$ and of $\tau(P)$. It becomes

$$d_X \mathbb{E}_\mu \rho(P) Y = \left\langle d_X \tau(P), Y \right\rangle_P, \tag{11}$$

with

$$\langle X, Y\rangle_P = \mathbb{E}_\mu \frac{\rho'(P)}{\tau'(P)} XY.$$

This relation should hold for any P in $\mathbb{M}$ and X in $T_P\mathbb{M}$, and for any random variable Y. The metric tensor $g_{XY} \equiv g(X, Y)$ becomes

$$\begin{aligned} g_{XY}(P) &= \left\langle d_X\tau(P), d_Y\tau(P)\right\rangle_P \\ &= \mathbb{E}_\mu \rho'(P)\tau'(P)d_X P d_Y P. \end{aligned} \tag{12}$$

This metric tensor is related to the divergence function introduced in the previous section by

$$d^P_Y d^Q_X \mathscr{D}_{\rho,\tau}(P, Q)\bigg|_{P=Q} = -g_{XY}(P),$$

where d^P is the derivative acting only on P and d^Q acts only on Q. See [21] for the derivation of the metric tensor in the form of (12) for the non-parametric setting.

In the case of a model p^θ which belongs to the exponential family the tangent plane can be identified with the coordinate space. The chart becomes $\chi_{p^\theta}(p^\zeta) = \zeta - \theta$ so that

$$d_\zeta p^\theta := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(p^{\theta+\varepsilon(\zeta-\theta)} - p^\theta\right).$$

If $(\zeta - \theta)_i = \delta_{i,j}$ then $d_\zeta p^\theta = \partial_j p^\theta$ follows and (11) reduces to (10).

5 Gauge Freedom

From (12) it is clear that the metric tensor depends only on the product $\rho'\tau'$ and not on ρ and τ separately. This implies that once the metric tensor is fixed there remains one function to be chosen freely, either the embedding ρ or the deformed logarithm τ, keeping $\rho'\tau'$ fixed. This is what we call the gauge freedom of the rho-tau formalism.

The notion of gauge freedom is common in Physics to mark the introduction of additional degrees of freedom which do not modify the model but control some of its appearances. Here, the Riemannian metric of the manifold is considered to be an essential feature while the different geometries such as the Riemannian geometry or Amari's dually flat geometries are attributes which give a further characterization.

It is known for long that distinct choices of the divergence function can lead to the same metric tensor. The present formalism offers the opportunity to profit from this freedom. Quantities such as the divergence function, the entropy or the alpha-family of connections depend on the specific choice of both ρ and τ. This is illustrated further on. Some examples are found in Table 1.

The simplest choice to fix the gauge is $\rho = \text{id}$. Several classes generalizing Bregman divergences found in the literature, e.g. [6, 10], belong to this case. The phi-divergence of [10] is obtained by choosing τ equal to the deformed logarithm $\log_\phi$ (see Sect. 8), the derivative of which is $1/\phi$. This implies $\rho'\tau' = 1/\phi$, which is also the condition for the deformed metric tensor of [10] to be conformally equivalent with

Table 1 Examples of ρ, τ combinations

$\rho(u)$	$\tau(u)$	$(\rho'\tau')(u)$	$f(u)$	$f^*(u)$
u	$\log u$	$\frac{1}{u}$	$u[\log u - 1]$	e^u
$2\sqrt{u}$	$2\sqrt{u}$	$\frac{1}{u}$	$\frac{1}{2}u^2$	$\frac{1}{2}u^2$
u	$\log_q(u)$	$\frac{1}{u^q}$	$\frac{u}{2-q}\left[\log_q(u) - 1\right]$	$\frac{1}{2-q}\left[\exp_q(u)\right]^{2-q}$
$\rho(u)$	$\log_\rho(u)$	$\frac{\rho'}{\rho}(u)$	$u[\log u - 1]$	e^u
u	$\log_\phi(u)$	$\frac{1}{\phi(u)}$	$u\log_\phi(u) - \int_1^u \frac{v}{\phi(v)}\mathrm{d}v$	$\int_1^{\exp_\phi(u)} \frac{v}{\phi(v)}\mathrm{d}v$

(12). The U-divergence of [6] is obtained by taking τ equal to the inverse function of U'. These were discussed in detail in [11, 12, 23].

Also of interest is the gauge defined by $\rho(u) = 1/\tau'(u)$. Let $\log_\rho$ be the corresponding deformed logarithm (see (21) below). It satisfies $\log_\rho(u) = \tau(u) - \tau(1)$. Hence, the entropy becomes

$$S_{\rho,\tau}(P) = -\mathbb{E}_\mu \rho(P)\tau(P) + \mathbb{E}_\mu P + \text{ constant.}$$

The divergence becomes

$$\mathscr{D}_{\rho,\tau}(P,Q) = \mathbb{E}_\mu \rho(P)\left[\log_\rho(P) - \log_\rho(Q)\right] - \mathbb{E}_\mu\left[P - Q\right].$$

This expression is an obvious generalization of the Kullback–Leibler divergence.

6 Induced Geometry

A divergence function not only fixes a metric tensor by taking two derivatives, it also fixes a pair of torsion-free connections by taking an extra derivative w.r.t. the first argument [4, 5]. In particular, the rho-tau-divergence (5) determines an alpha-family of connections [11, 19, 21].

A covariant derivative ∇_Z with respect to a vector field Z is defined by

$$\langle \nabla_Z d_X \tau(P), d_Y \tau(P)\rangle_P = -d_Z^P d_Y^P d_X^Q \mathscr{D}_{\rho,\tau}(P,Q)\Big|_{Q=P}.$$

A short calculation of the righthand side, with $\mathscr{D}_{\rho,\tau}$ defined by (4), gives

$$\langle \nabla_Z d_X \tau(P), d_Y \tau(P)\rangle_P = \mathbb{E}_\mu\left[d_X \tau(P)\right] d_Z d_Y \rho(P).$$

Let $\nabla_Z^{(1)} = \nabla_Z$ and let $\nabla_Z^{(-1)}$ be the operator obtained by interchanging ρ and τ. This is

$$\begin{aligned}\langle \nabla_Z^{(-1)} d_X \tau(P), d_Y \tau(P)\rangle_P &= \mathbb{E}_\mu \left[d_X \rho(P)\right] \, d_Z d_Y \tau(P) \\ &= \langle d_X \tau(P), d_Z d_Y \tau(P)\rangle_P. \end{aligned} \tag{13}$$

This shows that $\nabla_Z^{(-1)}$ is the adjoint of d_Z with respect to g. In addition one has

$$\langle \nabla_Z^{(1)} d_X \tau(P), d_Y \tau(P)\rangle_P + \langle d_X \tau(P), \nabla_Z^{(-1)} d_Y \tau(P)\rangle_P = d_Z \, g_{XY}(P). \tag{14}$$

The latter expression shows that the connections $\nabla^{(1)}$ and $\nabla^{(-1)}$ are the dual of each other with respect to g. The alpha-family of connections is then obtained by linear interpolation with $\alpha \in [-1, 1]$

$$\nabla_Z^{(\alpha)} = \frac{1+\alpha}{2} \nabla_Z^{(1)} + \frac{1-\alpha}{2} \nabla_Z^{(-1)}, \tag{15}$$

such that the covariant derivatives $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are mutually dual. In particular, $\nabla^{(0)}$ is self-dual and therefore coincides with the Levi-Civita connection. The family of α-connections (15) is induced by the divergence function $\mathscr{D}_{f,\rho}^{(\alpha)}(P, Q)$ given by (9), with corresponding α-values. Furthermore, upon switching $\rho \leftrightarrow \tau$ in the divergence function, the designation of 1-connection vs (-1)-connection also switches.

From (13) it is clear that the covariant derivative $\nabla_Z^{(-1)}$ vanishes on the tangent plane when

$$\langle d_X \tau(P), d_Z d_Y \tau(P)\rangle_P = 0 \quad \text{for all } X, Y \in T_P\mathbb{M}. \tag{16}$$

If this holds for all P in $\mathbb{M}$ then the $\nabla^{(-1)}$-geometry is flat. This implies that the dual geometry $\nabla^{(1)}$ is also flat — see Theorem 3.3 of [1]. The interpretation of (16) is that all second derivatives $d_Z d_Y \tau(P)$ are orthogonal to the tangent plane.

7 Parametric Models

The previous sections deal with the geometry of arbitrary manifolds consisting of random variables, without caring whether they possess special properties. Now parametric models with a Hessian metric g are considered.

From here on the random variables of the manifold $\mathbb{M}$ are probability distribution functions p^θ, labeled with coordinates θ belonging to some open convex subset U of $\mathbb{R}^n$. The manifold is assumed to be differentiable. In particular, the θ^i are covariant coordinates and the assumption holds that the derivatives $\partial_i p^\theta \equiv \partial p^\theta / \partial \theta^i$ form a basis for the tangent plane $T_\theta \mathbb{M} \equiv T_{p^\theta}\mathbb{M}$. The simplifications induced by this setting are that the tangent planes are finite-dimensional and that the dual coordinates belong again to $\mathbb{R}^n$. For general Banach manifolds both properties need not to hold. The assumptions imply that the metric tensor

$$g_{ij}(\theta) = \langle \partial_i \tau(p^\theta), \partial_j \tau(p^\theta) \rangle_\theta$$

is a strictly positive-definite matrix.

The metric g of the manifold $\mathbb{M}$ is said to be Hessian if there exists a strictly convex function $\Phi(\theta)$ with the property that $g_{ij}(\theta) = \partial_i \partial_j \Phi(\theta)$. See for instance [17]. Let $\Psi(\eta)$ denote the convex dual of $\Phi(\eta)$. This is

$$\Psi(\eta) = \sup_\theta \{\langle \eta, \theta \rangle - \Phi(\theta) : \theta \in U\}.$$

Let U^* denote the subset of $\mathbb{R}^n$ of η for which the maximum is reached at some θ in U. This θ is unique and defines a bijection $\theta \mapsto \eta$ between U and U^*. These η are dual coordinates for the manifold $\mathbb{M}$. Conversely [11], if there exist coordinates η_i for which $g_{ij}(\theta) = \partial_j \eta_i$ then the rho-tau metric tensor g is Hessian.

The condition (16) for $\nabla^{(-1)}$ to vanish can now be written as

$$\langle \partial_i \tau(p^\theta), \partial_k \partial_j \tau(p^\theta) \rangle_\theta = 0, \qquad \text{for all } \theta \in U \text{ and for all } i, j, k. \tag{17}$$

Theorem 2 *Assume that the θ^i are affine coordinates such that $\nabla^{(-1)} = 0$. Then*

(1) the metric tensor g is Hessian;
(2) the η_i are affine coordinates for the $\nabla^{(1)}$-geometry.

Proof (1) The metric tensor (12) becomes

$$g_{ij}(p^\theta) = \langle \partial_i \tau(p^\theta), \partial_j \tau(p^\theta) \rangle_\theta = \mathbb{E}_\mu \left(\partial_i \tau(p^\theta)\right) \partial_j \rho(p^\theta) = \mathbb{E}_\mu \left(\partial_j \tau(p^\theta)\right) \partial_i \rho(p^\theta).$$

This implies

$$\partial_k g_{ij}(p^\theta) = \mathbb{E}_\mu \left(\partial_k \partial_i \tau(p^\theta)\right) \partial_j \rho(p^\theta) + \mathbb{E}_\mu \left(\partial_i \tau(p^\theta)\right) \partial_k \partial_j \rho(p^\theta),$$

but also

$$\partial_k g_{ij}(p^\theta) = \mathbb{E}_\mu \left(\partial_k \partial_j \tau(p^\theta)\right) \partial_i \rho(p^\theta) + \mathbb{E}_\mu \left(\partial_j \tau(p^\theta)\right) \partial_k \partial_i \rho(p^\theta).$$

These equations simplify by means of (17). The result is

$$\partial_k g_{ij}(p^\theta) = \mathbb{E}_\mu \left(\partial_i \tau(p^\theta)\right) \partial_k \partial_j \rho(p^\theta) = \mathbb{E}_\mu \left(\partial_j \tau(p^\theta)\right) \partial_k \partial_i \rho(p^\theta).$$

This implies that $\partial_k g_{ij}(\theta) = \partial_i g_{kj}(\theta)$. Hence there exist functions $\eta_j(\theta)$ such that $g_{ij}(\theta) = \partial_i \eta_j(\theta)$. As remarked above, it is proved in [11] that this suffices to conclude that the metric g is Hessian.

(2) Let us show that

$$\eta(t) = (1-t)\eta^{(1)} + t\eta^{(2)}. \tag{18}$$

is a solution of the Euler-Lagrange equations

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta^i + \Gamma^i_{km}\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right) = 0. \tag{19}$$

Here, the Γ^i_{km} are the coefficients of the connection $\Gamma^{(1)}$ induced by the $\nabla^{(1)}$-geometry. They follow from

$$\Gamma_{ij,k} = \partial_i g_{jk}(\theta). \tag{20}$$

One has

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^i = \frac{\partial\theta^i}{\partial\eta_j}\frac{\mathrm{d}\eta_j}{\mathrm{d}t} = g^{ij}(\theta)\left[\eta^{(2)}_j - \eta^{(1)}_j\right]$$

and

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta^i &= \frac{\mathrm{d}}{\mathrm{d}t}g^{ij}(\theta)\left[\eta^{(2)}_j - \eta^{(1)}_j\right]\\ &= \left[\partial_k g^{ij}(\theta)\right]\frac{\mathrm{d}\theta^k}{\mathrm{d}t}\left[\eta^{(2)}_j - \eta^{(1)}_j\right]\\ &= \left[\partial_k g^{ij}(\theta)\right]g^{kl}(\theta)\left[\eta^{(2)}_l - \eta^{(1)}_l\right]\left[\eta^{(2)}_j - \eta^{(1)}_j\right]\\ &= \left[\partial_k g^{ij}(\theta)\right]g_{jm}(\theta)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right).\end{aligned}$$

The l.h.s. of (19) becomes

$$\text{l.h.s.} = \left\{\left[\partial_k g^{ij}(\theta)\right]g_{jm}(\theta) + \Gamma^i_{km}\right\}\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^k\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\theta^m\right).$$

This vanishes because (20) implies

$$\Gamma^i_{km} = -\left[\partial_k g^{ij}(\theta)\right]g_{jm}(\theta).$$

□

It is important to realize that the discussion in this section is generic for parametric models, without assuming particular parametric families.

8 The Deformed Exponential Family

A repeated measurement of n independent random variables $F_1, \dots, F_n$ results in a joint probability distribution $\pi(\zeta_1, \dots, \zeta_n)$, which describes the probability that the true value of the measured data equals ζ. More generally, the model can be taken to be a deformed exponential family, obtained by using a deformed exponential function $\exp_\phi$. Following [10], a deformed logarithm $\log_\phi$ is defined by

$$\log_\phi(u) = \int_1^u \mathrm{d}v \, \frac{1}{\phi(v)}, \tag{21}$$

where $\phi(v)$ is strictly positive and integrable on the open interval $(0, +\infty)$. The deformed exponential function $\exp_\phi(u)$ is the inverse function of $\log_\phi(u)$. It is defined on the range $\mathscr{R}$ of $\log_\phi(u)$, but is eventually extended with the value 0 if $u < \mathscr{R}$ and with the value $+\infty$ if $u > \mathscr{R}$.

The expression for the probability density function then becomes

$$p^\theta(x) = \exp_\phi\left(\sum_{k=1}^n \theta^k F_k(x) - \alpha(\theta)\right). \tag{22}$$

The function $\alpha(\theta)$ serves to normalize p^θ and is assumed to exist within the open convex domain $U \subset \mathbb{R}^n$ in which the model is defined. One can show [10] that it is a convex function. However, in general it does not coincide with the potential $\Phi(\theta)$ of the previous section. The explanation is that *escort probabilities* come into play. Indeed, from

$$0 = \partial_i \mathbb{E}_\mu p^\theta = \mathbb{E}_\mu \phi(p^\theta) \left[F_i - \partial_i \alpha\right]$$

follows that

$$\partial_i \alpha = \tilde{\mathbb{E}}_\theta F_i,$$

with the escort expectation $\tilde{\mathbb{E}}_\theta$ defined by

$$\tilde{\mathbb{E}}_\theta Y = \frac{\mathbb{E}_\mu \phi(p^\theta) Y}{\mathbb{E}_\mu \phi(p^\theta)}.$$

Only in the non-deformed case, when $\phi(u) = u$, the escort $\tilde{\mathbb{E}}_\theta$ coincides with the model expectation $\mathbb{E}_\theta$. Then the dual coordinates η_i satisfy $\eta_i = \mathbb{E}_\theta F_i = \partial_i \alpha(\theta)$.

In general, the rho-tau metric tensor g of the deformed exponential model is *not* Hessian.We have the following Theorem (see [12])

Theorem 3 *With respect to the (deformed) ϕ-exponential family p^θ obeying (22), the rho-tau metric tensor g is*

(a) conformal to Hessian if

$$\rho' \tau' \phi = \phi';$$

(b) Hessian if

$$\rho'\tau'\phi = id.$$

In case (a), the rho-tau metric tensor is conformally equivalent with the metric tensor obtained by taking the Hessian of the normalization function α; in case (b) the potential $\Phi(\theta)$ is constructed in [10]. However, there still leaves a gauge freedom. The question is then whether one can choose ρ and τ so that condition (16) for the dually flat geometry is satisfied. A sufficient condition is that $\rho = \text{id}$ and $\tau = \log_\phi$. This is the rho-affine gauge. In this gauge both the θ^i and the η_i coordinates are affine and the model has a dually flat structure.

Acknowledgements The research reported here is supported by DARPA/ARO Grant W911NF-16-1-0383 (PI: Jun Zhang).

References

1. Amari, S., Nagaoka, H.: Methods of Information Geometry. AMS Monograph. Oxford University Press, Oxford (2000). (Originally published in Japanese by Iwanami Shoten, Tokyo, Japan, 1993.)
2. Ay, N., Jost, J., LÊ, H.V., Schwachhöfer, L.: Information Geometry. Springer, Berlin (2017)
3. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Phys. **70**, 200–217 (1967)
4. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**, 793–803 (1983)
5. Eguchi, S.: A differential geometric approach to statistical inference on the basis of contrast functionals. Hiroshima Math. J. **15**, 341–391 (1985)
6. Eguchi, S.: Information geometry and statistical pattern recognition. Sugaku Expositions (Amer. Math. Soc.) **19**, 197–216 (2006). (originally Sūgaku 56 (2004) 380 in Japanese)
7. Lauritzen, S.: Statistical manifolds. In: Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R. (eds.) Differential Geometry in Statistical Inference. Lecture Notes, vol. 10, pp. 163–216. IMS, Hayward (1987)
8. Lê, H.V.: Statistical manifolds are statistical models. J. Geom. **84**, 83–93 (2005)
9. Montrucchio, L., Pistone, G.: Deformed exponential bundle: the linear growth case. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2017 LNCS Proceedings of Geometric Science of Information, pp. 239–246. Springer, Berlin (2017)
10. Naudts, J.: Estimators, escort probabilities, and phi-exponential families in statistical physics. J. Ineq. Pure Appl. Math. **5**, 102 (2004)
11. Naudts, J., Zhang, J.: Information geometry under monotone embedding. Part II: Geometry. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2017 LNCS Proceedings of Geometric Science of Information, pp. 215–222. Springer, Berlin (2017)
12. Naudts, J., Zhang J.: Information geometry under monotone embedding. Inf. Geom. (under review)
13. Newton, N.J.: An infinite-dimensional statistical manifold modeled on Hilbert space. J. Funct. Anal. **263**, 1661–1681 (2012)
14. Pistone, G., Sempi, C.: An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **33**, 1543–1561 (1995)
15. Pistone, G., Rogantin, M.P.: The exponential statistical manifold: mean parameters, orthogonality and space transformations. Bernoulli **5**, 721–760 (1999)

16. Pistone, G.: κ-exponential models from the geometrical viewpoint. Eur. Phys. J. B **70**, 29–37 (2009)
17. Shima, H.: The Geometry of Hessian Structures. World Scientific, Singapore (2007)
18. Vigelis, R.F., Cavalcante, C.C.: On ϕ-families of probability distributions. J. Theor. Probab. **26**, 870–884 (2013)
19. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**, 159–195 (2004)
20. Zhang, J.: Referential duality and representational duality on statistical manifolds. In: Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, Japan, pp. 58–67 (2005)
21. Zhang, J.: Nonparametric information geometry: from divergence function to referential-representational biduality on statistical manifolds. Entropy **15**, 1 (2013)
22. Zhang, J.: On monotone embedding in information geometry. Entropy **17**, 4485–4499 (2015)
23. Zhang, J., Naudts, J.: Information geometry under monotone embedding. Part I: Divergence functions. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2017 LNCS Proceedings of Geometric Science of Information, pp. 205–214. Springer, Berlin (2017)

A Class of Non-parametric Deformed Exponential Statistical Models

Luigi Montrucchio and Giovanni Pistone

Abstract We study the class on non-parametric deformed statistical models where the deformed exponential has linear growth at infinity and is sub-exponential at zero. This class generalizes the class introduced by N.J. Newton. We discuss the convexity and regularity of the normalization operator, the form of the deformed statistical divergences and their convex duality, the properties of the escort densities, and the affine manifold structure of the statistical bundle

1 Introduction

In this paper we study a geometry on the set $\mathscr{P}$ of strictly positive probability densities on a probability space $(\mathbb{X}, \mathscr{X}, \mu)$. In some cases one is led to consider the set $\overline{\mathscr{P}}$ of probability densities i.e., without the restriction of strict positivity. There is a considerable literature on the Information Geometry in the sense defined in the Amari and Nagaoka monograph [2] on $\mathscr{P}$. There is also a non-parametric approach i.e., we are not considering the geometry induced on the parameter set of a given statistical model but on the full set of densities. This was done in [21, 23] by using logarithmic chart to represent densities.

A different approach, that leads to the construction of an Hilbert manifold on $\mathscr{P}$, has been proposed by N.J. Newton in [18, 19]. It is based on the use of the chart $p \mapsto p - 1 - \log p$ instead of a purely logarithmic chart. This paper presents a variation on the same theme by enlarging the class of permitted charts.

Let $\mathscr{M} \subset \mathscr{P}$. At each $p \in \mathscr{M}$, the Hilbert space of square-integrable random variables $L^2(p)$ provides a fiber that sits at $p \in \mathscr{M}$, so we can define the *Hilbert bundle* with base $\mathscr{M}$. The Hilbert bundle, or similar bundles with fibers which are

L. Montrucchio
Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Turin, Italy
e-mail: luigi.montrucchio@unito.it

G. Pistone (✉)
de Castro Statistics, Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8,
10122 Turin, Italy
e-mail: giovanni.pistone@carloalberto.org

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_2

vector spaces of random variables, provides a convenient framework for Information Geometry, cf. [1, 12, 21].

If $\mathscr{M}$ is an exponential manifold in the sense of [23], there exists a splitting of each fiber $L^2(p) = \mathscr{H}_p \oplus \mathscr{H}_p^\perp$, such that each $\mathscr{H}_p$ contains a dense vector sub-space which is an expression of the tangent space $T_p\mathscr{M}$ of the manifold. Moreover, the manifold on $\mathscr{M}$ is an affine manifold (it can be defined by an atlas whose transition mapping are affine) and it is also an Hessian manifold (the inner product on each fiber is the second derivative of a potential function, [24]).

When the sample space is finite and $\mathscr{M}$ is the full set $\mathscr{P}$ of positive probability densities, then $\mathscr{H}_p$ is the space of centered square integrable random variables $L_0^2(p)$ and moreover there is an identification of the fiber with the tangent space $\mathscr{H}_p \simeq T_p\mathscr{P}$. A similar situation occurs even when $\mathscr{M}$ is a finite-dimensional exponential family. It is difficult to devise set-ups other than those mentioned above, where the identification of the Hilbert fiber with the tangent space holds true. In fact, a necessary condition would be the topological linear isomorphism among fibers. One possible option would be to take as fibers the spaces of bounded functions $L_0^\infty(p)$, see G. Loaiza and H.R. Quiceno [14].

This difficulty is overcome in the N.J. Newton's setting. On a probability space $(X, \mathscr{X}, \mu)$, he considers the "balanced chart" $\mathscr{M} \ni p \mapsto \log p + p - 1 \in L_0^2(\mu)$. In this chart, all the tangent spaces are identified with the fixed Hilbert space $L_0^2(\mu)$ so that the statistical Hilbert bundle is trivialized.

N.J. Newton balanced chart falls in a larger class of "deformation" of the usual logarithmic representation. It is in fact an instance of the class of "deformed logarithm" as defined by J. Naudts [17]. It is defined as $\log_A(x) = \int_1^x dt/A(t)$, where A is a suitable increasing function. If A is bounded, then a special class of deformed logarithms results. It includes N.J. Newton balanced chart as well as other deformed logarithms, notably the G. Kaniadakis logarithm [10, 11, 20].

In this paper, we try a mixture of the various approaches by considering deformed logarithms with linear growth as established by N.J. Newton, but we do not look for a trivialization of the Hilbert bundle. Instead we construct an affine atlas of charts, each one centered at a $p \in \mathscr{M}$. This is obtained by adapting the construction of the exponential manifold of [21] to the deformed exponential models as defined by J. Naudts [17]. Moreover, we allow for a form of general reference measure by using an idea introduced by R.F. Vigelis and C.C. Cavalcante [26]. That is, each density has the form $q = \exp_A(u - K_p(u) + \log_A p)$, where $\exp_A = \log_A^{-1}$ is an exponential-like function which has a linear growth at $+\infty$ and is dominated by an exponential at $-\infty$.

The formalism of deformed exponentials is discussed in Sect. 2. This section is intended to be self-contained and contains material from the references discussed above without an explicit mention. The following Sect. 3 is devoted to the study of non-parametric deformed exponential families. In Sect. 4 we introduce the formulation of the divergence, in accordance with our approach. In Sect. 5 the construction of the Hilbert statistical bundle is outlined.

A first version of this piece of research has been presented at the GSI 2017 Conference [16] and we refer to that paper for some of the proofs.

2 Deformed Exponential

Let us introduce a class of the deformed exponential, according to the formalism introduced by [17]. Assume to be given a function A from $]0,+\infty[$ onto $]0,a[$, strictly increasing, continuously differentiable and such that $\left\|A'\right\|_\infty < \infty$. This implies $a = \|A\|_\infty$ and $A(x) \leq \left\|A'\right\|_\infty x$, so that $\int_0^1 d\xi/A(\xi) \ = +\infty$.

The A-logarithm is the function

$$\log_A(x) = \int_1^x \frac{d\xi}{A(\xi)} , \quad x \in]0,+\infty[\ .$$

The A-logarithm is strictly increasing from $-\infty$ to $+\infty$, its derivative $\log_A'(x) = 1/A(x)$ is positive and strictly decreasing for all $x > 0$, hence $\log_A$ is strictly concave.

By inverting the A-logarithm, one obtains the A-exponential, $\exp_A = \log_A^{-1}$. The function $\exp_A :]-\infty,+\infty[\to]0,+\infty[$ is strictly increasing, strictly convex, and is the solution to the Cauchy problem

$$\exp_A'(y) = A(\exp_A(y)), \quad \exp_A(0) = 1 \ . \tag{1}$$

As a consequence, we have the linear bound

$$\left|\exp_A(y_1) - \exp_A(y_2)\right| \leq \|A\|_\infty |y_1 - y_2| \ . \tag{2}$$

The behavior of the A-logarithm is linear for large arguments and super-logarithmic for small arguments. To derive explicit bounds, set

$$\alpha_1 = \min_{x \leq 1} \frac{A(x)}{x} , \quad \alpha_2 = \max_{x \leq 1} \frac{A(x)}{x} ,$$

namely, they are the best constants such that $\alpha_1 x \leq A(x) \leq \alpha_2 x$ for $0 < x \leq 1$. Note that $\alpha_1 \geq 0$ while $\alpha_2 > 0$. If in addition also $\alpha_1 > 0$, then

$$\frac{1}{\alpha_2} \log x \leq \log_A x \leq \frac{1}{\alpha_1} \log x , \quad 0 < x \leq 1 \ . \tag{3}$$

If otherwise $\alpha_1 = 0$, the left inequality is true only.

For $x \geq 1$ we have $A(1) \leq A(x) < \|A\|_\infty$, hence

$$\frac{1}{\|A\|_\infty}(x-1) < \log_A x \leq \frac{1}{A(1)}(x-1) , \quad x \geq 1 \ . \tag{4}$$

Under the assumptions made on the function A, the coefficient $\alpha_1 > 0$, if and only if $A'(0+) > 0$.

2.1 Examples

The main example of A-logarithm is the N.J. Newton A-logarithm [18], with

$$A(\xi) = 1 - \frac{1}{1+\xi} = \frac{\xi}{1+\xi} \ ,$$

so that

$$\log_A(x) = \log x + x - 1 \ .$$

There is a simple algebraic expression for the product,

$$\log_A(x_1 x_2) = \log_A(x_1) + \log_A(x_2) + (x_1 - 1)(x_2 - 1) \ .$$

Other similar examples are available in the literature. One is a special case of the G. Kaniadakis' exponential of [9], generated by

$$A(\xi) = \frac{2\xi^2}{1+\xi^2} \ .$$

It turns out

$$\log_A x = \frac{x - x^{-1}}{2} \ ,$$

whose inverse provides

$$\exp_A(y) = y + \sqrt{1+y^2} \ .$$

A remarkable feature of the G. Kaniadakis' exponential is the relation

$$\exp_A(y)\exp_A(-y) = \left(y + \sqrt{1+y^2}\right)\left(-y + \sqrt{1+y^2}\right) = 1$$

Notice that the A function for N.J. Newton exponential is concave, while the A function of G. Kaniadakis exponential is not.

Another example is $A(\xi) = 1 - 2^{-\xi}$, which gives $\log_A(x) = \log_2(1 - 2^{-x})$ and $\exp_A(y) = \log_2(1 + 2^y)$.

Notable examples of deformed exponentials that do not fit into our set of assumptions are Tsallis q-logarithms, see [25]. For instance, for $q = 1/2$,

$$\log_{1/2} x = 2\left(\sqrt{x} - 1\right) = \int_1^x \frac{d\xi}{\sqrt{\xi}}.$$

In this case, $\log_{1/2}(0+) = -\int_0^1 d\xi/\sqrt{\xi} = -2$, so that the inverse is not defined for all real numbers. Tsallis logarithms provide models having heavy tails, which is not the case in our setting.

2.2 Superposition Operator

The deformed exponential will be employed to represent positive probability densities in the type $p(x) = \exp_A[u(x)]$, where u is a random variable on a probability space $(\mathbb{X}, \mathscr{X}, \mu)$. For this reason, we are interested in the properties of the *superposition operator*

$$S_A \colon u \mapsto \exp_A \circ u \tag{5}$$

defined in some convenient functional setting. About superposition operators, see e.g. [3, Ch. 1] and [4, Ch. 3].

It is clear from the Lipschitz condition (2) that $\exp_A(u) \le 1 + \|A\|_\infty |u|$, which in turn implies that the superposition operator S_A maps $L^\alpha(\mu)$ into itself for all $\alpha \in [1, +\infty]$ and the mapping is uniformly Lipschitz with constant $\|A\|_\infty$. Notice that we are assuming that μ is a finite measure.

The superposition operator $S_A \colon L^\alpha(\mu) \to L^\alpha(\mu)$ is 1-to-1 and its image consists of all positive random variables f such that $\log_A f \in L^\alpha(\mu)$. The following proposition intercepts a more general result [19]. We give a direct proof here for sake of completeness and because our setting includes deformed logarithms other than the case treated there.

Proposition 1 *1. For all $\alpha \in [1, \infty]$, the superposition operator S_A of Eq. (5) is Gateaux-differentiable with derivative*

$$dS_A(u)[h] = A(\exp_A(u))h\ . \tag{6}$$

2. S_A is Fréchet-differentiable from $L^\alpha(\mu)$ to $L^\beta(\mu)$, for all $\alpha > \beta \ge 1$.

Proof 1. Equation (1) implies that for each couple of random variables $u, h \in L^\alpha(\mu)$

$$\lim_{t \to 0} t^{-1}\left(\exp_A(u + th) - \exp_A(u)\right) - A(\exp_A(u))h = 0$$

holds point-wise. Moreover, if each $\alpha \in [1, \infty[$, by Jensen inequality we infer that if $t > 0$ then

$$\left|t^{-1}\left(\exp_A(u + th) - \exp_A(u)\right) - A(\exp_A(u))h\right|^\alpha \le \\ t^{-1} |h|^\alpha \int_0^t \left|A(\exp_A(u + rh)) - A(\exp_A(u))\right|^\alpha \, dr \le (2\,\|A\|_\infty)^\alpha\, |h|^\alpha\ .$$

Now, dominated convergence forces the limit to hold in $L^\alpha(\mu)$. If $t < 0$, it sufficies to replace h with $-h$.

Whenever $\alpha = \infty$, we can use the second-order bound

$$\left|t^{-1}\left(\exp_A(u+th)-\exp_A(u)\right)-A(\exp_A(u))h\right| = \\ |t|^{-1}h^2\left|\int_0^t (t-r)\frac{d}{dr}A(\exp_A(u+rh))\,dr\right| \le \frac{t}{2}\left\|h\right\|_\infty^2 \left\|A'\right\|_\infty \left\|A\right\|_\infty\ .$$

As $\left\|A'\cdot A\right\|_\infty < \infty$, the RHS goes to 0 as $t \to 0$ uniformly for each $h \in L^\infty(\mu)$.

2. Given $u, h \in L^\alpha(\mu)$, thanks again to Taylor formula,

$$\int \left|\exp_A(u+h)-\exp_A(u)-A(\exp_A(u))h\right|^\beta\ d\mu \le \\ \int |h|^\beta \int_0^1 \left|A(\exp_A(u+rh))-A(\exp_A(u))\right|^\beta\ dr\ d\mu\ .$$

By means of Hölder inequality, with conjugate exponents α/β and $\alpha/(\alpha-\beta)$, the RHS is bounded by

$$\left(\int |h|^\alpha\ d\mu\right)^{\frac{\beta}{\alpha}} \left(\iint \left|A(\exp_A(u+rh))-A(\exp_A(u))\right|^{\frac{\alpha\beta}{\alpha-\beta}}\ dr\ d\mu\right)^{\frac{\alpha-\beta}{\alpha}}\ .$$

Consequently,

$$\left\|h\right\|_{L^\alpha(\mu)}^{-1}\left\|\exp_A(u+h)-\exp_A(u)-A(\exp_A(u))h\right\|_{L^\beta(\mu)} \le \\ \left(\iint \left|A(\exp_A(u+rh))-A(\exp_A(u))\right|^{\frac{\alpha\beta}{\alpha-\beta}}\ dr\ d\mu\right)^{\frac{\alpha-\beta}{\alpha\beta}}\ .$$

In order to show that the RHS vanishes as $\left\|h\right\|_{L^\alpha(\mu)} \to 0$, observe that for all $\delta > 0$ we have

$$\left|A(\exp_A(u+rh))-A(\exp_A(u))\right| \le \begin{cases} 2\left\|A\right\|_\infty & \text{always,} \\ \left\|A'\right\|_\infty \left\|A\right\|_\infty \delta & \text{if } |h| \le \delta, \end{cases}$$

so that, decomposing the double integral as $\iint = \iint_{|h|\le\delta} + \iint_{|h|>\delta}$, we obtain

$$\iint \left|A(\exp_A(u+rh))-A(\exp_A(u))\right|^\gamma\ dr\ d\mu \le \\ (2\left\|A\right\|_\infty)^\gamma\ \mu\left\{|h|>\delta\right\} + \left(\left\|A'\right\|_\infty \left\|A\right\|_\infty \delta\right)^\gamma \le \\ (2\left\|A\right\|_\infty)^\gamma\ \delta^{-\alpha}\int |h|^\alpha\ d\mu + \left(\left\|A'\right\|_\infty \left\|A\right\|_\infty \delta\right)^\gamma\ ,$$

where $\gamma = \alpha\beta/(\alpha-\beta)$ and we have used Cebičev inequality. Now it is clear that the last bound implies the conclusion for each $\alpha < \infty$. The case $\alpha = \infty$ follows a fortiori. □

Remark 1 It is not generally true that the superposition operator S_A be Fréchet differentiable for $\alpha = \beta$, cf. [3, §1.2]. We repeat here the well known counter-example.

Assume μ is a non-atomic probability measure. For each $\lambda \in \mathbb{R}$ and $\delta > 0$ define the simple function

$$h_{\lambda,\delta}(x) = \begin{cases} \lambda & \text{if } |x| \leq \delta, \\ 0 & \text{otherwise.} \end{cases}$$

For each $\alpha \in [1, +\infty[$ we have

$$\lim_{\delta \to 0} \left\| h_{\lambda,\delta} \right\|_{L^\alpha(\mu)} = \lim_{\delta \to 0} |\lambda| \, \mu \left\{ |x| \leq \delta \right\}^{1/\alpha} = 0 \; .$$

Differentiability at 0 in $L^\alpha(\mu)$ would imply for all λ

$$0 = \lim_{\delta \to 0} \frac{\left\| \exp_A(h_{\lambda,\delta}) - 1 - A(1) h_{\lambda,\delta} \right\|_{L^\alpha(\mu)}}{\left\| h_{\lambda,\delta} \right\|_{L^\alpha(\mu)}} = \\ \lim_{\delta \to 0} \frac{\left| \exp_A(\lambda) - 1 - A(1)\lambda \right| \mu \left\{ x \,|\, |x| \leq \delta \right\}^{1/\alpha}}{|\lambda| \, \mu \left\{ x \,|\, |x| \leq \delta \right\}^{1/\alpha}} = \left| \frac{\exp_A(\lambda) - 1}{\lambda} - A(1) \right| \, ,$$

which is a contradiction.

Remark 2 Theorems about the differentiability of the deformed exponential are important because of computations like $\frac{d}{d\theta} \exp_A(v(\theta)) = \exp_A'(v(\theta))\dot{v}(\theta)$ are essential for the geometrical theory of statistical models. Several variations in the choice of the combination domain space - image space are possible. Also, one could look at a weaker differentiability property than Frechét differentiability. Our choice is motivated by the results of the following sections. A large class of cases is discussed in [19].

Remark 3 It would also be worth to study the action of the superposition operator on spaces of differentiable functions, for example Gauss-Sobolev spaces of P. Malliavin [15]. If μ is the standard Gaussian measure on $\mathbb{R}^n$, and u is a differentiable function such that $u, \frac{\partial}{\partial x_i} u \in L^2(\mu)$, $i = 1, \ldots, n$, then it follows that $\exp_A(u) \in L^2(\mu)$ as well as $\frac{\partial}{\partial x_i} \exp_A(u) \in L^2(\mu)$, since

$$\frac{\partial}{\partial x_i} \exp_A(u(x)) = A(\exp_A(u(x)) \frac{\partial}{\partial x_i} u(x) \; .$$

We do not pursue this line of investigation here.

3 Deformed Exponential Family Based on $\exp_A$

According to [5, 26], let us define the deformed exponential curve in the space of positive measures on $(\mathbb{X}, \mathscr{X})$ as follows

$$t \mapsto \mu_t = \exp_A(tu + \log_A p) \cdot \mu \ , \quad u \in L^1(\mu) \ .$$

We have the following inequality:

$$\exp_A(x + y) \le \|A\|_\infty x^+ + \exp_A(y).$$

Actually, it is true for $x \le 0$, as being $\exp_A$ increasing. For $x = x^+ > 0$ the inequality follows from Eq. (2). As a consequence, each μ_t is a finite measure, $\mu_t(\mathbb{X}) \le t\,\|A\|_\infty \int u^+ \ d\mu + 1$, with $\mu_0 = p \cdot \mu$. The curve is actually continuous and differentiable in $L^1(\mu)$ because the point-wise derivative of the density $p_t = \exp_A(tu + \log_A(p))$ is $\dot{p}_t = A(p_t)u$ so that $|\dot{p}_t| \le \|A\|_\infty |u|$. In conclusion $\mu_0 = p \cdot \mu$ and $\dot{\mu}_0 = A\,(p)\,u \cdot \mu$.

There are two ways to normalize the density p_t to total mass 1, either dividing by a normalizing constant $Z(t)$ to get the statistical model $t \mapsto \exp_A(tu + \log_A p)/Z(t)$ or, subtracting a constant $\psi(t)$ from the argument to get the model $t \mapsto \exp_A(tu - \psi(t) + \log_A(p))$. Unlike the standard exponential case, where these two methods lead to the same result, this is not the case for deformed exponentials where $\exp_A(\alpha + \beta) \ne \exp_A(\alpha)\exp_A(\beta)$. We choose in the present paper the latter option.

Here we use the ideas of [5, 17, 26] to construct deformed non-parametric exponential families. Recall that we are given: the probability space $(\mathbb{X}, \mathscr{X}, \mu)$; the set $\mathscr{P}$ of the positive probability densities and the function A satisfying the conditions set out in Sect. 2. Throughout this section, the density $p \in \mathscr{P}$ will be fixed.

The following proposition is taken from [16] where a detailed proof is given.

Proposition 2 *1. The mapping $L^1(\mu) \ni u \mapsto \exp_A(u + \log_A p) \in L^1(\mu)$ has full domain and is $\|A\|_\infty$-Lipschitz. Consequently, the mapping*

$$u \mapsto \int g \exp_A(u + \log_A p)\, d\mu$$

is $\|g\|_\infty \cdot \|A\|_\infty$-Lipschitz for each bounded function g.

2. *For each $u \in L^1(\mu)$ there exists a unique constant $K_p(u) \in \mathbb{R}$ such that $\exp_A(u - K_p(u) + \log_A p) \cdot \mu$ is a probability.*
3. *$K_p(u) = u$ if, and only if, u is constant. In such a case,*

$$\exp_A(u - K_p(u) + \log_A p) \cdot \mu = p \cdot \mu \ .$$

Otherwise, $\exp_A(u - K_p(u) + \log_A p) \cdot \mu \ne p \cdot \mu$.

4. *A density q is of the form $q = \exp_A(u - K_p(u) + \log_A p)$, with $u \in L^1(\mu)$ if, and only if, $\log_A q - \log_A p \in L^1(\mu)$.*

5. *If*
$$\exp_A(u - K_p(u) + \log_A p) = \exp_A(v - K_p(v) + \log_A p) ,$$
with $u, v \in L^1(\mu)$, then $u - v$ is constant.
6. *The functional $K_p \colon L^1(\mu) \to \mathbb{R}$ is translation invariant. More specifically,*
$$K_p(u + c) = K_p(u) + cK_p(1)$$
holds for all $c \in \mathbb{R}$.
7. *$K_p : L^1(\mu) \to \mathbb{R}$ is continuous and convex.*

3.1 Escort Density

For each positive density $q \in \overline{\mathscr{P}}$, its *escort density* is defined as

$$\text{escort}\,(q) = \frac{A(q)}{\int A(q)\,d\mu} ,$$

see [17]. Notice that $0 \leq A(q) \leq \|A\|_\infty$. In particular, $\widetilde{q} = \text{escort}\,(q)$ is a bounded positive density. Hence, $\text{escort}\left(\overline{\mathscr{P}}\right) \subseteq \overline{\mathscr{P}} \cap L^\infty(\mu)$. Clearly, the inclusion $\text{escort}\,(\mathscr{P}) \subseteq \mathscr{P} \cap L^\infty(\mu)$ is true as well.

Proposition 3 *1. The mapping* $\text{escort}\colon \overline{\mathscr{P}} \to \overline{\mathscr{P}} \cap L^\infty(\mu)$ *is a.s. injective.*

2. *A bounded positive density $\widetilde{q}$ is an escort density, i.e., $\widetilde{q} \in \text{escort}\left(\overline{\mathscr{P}}\right)$ if, and only if,*
$$\lim_{\alpha \uparrow \|A\|_\infty} \int A^{-1}\left(\alpha \frac{\widetilde{q}}{\|\widetilde{q}\|_\infty}\right) d\mu \geq 1 . \tag{7}$$
3. *Condition (7) is fulfilled if $\mu\left\{\widetilde{q} = \|\widetilde{q}\|_\infty\right\} > 0$. In particular, every density taking a finite number of different values, i.e., a simple density, is an escort density.*
4. *If $\widetilde{q}_1 = \text{escort}\,(q_1)$ is an escort density, and q_2 is a bounded positive density such that*
$$\mu\left\{\widetilde{q}_1 > t\,\|\widetilde{q}_1\|_\infty\right\} \leq \mu\left\{q_2 > t\,\|q_2\|_\infty\right\}, \quad t > 0 ,$$
then q_2 is an escort density as well.

Proof 1. Let $\text{escort}\,(q_1) = \text{escort}\,(q_2)$ for μ-almost all x. Say, $\int A \circ q_1 \; d\mu \geq \int A \circ q_2 \, d\mu$. Then $A(q_2(x)) \leq A(q_1(x))$, for μ-almost all x. Since A is strictly increasing, it follows $q_2(x) \leq q_1(x)$ for μ-almost all x, which, in turn, implies $q_1 = q_2$ μ-a.s. because both μ-integrals are equal to 1. Thus the escort mapping is a.s. injective.

2. Fix a $\widetilde{q} \in \overline{\mathscr{P}} \cap L^\infty(\mu)$, and define the function

$$f(\alpha) = \int A^{-1}\left(\alpha \frac{\widetilde{q}}{\|\widetilde{q}\|_\infty}\right) \, d\mu, \quad \alpha \in [0, \|A\|_\infty[\, .$$

It is finite, increasing, continuous and $f(0) = 0$. It is clear that the range condition (7) is necessary because $\widetilde{q} = \text{escort}\,(q)$ implies $q = A^{-1}\left(\left(\int A(q)\, d\mu\right) \widetilde{q}\right)$ and, in turn, $1 = \int A^{-1}\left(\left(\int A(q)\, d\mu\right) \widetilde{q}\right) \, d\mu$, given that q is a probability density. If we take $\alpha = \int A(q)\, d\mu \;\; \|\widetilde{q}\|_\infty \leq \|A\|_\infty$, the range condition is satisfied. Conversely, if the range condition holds, there exists $\alpha \leq \|A\|_\infty$ such that $q = A^{-1}\left(\alpha \frac{\widetilde{q}}{\|\widetilde{q}\|_\infty}\right)$ is a positive probability density whose escort is $\widetilde{q}$.

3. This is a special case of Item 2, in that

$$f(\alpha) = \int A^{-1}\left(\alpha \frac{\widetilde{q}}{\|\widetilde{q}\|_\infty}\right) \, d\mu \geq A^{-1}(\alpha)\mu\left\{\widetilde{q} = \|\widetilde{q}\|_\infty\right\} \, .$$

Therefore, $f(\alpha) \uparrow +\infty$, as $\alpha \uparrow \|A\|_\infty$.

4. For each bounded positive density q we have

$$\int A^{-1}\left(\frac{q}{\|q\|_\infty}\right) \, d\mu = \int_0^{+\infty} \mu\left\{\frac{q}{\|q\|_\infty} > A(t)\right\} \, dt = \int_0^{\|A\|_\infty} \mu\left\{\frac{q}{\|q\|_\infty} > s\right\} \frac{1}{A'\left(A^{-1}(s)\right)} \, ds \, .$$

Now the necessary condition of Item 3. Follows from Item 1. and our assumptions. □

The previous proposition shows that the range of the escort mapping is uniformly dense as it contains all simple densities. Moreover, in the partial order induced by the rearrangement of the normalized density (that is for each q the mapping $t \mapsto \mu\left\{\frac{q}{\|q\|_\infty} > t\right\}$), it contains the full right interval of each element. But the range of the escort mapping is not the full set of bounded positive densities, unless the σ-algebra $\mathscr{X}$ is generated by a finite partition. To provide an example, consider on the Lebesgue unit interval the densities $q_\delta(x) \propto (1 - x^{1/\delta})$, $\delta > 0$, and $A(x) = x/(1 + x)$. The density q_δ turns out to be an escort if, and only if, $\delta \leq 1$.

3.2 *Gradient of the Normalization Operator K_p*

Proposition 2 shows that the functional K_p is a global solution of an equation. We now study its local properties by the implicit function theorem as well as the related subgradients of the convex function K_p. We refer to [7, Part I] for the general theory of convex functions in infinite dimension.

For every $u \in L^1(\mu)$, let us write

$$q(u) = \exp_A(u - K_p(u) + \log_A p) \tag{8}$$

while $\widetilde{q}(u) = \text{escort}\,(q(u))$ denotes its escort density.

Proposition 4 *1. The functional $K_p \colon L^1(\mu) \to \mathbb{R}$ is Gateaux-differentiable with derivative*

$$\frac{d}{dt} K_p(u + tv)\bigg|_{t=0} = \int v\widetilde{q}(u)\, d\mu\ .$$

It follows that $K_p \colon L^1(\mu) \to \mathbb{R}$ is monotone and globally Lipschitz.

2. For every $u, v \in L^1(\mu)$, the inequality

$$K_p(u + v) - K_p(u) \geq \int v\widetilde{q}(u)\, d\mu$$

holds, i.e., the density $\widetilde{q}(u) \in L^\infty(\mu)$ is the unique subgradient of K_p at u.

Proof 1. Consider the equation

$$F(t, \kappa) = \int \exp_A(u + tv - \kappa + \log_A p)\, d\mu - 1 = 0, \quad t, \kappa \in \mathbb{R}\ ,$$

so that $\kappa = K_p(u + tv)$. Derivations under the integral hold by virtue of the bounds

$$\left|\frac{\partial}{\partial t} \exp_A(u + tv - \kappa + \log_A p)\right| = \\ \left|A(\exp_A(u + tv - \kappa + \log_A p))v\right| \leq \|A\|_\infty\, |v|$$

and

$$\left|\frac{\partial}{p}\partial\kappa \exp_A(u + tv - \kappa + \log_A p)\right| = \left|A(\exp_A(u + tv - \kappa + \log_A p))\right| \leq \|A\|_\infty\ .$$

Furthermore, the partial derivative with respect to κ is never zero. Thanks to the implicit function theorem, there exists the derivative $(d\kappa/dt)_{t=0}$ which is the desired Gateaux derivative. Since $\widetilde{q}(u)$ is positive and bounded, K_p is monotone and globally Lipschitz.

2. Thanks to the convexity of $\exp_A$ and the derivation formula, we have

$$\exp_A(u + v - K_p(u + v) + \log_A p) \geq q + A(q)(v - (K_p(u + v) - K_p(v)))\ ,$$

where $q = \exp_A(u - K_p(u) + \log_A p)$. If we take μ-integral of both sides,

$$0 \geq \int vA(q)\, d\mu - (K_p(u + v) - K_p(v)) \int A(q)\, d\mu\ .$$

Isolating the increment $K_p(u+v) - K_p(v)$, the desired inequality obtains. Therefore, $\widetilde{q}(u)$ is a subgradient of K_p at u. From Item 1. we deduce that $\widetilde{q}(u)$ is the unique subgradient and further $\widetilde{q}(u)$ is the Gateaux differential of K_p at u. □

We can also establish Fréchet-differentiability of the functional, under more stringent assumptions.

Proposition 5 *Let $\alpha \geq 2$.*

1. *The superposition operator*

$$L^\alpha(\mu) \ni v \mapsto \exp_A(v + \log_A p) \in L^1(\mu)$$

is continuously Fréchet differentiable with derivative

$$d\exp_A(v) = (h \mapsto A(\exp_A(v + \log_A p))h) \in \mathscr{L}(L^\alpha(\mu), L^1(\mu)) \ .$$

2. *The functional $K_p : L^\alpha(\mu) \to \mathbb{R}$, implicitly defined by the equation*

$$\int \exp_A(v - K_p(v) + \log_A p)\, d\mu = 1, \quad v \in L^\alpha(\mu)$$

is continuously Fréchet differentiable with derivative

$$dK_p(v) = (h \mapsto \int h\widetilde{q}(v)\, d\mu) \ ,$$

where $\widetilde{q}(u) = \text{escort}\,(q(u))$.

Proof 1. Setting $\beta = 1$ in Proposition 1, we get easily the assertion. It remains just to check that the Fréchet derivative is continuous, i.e., that the Fréchet derivative is a continuous map $L^\alpha(\mu) \to \mathscr{L}(L^\alpha(\mu), L^1(\mu))$. If $\|h\|_{L^\alpha(\mu)} \leq 1$ and $v, w \in L^\alpha(\mu)$ we have

$$\int \left|(A[\exp_A(v + \log_A p)] - A[\exp_A(w + \log_A p)])h\right|\, d\mu$$
$$\leq \|A[\exp_A(v + \log_A p) - A[\exp_A(w + \log_A p)]\|_{L^\sigma(\mu)} \ ,$$

where $\sigma = \alpha/\,(\alpha - 1)$ is the conjugate exponent ofα. On the other hand,

$$\|A[\exp_A(v + \log_A p) - A[\exp_A(w + \log_A p)]\|_{L^\sigma(\mu)}$$
$$\leq \left\|A'\right\|_\infty \|A\|_\infty \|v - w\|_{L^\sigma(\mu)}$$

and so the map $L^\alpha(\mu) \to \mathscr{L}(L^\alpha(\mu), L^1(\mu))$ is continuous whenever $\alpha \geq \sigma$, i.e., $\alpha \geq 2$.

2. Fréchet differentiability of K_p is a consequence of the Implicit Function Theorem in Banach spaces, see [6], applied to the C^1-mapping

$$L^{\alpha}(\mu) \times \mathbb{R} \ni (v, \kappa) \mapsto \int \exp_A(v - \kappa + \log_A p)\, d\mu\ .$$

The value of the derivative is given by Proposition 4. □

4 Deformed Divergence

In analogy with the standard exponential case, define the A-divergence between probability densities as

$$D_A(q\|p) = \int \left(\log_A q - \log_A p\right) \operatorname{escort}(q)\ d\mu, \quad \text{for } q, p \in \mathscr{P}\ .$$

Since $\log_A$ is strictly concave with derivative $1/A$, we have

$$\log_A(x) \le \log_A(y) + \frac{1}{A(y)}(x - y)$$

for all $x, y > 0$ and with equality if, and only if, $x = y$. Hence

$$A(y)\left(\log_A(y) - \log_A(x)\right) \ge y - x\ . \tag{9}$$

It follows in particular that $D_A(\cdot\|\cdot)$ is a well defined, possibly extended valued, function.

Observe further that by Proposition 2, $\log_A q - \log_A p \in L^1(\mu)$, and so $D_A(q\|p) < \infty$, whenever $q = q(u)$.

The binary relation D_A is a faithful divergence in that it satisfies the following Gibbs' inequality.

Proposition 6 *It holds $D_A(q\|p) \ge 0$ and $D_A(q\|p) = 0$ if and only if $p = q$.*

Proof From inequality (9) it follows

$$\begin{aligned} D_A(q\|p) &= \frac{1}{\int A(q)\, d\mu} \int \left(\log_A q - \log_A p\right) A(q)\ d\mu \\ &\ge \frac{1}{\int A(q)\, d\mu} \int (q - p)\ \ d\mu = 0. \end{aligned}$$

Moreover, equality holds if and only if $p = q$ μ-a.e. □

There are other alternative definitions that may fully candidate to be a divergence measure. For instance:

$$I_A(q\|p) = -\int \log_A(p/q)q\,d\mu.$$

or also

$$\widetilde{D}_A(q\|p) = \int A(q/p)\log_A(p/q)p\,d\mu.$$

By means of the concavity of $\log_A$, it is not difficult to check that both satisfy Gibbs' condition of Proposition 6, as well as they equal the Kullback–Leibner functional in the non-deformed case. Observe further that the functional $I_A(q\|p)$ is closely related to Tallis' divergence (see [25] and also [14]). In fact, if one replaces $\log_A$ with the q-logarithm, one gets just Tallis' q-divergence.

However our formulation for the divergence is motivated by the structure of the deformed exponential representation. As it will be now seen, our definition of divergence is more adapted to the present setting and it turns out be closely related to the normalizing operator.

In the equation

$$q = \exp_A(u - K_p(u) + \log_A p), \quad u \in L^1(\mu)\ ,\ q \in \mathscr{P}\ , \tag{10}$$

the random variable u is identified up to an additive constant for any fixed density q. There are at least two options for selecting an interesting representative member in the equivalence class.

One option is to impose the further condition $\int u\widetilde{p}\,d\mu = 0$, where $\widetilde{p} = \text{escort}\,(p)$, the integral being well defined, given that the escort density is bounded. This restriction provides a unique element u_q. On the other hand, if we solve Eq. (10) with respect to $u - K(u)$, we get the desired relation:

$$K_p(u_q) = E_{\widetilde{p}}\left[\log_A p - \log_A q\right] = D_A(p\|q), \tag{11}$$

where $u = u_q$ is uniquely characterized by the two equations: $E_{\widetilde{p}}\,[u] = 0$ and $q = \exp_A(u - K_p(u) + \log_A p)$.

Observe further that Eq. (11) entails the relation

$$K_p(u) = D_A(p\|q(u)) \quad \forall u \in L^1\,(\mu)\,.$$

The previous choice is that followed in the construction of the non-parametric exponential manifold, see [22, 23].

With regard to the non-deformed case, Eq. (11) yields the Kulback-Leibler divergence with p and q exchanged, with respect to what is considered more natural in Statistical Physics, see for example the comments [13].

For this purpose, we undertake another choice for the random variable in the equivalence class. More specifically, in Eq. (10) the random variable u will be now centered with respect to $\widetilde{q} = \text{escort}\,(q)$, i.e., $E_{\widetilde{q}}\,[u] = 0$.

To avoid confusion let us rewrite Eq. (10) as follows and where for convenience the function K_p is replaced with $H_p = -K_p$:

$$q = \exp_A(v + H_p(v) + \log_A p), \quad v \in L^1(\mu), \quad E_{\widetilde{q}}[v] = 0, \tag{12}$$

so that

$$H_p(v_q) = E_{\widetilde{q}}\left[\log_A q - \log_A p\right] = D_A(q \| p),$$

where $v = v_q$ is the solution to the two equations $E_{\widetilde{q}}[v] = 0$ and $q = \exp_A(v + H_p(v) + \log_A p)$. There are hence two notable representations of the same probability density q:

$$q = \exp_A(u - K_p(u) + \log_A p) = \exp_A(v + H_p(v) + \log_A p)$$

which implies $u_q - v_q = K_p(u_q) + H_p(v_q)$. This, in turn, leads to

$$-E_{\widetilde{p}}\left[v_q\right] = E_{\widetilde{q}}\left[u_q\right] = K_p(u_q) + H_p(v_q) = K_p(u_q) - K_p(v_q).$$

This provides the following remarkable relation

$$H_p(v_q) = E_{\widetilde{q}}\left[u_q\right] - K_p(u_q). \tag{13}$$

4.1 Variational Formula

We now present a variational formula in the spirit of the classical one by Donsker-Varadhan. Next proposition provides the convex conjugate of K_p, in the duality $L^\infty(\mu) \times L^1(\mu)$.

In what follows, the operator $\eta \mapsto \hat{\eta}$ denotes the inverse of the escort operator, i.e., $\eta = \text{escort}\left(\hat{\eta}\right)$. In the light of the results established in Sect. 3.1, this operator maps a dense subset of $\overline{\mathscr{P}} \cap L^\infty(\mu)$ onto $\overline{\mathscr{P}}$.

Proposition 7 *1. The convex conjugate function of K_p:*

$$K_p^*(w) = \sup_{u \in L^1(\mu)} \left(\int wu \, d\mu - K_p(u) \right), \quad w \in L^\infty(\mu) \tag{14}$$

has domain contained into $\overline{\mathscr{P}} \cap L^\infty(\mu)$. More precisely,

$$\text{escort}(\mathscr{P}) \subseteq dom K_p^* \subseteq \overline{\mathscr{P}} \cap L^\infty(\mu).$$

2. $K_p^*(w) \geq 0$ *for all* $w \in L^\infty(\mu)$. *For any* $\eta \in \text{escort}(\mathscr{P})$, *the conjugate* $K_p^*(\eta)$ *is given by the* Legendre transform*:*

$$K_p^*(\eta) = \int \eta\, u_{\hat{\eta}}\, d\mu - K_p(u_{\hat{\eta}})\ .$$

So that $K_p^*(\eta) = H_p(v_{\hat{\eta}}) = D_A(\hat{\eta}\|p)$*; equivalently:*

$$K_p^*(\text{escort}\,(q)) = D_A(q\|p) \quad \forall p, q \in L^1(\mu).$$

3. *It holds the inversion formula*

$$K_p\,(u) = \max_{\eta\in\text{escort}(\mathscr{P})} \left(\int \eta u\, d\mu - D_A(\hat{\eta}\|p)\right)$$
$$= \max_{q\in\mathscr{P}} \left(\int \text{escort}\,(q)\, u\, d\mu - D_A(q\|p)\right), \quad \forall u \in L^1(\mu).$$

Proof 1. It follows from the fact that K_p is monotone and translation invariant. Let us first suppose $w \notin L_+^\infty(\mu)$. That means that

$$\int w\chi_C\, d\mu < 0$$

is true for some indicator function χ_C. If we consider the cone generated by the function $-\chi_C$, we can write

$$K_p^*\,(w) \geq \sup_{u\in\ cone(-\chi_C)} \left(\int wu\, d\mu - K_p\,(u)\right) \geq \sup_{u\in\ cone(-\chi_C)} \int wu\, d\mu = +\infty,$$

since $K_p\,(u) \leq 0$ when $u \in\ cone(-\chi_C)$. Now consider the case in which $w \geq 0$. If we set $u = \lambda \in \mathbb{R}$, we have $K_p(\lambda) = \lambda$ and consequently

$$K_p^*(w) \geq \sup_{\lambda\in\mathbb{R}} \left(\lambda \int w\, d\mu - \lambda\right)\ . \tag{15}$$

This sup is $+\infty$, unless $\int w\, d\mu = 1$. Hence, $K_p^*\,(w) < \infty$ implies $w \in \overline{\mathscr{P}}$. Summarizing, the domain of K_p^* is contained into $\overline{\mathscr{P}} \cap L^\infty(\mu)$, and this proves one of the two claimed inclusions. The other one will be a direct consequence of the next point.

2. Equation (15) implies $K_p^* \geq 0$. By Proposition 4 the concave and Gateaux differentiable function $u \mapsto \int \eta u\ d\mu - K_p(u)$ has derivative at u given by $\eta - dK_p(u) = \eta - \text{escort}\,(q(u))$, where $q(u) = \exp_A(u - K_p(u) + \log_A p)$. Under our assumptions, the derivative vanishes at $u = u_{\hat{\eta}}$ and the sup in the definition of K_p^* is attained at that point. The maximum value is $K_p^*(\eta) = \int \eta u\, d\mu - K_p(u)$, by setting $u = u_{\hat{\eta}}$.

   The last formula follows straightforward from Eq. (13).

3. For a well-known property of Fenchel–Moreau duality theory, we have:

$$K_p(u) \geq \int wu\, d\mu - K_p^*(w) \quad \forall u \in L^1(\mu), \quad \forall w \in L^\infty(\mu)$$
$$K_p(u) = \int wu\, d\mu - K_p^*(w) \iff w \in \partial K_p(u)\,.$$

Clearly in our case $\partial K_p(u)$ is a singleton and the image of ∂K_p is the set escort ($\mathscr{P}$). Therefore

$$K_p(u) = \max_{w \in \operatorname{escort}(\mathscr{P})} \left(\int wu\, d\mu - K_p^*(w) \right).$$

By Item 2 the desired inversion formula obtains. □

5 Hilbert Bundle Based on $\exp_A$

We shall introduce the Hilbert manifold of probability densities as defined in [18, 19]. A slightly more general set-up than the one used in that reference will be introduced. By means of a general A function, we provide an atlas of charts, and define a linear bundle as an expression of the tangent space.

Let $\mathscr{P}(\mu)$ denote the set of all μ-densities on the probability space $(\mathbb{X}, \mathscr{X}, \mu)$ of the kind

$$q = \exp_A(u - K_1(u)), \quad u \in L^2(\mu), \quad E_\mu[u] = 0\,. \tag{16}$$

Notice that $1 \in \mathscr{P}(\mu)$ because we can take $u = 0$.

Proposition 8 *1. $\mathscr{P}(\mu)$ is the set of all densities q such that $\log_A q \in L^2(\mu)$, in which case $u = \log_A q - E_\mu\left[\log_A q\right]$.*

2. If in addition $A'(0+) > 0$, then $\mathscr{P}(\mu)$ is the set of all densities q such that both q and $\log q$ are in $L^2(\mu)$.

3. Let $A'(0+) > 0$. On a product space with reference probability measures μ_1 and μ_2, and densities respectively q_1 and q_2, we have $q_1 \in \mathscr{P}(\mu_1)$ and $q_2 \in \mathscr{P}(\mu_2)$ if, and only if, $q_1 \otimes q_2 \in \mathscr{P}(\mu_1 \otimes \mu_2)$.

Proof 1. From Eq. (16), it follows $\log_A q = u - K_1(u) \in L^2(\mu)$, provided $u \in L^2(\mu)$. Conversely, let $\log_A q \in L^2(\mu)$. Equation (16) yields

$$u = \log_A q - K_1(u) \quad \textit{and} \quad K_1(u) = -\log_A q.$$

Therefore $u = \log_A q - E_\mu\left[\log_A q\right]$ and $u \in L^2(\mu)$.

2. Write

$$\left|\log_A q\right|^2 = \left|\log_A q\right|^2 (q < 1) + \left|\log_A q\right|^2 (q \geq 1)$$

and use the bounds of Eqs. (3) and (4) to get

$$E_\mu\left[\left|\log_A q\right|^2\right] \leq \frac{1}{\alpha_2^2} E_\mu\left[|\log q|^2\,(q<1)\right] + \frac{1}{A(1)^2} E_\mu\left[|q-1|^2\,(q\geq 1)\right] \leq \\ \frac{1}{\alpha_2^2} E_\mu\left[|\log q|^2\right] + \frac{1}{A(1)^2} E_\mu\left[q^2\right] .$$

We deduce that the two conditions q and $\log q$ in $L^2(\mu)$ imply $\log_A q \in L^2(\mu)$. Conversely, let $\log_A q \in L^2(\mu)$. By means of the other two bounds (recall that $\alpha_1 > 0$) we have too

$$E_\mu\left[\left|\log_A q\right|^2\right] \geq \frac{1}{\alpha_1^2} E_\mu\left[|\log q|^2\,(q<1)\right] + \frac{1}{\|A\|_\infty^2} E_\mu\left[(q-1)^2(q\geq 1)\right] .$$

Consequently, $E_\mu\left[(q-1)^2(q\geq 1)\right] < +\infty$. This in turn gives $E_\mu\left[(q-1)^2\right] < +\infty$, and so $q \in L^2(\mu)$.
Once again, the previous inequality provides the condition $E_\mu\left[|\log q|^2\,(q<1)\right] < +\infty$. On the other hand, $E_\mu\left[|\log q|^2\,(q\geq 1)\right] < +\infty$ since $|\log q|^2\,(q\geq 1) \leq (q-1)^2(q\geq 1)$. Therefore, $\log q \in L^2(\mu)$.

3. We deduce by the previous item that: $q_1 \otimes q_2 \in \mathscr{P}(\mu_1 \otimes \mu_2)$ if and only if both $q_1 \otimes q_2$ and $\log(q_1 \otimes q_2)$ are in $L^2(\mu_1 \otimes \mu_2)$.

   The first condition is equivalent to both $q_1 \in L^2(\mu_1)$ and $q_2 \in L^2(\mu_2)$. The second one is equivalent to $\log q_1 + \log q_2 \in L^2(\mu_1 \otimes \mu_2)$. On the other hand, we have

$$E_{\mu_1\otimes\mu_2}\left[(\log q_1 + \log q_2)^2\right] = \\ E_{\mu_1}\left[\log^2 q_1\right] + E_{\mu_2}\left[\log^2 q_2\right] + 2\, E_{\mu_1}\left[\log q_1\right] E_{\mu_2}\left[\log q_2\right]. \tag{17}$$

   By Eq. (17), $q_1 \in \mathscr{P}(\mu_1)$ and $q_2 \in \mathscr{P}(\mu_2)$ imply $q_1 \otimes q_2 \in \mathscr{P}(\mu_1 \otimes \mu_2)$.
   Conversely, assume $q_1 \otimes q_2 \in \mathscr{P}(\mu_1 \otimes \mu_2)$. This implies that it holds, $E_{\mu_1\otimes\mu_2}\left[(\log q_1 + \log q_2)^2\right] < +\infty$. Since $E_{\mu_i}\left[\log q_i\right] \leq E_{\mu_1}\left[q_i - 1\right] = 0$. We have $E_{\mu_1}\left[\log q_1\right] E_{\mu_2}\left[\log q_2\right] \geq 0$. In view of Eq. (17), we can infer that $q_1 \in \mathscr{P}(\mu_1)$ and $q_2 \in \mathscr{P}(\mu_2)$ □

We proceed now to define an Hilbert bundle with base $\mathscr{P}(\mu)$. The notion of Hilbert bundle has been introduced in Information Geometry by [1]. We are here using an adaptation to the A-exponential of arguments elaborated by [8, 21]. Notice that the construction depends in a essential way on the specific conditions we are assuming for the present class of deformed exponential.

At each $q \in \mathscr{P}(\mu)$ the escort density $\widetilde{q}$ is bounded, so that we can define the fiber given by the Hilbert spaces

$$\mathscr{H}_q = \left\{u \in L^2(\mu) | E_{\widetilde{q}}\left[u\right] = 0\right\}$$

with scalar product $\langle u, v\rangle_q = \int uv\, d\mu$. The Hilbert bundle is

$$H\mathscr{P}(\mu) = \left\{(q, u)|q \in \mathscr{P}(\mu), u \in \mathscr{H}_q\right\} \ .$$

For each $p, q \in \mathscr{P}(\mu)$ the mapping $\mathbb{U}_p^q u = u - E_{\widetilde{q}}[u]$ is a continuous linear mapping from $\mathscr{H}_p$ to $\mathscr{H}_q$. Moreover, $\mathbb{U}_q^r \mathbb{U}_p^q = \mathbb{U}_p^r$. In particular, $\mathbb{U}_q^p \mathbb{U}_p^q$ is the identity on $\mathscr{H}_p$ and so $\mathbb{U}_p^q$ is an isomorphism of $\mathscr{H}_p$ onto $\mathscr{H}_q$.

In the next proposition an affine atlas of charts is constructed in order to define our Hilbert bundle which is an expression of the tangent bundle. The velocity of a curve $t \mapsto p(t) \in \mathscr{P}(\mu)$ is given in the Hilbert bundle by the so called A-score that, in our case, takes the form $A(p(t))^{-1}\dot{p}(t)$, where $\dot{p}(t)$ is computed in $L^1(\mu)$.

The following proposition is taken from [16] where a detailed proof is presented.

Proposition 9 *1. Fix $p \in \mathscr{P}(\mu)$. A positive density $q \in \mathscr{P}(\mu)$ if and only if*

$$q = \exp_A(u - K_p(u) + \log_A p), \ \ \textit{with } u \in L^2(\mu) \textit{ and } E_{\widetilde{p}}[u] = 0.$$

2. *For any fixed $p \in \mathscr{P}(\mu)$ the mapping $s_p \colon \mathscr{P}(\mu) \to \mathscr{H}_p$ defined by*

$$q \mapsto \log_A q - \log_A p + D_A(p\|q)$$

   is injective and surjective, with inverse $e_p(u) = \exp_A(u - K_p(u) + \log_A p)$.
3. *The atlas $\left\{s_p|p \in \mathscr{P}(\mu)\right\}$ is affine with transitions*

$$s_q \circ e_p(u) = \mathbb{U}_p^q u + s_p(q) \ .$$

4. *The velocity of the differentiable curve $t \mapsto p(t) \in \mathscr{P}(\mu)$ in the chart s_p is $ds_p(p(t))/dt \in \mathscr{H}_p$. Conversely, given any $u \in \mathscr{H}_p$, the curve*

$$p \colon t \mapsto \exp_A(tu - K_p(tu) + \log_A p)$$

   satisfies $p(0) = p$ and has velocity u at $t = 0$, expressed in the chart s_p. If the velocity of a curve is $t \mapsto \dot{u}(t)$, in a chart s_p, then $\mathbb{U}_p^q \dot{u}(t)$ is its velocity in the chart s_q.
5. *If $t \mapsto p(t) \in \mathscr{P}(\mu)$ is differentiable with respect to the atlas then it is differentiable as a mapping in $L^1(\mu)$. It follows that the A-score is well-defined and is the expression of the velocity of the curve $t \mapsto p(t)$ in the moving chart $t \mapsto s_{p(t)}$.*

We end here our discussion of the geometry of the Hilbert bundle, because our aim is limited to show the applicability of the analytic results obtained in the previous section. A detailed discussion of the relevant geometric objects e.g., the affine covariant derivative, is not attempted here.

6 Final Remarks

A non-parametric Hilbert manifold based on a deformed exponential representation of positive densities has been firstly introduced by N.J. Newton [18, 19]. We have derived regularity properties of the normalizing functional K_p and discussed the relevant Fenchel conjugation. In particular, we have discussed some properties of the escort mapping and a form of the divergence that appears to be especially adapted to our set-up. We have taken a path different from that of N.J. Newton original presentation. We allow for a manifold defined by an atlas containing charts centered at each density in the model. In conclusion, we have discussed explicitly a version of the Hilbert bundle as a family of codimension 1 sub-vector spaces of the basic Hilbert space.

Acknowledgements The Authors wish to thank the anonymous referees whose comments have led to a considerable improvement of the paper. L. Montrucchio is Honorary Fellow of the Collegio Carlo Alberto Foundation. G. Pistone is a member of GNAMPA-INdAM and acknowledges the support of de Castro Statistics and Collegio Carlo Alberto.

References

1. Amari, S.: Dual connections on the Hilbert bundles of statistical models. In: Geometrization of Statistical Theory (Lancaster, 1987), pp. 123–151. ULDM Publ. (1987)
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence (2000). Translated from the 1993 Japanese original by Daishi Harada
3. Ambrosetti, A., Prodi, G.: A Primer of Nonlinear Analysis. Cambridge Studies in Advanced Mathematics, vol. 34. Cambridge University Press, Cambridge (1993)
4. Appell, J., Zabrejko, P.P.: Nonlinear Superposition Operators. Cambridge Tracts in Mathematics, vol. 95. Cambridge University Press, Cambridge (1990). https://doi.org/10.1017/CBO9780511897450
5. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry. Springer, Berlin (2017)
6. Dieudonné, J.: Foundations of Modern Analysis. Academic Press, New York (1960)
7. Ekeland, I., Témam, R.: Convex Analysis and Variational Problems. Classics in Applied Mathematics, vol. 28, English edn. Society for Industrial and Applied Mathematics (SIAM) (1999). https://doi.org/10.1137/1.9781611971088. Translated from the French
8. Gibilisco, P., Pistone, G.: Connections on non-parametric statistical manifolds by Orlicz space geometry. IDAQP **1**(2), 325–347 (1998)
9. Kaniadakis, G.: Non-linear kinetics underlying generalized statistics. Phys. A **296**(3–4), 405–425 (2001)
10. Kaniadakis, G.: Statistical mechanics in the context of special relativity. Phys. Rev. E **66**, 056, 125 1–17 (2002)
11. Kaniadakis, G.: Statistical mechanics in the context of special relativity. ii. Phys. Rev. E **72**(3), 036,108 (2005). https://doi.org/10.1103/PhysRevE.72.036108
12. Kass, R.E., Vos, P.W.: Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley, New York (1997). https://doi.org/10.1002/9781118165980. A Wiley-Interscience Publication
13. Landau, L.D., Lifshits, E.M.: Course of Theoretical Physics. Statistical Physics, vol. V, 3rd edn. Butterworth-Heinemann, Oxford (1980)

14. Loaiza, G., Quiceno, H.R.: A q-exponential statistical Banach manifold. J. Math. Anal. Appl. **398**(2), 466–476 (2013). https://doi.org/10.1016/j.jmaa.2012.08.046
15. Malliavin, P.: Integration and Probability. Graduate Texts in Mathematics, vol. 157. Springer-Verlag (1995). With the collaboration of Hlne Airault, Leslie Kay and Grard Letac. Edited and translated from the French by Kay, With a foreword by Mark Pinsky
16. Montrucchio, L., Pistone, G.: Deformed exponential bundle: the linear growth case. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information. LNCS, vol. 10589, pp. 239–246. Springer (2017). Proceedings of the Third International Conference, GSI 2017, Paris, France, November 7–9
17. Naudts, J.: Generalised Thermostatistics. Springer-Verlag London Ltd. (2011). https://doi.org/10.1007/978-0-85729-355-8
18. Newton, N.J.: An infinite-dimensional statistical manifold modelled on Hilbert space. J. Funct. Anal. **263**(6), 1661–1681 (2012). https://doi.org/10.1016/j.jfa.2012.06.007
19. Newton, N.J.: Infinite-dimensional statistical manifolds based on a balanced chart. Bernoulli **22**(2), 711–731 (2016). https://doi.org/10.3150/14-BEJ673
20. Pistone, G.: κ-exponential models from the geometrical viewpoint. Eur. Phys. J. B Condens. Matter Phys. **71**(1), 29–37 (2009). https://doi.org/10.1140/epjb/e2009-00154-y
21. Pistone, G.: Nonparametric information geometry. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information. Lecture Notes in Computer Science, vol. 8085, pp. 5–36. Springer, Heidelberg (2013). Proceedings First International Conference, GSI 2013 Paris, France, August 28–30
22. Pistone, G., Rogantin, M.: The exponential statistical manifold: mean parameters, orthogonality and space transformations. Bernoulli **5**(4), 721–760 (1999)
23. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **23**(5), 1543–1561 (1995)
24. Shima, H.: The Geometry of Hessian Structures. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2007). https://doi.org/10.1142/9789812707536
25. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. **52**(1–2), 479–487 (1988)
26. Vigelis, R.F., Cavalcante, C.C.: On ϕ-families of probability distributions. J. Theor. Probab. **26**, 870–884 (2013)

Statistical Manifolds Admitting Torsion and Partially Flat Spaces

Masayuki Henmi and Hiroshi Matsuzoe

Abstract It is well-known that a contrast function defined on a product manifold $M \times M$ induces a Riemannian metric and a pair of dual torsion-free affine connections on the manifold M. This geometrical structure is called a statistical manifold and plays a central role in information geometry. Recently, the notion of pre-contrast function has been introduced and shown to induce a similar differential geometrical structure on M, but one of the two dual affine connections is not necessarily torsion-free. This structure is called a statistical manifold admitting torsion. The notion of statistical manifolds admitting torsion has been originally introduced to study a geometrical structure which appears in a quantum statistical model. However, it has been shown that an estimating function which is used in "classical" statistics also induces a statistical manifold admitting torsion through its associated pre-contrast function. The aim of this paper is to summarize such previous results. In particular, we focus on a partially flat space, which is a statistical manifold admitting torsion where one of its dual connections is flat. In this space, it is possible to discuss some properties similar to those in a dually flat space, such as a canonical pre-contrast function and a generalized projection theorem.

1 Introduction

A statistical manifold is a Riemannian manifold with a pair of dual torsion-free affine connections and it plays a central role in information geometry. This geometrical structure is induced from an asymmetric (squared) distance-like smooth function called a contrast function by taking its second and third derivatives [1, 2]. The Kullback–Leibler divergence on a regular parametric statistical model is a typical example of contrast functions and its induced geometrical objects are the Fisher met-

M. Henmi (✉)
The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
e-mail: henmi@ism.ac.jp

H. Matsuzoe
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555, Japan
e-mail: matsuzoe@nitech.ac.jp

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_3

ric, the exponential and mixture connections. The geometrical structure determined by these objects plays an important role in the geometry of statistical inference, as is widely known [3, 4].

A statistical manifold admitting torsion (SMAT) is a Riemannian manifold with a pair of dual affine connections, where only one of them must be torsion-free but the other is *not* necessarily so. This geometrical structure naturally appears in a quantum statistical model (i.e. a set of density matrices representing quantum states) [3] and the notion of SMAT was originally introduced to study such a geometrical structure from a mathematical point of view [5]. A pre-contrast function was subsequently introduced as a generalization for the first derivative of a contrast function and it was shown that an pre-contrast function induces a SMAT by taking its first and second derivatives [6].

In statistics, an estimating function is a function defined on a direct product of parameter and sample spaces, and it is used to obtain an estimator by solving its corresponding estimating equation. Henmi and Matsuzoe [7] showed that a SMAT also appears in “classical" statistics through an estimating function. More precisely, an estimating function naturally defines a pre-contrast function on a parametric statistical model and a SMAT is induced from it.

This paper summarizes such previous results, focusing on a SMAT where one of its dual connections is flat. We call this geometrical structure a partially flat space. Although this space is different from a dually flat space in general since one of the dual connections in a SMAT possibly has torsion, some similar properties hold. For example, the canonical pre-contrast function can be naturally defined on a partially flat space, which is an analog of the canonical contrast function (or canonical divergence) in a dually flat space. In addition, a generalized projection theorem holds with respect to the canonical pre-contrast function. This theorem can be seen as a generalization of the projection theorem in a dually flat space. This paper is an extended version of the conference proceedings [8]. We consider a statistical problem to see an example of statistical manifolds admitting torsion induced from estimating functions and discuss some future problems, neither of which were included in [8].

2 Statistical Manifolds and Contrast Functions

Through this paper, we assume that all geometrical objects on differentiable manifolds are smooth and restrict our attention to Riemannian manifolds, although the most of the concepts can be defined for semi-Riemannian manifolds.

Let (M, g) be a Riemannian manifold and ∇ be an affine connection on M. The *dual connection* ∇^* of ∇ with respect to g is defined by

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z) \quad (\forall X, \forall Y, \forall Z \in \mathscr{X}(M)),$$

where $\mathscr{X}(M)$ is the set of all vector fields on M.

For an affine connection ∇ on M, its curvature tensor field R and torsion tensor field T are defined by the following equations as usual:

$$R(X, Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,$$
$$T(X, Y) := \nabla_X Y - \nabla_Y X - [X, Y]$$

$(\forall X, \forall Y, \forall Z \in \mathscr{X}(M))$. It is said that an affine connection ∇ is *torsion-free* if $T = 0$. Note that for a torsion-free affine connection ∇, $\nabla^* = \nabla$ implies that ∇ is the Levi-Civita connection with respect to g. Let R^* and T^* be the curvature and torsion tensor fields of ∇^*, respectively. It is easy to see that $R = 0$ always implies $R^* = 0$, but $T = 0$ does not necessarily imply $T^* = 0$.

Let ∇ be a torsion-free affine connection on a Riemannian manifold (M, g). Following [9], we say that (M, g, ∇) is a *statistical manifold* if and only if ∇g is a symmetric (0, 3)-tensor field, that is

$$(\nabla_X g)(Y, Z) = (\nabla_Y g)(X, Z) \quad (\forall X, \forall Y, \forall Z \in \mathscr{X}(M)). \tag{1}$$

This condition is equivalent to $T^* = 0$ under the condition that ∇ is a torsion-free. If (M, g, ∇) is a statistical manifold, so is (M, g, ∇^*) and it is called the *dual statistical manifold* of (M, g, ∇). Since ∇ and ∇^* are both torsion-free for a statistical manifold (M, g, ∇), $R = 0$ implies that ∇ and ∇^* are both flat. In this case, (M, g, ∇, ∇^*) is called a *dually flat space* [3].

Let ϕ be a real-valued function on the direct product $M \times M$ of a manifold M and $X_1, \ldots, X_i, Y_1, \ldots, Y_j$ be vector fields on M. The functions $\phi[X_1, \ldots, X_i | Y_1, \ldots, Y_j]$, $\phi[X_1, \ldots, X_i | \,]$ and $\phi[\, | Y_1, \ldots, Y_j]$ on M are defined by the equations

$$\phi[X_1, \ldots, X_i | Y_1, \ldots, Y_j](r) := (X_1)_p \cdots (X_i)_p (Y_1)_q \cdots (Y_j)_q \phi(p, q)|_{p=r, q=r}, \tag{2}$$

$$\phi[X_1, \ldots, X_i | \,](r) := (X_1)_p \cdots (X_i)_p \phi(p, r)|_{p=r}, \tag{3}$$

$$\phi[\, | Y_1, \ldots, Y_j](r) := (Y_1)_q \cdots (Y_j)_q \phi(r, q)|_{q=r} \tag{4}$$

for any $r \in M$, respectively [1]. Using these notations, a *contrast function* ϕ on M is defined to be a real-valued function on $M \times M$ which satisfies the following conditions [1, 2]:

(a) $\phi(p, p) = 0 \quad (\forall p \in M)$,

(b) $\phi[X | \,] = \phi[\, | X] = 0 \quad (\forall X \in \mathscr{X}(M))$,

(c) $g(X, Y) := -\phi[X | Y] \quad (\forall X, \forall Y \in \mathscr{X}(M))$ is a Riemannian metric on M.

Note that these conditions imply that

$$\phi(p, q) \geq 0, \quad \phi(p, q) = 0 \Longleftrightarrow p = q$$

in some neighborhood of the diagonal set $\{(r, r) | r \in M\}$ in $M \times M$. Although a contrast function is not necessarily symmetric, this property means that a contrast function measures some discrepancy between two points on M (at least locally). For a given contrast function ϕ, the two affine connections ∇ and ∇^* are defined by

$$g(\nabla_X Y, Z) = -\phi[XY|Z], \quad g(Y, \nabla^*_X Z) = -\phi[Y|XZ]$$

$(\forall X, \forall Y, \forall Z \in \mathscr{X}(M))$. In this case, ∇ and ∇^* are both torsion-free and dual to each other with respect to g. This means that both of (M, g, ∇) and (M, g, ∇^*) are statistical manifolds. In particular, (M, g, ∇) is called the statistical manifold induced from the contrast function ϕ.

A typical example of contrast functions is the Kullback–Leibler divergence on a statistical model. Let $S = \{p(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta^1, \ldots, \theta^d) \in \Theta \subset \boldsymbol{R}^d\}$ be a regular parametric statistical model, which is a set of probability density functions with respect to a dominating measure ν on a sample space Ω. Each element is indexed by a parameter (vector) $\boldsymbol{\theta}$ in an open subset Θ of $\boldsymbol{R}^d$ and the set S satisfies some regularity conditions, under which S can be seen as a differentiable manifold. The Kullback–Leibler divergence of the two density functions $p_1(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_2)$ in S is defined to be

$$\phi_{KL}(p_1, p_2) := \int_\Omega p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})} \nu(d\boldsymbol{x}).$$

It is easy to see that the Kullback–Leibler divergence satisfies the conditions (a), (b) and (c), and so it is a contrast function on S. Its induced Riemannian metric and dual connections are Fisher metric g^F, the exponential connection $\nabla^{(e)}$ and mixture connection $\nabla^{(m)}$, respectively. They are given as follows:

$$g^F_{jk}(\boldsymbol{\theta}) := g^F(\partial_j, \partial_k) = E_{\boldsymbol{\theta}}\{s^j(\boldsymbol{x}, \boldsymbol{\theta}) s^k(\boldsymbol{x}, \boldsymbol{\theta})\},$$
$$\begin{cases} \Gamma^{(e)}_{ij,k}(\boldsymbol{\theta}) := g^F(\nabla^{(e)}_{\partial_i} \partial_j, \partial_k) = E_{\boldsymbol{\theta}}[\{\partial_i s^j(\boldsymbol{x}, \boldsymbol{\theta})\} s^k(\boldsymbol{x}, \boldsymbol{\theta})] \\ \Gamma^{(m)}_{ik,j}(\boldsymbol{\theta}) := g^F(\partial_j, \nabla^{(m)}_{\partial_i} \partial_k) = \int_\Omega s^j(\boldsymbol{x}, \boldsymbol{\theta}) \partial_i \partial_k p(\boldsymbol{x}; \boldsymbol{\theta}) \nu(d\boldsymbol{x}) \end{cases},$$

where $E_{\boldsymbol{\theta}}$ indicates that the expectation is taken with respect to $p(\boldsymbol{x}; \boldsymbol{\theta})$, $\partial_i = \frac{\partial}{\partial \theta^i}$ and $s^i(\boldsymbol{x}; \boldsymbol{\theta}) = \partial_i \log p(\boldsymbol{x}; \boldsymbol{\theta})$ $(i = 1, \ldots, d)$. As is widely known, this geometrical structure plays the most fundamental and important role in the differential geometry of statistical inference [3, 4].

3 Statistical Manifolds Admitting Torsion and Pre-contrast Functions

A statistical manifold admitting torsion is an abstract notion for the geometrical structure where only one of the dual connections is allow to have torsion, which naturally appears in a quantum statistical model [3]. The definition is obtained by generalizing (1) in the definition of statistical manifold as follows [5].

Let (M, g) be a Riemannian manifold and ∇ be an affine connection on M. We say that (M, g, ∇) is a *statistical manifold admitting torsion* (SMAT for short) if and only if

$$(\nabla_X g)(Y, Z) - (\nabla_Y g)(X, Z) = -g(T(X, Y), Z) \quad (\forall X, \forall Y, \forall Z \in \mathscr{X}(M)). \quad (5)$$

This condition is equivalent to $T^* = 0$ in the case where ∇ possibly has torsion, and it reduces to (1) if ∇ is torsion-free. Note that (M, g, ∇^*) is not necessarily a statistical manifold although ∇^* is torsion-free. It should be also noted that (M, g, ∇^*) is a SMAT whenever a torsion-free affine connection ∇ is given on a Riemannian manifold (M, g).

For a SMAT (M, g, ∇), $R = 0$ does not necessarily imply that ∇ is flat, but it implies that ∇^* is flat since $R^* = 0$ and $T^* = 0$. In this case, we call (M, g, ∇, ∇^*) a *partially flat space*.

Let ρ be a real-valued function on the direct product $TM \times M$ of a manifold M and its tangent bundle TM, and $X_1, \ldots, X_i, Y_1, \ldots, Y_j, Z$ be vector fields on M. The function $\rho[X_1, \ldots, X_i Z | Y_1, \ldots, Y_j]$ on M is defined by

$$\rho[X_1, \ldots, X_i Z | Y_1, \ldots, Y_j](r) := (X_1)_p \cdots (X_i)_p (Y_1)_q \cdots (Y_j)_q \rho(Z_p, q)|_{p=r, q=r}$$

for any $r \in M$. Note that the role of Z is different from those of the vector fields in the notation of (2). The functions $\rho[X_1, \ldots, X_i Z | \,]$ and $\rho[\, | Y_1, \ldots, Y_j]$ are also defined in the similar way to (3) and (4).

We say that ρ is a *pre-contrast function* on M if and only if the following conditions are satisfied [6, 7]:

(a) $\rho(f_1 X_1 + f_2 X_2, q) = f_1 \rho(X_1, q) + f_2 \rho(X_2, q)$
$(\forall f_1, \forall f_2 \in C^\infty(M),\ \forall X_1, \forall X_2 \in \mathscr{X}(M),\ \forall q \in M)$.

(b) $\rho[X | \,] = 0 \ \ (\forall X \in \mathscr{X}(M)) \ \ \big(i.e.\ \rho(X_p, p) = 0 \ \ (\forall p \in M)\big)$.

(c) $g(X, Y) := -\rho[X | Y] \ \ (\forall X, \forall Y \in \mathscr{X}(M))$ is a Riemannian metric on M.

Note that for any contrast function ϕ on M, the function ρ_ϕ which is defined by

$$\rho_\phi(X_p, q) := X_p \phi(p, q) \quad (\forall p, \forall q \in M, \ \forall X_p \in T_p(M))$$

is a pre-contrast function on M. The notion of pre-contrast function is obtained by taking the fundamental properties of the first derivative of a contrast function as axioms. For a given pre-contrast function ρ, two affine connections ∇ and ∇^* are defined by

$$g(\nabla_X Y, Z) = -\rho[XY | Z], \quad g(Y, \nabla^*_X Z) = -\rho[Y | XZ]$$

$(\forall X, \forall Y, \forall Z \in \mathscr{X}(M))$ in the same way as for a contrast function. In this case, ∇ and ∇^* are dual to each other with respect to g and ∇^* is torsion-free. However, the affine connection ∇ possibly has torsion. This means that (M, g, ∇) is a SMAT and it is called the SMAT induced from the pre-contrast function ρ.

4 Canonical Pre-contrast Functions in Partially Flat Spaces

In a dually flat space (M, g, ∇, ∇^*), it is well-known that the canonical contrast functions (called ∇ and ∇^*- divergences) are naturally defined, and the Pythagorean theorem and the projection theorem are stated in terms of the ∇ and ∇^*- geodesics and the canonical contrast functions [3, 4]. In a partially flat space (M, g, ∇, ∇^*), where $R = R^* = 0$ and $T^* = 0$, it is possible to define a pre-contrast function which can be seen as canonical, and a projection theorem holds with respect to the “canonical" pre-contrast function and the ∇^*-geodesic.

Proposition 1 (Canonical Pre-contrast Functions) *Let (M, g, ∇, ∇^*) be a partially flat space (i.e. (M, g, ∇) is a SMAT with $R = R^* = 0$ and $T^* = 0$) and (U, η_i) be an affine coordinate neighborhood with respect to ∇^* in M. The function ρ on $TU \times U$ defined by*

$$\rho(Z_p, q) := -g_p(Z_p, \dot{\gamma}^*(0)) \;\; (\forall p, \forall q \in U, \forall Z_p \in T_p(U)), \tag{6}$$

is a pre-contrast function on U, where $\gamma^* : [0, 1] \to U$ is the ∇^*-geodesic such that $\gamma^*(0) = p, \gamma^*(1) = q$ and $\dot{\gamma}^*(0)$ is the tangent vector of γ^* at p. Furthermore, the pre-contrast function ρ induces the original Riemannian metric g and the dual connections ∇ and ∇^* on U.

Proof For the function ρ defined as (6), the condition (a) in the definition of pre-contrast functions follows from the bilinearity of the inner product g_p. The condition (b) immediately follows from $\dot{\gamma}^*(0) = 0$ when $p = q$. By calculating the derivatives of ρ with the affine coordinate system (η_i), it can be shown that the condition (c) holds and that the induced Riemannian metric and dual affine connections coincide with the original g, ∇ and ∇^*. □

In particular, if (U, g, ∇, ∇^*) is a dually flat space, the pre-contrast function ρ defined in (6) coincides with the directional derivative $Z_p\phi^*(\cdot, q)$ of ∇^*-divergence $\phi^*(\cdot, q)$ with respect to Z_p (cf. [10, 11]). Hence, the definition of (6) seems to be natural one and we call the function ρ in (6) the *canonical pre-contrast function* in a partially flat space (U, g, ∇, ∇^*).

From the definition of the canonical pre-contrast function, we can immediately obtain the following theorem.

Corollary 1 (Generalized Projection Theorem) *Let (U, η_i) be an affine coordinate neighborhood in a partially flat space (M, g, ∇, ∇^*) and ρ be the canonical pre-contrast function on U. For any submanifold N in U, the following conditions are* *equivalent:*

(i) *The* $\nabla^* -$ *geodesic starting at* $q \in U$ *is perpendicular to* N *at* $p \in N$.

(ii) $\rho(Z_p, q) = 0$ *for any* Z_p *in* $T_p(N)$.

If (U, g, ∇, ∇^*) is a dually flat space, this theorem reduces to the projection theorem with respect to the ∇^*-divergence ϕ^*, since $\rho(Z_p, q) = Z_p\phi^*(p, q)$. In this sense, it can be seen as a generalized version of the projection theorem in dually flat spaces, and this is also one of the reasons why we consider the pre-contrast function ρ defined in (6) as canonical.

5 Statistical Manifolds Admitting Torsion Induced from Estimating Functions

As we mentioned in Introduction, a SMAT naturally appears through an estimating function in a “classical" statistical model as well as in a quantum statistical model. In this section, we briefly explain how a SMAT is induced on a parametric statistical model from an estimating function. See [7] for more details.

Let $S = \{p(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta^1, \ldots, \theta^d) \in \Theta \subset \boldsymbol{R}^d\}$ be a regular parametric statistical model. An estimating function on S, which we consider here, is a $\boldsymbol{R}^d$-valued function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ satisfying the following conditions:

$$E_{\boldsymbol{\theta}}\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\} = \boldsymbol{0}, \quad E_{\boldsymbol{\theta}}\{\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\|^2\} < \infty, \quad \det\left[E_{\boldsymbol{\theta}}\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta})\right\}\right] \neq 0 \quad (\forall \boldsymbol{\theta} \in \Theta).$$

The first condition is called the unbiasedness of estimating functions, which is important to ensure the consistency of the estimator obtained from an estimating function. Let $X_1, \ldots, X_n$ be a random sample from an unknown probability distribution $p(\boldsymbol{x}; \boldsymbol{\theta}_0)$ in S. The estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ is called an M-estimator if it is obtained as a solution to the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \boldsymbol{0}. \tag{7}$$

The M-estimator $\hat{\boldsymbol{\theta}}$ has the consistency

$$\hat{\boldsymbol{\theta}} \longrightarrow \boldsymbol{\theta}_0 \quad \text{(in probability)}$$

as $n \to \infty$ and the asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N\left(\boldsymbol{0}, \mathrm{Avar}\left(\hat{\boldsymbol{\theta}}\right)\right) \quad \text{(in distribution)}$$

as $n \to \infty$ under some additional regularity conditions [12], which are also assumed in the following discussion. The matrix Avar$(\hat{\boldsymbol{\theta}})$ is the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ and is given by

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \{A(\boldsymbol{\theta}_0)\}^{-1} B(\boldsymbol{\theta}_0) \{A(\boldsymbol{\theta}_0)\}^{-T}, \tag{8}$$

where $A(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left\{(\partial \boldsymbol{u}/\partial \boldsymbol{\theta})(\boldsymbol{x}, \boldsymbol{\theta})\right\}$, $B(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}$ and $-T$ means transposing an inverse matrix (or inverting a transposed matrix).

In order to induce the structure of SMAT on S from an estimating function, we consider the notion of *standardization* of estimating functions. For an estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$, its standardization (or *standardized estimating function*) is defined by

$$\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left\{\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}\left[E_{\boldsymbol{\theta}}\left\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}\right]^{-1}\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}),$$

where $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta}) = (\partial/\partial \boldsymbol{\theta}) \log p(\boldsymbol{x}; \boldsymbol{\theta})$ is the score function for $\boldsymbol{\theta}$ [13]. Geometrically, the i-th component of the standardized estimating function $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ is the orthogonal projection of the i-th component of the score function $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})$ onto the linear space spanned by all components of the estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ in the Hilbert space

$$\mathscr{H}_{\boldsymbol{\theta}} := \{a(\boldsymbol{x}) \mid E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})\} = 0,\ E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})^2\} < \infty\}$$

with the inner product $< a(\boldsymbol{x}), b(\boldsymbol{x}) >_{\boldsymbol{\theta}} := E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})b(\boldsymbol{x})\}$ $(\forall a(\boldsymbol{x}), \forall b(\boldsymbol{x}) \in \mathscr{H}_{\boldsymbol{\theta}})$. The standardization $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ of $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ does not change the estimator since the estimating equation obtained from $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ is equivalent to the original estimating equation (7). In terms of the standardization, the asymptotic variance-covariance matrix (8) can be rewritten as

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \{G(\boldsymbol{\theta}_0)\}^{-1},$$

where $G(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left\{\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}$. The matrix $G(\boldsymbol{\theta})$ is called a Godambe information matrix [14], which can be seen as a generalization of the Fisher information matrix.

As we have seen in Sect. 2, the Kullback–Leibler divergence ϕ_{KL} is a contrast function on S. Hence, the first derivative of ϕ_{KL} is a pre-contrast function on S and given by

$$\rho_{KL}((\partial_j)_{p_1}, p_2) := (\partial_j)_{p_1}\phi_{KL}(p_1, p_2) = -\int_{\Omega} s^j(\boldsymbol{x}, \boldsymbol{\theta}_1)p(\boldsymbol{x}; \boldsymbol{\theta}_2)\nu(d\boldsymbol{x})$$

for any two probability distributions $p_1(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_2)$ in S and $j = 1, \ldots, d$. This observation leads to the following proposition [7].

Proposition 2 (Pre-contrast Functions from Estimating Functions) *For an estimating function* $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ *on the parametric model* S*, a pre-contrast function* $\rho_{\boldsymbol{u}}$:

$TS \times S \to \boldsymbol{R}$ is defined by

$$\rho_{\boldsymbol{u}}((\partial_j)_{p_1}, p_2) := -\int_{\Omega} u_*^j(\boldsymbol{x}, \boldsymbol{\theta}_1) p(\boldsymbol{x}; \boldsymbol{\theta}_2) \nu(d\boldsymbol{x}) \tag{9}$$

for any two probability distributions $p_1(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_2)$ in S and $j = 1, \ldots, d$, where $u_*^j(\boldsymbol{x}, \boldsymbol{\theta})$ is the j-th component of the standardization $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ of $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$.

The use of the standardization $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ instead of $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ ensures that the definition of the function $\rho_{\boldsymbol{u}}$ does not depend on the choice of coordinate system (parameter) of S. In fact, for a coordinate transformation (parameter transformation) $\boldsymbol{\eta} = \Phi(\boldsymbol{\theta})$, the estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ is changed into $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\eta}) = \boldsymbol{u}(\boldsymbol{x}, \Phi^{-1}(\boldsymbol{\eta}))$ and we have

$$\boldsymbol{v}_*(\boldsymbol{x}, \boldsymbol{\eta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}\right)^T \boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta}).$$

This is the same as the transformation rule of coordinate bases on a tangent space of a manifold. The set of all components of the standardized estimating function $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ can be seen as a representation of the coordinate basis $\{(\partial_1)_p, \ldots, (\partial_d)_p\}$ on the tangent space $T_p(S)$ of S, where $p(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta})$.

The proof of Proposition 2 is straightforward. In particular, the condition (b) in the definition of pre-contrast function follows from the unbiasedness of the (standardized) estimating function. The Riemannian metric g, dual connections ∇ and ∇^* induced from the pre-contrast function $\rho_{\boldsymbol{u}}$ are given as follows:

$$g_{jk}(\boldsymbol{\theta}) := g(\partial_j, \partial_k) = E_{\boldsymbol{\theta}}\{u_*^j(\boldsymbol{x}, \boldsymbol{\theta}) u_*^k(\boldsymbol{x}, \boldsymbol{\theta})\} = G(\boldsymbol{\theta})_{jk},$$
$$\begin{cases} \Gamma_{ij,k}(\boldsymbol{\theta}) := g(\nabla_{\partial_i}\partial_j, \partial_k) = E_{\boldsymbol{\theta}}[\{\partial_i u_*^j(\boldsymbol{x}, \boldsymbol{\theta})\} s^k(\boldsymbol{x}, \boldsymbol{\theta})] \\ \Gamma^*_{ik,j}(\boldsymbol{\theta}) := g(\partial_j, \nabla^*_{\partial_i}\partial_k) = \int_{\Omega} u_*^j(\boldsymbol{x}, \boldsymbol{\theta}) \partial_i \partial_k p(\boldsymbol{x}; \boldsymbol{\theta}) \nu(d\boldsymbol{x}) \end{cases},$$

where $G(\boldsymbol{\theta})_{jk}$ is the (j, k) component of the Godambe information matrix $G(\boldsymbol{\theta})$. Note that ∇^* is always torsion-free since $\Gamma^*_{ik,j} = \Gamma^*_{ki,j}$, whereas ∇ is not necessarily torsion-free unless $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ is integrable with respect to $\boldsymbol{\theta}$ (*i.e.* there exists a function $\psi(\boldsymbol{x}, \boldsymbol{\theta})$ satisfying $\partial_j \psi(\boldsymbol{x}, \boldsymbol{\theta}) = u_*^j(\boldsymbol{x}, \boldsymbol{\theta})$ $(j = 1, \ldots, d)$).

If it is integrable and ∇ is torsion-free, it is possible to construct a contrast function on S, from which the pre-contrast function $\rho_{\boldsymbol{u}}$ in (9) is obtained by taking its first derivative, as follows:

$$\phi_{\boldsymbol{u}}(p_1, p_2) = \int_{\Omega} \{\psi(\boldsymbol{x}, \boldsymbol{\theta}_1) - \psi(\boldsymbol{x}, \boldsymbol{\theta}_2)\}\, p(\boldsymbol{x}; \boldsymbol{\theta}_2) \nu(d\boldsymbol{x}),$$

where $\partial_j \psi(\boldsymbol{x}, \boldsymbol{\theta}) = u_*^j(\boldsymbol{x}, \boldsymbol{\theta})$ $(j = 1, \ldots, d)$ and $p_l(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_l)$ $(l = 1, 2)$.

Table 1 Votes cast in the n-th constituency ($n = 1, \ldots, N$)

Party	C	L	Total
C	X_{1n}	$m_{1n} - X_{1n}$	m_{1n}
L	X_{2n}	$m_{2n} - X_{2n}$	m_{2n}
Total	X_n	$m_n - X_n$	m_n

6 Example

In this section, we consider the estimation problem of voter transition probabilities described in [15] to see an example of statistical manifolds admitting torsion (SMAT) induced from estimation functions.

Suppose that we had two successive elections which were carried out in N constituencies, and that the two political parties C and L contended in each election. The table below summarizes the numbers of voters in the n-th constituency for the respective elections. It is assumed that we can observe only the marginal totals m_{1n}, m_{2n}, X_n and $m_n - X_n$, where X_n is a random variable and the others are treated as fixed constants. Let θ^1 and θ^2 be the probabilities that a voter who votes for the parties C and L in Election 1 votes for C in Election 2, respectively. They are the parameters of interest here. Then, the random variables X_{1n} and X_{2n} in Table 1 are assumed to independently follow the binomial distributions $B(m_{1n}, \theta^1)$ and $B(m_{2n}, \theta^2)$, respectively.

In the n-th constituency, the probability function of the observation $X_n = X_{1n} + X_{2n}$ is given by

$$p_n(x_n; \boldsymbol{\theta}) = \sum_{x_{1n}=0}^{m_{1n}} \binom{m_{1n}}{x_{1n}} \binom{m_{2n}}{x_n - x_{1n}} \left(\theta^1\right)^{x_{1n}} \left(1-\theta^1\right)^{m_{1n}-x_{1n}} \left(\theta^2\right)^{x_n - x_{1n}} \left(1-\theta^2\right)^{m_{2n}-x_n+x_{1n}},$$

where $\boldsymbol{\theta} = \left(\theta^1, \theta^2\right)$. The statistical model S in this problem consists of all possible probability functions of the observed data $\boldsymbol{X} = (X_1, \ldots, X_N)$ as follows:

$$S = \left\{ p(\boldsymbol{x}; \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta} = \left(\theta^1, \theta^2\right) \in (0, 1) \times (0, 1) \right\},$$

where $p(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{n=1}^{N} p_n(x_n; \boldsymbol{\theta})$ $(\boldsymbol{x} = (x_1, \ldots, x_N))$ since $X_1, \ldots, X_N$ are independent.

Although the maximum likelihood estimation for $\boldsymbol{\theta}$ is possible based on the likelihood function $L(\boldsymbol{\theta}) = p(\boldsymbol{X}; \boldsymbol{\theta})$, it is a little complicated since X_{1n} and X_{2n} are not observed in each n-th constituency. An alternative approach for estimating $\boldsymbol{\theta}$ is to use the quasi-score function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta}) = (q^1(\boldsymbol{x}, \boldsymbol{\theta}), q^2(\boldsymbol{x}, \boldsymbol{\theta}))^T$ [15] as an estimating function, where

$$q^1(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{m_{1n}\{x_n - \mu_n(\boldsymbol{\theta})\}}{V_n(\boldsymbol{\theta})}, \quad q^2(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{m_{2n}\{x_n - \mu_n(\boldsymbol{\theta})\}}{V_n(\boldsymbol{\theta})}.$$

Here, $\mu_n(\boldsymbol{\theta})$ and $V_n(\boldsymbol{\theta})$ are the mean and variance of X_n, respectively, i.e.

$$\begin{aligned}\mu_n(\boldsymbol{\theta}) &= E(X_n) = m_{1n}\theta^1 + m_{2n}\theta^2 \\ V_n(\boldsymbol{\theta}) &= V(X_n) = m_{1n}\theta^1\left(1-\theta^1\right) + m_{2n}\theta^2\left(1-\theta^2\right).\end{aligned} \tag{10}$$

In this example, the random variables $X_1, \dots, X_N$ in the observed data are independent, but not identically distributed. However, it is possible to apply the results in Sect. 5 by considering the whole of the left hand side of (7) as an estimating function and modifying the results in this case. Note that the estimating function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ is already standardized since the i-th component $q^i(\boldsymbol{x}, \boldsymbol{\theta})$ of $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ is obtained by the orthogonal projection of the i-th component of the score function $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})$ for $\boldsymbol{\theta}$ onto the linear space spanned by $\{x_1 - \mu_1(\boldsymbol{\theta}), \dots, x_N - \mu_N(\boldsymbol{\theta})\}$. In fact, the orthogonal projection is calculated as follows:

$$\begin{aligned}& E_{\boldsymbol{\theta}}\left\{\boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\right\}\left[E_{\boldsymbol{\theta}}\left\{(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\right\}\right]^{-1}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta})) \\ &= -E_{\boldsymbol{\theta}}\left\{\frac{\partial}{\partial\boldsymbol{\theta}^T}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right\}\left[E_{\boldsymbol{\theta}}\left\{(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\right\}\right]^{-1}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta})) \\ &= \begin{pmatrix} m_{11} & \cdots & m_{1N} \\ m_{21} & \cdots & m_{2N}\end{pmatrix}\begin{pmatrix} V_1(\boldsymbol{\theta}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_n(\boldsymbol{\theta})\end{pmatrix}^{-1}\begin{pmatrix} x_1-\mu_1(\boldsymbol{\theta}) \\ \vdots \\ x_N-\mu_N(\boldsymbol{\theta})\end{pmatrix} = \begin{pmatrix} q^1(\boldsymbol{x},\boldsymbol{\theta}) \\ q^2(\boldsymbol{x},\boldsymbol{\theta})\end{pmatrix},\end{aligned}$$

where $\boldsymbol{x} = (x_1, \dots, x_N)^T$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \dots, \mu_N(\boldsymbol{\theta}))^T$. In addition, the estimating function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ is not integrable with respect to $\boldsymbol{\theta}$ since $\partial q^1/\partial\theta^2 \neq \partial q^2/\partial\theta^1$. From Proposition 2 and the fact that $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ itself is a standardized estimating function, we immediately obtain the pre-contrast function $\rho_q : TS \times S \to \boldsymbol{R}$ defined by $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$, where

$$\rho_q((\partial_i)_{p_1}, p_2) = -\sum_{\boldsymbol{x}} q^i(\boldsymbol{x},\boldsymbol{\theta}_1)p(\boldsymbol{x};\boldsymbol{\theta}_2) = \sum_{n=1}^{N}\frac{m_{in}\{\mu_n(\boldsymbol{\theta}_1)-\mu_n(\boldsymbol{\theta}_2)\}}{V_n(\boldsymbol{\theta}_1)}$$

with $p_l(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_l) \in S$ $(l = 1, 2)$. The pre-contrast function ρ_q induces the statistical manifold admitting torsion as follows.
Riemannian metric g:

$$g_{ij}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} q^i(\boldsymbol{x},\boldsymbol{\theta})q^j(\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{n=1}^{N}\frac{1}{V_n(\boldsymbol{\theta})}m_{in}m_{jn}.$$

Dual affine connections ∇^* and ∇:

$$\begin{aligned}\Gamma^*_{ij,k}(\boldsymbol{\theta}) &= \sum_{\boldsymbol{x}}\{\partial_i\partial_j p(\boldsymbol{x};\boldsymbol{\theta})\}q^k(\boldsymbol{x},\boldsymbol{\theta}) \\ &= \sum_{n=1}^{N}\frac{m_{kn}}{V_n(\boldsymbol{\theta})}\left[\sum_{\boldsymbol{x}} x_n\{\partial_i\partial_j p(\boldsymbol{x};\boldsymbol{\theta})\} - \mu_n(\boldsymbol{\theta})\sum_{\boldsymbol{x}}\{\partial_i\partial_j p(\boldsymbol{x};\boldsymbol{\theta})\}\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} \frac{m_{kn}}{V_n(\boldsymbol{\theta})} \left[\partial_i \partial_j \sum_{\boldsymbol{x}} x_n p(\boldsymbol{x}; \boldsymbol{\theta}) - \mu_n(\boldsymbol{\theta}) \partial_i \partial_j \sum_{\boldsymbol{x}} p(\boldsymbol{x}; \boldsymbol{\theta}) \right] \\
&= \sum_{n=1}^{N} \frac{m_{kn}}{V_n(\boldsymbol{\theta})} \partial_i \partial_j \mu_n(\boldsymbol{\theta}) = 0 \ \ \text{(from (10))} \\
\Gamma_{ij,k}(\boldsymbol{\theta}) &= \Gamma^*_{ij,k}(\boldsymbol{\theta}) - \partial_i g_{jk}(\boldsymbol{\theta}) \ \ \text{(from the duality between } \nabla \text{ and } \nabla^*) \\
&= \sum_{n=1}^{N} \frac{1 - 2\theta^i}{V_n(\boldsymbol{\theta})^2} m_{in} m_{jn} m_{kn}.
\end{aligned}
$$

In this example, the statistical model S is ∇^*-flat since the coefficient of ∇^* with respect to the parameter $\boldsymbol{\theta}$ is equal to zero. Furthermore, this shows that $\boldsymbol{\theta}$ provides an affine coordinate system for ∇^*. Although the curvature tensor of ∇ vanishes because the curvature tensor of ∇^* vanishes and ∇ is dual to ∇^*, the statistical model S is not ∇-flat because ∇ is not torsion-free, which comes from the non-integrability of the estimating function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$. Hence, this geometrical structure provides an example of partially flat spaces, which was discussed in Sect. 4.

7 Future Problems

In this paper, we have summarized existing results on statistical manifolds admitting torsion, especially focusing on partially flat spaces. Although some results that are not seen in the standard theory of information geometry have been obtained, including a generalized projection theorem in partially flat spaces and statistical manifolds admitting torsion induced from estimating functions in statistics, a lot of (essential) problems have been unsolved. We discuss some of them to conclude this paper.

(1) The canonical pre-contrast function and the generalized projection theorem in a partially flat spase (M, g, ∇, ∇^*) are described only in terms of the flat connection ∇^*. In this sense, it can be said that these are a concept and a theorem for the Riemannian manifold (M, g) with the flat connection ∇^*. What is the role of the affine connection ∇ in the partially flat space (M, g, ∇, ∇^*), especially when ∇ is not torsion-free?

(2) The canonical pre-contrast function is defined in terms of the Riemannian metric g and the ∇^*-geodesic in a partially flat space (U, g, ∇, ∇^*) *without* using the affine coordinate system (η_i) on U. Hence, this function can be defined in a general statistical manifold admitting torsion (M, g, ∇) as long as the ∇^*-geodesic uniquely exists. What is the condition under which this function is a pre-contrast function that induces the original Riemannian metric g, dual affine connections ∇ and ∇^*? What properties does the (canonical) pre-contrast function have in this case? These problems are closely related to the works by [10, 11], who try to define a canonical divergence (canonical contrast function) on a general statistical manifold beyond a dually flat space.

(3) The definition of pre-contrast functions from estimating functions is obtained by replacing the score function which appears in the pre-contrast function as the derivative of Kullback–Leibler divergence with the standardized estimating functions. However, this is not the unique way to obtain a pre-contrast function from an estimating function. For example, if we consider the β-divergence [16] (or density power divergence [17]) as a contrast function, its first derivative is also a pre-contrast function and takes the same form as (9) in Proposition 2. However, the estimating function which appears in the pre-contrast function is not standardized. Although the standardization seems to be natural, further consideration is necessary on how to define a pre-contrast function from a given estimating function.

(4) For the example considered in Sect. 6, we can show that the pre-contrast function ρ_q coincides with the canonical pre-contrast function in the partially flat space (S, g, ∇, ∇^*) and the generalized projection theorem (Corollary 1 in Sect. 4) can be applied. However, its statistical meaning has not been clarified yet. Although it is expected that the SMAT induced from an estimating function has something to do with statistical inference based on the estimating function, the clarification on it is a future problem.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers JP15K00064, JP15K04842.

References

1. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**, 631–647 (1992)
2. Matsuzoe, H.: Geometry of contrast functions and conformal geometry. Hiroshima Math. J. **29**, 175–191 (1999)
3. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence; Oxford University Press, Oxford (2000)
4. Amari, S.: Information Geometry and Its Applications. Springer, Berlin (2016)
5. Kurose, T.: Statistical manifolds admitting torsion. Geometry and Something, Fukuoka University (2007)
6. Matsuzoe, H.: Statistical manifolds admitting torsion and pre-contrast functions. Information Geometry and Its Related Fields, Osaka City University (2010)
7. Henmi, M., Matsuzoe, H.: Geometry of pre-contrast functions and non-conservative estimating functions. AIP Conf. Proc. **1340**, 32–41 (2011)
8. Henmi, M.: Statistical manifolds admitting torsion, pre-contrast functions and estimating functions. Lect. Notes Comput. Sci. **10589**, 153–161 (2017)
9. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tohoku Math. J. **46**, 427–433 (1994)
10. Henmi, M., Kobayashi, R.: Hooke's law in statistical manifolds and divergences. Nagoya Math. J. **159**, 1–24 (2000)
11. Ay, N., Amari, S.: A novel approach to canonical divergences within information geometry. Entropy **17**, 8111–8129 (2015)
12. van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (2000)
13. Heyde, C.C.: Quasi-Likelihood and Its Application. Springer, Berlin (1997)
14. Godambe, V.: An optimum property of regular maximum likelihood estimation. Ann. Math. Stat. **31**, 1208–1211 (1960)

15. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, Boca Raton (1989)
16. Eguchi, S., Kano, Y.: Robustifying maximum likelihood estimation. In: Research Memorandum of the Institute of Statistical Mathematics, vol. 802 (2001)
17. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimizing a density power divergence. Biometrika **85**, 549–559 (1998)

Conformal Flattening on the Probability Simplex and Its Applications to Voronoi Partitions and Centroids

Atsumi Ohara

Abstract A certain class of information geometric structure can be conformally transformed to dually flat one. This paper studies the transformation on the probability simplex from a viewpoint of *affine differential geometry* and provides its applications. By restricting affine immersions with certain conditions, the probability simplex is realized to be 1-conformally flat statistical manifolds immersed in $\mathbf{R}^{n+1}$. Using this fact, we introduce a concept of *conformal flattening* for such manifolds in order to obtain the corresponding dually flat statistical (Hessian) ones with conformal divergences, and show explicit forms of potential functions and affine coordinates. Finally, we demonstrate applications of the flattening to nonextensive statistical physics, Voronoi partitions and weighted centroids on the probability simplex with respect to *geometric divergences*, which are not necessarily of Bregman type.

Keywords Conformal flattening · Affine differential geometry · Escort probability · Geometric divergence · Conformal divergence

1 Introduction

In the theory of information geometry for statistical models, the logarithmic function is crucially significant to give a standard information geometric structure for exponential family [1, 2]. By changing the logarithmic function to another one we can deform the standard structure to a new one while keeping its basic property as a statistical manifold, which consists of a pair of mutually dual affine connections (∇, ∇^*) with respect to Riemannian metric g. There exist several ways [3–6] to introduce functions to deform a statistical manifold structure and these functions are sometimes called *embedding* or *representing functions*.

A. Ohara (✉)
Department of Electrical and Electronics Engineering,
University of Fukui, Bunkyo 3-9-1, Fukui 910-8507, Japan
e-mail: ohara@fuee.u-fukui.ac.jp

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_4

Affine immersion [7] can be regarded as one of possible ways. Further, Kurose [8] has proved that *1-conformally flat* statistical manifolds (See Appendix) realized by a certain class of affine immersions can be conformally transformed to dually flat ones, which are the most fruitful information geometric structures.

In this paper we call the transformation *conformal flattening* and give its explicit formula in order to elucidate the relations between representing functions and realized information geometric structures. We also discuss its applicability to computational geometric topics. These are interpreted as generalizations of the results in [9, 10], where the arguments are limited to conformal flattening of the alpha-geometry [1, 2] (See also Sect. 2.4).

The paper is organized as follows: In Sect. 2 we first discuss the affine immersion of the probability simplex and its geometric structure realized by the associated *geometric divergence*. Next, the conformally flattening transformation is given and the obtained dually flat structure with the associated *conformal divergence* is investigated. Section 3 describes applications of the conformal flattening. We consider a Voronoi partition and a weighted centroid with respect to the geometric divergence on the probability simplex. While geometric divergences are not of Bregman type in general, geometric properties such as conformality and projectivity are well utilized in these topics. We also see that *escort probabilities*, which are interpreted as the dual affine coordinates for the flattened geometry, play important roles. Section 4 includes concluding remarks. Finally, a short review on statistical manifolds and affine differential geometry is given in Appendix.

2 Affine Immersion of the Probability Simplex

Let $\mathcal{S}^n$ be the relative interior of the probability simplex, i.e.,

$$\mathcal{S}^n := \left\{ p = (p_i) \,\middle|\, p_i \in \mathbf{R}_+, \ \sum_{i=1}^{n+1} p_i = 1 \right\},$$

where $\mathbf{R}_+$ denotes the set of positive numbers.

Consider an affine immersion [7] (f, ξ) of the simplex $\mathcal{S}^n$ (see also Appendix). Let D be the canonical flat affine connection on $\mathbf{R}^{n+1}$. Further, let f be an immersion from $\mathcal{S}^n$ into $\mathbf{R}^{n+1}$ and ξ be a transversal vector field on $\mathcal{S}^n$ (cf. Fig. 1). For a given *affine immersion* (f, ξ) of $\mathcal{S}^n$, the induced torsion-free connection ∇ and the affine fundamental form h are defined from the Gauss formula by

$$D_X f_*(Y) = f_*(\nabla_X Y) + h(X, Y)\xi, \quad X, Y \in \mathcal{X}(\mathcal{S}^n), \tag{1}$$

where f_* is the differential of f and $\mathcal{X}(\mathcal{S}^n)$ is the set of vector fields on $\mathcal{S}^n$.

It is well known [7, 8] that the realized geometric structure $(\mathcal{S}^n, \nabla, h)$ is a statistical manifold if and only if (f, ξ) is nondegenerate and equiaffine, i.e., h is nondegenerate

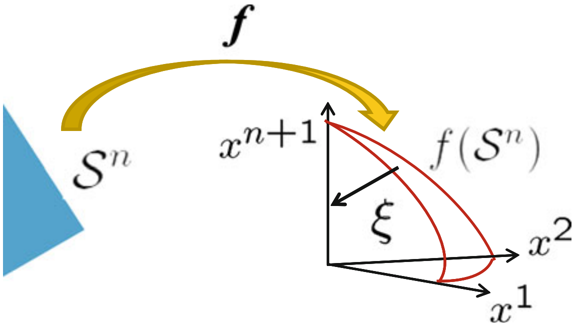


Fig. 1 An affine immersion (f, ξ) from $\mathcal{S}^n$ to $\mathbf{R}^{n+1}$

and $D_X\xi$ is tangent to $\mathcal{S}^n$ for any $X \in \mathcal{X}(\mathcal{S}^n)$. Furthermore, the statistical manifold $(\mathcal{S}^n, \nabla, h)$ is 1-conformally flat [8] (but not necessarily dually flat nor of constant curvature).

Now we consider the affine immersion with the following assumptions.

Assumptions 1. The affine immersion (f, ξ) is nondegenerate and equiaffine,

2. Let $\{x^i\}$ be an affine coordinate system for D on $\mathbf{R}^{n+1}$. The immersion f is given by the component-by-component and a common representing function L, i.e.,

$$f : \mathcal{S}^n \ni p = (p_i) \mapsto x = (x^i) \in \mathbf{R}^{n+1}, \quad x^i = L(p_i), \; i = 1, \ldots, n+1,$$

3. The representing function $L : (0, \; 1) \to \mathbf{R}$ is sign-definite (or non-zero), concave with $L'' < 0$ and strictly increasing, i.e., $L' > 0$, Hence, the inverse of L denoted by E exists, i.e., $E \circ L = \text{id}$.
4. Each component of ξ satisfies $\xi^i < 0, \; i = 1, \ldots, n+1$ on $\mathcal{S}^n$.

Remark 1 From the assumption 3, it follows that $L'E' = 1$, $E' > 0$ and $E'' > 0$. Regarding sign-definiteness of L, note that we can adjust $L(u)$ to $L(u) + c$ by a suitable constant c without loss of generality since the resultant geometric structure is unchanged (See Proposition 1) by the adjustment. For a fixed L satisfying the assumption 3, we can choose ξ that meets the assumptions 1 and 4. For example, if we take $\xi^i = -|L(p_i)|$ then (f, ξ) is called *centro-affine*, which is known to be equiaffine [7]. The assumptions 3 and 4 also assure positive definiteness of h (The details are described in the proof of Proposition 1). Hence, (f, ξ) is non-degenerate and we can regard h as a Riemannian metric on $\mathcal{S}^n$.

2.1 Conormal Vector and the Geometric Divergence

Define a function Ψ on $\mathbf{R}^{n+1}$ by

$$\Psi(x) := \sum_{i=1}^{n+1} E(x^i),$$

then $f(\mathcal{S}^n)$ immersed in $\mathbf{R}^{n+1}$ is expressed as a level surface of $\Psi(x) = 1$. Denote by $\mathbf{R}_{n+1}$ the dual space of $\mathbf{R}^{n+1}$ and by $\langle \nu, x \rangle$ the pairing of $x \in \mathbf{R}^{n+1}$ and $\nu \in \mathbf{R}_{n+1}$. The conormal vector [7] $\nu : \mathcal{S}^n \to \mathbf{R}_{n+1}$ for the affine immersion (f, ξ) is defined by

$$\langle \nu(p), f_*(X) \rangle = 0, \ \forall X \in T_p\mathcal{S}^n, \qquad \langle \nu(p), \xi(p) \rangle = 1 \tag{2}$$

for $p \in \mathcal{S}^n$. Using the assumptions and noting the relations:

$$\frac{\partial \Psi}{\partial x^i} = E'(x^i) = \frac{1}{L'(p_i)} > 0, \quad i = 1, \dots, n+1,$$

we have

$$\nu_i(p) := \frac{1}{\Lambda} \frac{\partial \Psi}{\partial x^i} = \frac{1}{\Lambda(p)} E'(x^i) = \frac{1}{\Lambda(p)} \frac{1}{L'(p_i)}, \quad i = 1, \dots, n+1, \tag{3}$$

where Λ is a normalizing factor defined by

$$\Lambda(p) := \sum_{i=1}^{n+1} \frac{\partial \Psi}{\partial x^i} \xi^i = \sum_{i=1}^{n+1} \frac{1}{L'(p_i)} \xi^i(p). \tag{4}$$

Then we can confirm (2) using the relation $\sum_{i=1}^{n+1} X^i = 0$ for $X = (X^i) \in \mathcal{X}(\mathcal{S}^n)$. Note that $v : \mathcal{S}^n \to \mathbf{R}_{n+1}$ defined by

$$v_i(p) := \Lambda(p)\nu_i(p) = \frac{1}{L'(p_i)}, \quad i = 1, \dots, n+1,$$

also satisfies

$$\langle v(p), f_*(X) \rangle = 0, \ \forall X \in T_p\mathcal{S}^n. \tag{5}$$

Further, it follows, from (3), (4) and the assumption 4, that

$$\Lambda(p) < 0, \quad \nu_i(p) < 0, \quad i = 1, \dots, n+1,$$

for all $p \in \mathcal{S}^n$.

It is known [7] that the affine fundamental form h can be represented by

$$h(X,Y) = -\langle \nu_*(X), f_*(Y)\rangle, \quad X, Y \in T_p\mathcal{S}^n.$$

In our case, it is calculated via (5) as

$$\begin{aligned} h(X,Y) &= -\Lambda^{-1}\langle v_*(X), f_*(Y)\rangle - X(\Lambda^{-1})\langle v, f_*(Y)\rangle \\ &= -\frac{1}{\Lambda}\sum_{i=1}^{n+1}\left(\frac{1}{L'(p_i)}\right)' L'(p_i)X^iY^i = \frac{1}{\Lambda}\sum_{i=1}^{n+1}\frac{L''(p_i)}{L'(p_i)}X^iY^i. \end{aligned}$$

Since h is positive definite from the assumptions 3 and 4, we can regard it as a Riemannian metric.

Utilizing these notions from affine differential geometry, we can introduce the function ρ on $\mathcal{S}^n \times \mathcal{S}^n$, which is called a *geometric divergence* [8], as follows:

$$\begin{aligned} \rho(p,r) &= \langle \nu(r), f(p) - f(r)\rangle = \sum_{i=1}^{n+1}\nu_i(r)(L(p_i) - L(r_i)) \\ &= \frac{1}{\Lambda(r)}\sum_{i=1}^{n+1}\frac{L(p_i)-L(r_i)}{L'(r_i)}, \qquad p, r \in \mathcal{S}^n. \end{aligned} \tag{6}$$

We can easily see that ρ is a contrast function [2, 11] of the geometric structure $(\mathcal{S}^n, \nabla, h)$ because it holds that

$$\rho[X|] = 0, \quad h(X,Y) = -\rho[X|Y], \tag{7}$$

$$h(\nabla_X Y, Z) = -\rho[XY|Z], \quad h(Y, \nabla^*_X Z) = -\rho[Y|XZ], \tag{8}$$

where $\rho[X_1 \dots X_k | Y_1 \dots Y_l]$ stands for

$$\rho[X_1 \dots X_k|Y_1 \dots Y_l](p) := (X_1)_p \dots (X_k)_p (Y_1)_r \dots (Y_l)_r \rho(p,r)|_{p=r}$$

for $p, r \in \mathcal{S}^n$ and $X_i, Y_j \in \mathcal{X}(\mathcal{S}^n)$.

2.2 Conformal Divergence and 1-Conformal Transformation

Let σ be a positive function on $\mathcal{S}^n$. Associated with the geometric divergence ρ, the *conformal divergence* [8] of ρ with respect to a conformal factor $\sigma(r)$ is defined by

$$\tilde\rho(p,r) = \sigma(r)\rho(p,r), \qquad p, r \in \mathcal{S}^n. \tag{9}$$

The divergence $\tilde{\rho}$ can be proved to be a contrast function for $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$, which is 1-conformally transformed geometric structure from $(\mathcal{S}^n, \nabla, h)$, where $\tilde{h}$ and $\tilde{\nabla}$ are given by

$$\tilde{h} = \sigma h, \tag{10}$$

$$h(\tilde{\nabla}_X Y, Z) = h(\nabla_X Y, Z) - d(\ln \sigma)(Z)h(X, Y). \tag{11}$$

When there exists such a positive function σ that relates $(\mathcal{S}^n, \nabla, h)$ with $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ as in (10) and (11), they are called 1-conformally equivalent and $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ is also a statistical manifold [8].

2.3 Main Result

Generally, the induced structure $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ from the conformal divergence $\tilde{\rho}$ is not also dually flat, which is the most abundant structure in information geometry. However, by choosing the conformal factor σ carefully, we can demonstrate that $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ is dually flat. Hereafter, we call such a transformation as *conformal flattening*.

Define

$$Z(p) := \sum_{i=1}^{n+1} \nu_i(p) = \frac{1}{\Lambda(p)} \sum_{i=1}^{n+1} \frac{1}{L'(p_i)},$$

then it is negative because each $\nu_i(p)$ is. The conformal divergence of ρ with respect to the conformal factor $\sigma(r) := -1/Z(r)$ is

$$\tilde{\rho}(p, r) = -\frac{1}{Z(r)} \rho(p, r).$$

Proposition 1 *If the conformal factor is given by $\sigma = -1/Z$, then the statistical manifold $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ that is 1-conformally transformed from $(\mathcal{S}^n, \nabla, h)$ via (10) and (11) is dually flat. Further, $\tilde{\rho}$ is the canonical divergence where mutually dual pair of affine coordinates (θ^i, η_i) and a pair of potential functions (ψ, φ) are explicitly given by*

$$\theta^i(p) = x^i(p) - x^{n+1}(p) = L(p_i) - L(p_{n+1}), \quad i = 1, \dots, n \tag{12}$$

$$\eta_i(p) = \frac{\nu_i(p)}{Z(p)} =: P_i(p), \quad i = 1, \dots, n, \tag{13}$$

$$\psi(p) = -x_{n+1}(p) = -L(p_{n+1}), \tag{14}$$

$$\varphi(p) = \frac{1}{Z(p)} \sum_{i=1}^{n+1} \nu_i(p) x^i(p) = \sum_{i=1}^{n+1} P_i(p) L(p_i). \tag{15}$$

Proof Using given relations, we first show that the conformal divergence $\tilde{\rho}$ is the canonical divergence [2] for $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$:

$$\begin{aligned}
\tilde{\rho}(p, r) &= -\frac{1}{Z(r)}\langle \nu(r), f(p) - f(r)\rangle = \langle P(r), f(r) - f(p)\rangle \\
&= \sum_{i=1}^{n+1} P_i(r)(x^i(r) - x^i(p)) \\
&= \sum_{i=1}^{n+1} P_i(r)x^i(r) - \sum_{i=1}^{n} P_i(r)(x^i(p) - x^{n+1}(p)) - \left(\sum_{i=1}^{n+1} P_i(r)\right) x^{n+1}(p) \\
&= \varphi(r) - \sum_{i=1}^{n} \eta_i(r)\theta^i(p) + \psi(p). \qquad (16)
\end{aligned}$$

Next, let us confirm that $\partial\psi/\partial\theta^i = \eta_i$. Since $\theta^i(p) = L(p_i) + \psi(p),\ i = 1, \ldots, n$, we have

$$p_i = E(\theta^i - \psi), \quad i = 1, \ldots, n+1,$$

by setting $\theta^{n+1} := 0$. Hence, we have

$$1 = \sum_{i=1}^{n+1} E(\theta^i - \psi).$$

Differentiating by θ^j, we have

$$\begin{aligned}
0 = \frac{\partial}{\partial\theta^j}\sum_{i=1}^{n+1} E(\theta^i - \psi) &= \sum_{i=1}^{n+1} E'(\theta^i - \psi)\left(\delta^i_j - \frac{\partial\psi}{\partial\theta^j}\right) \\
&= E'(x^j) - \left(\sum_{i=1}^{n+1} E'(x^i)\right)\frac{\partial\psi}{\partial\theta^j}.
\end{aligned}$$

This implies that

$$\frac{\partial\psi}{\partial\theta^j} = \frac{E'(x^j)}{\sum_{i=1}^{n+1} E'(x^i)} = P_j = \eta_j.$$

Together with (16) and this relation, φ is confirmed to be the Legendre transform of ψ.

The dual relation $\partial\varphi/\partial\eta_i = \theta^i$ follows automatically from the property of the Legendre transform. □

Remark 2 Since the conformal metric is $\tilde{h} = -h/Z$, it is also positive definite. The dual affine connections ∇^* and $\tilde{\nabla}^*$ are known to be projectively equivalent [8]. Hence, ∇^* is projectively (or -1-conformally) flat. Further, the following corollary implies that the realized affine connection ∇ is also projectively equivalent to the flat connection $\tilde{\nabla}$ if we use the centro-affine immersion, i.e., $\xi^i = -L(p_i)$ [7, 8] (See also Appendix). Note that the expressions of the dual coordinates $\eta_i(p) = P_i(p)$ can be interpreted as a generalization of the *escort probability* [12] because it is a normalization of deformed probabilities $1/L'(p_i)$ (see the following subsection).

Corollary 1 *The choice of ξ does not affect the obtained dually flat structure $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$.*

Proof We have the following alternative expressions of $\eta_i = P_i$ with respect to L and E:

$$P_i(p) = \frac{1/L'(p_i)}{\sum_{k=1}^{n+1} 1/L'(p_k)} = \frac{E'(x_i)}{\sum_{i=1}^{n+1} E'(x_i)} > 0, \quad i = 1, \ldots, n.$$

Hence, all the expressions in Proposition 1 does not depend on ξ, and the statement follows. □

2.4 Examples

Ex.(1) If we take L to be the logarithmic function $L(t) = \ln(t)$, the conformally flattened geometry immediately defines the standard dually flat structure (g^{F}, $\nabla^{(1)}$, $\nabla^{(-1)}$) on the simplex $\mathcal{S}^n$, where g^F denotes the Fisher metric. We see that $-\varphi(p)$ is the entropy, i.e., $\varphi(p) = \sum_{i=1}^{n+1} p_i \ln p_i$ and the conformal divergence is the KL divergence (relative entropy), i.e., $\tilde{\rho}(p, r) = D^{(\mathrm{KL})}(r||p) = \sum_{i=1}^{n+1} r_i(\ln r_i - \ln p_i)$.

Ex.(2) Next let the affine immersion (f, ξ) be defined by the following L and ξ:

$$L(t) := \frac{1}{1-q} t^{1-q}, \quad x^i(p) = \frac{1}{1-q}(p_i)^{1-q},$$

and

$$\xi^i(p) = -q(1-q)x^i(p),$$

with $0 < q$ and $q \neq 1$, then it realizes the alpha-geometry [2] ($\mathcal{S}^n$, $\nabla^{(\alpha)}$, g^F) with $q = (1+\alpha)/2$. Since the immersion (f, ξ) is centro-affine and the length of ξ is suitably scaled, ($\mathcal{S}^n$, $\nabla^{(\alpha)}$, g^F) is of constant curvature $\kappa = (1-\alpha^2)/4$. The associated geometric divergence is the alpha-divergence, i.e.,

$$\rho(p, r) = D^{(\alpha)}(p, r) = \frac{4}{1-\alpha^2}\left(1 - \sum_{i=1}^{n+1}(p_i)^{(1-\alpha)/2}(r_i)^{(1+\alpha)/2}\right). \tag{17}$$

Following the procedure of conformal flattening described in the above, we have [9]

$$\Psi(x) = \sum_{i=1}^{n+1} ((1-q)x^i)^{1/1-q}, \quad \Lambda(p) = -q, \ (constant)$$

$$\nu_i(p) = -\frac{1}{q}(p_i)^q, \quad -\frac{1}{Z(p)} = \frac{q}{\sum_{k=1}^{n+1}(p_i)^q},$$

and obtain dually flat structure $(\tilde{h}, \tilde{\nabla}, \tilde{\nabla}^*)$ via the formulas in Proposition 1:

$$\eta_i = \frac{(p_i)^q}{\sum_{k=1}^{n+1}(p_k)^q}, \quad \theta^i = \frac{1}{1-q}(p_i)^{1-q} - \frac{1}{1-q}(p_{n+1})^{1-q} = \ln_q(p_i) - \psi(p),$$

$$\psi(p) = -\ln_q(p_{n+1}), \quad \varphi(p) = \ln_q\left(\frac{1}{\exp_q(S_q(p))}\right), \quad \tilde{h}(p) = -\frac{1}{Z(p)}g^F(p).$$

Here, $\ln_q$ and $S_q(p)$ are the *q-logarithmic function* and the *Tsallis entropy* [12], respectively defined by

$$\ln_q(t) = \frac{t^{1-q}-1}{1-q}, \quad S_q(p) = \frac{\sum_{i=1}^{n+1}(p_i)^q - 1}{1-q}.$$

3 Construction of Voronoi Partitions and Centroids with Respect to Geometric Divergences

In the previous section we have seen that various geometric divergences ρ can be constructed on the statistical manifold $\mathcal{S}^n$ by changing the representing function L and the transversal vector field ξ.

We demonstrate interesting applications of the conformal flattening to topics related with computational geometry, which are Voronoi partitions and centroids for the geometric divergence on a 1-conformally flat statistical manifold. We find that escort probabilities P_i (dual coordinates η_i) play important roles.

In this section, subscripts by Greek letters such as p_λ are used to denote the λ-th point in $\mathcal{S}^n$ among given ones while subscripts by Roman letters such as p_i denote the ith coordinate of a point $p = (p_i) \in \mathcal{S}^n$.

3.1 Voronoi Partitions

Let ρ be a geometric divergence defined in (6) on a 1-conformal statistical manifold $(\mathcal{S}^n, \nabla, h)$. For given m points $p_\lambda, \ \lambda = 1, \ldots, m$ on $\mathcal{S}^n$ we define *Voronoi regions* on $\mathcal{S}^n$ with respect to the geometric divergence ρ as follows:

$$\mathrm{Vor}^{(\rho)}(p_\lambda) := \bigcap_{\mu \neq \lambda} \{r \in \mathcal{S}^n | \rho(p_\lambda, r) < \rho(p_\mu, r)\}, \quad \lambda = 1, \ldots, m.$$

An *Voronoi partition (diagram)* on $\mathcal{S}^n$ is a collection of the Voronoi regions and their boundaries. For example, if we take $L(t) = t^{1-q}/(1-q)$ as in Sect. 2.4, the corresponding Voronoi partition is the one with respect to the alpha-divergence $D^{(\alpha)}$ in (17) on $(\mathcal{S}^n, \nabla^{(\alpha)}, g^F)$ (Cf. the figures in [10]). Note that $D^{(\alpha)}$ approaches the Kullback–Leibler (KL) divergence if $\alpha \to -1$, and $D^{(0)}$ is called the Hellinger distance. Further, the partition is also equivalent to that with respect to *Rényi divergence* [13] defined by

$$D_\alpha(p, r) := \frac{1}{\alpha - 1} \ln \sum_{i=1}^{n+1} (p_i)^\alpha (r_i)^{1-\alpha}$$

because of their one-to-one functional relationship.

The acclaimed algorithm using projection of a convex polyhedron [14, 15] has been known to commonly work well to construct Voronoi partitions for the KL divergence [16–18] as well as the Euclidean distance. Furthermore, the algorithm is generally applicable if a divergence function δ is of *Bregman type* [19], which is represented by the remainder of the first order Taylor expansion of a convex potential function in a suitable coordinate system. Geometrically speaking, this implies that

(i) the divergence δ is a *canonical divergence* [2] associated with a dually flat structure, i.e, it is of Bregman type:

$$\begin{aligned} \delta(p, r) &= \psi(\theta(r)) + \varphi(\eta(r)) - \sum_{i=1}^{n} \theta^i(p)\eta_i(r) \\ &= \varphi(\eta(r)) - \left\{ \varphi(\eta(p)) + \sum_{i=1}^{n} \theta^i(p) \left(\eta_i(r) - \eta_i(p)\right) \right\}, \qquad (18) \\ \theta^i &= \frac{\partial \varphi(\eta)}{\partial \eta_i}, \quad i = 1, \ldots, n, \end{aligned}$$

(ii) its affine coordinate system $\eta = (\eta_i)$ is chosen to realize the corresponding Voronoi partitions. In this coordinate system with one extra complementary coordinate the polyhedron is expressed as the upper envelope of m hyperplanes tangent to the potential function $\varphi(\eta)$ at $\eta(p_\lambda), \ \lambda = 1, \ldots, m$.

Unfortunately a problem for the case of our Voronoi partition is that the geometric divergences ρ on $\mathcal{S}^n$ is *not* of Bregman type generally, i.e., they *cannot* be represented as a remainder of any convex potentials as in (18).

The following theorem, however, claims that the problem is resolved via Proposition 1. In other words, we can still apply the projection algorithm by conformally flattening a statistical manifold $(\mathcal{S}, \nabla, h)$ to a dually flat structure $(\mathcal{S}, \tilde{\nabla}, \tilde{h})$ and by invoking the conformal divergence $\tilde{\rho}$, which is always of Bregman type, and escort probabilities $\eta_i(p) = P_i(p)$ as a coordinate system.

The similar result is proved in [10] for the case of the α-divergence $D^{(\alpha)}$. However, the proof there was based on the fact that $(\mathcal{S}^n, \nabla^{(\alpha)}, g^F)$ is a statistical manifold *of constant curvature* in order to use the *modified Pythagorean relation* (See Appendix). In the following theorem, the assumption is relaxed to a 1-conformally flat statistical manifold $(\mathcal{S}, \nabla, h)$ and we prove with the usual Pythagorean relation on dually flat space.

Here, we denote the space of escort distributions by $\mathcal{E}^n$ and represent the point on $\mathcal{E}^n$ by $P = (P_1, \dots, P_n)$ because $P_{n+1} = 1 - \sum_{i=1}^n P_i$ and $\mathcal{E}^n$ is also the probability simplex.

Theorem 1 *(i) The bisector of two points p_λ and p_μ defined by $\{r|\rho(p_\lambda, r) = \rho(p_\mu, r)\}$ is a simultaneously ∇^*- and $\tilde{\nabla}^*$-autoparallel hypersurface on $\mathcal{S}^n$.*

(ii) Let $\mathcal{H}_\lambda, \lambda = 1, \dots, m$ be the hyperplane in $\mathcal{E}^n \times \mathbf{R}$ which is respectively tangent at $(P(p_\lambda), \varphi(p_\lambda))$ to the hypersurface $\{(P, y) = (P(p), \varphi(p))|p \in \mathcal{S}^n\}$. The Voronoi partition with respect to ρ can be constructed on $\mathcal{E}^n$ by projecting the upper envelope of all $\mathcal{H}_\lambda$'s along the y-axis.

Proof (i) We construct a bisector for points p_λ and p_μ. Consider the $\tilde{\nabla}$-geodesic $\tilde{\gamma}$ connecting p_λ and p_μ, and let $\bar{p}$ be the midpoint on $\tilde{\gamma}$ satisfying $\tilde{\rho}(p_\lambda, \bar{p}) = \tilde{\rho}(p_\mu, \bar{p})$. Note that the point $\bar{p}$ satisfies $\rho(p_\lambda, \bar{p}) = \rho(p_\mu, \bar{p})$ by the conformal relation (9). Denote by $\mathcal{B}$ the $\tilde{\nabla}^*$-autoparallel hypersurface that is orthogonal to $\tilde{\gamma}$ at $\bar{p}$ with respect to the conformal metric $\tilde{h}$. Note that $\mathcal{B}$ is simultaneously ∇^*-autoparallel because of the projective equivalence of ∇^* and $\tilde{\nabla}^*$ as is mentioned in Remark 2.

Using these setup and the fact that $(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ is dually flat, we have the following relation from the Pythagorean theorem [2]

$$\tilde{\rho}(p_\lambda, r) = \tilde{\rho}(p_\lambda, \bar{p}) + \tilde{\rho}(\bar{p}, r) = \tilde{\rho}(p_\mu, \bar{p}) + \tilde{\rho}(\bar{p}, r) = \tilde{\rho}(p_\mu, r),$$

for all $r \in \mathcal{B}$. Using the conformal relation (9) again, we have $\rho(p_\lambda, r) = \rho(p_\mu, r)$ for all $r \in \mathcal{B}$. Hence, $\mathcal{B}$ is a bisector of p_λ and p_μ.

(ii) Recall the conformal relation (9) between ρ and $\tilde{\rho}$, then we see that $\mathrm{Vor}^{(\rho)}(p_\lambda) = \mathrm{Vor}^{(\tilde{\rho})}(p_\lambda)$ holds on $\mathcal{S}^n$, where

$$\mathrm{Vor}^{(\tilde{\rho})}(p_\lambda) := \bigcap_{\mu \neq \lambda} \{r \in \mathcal{S}^n | \tilde{\rho}(p_\lambda, r) < \tilde{\rho}(p_\mu, r)\}.$$

Proposition 1 and the Legendre relations (16) imply that $\tilde{\rho}(p_\lambda, r)$ is represented with the escort probabilities, i.e., the dual coordinates $(P_i) = (\eta_i)$ by

$$\tilde{\rho}(p_\lambda, r) = \varphi(P(r)) - \left(\varphi(P(p_\lambda)) + \sum_{i=1}^n \frac{\partial \varphi}{\partial P_i}(p_\lambda)\{P_i(r) - P_i(p_\lambda)\} \right),$$

By definition the hyperplane $\mathcal{H}_\lambda$ is expressed by

$$\mathcal{H}_\lambda = \left\{ (P(r), y(r)) \,\middle|\, y(r) = \varphi(P(p_\lambda)) + \sum_{i=1}^{n} \frac{\partial \psi^*}{\partial P_i}(p_\lambda)\{P_i(r) - P_i(p_\lambda)\},\ r \in \mathcal{S}^n \right\}.$$

Hence, we have $\tilde{\rho}(p_\lambda, r) = \varphi(P(r)) - y(r)$. Thus, we see, for example, that the bisector on $\mathcal{E}^n$ for p_λ and p_μ is represented as a projection of $\mathcal{H}_\lambda \cap \mathcal{H}_\mu$. Thus, the statement follows. □

As a special case of the above theorem for $\rho = D^{(\alpha)}$, examples of Voronoi partitions with respect to $D^{(\alpha)}$ on usual probability simplex $\mathcal{S}^n$ and escort probability simplex $\mathcal{E}^n$ are given with their figures in [10].

Remark 3 Voronoi partitions for broader class of divergences that are not necessarily associated with any convex potentials are theoretically studied [20] from more general affine differential geometric points of views.

On the other hand, if the domain is extended from $\mathcal{S}^n$ to the positive orthant $\mathbf{R}_+^{n+1}$, then the α-divergence there can be expressed as a Bregman divergence [1, 2, 21]. Hence, the α-geometry on $\mathbf{R}_+^{n+1}$ is dually flat. Using this property, α-Voronoi partitions on $\mathbf{R}_+^{n+1}$ is discussed in [22].

However, while both of the above mentioned methods require constructions of the convex polyhedrons in the space of dimension $d = n + 2$, the new one proposed in this paper does in the space of dimension $d = n + 1$. Since it is known [23] that the optimal computational time of polyhedrons depends on the dimension d by $O(m \log m + m^{\lfloor d/2 \rfloor})$, the new one is slightly better when n is even and m is large.

3.2 Weighted Centroids

Let $p_\lambda,\ \lambda = 1, \ldots, m$ be given m points on $\mathcal{S}^n$ and $w_\lambda > 0,\ \lambda = 1, \ldots, m$ be their weights. Define the *weighted ρ-centroid $c^{(\rho)}$* $\in \mathcal{S}^n$ by the minimizer of the following problem:

$$\min_{p \in \mathcal{S}^n} \sum_{\lambda=1}^{m} w_\lambda \rho(p, p_\lambda).$$

Theorem 2 *The weighted ρ-centroid $c^{(\rho)}$ for given m points $p_1, \ldots, p_m$ on $\mathcal{S}^n$ is expressed by*

$$P_i(c^{(\rho)}) = \frac{1}{\sum_{\lambda=1}^{m} w_\lambda Z(p_\lambda)} \sum_{\lambda=1}^{m} w_\lambda Z(p_\lambda) P_i(p_\lambda), \quad i = 1, \ldots, n+1,$$

with weights w_λ, escort probabilities $P(p_\lambda)$ and the conformal factors $\sigma(p_\lambda) = -1/Z(p_\lambda) > 0$ for $p_\lambda,\ \lambda = 1, \ldots, m$.

Proof Denote $\theta^i(p)$ by θ^i simply. Using (9), we have

$$\begin{aligned}\sum_{\lambda=1}^m w_\lambda \rho(p, p_\lambda) &= -\sum_{\lambda=1}^m w_\lambda Z(p_\lambda)\tilde{\rho}(p, p_\lambda)\\ &= -\sum_{\lambda=1}^m w_\lambda Z(p_\lambda)\left\{\psi(\theta) + \psi^*(\eta(p_\lambda)) - \sum_{i=1}^n \theta^i \eta_i(p_\lambda)\right\}.\end{aligned}$$

Then the optimality condition is

$$\frac{\partial}{\partial\theta^i}\sum_{\lambda=1}^m w_\lambda \rho(p, p_\lambda) = -\sum_{\lambda=1}^m w_\lambda Z(p_\lambda)\{\eta_i - \eta_i(p_\lambda)\} = 0, \quad i = 1, \ldots, n,$$

where $\eta_i = \eta_i(p)$. Thus, the statements for $i = 1, \ldots, n$ hold from $\eta_i = P_i$ in Proposition 1. For $i = n + 1$, we have as follows:

$$\begin{aligned}P_{n+1}(c^{(\rho)}) &= 1 - \sum_{i=1}^n P_i(c^{(\rho)})\\ &= \frac{1}{\sum_{\lambda=1}^m w_\lambda Z(p_\lambda)}\sum_{\lambda=1}^m w_\lambda Z(p_\lambda)\left\{1 - \sum_{i=1}^n P_i(p_\lambda)\right\}\\ &= \frac{1}{\sum_{\lambda=1}^m w_\lambda Z(p_\lambda)}\sum_{\lambda=1}^m w_\lambda Z(p_\lambda) P_{n+1}(p_\lambda).\end{aligned}$$

□

4 Concluding Remarks

We have realized 1-conformally flat structures $(\mathcal{S}^n, \nabla, h)$ by changing affine immersions (f, ξ) or representing functions L, considered their conformal flattening and explicitly derived the corresponding dually flat structure, i.e., mutually dual potentials and affine coordinate systems.

Applications of the conformal flattening to topics in computational geometry are also demonstrated. As a result the geometric divergence, which is not generally of Bregman type, can be easily treated via the traditional computation algorithm. Recently, conformal divergences for Bregman-type divergences are proposed from different viewpoints and their properties are exploited [24, 25].

Extensions of the conformal flattening to other non-flat statistical manifolds or families of continuous probability distributions are left in the future work. While relations with the gradient flows (replicator flows, in a special case) on $(\mathcal{S}^n, \nabla, h)$ or

$(\mathcal{S}^n, \tilde{\nabla}, \tilde{h})$ can be found in [26], searching for the other applications of the technique would be also of interest.

Acknowledgements A. O. is partially supported by JSPS Grant-in-Aid (C) 15K04997.

Appendix: A Short Review of Statistical Manifolds and Affine Differential Geometry

We shortly summarize the basic notions and results in information geometry [1, 2], Hessian domain [27] and affine differential geometry [7, 8], which are used in this paper. See for the details and proofs in the literature.

Statistical Manifolds

For a torsion-free affine connection ∇ and a pseudo-Riemannian metric g on a manifold $\mathcal{M}$, the triple $(\mathcal{M}, \nabla, g)$ is called a *statistical (Codazzi) manifold* if it admits another torsion-free affine connection ∇^* satisfying

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z) \tag{19}$$

for arbitrary X, Y and Z in $\mathcal{X}(\mathcal{M})$, where $\mathcal{X}(\mathcal{M})$ is the set of all tangent vector fields on $\mathcal{M}$. We say that ∇ and ∇^* *duals* of each other with respect to g, and (g, ∇, ∇^*) is called *dualistic structure* on $\mathcal{M}$.

A statistical manifold $(\mathcal{M}, \nabla, g)$ is said to be of *constant curvature* $\kappa \in \mathbf{R}$ if the curvature tensor R of ∇ satisfies

$$R(X,Y)Z = \kappa\{g(Y,Z)X - g(X,Z)Y\}. \tag{20}$$

When the constant κ is zero, the statistical manifold is called *flat*, or *dually flat*, because the dual curvature tensor R^* of ∇^* also vanishes automatically [2, 27].

For $\alpha \in \mathbf{R}$, statistical manifolds $(\mathcal{M}, \nabla, g)$ and $(\mathcal{M}, \tilde{\nabla}, \tilde{g})$ are said to be α-*conformally equivalent* [8] if there exists a positive function σ on $\mathcal{M}$ satisfying

$$\begin{aligned}\tilde{g}(X,Y) &= \sigma g(X,Y)\\ g(\tilde{\nabla}_X Y, Z) &= g(\nabla_X Y, Z) - \frac{1+\alpha}{2}(d\ln\sigma)(Z)g(X,Y)\\ &\quad + \frac{1-\alpha}{2}\{(d\ln\sigma)(X)g(Y,Z) + (d\ln\sigma)(Y)g(X,Z)\}.\end{aligned}$$

Statistical manifolds $(\mathcal{M}, \nabla, g)$ and $(\mathcal{M}, \tilde{\nabla}, \tilde{g})$ are α-conformally equivalent if and only if $(\mathcal{M}, \nabla^*, g)$ and $(\mathcal{M}, \tilde{\nabla}^*, \tilde{g})$ are $-\alpha$-conformally equivalent. In particular, -1-conformal equivalence means *projective equivalence* of ∇ and $\tilde{\nabla}$, which implies that a ∇-pregeodesic curve is simultaneously $\tilde{\nabla}$-pregeodesic [7]. A statistical manifold $(\mathcal{M}, \nabla, g)$ is called α-*conformally flat* if it is locally α-conformally equivalent to a flat statistical manifold. It is known that a statistical manifold is of constant curvature if and only if it is ± 1-conformally flat, when $\dim \mathcal{M} \geq 3$ [8].

Affine Differential Geometry

Let $\mathcal{M}$ be an n-dimensional manifold and consider an *affine immersion* [7] (f, ξ), which is the pair of an immersion f from $\mathcal{M}$ into $\mathbf{R}^{n+1}$ and a transversal vector field ξ along $f(\mathcal{M})$. By a given affine immersion (f, ξ) of $\mathcal{M}$ and the usual flat affine connection D of $\mathbf{R}^{n+1}$, the Gauss and Weingarten formulas are respectively obtained as follows:

$$D_X f_*(Y) = f_*(\nabla_X Y) + h(X, Y)\xi,$$
$$D_X \xi = -f_*(SX) + \tau(X)\xi.$$

Here, ∇, h, S and τ are called, respectively, *induced connection, affine fundamental form, affine shape operator* and *transversal connection form*. In this case, we say that the affine immersion realizes $(\mathcal{M}, \nabla, h)$ in $\mathbf{R}^{n+1}$. If h is non-degenerate (resp. $\tau = 0$ on $\mathcal{M}$), the affine immersion (f, ξ) is called *non-degenerate* (resp. *equiaffine*). It is known that non-degenerate and equiaffine (f, ξ) realizes a statistical manifold $(\mathcal{M}, \nabla, h)$ by regarding h as a pseudo-Riemannian metric g.

Such a statistical manifold is characterized as follows:

Proposition 2 *[8] A simply connected statistical manifold* $(\mathcal{M}, \nabla, g)$ *can be realized by a non-degenerate and equiaffine immersion if and only if it is 1-conformally flat.*

Let a point o in $\mathbf{R}^{n+1}$ be chosen as origin and consider an immersion f from $\mathcal{M}$ to $\mathbf{R}^{n+1} \backslash \{o\}$ so that $\xi = -\overrightarrow{of(p)}$ is transversal to $f(\mathcal{M})$ for $p \in \mathcal{M}$. Such an affine immersion (f, ξ) is called *centro-affine*, where the Weingarten formula is $D_X \xi = -f_*(X)$, or $S = I$ and $\tau = 0$. This implies that a centro-affine immersion, if it is non-degenerate, realizes a statistical manifold of constant curvature because of the Gauss equation:

$$R(X, Y)Z = h(X, Z)SX - h(X, Z)SY.$$

Further, the realized affine connection ∇ is projectively flat [7].

Denote the dual space of $\mathbf{R}^{n+1}$ by $\mathbf{R}_{n+1}$ and the pairing of $x \in \mathbf{R}^{n+1}$ and $y \in \mathbf{R}_{n+1}$ by $\langle y, x\rangle$. Define a map $\nu : \mathcal{M} \to \mathbf{R}_{n+1}\backslash\{o\}$ as follows:

$$\langle \nu_p, \xi_p\rangle = 1, \quad \langle \nu_p, f_*(X)\rangle = 0 \quad (\forall X \in T_p\mathcal{M}).$$

Such ν_p is uniquely defined and is called the *conormal vector*.

The pair $(\nu, -\nu)$ can be regarded as a centro-affine immersion from $\mathcal{M}$ into the dual space $\mathbf{R}_{n+1}$ equipped with the usual flat connection D^*. The formulas are

$$\begin{aligned} D^*_X(\nu_* Y) &= \nu(\nabla^*_X Y) + h^*(X, Y)(-\nu), \\ D^*_X(-\nu) &= -\nu_*(X), \end{aligned}$$

where $h^*(X, Y) = h(SX, Y)$, and ∇^* is dual of ∇ with respect to h. Hence, when (f, ξ) realizes a statistical manifold $(\mathcal{M}.\nabla, h)$ with $S = I$, then $(\nu, -\nu)$ realizes its dual statistical manifold $(\mathcal{M}, \nabla^*, h)$ [7]. Both manifolds are of constant curvature.

For a statistical manifold $(\mathcal{M}, \nabla, h)$ realized by a non-degenerate and equiaffine immersion (f, ξ), we can define a *contrast function* ρ that induces the structure $(\mathcal{M}, \nabla, h)$

$$\rho(p, q) = \langle \nu(q), f(p) - f(q)\rangle, \quad (p, q \in \mathcal{M}).$$

The function ρ is called the *geometric divergence* of $(\mathcal{M}, \nabla, h)$ [8]. For a statistical manifold $(\mathcal{M}, \tilde{\nabla}, \tilde{h})$ that is 1-conformally equivalent to $(\mathcal{M}, \nabla, h)$, one of its contrast function is given by $\tilde{\rho}(p, q) = \sigma(q)\rho(p, q)$ for a certain positive function σ. The contrast function $\tilde{\rho}$ is called the *conformal divergence* [8].

A statistical manifold $(\mathcal{M}, \nabla, g)$ of constant curvature κ is studied from a viewpoint of affine differential geometry [8]. It is known that $(\mathcal{M}, \nabla, g)$ realized in $\mathbf{R}^{n+1}$ has the following geometric properties:

P1 For three points p, q and r in $\mathcal{M}$ let the ∇-geodesic connecting p and q and the ∇^*-geodesic connecting q and r are orthogonal at q. Then the following modified Pythagorean relation holds:

$$\rho(p, r) = \rho(p, q) + \rho(q, r) - \kappa\rho(p, q)\rho(q, r),$$

P2 An arbitrary ∇-geodesic on $\mathcal{M}$ is the intersection of a two-dimensional subspace in $\mathbf{R}^{n+1}$ and $\mathcal{M}$,

P3 The volume element θ on $\mathcal{M}$ induced from $\mathbf{R}^{n+1}$ satisfies $\nabla\theta = 0$,

and so on. A typical example of the statistical manifold of non-zero constant curvature is the alpha-geometry $(\mathcal{S}^n, \nabla^{(\alpha)}, g^F)$, where $\kappa = (1 - \alpha^2)/4$. In this case, the modified Pythagorean relation induces the widely-known nonextensivity relation of Tsallis entropy [21, Remark 2].

References

1. Amari, S.-I.: Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics, vol. 28. Springer, New York (1985)
2. Amari, S-I., Nagaoka, H.: Methods of Information Geometry. Translations of Mathematical Monographs, vol. 191. American Mathematical Society and Oxford University Press, Oxford (2000)
3. Zhang, J.: Divergence function, duality, and convex analysis. Neural Comput. **16**, 159–195 (2004)
4. Eguchi, S.: Information geometry and statistical pattern recognition. Sugaku Expos. **19**, 197–216 (2006). (originally 2004 Sugaku 56 380–399 in Japanese)
5. Naudts, J.: Continuity of a class of entropies and relative entropies. Rev. Math. Phys. **16**(6), 809–822 (2004)
6. Harsha, K.V., Subrahamanian Moosath, K.S.: F-geometry and Amari's α-geometry on a statistical manifold. Entropy **16**(5), 2472–2487 (2014)
7. Nomizu, K., Sasaki, T.: Affine Differential Geometry. Cambridge University Press, Cambridge (1993)
8. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tohoku Math. J. **46**, 427–433 (1994)
9. Ohara, A., Matsuzoe, H., Amari, S.-I.: A dually flat structure on the space of escort distributions. J. Phys. Conf. Ser. **201**, 012012 (2010)
10. Ohara, A., Matsuzoe, H., Amari, S.: Conformal geometry of escort probability and its applications. Mod. Phys. Lett. B **26**(10), 1250063-1–1250063-14 (2012)
11. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. 631–647 (1992)
12. Tsallis, C.: Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World. Springer, New York (2009)
13. Rényi, A.: On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 547–561. University of California Press, Berkeley (1961)
14. Edelsbrunner, H., Seidel, R.: Voronoi diagrams and arrangements. Discrete Comput. Geom. **1**, 25–44 (1986)
15. Edelsbrunner, H.: Algorithms in Combinatorial Geometry. Springer, Berlin (1987)
16. Onishi, K., Takayama, N.: Construction of Voronoi diagram on the upper half-plane. IEICE Trans. Fundam. **E79-A**, 533–539 (1996)
17. Onishi, K., Imai, H.: Voronoi diagram in statistical parametric space by Kullback-Leibler divergence. In: Proceedings of the 13th ACM-SIAM Symposium on Computational Geometry, pp. 463–465 (1997)
18. Imai, H., Inaba, M.: Divergence-based geometric clustering and its underlying discrete proximity structures. IEICE Trans. Inf. Syst. **E83-D**, 27–35 (2000)
19. Boissonnat, J.-D., Nielsen, F., Nock, N.: Bregman Voronoi diagram. Discrete Comput. Geom. **44**, 281–307 (2010)
20. Matsuzoe, H.: Computational Geometry from the Viewpoint of Affine Differential Geometry. In: Nielsen, F. (eds.) Emerging Trends in Visual Computing, pp. 103–123. Springer, Berlin (2009)
21. Ohara, A.: Geometry of distributions associated with Tsallis statistics and properties of relative entropy minimization. Phys. Lett. A **370**, 184–193 (2007)
22. Nielsen, F., Nock, R.: The dual Voronoi diagrams with respect to representational Bregman divergences. In: International Symposium on Voronoi Diagrams (ISVD), DTU Lyngby, Denmark. IEEE Press, New York (2009)
23. Chazelle, B.: An optimal convex hull algorithm in any fixed dimension. Discrete Comput. Geom. **10**, 377–409 (1993)
24. Nielsen, F., Nock, R.: Total Jensen divergences: definition, properties and clustering. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015)

25. Nock, R., Nielsen, F., Amari, S.-I.: On conformal divergences and their population minimizers. IEEE Trans. IT **62**(1), 527–538 (2016)
26. Ohara, A.: Conformal flattening for deformed information geometries on the probability simplex. Entropy **20**, 186 (2018)
27. Shima, H.: The Geometry of Hessian Structures. World Scientific, Singapore (2007)

Monte Carlo Information-Geometric Structures

Frank Nielsen and Gaëtan Hadjeres

Abstract Exponential families and mixture families are parametric probability models that can be geometrically studied as smooth statistical manifolds with respect to any statistical divergence like the Kullback–Leibler (KL) divergence or the Hellinger divergence. When equipping a statistical manifold with the KL divergence, the induced manifold structure is dually flat, and the KL divergence between distributions amounts to an equivalent Bregman divergence on their corresponding parameters. In practice, the corresponding Bregman generators of mixture/exponential families require to perform definite integral calculus that can either be too time-consuming (for exponentially large discrete support case) or even do not admit closed-form formula (for continuous support case). In these cases, the dually flat construction remains theoretical and cannot be used by information-geometric algorithms. To bypass this problem, we consider performing stochastic Monte Carlo (MC) estimation of those integral-based mixture/exponential family Bregman generators. We show that, under natural assumptions, these MC generators are almost surely Bregman generators. We define a series of dually flat information geometries, termed Monte Carlo Information Geometries, that increasingly-finely approximate the untractable geometry. The advantage of this MCIG is that it allows a practical use of the Bregman algorithmic toolbox on a wide range of probability distribution families. We demonstrate our approach with a clustering task on a mixture family manifold. We then show how to generate MCIG for arbitrary separable statistical divergence between distributions belonging to a same parametric family of distributions.

F. Nielsen (✉)
Sony Computer Science Laboratories, Tokyo, Japan
e-mail: Frank.Nielsen@acm.org

G. Hadjeres
Sony Computer Science Laboratory, Paris, France
e-mail: Gaetan.Hadjeres@sony.com

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_5

Keywords Computational information geometry · Statistical manifold
Dually flat information geometry · Bregman generator
Stochastic Monte Carlo integration · Mixture family · Exponential family
Clustering

1 Introduction

We concisely describe the construction and properties of dually flat spaces [1, 8] in Sect. 1.1, define the statistical manifolds of exponential families and mixture families in Sect. 1.2, and discuss about the computational tractability of Bregman algorithms in dually flat spaces in Sect. 1.3.

1.1 Dually Flat Space: Bregman Geometry

A smooth (potentially asymmetric) distance $D(\cdot,\cdot)$ is called a *divergence* in information geometry [1, 8], and induces a differential-geometric dualistic structure [1, 2, 8, 17]. In particular, a strictly convex and twice continuously differentiable D-dimensional real-valued function F, termed a *Bregman generator*, induces a dually connection-flat structure via a corresponding Bregman Divergence (BD) [4] $B_F(\cdot,\cdot)$ given by:

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2)\rangle, \tag{1}$$

where $\langle y, x\rangle := y^\top x$ denotes the inner product, and $\nabla F(\theta) := (\partial_i F(\theta))_i$ denotes the gradient vector of partial first-order derivatives with respect to vector parameter θ. We use the standard notational convention of information geometry [1, 8]: $\partial_i :=: \frac{\partial}{\partial \theta^i}$ to indicate[1] a contravariant vector [18] $\theta = (\theta^i)_i$.

The Legendre–Fenchel transformation [30] :

$$F^*(\eta) = \sup_\theta\{\langle \theta, \eta\rangle - F(\theta)\}, \tag{2}$$

is at the heart of the duality of flat structures by defining two global affine coordinate systems: The *primal affine θ-coordinate system* and the *dual affine η-coordinate system*, so that any point P of the manifold $\mathcal{M}$ can either be accessed by its *primal* $\theta(P)$ coordinates or equivalently by its *dual* $\eta(P)$ coordinates. We can switch between these two dual coordinates as follows:

[1] The $:=:$ symbol means it is a notational convention equality, like $\sum_{i=1}^k x_i :=: x_1 + \ldots x_k$. It differs from $a := b$ which denotes the symbol of a quantity equality by definition.

$$\eta = \eta(\theta) = \nabla F(\theta) = (\partial_i F(\theta))_i, \tag{3}$$
$$\theta = \theta(\eta) = \nabla F^*(\eta) = (\partial^i F^*(\eta))_i, \tag{4}$$

with reciprocal gradients $\nabla F^* := (\nabla F)^{-1}$. We used the notational convention $\partial^i :=: \frac{\partial}{\partial \eta_i}$ which indicates the covariant vector [18] $\eta = (\eta_i)_i$.

The metric tensor g of the dually flat structure $(\mathcal{M}, F)$ can either be expressed using the θ- or η-coordinates using the Hessians of the potential functions [53]:

$$G(\theta) = \nabla^2 F(\theta), \tag{5}$$
$$G^*(\eta) = \nabla^2 F^*(\eta). \tag{6}$$

It defines a smooth bilinear form $\langle v, v' \rangle_g$ on $\mathcal{M}$ so that for two vectors v, w of a tangent plane T_P:

$$\langle v, v' \rangle_g = \theta(v)^\top G(\theta)\theta(w), \tag{7}$$
$$= \eta(v)^\top G^*(\eta)\eta(w), \tag{8}$$

where $\theta(v) = (v^i)_i$ and $\eta(v) = (v_i)_i$ denote the contravariant coefficients and covariant coefficients of a vector v, respectively. That is, any vector $v \in T_P$ can be written either as $v = \sum_i v^i e_i$ or as $\sum_i v_i e^{*i}$, where $\{e_i\}_i$ and $\{e^{*i}\}_i$ are the dual basis [18] of the vector space structure of T_P.

Matrices $G(\theta)$ and $G^*(\eta)$ are symmetric positive definite (SPD, denoted by $G(\theta) \succ 0$ and $G^*(\eta) \succ 0$), and they satisfy the Crouzeix identity [13]:

$$G(\theta) G^*(\eta) = I, \tag{9}$$

where I stands for the $D \times D$ identity matrix. This indicates that at each tangent plane T_P, the dual coordinate systems are biorthogonal [57] (with $\{e_i\}_i$ and $\{e^{*i}\}_i$ forming a dual basis [18] of the vector space structure of T_P):

$$\langle e_i, e^{*j} \rangle = \delta_i^j, \tag{10}$$

with δ_i^j the Krönecker symbol: $\delta_i^j = 1$ if and only if (iff) $i = j$, and 0 otherwise. We have:

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij}(\theta) = \langle e_i, e_j \rangle, \tag{11}$$
$$\frac{\partial \theta^i}{\partial \eta_j} = g^{ij}(\eta) = \langle e^{*i}, e^{*j} \rangle. \tag{12}$$

The convex conjugate functions $F(\theta)$ and $F^*(\eta)$ are called *dual potential functions*, and define the global metric [53].

Table 1 Overview of the dually differential-geometric structure $(\mathcal{M}, F)$ induced by a Bregman generator F. Notice that if F and ∇F^* are available in closed-form then so are ∇F and F^*

Manifold $(\mathcal{M}, F)$	Primal structure	Dual structure
Affine coordinate system	$\theta(\cdot)$	$\eta(\cdot)$
Conversion $\theta \leftrightarrow \eta$	$\theta(\eta) = \nabla F^*(\eta)$	$\eta(\theta) = \nabla F(\theta)$
Potential function	$F(\theta) =$ $\langle \theta, \nabla F(\theta)\rangle - F^*(\nabla F(\theta))$	$F^*(\eta) =$ $\langle \eta, \nabla F^*(\eta)\rangle - F(\nabla F^*(\eta))$
Metric tensor g	$G(\theta) = \nabla^2 F(\theta)$	$G^*(\eta) = \nabla^2 F^*(\eta)$
	$g_{ij} = \partial_i \partial_j F(\theta)$	$g^{ij} = \partial^i \partial^j F^*(\eta)$
Geodesic ($\lambda \in [0, 1]$)	$\gamma(P, Q) := \{(PQ)_\lambda =$ $(1-\lambda)\theta(P) + \lambda\theta(Q)\}_\lambda$	$\gamma^*(P, Q) := \{(PQ)^*_\lambda =$ $(1-\lambda)\eta(P) + \lambda\eta(Q)\}_\lambda$

Table 1 summarizes the differential-geometric structures of dually flat spaces. Since Bregman divergences are *canonical divergences*[2] of dually flat spaces [1], the geometry of dually flat spaces is also referred to the *Bregman geometry* [15] in the literature.

Definition 1 (*Bregman generator*) A Bregman generator is a strictly convex and twice continuously differentiable real-valued function $F : \mathbb{R}^D \rightarrow \mathbb{R}$.

Let us cite the following well-known properties of Bregman generators [4]:

Property 1 *(Bregman generators are equivalent up to modulo affine terms) The Bregman generator $F_2(\theta) = F_1(\theta) + \langle a, \theta\rangle + b$ (with $a \in \mathbb{R}^D$ and $b \in \mathbb{R}$) yields the same Bregman divergence as the Bregman divergence induced by F_1, $B_{F_2}(\theta_1 : \theta_2) = B_{F_1}(\theta_1 : \theta_2)$, and therefore the same dually flat space $(\mathcal{M}, F_2) \cong (\mathcal{M}, F_1)$.*

Property 2 *(Linearity rule of Bregman generators) Let F_1, F_2 be two Bregman generators and $\lambda_1, \lambda_2 > 0$. Then $B_{\lambda_1 F_1 + \lambda_2 F_2}(\theta : \theta') = \lambda_1 B_{F_1}(\theta : \theta') + \lambda_2 B_{F_2}(\theta : \theta')$.*

In practice, the algorithmic toolbox in dually flat spaces (e.g., clustering [4], minimum enclosing balls [39], hypothesis testing [31] and Chernoff information [32], Voronoi diagrams [6, 34], proximity data-structures [45, 46], etc.) can be used whenever the dual Legendre convex conjugates F and F^* are both available in closed-form (see Type 1 of Table 4). In that case, both the primal $\gamma(P, Q) := \{(PQ)_\lambda\}_\lambda$ and dual $\gamma^*(P, Q) := \{(PQ)^*_\lambda\}_\lambda$ geodesics are available in closed form. These dual geodesics can either be expressed using the θ or η-coordinate systems as follows:

$$(PQ)_\lambda = \begin{cases} \theta((PQ)_\lambda) = \theta(P) + \lambda(\theta(Q) - \theta(P)), \\ \eta((PQ)_\lambda) = \nabla F(\theta((PQ)_\lambda)) = \nabla F(\nabla F^*(\eta(P)) + \lambda(\nabla F^*(\eta(Q)) - \nabla F^*(\eta(P)))), \end{cases} \tag{13}$$

[2] That is, we can associate to any dually flat manifold a divergence that amounts to a Bregman divergence [1].

Table 2 Some fundamental Bregman clustering algorithms [4, 22, 41] (of the Bregman algorithmic toolbox) that illustrate which closed-form are required to be run in practice

Algorithm	$F(\theta)$	$\eta(\theta)=\nabla F(\theta)$	$\theta(\eta)=\nabla F^*(\eta)$	$F^*(\eta)$
Right-sided Bregman clustering	✓	✓	×	×
Left-sided Bregman clustering	×	×	✓	✓
Symmetrized Bregman centroid	✓	✓	✓	✓
Mixed Bregman clustering	✓	✓	✓	✓
Maximum Likelihood Estimator for EFs	×	×	✓	×
Bregman soft clustering ($\equiv$ EM)	×	✓	✓	✓

$$(PQ)^*_\lambda = \begin{cases} \eta((PQ)^*_\lambda) = \eta(P) + \lambda(\eta(Q) - \eta(P)), \\ \theta((PQ)^*_\lambda) = \nabla F^*(\eta((PQ)^*_\lambda)) = \nabla F^*(\nabla F(\theta(P)) + \lambda(\nabla F(\theta(Q)) - \nabla F(\theta(P)))) \end{cases} \tag{14}$$

That is, the primal geodesic corresponds to a straight line in the primal coordinate system while the dual geodesic is a straight line in the dual coordinate system. However, in many interesting cases, the convex generator F or its dual F^* (or both) are not available in closed form or are computationally intractable, and the above Bregman toolbox cannot be used. Table 2 summarizes the closed-form formulas required to execute some fundamental clustering algorithms [4, 22, 41] in a Bregman geometry.

Let us notice that so far the points $P \in \mathcal{M}$ in the dually flat manifold have no particular meaning, and that the dually flat space structure is generic, not necessarily related to a statistical flat manifold. We shall now quickly review the dualistic structure of statistical manifolds [24].

1.2 Geometry of Statistical Manifolds

Let $I_1(x; y)$ denote a *scalar divergence*. A *statistical divergence* between two probability distributions P and Q, with Radon-Nikodym derivatives $p(x)$ and $q(x)$ with respect to (w.r.t.) a base measure μ defined on the support $\mathcal{X}$, is defined as:

$$I(P:Q) = \int_{x\in\mathcal{X}} I_1\left(p(x) : q(x)\right) \mathrm{d}\mu(x). \tag{15}$$

A statistical divergence is a measure of dissimilarity/discrimination that satisfies $I(P : Q) \geq 0$ with equality iff. $P = Q$ (a.e., reflexivity property) . For example, the Kullback–Leibler divergence is a statistical divergence:

$$\mathrm{KL}(P:Q) := \int_{x\in\mathcal{X}} \mathrm{kl}(p(x) : q(x))\mathrm{d}\mu(x), \tag{16}$$

with corresponding scalar divergence:

$$\mathrm{kl}(x : y) := x \log \frac{x}{y}. \tag{17}$$

The KL divergence between P and Q is also called the *relative entropy* [11] because it is the difference of the *cross-entropy* $h^\times(P : Q)$ between P and Q with the Shannon entropy $h(P)$ of P:

$$\begin{aligned} \mathrm{KL}(P : Q) &= h^\times(P : Q) - h(P), && (18)\\ h^\times(P : Q) &:= \int_{x\in\mathcal{X}} p(x) \log \frac{1}{q(x)} \mathrm{d}\mu(x), && (19)\\ h(P) &:= \int_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} \mathrm{d}\mu(x) = h^\times(P : P). && (20) \end{aligned}$$

Thus we distinguish a statistical divergence from a parameter divergence by stating that a statistical divergence is a separable divergence that is the definite integral on the support of a scalar divergence.

In information geometry [1, 8], we equip a probability manifold $\mathcal{M} = \{p(x;\theta) : \theta \in \Theta\}$ with a *metric tensor* g (for measuring angles between vectors and lengths of vectors in tangent planes) and a *pair of dual torsion-free connections* ∇ and ∇^* (for defining parallel transports and geodesics) that are defined by their Christoffel symbols Γ_{ijk} and Γ^*_{ijk}. These geometric structures $(\mathcal{M}, D) := (\mathcal{M}, g_D, \nabla_D, \nabla^*_D)$ can be induced by *any smooth* C^∞ divergence $D(\cdot : \cdot)$ [1, 2, 8, 17] as follows:

$$g_{ij}(x) = \left.\frac{\partial^2}{\partial x_i \partial x_j} D(x : y)\right|_{y=x}, \tag{21}$$

$$\Gamma_{ijk}(x) = -\left.\frac{\partial^3}{\partial x_i \partial x_j \partial y_k} D(x : y)\right|_{y=x}. \tag{22}$$

The *dual divergence* $D^*(p : q) := D(q : p)$ highlights the *reference duality* [57], and the dual connection ∇^* is induced by the dual divergence $D^*(\cdot : \cdot)$ (∇^* is defined by $\Gamma^*_{ijk}(x) = -\frac{\partial^3}{\partial x_i \partial x_j \partial y_k} D^*(x : y)\Big|_{y=x}$). Observe that the metric tensor is self-dual: $g^* = g$.

Let us give some examples of parametric probability families and their statistical manifolds induced by the Kullback–Leibler divergence.

1.2.1 Exponential Family Manifold (EFM)

We start by a definition:

Definition 2 (*Exponential family*) Let μ be a prescribed base measure and $t(x)$ a sufficient statistic vector. We can build a corresponding exponential family:

$$\mathcal{E}_{t,\mu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle)\}_\theta, \tag{23}$$

where $p(x;\theta) := \frac{\mathrm{d}P(\theta)}{\mathrm{d}\mu}(x)$.

The densities are normalized by the cumulant function F:

$$F(\theta) := \log\left(\int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu(x)\right), \tag{24}$$

so that:

$$p(x;\theta) = \exp(\langle t(x), \theta\rangle - F(\theta)). \tag{25}$$

The function F is a Bregman generator on the natural parameter space:

$$\Theta := \left\{\theta : \int_{x\in\mathcal{X}} \exp(\langle t(x), \theta\rangle)\mathrm{d}\mu(x) < \infty\right\}. \tag{26}$$

If we add an extra carrier term $k(x)$ and consider the measure $\nu(x) := \frac{\mu(x)}{\exp(k(x))}$, we get the generic form of an exponential family [36]:

$$\mathcal{E}_{t,k,\nu} := \{p(x;\theta) \propto \exp(\langle t(x), \theta\rangle + k(x)) : \theta \in \Theta\}. \tag{27}$$

We call the function F the *Exponential Family Bregman Generator*, or EFBG for short in the remainder.

It turns out that $(\mathcal{E}_{t,\mu}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) \cong (\mathcal{M}, F)$ (meaning the information-geometric structure of the statistical manifold is isomorphic to the information-geometry of a dually flat manifold) so that:

$$\begin{aligned} \mathrm{KL}(p(x;\theta_1) : p(x;\theta_2) &= B_F(\theta_2 : \theta_1), & (28)\\ &= B_{F^*}(\eta_1 : \eta_2), & (29) \end{aligned}$$

with $\eta = E_{p(x;\theta)}[t(x)]$ the dual parameter called the expectation parameter or moment parameter.

1.2.2 Mixture Family Manifold (MFM)

Another important type of families of probability distributions are the mixture families:

Definition 3 (*Mixture family*) Given a set of k prescribed statistical distributions $p_0(x), \ldots, p_{k-1}(x)$, all sharing the same support $\mathcal{X}$ (say, $\mathbb{R}$), a *mixture family* $\mathcal{M}$ of order $D = k - 1$ consists of all *strictly convex combinations* of these component distributions [43, 44]:

$$\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x) : \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}. \tag{30}$$

It shall be understood from the context that $\mathcal{M}$ is a shorthand for $\mathcal{M}_{p_0(x),\ldots,p_D}$.

It turns out that $(\mathcal{M}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) \cong (\mathcal{M}, G)$ so that:

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta'), \tag{31}$$

for the Bregman generator being the Shannon negative entropy (also called Shannon information):

$$G(\eta) = -h(m(x;\eta)) = \int_{x \in \mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x). \tag{32}$$

We call function G the *Mixture Family Bregman Generator*, or MFBG for short in the remainder.

For a mixture family, we prefer to use the notation η instead of θ for indexing the distribution parameters as it is customary in textbooks of information geometry [1, 8]. One reason comes from the fact that the KL divergence between two mixtures amounts to a BD on their respective parameters (Eq. 31) while the KL divergence between exponential family distributions is equivalent to a BD on the swapped order of their respective parameters (Eq. 28), see [3, 19]. Thus in order to get the same order of arguments for the KL between two exponential family distributions, we need to use the dual Bregman divergence on the dual η parameter, see Eq. 29.

1.2.3 Cauchy Family Manifold (CFM)

This example is only given to emphasize that probability families may neither be exponential nor mixture families [28].

A Cauchy distribution has probability density defined on the support $\mathcal{X} = \mathbb{R}$ by:

$$p(x;\mu,\sigma) = \frac{1}{\pi\sigma\left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)}. \tag{33}$$

The space of all Cauchy distributions:

$$\mathcal{C} = \{p(x;\mu,\sigma) : \mu \in \mathbb{R}, \sigma > 0\}. \tag{34}$$

is a location-scale family [23]. It is not an exponential family nor a mixture family.

Table 3 compares the dually flat structures of mixture families with exponential families. In information geometry, $(\mathcal{E}_{t,k,\mu}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) = (\mathcal{E}_{t,k,\mu}, g, \nabla^e, \nabla^m)$ and $(\mathcal{M}, \mathrm{KL}, \nabla_{\mathrm{KL}}, \nabla^*_{\mathrm{KL}}) = (\mathcal{M}, g, \nabla^m, \nabla^e)$ where g is the *Fisher information metric*

Table 3 Characteristics of the dually flat geometries of Exponential Families (EFs) and Mixture Families (MFs)

	Exponential Family	Mixture Family
Density	$p(x;\theta) = \exp(\langle\theta, x\rangle - F(\theta))$	$m(x;\eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x)$
		$f_i(x) = p_i(x) - p_0(x)$
Family/Manifold	$\mathcal{M} = \{p(x;\theta)\ :\ \theta \in \Theta^\circ\}$	$\mathcal{M} = \{m(x;\eta)\ :\ \eta \in H^\circ\}$
Convex function ($\equiv ax + b$)	F: cumulant	F^*: negative entropy
Dual coordinates	moment $\eta = E[t(x)]$	$\theta^i = h^\times(p_0 : m) - h^\times(p_i : m)$
Fisher Information $g = (g_{ij})_{ij}$	$g_{ij}(\theta) = \partial_i \partial_j F(\theta)$ $g = \mathrm{Var}[t(X)]$	$g_{ij}(\eta) = \int_{\mathcal{X}} \frac{f_i(x) f_j(x)}{m(x;\eta)} \mathrm{d}\mu(x)$
Christoffel symbol	$\Gamma_{ij,k} = \frac{1}{2}\partial_i \partial_j \partial_k F(\theta)$	$g_{ij}(\eta) = -\partial_i \partial_j h(\eta)$ $\Gamma_{ij,k} = -\frac{1}{2} \int_{\mathcal{X}} \frac{f_i(x) f_j(x) f_k(x)}{m^2(x;\eta)} \mathrm{d}\mu(x)$
Entropy	$-F^*(\eta)$	$-F^*(\eta)$
Kullback–Leibler divergence	$B_F(\theta_2 : \theta_1)$	$B_{F^*}(\eta_1 : \eta_2)$
	$= B_{F^*}(\eta_1 : \eta_2)$	$= B_F(\theta_2 : \theta_1)$

tensor and ∇^e and ∇^m are the exponential and mixture connections, respectively. These connections are dual to each others, see [8].

1.3 Computational Tractability of Dually Flat Statistical Manifolds

The previous section explained the dually flat structures (i.e., Bregman geometry) of the exponential family manifold and of the mixture family manifold. However these geometries may be purely theoretical as the Bregman generator F may not be available in closed form so that the Bregman toolbox cannot be used in practice. This work tackles this problem faced in exponential and mixture family manifolds by proposing the novel framework of *Monte Carlo Information Geometry* (MCIG). MCIG approximates the untractable Bregman geometry by considering the Monte Carlo stochastic integration of the definite integral-based ideal Bregman generator.

But first, let us quickly review the five types of tractability of Bregman geometry in the context of statistical manifolds by giving an illustrating family example for each type:

Type 1. F and ∇F^* are both available in closed-form, and so are ∇F and F^*. For example, this is the case of the *the Gaussian exponential family*. The normal distribution [36] has sufficient statistic vector $t(x) = (x, x^2)$ so that its log-normalizer is

$$F(\theta) = \log\left(\int_{-\infty}^{+\infty} \exp(\theta_1 x + \theta_2 x^2)\mathrm{d}x\right). \tag{35}$$

Since $\int_{-\infty}^{\infty} \exp(\theta_1 x + \theta_2 x^2) = \sqrt{\frac{\pi}{-\theta_2}} \exp(-\frac{\theta_1^2}{4\theta_2})$ for $\theta_2 < 0$, we find:

$$F(\theta) = \log\left(\int \exp(\theta_1 x + \theta_2 x^2)\mathrm{d}x\right) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{\pi}{-\theta_2}. \tag{36}$$

This is in accordance with the direct canonical decomposition [36] of the density $p(x;\theta) = \exp(\langle t(x), \theta\rangle - F(\theta))$ of the normal density $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}$ $\exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

Remark 1 When $F(\theta)$ can be expressed using the canonical decomposition of exponential families, this means that the definite integral $\log(\int \exp(\langle t(x), \theta\rangle + k(x))\mathrm{d}x)$ is available in closed form, and vice-versa.

Type 2. F is available in closed form (and so is ∇F) but ∇F^* is not available in closed form (and therefore F^* is not available too). This is for example the *Beta exponential family*. A Beta distribution $\mathrm{Be}(\alpha, \beta)$ has density on support $x \in (0, 1)$:

$$p(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{37}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and $(\alpha > 0, \beta > 0)$ are the shape parameters. The Beta family of distributions is an exponential family with $\theta = (\alpha, \beta)$, $t(x) = (\log(x), \log(1-x))$, $k(x) = -\log(x) - \log(1-x)$ and $F(\theta) = \log B(\theta_1, \theta_2) = \log\Gamma(\theta_1) + \log\Gamma(\theta_2) - \log\Gamma(\theta_1+\theta_2)$. Note that we could also have chosen $\theta = (\alpha-1, \beta-1)$ and $k(x) = 0$. Thus $\nabla F(\theta) = (\psi(\theta_1) - \psi(\theta_1+\theta_2), \psi(\theta_2) - \psi(\theta_1+\theta_2))$ where $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function. Inverting the gradient $\nabla F(\theta) = \eta$ to get $\eta = \nabla F^*(\theta)$ is not available in closed-form.[3]

Type 3. This type of families has discrete support $\mathcal{X}$ and thus requires an exponential time to compute the log-normalizer. For example, consider the Ising models [5, 9, 21]: Let $G = (V, E)$ be an undirected graph of $|V|$ nodes and $|E|$ edges. Each node $v \in V$ is associated with a binary random variable $x_v \in \{0, 1\}$. The probability of an Ising model is defined as follows:

$$p(x;\theta) = \exp\left(\sum_{v\in V}\theta_v x_v + \sum_{(v,w)\in E}\theta_{vw}x_v x_w - F(\theta)\right). \tag{38}$$

[3] To see this, consider the digamma difference property: $f_\Delta(\theta) = \psi(\theta) - \psi(\theta+\Delta) = -\sum_{i=0}^{\Delta-1}\frac{1}{x+i}$ for $\Delta \in \mathbb{N}$. We cannot invert $f_\Delta(\theta)$ since it involves solving the root of a high-degree polynomial.

The vector $t(x) = (\ldots, x_v, \ldots, x_{vw}, \ldots)$ of sufficient statistics is D-dimensional with $D = |V| + |E|$. The log-normalizer is:

$$F(\theta) = \log\left(\sum_{(x_v)_v \in \{0,1\}^{|V|}} \left(\exp \sum_{v \in V} \theta_v x_v + \sum_{(v,w) \in E} \theta_{vw} x_v x_w\right)\right). \tag{39}$$

It requires to sum up $2^{|V|}$ terms.

Type 4. This type of families has a Bregman generator which is not available in closed-form. For example, this is the case of the *Polynomial Exponential Family* [10, 42] (PEF) which are helpful to model a multimodal distribution (instead of using a statistical mixture). Consider the following vector of sufficient statistics $t(x) = (x, x^2, \ldots, x^D)$ for defining an exponential family:

$$\mathcal{E}_{t(x),\mu} = \left\{ p(x;\theta) = \exp\left(\sum_{i=1}^{D} \theta_i x^i - F(\theta)\right) : \theta \in \Theta \right\}. \tag{40}$$

(Beware that here, $x^i = \mathrm{Pow}(x, i) := \underbrace{x \times \cdots \times x}_{i \text{ times}}$ denotes the ith power of x (monomial of degree i), and not a contravariant coefficient of a vector x.)

In general, the definite integral of the cumulant function (the Exponential Family Bregman Generator, EFBG) of Eq. 24 does not admit a closed form, but is analytic. For example, choosing $t(x) = x^8$, we have:

$$F(\theta) = \log \int_{-\infty}^{\infty} \exp(\theta x^8) \mathrm{d}x = \log 2 + \log \Gamma(9/8) - \frac{1}{8} \log(-\theta), \tag{41}$$

for $\theta < 0$. But $\int_{-\infty}^{\infty} \exp(-x^8 - x^4 - x^2)\mathrm{d}x \simeq 1.295$ is not available in closed form.

Type 5. This last category is even more challenging from a computational point of view because of log-sum terms. For example, the *mixture family*. As already stated, the negative Shannon entropy (i.e., the Mixture Family Bregman Generator, MFBG) is not available in closed form for statistical mixture models [43]. It is in fact even worse, as the Shannon entropy of mixtures is not analytic [56].

This paper considers approximating the computationally untractable generators of statistical exponential/mixture families (type 4 and type 5) using stochastic Monte Carlo approximations.

In [12], Critchley et al. take a different approach of the computational tractability by discretizing the support $\mathcal{X}$ into a finite number of bins, and considering the corresponding discrete distribution. However, this approach does not scale well with the dimension of the support. Our Monte Carlo Information Geometry scales to arbitrary high dimensions because it relies on the fact that the Monte Carlo stochastic estimator is independent of the dimension [52].

1.4 Paper Organization

In Sect. 2, we consider the MCIG structure of mixture families: Namely, Sect. 2.1 considers first the uni-order families to illustrate the basic principle. It is followed by the general case in Sect. 2.2. Similarly, Sect. 3 handles the exponential family case by first explaining the uni-order case in Sect. 3.1 before tackling the general case in Sect. 3.2. Sect. 4 presents an application of the computationally-friendly MCIG structures for clustering distributions in dually flat statistical mixture manifolds. In Sect. 5, we show how to construct non-flat MCIG structures of a parametric family of distributions given by a statistical separable divergence. Finally, we conclude and discuss several perspectives in Sect. 6.

2 Monte Carlo Information Geometry of Mixture Families

Recall the definition of a statistical mixture model (Definition 3): Given a set of k prescribed statistical distributions $p_0(x), \dots, p_{k-1}(x)$, all sharing the same support $\mathcal{X}$, a *mixture family* $\mathcal{M}$ of order $D = k - 1$ consists in all *strictly convex combinations* of the $p_i(x)$'s [43]:

$$\mathcal{M} := \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x) : \eta_i > 0, \sum_{i=1}^{k-1} \eta_i < 1 \right\}. \tag{42}$$

The differential-geometric structure of $\mathcal{M}$ is well studied in information geometry [1, 8] (although much less than for the exponential families), where it is known that:

$$\mathrm{KL}(m(x;\eta) : m(x;\eta')) = B_G(\eta : \eta'), \tag{43}$$

for the Bregman generator being the Shannon negative entropy (MFBG):

$$G(\eta) = -h(m(x;\eta)) = \int_{x \in \mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x). \tag{44}$$

The negative entropy $G(\eta) = \int_{x \in \mathcal{X}} m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x)$ is a smooth and strictly convex function which induces a dually flat structure with Legendre convex conjugate:

$$F(\theta) = G^*(\theta) = -\int_{x \in \mathcal{X}} p_0(x) \log m(x;\eta) \mathrm{d}\mu(x) = h^\times(p_0(x) : m(x;\eta)), \tag{45}$$

interpretable as the cross-entropy of $p_0(x)$ with the mixture $m(x;\eta)$ [43].

Notice that the component distributions may be heterogeneous like $p_0(x)$ being a fixed Cauchy distribution, $p_1(x)$ being a fixed Gaussian distribution, $p_2(x)$ a Laplace distribution, etc. Except for the case of the finite categorical distributions (that are both interpretable as either a mixture family and an exponential family, see [1]), $G(\eta)$ provably does not admit a closed form [56] (i.e., meaning that the definite integral of Eq. 32 does not admit a simple formula using common standard functions). Thus the dually-flat geometry $(\mathcal{M}, G)$ is a theoretical construction which cannot be explicitly used by Bregman algorithms.

One way to tackle the lack of closed form in Eq. 32, is to approximate the definite integrals whenever they are used by using Monte Carlo stochastic integration. However, this is computationally very expensive, and, even worse, it cannot guarantee that the overall computation is consistent.

Let us briefly explain the meaning of *consistency*: We can estimate the KL between two distributions p and q by drawing m variates $x_1, \ldots, x_m \sim p(x)$, and use the following MC KL estimator:

$$\widehat{\mathrm{KL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m} \log \frac{p(x_i)}{q(x_i)}. \tag{46}$$

Now, suppose we have $\mathrm{KL}(p:q) \leq \mathrm{KL}(q:r)$, then their MC estimates may not satisfy $\widehat{\mathrm{KL}}_m(p:q) < \widehat{\mathrm{KL}}_m(q:r)$ (since each time we evaluate a $\widehat{\mathrm{KL}}_m$ we draw different samples). Thus when running a KL/Bregman algorithm, the more MC stochastic approximations of integrals are performed in the algorithm, the less likely is the output consistent. For example, consider computing the Bregman Voronoi diagram [34] of a set of n mixtures belonging to a mixture family manifold (say, with $D = 2$) using the algorithm explained in [34]: Since we use for each BD calculation or predicate evaluation relying on F or F^* stochastic Monte Carlo integral approximations, this MC algorithm may likely not deliver a proper combinatorial structure of the Voronoi diagram: The Voronoi structure is likely to be inconsistent.

Let us now show how Monte Carlo Information Geometry (MCIG) approximates this computationally untractable $(\mathcal{M}, G)$ geometric structure by defining a consistent and computationally-friendly dually-flat information geometry $(\mathcal{M}, \tilde{G}_{\mathcal{S}})$ for a finite number m of identically and independently distributed (iid) random samples $\mathcal{S}$.

2.1 MCIG of Order-1 Mixture Family

In order to highlight the principle of MCIGs, let us first consider a mixture family of order $D = 1$. That is, we consider a set of mixtures of $k = 2$ components with density:

$$m(x;\eta) = \eta p_1(x) + (1-\eta)p_0(x) = p_0(x) + \eta(p_1(x) - p_0(x)), \tag{47}$$

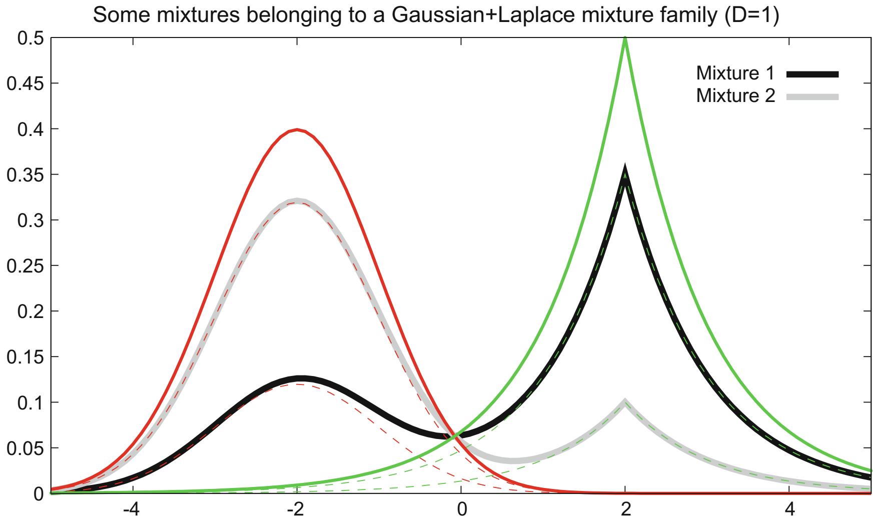


Fig. 1 Example of a mixture family of order $D = 1$ ($k = 2$): $p_0(x) \sim \text{Gaussian}(-2, 1)$ (red) and $p_1(x) \sim \text{Laplace}(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = 0.7$ and $m_2(x) = m(x; \eta_2)$ (grey) with $\eta_2 = 0.2$. Weighted component distributions are displayed in dashed

with parameter η ranging in $(0, 1)$. The two prescribed component densities $p_0(x)$ and $p_1(x)$ (with respect to a base measure μ, say the Lebesgue measure) are defined on a common support $\mathcal{X}$. Densities $p_0(x)$ and $p_1(x)$ are assumed to be linearly independent [8].

Figure 1 displays an example of uni-order mixture family with heterogeneous components: $p_0(x)$ is chosen as a Gaussian distribution while $p_1(x)$ is taken as a Laplace distribution. A mixture $m(x; \eta)$ of $\mathcal{M}$ is visualized as a point P (here, one-dimensional) with $\eta(P) = \eta$.

Let $\mathcal{S} = \{x_1, \ldots, x_m\}$ denote a iid sample from a fixed *proposal distribution* $q(x)$ (with $q(x) > 0$ for $x \in \mathcal{X}$, and $q(x)$ independent of η). We approximate the Bregman generator $G(\eta)$ using Monte Carlo stochastic integration with importance sampling as follows:

$$G(\eta) \simeq \tilde{G}_{\mathcal{S}}(\eta) := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{q(x_i)} m(x_i; \eta) \log m(x_i; \eta). \tag{48}$$

Let us prove that the Monte Carlo function $\tilde{G}_{\mathcal{S}}(\eta)$ is a proper Bregman generator. That is, that $\tilde{G}_{\mathcal{S}}(\eta)$ is strictly convex and twice continuously differentiable (Definition 1).

Write for short $m_x(\eta) := m(x;\eta)$ so that $G(\eta) = \int_{x\in\mathcal{X}} m_x(\eta)\log m_x(\eta)\mathrm{d}\mu(x)$ is approximated by $\frac{1}{m}\sum_{i=1}^m \frac{1}{q(x_i)} m_{x_i}(\eta)\log m_{x_i}(\eta)$. Since $\frac{1}{m}\frac{1}{q(x_i)} > 0$, it suffices to prove that the function $g_x(\eta) = m_x(\eta)\log m_x(\eta)$ is strictly convex wrt parameter η. Then we shall conclude that $\tilde{G}_{\mathcal{S}}(\eta)$ is strictly convex because it is a finite positively weighted sum of strictly convex functions.

Let us write the first and second derivatives of $g_x(\eta)$ as follows:

$$g_x(\eta)' = m_x(\eta)'(\log m_x(\eta) + 1), \tag{49}$$

$$g_x(\eta)'' = m_x(\eta)''(\log m_x(\eta) + 1) + \frac{(m_x(\eta)')^2}{m_x(\eta)}. \tag{50}$$

Since $m_x'(\eta) = p_1(x) - p_0(x)$ and $m_x''(\eta) = 0$, we get:

$$g_x(\eta)'' = \frac{(p_1(x) - p_0(x))^2}{m_x(\eta)}. \tag{51}$$

Thus it follows that:

$$\tilde{G}_{\mathcal{S}}''(\eta) = \frac{1}{m}\sum_{i=1}^m \frac{1}{q(x_i)}\frac{(p_1(x_i) - p_0(x_i))^2}{m(x_i;\eta)} \geq 0. \tag{52}$$

It is strictly convex provided that there exists at least one x_i such that $p_1(x_i) \neq p_0(x_i)$.

Let $\mathcal{D} \subset \mathcal{X}$ denote the degenerate set $\mathcal{D} = \{x \in \mathcal{X} : p_1(x) = p_0(x)\}$. For example, if $p_0(x)$ and $p_1(x)$ are two distinct univariate normal distributions, then $|\mathcal{D}| = 2$ (roots of a quadratic equation), and

$$\mu_q(\mathcal{D}) := \int_{x\in\mathcal{X}} 1_{[p_0(x)=p_1(x)]} q(x)\mathrm{d}\mu(x) = 0. \tag{53}$$

Assumption 1 *(AMF1D)* We assume that $p_0(x)$ and $p_1(x)$ are linearly independent (non-singular statistical model, see [8]), and that $\mu_q(\mathcal{D}) = 0$.

Lemma 1 (Monte Carlo Mixture Family Function is a Bregman generator) *The Monte Carlo Mixture Family Function (MCMFF) $\tilde{F}_{\mathcal{S}}(\theta)$ is a Bregman generator almost surely.*

Proof When there exists a sample $x \in \mathcal{S}$ with two distinct densities $p_0(x)$ and $p_1(x)$, we have $(p_1(x_i) - p_0(x_i))^2 > 0$ and therefore $\tilde{G}_{\mathcal{S}}''(\eta) > 0$. The probability to get a degenerate sample is almost zero.

To recap, the MCMFF of the MCIG of uni-order family has the following characteristics:

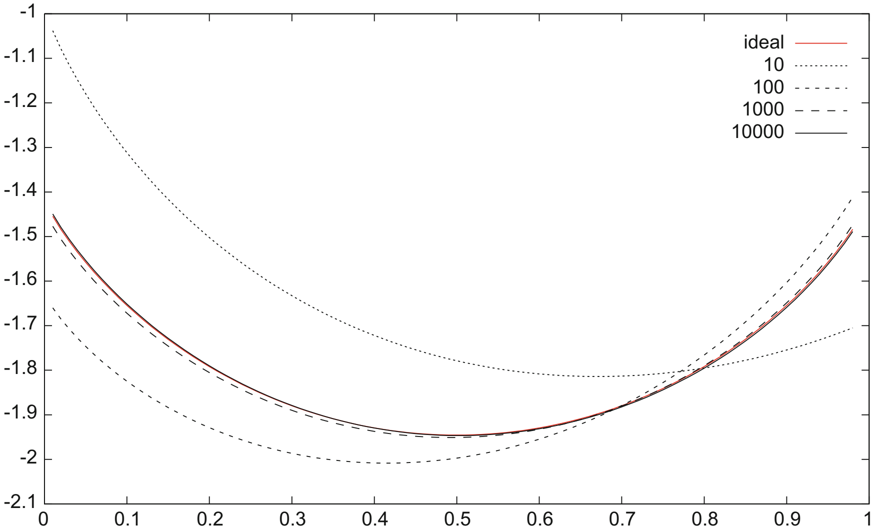


Fig. 2 A series $G_{\mathcal{S}}(\eta)$ of Bregman Monte Carlo Mixture Family generators (for $m = |\mathcal{S}| \in \{10, 100, 1000, 10000\}$) approximating the untractable ideal negentropy generator $G(\eta) = -h(m(x;\eta))$ (red) of a mixture family with prescribed Gaussian distributions $m(x;\eta) = (1-\eta)p(x;0,3) + \eta p(x;2,1)$ for the proposal distribution $q(x) = m(x;\frac{1}{2})$

Monte Carlo Mixture Family Generator 1D:

$$\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{q(x_i)} m(x_i;\eta)\log m(x_i;\eta), \tag{54}$$

$$\tilde{G}'_{\mathcal{S}}(\eta) = \theta = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{q(x_i)}(p_1(x_i) - p_0(x_i))(1 + \log m(x_i;\eta)), \tag{55}$$

$$\tilde{G}''_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{q(x_i)} \frac{(p_1(x_i) - p_0(x_i))^2}{m(x_i;\eta)}. \tag{56}$$

Note that $(G^*)'$ and G^* may be calculated numerically but not in closed-form. We may also MC approximate ∇G^* since $\theta = (h^\times(p_0 : m) - h^\times(p_i : m))_i$.

Thus we change from type 5 to type 2 the computational tractability of mixtures by adopting the MCIG approximation.

Figure 2 displays a series of Bregman mixture family MC generators for a mixture family for different values of $|\mathcal{S}| = m$.

As we increase the sample size of $\mathcal{S}$, the MCMFF Bregman generator tends to the ideal mixture family Bregman generator.

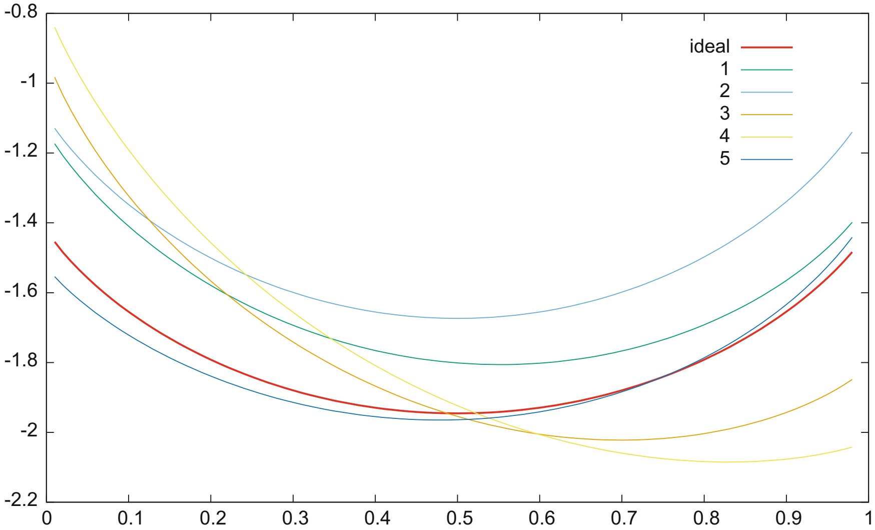


Fig. 3 The Monte Carlo Mixture Family Generator $\hat{G}_{10}$ (MCMFG) considered as a random variable: Here, we show five realizations (i.e., $\mathcal{S}_1, \ldots, \mathcal{S}_5$) of the randomized generator for $m = 5$. The ideal generator is plot in thick red

Theorem 1 (Consistency of MCIG) *Almost surely,* $\lim_{m\to\infty}(\mathcal{M}, \tilde{G}_{\mathcal{S}}) = (\mathcal{M}, G)$ *when* $\mu_q(\mathcal{D}) = 0$.

Proof It suffices to prove that $\lim_{m\to\infty} \tilde{G}_{\mathcal{S}}(\eta) = G(\eta)$. The general theory of Monte Carlo stochastic integration yields a consistent estimator provided that the following variance is bounded

$$\mathrm{Var}_q\left[\frac{m(x;\eta)\log m(x;\eta)}{q(x)}\right] < \infty. \tag{57}$$

For example, when $m(x;\eta)$ is a mixture of prescribed isotropic gaussians (say, from a KDE), and $q(x)$ is also an isotropic Gaussian, the variance is bounded. Note that q is the proposal density wrt the base measure μ.

In practice, the proposal distribution $q(x)$ can be chosen as the uniform mixture of the fixed component distributions:

$$q(x) = \frac{1}{m}\sum_{i=0}^{D} p_i(x). \tag{58}$$

Notice that the Monte Carlo Mixture Family Function is a random variable (r.v. for short) estimator itself by considering a vector of iid variables instead of a sample variate: $\hat{G}_m(\eta)$. Figure 3 displays five realizations of the random variable $\hat{G}_m(\eta)$ for $m = 10$.

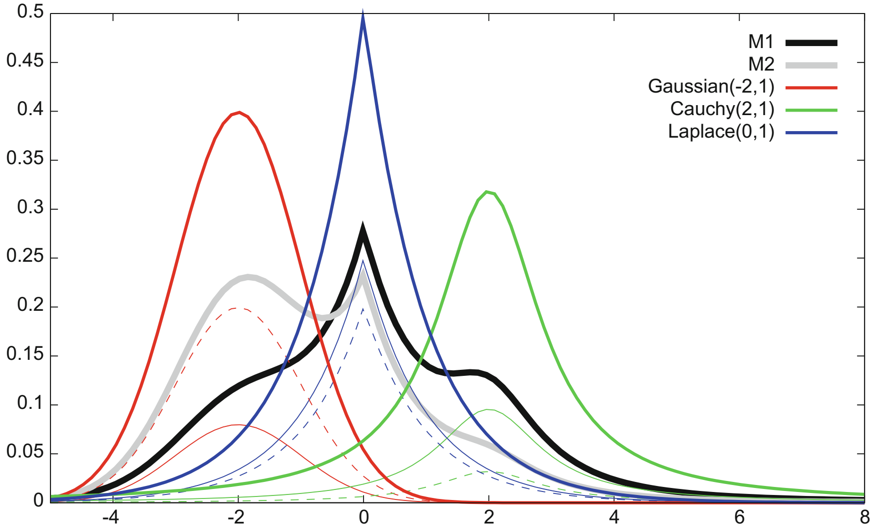


Fig. 4 Example of a mixture family of order $D = 2$ ($k = 3$): $p_0(x) \sim$ Gaussian$(-2, 1)$ (red), $p_1(x) \sim$ Laplace$(0, 1)$ (blue) and $p_2(x) \sim$ Cauchy$(2, 1)$ (green). The two mixtures are $m_1(x) = m(x; \eta_1)$ (black) with $\eta_1 = (0.3, 0.5)$ and $m_2(x) = m(x; \eta)$ (gray) with $\eta = (0.1, 0.4)$

2.2 General D-Order Mixture Case

Here, we consider statistical mixtures with $k = D + 1 > 2$ prescribed distributions $p_0(x), \ldots, p_D(x)$. The component distributions are linearly independent so that they define a non-singular statistical model [8].

We further strengthen conditions on the prescribed distributions as follows:

Assumption 2 *(AMF)* We assume that the linearly independent prescribed distributions further satisfy:

$$\sup_{B \in \mathcal{B}} \left\{ \mu_q(B) : \exists \lambda \neq (0), \sum_{i \neq j} \lambda_i \left(p_i|_B - p_j\big|_B \right) = 0 \right\} = 0, \quad \forall j, \tag{59}$$

where the supremum is over all subsets B of the σ-algebra $\mathcal{B}$ of the probability space with support $\mathcal{X}$ and measure μ, with $p_i|_B$ denoting the restriction of p_i to subset B. In other words, we impose that the components $(p_i)_i$ still constitute an affinely independent family when restricted to any subset of positive measure.

For example, Figure 4 displays two mixture distributions belonging to a 2D mixture family with Gaussian, Laplace and Cauchy component distributions.

Recall that the mixture family Monte Carlo generator is:

$$\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}m(x_i;\eta)\log m(x_i;\eta). \tag{60}$$

In order to prove that G is strictly convex, we shall prove that $\nabla^2\tilde{G}_{\mathcal{S}}(\eta) \succ 0$ almost surely. It suffices to consider the basic Hessian matrix $\nabla^2 g_x = (\partial^i\partial^j g_x(\eta))_{ij}$ of $g_x(\eta) = m_x(\eta)\log m_x(\eta)$. We have the partial first derivatives:

$$\partial^i g_x(\eta) = (p_i(x) - p_0(x))(1 + \log m(x;\eta)), \tag{61}$$

and the partial second derivatives:

$$\partial^i\partial^j g_x(\eta) = \frac{(p_i(x) - p_0(x))(p_j(x) - p_0(x))}{m(x;\eta)}, \tag{62}$$

so that

$$\partial^i\partial^j\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{q(x_l)}\frac{(p_i(x_l) - p_0(x_l))(p_j(x_l) - p_0(x_l))}{m(x_l;\eta)}. \tag{63}$$

Theorem 2 (Monte Carlo Mixture Family Function is a Bregman generator) *The Monte Carlo multivariate function $\tilde{G}_{\mathcal{S}}(\eta)$ is always convex and twice continuously differentiable, and strictly convex almost surely.*

Proof Consider the D-dimensional vector:

$$v_l = \begin{bmatrix} \frac{p_1(x_l)-p_0(x_l)}{\sqrt{q(x_l)m(x_l;\eta)}} \\ \vdots \\ \frac{p_D(x_l)-p_0(x_l)}{\sqrt{q(x_l)m(x_l;\eta)}} \end{bmatrix}. \tag{64}$$

Then we rewrite the Monte Carlo generator $\tilde{G}_{\mathcal{S}}(\eta)$ as:

$$\partial^i\partial^j\tilde{G}_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{l=1}^{m} v_l v_l^\top. \tag{65}$$

Since $v_l v_l^\top$ is always a symmetric positive semidefinite matrix of rank one, we conclude that $\tilde{G}_{\mathcal{S}}(\eta)$ is a symmetric positive semidefinite matrix when $m < D$ (rank deficient) and a symmetric positive definite matrix (full rank) almost surely when $m \geq D$.

3 Monte Carlo Information Geometry of Exponential Families

We follow the same outline as for mixture familes: Sect. 3.1 first describes the univariate case. It is then followed by the general multivariate case in Sect. 3.1.

3.1 MCIG of Order-1 Exponential Family

We consider the order-1 exponential family of parametric densities with respect to a base measure μ:

$$\mathcal{E} := \{p(x;\theta) = \exp(t(x)\theta - F(\theta) + k(x)) : \theta \in \Theta\}, \tag{66}$$

where Θ is the natural parameter space, such that the log-normalizer/cumulant function [1] is

$$F(\theta) = \log\left(\int \exp(t(x)\theta + k(x))\mathrm{d}\mu(x)\right). \tag{67}$$

The sufficient statistic function $t(x)$ and 1 are linearly independent [8].

We perform Monte Carlo stochastic integration by sampling a set $\mathcal{S} = \{x_1, \ldots, x_m\}$ of m iid variates from a proposal distribution $q(x)$ to get:

$$F(\theta) \simeq \tilde{F}_{\mathcal{S}}^{\dagger}(\theta) := \log\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\exp(t(x_i)\theta + k(x_i))\right). \tag{68}$$

Without loss of generality, assume that x_1 is the element that minimizes the sufficient statistic $t(x)$ among the elements of $\mathcal{S}$, so that $a_i = t(x_i) - t(x_1) \geq 0$ for all $x_i \in \mathcal{S}$.

Let us factorize $\frac{1}{q(x_1)}\exp(t(x_1)\theta + k(x_1))$ in Eq. 68 and remove an affine term from the generator $\tilde{F}_{\mathcal{S}}(\theta)$ to get the equivalent generator (see Property 1):

$$\begin{aligned}
\tilde{F}_{\mathcal{S}}^{\dagger}(\theta) &\equiv \tilde{F}_{\mathcal{S}}(\theta), && (69)\\
\tilde{F}_{\mathcal{S}}(\theta) &= \log\left(1 + \sum_{i=2}^{m}\exp((t(x_i) - t(x_1))\theta + k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1))\right), && (70)\\
&= \log\left(1 + \sum_{i=2}^{m}\exp(a_i\theta + b_i)\right), && (71)\\
&:= \mathrm{lse}_0^{+}(a_2\theta + b_2, \ldots, a_m\theta + b_m), && (72)
\end{aligned}$$

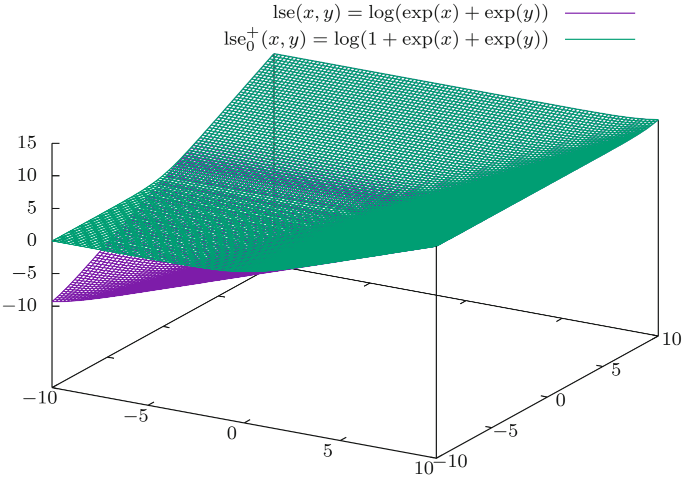


Fig. 5 Graph plots of the lse and lse_0^+ functions: The lse function (violet) is only convex while the lse_0^+ function (green) is always guaranteed to be strictly convex

with $a_2, \ldots, a_m > 0$ and $b_i = k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1)$. Function $\mathrm{lse}_0^+(x_1, \ldots, x_m) = \mathrm{lse}(0, x_1, \ldots, x_m)$ is the log-sum-exp function [20, 47] $\mathrm{lse}(x_1, \ldots, x_m) = \log \sum_{i=1}^n \exp(x_i)$ with an additional argument set to zero.

Let us notice that the lse_0^+ function is always *strictly convex* while the lse function is only convex[4] [7], p. 74. Figure 5 displays the graph plots of the lse and lse_0^+ functions. Let us clarify this point with a usual exponential family: The binomial family. The binomial distribution is a categorical distribution with $D = 1$ (and 2 bins). We have $F(\theta) = \log(1 + \exp(\theta)) = \mathrm{lse}(0, \theta) := \mathrm{lse}_0^+(\theta)$. We check the strict convexity of $F(\theta)$: $F'(\theta) = \frac{e^\theta}{1+e^\theta}$ and $F''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} > 0$.

We write for short $\mathrm{lse}_0^+(x) = \mathrm{lse}_0^+(x_1, \ldots, x_d)$ for a d-dimensional vector x.

Theorem 3 (lse_0^+ is a Bregman generator) *Multivariate function* $\mathrm{lse}_0^+(x)$ *is a Bregman generator.*

Proof is deferred to Appendix 7.

Lemma 2 (Univariate Monte Carlo Exponential Family Function is a Bregman generator) *Almost surely, the univariate function $\tilde{F}_{\mathcal{S}}(\theta)$ is a Bregman generator.*

Proof The first derivative is:

$$\eta = \tilde{F}'_{\mathcal{S}}(\theta) = \frac{\sum_{i=2}^m a_i \exp(a_i\theta + b_i)}{1 + \sum_{i=2}^m \exp(a_i\theta + b_i)} \geq 0, \tag{73}$$

[4] Function lse can be interpreted as a vector function, and is C^2, convex but not strictly convex on $\mathbb{R}^m$. For example, lse is affine on lines since $\mathrm{lse}(x + \lambda 1) = \mathrm{lse}(x) + \lambda$ (or equivalently $\mathrm{lse}(x_1, \ldots, x_m) = \lambda + \mathrm{lse}(x_1 - \lambda, \ldots, x_m - \lambda)$). It is affine only on lines passing through the origin.

and is strictly greater than 0 when there exists at least two elements with distinct sufficient statistics (i.e., $t(x_i) \neq t(x_j)$) so that at least one $a_i > 0$.

The second derivative is:

$$\tilde{F}''_{\mathcal{S}}(\theta) = \frac{\left(\sum_{i=2}^m a_i^2 \exp(a_i\theta + b_i)\right)\left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i)\right) - \left(\sum_{i=2}^m a_i \exp(a_i\theta + b_i)\right)^2}{(1 + \sum_{i=2}^m \exp(a_i\theta + b_i))^2} =: \frac{\text{Num}}{\text{Den}} \tag{74}$$

For each value of $\theta \in \Theta$, we shall prove that $\tilde{F}''_{\mathcal{S}}(\theta) > 0$. Let $c_i = c_i(\theta) = \exp(a_i\theta + b_i) > 0$ for short (θ being fixed, we omit it in the c_i notation in the calculus derivation). Consider the numerator Num since the denominator Den is a non-zero square, hence strictly positive. We have:

$$\text{Num} > \left(\sum_{i=2}^m a_i^2 c_i\right)\left(\sum_{i=2}^m c_i\right) - \left(\sum_{i=2}^m a_i c_i\right)^2, \tag{75}$$

$$\text{Num} > \sum_{ij} a_i^2 c_i c_j - \sum_i a_i^2 c_i^2 - 2\sum_{i<j} a_i a_j c_i c_j, \tag{76}$$

$$\text{Num} > \sum_{i=j} a_i^2 c_i^2 + \sum_{i\neq j} a_i^2 c_i c_j - \sum_i a_i^2 c_i^2 - 2\sum_{i<j} a_i a_j c_i c_j, \tag{77}$$

$$\text{Num} > \sum_{i<j} a_i^2 c_i c_j + \sum_{i>j} a_i^2 c_i c_j - 2\sum_{i<j} a_i a_j c_i c_j, \tag{78}$$

$$\text{Num} > \sum_{i<j} a_i^2 c_i c_j + \sum_{i<j} a_j^2 c_i c_j - 2\sum_{i<j} a_i a_j c_i c_j, \tag{79}$$

$$\text{Num} > \sum_{i<j} (a_i^2 + a_j^2 - 2a_i a_j) c_i c_j, \tag{80}$$

$$\text{Num} > \sum_{i<j} (a_i - a_j)^2 c_i c_j > 0. \tag{81}$$

Therefore the numerator is strictly positive if at least two a_i's are distinct.

Thus we add the following assumption:

Assumption 3 *(AEF1D)* For all $y \in \text{dom}(t)$, $E_q[1_{t(x)=y}] = 0$.

To recap, the MCEFF of the MCIG of uni-order family has the following characteristics:

Monte Carlo Mixture Family Generator 1D:

$$\tilde{F}_{\mathcal{S}}(\theta) = \mathrm{lse}_0^+(a_2\theta + b_2, \ldots, a_m\theta + b_m), \tag{82}$$

$$a_i = t(x_i) - t(x_1), \tag{83}$$

$$b_i = k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1), \tag{84}$$

$$\tilde{F}'_{\mathcal{S}}(\theta) = \frac{\sum_{i=2}^m a_i \exp(a_i\theta + b_i)}{1 + \sum_{i=2}^m \exp(a_i\theta + b_i)} =: \eta, \tag{85}$$

$$\tilde{F}''_{\mathcal{S}}(\theta) = \frac{\left(\sum_{i=2}^m a_i^2 \exp(a_i\theta + b_i)\right)\left(1 + \sum_{i=2}^m \exp(a_i\theta + b_i)\right) - \left(\sum_{i=2}^m a_i \exp(a_i\theta + b_i)\right)^2}{(1 + \sum_{i=2}^m \exp(a_i\theta + b_i))^2} \tag{86}$$

3.2 *The general D-Order Case*

The difference of sufficient statistics $a_i = t(x_i) - t(x_1)$ is now a vector of dimension D:

$$a_i = \begin{bmatrix} a_i^1 \\ \vdots \\ a_i^D \end{bmatrix}. \tag{87}$$

We replace the scalar multiplication $a_i\theta$ by an inner product $\langle a_i, \theta\rangle$ in Eq. 72, and let $c_i(\theta) = \exp(\langle a_i, \theta\rangle + b_i)$ with $b_i = k(x_i) - k(x_1) - \log q(x_i) + \log q(x_1)$. Then the Monte Carlo Exponential Family Function (MCEFF) writes concisely as:

$$\tilde{F}_{\mathcal{S}}(\theta) = \log\left(1 + \sum_{l=2}^m c_l(\theta)\right), \tag{88}$$

$$:= \mathrm{lse}_0^+(c_2(\theta), \ldots, c_m(\theta)), \tag{89}$$

Theorem 4 (Monte Carlo Exponential Family Function is a Bregman Generator) *Almost surely, the function $\tilde{F}_{\mathcal{S}}(\theta)$ is a proper Bregman generator.*

Proof We have the gradient of first-order partial derivatives:

$$\eta_i = \partial_i \tilde{F}_{\mathcal{S}}(\theta) = \frac{\sum_{l=2}^m a_l^i c_l(\theta)}{1 + \sum_{l=2}^m c_l(\theta)}, \tag{90}$$

and the Hessian matrix of second-order partial derivatives:

$$\partial_i \partial_j \tilde{F}_{\mathcal{S}}(\theta) = \frac{(\sum_{l=2}^m a_l^i a_l^j c_l(\theta))(1 + \sum_{l=2}^m c_l(\theta)) - (\sum_{l=2}^m a_l^i c_l(\theta))(\sum_{l=2}^m a_l^j c_l(\theta))}{(1 + \sum_{l=2}^m c_l(\theta))^2} =: \frac{\text{Num}}{\text{Den}}. \tag{91}$$

Let us prove that the Hessian matrix $\nabla^2 \tilde{F}_{\mathcal{S}}(\theta) = (\partial_i \partial_j \tilde{F}_{\mathcal{S}}(\theta))_{ij}$ is always symmetric positive semi-definite, and symmetric positive definite almost surely.

Indeed, we have:

$$\text{Num} = \underbrace{\sum_k a_k^i a_k^j c_k}_{:=D} + \underbrace{\sum_{k,l} a_k^i a_k^j c_k c_l - \sum_{k,l} a_k^i c_k a_l^j c_l}_{:=E}. \tag{92}$$

Let us rewrite D as $D = CA^\top A$ with $C = \operatorname{diag}(c_1, \ldots, c_D)$. It follows that matrix D is symmetric positive definite. Let us prove that matrix E is also SPD:

$$\begin{aligned} E &\overset{\star}{=} \sum_{k<l} a_k^i a_k^j c_k c_l + \sum_{l<k} a_k^i z_k^j c_k c_l - \sum_{k<l} a_k^i a_l^j c_k c_l - \sum_{l<k} a_k^i a_l^j c_k c_l, && (93)\\ &\overset{\star\star}{=} \sum_{k<l} \left(a_k^i a_k^j + a_l^i a_l^j - a_k^i a_l^j - a_l^i a_k^j \right) c_k c_l, && (94)\\ &= \sum_{k<l} (a_k^i - a_l^i)(a_k^j - a_l^j) c_k c_l. && (95) \end{aligned}$$

$\star$: The terms $l = k$ vanish
$\star\star$: After a change of variable $l \leftrightarrow k$ in the second and fourth sums of Eq. 93.

Thus Eq. 95 can be rewritten as $(a_k - a_l)(a_k - a_l)^\top c_k c_l$ where $a_k = \begin{bmatrix} a_k^1 \\ \vdots \\ a_k^D \end{bmatrix}$. It follows that E is a positively weighted sum of rank-1 symmetric positive semi-definite matrices, and is therefore symmetric positive semi-definite.

We want $y^T E y > 0$ for all $y \neq 0 \in \mathbb{R}^D$. Suppose that there exists $y \neq 0 \in \mathbb{R}^D$ such that $y^T E y = 0$. Noting that $a_k^i - a_l^i = t_i(x_k) - t_i(x_l)$, we can write this as

$$\sum_{k<l} \left(\sum_i y_i c_i (t_i(x_k) - t_i(x_l)) \sum_j y_j c_j (t_j(x_k) - t_j(x_l)) \right) = 0, \tag{96}$$

which implies

$$\sum_i y_i c_i \, (t_i(x_k) - t_i(x_l)) \sum_j y_j c_j \, \big(t_j(x_k) - t_j(x_l)\big) = 0, \quad \forall k < l, \tag{97}$$

since each of these terms is non negative. In particular, we have the existence of a $y \neq 0 \in \mathbb{R}^D$ such that

$$\sum_i y_i t_i(x_k) = \sum_i y_i t_i(x_l), \quad \forall y \neq 0, \quad \forall k < l. \tag{98}$$

To get almost surely a Monte Carlo Bregman generator, we introduce the following assumption:

Assumption 4 (*AEF*) The sufficient statistics (t_i) verify that for all $\lambda \neq 0$ and all $y \in dom(\sum_i \lambda_i t_i)$:

$$E_q\left[1_{\sum_i \lambda_i t_i(x)=y}\right] = 0.$$

4 Application to Clustering

In this section, we demonstrate the practical use of MCIG to cluster a set of mixtures in Sect. 4.1, and consider in Sect. 4.2 parallel calculations/aggregations of Monte Carlo Exponential/Mixture Functions.

4.1 *Clustering Mixtures on the Mixture Family Manifold*

Consider clustering a set of n mixtures $m(x;\eta_1),\ldots,m(x;\eta_n)$ of the mixture family manifold. Prior work considered clustering the mixture components (e.g., Gaussian components) to simplify mixtures by using the Bregman k-means [14, 37]. This prior work can be interpreted as a Gaussian component quantization procedure.

Here, we address the different problem of clustering the mixtures themselves, not their components.

Since $\mathrm{KL}(m(x;\eta_i) : m(x;\eta_j)) = B_G(\eta_i : \eta_j)$ for $G(\eta) = -h(m(x;\eta))$ (Shannon information), we may approximate the KL divergence from the MC Bregman Divergence (MCBD) $\tilde{G}_{\mathcal{S}}$ as follows:

$$\mathrm{KL}(m(x;\eta_i) : m(x;\eta_j)) = B_G(\eta_i : \eta_j), \tag{99}$$

$$\simeq B_{\tilde{G}_{\mathcal{S}}}(\eta_i : \eta_j). \tag{100}$$

One advantage of using a MCIG is that all divergence computations $B_{\tilde{G}_{\mathcal{S}}}$ performed during the execution of a Bregman algorithm are consistent by reusing the same variates of $\mathcal{S}$. In particular, this also guarantees to always have nonnegative estimated KL divergences.

The traditional way to MC estimate the KL divergence is to consider the MC stochastic integration of the extended Kullback–Leibler divergence [4]:

$$\widehat{\mathrm{eKL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m}\left(\log\frac{p(x_i)}{q(x_i)} + \frac{q(x_i)}{p(x_i)} - 1\right), \tag{101}$$

for $x_1, \ldots, x_m \sim p(x)$. Indeed, if we just used the MC KL estimator:

$$\widehat{\mathrm{KL}}_m(p:q) := \frac{1}{m}\sum_{i=1}^{m}\log\frac{p(x_i)}{q(x_i)}, \tag{102}$$

we may endup with negative values to our estimated KL, depending on the sample variates! This never happens for eKL which is a statistical divergence for the scalar divergence $\mathrm{ekl}(p:q) = p\log\frac{p}{q} + q - p \geq 0$.

Bregman k-means [4, 22] can be applied using either the sided or their symmetrized centroid [40]: The right-sided centroid is always the center of mass of the parameters. The left-sided centroid requires to compute $F'(\theta)$ and its reciprocal inverse function $(F'(\theta))^{-1}$ (wlog, assuming $D = 1$ for simplicity[5]). Although $F'(\theta)$ is available in closed form (and define the dual parameter θ):

$$\tilde{G}'_{\mathcal{S}}(\eta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{q(x_i)}\left(p_1(x_i) - p_0(x)\right)\left(1 + \log m(x;\eta)\right) = \theta, \tag{103}$$

the dual parameter of $(\mathcal{M}, G)$ cannot be written as a simple function $\eta = F^{*\prime}(\eta)$. Notice that $\theta = \tilde{G}'_{\mathcal{S}}(\eta)$ is an increasing function of η and that the inverting operation can be performed numerically. Indeed, we can compute $\eta = (\tilde{G}'_{\mathcal{S}})^{-1}(\theta) = \tilde{G}^*_{\mathcal{S}}(\theta)$ using a numerical scheme (e.g., bisection search).

The symmetric Jeffreys divergence is:

$$\begin{aligned} J(m(x;\eta_i):m(x;\eta_j)) &= \mathrm{KL}(m(x;\eta_i):m(x;\eta_j)) + \mathrm{KL}(m(x;\eta_j):m(x;\eta_i)), && (104)\\ &= B_G(\eta_i:\eta_j) + B_G(\eta_j:\eta_i), && (105)\\ &= B_G(\eta_i:\eta_j) + B_{G^*}(\theta_i:\theta_j), && (106)\\ &= \langle \Delta\theta_{ij}, \Delta\eta_{ij}\rangle, && (107) \end{aligned}$$

where $\Delta\theta_{ij} = \theta_i - \theta_j$ and $\Delta\eta_{ij} = \eta_i - \eta_j$.

We may approximate the J divergence by considering the Monte Carlo Bregman generator in Eq. 105:

$$J(m(x;\eta_i):m(x;\eta_j)) \simeq B_{\tilde{G}_{\mathcal{S}}}(\eta_i:\eta_j) + B_{\tilde{G}_{\mathcal{S}}}(\eta_j:\eta_i). \tag{108}$$

We can then apply the technique of mixed Bregman clustering [49] that considers two centers per cluster. Moreover a fast probabilistic initialization, called *mixed Bregman k-means++* [49], allows one to guarantee a good initialization with high probability (without computing centroids but requiring to compute divergences).

[5] Otherwise, we need to consider monotone operator theory [25] to invert $\nabla F(\theta)$.

Another technique to bypass the computation of the gradient $\nabla\tilde{G}_{\mathcal{S}}$ in the BD consists in taking the scaled skew α-Jensen divergence [35] for an infinitesimal value of α. Indeed, we have the α-Jensen divergence defined by:

$$J_F^{\alpha}(p:q) = (1-\alpha)F(p) + \alpha F(q) - F((1-\alpha)p + \alpha q), \tag{109}$$

and asymptotically this skewed Jensen divergences yield the sided Bregman divergences [35] as follows:

$$\lim_{\alpha\to 0^+} \frac{J_F^{\alpha}(p:q)}{\alpha} = B_F(q:p), \tag{110}$$

$$\lim_{\alpha\to 1^-} \frac{J_F^{\alpha}(p:q)}{1-\alpha} = B_F(p:q), \tag{111}$$

Thus we have for small values of $\alpha > 0$ (say, $\alpha = 0.001$):

$$J(m(x;\eta_i):m(x;\eta_j)) = B_G(\eta_i:\eta_j) + B_G(\eta_j:\eta_i), \tag{112}$$

$$\simeq \frac{1}{\alpha} J_{\tilde{G}_{\mathcal{S}}}^{\alpha}(\eta_i:\eta_j) + \frac{1}{1-\alpha} J_{\tilde{G}_{\mathcal{S}}}^{1-\alpha}(\eta_i:\eta_j). \tag{113}$$

The last equation Eq.113 is the symmetrized skew Jensen divergence studied in [29].

Figure 6 plots the result of a 2-cluster clustering wrt the Jeffreys' divergence for a set of $n = 8$ mixtures.

4.2 Parallelizing Information Geometry

We can distribute the Monte Carlo information geometry either on a multicore machine with l cores with shared memory or on a cluster of l machines with distributed memory, or even consider hybrid architectures.

Let $(M, \tilde{F}_{\mathcal{S}_1}), \ldots, (M, \tilde{F}_{\mathcal{S}_l})$ be a set of l information-geometric manifolds obtained from iid sample sets $\mathcal{S}_1, \ldots, \mathcal{S}_l$. Let $\oplus_{i=1}^{s}\mathcal{S}_i$ be a partition of $\mathcal{S}$.

4.2.1 Multicore Architectures

On a multicore architecture, we may evaluate the mixture family Bregman divergence $B_{\tilde{G}_{\mathcal{S}}}(\eta:\eta')$ by evaluating $B_{\tilde{G}_{\mathcal{S}_i}}(\theta:\theta')$, and using the compositionality rule of Bregman generators in BDs (Property 2) with:

$$\tilde{G}_{\mathcal{S}}(\theta) = \sum_{i=1}^{l} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \tilde{G}_{\mathcal{S}_i}(\eta). \tag{114}$$

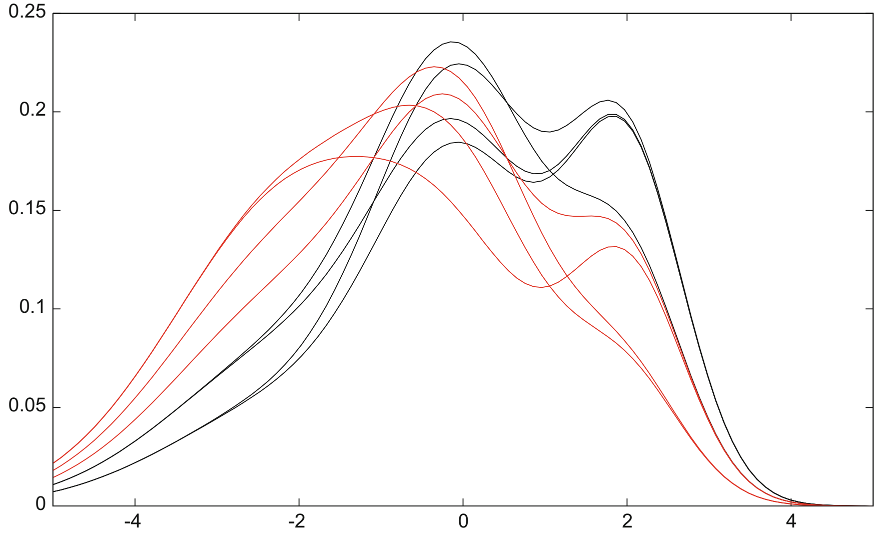


Fig. 6 Clustering a set of $n = 8$ statistical mixtures of order $D = 2$ with $K = 2$ clusters: Each mixture is represented by a 2D point on the mixture family manifold. The Kullback–Leibler divergence is equivalent to an integral-based Bregman divergence that is computationally untractable: The Bregman generator is stochastically approximated by Monte Carlo sampling

That is, $\tilde{G}_{\mathcal{S}}(\eta)$ is the *arithmetic weighted mean* of the mixture sub-generators.

For the exponential families, recall that we have:

$$\tilde{F}_{\mathcal{S}}(\theta) = \log\left(\sum_{i=1}^{s} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \exp(\tilde{F}_{\mathcal{S}_i})\right). \tag{115}$$

That is, $\tilde{F}_{\mathcal{S}}(\theta)$ can be interpreted as an *exponential mean* (quasi-arithmetic mean, called f-mean [35] for the monotonically increasing function $f(x) = \exp(x)$) of the sub-generators. Thus we can perform the computation of the MC Bregman generators on multi-core architectures easily with a MapReduce strategy [33].

Fact 1 (MapReduce evaluation of MC Bregman generators) *The MCMF or MCEF functions can be computed in parallel using a quasi-arithmetic mean MapReduce operation.*

4.2.2 Cluster Architectures

Since the MC Bregman generators can be interpreted as random variables $\tilde{G}_m(\theta)$ and $\tilde{F}_m(\theta)$, we may obtain robust estimate [51] by carrying the calculations on l MCIGs on a cluster architecture, and then integrate those l geometries.

Given a sequence of matching parameters $\theta_1 \in (M, \tilde{F}_{s_1}), \dots, \theta_l \in (M, \tilde{F}_{s_l})$, we aggregate these parameters by doing the *KL-averaging* method [26]. This amounts to compute a sided centroid for θ.

5 Information-Geometric Structures Induced by Statistical Separable Divergences

In this section, we consider Monte Carlo sampling to define a (tractable) statistical divergence that approximates another (untractable) statistical divergence, and uses this MC statistical divergence to define an information-geometric manifold.

The core structure of information geometry [1] is a manifold M equipped with a pair of dual connections, ∇ and ∇^* coupled to the metric tensor g: (M, g, ∇, ∇^*). In terms of differential geometry, the definition of this coupling is expressed as

$$X\langle Y, Z\rangle_g = \langle \nabla_X Y, Z\rangle_g + \langle Y, \nabla^*_X Z\rangle_g, \tag{116}$$

where X, Y and Z are smooth vector fields on M. The coupling of connections to the metric tensor means that the dual parallel transport is compatible with the metric:

$$\langle u, v\rangle_{c(0)} = \left\langle \prod^{\nabla}_{c(0)\to c(t)} u, \prod^{\nabla^*}_{c(0)\to c(t)} v \right\rangle_{c(t)}, \tag{117}$$

where c is a smooth curve (parallel transport is path dependent, except for dually flat connections). The notation $\prod^{\nabla}_{c(0)\to c(t)} u$ means that vector $u \in T_{c(0)} = T_p$ is parallel transported along smooth curve c to tangent plane $T_{c(t)}$ with respect to the affine connection ∇. From this (M, g, ∇, ∇^*) structure, a statistical manifold [24] (M, g, C) can be defined, where $C(X, Y, Z) = \langle \nabla_X Y - \nabla^*_X Y, Z\rangle$ is a totally symmetric cubic tensor, termed the Amari–Chentsov cubic tensor. It follows a one-parameter family of dual connections [1] (with ∇^0 being the Levi-Civita metric connection): $(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$ so that if connection ∇^{α} has constant curvature κ then its dual connection has also the same curvature. Furthermore, one can build [1, 16, 17] a pair of dual connections coupled to a metric from any smooth divergence D: $(M, {}^D g, {}^D\nabla, {}^D\nabla^*)$. Figure 7 summarizes the fundamental structures of parametric information geometry and their relationships.

Let us consider a separable statistical divergence:

$$D[p : q] := \int d(p(x) : q(x))\mathrm{d}\mu(x), \tag{118}$$

where $d(x : y)$ is a scalar divergence. For example, the f-divergences [1] are obtained for $i_f(x : y) = xf(x/y)$:

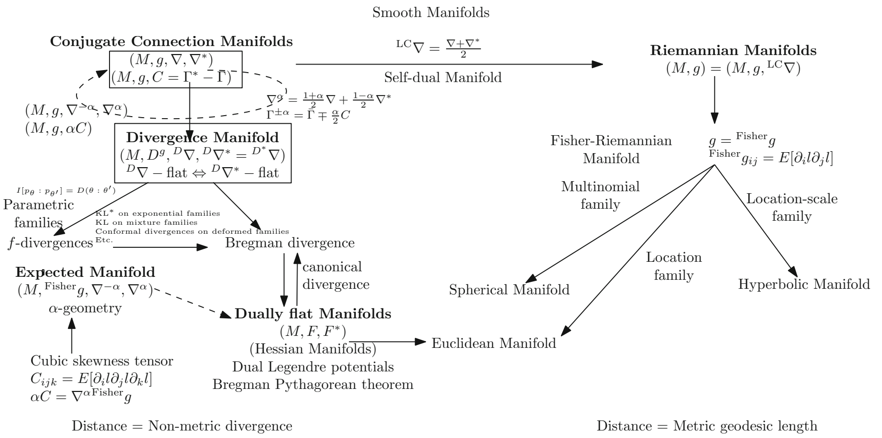

Smooth Manifolds
Conjugate Connection Manifolds
(M, g, ∇, ∇^*)
$(M, g, C = \Gamma^* - \bar{\Gamma})$
${}^{\mathrm{LC}}\nabla = \frac{\nabla + \nabla^*}{2}$
Self-dual Manifold
Riemannian Manifolds
$(M, g) = (M, g, {}^{\mathrm{LC}}\nabla)$
$(M, g, \nabla^{-\alpha}, \nabla^{\alpha})$
$(M, g, \alpha C)$
$\nabla^{\alpha} = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\nabla^*$
$\Gamma^{\pm\alpha} = \bar{\Gamma} \mp \frac{\alpha}{2} C$
Divergence Manifold
$(M, D^g, {}^D\nabla, {}^D\nabla^* = {}^{D^*}\nabla)$
${}^D\nabla$ – flat ⇔ ${}^D\nabla^*$ – flat
Fisher-Riemannian Manifold
$g = {}^{\mathrm{Fisher}}g$
${}^{\mathrm{Fisher}}g_{ij} = E[\partial_i l \partial_j l]$
$I[p_\theta : p_{\theta'}] = D(\theta : \theta')$
Parametric families
Multinomial family
Location-scale family
KL^* on exponential families
KL on mixture families
Conformal divergences on deformed families
Etc.
f-divergences
Bregman divergence
Location family
Expected Manifold
$(M, {}^{\mathrm{Fisher}}g, \nabla^{-\alpha}, \nabla^{\alpha})$
α-geometry
canonical divergence
Spherical Manifold
Hyperbolic Manifold
Dually flat Manifolds
(M, F, F^*)
(Hessian Manifolds)
Dual Legendre potentials
Bregman Pythagorean theorem
Euclidean Manifold
Cubic skewness tensor
$C_{ijk} = E[\partial_i l \partial_j l \partial_k l]$
$\alpha C = \nabla^{\alpha\mathrm{Fisher}} g$
Distance = Non-metric divergence
Distance = Metric geodesic length

Fig. 7 The web of fundamental information-geometric structures. An arrow $a \to b$ means that geometric structure b is a special case of the (meta-)structure a

$$I_f[p : q] := \int p(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x) = \int i_f(p(x) : q(x)) \mathrm{d}\mu(x), \tag{119}$$

The f-divergences are the only statistical separable divergences that satisfy the information monotonicity property [1]. On a parametric family of distributions $\{p_\theta\}$, the statistical f-divergences amount to equivalent parameter divergences:

$$D_f(\theta_1 : \theta_2) := I_f[p_{\theta_1} : p_{\theta_2}] \tag{120}$$

The information-geometric structure induced by this (parameter) divergence is $(M, {}^{D_f}g, {}^{D_f}\nabla, {}^{D_f}\nabla^*)$, and the dual connections correspond to the expected α-connections[1] for f-divergences.

It may happen that D_f, although well-defined, may not be available in closed form. In that case, we approximate the divergence by Monte Carlo stochastic integration by drawing a set $\mathcal{S}_m = \{x_1, \ldots, x_m\}$ of m iid variates from p_{θ_1}:

$$\tilde{D}_{\mathcal{S}_m}(\theta_1 : \theta_2) := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{p_{\theta_1}(x_i)} d(p_{\theta_1}(x_i) : p_{\theta_2}(x_i)). \tag{121}$$

We need to assert that $\tilde{D}_{\mathcal{S}_m}$ is a smooth divergence: The smoothness of the divergence $\tilde{D}_{\mathcal{S}_m}$ follows from the smoothness divergence of the corresponding scalar divergence d. Then we need to guarantee that $\tilde{D}_{\mathcal{S}_m}(\theta_1 : \theta_2) = 0$ iff $\theta_1 = \theta_2$. Since $d(p_{\theta_1}(x) : p_{\theta_2}(x)) = 0$ if and only if $p_{\theta_1}(x) = p_{\theta_2}(x)$, we need to assert that with high probability $p_{\theta_1}(x) \neq p_{\theta_2}(x)$ when $\theta_1 \neq \theta_2$. Let $I = \max_{\theta_1, \theta_2} \mu(\{p_{\theta_1}(x) = p_{\theta_2}(x), x \in \mathcal{X}\})$. When $I = 0$, then almost surely $\tilde{D}_{\mathcal{S}}$ is a divergence. This condition holds when the probability densities intersect in at most a finite number of points. It follows the corresponding information-geometric structure $(M, {}^{\tilde{D}_{\mathcal{S}_m}}g, {}^{\tilde{D}_{\mathcal{S}_m}}\nabla, {}^{\tilde{D}_{\mathcal{S}_m}}\nabla^*)$ (with its associated one-family of α-connections) such that asymptotically, we have:

$$\lim_{m\to\infty}(M, {}^{\tilde{D}_{\mathcal{S}_m}}g, {}^{\tilde{D}_{\mathcal{S}_m}}\nabla, {}^{\tilde{D}_{\mathcal{S}_m}}\nabla^*) = (M, {}^{D}g, {}^{D}\nabla, {}^{D}\nabla^*), \tag{122}$$

as desired.

Let us quickly report two examples to illustrate these divergence-based sequences of information-geometric structures:

- Polynomial Exponential Families (PEFs) of order D with the γ-divergences [50]: Let us notice that we do not need to normalize the PEF distributions in order to sample variates, and that the γ-divergence D_γ is a projective divergence [48] (invariant by positive rescaling of the distributions) which tends to the KL divergence when $\gamma \to 0$. Since the densities of any two distinct PEF distributions of order D intersect in at most $D + 1$ points, we check that $I = 0$. Thus for $\gamma \to 0$ and $m \to \infty$, we tend to a dually flat manifold. As an application, we can consider clustering these PEFs on a MCIG manifold.
- Consider a mixture family $\{m_\eta(x) = (1 - \eta)p_1(x) + \eta p_2(x), \eta \in (0, 1)\}$ of order $D = 1$ for the two mixture component distributions p_1 and p_2, linearly independent. We have $m_{\eta_1}(x) = m_{\eta_2}(x)$ iff $p_1(x) = p_2(x)$ (holds only for this particular case of $D = 1$). Assume $I = 0$ for the component distributions, then we obtain a sequence of Monte Carlo information-geometric structures that tend asymptotically to the dually flat mixture manifold.

In the later case, we consider the MCIG manifold for a 1D mixture manifold with respect to an arbitrary divergence. Notice that the divergence-based MCIG for the exponential/mixture manifold may not be flat for KL. In Sect. 2.2, we took the different approach of approximating the negative differential entropy via Monte-Carlo, ensuring that all sequence of MCIG manifolds are dually flat.

6 Conclusion and Perspectives

In this work, we proposed a new type of *randomized information-geometric structure* to cope with computationally untractable information-geometric structures (types 4 and 5 in the classification of Table 4): Namely, the Monte Carlo Information Geometry [38] (MCIG). MCIG performs stochastic integration of the ideal but computationally intractable definite integral-based Bregman generator (e.g. Eq. 32 for mixture family) for mixture family and Eq. 24 for exponential family). We proved that the MC Bregman generators for the mixture family and the exponential family are almost surely strictly convex and differentiable (Theorem 2 and Theorem 4, respectively), and therefore yield a computationally tractable information-geometric structure (Type 2 in the classification of Table 4). Thus we can get a series of *consistent* and *computationally-friendly* information-geometric structures that tend asymptotically

Table 4 A smooth and strictly convex function F induces a dually flat structure: We classify those structures according to their computational tractability properties

Type	F	∇F^*	Example
Type 1	Closed-form	Closed-form	Gaussian (exponential) family
Type 2	Closed-form	Not closed-form	Beta (exponential) family
Type 3	Comp. intractable	Not closed-form	Ising family [54]
Type 4	Not closed-form	Not closed-form	Polynomial exponential family [42]
Type 5	Not analytic	Not analytic	Mixture family

to the untractable ideal information geometry. We have demonstrated the usefulness of our technique for a basic Bregman k-means clustering technique: Clustering statistical mixtures on a mixture family manifold. Although the MCIG structures are computationally convenient, we do not have in closed-form ∇F^* (nor F^*) because our Bregman generators are the sum of basic generators whose gradients are the sum of elementary gradients that cannot be inverted easily.[6] This step requires a numerical or symbolic technique [25].

We note that in the recent work of [27], Matsuzoe et al. defined a sequence of statistical manifolds relying on a sequential structure of escort expectations for non-exponential type statistical models.

Codes for reproducible results are available at:

https://franknielsen.github.io/MCIG/

7 Function $\mathrm{lse}_0^+(x)$ is a Bregman Generator

We give the proof of Theorem 3:

Proof Since $\mathrm{lse}_0^+(x_1,\ldots,x_d) = \log\left(1+\sum_{i=1}^d \exp(x_i)\right)$ is twice continuously differentiable, it suffices to prove that $\nabla^2 \mathrm{lse}_0^+(x) \succ 0$. We have:

$$\partial_i \mathrm{lse}_0^+(x) = \frac{e^{x_i}}{1+\sum_k e^{x_k}}, \tag{123}$$

$$\partial_j\partial_i \mathrm{lse}_0^+(x) \overset{j\neq i}{=} \frac{-e^{x_i}e^{x_j}}{(1+\sum_k e^{x_k})^2}, \tag{124}$$

$$\partial_i\partial_i \mathrm{lse}_0^+(x) = \frac{e^{x_i}(1+\sum_k e^{x_k}) - e^{x_i}e^{x_j}}{(1+\sum_k e^{x_k})^2}. \tag{125}$$

[6] The Legendre conjugate of an infimal convolution of elementary functions is the sum of the elementary conjugate functions.

It follows that the Hessian $(\partial_j \partial_i \mathrm{lse}_0^+(x))_{ij}$ is a diagonally dominant matrix since:

$$e^{x_i}\left(1+\sum_k e^{x_k}\right) = e^{x_i} + e^{x_i}\sum_k e^{x_k} > \sum_{j\neq i}\left|-e^{x_i}e^{x_j}\right| = e^{x_i}\sum_{j\neq i} e^{x_j}. \tag{126}$$

To conclude that the Hessian matrix is SPD, we use Gershgorin circle theorem [55] to bound the spectrum of a square matrix: The eigenvalues of the Hessian matrix are thus real and fall inside a disk of center $(e^{x_i}(1+\sum_k e^{x_k}))_i$ and radius $e^{x_i}\sum_{j\neq i} e^{x_j}$. Therefore all eigenvalues are positive, and the Hessian matrix is positive definite.

For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we have:

$$\nabla \mathrm{lse}(x) = \sigma(x), \tag{127}$$

where $\sigma(x)$ is the *softmax* function:

$$\sigma(x) := \left(\frac{e^{x_i}}{\sum_{k=1}^d e^{x_k}}\right)_{i\in\{1,\ldots,d\}}. \tag{128}$$

By analogy, we may define for $x \in \mathbb{R}^d$:

$$\sigma_0^+(x) := \left(\frac{e^{x_i}}{1+\sum_k e^{x_k}}\right)_{i\in\{1,\ldots,d\}}, \tag{129}$$

so that $\nabla \mathrm{lse}_0^+(x) = \sigma_0^+(x)$.

References

1. Amari, S.: Information Geometry and Its Applications. Applied Mathematical Sciences. Springer, Japan (2016)
2. Amari, Si, Cichocki, A.: Information geometry of divergence functions. Bull. Polish Acad. Sci.: Tech. Sci. **58**(1), 183–195 (2010)
3. Azoury, K.S., Warmuth, M.K.: Relative loss bounds for on-line density estimation with the exponential family of distributions. Mach. Learn. **43**(3), 211–246 (2001)
4. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**(Oct), 1705–1749 (2005)
5. Bhattacharya, B.B., Mukherjee, S., et al.: Inference in Ising models. Bernoulli **24**(1), 493–525 (2018)
6. Boissonnat, J.D., Nielsen, F., Nock, R.: Bregman Voronoi diagrams. Discret. Comput. Geom. **44**(2), 281–307 (2010)
7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
8. Calin, O., Udriste, C.: Geometric Modeling in Probability and Statistics. Mathematics and Statistics. Springer International Publishing, Berlin (2014)

9. Cipra, B.A.: The Ising model is NP-complete. SIAM News **33**(6), 1–3 (2000)
10. Cobb, L., Koppstein, P., Chen, N.H.: Estimation and moment recursion relations for multimodal distributions of the exponential family. J. Am. Stat. Assoc. **78**(381), 124–130 (1983)
11. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2012)
12. Critchley, F., Marriott, P.: Computational information geometry in statistics: theory and practice. Entropy **16**(5), 2454–2471 (2014)
13. Crouzeix, J.P.: A relationship between the second derivatives of a convex function and of its conjugate. Math. Programm. **13**(1), 364–365 (1977)
14. Davis, J.V., Dhillon, I.S.: Differential entropic clustering of multivariate gaussians. In: Advances in Neural Information Processing Systems, pp. 337–344 (2007)
15. Dawid, A.P.: The geometry of proper scoring rules. Ann. Inst. Stat. Math. **59**(1), 77–93 (2007)
16. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**, 793–803 (1983)
17. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**(3), 631–647 (1992)
18. Fleisch, D.A.: A Student's Guide to Vectors and Tensors. Cambridge University Press, Cambridge (2011)
19. Frongillo, R., Reid, M.D.: Convex foundations for generalized MaxEnt models. In: AIP Conference Proceedings, vol. 1636, pp. 11–16. AIP (2014)
20. Gao, B., Pavel, L.: On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning. ArXiv e-prints (2017)
21. Geman, S., Graffigne, C.: Markov random field image models and their applications to computer vision. In: Proceedings of the International Congress of Mathematicians, vol. 1, p. 2 (1986)
22. Grønlund, A., Larsen, K.G., Mathiasen, A., Nielsen, J.S.: Fast exact k-means, k-medians and Bregman divergence clustering in 1D (2017). arXiv:1701.07204
23. Kass, R.E., Vos, P.W.: Geometrical Foundations of Asymptotic Inference. Fisher-Rao metric of location-scale family is hyperbolic (and can be diagonalized), pp. 192–193. Wiley-Interscience (1997)
24. Lauritzen, S.L.: Statistical manifolds. Differential Geometry in Statistical Inference, p. 164 (1987)
25. Lauster, F., Luke, D.R., Tam, M.K.: Symbolic computation with monotone operators. Set-Valued and Variational Analysis, pp. 1–16 (2017)
26. Liu, Q., Ihler, A.T.: Distributed estimation, information loss and exponential families. In: Advances in Neural Information Processing Systems, pp. 1098–1106 (2014)
27. Matsuzoe, H., Scarfone, A.M., Wada, T.: A sequential structure of statistical manifolds on deformed exponential family. In: International Conference on Geometric Science of Information, pp. 223–230. Springer (2017)
28. Mitchell, A.F.S.: Statistical manifolds of univariate elliptic distributions. International Statistical Review/Revue Internationale de Statistique, pp. 1–16 (1988)
29. Nielsen, F.: A family of statistical symmetric divergences based on Jensen's inequality (2010). arXiv:1009.4004
30. Nielsen, F.: Legendre transformation and information geometry (2010)
31. Nielsen, F.: Hypothesis testing, information divergence and computational geometry. In: Geometric Science of Information, pp. 241–248. Springer (2013)
32. Nielsen, F.: An information-geometric characterization of Chernoff information. IEEE Signal Process. Lett. **20**(3), 269–272 (2013)
33. Nielsen, F.: Introduction to HPC with MPI for Data Science. Undergraduate Topics in Computer Science. Springer (2016). https://doi.org/10.1007/978-3-319-21903-5. https://doi.org/10.1007/978-3-319-21903-5
34. Nielsen, F., Boissonnat, J.D., Nock, R.: On Bregman Voronoi diagrams. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp. 746–755 (2007)
35. Nielsen, F., Boltz, S.: The Burbea-Rao and Bhattacharyya centroids. IEEE Trans. Inf. Theory **57**(8), 5455–5466 (2011)

36. Nielsen, F., Garcia, V.: Statistical exponential families: a digest with flash cards (2009) arXiv:0911.4863
37. Nielsen, F., Garcia, V., Nock, R.: Simplifying Gaussian mixture models via entropic quantization. In: 17th European Conference on Signal Processing (EUSIPCO), pp. 2012–2016. IEEE (2009)
38. Nielsen, F., Hadjeres, G.: Monte Carlo information geometry: the dually flat case (2018). CoRR arXiv:1803.07225
39. Nielsen, F., Nock, R.: On the smallest enclosing information disk. Inf. Process. Lett. **105**(3), 93–97 (2008)
40. Nielsen, F., Nock, R.: Sided and symmetrized Bregman centroids. IEEE Trans. Inf. Theory **55**(6), 2882–2904 (2009)
41. Nielsen, F., Nock, R.: Optimal interval clustering: application to Bregman clustering and statistical mixture learning. IEEE Signal Process. Lett. **21**(10), 1289–1292 (2014)
42. Nielsen, F., Nock, R.: Patch matching with polynomial exponential families and projective divergences. In: International Conference on Similarity Search and Applications, pp. 109–116. Springer (2016)
43. Nielsen, F., Nock, R.: On w-mixtures: finite convex combinations of prescribed component distributions (2017). CoRR arXiv:1708.00568
44. Nielsen, F., Nock, R.: On the geometric of mixtures of prescribed distributions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
45. Nielsen, F., Piro, P., Barlaud, M.: Bregman vantage point trees for efficient nearest neighbor queries. In: 2009 IEEE International Conference on Multimedia and Expo, ICME 2009, pp. 878–881. IEEE (2009)
46. Nielsen, F., Piro, P., Barlaud, M.: Tailored Bregman ball trees for effective nearest neighbors. In: Proceedings of the 25th European Workshop on Computational Geometry (EuroCG), pp. 29–32 (2009)
47. Nielsen, F., Sun, K.: Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. Entropy **18**(12), 442 (2016)
48. Nielsen, F., Sun, K., Marchand-Maillet, S.: On hölder projective divergences. Entropy **19**(3), 122 (2017)
49. Nock, R., Luosto, P., Kivinen, J.: Mixed Bregman clustering with approximation guarantees. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 154–169. Springer (2008)
50. Notsu, A., Komori, O., Eguchi, S.: Spontaneous clustering via minimum γ-divergence. Neural Comput. **26**(2), 421–448 (2014)
51. Pelletier, B.: Informative barycentres in statistics. Ann. Inst. Stat. Math. **57**(4), 767–780 (2005)
52. Robert, C.P.: Monte Carlo methods. Wiley Online Library (2004)
53. Shima, H.: The Geometry of Hessian Structures. World Scientific, Singapore (2007)
54. Tang, Y., Salakhutdinov, R.R.: Learning stochastic feedforward neural networks. In: Advances in Neural Information Processing Systems, pp. 530–538 (2013)
55. Varga, R.S.: Geršgorin and His Circles, vol. 36. Springer Science & Business Media, Berlin (2010)
56. Watanabe, S., Yamazaki, K., Aoyagi, M.: Kullback information of normal mixture is not an analytic function. Technical report of IEICE (in Japanese) (2004-0), pp. 41–46 (2004)
57. Zhang, J.: Reference duality and representation duality in information geometry. In: AIP Conference Proceedings, vol. 1641, pp. 130–146. AIP (2015)

Information Geometry in Portfolio Theory

Ting-Kam Leonard Wong

Abstract We review some recent developments in stochastic portfolio theory motivated by information geometry, present illustrative examples and an extension of functional portfolio generation. Several problems are suggested for further study.

1 Introduction

In the first chapter of their influential monograph [1, p.1] Amari and Nagaoka explained the key idea of information geometry as follows:

> Information geometry ... allow[s] us to take problems from a variety of fields: statistics, information theory, and control theory; visualize them geometrically; and from this develop novel tools with which to extend and advance these fields.

In this chapter we show that this principle can be fruitfully applied to financial problems. We review some recent development in the field of stochastic portfolio theory (SPT) motivated by information geometry, present illustrative examples and an extension of functional portfolio generation (announced in [2] which is an early version of this paper), and suggest several problems for further study. It is hoped that the materials will be of interest to researchers in both information geometry and mathematical finance. The topics discussed are heavily influenced by the author's research interests. Other financial applications of information geometry are briefly reviewed in Sect. 1.2.

T.-K. L. Wong (✉)
University of Toronto, Toronto, Canada
e-mail: tkl.wong@utoronto.ca

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication
Technology, https://doi.org/10.1007/978-3-030-02520-5_6

1.1 Main Ideas: Market Diversity and Volatility

To set the stage let us consider a universe of stocks represented by a capitalization-weighted index. A typical example is the S&P 500 Index which includes a significant portion of the US stock market. In a capitalization-weighted index, the influence of a stock is proportional to its market capitalization. Following the standard set up of stochastic portfolio theory (see for example [3, 4]), if we let $X_i(t) > 0$ be the market capitalization of stock i at time t, then

$$\mu_i(t) := \frac{X_i(t)}{X_1(t) + \cdots + X_n(t)} \tag{1}$$

is the *market weight* of the stock. Throughout this chapter we let time be discrete. Let $n \geq 2$ be the number of stocks in the market. The vector $\mu(t) = (\mu_1(t), \ldots, \mu_n(t))$ takes values in the open unit simplex Δ_n given by

$$\Delta_n := \{p = (p_1, \ldots, p_n) \in (0, 1)^n : p_1 + \cdots + p_n = 1\}.$$

Its closure in $\mathbb{R}^n$ is denoted by $\overline{\Delta_n}$. The vector $\mu(t)$ may also be regarded as the portfolio weights of the *market portfolio*.

A major objective of SPT is to construct investment strategies that beat the market under realistic conditions on market behaviors. These portfolios are called *relative arbitrages* with respect to the market portfolio. For more details about this important problem see for example [3–5] and their references. As a simple example, Fig. 1 plots the path of $\{\mu(t)\}$ for a hypothetical 3-stock market consisting of the US stocks Ford, IBM and Walmart. It should not be surprising that the relative performance of portfolios with respect to the market can be analyzed using appropriate geometries on the simplex.[1]

Regarding the market as a process in the simplex Δ_n, there are two natural quantities an investor may want to keep track of: *diversity* and *volatility*. Diversity refers to the degree of capital concentration in the equity market. For example, in Fig. 1, the market moves towards the vertex representing Walmart and thus becomes more concentrated. According to [6], many mutual funds tend to overweight small stocks and underweight large stocks (relative to the market index), so the change in market diversity is a significant predictor of their relative performance. To quantify diversity one introduces a positive *concave* function $\Phi : \Delta_n \to (0, \infty)$, and we say that the market is more diverse when $\Phi(\mu(t))$ is large. Typical examples include the Shannon entropy

$$\Phi(p) = -\sum_{j=1}^{n} p_j \log p_j$$

[1] In practice the number of stocks changes with time, and the market capitalization may fluctuate due to public offerings and other events. For simplicity these complications are neglected here.

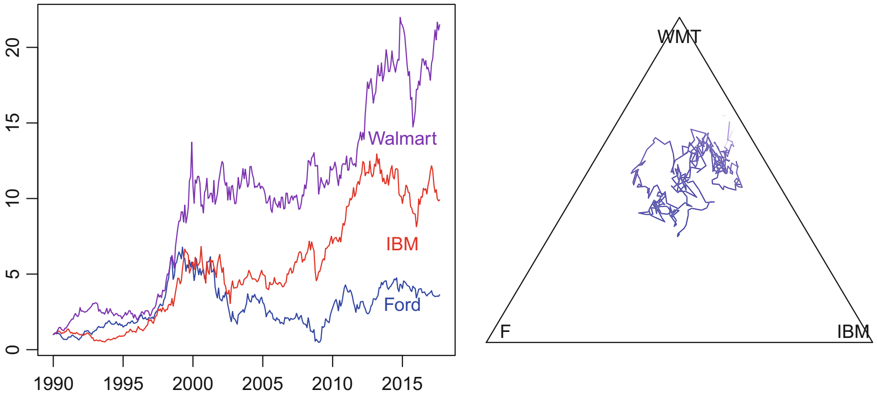


Fig. 1 Left: Monthly prices of the stocks (regarded as the capitalization and normalized to be 1 from January 1990). Right: Path of the market weight vector $\mu(t)$ in the simplex Δ_3

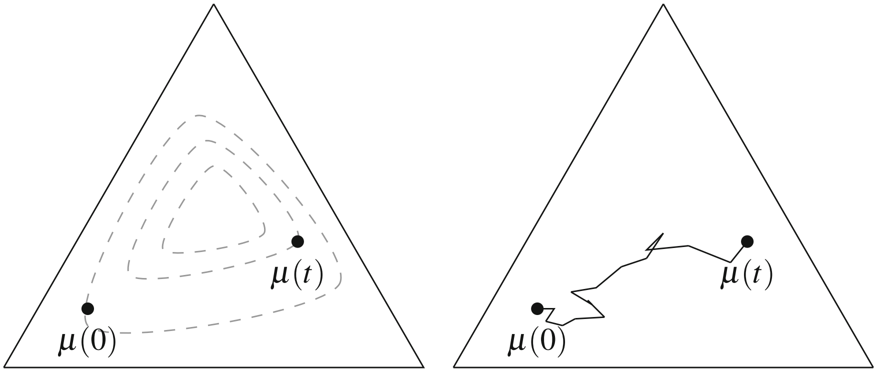


Fig. 2 Left: Change in market diversity measured by φ. The dashed curves represent level sets of φ. Right: Market volatility can be measured by the L-divergence $\mathbf{D}[\cdot \mid \cdot]$

as well as the λ-diversity function

$$\Phi(p) = \left(\sum_{j=1}^{n} (p_j)^{\lambda} \right)^{1/\lambda}, \tag{2}$$

where $0 < \lambda < 1$ is a parameter. More examples of measures of diversity can be found in [3, Chapter 3]. Note that we allow Φ to be asymmetric, so it can attain its maximum value at a point other than the barycenter $\overline{e} := \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. As it turns out, it is mathematically more convenient to consider its logarithm $\varphi := \log \Phi$. Since $e^{\varphi} = \Phi$ is concave, we say that φ is *exponentially concave*. We remark that φ, being the logarithm of a concave function, is itself a concave function. Given φ, the time series of $\{\varphi(\mu(t))\}$ is an indicator of market diversity (see Fig. 2 (left)).

The volatility of the market weight $\mu(t)$ refers to the volatility of the stocks *relative to each other*. Anticipating the use of information geometry, note that the Euclidean quadratic variation (in discrete time) given by

$$\sum_{s=0}^{t-1} |\mu(s+1) - \mu(s)|^2 \tag{3}$$

may not be appropriate because the Euclidean norm on the simplex may not have a financial meaning (see however Example 3). In particular, the same displacement $v = \mu(s+1) - \mu(s)$ (approximated by a tangent vector) should have different norms on different portions of the simplex. Depending on the application, market volatility along a path should be quantified by a sum like

$$\sum_{s=0}^{t-1} \mathbf{D}\left[\mu(s+1) \mid \mu(s)\right],$$

where $\mathbf{D}[\cdot \mid \cdot] : \Delta_n \times \Delta_n \to [0, \infty)$ is possibly asymmetric in its arguments. In information geometry we say that $\mathbf{D}[\cdot \mid \cdot]$ a *divergence* (a rigorous definition is given in Definition 5). Intuitively, the asymmetry of $\mathbf{D}[\cdot \mid \cdot]$ reflects the effect of time: the time-reversed path $\tilde{\mu}(t) = \mu(T - t)$ probably has different impacts on the portfolio.

The main idea of this chapter is the following. A differentiable, exponentially concave function $\varphi : \Delta_n \to \mathbb{R}$ defines an *L-divergence* (*L* stands for logarithmic)

$$\mathbf{D}^{(1)}\left[q \mid p\right] := \log\left(1 + \nabla\varphi(p) \cdot (q - p)\right) - \left(\varphi(q) - \varphi(p)\right) \geq 0, \quad p, q \in \Delta_n, \tag{4}$$

which can be used to quantify market volatility. (The superscript will become clear in Definition 1.) An important example of L-divergence is the *excess growth rate* (also known as the *diversification return* [7]) defined for a fixed portfolio vector $\pi \in \overline{\Delta_n}$ by

$$\mathbf{T}_\pi\left[q \mid p\right] := \log\left(\sum_{i=1}^n \pi_i \frac{q_i}{p_i}\right) - \sum_{i=1}^n \pi_i \log\frac{q_i}{p_i}. \tag{5}$$

The corresponding exponentially concave function is $\varphi(p) = \sum_{i=1}^n \pi_i \log p_i$. Note that the L-divergence is different from the classical Bregman divergence of a concave function:

$$\mathbf{D}^{(0)}\left[q \mid p\right] := \nabla\varphi(p) \cdot (q - p) - \left(\varphi(q) - \varphi(p)\right). \tag{6}$$

The most important example of Bregman divergence is the relative entropy

$$\mathbf{H}\left(q \mid p\right) := \sum_{i=1}^n q_i \log\frac{q_i}{p_i}, \tag{7}$$

where the potential function φ is the Shannon entropy. Comparing (5) and (7), we see that the excess growth rate involves a nonlinear transformation of an integral, whereas the relative entropy is itself an integral. In Example 2 we show that the Rényi entropy generates the Rényi divergence in the sense of L-divergence and is related to the diversity function (2).

In Sect. 3, we will show that the L-divergence determines uniquely an investment strategy, called a *multiplicatively generated portfolio*, whose performance $V(t)$ relative to the market has the pathwise decomposition

$$\log V(t) - \log V(0) = \varphi(\mu(t)) - \varphi(\mu(0)) + \sum_{s=0}^{t-1} \mathbf{D}^{(1)}\left[\mu(s+1) \mid \mu(s)\right]. \tag{8}$$

From this decomposition, as long as $\varphi(\mu(t))$ remains bounded and the cumulative volatility grows at a steady rate, the portfolio will outperform the market in the long run (see Proposition 1). Furthermore, the dualistic geometry induced by the L-divergence (in the sense of [8, 9]; also see [10]) has interesting financial applications (see Sect. 5.2). In a continuous time framework and without using geometric concepts, these portfolios were first introduced by Fernholz [3, 11]. Here we adopt the discrete time, geometric approach established in [12, 13].

Following [2, 14], in Sect. 4 we will generalize the portfolio construction using the $L^{(\alpha)}$-divergence:

Definition 1 ($L^{(\alpha)}$*-divergence*) Let $\varphi : \Delta_n \to \infty$ be differentiable and α-exponentially concave, i.e, $e^{\alpha\varphi}$ is concave. The $L^{(\alpha)}$-divergence of φ is defined for $p, q \in \Delta_n$ by

$$\mathbf{D}^{(\alpha)}\left[q \mid p\right] := \frac{1}{\alpha} \log\left(1 + \alpha \nabla\varphi(p) \cdot (q - p)\right) - \left(\varphi(q) - \varphi(p)\right), \tag{9}$$

Our framework covers also the *additively generated portfolio* introduced recently in [15]. Note that the L-divergence is the $L^{(1)}$-divergence, and the Bregman divergence is equal to $\mathbf{D}^{(\alpha)}$ as $\alpha \downarrow 0$, so we will also call it the $L^{(0)}$-divergence. Using obvious notations, we have the identity

$$\mathbf{D}_{\varphi}^{(\alpha)}\left[\cdot \mid \cdot\right] \equiv \frac{1}{\alpha} \mathbf{D}_{\alpha\varphi}^{(1)}\left[\cdot \mid \cdot\right]. \tag{10}$$

According to the results obtained recently in [14], it appears that the $L^{(\alpha)}$-divergence is the canonical interpolation between the Bregman divergence and the L-divergence, and plays a fundamental role in information geometry.

1.2 Financial Applications of Information Geometry

Instead of conducting an extensive literature review, we content with giving a sample of other financial applications of information geometry. There are mainly two (overlapping) directions: (i) geometries on the state space of financial dynamics, and (ii) optimization using information-geometric quantities such as entropy and divergence.

Regarding the first direction, the paper [16] identifies a yield curve with a distribution function and studies its corresponding dynamics using the Fisher information metric. In option pricing, [17] applies Tsallis's deformed exponentials and generalized the Black-Scholes model to fat-tailed distributions. Though not directly related to finance, the paper [18] generalizes the concept of multiplicatively generated portfolio map (see Sect. 3) to a large class of optimal transport problems. The recent work [19] applies the Fisher metric in the study of systemic risk.

On the other hand, divergences are frequently useful as objective/cost functionals. In [20, 21], the authors generalize Markowitz's mean-variance model to a mean-divergence model, and show that the resulting portfolios have superior performance. Optimization of probabilistic functionals under a divergence constraint is studied [22] and applied to model risk.

1.3 Outline of the Paper

In Sect. 2 we present the market model and introduce two ways of representing a trading strategy and the associated value process. Section 3 reviews known results about multiplicatively generated portfolios with an emphasis on ideas and clarity. Motivated by these results and the recently introduced additively generated portfolio, in Sect. 4 we introduce a general framework for functional portfolio generation, where the $L^{(\alpha)}$-divergence arises naturally. Further properties of the L-divergence are discussed in Sect. 5 and several related problems are stated.

2 The Market Model

We work under the discrete time, pathwise setting used in our previous papers [12, 23, 24]. Let $n \geq 2$, the number of stocks in the market, be fixed. The data of our model is a sequence $\{\mu(t) = (\mu_1(t), \ldots, \mu_n(t))\}_{t=0}^{\infty}$ with values in the open unit simplex Δ_n. We regard $\mu(t)$ as the vector of market weights at time t. At this point we do not impose any condition on the sequence $\{\mu(t)\}_{t=0}^{\infty}$. In Proposition 1 we will give examples of path properties that lead to relative arbitrages.

We consider self-financing trading strategies in this market model. Let us express a strategy in terms of the number of shares held at each point in time. Furthermore, we use the market portfolio as the numéraire (i.e., unit of price). This means that the (relative) value of stock i is simply the market weight $\mu_i(t)$. We assume that trading is frictionless.

Definition 2 (*Self-financed trading strategy*) A self-financing trading strategy is a sequence $\eta = \{\eta(t)\}_{t=0}^{\infty}$, with values in $\mathbb{R}^n$, such that the self-financing identity

$$\sum_{i=1}^{n} \eta_i(t)\mu_i(t+1) \equiv \sum_{i=1}^{n} \eta_i(t+1)\mu_i(t+1) \tag{11}$$

holds for all time t. We always assume η is adapted in the sense that for each $t \geq 0$, $\eta(t)$ is a deterministic function of $\{\mu(s)\}_{0 \leq s \leq t}$. The (relative) value process of η is defined by

$$V_\eta(t) = V_\eta(0) + \sum_{s=0}^{t-1} \eta(s) \cdot (\mu(s+1) - \mu(s)), \tag{12}$$

where $V_\eta(0) := \eta(0) \cdot \mu(0)$ and $a \cdot b$ is the Euclidean inner product.

In this paper all trading strategies are self-financed. Also, we only study the value of portfolios relative to the market portfolio, so for simplicity we may omit the words 'self-financed' and 'relative'. By the identity (11), the value (12) of the portfolio is equal to

$$V_\eta(t) = \eta(t) \cdot \mu(t).$$

Note that because we allow both long and short positions in the portfolio, the value $V_\eta(t)$ may take negative values. The self-financing identity (11) means that all changes in the portfolio value are due to price changes (but not addition or withdrawal of capital).

If the portfolio value $V_\eta(t)$ is strictly positive for all t, we may define the *portfolio weight vector* at time t by

$$\pi(t) = (\pi_1(t), \ldots, \pi_n(t)) = \left(\frac{\eta_1(t)\mu_1(t)}{V_\eta(t)}, \ldots, \frac{\eta_n(t)\mu_n(t)}{V_\eta(t)}\right). \tag{13}$$

The components of $\pi(t)$ represent the percentages of current capital invested in each of the stocks; clearly $\sum_{i=1}^{n} \pi_i(t) \equiv 1$. In this case, the value $V_\eta(t)$ can be expressed *multiplicatively* in the form

$$V_\eta(t) = V_\eta(0) \prod_{s=0}^{t-1} \left(\pi(s) \cdot \frac{\mu(s+1)}{\mu(s)}\right), \tag{14}$$

where $\frac{\mu(s+1)}{\mu(s)} = \left(\frac{\mu_i(s+1)}{\mu_i(s)}\right)_{1 \leq i \leq n}$ is the vector of componentwise ratios. Compare this with the *additive* representation (12). If $\pi_i(t) \geq 0$ for all i and t, we say that the portfolio is *all-long*. It is clear that an adaptive sequence (in the sense of Definition 2) of portfolio weight vectors defines a self-financing all-long trading strategy for each initial value. The *market portfolio* corresponds to $\pi(t) \equiv \mu(t)$.

3 Multiplicatively Generated Portfolio

In this section we review the definition and main results of multiplicative functional generation using the approach of [12]. For simplicity of exposition we assume that the generating functions are smooth.

3.1 Pathwise Decomposition and Relative Arbitrage

Definition 3 (*Multiplicatively generated portfolio*) Let $\varphi : \Delta_n \to \mathbb{R}$ be smooth and exponentially concave, to be called a generating function. Given φ, we define a mapping $\boldsymbol{\pi} : \Delta_n \to \overline{\Delta_n}$, called the portfolio map, by

$$\boldsymbol{\pi}_i(p) = p_i \left(1 + D_{e_i - p}\varphi(p)\right), \quad i = 1, \ldots, n, \tag{15}$$

where $(e_1, \ldots, e_n)$ is the standard Euclidean basis and $D_{e_i - p}$ is the directional derivative along the tangent vector $e_i - p$. It defines an all-long trading strategy η such that the portfolio weight at time t is

$$\pi(t) = \left(\frac{\eta_1(t)\mu_1(t)}{V_\eta(t)}, \ldots, \frac{\eta_n(t)\mu_n(t)}{V_\eta(t)}\right) = \boldsymbol{\pi}(\mu(t)). \tag{16}$$

We say that η (and $\boldsymbol{\pi}$) are generated multiplicatively by φ.

Here is a geometric interpretation of the formula (15). Consider the graph of the positive concave function $\Phi = e^{\varphi}$. Given $p \in \Delta_n$, let the tangent hyperplane to Φ at p be given by $q \mapsto \sum_{i=1}^n c_i q_i$ (see Fig. 3). We have

$$c_i = \Phi(p) + D_{e_i - p}\Phi(p) = \Phi(p)\left(1 + D_{e_i - p}\varphi(p)\right).$$

Since $\sum_{i=1}^n p_i(e_i - p) = 0$, the portfolio vector $\boldsymbol{\pi}(p)$ is given by

$$\boldsymbol{\pi}_i(p) = \frac{c_i p_i}{c_1 p_1 + \cdots + c_n p_n}, \quad i = 1, \ldots, n. \tag{17}$$

In particular, $\boldsymbol{\pi}(p)$ is an element of the closed simplex $\overline{\Delta_n}$ (so the portfolio is all-long), and the weight ratio $\boldsymbol{\pi}_i(p)/p_i$ is proportional to c_i. We say that the trading strategy is generated *multiplicatively* because the generating function φ specifies the weight ratios through its derivatives.

The following is the main result about multiplicatively generated portfolios. As this result is fundamental let us give a complete proof which also motivates the definition of the L-divergence. We also note that this proof is more transparent than the original proof (see [3, Theorem 3.1.5] which is formulated in continuous time).

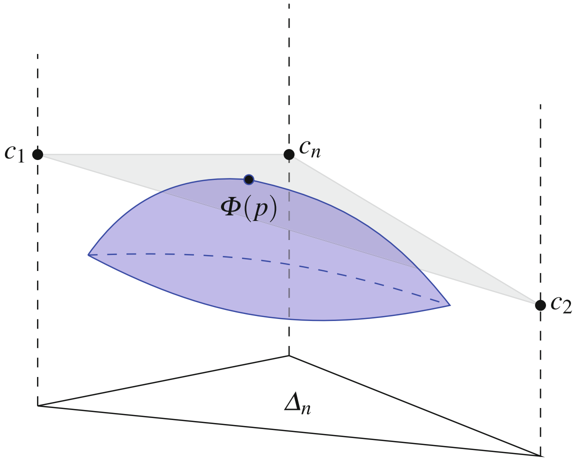


Fig. 3 Geometric interpretation of multiplicatively generated portfolio

Theorem 1 (Multiplicative decomposition [11, 12]) *Let η be the trading strategy generated multiplicatively by the exponentially concave function φ as in Definition 3. Then the value process of η satisfies the decomposition*

$$\log V_\eta(t) - \log V_\eta(0) = \varphi(\mu(t)) - \varphi(\mu(0)) + \sum_{s=0}^{t-1} \mathbf{D}^{(1)}\left[\mu(s+1) \mid \mu(s)\right], \quad (18)$$

where $\mathbf{D}^{(1)}[\cdot \mid \cdot]$ is the $L^{(1)}$-divergence of φ defined by (4).

Proof Using the multiplicative representation (14), we have

$$\frac{V_\eta(s+1)}{V_\eta(s)} = \sum_{i=1}^n \boldsymbol{\pi}_i(\mu(s)) \frac{\mu_i(s+1)}{\mu_i(s)}.$$

From (15), we have

$$\frac{\boldsymbol{\pi}_i(\mu(s))}{\mu_i(s)} = 1 + D_{e_i - \mu(s)}\varphi(\mu(s)),$$

so we get the useful identity

$$\begin{aligned} \frac{V_\eta(s+1)}{V_\eta(s)} &= 1 + \sum_{i=1}^n \mu_i(s+1) D_{e_i - \mu(s)}\varphi(\mu(s)) \\ &= 1 + D_{\mu(s+1)-\mu(s)}\varphi(\mu(s)) \\ &= 1 + \nabla\varphi(\mu(s)) \cdot (\mu(s+1) - \mu(s)). \end{aligned} \quad (19)$$

Here we think of the gradient $\nabla\varphi(\mu(s))$ as operating on tangent vectors of Δ_n. Financially, (19) says that the relative return $(V_\eta(s+1) - V_\eta(s))/V_\eta(s)$ of the portfolio is nothing but the directional derivative of φ.

By the concavity of $\Phi = e^\varphi$, for any $p, q \in \Delta_n$ we have

$$\Phi(p) + \nabla\Phi(p) \cdot (q - p) \geq \Phi(q).$$

Rewriting the inequality in terms of φ and taking logarithm on both sides, we have

$$\mathbf{D}^{(1)}[q \mid p] = \log(1 + \nabla\varphi(p) \cdot (q - p)) - (\varphi(q) - \varphi(p)) \geq 0.$$

Taking logarithm on both sides of (19) and rearranging, we get

$$\log V_\eta(s+1) - \log V_\eta(s) = \varphi(\mu(s+1)) - \varphi(\mu(s)) + \mathbf{D}^{(1)}[\mu(s+1)|\mu(s)].$$

Summing over time gives the decomposition (18).

From (18), the performance of the portfolio relative to the market can be attributed to two quantities. The first is the change in the market diversity $\varphi(\mu)$. It depends only on the beginning location $\mu(0)$ and the current location $\mu(t)$ of the market. Note that a change in $\varphi(\mu(t))$ is only caused by the component of market movement along the direction of $\nabla\varphi(\mu(t))$ which is perpendicular to the level set of φ. In particular, displacement along the same level set is not visible in this first term. The second term in (18) measures the volatility of the market, as it travels from $\mu(0)$ to $\mu(t)$, by the sum of $\mathbf{D}^{(1)}\,[\mu(s+1) \mid \mu(s)]$ over time. Intuitively, the functionally generated trading strategy η outperforms the market if and only if the volatility is greater than the change in market diversity. In SPT, this decomposition allows one to formulate conditions under which relative arbitrage (with respect to the market portfolio) exists. Here is the simplest version of this idea:

Proposition 1 (Relative arbitrage) *Fix a smooth, exponentially concave function $\varphi : \Delta_n \to \mathbb{R}$. Let $M > 0$ be a constant and let $T > 0$ be a finite time horizon. We say that a market weight sequence $\{\mu(t)\}_{t=0}^\infty$ satisfies property $\mathscr{P}$ if $\varphi(\mu(t)) - \varphi(\mu(0)) > -M$ for all t and $\sum_{s=0}^{T-1} \mathbf{D}^{(1)}[\mu(s+1)|\mu(s)] > M$, where $\mathbf{D}^{(1)}$ is the $L^{(1)}$-divergence of φ. Then there exists an all-long trading strategy η such that $V_\eta(T)/V_\eta(0) > 1$ (i.e., the strategy outperforms the market over the horizon $[0, T]$) for all market weight sequences satisfying property $\mathscr{P}$.*

Proof Let η be the trading strategy generated multiplicatively by φ. By Theorem 1, if the market weight sequence satisfies property $\mathscr{P}$, we have

$$\begin{aligned}\log V_\eta(T) - \log V_\eta(0) &= \varphi(\mu(t)) - \varphi(\mu(0)) + \sum_{s=0}^{t-1} \mathbf{D}^{(1)}\,[\mu(s+1) \mid \mu(s)] \\ &> -M + M = 0.\end{aligned}$$

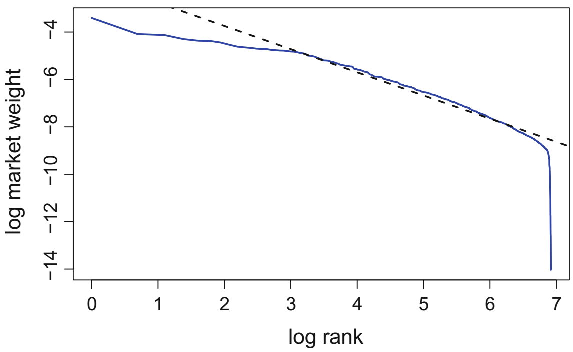


Fig. 4 Capital distribution of the Russel 1000 Index in June 2015 (taken from [25])

The proof of Proposition 1 is almost trivial because we already have the concept of multiplicatively generated portfolio. Without knowing this construction, it is not immediate why a relative arbitrage exists and how it is constructed. The usefulness of this result comes from the following observations (see [3]). Consider the *capital distribution* of the market defined by the reversed order statistics of the components of $\mu(t)$:

$$\mu_{(1)}(t) \geq \mu_{(2)}(t) \geq \cdots \geq \mu_{(n)}(t).$$

Empirically, it is found that if one plots $\log \mu_{(k)}(t)$ against $\log k$ (log of the rank), one gets an approximately linear curve (except the tail) which is relatively stable over time (see Fig. 4). This means that the capital distribution has an approximate Pareto distribution, and the market diversity $\varphi(\mu(t))$ is mean-reverting for φ suitably chosen. On the other hand, for a typical φ the market volatility $\sum_{s=0}^{T-1} \mathbf{D}[\mu(s+1)|\mu(s)]$ grows roughly linearly in time [26]. Thus, it appears that the market satisfies the conditions of (1).

The stability of the capital distribution has inspired many works on the construction and analysis of market models that exhibit such behaviors. Mathematically, these are systems of Brownian particles (representing the market capitalizations) where the drift and volatility coefficients depend on their relative rankings, and so they are called *rank-based models*. For more details we refer the reader to the papers [27–33] and their references.

Proposition 1 only addresses long term relative arbitrages. In practice, *short term* relative arbitrages are much more relevant and interesting. Naturally their constructions require more work and conditions (see for example [25, 34–36]). In particular, the paper [5] proves that market volatility alone does not imply the existence of short term relative arbitrage (this problem had been open in SPT for more than 10 years).

3.2 Examples

We give some examples of exponentially concave functions on Δ_n, the portfolios they generate as well as the corresponding $L^{(1)}$-divergences.

Example 1 (*Constant-weighted portfolio*) Fix a probability vector $\pi \in \overline{\Delta_n}$ and consider the function

$$\varphi(p) = \sum_{j=1}^{n} \pi_j \log p_j.$$

It is exponentially concave since $\Phi = e^{\varphi} = (p_1)^{\pi_1} \cdots (p_n)^{\pi_n}$, the geometric mean with weights $\pi_1, \ldots, \pi_n$, is concave on Δ_n. This function generates the constant-weighted portfolio $\boldsymbol{\pi}(p) \equiv \pi$, and the $L^{(1)}$-divergence is the excess growth rate $\mathbf{T}_{\pi}\left[q \mid p\right]$ given by (5). We also observe that

$$\begin{aligned} \varphi(\mu(t)) - \varphi(\mu(0)) &= \sum_{j=1}^{n} \pi_j \log \frac{\pi_j}{\mu_j(0)} - \sum_{j=1}^{n} \pi_j \log \frac{\pi_j}{\mu_j(t)} \\ &= H\left(\pi \mid \mu(0)\right) - H\left(\pi \mid \mu(t)\right) \end{aligned}$$

is the negative of the change in the relative entropy $H\left(\pi \mid \cdot\right)$. In [23], we call the decomposition (18) for this portfolio the *energy-entropy decomposition*.

Example 2 (*Diversity-weighted portfolio*) For $\lambda \in (0, 1)$ fixed, let φ be the function

$$\varphi(p) = \frac{1}{\lambda} \log \sum_{j=1}^{n} (p_j)^{\lambda},$$

which is the logarithm of the function Φ given by (2). Then φ is exponentially concave and generates the *diversity-weighted portfolio* where

$$\boldsymbol{\pi}_i(p) = \frac{(p_i)^{\lambda}}{\sum_{j=1}^{n} (p_j)^{\lambda}}, \quad i = 1, \ldots, n. \tag{20}$$

Note that this portfolio interpolates between the equal-weighted portfolio $\boldsymbol{\pi}(p) \equiv \overline{e}$ (when $\lambda \downarrow 0$) and the market portfolio $\boldsymbol{\pi}(p) \equiv p$ (when $\lambda \uparrow 1$). See [37] where the diversity-weighted portfolio is studied for negative values of λ. We remark that these cover, except for the log case, portfolios constructed using *Tukey's transformation ladder*; for more details see [38] where an extensive empirical study is given.

For $\lambda \in (0, 1)$ fixed, let $p^{(\lambda)} \in \Delta_n$ be given by $\boldsymbol{\pi}(p)$ as in (20). In information geometry, $p^{(\lambda)}$ is called the λ-*escort distribution* corresponding to the distribution p (see [10, Section 4.3]). Using this terminology we may interpret the portfolio in terms of the Rényi entropy and divergence.

Proposition 2 *Let* $0 < \lambda < 1$. *For* $p \in \Delta_n$, *we have*

$$\varphi(p) = \frac{1}{\lambda} \log \left(\sum_{j=1}^{n} (p_j)^{\lambda} \right) = (\alpha - 1)\mathbf{H}_{\alpha}(p^{(\lambda)}),$$

where $\alpha := \frac{1}{\lambda} \in (1, \infty)$, $p^{(\lambda)}$ *is the* λ*-escort distribution corresponding to* p, *and*

$$\mathbf{H}_{\alpha}(r) := \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{n} (r_j)^{\alpha} \right)$$

is the Rényi entropy of order α.

Moreover, the $L^{(1)}$*-divergence of* φ *is given by*

$$\mathbf{D}^{(1)}[q \mid p] = (\alpha - 1)\mathbf{D}_{\alpha}(q^{(\lambda)}||p^{(\lambda)}), \tag{21}$$

where $\mathbf{D}_{\alpha}(\cdot||\cdot)$ *is the Rényi divergence of order* α *defined by*

$$\mathbf{D}_{\alpha}(p||q) := \frac{1}{\alpha - 1} \log \left(\sum_{j=1}^{n} (p_j)^{\alpha} (q_j)^{1-\alpha} \right). \tag{22}$$

Proof This is a direct computation and we only give the proof of the first statement. Using the fact that $p = \left(p^{(\lambda)}\right)^{(1/\lambda)}$, we have

$$\begin{aligned}
\frac{1}{\lambda} \log \left(\sum_{i=1}^{n} p_j^{\lambda} \right) &= \frac{1}{\lambda} \log \left(\sum_{i=1}^{n} \left(\frac{\left(p_i^{(\lambda)}\right)^{1/\lambda}}{\sum_{j=1}^{n} \left(p_j^{(\lambda)}\right)^{1/\lambda}} \right)^{\lambda} \right) \\
&= -\log \left(\sum_{j=1}^{n} \left(p_j^{(\lambda)}\right)^{\frac{1}{\lambda}} \right) \\
&= (\alpha - 1) \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{n} \left(p_j^{(\lambda)}\right)^{\alpha} \right),
\end{aligned}$$

which is $\alpha - 1$ times the Rényi entropy.

Corollary 1 *Let* $\lambda \in (0, 1)$ *and* $\alpha := \frac{1}{\lambda} \in (1, \infty)$. *The relative value of the diversity-weighted portfolio with parameter* λ *satisfies*

$$\begin{aligned}
&\log V_{\eta}(t) V_{\eta}(0) \\
&= (\alpha - 1) \left[\mathbf{H}_{\alpha}\left(\mu^{(\lambda)}(t)\right) - \mathbf{H}_{\alpha}\left(\mu^{(\lambda)}(0)\right) + \sum_{s=0}^{t-1} \mathbf{D}_{\alpha}\left(\mu^{(\lambda)}(t+1)\big|\big|\mu^{(\lambda)}(t)\right) \right].
\end{aligned}$$

Consider a portfolio manager who tries to optimize over the parameter λ of a diversity-weighted portfolio. By Corollary 1, this means comparing the dynamics of the market weight $\mu(t)$ using different λ-escort geometries of the simplex. In particular, it is well-known that the Rényi divergence satisfies

$$\mathbf{D}_{\alpha}\left(r + tv || r\right) = \frac{\alpha}{2} t^2 \|v\|_r^2 + O(|t|^3), \quad t \to 0, \tag{23}$$

where $\|v\|_r^2 := \sum_{i=1}^n v_i^2 / r_i$ is the Fisher information metric at the point r. This geometric viewpoint may lead to new statistical methods and algorithms. In this regard, let us mention the recent work [39] which studies the dynamics of market diversity in the context of large rank-based models, as well as the paper [40] which proposes a model for predicting change in market diversity. More generally, optimization of functionally generated portfolio amounts to finding the geometry in which $\mu(t)$ has the least change in diversity and has the greatest cumulated volatility.

3.3 Multiplicative Cyclical Monotonicity

In this subsection we provide a financial argument, given in [12], that motivates the definition of multiplicatively generated portfolio.

Let us restrict to all-long trading strategies defined by portfolio maps, i.e., the portfolio weights satisfies $\pi(t) = \boldsymbol{\pi}(\mu(t))$ where $\boldsymbol{\pi} : \Delta_n \to \overline{\Delta_n}$ is a fixed deterministic function. When is $\boldsymbol{\pi}$ able to profit from market volatility? Intuitively it should satisfy the following property. Let O be a (small) neighborhood in the simplex Δ_n, and suppose $\mu(t) \in O$ for all t. From the discussion in Sect. 1.1 the capital distribution is stable. Then, we expect that the portfolio will outperform the market asymptotically as long as there is enough volatility. Specifically, it should outperform the market whenever it is periodic. This idea leads to the following definition.

Definition 4 (*multiplicative cyclical monotonicity (MCM)*) A portfolio map $\boldsymbol{\pi} : \Delta_n \to \overline{\Delta_n}$ is multiplicatively cyclical monotone if for any integer $m \geq 0$ and any cycle $\{\mu(t)\}_{t=0}^m$ with $\mu(0) = \mu(m)$ we have $V_\eta(m) \geq 1$, i.e.,

$$\prod_{t=0}^{m-1} \left(\boldsymbol{\pi}(\mu(t)) \cdot \frac{\mu(t+1)}{\mu(t)} \right) \geq 1. \tag{24}$$

In [12] we observed that this property characterizes the multiplicatively generated portfolio. The following result is the multiplicative analogue of Rockafellar's theorem which characterizes the subdifferentials of convex functions in terms of cyclical monotonicity [41, Section 24].

Theorem 2 *Suppose the portfolio map $\boldsymbol{\pi} : \Delta_n \to \overline{\Delta_n}$ is continuous. Then it is multiplicatively cyclical monotone if and only if there exists a differentiable, exponentially concave function $\varphi : \Delta_n \to \mathbb{R}$ which generates $\boldsymbol{\pi}$ in the sense of* (15).

Proof Let us provide a sketch of proof to illustrate the main idea. Continuity of $\boldsymbol{\pi}$ is included here only to simplify the statement (in the general case φ is not necessarily differentiable and we need to use supergradients). Suppose $\boldsymbol{\pi}$ is generated by φ. Consider a market weight sequence with $\mu(m) = \mu(0)$. By the decomposition (18), we have

$$\log V_\eta(m) - \log V_\eta(0) = \sum_{t=0}^{m-1} \mathbf{D}^{(1)}\left[\mu(t+1) \mid \mu(t)\right] \geq 0.$$

Thus $V_\eta(m) \geq 1$ and $\boldsymbol{\pi}$ is MCM.

Conversely, suppose that $\boldsymbol{\pi}$ is MCM. Consider the function φ defined by

$$\begin{aligned}\varphi(p) &= \varphi(p_0) + \inf\left\{\log V_\eta(t) - \log V_\eta(0)\right\}\\ &= \varphi(p_0) + \inf\left\{\sum_{s=0}^{t-1} \log\left(\sum_{i=1}^{n} \boldsymbol{\pi}_i(\mu(s)) \frac{\mu_i(s+1)}{\mu_i(s)}\right)\right\},\end{aligned}$$

where $p_0 \in \Delta_n$ is fixed, $\varphi(p_0) \in \mathbb{R}$ is arbitrary, and the infimum is taken over $t \geq 0$ and all market weight sequences $\{\mu(s)\}_{s=0}^{t}$ for which $\mu(0) = p_0$ and $\mu(t) = p$. Then it can be shown that φ is differentiable, exponentially concave, and generates the given portfolio map $\boldsymbol{\pi}$. It can also be shown that the function φ is unique up to an additive constant.

Using this characterization, in [12] we introduced a Monge-Kantorovich optimal transport problem and showed that the optimal coupling can be represented using exponentially concave functions and the portfolios they generate.

4 Generalized Functional Portfolio Generation

4.1 Motivations

As it turns out, Theorem 1 is not the only way to generate a portfolio such that a pathwise decomposition holds. In [15] Karatzas and Ruf introduced a novel *additive* generation and used it to construct relative arbitrages (see [42] for another extension involving an additional finite variation process). The following result uses the terminology of [43, Section 3.3] and adapts the construction to our discrete time setting. We omit the proof as it is contained (in the limit) in Theorem 5 below.

Theorem 3 (Additively generated portfolio [15]) *Let $\varphi : \Delta_n \to (0, \infty)$ be a smooth concave function and let $v_0 \in \mathbb{R}$ be an initial portfolio value. Then there is a self-financing trading strategy η satisfying $V_\eta(0) = v_0$ and*

$$\eta_i(t) = D_{e_i - \mu(t)}\varphi(\mu(t)) + V_\eta(t), \quad i = 1, \ldots, n. \tag{25}$$

Its relative value satisfies the decomposition

$$V_\eta(t) - V_\eta(0) = \varphi(\mu(t)) - \varphi(\mu(0)) + \sum_{s=0}^{t-1} \mathbf{D}^{(0)}\left[\mu(t+1) \mid \mu(t)\right], \tag{26}$$

where $\mathbf{D}^{(0)}[\cdot|\cdot]$ *is the Bregman (or* $L^{(0)}$*) divergence of* φ *as in* (6)*. We call* η *the strategy generated additively by* φ*.*

Note that the additively generated portfolio involves a concave rather than exponentially concave function.

Example 3 Consider the function

$$\varphi(p) = \frac{-1}{2}|p|^2 = \frac{-1}{2}\left(p_1^2 + \cdots + p_n^2\right).$$

It generates the trading strategy $\eta(t)$ given by $\eta_i(t) = |p|^2 - p_i + V_\eta(t)$. It is interesting to note that the Bregman divergence of φ is half of the squared Euclidean distance:

$$\mathbf{D}^{(0)}\left[q \mid p\right] = \frac{1}{2}|p-q|^2.$$

Thus the squared Euclidean distance indeed has a financial meaning for this specific trading strategy.

Observe that both the multiplicative and additive decompositions (18) and (26) can be written in the form

$$g(V_\eta(t)) - g(V_\eta(0)) = \varphi(\mu(t)) - \varphi(\mu(0)) + \mathbf{D}\left[\mu(t+1) \mid \mu(t)\right], \tag{27}$$

where g, φ and $\mathbf{D}[\cdot \mid \cdot]$ are suitable functions:

- (Multiplicative generation) $g(x) = \log x$ and $\mathbf{D}[\cdot \mid \cdot]$ is the $L^{(1)}$-divergence of the exponentially concave function φ.
- (Additive generation) $g(x) = x$ and $\mathbf{D}[\cdot \mid \cdot]$ is the $L^{(0)}$-divergence of the concave function φ.

It is natural to ask if there exist other portfolio constructions that admit pathwise decompositions of the form (27). To formulate this question rigorously we introduce the general concept of divergence.

Definition 5 (*Divergence on* Δ_n) A divergence on Δ_n is a non-negative functional $\mathbf{D}[\cdot \mid \cdot] : \Delta_n \times \Delta_n \to [0, \infty)$ satisfying the following conditions:

(i) $\mathbf{D}[q \mid p] = 0$ if and only if $p = q$.
(ii) It admits a quadratic approximation of the form

$$\mathbf{D}\left[p + \Delta p \mid p\right] = \frac{1}{2}\sum_{i,j=1}^{n} g_{ij}(p)\Delta p_i \Delta p_j + O(|\Delta p|^3) \tag{28}$$

as $|\Delta p| \to 0$, and the matrix $G(p) = \big(g_{ij}(p)\big)$ varies smoothly in p and is strictly positive definite in the sense that

$$\sum_{i,j=1}^{n} g_{ij}(p) v_i v_j > 0 \tag{29}$$

for all vectors $v \in \mathbb{R}^n$ that are tangent to Δ_n, i.e., $v_1 + \cdots + v_n = 0$.

If condition (i) is dropped and in (29) we do not strict inequality, we call $\mathbf{D}\,[\cdot \mid \cdot]$ a pseudo-divergence.

Example 4 Let $\operatorname{Hess} \varphi$ denote the Euclidean Hessian of φ. If $\alpha > 0$ and φ is α-exponentially concave, then its $L^{(\alpha)}$-divergence satisfies

$$\begin{aligned} &\mathbf{D}^{(\alpha)}\,[p + \Delta p \mid p] \\ &= \frac{-1}{2}(\Delta p)^{\top} \big(\operatorname{Hess} \varphi(p) + \alpha(\nabla \varphi(p))(\nabla \varphi(p))^{\top}\big)\,(\Delta p) + O(|\Delta p|^3). \end{aligned} \tag{30}$$

If φ is concave, then its $L^{(0)}$-divergence satisfies

$$\mathbf{D}^{(0)}\,[p + \Delta p \mid p] = \frac{-1}{2}(\Delta p)^{\top} \operatorname{Hess} \varphi(p)(\Delta p) + O(|\Delta p|^3).$$

It is easy to verify that the corresponding matrix $G(p)$ is semi-positive definite. They become true divergences if $\operatorname{Hess} e^{\alpha\varphi}$ and $\operatorname{Hess} \varphi$ respectively are strictly positive definite.

Definition 6 (*General functional portfolio construction*) Let $\eta = \{\eta(t)\}_{t=0}^{\infty}$ be a self-financing trading strategy whose relative value process is $\{V_\eta(t)\}$, and let $\varphi, g : \Delta_n \to \mathbb{R}$ be functions on Δ_n where g is strictly increasing. We say that η is generated by φ with scale function g if there exists a pseudo-divergence $\mathbf{D}[\cdot : \cdot]$ on Δ_n such that (27) holds for all market sequences $\{\mu(t)\}_{t=0}^{\infty}$.

In this section we will introduce a new (α, C)-generation, and, after giving an empirical example, show that it characterizes all functional portfolio generation in the sense of Definition 6. For expositional convenience we always assume that the generating function is smooth. Extension of this construction to continuous time is left for future research.

4.2 A New Functional Portfolio Generation

Theorem 4 ((α, C)-generation) *Let $\alpha > 0$ and $C \geq 0$ be fixed parameters, and let $\varphi : \Delta_n \to \mathbb{R}$ be smooth and α-exponentially concave. Then, for any given initial* *value $v_0 \in \mathbb{R}$, there exists a unique self-financing trading strategy $\eta(t)$ such that $V_\eta(0) = v_0$ and*

$$\eta_i(t) = \alpha(C + V_\eta(t))D_{e_i-\mu(t)}\varphi(\mu(t)) + V_\eta(t), \quad i = 1, \ldots, n, \tag{31}$$

for all t. We call η the strategy which is (α, C)-generated by φ.

Proof We will prove by induction on $T \geq 0$ that the statement holds on the time interval $[0, T]$. Consider $T = 0$. Let $V_\eta(0) = v_0$. If we define

$$\eta_i(0) = \alpha(C + v_0)D_{e_i-\mu(0)}\varphi(\mu(t)) + v_0,$$

then, since $\sum_{i=1}^n \mu_i(e_i - \mu) = 0$, we have

$$\eta(0) \cdot \mu(0) = \alpha(C + v_0)\sum_{i=1}^n \mu_i(0)D_{e_i-\mu(0)}\varphi(\mu(t)) + \sum_{i=1}^n \mu_i(0)v_0 = v_0.$$

Thus there is a unique strategy η which satisfies (31) at time 0.

Suppose by the induction hypothesis that the statement holds up to time T. Then $\eta(T)$ and $V_\eta(T)$ are uniquely defined, and the portfolio value at time $T + 1$ is given uniquely by

$$V_\eta(T + 1) := V_\eta(T) + \eta(T) \cdot (\mu(T + 1) - \mu(T)). \tag{32}$$

Thus there is a unique vector $\eta(T + 1)$ satisfying (31).

It remains to show that the strategy is self-financing at time $T + 1$, i.e., $\eta(T) \cdot \mu(T + 1) = \eta(T + 1) \cdot \mu(T + 1)$ (see (11)). Using (31), we have

$$\begin{aligned}
\eta(T + 1) \cdot \mu(T + 1) &= \sum_{i=1}^n \mu_i(T + 1)\alpha(C + V_\eta(T + 1))D_{e_i-\mu(T+1)}\varphi(\mu(T + 1)) \\
&\quad + \sum_{i=1}^n \mu_i(T + 1)V_\eta(T + 1) \\
&= V_\eta(T + 1) \quad (\text{since} \sum_{i=1}^n \mu_i(T + 1)D_{e_i-\mu(T+1)} = 0) \\
&= V_\eta(T) + \eta(T) \cdot (\mu(T + 1) - \mu(T)) \quad (\text{by (32)}) \\
&= \eta(T) \cdot \mu(T) + \eta(T) \cdot (\mu(T + 1) - \mu(T)) \\
&= \eta(T) \cdot \mu(T + 1).
\end{aligned}$$

In the second last equality we used the self-financing property up to time T. This proves that the strategy is uniquely defined at all times.

In Theorem 5 we show that this trading strategy corresponds to the scale function given by

$$g(x) = \frac{1}{\alpha} \log(C + x). \tag{33}$$

Moreover, in Sect. 4.4 we show that up to an additive constant this function (together with $g(x) = x$) is the most general scale function. Comparing (31) with (15) and (25), we see that multiplicative generation corresponds to the case $C = 0$ and $\alpha = 1$, and additive generation corresponds to the limit when $\alpha = \frac{1}{C} \to 0$.

The trading strategy η given by (31) can be interpreted as follows.

Lemma 1 (Portfolio weight of η) *Let $\pi^{(\alpha)}$ be the portfolio process generated multiplicatively by the 1-exponentially concave function $\alpha\varphi$. If $V_\eta(t) > 0$, the portfolio weight vector $\pi(t)$ of the (α, C)-generated trading strategy η is given by*

$$\pi(t) = \left(\frac{\eta_1(t)\mu_1(t)}{V_\eta(t)}, \ldots, \frac{\eta_n(t)\mu_n(t)}{V_\eta(t)} \right) = \frac{C + V_\eta(t)}{V_\eta(t)} \pi^{(\alpha)}(t) - \frac{C}{V_\eta(t)} \mu(t). \tag{34}$$

In particular, $\eta(t)$ longs the multiplicatively generated portfolio $\pi^{(\alpha)}$ and shorts the market portfolio with weights depending on $V_\eta(t)$ and C.

Proof Direct computation using (31).

By increasing C, we may construct portfolios that are more aggressive than the multiplicatively generated portfolio. Note that we keep the parameter α so that we can generate different portfolios with the same generating function φ (as long as $e^{\alpha\varphi}$ is concave).

Next we show that the new portfolio generation admits a pathwise decomposition for the portfolio value.

Theorem 5 (Pathwise decomposition) *Consider an (α, C)-generated trading strategy η as in Theorem 4. If $V_\eta(\cdot) > -C$, then the value process satisfies the pathwise decomposition*

$$\frac{1}{\alpha} \log \frac{C + V_\eta(t)}{C + V_\eta(0)} = \varphi(\mu(t)) - \varphi(\mu(0)) + \sum_{s=0}^{t-1} \mathbf{D}^{(\alpha)} \left[\mu(s+1) \mid \mu(s) \right], \tag{35}$$

where $\mathbf{D}^{(\alpha)}$ is the $L^{(\alpha)}$-divergence of φ.

Proof The proof is similar to that of Theorem 1. By (31), for each time t we have

$$\begin{aligned}
&\frac{1}{\alpha} \log(C + V_\eta(t+1)) - \frac{1}{\alpha} \log(C + V_\eta(t)) \\
&= \frac{1}{\alpha} \log \frac{C + V_\eta(t) + \alpha(C + V_\eta(t))\nabla\varphi(\mu(t)) \cdot (\mu(t+1) - \mu(t))}{C + V_\eta(t)} \\
&= \frac{1}{\alpha} \log\left(1 + \alpha\nabla\varphi(\mu(t)) \cdot (\mu(t+1) - \mu(t))\right) \\
&= \varphi(\mu(t+1)) - \varphi(\mu(t)) + \mathbf{D}^{(\alpha)} \left[\mu(t+1) \mid \mu(t) \right].
\end{aligned} \tag{36}$$

This yields the desired decomposition. The condition $V_\eta(\cdot) > -C$ is imposed so that the logarithms make sense.

4.3 An Empirical Example

Consider a smooth and exponentially concave function φ. It is α-exponentially concave for all $0 < \alpha \leq 1$ and is concave (which corresponds to the case $\alpha \downarrow 0$). Thus both the additive and multiplicatively generated portfolios are well-defined. Unfortunately, while the $L^{(\alpha)}$-divergence is a natural interpolation, there does not seem to be a canonical choice for the constant C that connects the two basic cases.

In this example we consider instead the parameterized family $\{\eta^{(\alpha)}\}_{0\leq\alpha\leq1}$ where $\eta^{(\alpha)}$ is the trading strategy $(\alpha, \frac{1}{\alpha})$-generated by φ (so when $\alpha = 0$ it is the additively generated portfolio), and compare their empirical performance. We set $V_{\eta^{(\alpha)}}(0) = 1$ for all α. Note that $\eta^{(1)}$ is not the multiplicatively generated portfolio as it also shorts the market portfolio.

Consider as in Fig. 1 the (beginning) monthly stock prices of the US companies Ford, Walmart and IBM from January 1990 ($t = 0$) to September 2017 ($t = 332$). We normalize the prices so that at $t = 0$ the market weight is at the barycenter $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. The path of the market weight $\mu(t)$ in the simplex Δ_3 is plotted in Fig. 1 (right).

We consider the 1-exponentially concave function

$$\varphi(p) = \sum_{i=1}^{3} \frac{1}{3} \log p_i \tag{37}$$

which generates multiplicatively the equal-weighted portfolio $\boldsymbol{\pi}(p) \equiv \overline{e} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. By (31), for each $\alpha \in [0, 1]$ the trading strategy is given by

$$\eta_i^{(\alpha)}(t) = \left(1 + \alpha V_\eta(t)\right)\left(\frac{1}{3\mu_i(t)} - 1\right) + V_\eta(t).$$

In terms of portfolio weights, we have

$$\pi^{(\alpha)}(t) = \frac{1 + \alpha V_\eta(t)}{V_\eta(t)}\overline{e} - \frac{1 + \alpha V_\eta(t) - V_\eta(t)}{V_\eta(t)}\mu(t).$$

Thus the portfolio longs more and more the equal-weighted portfolio as α increases. The corresponding $L^{(\alpha)}$-divergence is given by

$$\mathbf{D}^{(\alpha)}\left[q \mid p\right] = \frac{1}{\alpha}\log\left(1 + \alpha\sum_{i=1}^{n}\frac{1}{np_i}(q_i - p_i)\right) - \sum_{i=1}^{n}\frac{1}{n}\log\frac{q_i}{p_i}.$$

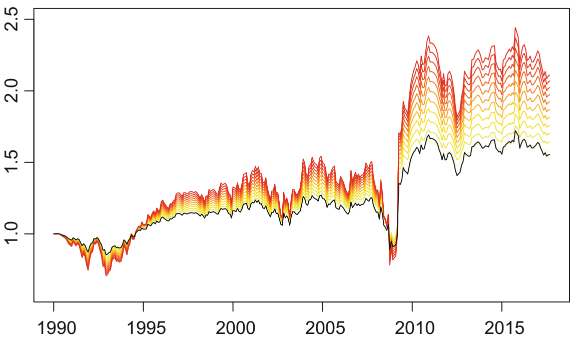


Fig. 5 Time series of the relative portfolio value $V_{\eta^{(\alpha)}}(t)$, from $\alpha = 0$ (yellow) to $\alpha = 1$ (red). The value of the equal-weighted portfolio is shown in black

The relative values of the simulated portfolios are plotted in Fig. 5. At the end of the period the portfolio value is increasing in α, and the additive portfolio ($\alpha = 0$) has the smallest value. It is interesting to note that the reverse is true at the beginning. Note that the values fluctuate widely in the period 2008–2009 corresponding to the financial crisis. For comparison, we also simulate the multiplicatively generated equal-weighted portfolio (i.e., $(\alpha, C) = (1, 0)$) and plot the result in Fig. 5. Surprisingly the additive and multiplicative portfolios have similar behaviors here. In this period, shorting the market by using a positive value for C gives significant advantage over both the additive and multiplicative portfolios. Dynamic optimization over our extended functionally generated portfolios is an interesting problem.

4.4 Characterizing Functional Portfolio Generation

Now we show that our (α, C)-generation is the most general one. Throughout this subsection we let η be a functionally generated trading strategy as in Definition 6. We assume that the scale function g is smooth and $g'(x) > 0$ for all x. We also require that the domain of g contains the positive real line $(0, \infty)$. Furthermore, we assume that φ is smooth, and η is non-trivial in the sense that for all $t \geq 0$ and all market weight paths $\{\mu(s)\}_{s=0}^{t}$ up to time t, the profit-or-loss

$$V_\eta(t+1) - V_\eta(t) = \eta(t) \cdot (\mu(t+1) - \mu(t))$$

is not identically zero as a function of $\mu(t+1) \in \Delta_n$. For technical reasons we also assume that for any $x > 0$, there exists $t \geq 0$ and a sequence $\{\mu(s)\}_{s=0}^{t}$ such that $V_\eta(t) = x$.

Theorem 6 *Under the above conditions, the scale function has one of the following forms. Either*

$$g(x) = c_1 x + c_2 \tag{38}$$

where $c_1 > 0$ and $c_2 \in \mathbb{R}$, or

$$g(x) = c_2 \log(c_1 + x) + c_3 \tag{39}$$

where $c_1 \geq 0$, $c_2 > 0$ and $c_3 \in \mathbb{R}$. In the first case φ is concave and η is additively generated by φ, whereas in the second case φ is α-exponentially concave with $c_2 = \frac{1}{\alpha}$ and η is (α, c_1)-generated by φ. The corresponding pseudo-divergence is the $L^{(\alpha)}$-divergence of φ.

Note that in (38) and (39) the additive constants are irrelevant and may be discarded. We will prove Theorem 6 with several lemmas. First we observe that the decomposition (27) already implies a formula of the trading strategy.

Lemma 2 *For any t and any tangent vector v of Δ_n (i.e., $v_1 + \cdots + v_n = 0$) we have*

$$\eta(t) \cdot v = \frac{1}{g'(V_\eta(t))} \nabla\varphi(\mu(t)) \cdot v. \tag{40}$$

In particular, for each $i = 1, \ldots, n$ we have

$$\eta_i(t) = \frac{1}{g'(V_\eta(t))} D_{e_i - \mu(t)} \varphi(\mu(t)) + V_\eta(t). \tag{41}$$

Proof From (27) we have the identity

$$g(V_\eta(t+1)) - g(V_\eta(t)) = \varphi(\mu(t+1)) - \varphi(\mu(t)) + \mathbf{D}\left[\mu(t+1) \mid \mu(t)\right] \tag{42}$$

which holds for all values of $\mu(t+1)$. Write

$$V_\eta(t+1) = V_\eta(t) + \eta(t) \cdot (\mu(t+1) - \mu(t))$$

and let $\mu(t+1) - \mu(t) = \delta v$, $\delta > 0$ sufficiently small, and compute the first order approximation of both sides of (42). Since $\mathbf{D}[\cdot \mid \cdot]$ is a pseudo-divergence, by (28) its first order approximation vanishes. Evaluating the derivatives and dividing by $\delta > 0$, we obtain (40).

Letting $v = e_i - \mu(t)$ in (41) for $i = 1, \ldots, n$, we get the formula (41).

Observe that (40) reduces to (25) when $g(x) = x$, and to (15) when $g(x) = \log x$. Also, we note that $\eta(t)$ depends only on $\mu(t)$ and the current portfolio value $V_\eta(t)$. Putting $v = \mu(t+1) - \mu(t)$ in (40), we have

$$V_\eta(t+1) - V_\eta(t) = \frac{1}{g'(V_\eta(t))} \nabla\varphi(\mu(t)) \cdot (\mu(t+1) - \mu(t)). \tag{43}$$

Consider the expression

$$\begin{aligned} & g(V_\eta(t+1)) - g(V_\eta(t)) \\ & = g\left(V_\eta(t) + \left[V_\eta(t+1) - V_\eta(t)\right]\right) - g(V_\eta(t)) \\ & = g\left(V_\eta(t) + \frac{1}{g'(V_\eta)} \nabla \varphi(\mu(t)) \cdot (\mu(t+1) - \mu(t))\right) - g(V_\eta(t)). \end{aligned} \tag{44}$$

By (27), this equals

$$\varphi(\mu(t+1)) - \varphi(\mu(t)) + \mathbf{D}[\mu(t+1) | \mu(t)],$$

which is a function of $\mu(t)$ and $\mu(t+1)$ only. Thus, the expression in (44) does not depend on the current portfolio value $V_\eta(t)$. From this observation we will derive a differential equation satisfied by g.

Lemma 3 *The scale function g satisfies the third order nonlinear ODE*

$$g' g''' = 2(g'')^2 \tag{45}$$

on the positive real line $(0, \infty)$.

Proof Given $x > 0$, write $x = V_\eta(t)$ for some $t \geq 0$ and market sequence $\{\mu(s)\}_{s=0}^t$. Let $\delta = \nabla \varphi(\mu(t)) \cdot (\mu(t+1) - \mu(t))$. From (44), for any δ, the expression

$$g(x + \frac{1}{g'(x)} \delta) - g(x) \tag{46}$$

does not depend on x.

Differentiating (46) with respect to x, we have

$$g'(x + \frac{1}{g'(x)}\delta)\left(1 - \delta \frac{g''(x)}{(g'(x))^2}\right) - g'(x) = 0.$$

Next we differentiate with respect to δ (since η is assumed to be non-trivial, this can be done by varying $\mu(t+1)$):

$$g''(x + \frac{1}{g'(x)}\delta)\frac{1}{g'(x)}\left(1 - \delta \frac{g''(x)}{(g'(x))^2}\right) + g'(x + \frac{1}{g'(x)}\delta)\frac{-g''(x)}{(g'(x))^2} = 0.$$

Differentiating one more time with respect to δ, we have

$$\begin{aligned} & g'''(x + \frac{1}{g'(x)}\delta)\frac{1}{(g'(x))^2}\left(1 - \delta \frac{g''(x)}{(g'(x))^2}\right) \\ & \quad + g''(x + \frac{1}{g'(x)}\delta)\frac{-g''(x)}{(g'(x))^3} - \frac{g''(x + \frac{1}{g'(x)}\delta) g''(x)}{(g'(x))^3} = 0. \end{aligned}$$

Setting $\delta = 0$, we get $\frac{g'''(x)}{(g'(x))^2} - 2\frac{(g''(x))^2}{(g'(x))^3} = 0$ which gives the ODE (45). (Note that we assumed that $g'(x) > 0$ for all x.)

With the differential equation in hand, it is not difficult to find the general solutions. They can be verified using direct substitution and the local uniqueness of the autonomous equation.

Lemma 4 *All solutions to the ODE (45) can be written in the form*

$$g(x) = c_0 + c_1 x \quad \textit{or} \quad g(x) = c_2 \log(c_1 + x) + c_3, \tag{47}$$

where the c_i's are real constants. The constraints on the constants stated in Theorem 6 follow from our assumptions of g.

Now we are ready to complete the proof of Theorem 6. By Lemma 4, the scale function (up to an additive constant which is irrelevant) has the form (38) or (39). Consider the second case (the first case is similar). Then by (43), (44) and the third equality of (36) (which does not depend on α-exponential concavity of φ), for any $p = \mu(s)$ and $q = \mu(s+1)$, we have

$$\mathbf{D}\left[q \mid p\right] = \frac{1}{\alpha} \log\left(1 + \alpha \nabla\varphi(p) \cdot (q - p)\right) - (\varphi(q) - \varphi(p))$$

which is exactly the expression of the $L^{(\alpha)}$-divergence. By assumption $\mathbf{D}\left[\cdot \mid \cdot\right]$ is a pseudo-divergence so it is non-negative for all p, q. It is easy to check that this implies that φ is α-exponentially concave, and so η is the (α, c_1)-generated trading strategy.

5 Further Properties of L-divergence

In this section we gather some further properties of L-divergence and describe some related problems. For simplicity we focus on the $L^{(1)}$-divergence, and refer the reader to [14] for a systematic study of the $L^{(\alpha)}$-divergence. We always assume that the generating functions are smooth. It is clear that some of the problems make sense on domains other than the unit simplex.

5.1 Interpolation and Comparison

If $\varphi^{(0)}$ and $\varphi^{(1)}$ are exponentially concave functions on Δ_n, by the inequality of the arithmetic and geometric means, we have that

$$\varphi^{(\lambda)} := (1 - \lambda)\varphi^{(0)} + \lambda\varphi^{(1)}$$

is exponentially concave for any $0 < \lambda < 1$. If $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\pi}^{(1)}$ are the portfolio maps generated multplicatively by $\varphi^{(0)}$ and $\varphi^{(1)}$ respectively, then $\varphi^{(\lambda)}$ generates the portfolio map

$$\boldsymbol{\pi}^{(\lambda)}(\cdot) \equiv (1-\lambda)\boldsymbol{\pi}^{(0)}(\cdot) + \lambda\boldsymbol{\pi}^{(1)}(\cdot),$$

which is a constant-weighted portfolio of $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\pi}^{(1)}$ [24, Lemma 4.4]. Thus, the spaces of exponentially concave functions and MCM portfolio maps (see Definition 3) are convex. Moreover, the L-divergence $\mathbf{D}^{(1)}[\cdot \mid \cdot]$ is concave in the function φ. In [13], this interpolation provides a new displacement interpolation for a logarithmic optimal transport problem.

Let $\mathbf{D}_\varphi$ and $\mathbf{D}_\psi$ be the L-divergences generated respectively by the exponentially concave functions φ and ψ, we say that $\mathbf{D}_\psi$ *dominates* $\mathbf{D}_\varphi$ if

$$\mathbf{D}^{(1)}_\psi[q \mid p] \geq \mathbf{D}^{(1)}_\varphi[q \mid p] \tag{48}$$

for all $p, q \in \Delta_n$. Financially, this means that the portfolio generated by ψ captures more volatility than the one generated by φ. An interesting problem is to find the maximal elements in this partial order, and the following result is obtained in [24] using the relative convexity lemma in [44].

Theorem 7 *Suppose φ is symmetric, i.e., $\varphi(p_1, \ldots, p_n) = \varphi(p_{\sigma(1)}, \ldots, p_{\sigma(n)})$ for any permutation σ of the coordinates. If*

$$\int_0^1 e^{-2\varphi((1-t)e_1 + t\overline{e})} dt = \infty,$$

then $\mathbf{D}_\varphi$ is maximal in the partial order (48): if $\mathbf{D}_\varphi$ is dominated by $\mathbf{D}_\psi$, then $\varphi - \psi$ is constant on Δ_n and $\mathbf{D}_\psi \equiv \mathbf{D}_\varphi$.

As an example, the function $\varphi(p) = \frac{1}{n}\sum_{i=1}^n \log p_i$ (which generates the equal-weighted portfolio) is maximal. Another example is $\varphi(p) = \log\left(-\sum_{i=1}^n p_i \log p_i\right)$, the logarithm of the Shannon entropy.

Note that Theorem 7 is concerned with the global properties of the generating function. One can also study maximal exponentially concave functions over a local neighborhood; this idea is used in [18] to construct short term relative arbitrages.

5.2 *Dualistic Geometry and the Generalized Pythagorean Theorem*

Consider the $L^{(1)}$-divergence $\mathbf{D}[\cdot \mid \cdot]$ of an exponentially concave function φ on the simplex Δ_n. It defines a Riemannian metric g (as in (30)) and a dual pair of torsion-free affine connections (∇, ∇^*) (see [10, Chapter 6]). In [13] this geometry is derived and many interesting properties are shown. It generalizes the dually flat geometry of Bregman divergence (a unified framework is established recently in [14]).

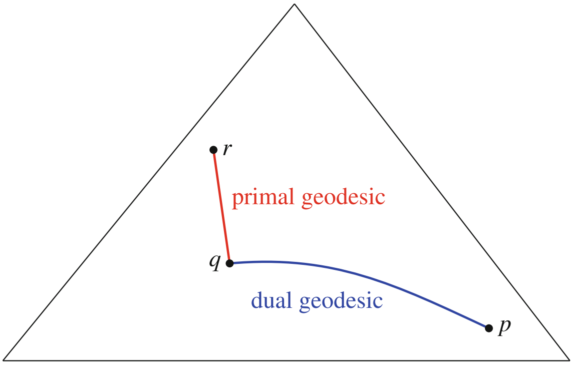


Fig. 6 Generalized Pythagorean theorem for L-divergence

As before we let p denote a generic element of Δ_n. Now we regard this a global coordinate system of the manifold $M = \Delta_n$, and call it the *primal* coordinate system. Let $\boldsymbol{\pi}$ be the portfolio map generated by φ. It defines a *dual* coordinate system

$$p^* := \left(\frac{\pi_1(p)/p_1}{\sum_{j=1}^n \pi_j(p)/p_j}, \ldots, \frac{\pi_n(p)/p_n}{\sum_{j=1}^n \pi_j(p)/p_j} \right)$$

which also takes values in the unit simplex. The main properties of the geometry are summarized in the following theorem, and we refer the reader to [13, 14] for further properties related to the geodesic equations, gradient flows and connections with optimal transport.

Theorem 8 [13] *Consider the dualistic geometry induced by* $\mathbf{D}[\cdot \mid \cdot]$.

(i) The trace of a primal geodesic is a straight line under the primal coordinate system.
(ii) The trace of a dual geodesic is a straight line under the dual coordinate system.
(iii) The geometry has constant (primal and dual) sectional curvature -1 *(when* $n \geq 3$*) with respect to the induced Riemannian metric.*

In particular, the induced geometry is dually *projectively* flat but not flat. Furthermore, the $L^{(1)}$-divergence satisfies a generalized Pythagorean theorem.

Theorem 9 (Generalized Pythagorean theorem) *Given* $(p, q, r) \in (\Delta_n)^3$, *consider the dual geodesic joining* q *and* p *and the primal geodesic joining* q *and* r. *Consider the Riemannian angle between the geodesics at* q. *Then the difference*

$$\mathbf{D}[q \mid p] + \mathbf{D}[r \mid q] - \mathbf{D}[r \mid p] \tag{49}$$

is positive, zero or negative depending on whether the angle is less than, equal to, or greater than 90 *degrees (see Fig. 6).*

Further properties of the Pythagorean theorem can be studied. To give a flavor we present an interesting result for the excess growth rate $\mathbf{T}_\pi[\cdot \mid \cdot]$ defined by (5).

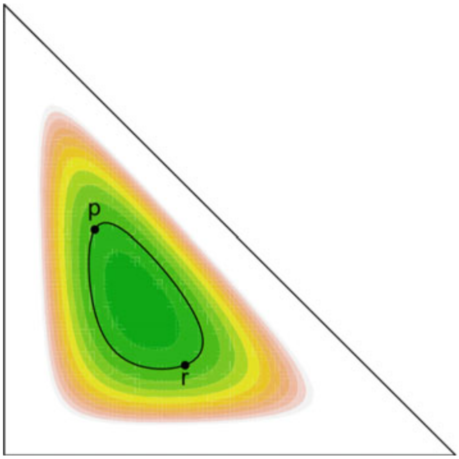


Fig. 7 The level sets of the map $q \mapsto \mathbf{T}_\pi [q \mid p] + \mathbf{T}_\pi [r \mid q] - \mathbf{T}_\pi [r \mid p]$ where p and r are fixed. Here $\pi = \overline{e}$ is the equal-weighted portfolio

Proposition 3 *Consider the excess growth rate $\mathbf{T}_\pi [\cdot \mid \cdot]$ for a fixed portfolio weight vector $\pi \in \overline{\Delta_n}$. For $p, r \in \Delta_n$ fixed, the mapping*

$$q \in \Delta_n \mapsto f(q) := \mathbf{T}_\pi [q \mid p] + \mathbf{T}_\pi [r \mid q] - \mathbf{T}_\pi [r \mid p] \tag{50}$$

is quasiconvex, i.e, the sublevel sets $\{q : f(q) \leq \lambda\}$ are convex (see Fig. 7).

Proof It suffices to show that the map

$$g(x) := \log\left(\pi \cdot \frac{x}{p}\right) + \log\left(\pi \cdot \frac{r}{x}\right)$$

is quasiconvex on Δ_n. We will use the following characterization of quasiconvex functions (see [45, Section 3.4.3]): g is quasiconvex if and only if

$$g(y) \leq g(x) \Rightarrow \nabla g(x) \cdot (y - x) \leq 0. \tag{51}$$

for any x and y.

Let $x, y \in \Delta_n$ be such that $g(y) \leq g(x)$. We have

$$\partial_i g(x) = \frac{\frac{\pi_i}{p_i}}{\pi \cdot \frac{x}{p}} + \frac{-\frac{\pi_i r_i}{x_i^2}}{\pi \cdot \frac{r}{x}}.$$

After some simplifications, we have

$$
\begin{aligned}
\nabla g(x) \cdot (y - x) &= \sum_{i=1}^{n} \left(\frac{\frac{\pi_i}{p_i}}{\pi \cdot \frac{x}{p}} + \frac{-\frac{\pi_i r_i}{x_i^2}}{\pi \cdot \frac{r}{x}} \right) (y_i - x_i) \\
&= \frac{\pi \cdot \frac{y}{p}}{\pi \cdot \frac{x}{p}} - \frac{\sum_{i=1}^{n} \pi_i \frac{r_i}{x_i} \frac{y_i}{x_i}}{\pi \cdot \frac{r}{x}}.
\end{aligned} \tag{52}
$$

Since $g(y) \leq g(x)$, we have

$$
\frac{\pi \cdot \frac{x}{p}}{\pi \cdot \frac{x}{p}} \leq \frac{\pi \cdot \frac{r}{x}}{\pi \cdot \frac{r}{y}}.
$$

Substituting this into (52), we get

$$
\begin{aligned}
\nabla g(x) \cdot (y - x) &\leq \frac{\pi \cdot \frac{r}{x}}{\pi \cdot \frac{r}{y}} - \frac{\sum_{i=1}^{n} \pi_i \frac{r_i}{x_i} \frac{y_i}{x_i}}{\pi \cdot \frac{r}{x}} \\
&= \frac{1}{\left(\pi \cdot \frac{r}{x}\right)\left(\pi \cdot \frac{r}{y}\right)} \left[\left(\pi \cdot \frac{r}{x}\right)^2 - \sum_{i,j=1}^{n} \pi_i \pi_j r_i r_j \frac{1}{x_i x_j} \frac{y_j}{x_j} \frac{x_i}{y_i} \right] \\
&= \frac{1}{\left(\pi \cdot \frac{r}{x}\right)\left(\pi \cdot \frac{r}{y}\right)} \sum_{i,j=1}^{n} \frac{\pi_i r_i}{x_i} \frac{\pi_j r_j}{x_j} \left(1 - \frac{x_i}{y_i} \frac{y_j}{x_j}\right).
\end{aligned} \tag{53}
$$

Let $A = \sum_{i=1}^{n} \frac{\pi_i r_i}{x_i}$ and let $\alpha_i = \frac{\pi_i r_i}{x_i} / A$. Note that α is a probability vector. Now we may write (53) in the form

$$
C \left(1 - \sum_{i,j=1}^{n} \alpha_i \alpha_j \frac{x_i}{y_i} \frac{y_j}{x_j} \right),
$$

where $C > 0$ is a constant. Let X and Y be independent and identically distributed random variables such that

$$
\mathbb{P}\left(X = \frac{x_i}{y_i}\right) = \mathbb{P}\left(Y = \frac{x_i}{y_i}\right) = \alpha_i, \quad i = 1, \ldots, n.
$$

By Jensen's inequality, we have

$$
\sum_{i,j=1}^{n} \alpha_i \alpha_j \frac{x_i}{y_i} \frac{y_j}{x_j} = \mathbb{E}\left[X \cdot \frac{1}{Y}\right] = \mathbb{E}[X]\mathbb{E}\left[\frac{1}{Y}\right] \geq \mathbb{E}[X] \frac{1}{\mathbb{E}[X]} = 1.
$$

Thus $\nabla g(\mathbf{x}) \cdot (y - x) \leq 0$ and we have proved that g is quasiconvex.

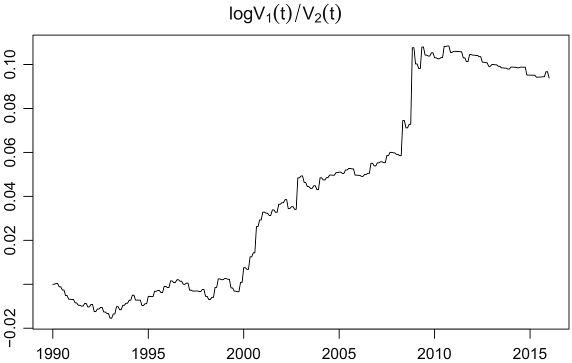


Fig. 8 Relative performance of the portfolio V_1 (rebalanced every month) versus V_2 (rebalanced every two months)

5.3 *Optimal Rebalancing Frequency*

Finally we mention a practical problem related to the L-divergence. The generalized Pythagorean theorem gives a geometric way to study the rebalancing frequency of a functionally generated portfolio. To give a simple example, consider the empirical example as in Fig. 1 and the equal-weighted portfolio $\boldsymbol{\pi}(p) \equiv \overline{e} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

Consider two ways of implementing this portfolio: (i) rebalance every month; (ii) rebalance every two months. If V_1 and V_2 denote respectively the values of these implementations, Fig. 8 plots the time series of the relative performance $\log V_1(t)/V_2(t)$. For this data set, rebalancing every month boost the return by about 10% in log scale. Using the decomposition (18), we see that if T is the terminal time, then

$$\log \frac{V_1(T)}{V_2(T)}$$
$$= \sum_k \left(\mathbf{D}\left[\mu(2k+1) \mid \mu(2k)\right] + \mathbf{D}\left[\mu(2k+1) \mid \mu(2k+2)\right] - \mathbf{D}\left[\mu(2k+2) \mid \mu(k)\right]\right).$$

By Theorem 9, the sign of each term is determined by the Riemannian angle of the geodesic triangle. This angle summarizes in a single number the correlation among the stock returns that is relevant to the rebalancing frequency. Further work should study the joint relationship between the angle and the *size* of the geodesic triangle which determines the magnitude of (49).

In practice trading incurs transaction costs which have been neglected so far. Transaction costs create a drag of the portfolio value. A common setting is that the transaction cost is *proportional*, i.e., we pay a fixed percentage of the value exchanged. In our model, the transaction cost may be approximated by a functional $\mathbf{C}\left[q \mid p\right] \geq 0$ where p and q are the beginning and ending market weights of the holding period (Fig. 9). Since the transaction cost is proportional, the cost is of linear order when $q \approx p$. On the other hand, the L-divergence is approximately quadratic when $q \approx p$. Thus the net difference $\mathbf{D}\left[q \mid p\right] - \mathbf{C}\left[q \mid p\right]$ is negative when q is sufficiently close to p. Financially, this means that the investor should not rebalance

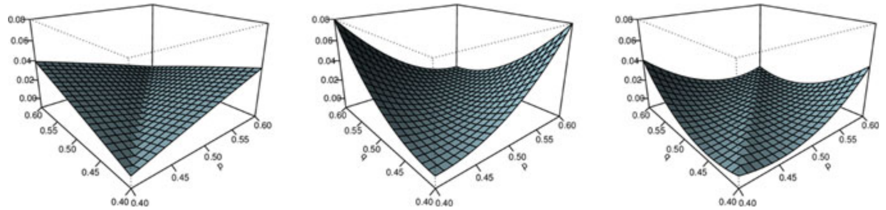


Fig. 9 Left: Transaction cost $\mathbf{C}\,[q \mid p]$. Middle: The L-divergence $\mathbf{D}\,[q \mid p]$. Right: The difference $\mathbf{D} - \mathbf{C}$. In these figures $n = 2$ and the x and y-axes are the first coordinates of p and q respectively

too often – at least when the increments of the market weights are 'small'. We end this paper with the following problem:

Problem 1 Design a robust strategy for rebalancing a given trading strategy.

References

1. Amari, S., Nagaoka, H., Harada, D.: Methods of Information Geometry (Translations of Mathematical Monographs). American Mathematical Society, Providence (2002)
2. Wong, T.-K.L.: On portfolios generated by optimal transport. arXiv preprint arXiv:1709.03169 (2017)
3. Fernholz, E.R.: Stochastic Portfolio Theory. Springer, Berlin (2002)
4. Fernholz, E.R., Karatzas, I.: Stochastic portfolio theory: an overview. In: Ciarlet, P.G. (ed.) Handbook of Numerical Analysis, vol. 15, pp. 89–167. Elsevier, Amsterdam (2009)
5. Fernholz, E.R., Karatzas, I., Ruf, J.: Volatility and arbitrage. Ann. Appl. Probab. **28**(1), 378–417 (2018)
6. Fernholz, R., Garvy, R., Hannon, J.: Diversity-weighted indexing. J. Portf. Manag. **24**(2), 74–82 (1998)
7. Booth, D.G., Fama, E.F.: Diversification returns and asset contributions. Financ. Anal. J. **48**(3), 26–32 (1992)
8. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**(3), 793–803 (1983)
9. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**(3), 631–647 (1992)
10. Amari, S.-I.: Information Geometry and Its Applications. Springer, Berlin (2016)
11. Fenholz, R.: Portfolio generating functions. Quantitative Analysis in Financial Markets: Collected Papers of the New York University Mathematical Finance Seminar, vol. 1, pp. 344–367. World Scientific, Singapore (1999)
12. Pal, S., Wong, T.-K.L.: The geometry of relative arbitrage. Math. Financ. Econ. **10**(3), 263–293 (2016)
13. Pal, S., Wong, T.-K.L.: Exponentially concave functions and a new information geometry. Ann. Probab. **46**(2), 1070–1113 (2018)
14. Wong, T.-K.L.: Logarithmic divergences from optimal transport and Rényi geometry. Inf. Geom. **1**(1), 39–78 (2018)
15. Karatzas, I., Ruf, J.: Trading strategies generated by Lyapunov functions. Financ. Stoch. **21**(3), 753–787 (2017)
16. Brody, D.C., Hughston, L.P.: Interest rates and information geometry. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 457, pp. 1343–1363. The Royal Society (2001)

17. Trivellato, B.: Deformed exponentials and applications to finance. Entropy **15**(9), 3471–3489 (2013)
18. Pal, S.: Embedding optimal transports in statistical manifolds. Indian J. Pure Appl. Math. **48**(4), 541–550 (2017)
19. Khashanah, K., Yang, H.: Evolutionary systemic risk: fisher information flow metric in financial network dynamics. Phys. A Stat. Mech. Appl. **445**, 318–327 (2016)
20. Nock, R., Magdalou, B., Briys, E., Nielsen, F.: On tracking portfolios with certainty equivalents on a generalization of Markowitz model: the fool, the wise and the adaptive. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 73–80 (2011)
21. Nock, R., Magdalou, B., Briys, E., Nielsen, F.: Mining matrix data with Bregman matrix divergences for portfolio selection. Matrix Information Geometry, pp. 373–402. Springer, Berlin (2013)
22. Breuer, T., Csiszár, I.: Information geometry in mathematical finance: model risk, worst and almost worst scenarios. In: 2013 IEEE International Symposium on Information Theory Proceedings (ISIT). IEEE (2013)
23. Pal, S., Wong, T.-K.L.: Energy, entropy, and arbitrage. arXiv preprint arXiv:1308.5376 (2013)
24. Wong, T.-K.L.: Optimization of relative arbitrage. Ann. Financ. **11**(3–4), 345–382 (2015)
25. Pal, S.: Exponentially concave functions and high dimensional stochastic portfolio theory. arXiv preprint arXiv:1603.01865 (2016)
26. Fernholz, R., Karatzas, I.: Relative arbitrage in volatility-stabilized markets. Ann. Financ. **1**(2), 149–177 (2005)
27. Banner, A.D., Fernholz, R., Karatzas, I.: Atlas models of equity markets. Ann. Appl. Probab. **15**(4), 2296–2330 (2005)
28. Dembo, A., Tsai, L.-C.: Equilibrium fluctuation of the Atlas model. Ann. Probab. **45**(6b), 4529–4560 (2017)
29. Fernholz, R., Ichiba, T., Karatzas, I.: A second-order stock market model. Ann. Financ. **9**(3), 439–454 (2013)
30. Ichiba, T., Pal, S., Shkolnikov, M.: Convergence rates for rank-based models with applications to portfolio theory. Probab. Theory Relat. Fields **156**(1–2), 415–448 (2013)
31. Ichiba, T., Papathanakos, V., Banner, A., Karatzas, I., Fernholz, R.: Hybrid Atlas models. Ann. Appl. Probab. **21**(2), 609–644 (2011)
32. Jourdain, B., Reygner, J.: Capital distribution and portfolio performance in the mean-field atlas model. Ann. Financ. **11**(2), 151–198 (2015)
33. Pal, S.: Analysis of market weights under volatility-stabilized market models. Ann. Appl. Probab. **21**(3), 1180–1213 (2011)
34. Banner, A.D., Fernholz, D.: Short-term relative arbitrage in volatility-stabilized markets. Ann. Financ. **4**(4), 445–454 (2008)
35. Fernholz, D., Karatzas, I.: On optimal arbitrage. Ann. Appl. Probab. **20**(4), 1179–1204 (2010)
36. Fernholz, R., Karatzas, I., Kardaras, C.: Diversity and relative arbitrage in equity markets. Financ. Stoch. **9**(1), 1–27 (2005)
37. Vervuurt, A., Karatzas, I.: Diversity-weighted portfolios with negative parameter. Ann. Financ. **11**(3–4), 411–432 (2015)
38. Ernst, P.A., Thompson, J.R., Miao, Y.: Tukeys transformational ladder for portfolio management. Financ. Mark. Portf. Manag. **31**(3), 317–355 (2017)
39. Monter, S.A.A., Shkolnikov, M., Zhang, J.: Dynamics of observables in rank-based models and performance of functionally generated portfolios. arXiv preprint arXiv:1802.03593 (2018)
40. Audrino, F., Fernholz, R., Ferretti, R.G.: A forecasting model for stock market diversity. Ann. Financ. **3**(2), 213–240 (2007)
41. Rockafellar, R.T.: Convex Analysis. Princeton Landmarks in Mathematics. Princeton University Press, New Jercy (1997)
42. Johannes, R. Kangjianan, X.: Generalised lyapunov functions and functionally generated trading strategies. arXiv preprint arXiv:1801.07817 (2018)
43. Vervuurt, A.: On portfolio construction through functional generation. Ph.D. thesis, Oxford Unviersity (2016)

44. Chuaqui, M., Duren, P., Osgood, B.: Schwarzian derivative criteria for valence of analytic and harmonic mappings. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 143, pp. 473–486. Cambridge University Press, Cambridge (2007)
45. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)

Generalising Frailty Assumptions in Survival Analysis: A Geometric Approach

Vahed Maroufy and Paul Marriott

Abstract This paper uses Information Geometry in a practical and important applied statistical context: Cox regression in survival analysis. We explore the geometry of the corresponding model space including its potentially complex boundary. The exact manner that frailty terms in Cox's hazard model are specified has important implications for modelling. For example, it is very common to assume a gamma frailty for reasons of mathematical tractability and convenience. In this paper, we examine if there is a cost to having the gamma as a default option, without further scientific justification. We take a geometric approach to understanding the effect of precise model specification. We use a new, highly flexible but statistically well-behaved, way of specifying the frailty to calibrate modelling assumptions that are very commonly used in practice. We show that the gamma frailty assumption has the effect of considerably under-estimating standard errors when compared to our more general assumptions and, potentially, introducing bias. We comment on the implications of this. The survival times of adult acute myeloid leukaemia patients in northwest England are analyzed.

1 Introduction

One of the key insights of Information Geometry is the duality which links the exponential and mixture affine structures, [10]. Finite dimensional affine structures in the first are exponential families and affine structures in the second determine identification conditions in mixture models, [22]. This paper takes this fundamental insight and explores its implications in a practical and important applied statistical context.

V. Maroufy
Department of Biostatistics and Data Science, School of Public Health, University of Texas Health, Houston, USA
e-mail: vahed.maroufy@uth.tmc.edu

P. Marriott (✉)
Department of Statistics, University of Waterloo, Waterloo, ON, Canada
e-mail: pmarriot@uwaterloo.ca

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_7

The motivating example of this paper is a study of the factors which affect the survival times of adult acute myeloid leukaemia patients. This is an example of a problem in lifetime data analysis, [20]. One of the most popular tools here is Cox's proportional hazard model, [9]. This models the time dependent hazard function, for subject i with covariates X_i, as a product

$$h_i(t) = h_0(t) \exp\{X_i \beta\},$$

where $h_0(t)$ is the time dependent baseline hazard common to all subjects. This will be treated as a non-parametric nuisance function. The second term is a time independent function of the explanatory variables, which is modelled parametrically. Under the proportional hazard assumption the interest parameters can be estimate using a partial likelihood. Often though there are unmeasured individual level covariates which need to be taken into account in the analysis. This is commonly done by adding subject level terms θ_i in the adapted hazard given by (1). Since this will give a parameter for each subject it is convenient to treat these as unobserved random effects, or *frailties*, and simply estimate parameters associated with the distribution of these terms.

This means that we work by mixing over different hazard functions and the key model choice issue is the specification of the frailty distribution. It is here that Information Geometry plays a role through consideration of the fundamental mixture geometry and its impact on inference of the parameters of interest. In particular we use tools associated with the *local mixture model*, see [2, 22]. Of special geometric interest is the potentially complex boundary that these models can have. For example [24] shows that local mixture models can have boundaries which are boundaries of polytopes or can be a non-smooth union of a finite number of smooth components.

Frailty models have been studied by many researchers; for example, [6, 12, 15, 19]. Various hazard models, including Cox's regression model, have been generalised by assuming a random frailty variable. The frailty model is commonly chosen to return a tractable marginal likelihood function; hence, gamma, inverse Gaussian and positive stable distributions with closed-form Laplace transformations are regular choices [7, p.77]. In [16] a discrete mixture model is form a gamma and an Inverse Gaussian, leading to a three-parameter model, is considered for the frailty. Among these the gamma is the most popular model choice, [19, 27, 28].

The goal of this paper is to investigate what might be the cost of specifying a particular parametric form, specifically the gamma, for the frailty distribution when this choice has been made purely for mathematical convenience. Model uncertainty is a critical problem in applied statistics, and – as we do here – its analysis can be treated in a geometric way. The use of geometry in the area is not new, for example the paper [8] provides an intriguing solution by proposing the 'double the variance' method for addressing the possibility of undetectably small departures from the model. Much more detail on the geometry of model specification can be found in [4].

We investigate the choice of a gamma frailty by calibrating it to a much more general space of mixtures, which can be used as a bench-marking tool. We use the excellent study of [13] to illustrate these effects in a real context. We show that making the gamma frailty assumption can have the effect of considerably under-estimating standard errors and its potential misspecification gives rise to important biases.

In general with mixture models, and specifically with the frailty models considered here, the analyst has to balance flexibility with inferential and computational tractability. In this paper we use the recently defined *discrete mixture of local mixture models*, introduced in [25], to achieve the calibration. This new model generates a finite dimensional and suitably parameterized space as a high quality approximation to a general mixture model with unknown mixing mechanism. The model is always identifiable and estimable, and its geometric and inferential properties allow for fast and efficient estimation algorithms. These properties make it suitable as a calibration tool, and we are currently exploring using it as a frailty model in its own right.

Notation, motivation and the main results of the paper, including our proposed method are presented in Sect. 2. Section 3 looks at the calibration, with Sect. 3.1 showing a simulation study, illustrating that the local mixture method returns similar biases, but larger – often more than double – standard deviations for the estimates compared to the Expected-Maximization method of [19] where a gamma frailty is assumed. This shows the considerable impact that fixing on a particular parametric form has on inference, and clearly illustrates that the reduction in standard error may be just an artefact of modelling choice rather than being real. In Sect. 3.2, the survival time of 1043 adults acute myeloid leukemia patients, recorded between 1982 and 1998 in northwest England, is analyzed, again with important differences in inferential conclusions being found. The paper closes with a short discussion in Sect. 4.

2 Methodology

Throughout this section, we follow the notation and definitions in [12, 20]. Let (T_i^0, C_i), for $i = 1, \dots, n$, be the failure time and censoring time of the ith individual, and also let X be the $n \times p$ design matrix of the covariate vectors. Define $T_i = \min(T_i^0, C_i)$ and $\delta_i = I(T_i^0 < C_i)$, where $I(\cdot)$ is an indicator function. In addition, associated with the ith individual, an unobservable covariate θ_i, the frailty, is assumed, where θ_i's follow some distribution, Q. Adapting the proportional hazard model of [9] for the ith individual, conditional on the frailty θ_i, the hazard function is,

$$h_i(t) = \theta_i \, h_0(t) \exp\{X_i \beta\}, \tag{1}$$

where $h_0(t)$ is the base hazard function, X_i is the ith row of X and $\beta = (\beta_0, \dots, \beta_{p-1})^T$ is a p-vector of regression coefficients. The cumulative hazard and survival functions are, respectively, defined as $H_i(t) = \int_0^t h_i(u)\, du$, and $S_i(t) = \exp\{-H_i(t)\}$, with the base cumulative hazard function $H_0(t)$. We also assume that the frailty θ is independent of X, and further that, given X and θ, censoring is independent and noninformative for θ and (h_0, β), see [12]. The full likelihood function for the parameter vector (β, H_0) is written as

$$L(\beta, H_0, Q) = \prod_{i=1}^{n} \int \left[\left(\theta\, h_0(T_i)\, e^{X_i\beta}\right)^{\delta_i} \exp\left\{-\theta\, H_0(T_i)\, e^{X_i\beta}\right\}\right] dQ(\theta). \tag{2}$$

In a similar way to a general mixture model problem, frailty survival models with an unknown frailty distribution, suffer from identification and estimability issues. Although, when all the covariates variables are continuous with a continuous distribution, [11], show that, given the distribution of the time duration variable, all the three multiplicative factors are, at least formally, identified. This theoretical result does not solve the identifiability issue in the general sense. For instance, when there is a discrete covariate then identifiability requires the corresponding regression coefficient to be limited to a known compact set [14, Ch. 2].

2.1 Local and Global Mixture Models

To allow full generality of the mixture structure it is tempting to only restrict Q to be a finite discrete distribution with an unknown number of components., i.e.

$$Q(\theta) = \sum_{i=1}^{N} \rho_i \delta_{\theta_i}(\theta),$$

where δ is the indicator function, and $\sum_{i=1}^{N} \rho_i = 1$ and $\rho_i > 0$. Indeed, as shown by [21], the non-parametric maximum likelihood estimate of Q lies in such a family. In that case the perturbation space would be 'parameterised' by N, the number of components, $(\theta_1, \ldots, \theta_N)$, the components, and $(\rho_1, \ldots, \rho_N)$, the mixing weights. However, this parameterization has many problems in implementation. Specifically, it is poorly identified and has complex boundaries. A key problem is that mixture components may be too close to one another to be resolved with a given set of data and so the order of the finite mixture is essentially not estimable, see [25].

The case where there is a single set of closely grouped components – or the much more general situation where Q is any small-variance distribution – is exactly the case which motivated the design of the local mixture model (LMM), see [1, 22]. The key idea is, essentially, to replace the unknown number of clustered mixing components with a fixed number of low order moments parameterised in an inferentially 'nice' way. The geometric intuition is that for a local perturbation all the mixing component distributions will lie close to a low dimensional linear (affine) space. This space is spanned by derivatives of the prior and parameterised with a small number of identified parameters.

Definition 1 For a density function, $f(y; \theta)$, belonging to the exponential family, the local mixture model of order k, centred at ϑ, is defined as

$$g_{\vartheta}(y; \lambda) = f(y; \vartheta) + \sum\nolimits_{j=1}^{k} \lambda_j f^{(j)}(y; \vartheta), \tag{3}$$

where $\lambda := (\lambda_1, \ldots, \lambda_k) \in \Lambda_\vartheta, f^{(j)}(y; \vartheta) = \frac{\partial^j f(y;\theta)}{\partial \theta^j}\ |_{\theta=\vartheta}$. The parameter space, Λ_ϑ, is defined by

$$\Lambda_\vartheta := \{\lambda \mid g_\vartheta(y; \lambda) \geq 0,\ \forall y\}$$

is convex with a boundary determined by the non-negativity of (3). The structure of this boundary is the union of smooth manifolds and is studied in [24]. Also (3) is parameter invariant, and mean parameter is chosen only because of its clear interpretation, see [22].

In [25], local mixtures are generalised to discrete mixtures of local mixture models. In Definition 1 the point ϑ is fixed, selected by the analyst. This has the natural generalisation to have a set of possible centres, $\vartheta_1, \ldots \vartheta_L$, selected independently of the data. This gives rise to the following.

Definition 2 The discrete mixture of local mixture models is defined by

$$g(y; \underline{\vartheta}, \underline{\lambda}) = \sum\nolimits_{l=1}^{L} \rho_l\, g_{\vartheta_l}(y; \lambda^l), \tag{4}$$

where $\underline{\lambda} = (\lambda^1, \ldots, \lambda^L)$, $\underline{\vartheta} = (\vartheta_1, \ldots, \vartheta_L)$, $\rho_l \geq 0$ and $\sum_{l=1}^{L} \rho_l = 1$.

The model selection task with such models is to select k, L and $\vartheta_1, \ldots \vartheta_L$ to balance estimability with the quality of approximation to a completely general mixture. We use the results of [25, Section 2.1] which show, for a given value of k, how to select the points $\vartheta_1 < \cdots < \vartheta_L$ to have a uniform L^1-bound on the difference between $g(y; \underline{\vartheta}, \underline{\lambda})$ and any mixture with mixing distribution Q whose support lies in a compact region containing $\{\vartheta_1, \ldots, \vartheta_L\}$.

We emphasis, before we show some of the details in an example, that in this paper, we are using the LMM structure as a calibration tool in order to evaluate the effect of selecting a particular parametric frailty model. We are generating a rich, but well behaved, class of mixtures which subsume the gamma assumption.

Example 1 To illustrate the flexibility of this framework, Fig. 1 shows some discrete mixtures of local mixture models for the exponential distribution on both the density and hazard scales. Using the methods of [25, Section 2.1] some straightforward numerical analysis shows that selecting points, in the rate parametrisation, at $\vartheta_1 = 0.5, \vartheta_2 = 1, \vartheta_3 = 2$ and $\vartheta_4 = 3.5$ and fixing $k = 4$ gives excellent L^1-approximations for all mixtures Q with support in $[0.3, 5]$. In the figure Model 1 is the unmixed exponential model, Model 2 is a single component local mixture centered at $\vartheta_2 = 1$, Models 3 and 4 are two components mixtures of local mixture centred at $(1, 3.5)$ and $(0.5, 1)$, respectively.

In [25] it was shown that an adaptation of the EM algorithm works well for fitting a finite mixture of local mixtures. We only consider here local mixture models of order $k = 4$. Increasing this degree – while mathematically possible – only adds a small marginal improvement to the local modelling performance, [23], at the cost of extra parameters.

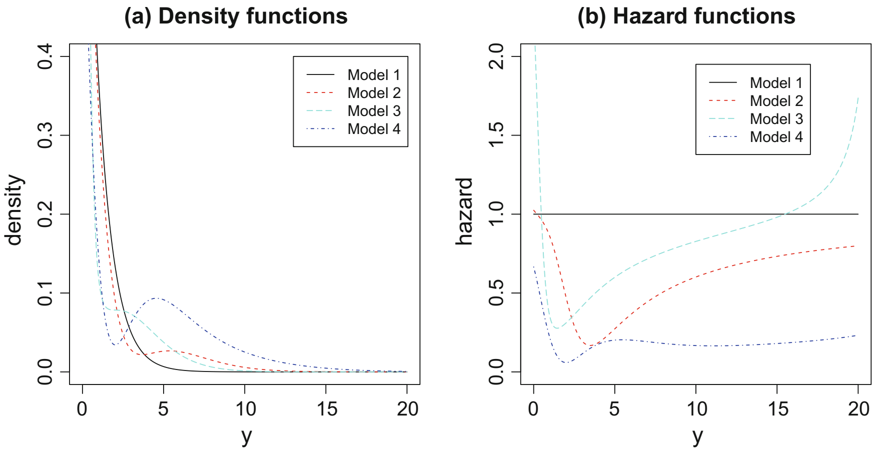


Fig. 1 The flexibility of the model class: **a** four models on a density scale, **b** the same models' hazard functions

The key, novel, issue in the estimation is to deal with the fact that the parameter space, Λ_ϑ, has a complex boundary. For this paper, for the Cox model, we need to consider the case where $f(y;\theta) = \theta\exp(-\theta y)$ and then the parameter space is characterized as the set of all λ's such that, for all $y > 0$,

$$\vartheta\lambda_4 y^4 - (4\lambda_4 + \vartheta\lambda_3)y^3 + (3\lambda_3 + \vartheta\lambda_2)y^2 - (2\lambda_2 + \vartheta\lambda_1)y + \lambda_1 + \vartheta \geq 0. \quad (5)$$

The boundary is characterised as being the cases where (5) has only repeated positive roots, and it can be shown, by direct calculation, that the parameter space is the union of smooth manifolds. This characterisation allows explicit parameterisations of the component manifolds to be computed and these are exploited in the numerical studies of this paper. See Appendix for details.

A cost of this boundary is that the maximum likelihood estimate may not exist in the regular sense as a turning point. Furthermore, asymptotic approximations in sample size, similar to that of [29], ultimately break down close to the boundary, irregardless of their order. In [3] a diagnostic approach is proposed which identifies when a first order asymptotic is appropriate, depending on the square distance of the MLE from the boundary, measured using the Fisher information.

3 Application to Frailty Modelling

In this section, we investigate the size of the effect on inference associated with fixing on a particular frailty distribution by comparing a standard gamma frailty model to our more general local mixture approach. In cases where the only reason for selecting a gamma frailty was convenience we argue differences, such as a reduction in standard error, are artefacts of model choice and not real.

In order to investigate the best possible case for the gamma frailty assumption we simulate, in Sect. 3.1, from that frailty distribution, and we only make the comparison to a one component local mixture with $k = 4$. With these restrictions we see that standard errors associated with the gamma assumption are considerably smaller than those of the LMM. Furthermore, in the real data example of Sect. 3.2, it turns out that the $k = 4$, one component local mixture is the appropriate choice for that data and we find important inferential differences which might be due to either misspecification bias, or artificially reduced standard errors, in the gamma based model.

Intuitively, for any fixed ϑ, the finite dimensional parameter vector λ represents the frailty distribution through its central moments. Local mixing also can be seen as a mechanism that extends a parametric model to a larger and more flexible space of densities which has nice geometric and inferential properties, allowing calibration of a particular parametric assumption. Substituting the local mixture expansion of Eq. (3) we obtain

$$\begin{aligned} l(\beta, H_0, \lambda) = &\sum\nolimits_{i=1}^{n} (\delta_i[\log h_0(T_i) + X_i\beta] + \log f(T_i, \beta, \vartheta)) \\ &+ \sum\nolimits_{i=1}^{n} \log\left(1 + \sum\nolimits_{j=1}^{k} \lambda_j A_j(\delta_i, y_i)\right), \qquad \lambda \in \Lambda_\vartheta \end{aligned} \tag{6}$$

in which $A_j(\delta_i, y_i) = \frac{f^{(j)}(T_i,\beta,\vartheta)}{f(T_i,\beta,\vartheta)}$, and $y_i = H_0(T_i)e^{X_i\beta}$. We maximize Eq. (6) when estimating β, where H_0 and λ are considered as nuisance parameters. Thus, a profile likelihood optimization method is employed. That is, we first maximize for λ over Λ_ϑ to obtain $\hat{\lambda}$ and impute $\hat{H}_0$ for H_0, then maximize $l_p(\beta) = l(\beta, \hat{H}_0, \hat{\lambda})$ to estimate β.

To impute $h_0(t)$ and $H_0(t)$ we use the arguments in [12] to provide a recursive estimate of the cumulative hazard function using the fact that for two consecutive failure times, $T_{(i)}$ and $T_{(i+1)}$, we have $H_0(T_{(i+1)}) = H_0(T_{(i)}) + \Delta H_i$. Substituting this recursive equation into the log-likelihood function in (6), considering the conventions in [5] and taking partial derivative with respect to ΔH_i, we obtain

$$\frac{\partial l}{\partial \Delta H_i} = \frac{1}{\Delta H_i} - \sum_{\ell=i}^{n} e^{X_\ell \beta} + \frac{P'(e^{X_i\beta}[H_0(T_{(i)}) + \Delta H_i])}{P(e^{X_i\beta}[H_0(T_{(i)}) + \Delta H_i])}, \tag{7}$$

which is a function of just ΔH_i when $\hat{H}_0(T_{(i)})$ is given at time $T_{(i+1)}$, $P(\cdot)$ is a polynomial of degree four with its coefficients linear functions of $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ and $P'(\cdot)$ is its derivative with respect to ΔH_i. When the denominator is not zero, Eq. (7) is a polynomial of degree five which can be solved numerically for ΔH_i. Note that when there is no frailty factor – that is $\lambda = (0, 0, 0, 0)$ – then the last term in Eq. (7) is zero, and the estimate of the cumulative hazard function reduces to the form in [18] which is the estimate in [19] with $\hat{\omega} = 1$.

Table 1 Bias and standard errors of coefficient estimates, when frailty is generated from $\Gamma(\frac{1}{\eta}, \eta)$. LMM is the local mixture method; Gamma the Expectation–Maximization method for the gamma frailty

			Gamma		LMM	
n	η	β	*bias*	*std*	*bias*	*std*
200	0.5	$\log 3$	−0.057	0.18	−0.048	0.43
200	0.7	$\log 3$	−0.056	0.21	−0.038	0.40
200	1	$\log 3$	−0.117	0.22	−0.094	0.41

3.1 Simulation Study

In this section, a simulation study illustrates the effect of making the gamma frailty assumption when compared to the much richer local mixture method. We adapt the method in [19], which assumes a gamma model with mean 1 and variance η for the frailty, and apply the Expectation–Maximization algorithm.

We let $C = 0.01$, $\tau = 4.6$ and we follow a similar set-up as found in [17]. For each individual the event time is $T = [-\log(1-U)\{\theta \exp\{\beta X\}\}^{-1}]^{-1/\tau} C^{-1}$, where $X \sim N(0, 1)$, $U \sim \text{uniform}[0, 1]$. The censoring distribution is $N(100, 15)$, and frailty is assumed to follow a gamma distribution with mean 1 and variance η. Table 1 shows the bias and standard error for the estimates of the regression coefficient using both methods for three different values of η. It is clear that the local mixture method, which does not use any information about the frailty model, returns very similar biases as the Expectation–Maximization method for the gamma frailty. However, the standard deviation for the estimates in the local mixture method are almost twice as large as these for the Expectation–Maximization method.

Similar simulation studies are presented in Tables 2 and 3 where the frailty is generated from different models than the gamma distribution discussed above. In Table 2, the frailty is generated from uniform distributions with means set to one and different variances. The standard deviation for the LMM model is twice as large as that for Gamma model, similar to that in Table 1. In addition, since the gamma model is now misspecification, as the variation of the frailty gets larger, the LMM slightly beats the Gamma approach with respect to estimation bias.

Table 3 considers two different models for the frailty: the Inverse Gaussian and Inverse Gamma models. All have their mean set to one but with different variances. Although the LMM model still returns a bigger standard deviation, it is no longer twice as big as that for the Gamma model.

The results in Tables 1, 2 and 3 conveys three important messages. First, without assuming any specific model form for the unobserved frailty, our methodology does equally well in terms of the bias of estimation. Second, our method automatically addresses the possibility of departures from the gamma assumption by returning a bigger standard deviation, double in some cases. We argue that the apparent reduction of standard error, associated with the gamma, is an artifact of model choice which

Table 2 Bias and standard errors of coefficient estimates, when frailty is generated from uniform $U(1-\eta, 1+\eta)$ distributions. LMM is the local mixture method; Gamma the Expectation–Maximization method for the gamma frailty

			Gamma		LMM	
n	η	β	*bias*	*std*	*bias*	*std*
200	0.5	$\log 3$	0.003	0.104	0.02	0.205
200	0.7	$\log 3$	0.067	0.113	0.011	0.237
200	1	$\log 3$	−0.23	0.144	−0.18	0.264

Table 3 Bias and standard errors of coefficient estimates, when frailty is generated from Inverse-gamma and Inverse-Gaussian. LMM is the local mixture method; Gamma the Expectation–Maximization method for the gamma frailty

			Gamma		LMM		
n	*Model*	$(mean, var)$	β	*bias*	*std*	*bias*	*std*
200	$IGamma(3, 2)$	$(1, 1)$	$\log 3$	−0.106	0.112	−0.094	0.172
200	$IGamma(2, 1)$	$(1, 2)$	$\log 3$	0.179	0.110	0.156	0.166
200	$IGauss(1, 0.5)$	$(1, 0.5)$	$\log 3$	−0.130	0.115	−0.115	0.151
200	$IGauss(1, 1)$	$(1, 1)$	$\log 3$	−0.198	0.104	−0.181	0.161

can only be justified if there were good, extra-data, reasons for trusting the gamma assumption. Third, in the case of model misspecification, the LMM dominates the Gamma model, as amount of frailty gets larger.

3.2 Data Example

We compare the two methods in the context of a study of the survival time of 1043 adult acute myeloid leukemia patients, recorded between 1982 and 1998 in northwest England [13]. In this example 16% of the survival times are censored and complete information is available for four covariates; age, sex, white blood cell count (WBC), and a measure of deprivation for the enumeration district of residence (Dep). [13], studied the data to investigate possible spatial variation in the survival time assuming a gamma marginal frailty with covariance structure for the frailty among the 24 districts. Although they find some indication of spatial variation between the districts, their analysis illustrates that, assuming covariance structure in the frailty variable does not affect the inference on the regression coefficients, and differences in the coefficient estimates are not significantly bigger than one standard deviation. The estimates and standard errors of the regression coefficients, $(\beta_1, \ldots, \beta_4)$, based on the independent gamma frailty are shown in the first two rows of Table 4. The frailty variance is estimated as $\hat{\eta} = 0.772$, indicating the existence of unobserved variation in the patients' survival time.

Table 4 Estimates and standard errors for covariate coefficient are presented. LMM, the local mixture method; Gamma; the estimates in *Henderson et al. (2002)* for the gamma frailty. Std are the associated standard errors for the gamma model

	Age	Sex	WBC	Dep
Gamma	0.0470	0.0563	0.0057	0.0547
Std	0.0045	0.0505	0.0008	0.0187
LMM	0.0422	0.0156	0.0114	0.0306

Applying the local mixture method we obtain $\hat{\lambda} = (-0.687, 0.017, 0.056, 0.023)$, also reflecting the existence of some unobserved variation. Furthermore, $\hat{\lambda}$ lie in the relative interior of Λ_{ϑ}, indicating that the one component local mixture is adequate. We note that if this estimate were to lay on the boundary the one component LMM approximation used here might still be a good L^1 approximation but perhaps poor in a KL-divergence sense. If this were to happen we would recommend trying a larger number of components.

The estimates for $(\beta_1, \ldots, \beta_4)$, using the LMM approach, are shown in the third row of Table 4. Comparing the differences between the estimates with the standard errors, we realize that all differences are inside the one standard deviation bound except for β_3, the coefficient for WBC, where the ratio is 7.125. This important difference in the estimate is interpreted as the result of a possible misspecification of the frailty model when a gamma distribution is imposed. It could be due to either misspecification bias or an under estimate of the standard error in the gamma frailty model. We note that our simulation results point to the fact that the bias in the LMM might be expected to be small when there is, in fact, a gamma frailty.

4 Discussion

In the context of frailty survival analysis, this paper applies an approximation of a general mixture model – with a completely unspecified mixing mechanism – by a novel family of models, the discrete mixture of local mixture models. These models are built in a geometrically and inferentially convenient way and have many advantages compared to finite mixture models. We do this to generate a calibration tool with which we can evaluate the effect of specifying a particular choice of frailty model. These geometric properties lead to inferential properties such as concavity of the log-likelihood function, identifiability and orthogonality in parametrization.

Our novel, highly flexible, but inferentially well-behaved, family gives a way of calibrating the effect of making a closed form parametric assumption on the frailty term. We see considerable differences in both the simulation and real data examples. Hence we conclude that making the gamma frailty assumption can have the effect of considerably under-estimating standard errors.

We note that, in this manuscript, we only use a one component LMM for the simulation and data example, and work under the assumption that the frailty is only a small component of the total variation. Nevertheless, the theory is quite general and allows for all mixing distributions in principle. For example we have recently applied it to an analysis of robustness in Bayesian conjugate and sparsity priors [26]. To apply Definition 2 here to its full potential, requires finding a good estimator for the cumulative hazard function and this is still an open problem.

Appendix

For the exponential distribution, the density of a single component local mixture has the form

$$\left\{\vartheta\lambda_4 y^4 - (4\lambda_4 + \vartheta\lambda_3)y^3 + (3\lambda_3 + \vartheta\lambda_2)y^2 - (2\lambda_2 + \vartheta\lambda_1)y + \lambda_1 + \vartheta\right\} \exp(-\vartheta y), \tag{8}$$

which is a density when the polynomial is non-negative for all $y > 0$. It will lie on the boundary of the parameter space when all the positive roots of the polynomial are of order 2 or 4 and $\lambda_4 > 0$. Polynomials of this form can be written as

$$\left\{a(y-r)^2(y^2 + 2by + c)\right\}, \tag{9}$$

then the roots of $(y^2 + 2by + c)$ must be repeated when they are positive, and a, b, c and r satisfy

$$\vartheta^5 - \vartheta^4 ar^2c - (2ar^2b - 2arc)\vartheta^3 - (2ac + 2ar^2 - 8arb)\vartheta^2 - (12ab - 12ar)\vartheta - 24a = 0. \tag{10}$$

This follows since a density integrates to one. Comparing (8) with (9) and imposing (10) results in an explicit parameterisation of the boundary in terms of b, c and r. The Jacobean of this has full rank as long as $r \neq 0, -b \pm \sqrt{b^2 - c}$. So, under these conditions the boundary is a smooth manifold. In the singular case where $r = -b \pm \sqrt{b^2 - c}$, since roots must have even order, we have $r = -b$ and $b < 0$. This singular set is given by a one dimensional manifold, again with an explicit parameterisation.

References

1. Anaya-Izquierdo, K., Marriott, P.: Local mixture models of exponential families. Bernoulli **13**, 623–640 (2007)
2. Anaya-Izquierdo, K., Marriott, P.: Local mixture of the exponential distribution. Ann. Inst. Math. Stat. **59**(1), 111–134 (2007)
3. Anaya-Izquierdo, K., Critchley, F., Marriott, P.: When are first order asymptotics adequate? a diagnostic. Stat **3**(1), 17–22 (2013)

4. Anaya-Izquierdo, K., Critchley, F., Marriott, P., Vos, P.: The geometry of model sensitivity: an illustration. In: Computational Information Geometry: For Image and Signal Processing. Springer, Berlin (2016)
5. Breslow, N.: Contribution to the discussion on the paper of Cox. Biom **34**, 216–217 (1972)
6. Clayton, D.G.: A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika **65**(1), 141–151 (1978)
7. Cook, R.J., Lawless, J.F.: The statistical analysis of recurrent events (2007)
8. Copas, J., Eguchi, S.: Local model uncertainty and incomplete-data bias (with discussion). J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**(4), 459–513 (2005)
9. Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. B **34**, 187–200 (1972)
10. Critchley, F., Marriott, P.: Information geometry and its applications: an overview. In: Computational Information Geometry: For Image and Signal Processing, pp. 1–31. Springer, Berlin (2017)
11. Eleber, C., Ridder, G.: True and spurious duration dependence: the identifiability of the proportional hazard model. Rev. Econ. Stud. **49**, 403–409 (1982)
12. Gorfine, M., Zucher, D.M., Hsu, L.: Prospective survival analysis with general semiparametric shared frialty model: a pseudo full likelihood approach. Biometrika **93**(3), 735–741 (2006)
13. Henderson, R., Shimakura, S., Gorst, D.: Modeling partial variation in Leukemia survival data. J. Am. Stat. Assoc. **97**(460), 965–972 (2002)
14. Horowitz, J.L.: Semiparametric and Nonparametric Methods in Econometrics. Springer, Berlin (2010)
15. Hougaard, P.: Survival models for heterogeneous population derived from stable distributions. Biometrika **73**(2), 387–396 (1986)
16. Hougaard, P.: Frailty models for survival data. Lifetime Data Anal. **1**(3), 255–273 (1995)
17. Hsu, L., Chen, L., Gorfine, M., Malone, K.: Semiparametric estimation of marginal hazard function from case-control family studies. Biometrics **60**, 936–944 (2004)
18. Johansen, S.: An extension of Cox's regression model. Int. Stat. Rev. **51**, 258–262 (1983)
19. Klein, J.P.: Semiparametric estimation of random effects using cox model based on the EM algorithm. Biometrics **48**, 795–806 (1992)
20. Lawless, J.F.: Statistical Models and Methods for Lifetime Data. Wiley, New York (1981)
21. Lindsay, B.G.: Mixture Models: Theory, Geometry and Applications. Institute of Mathematical Statistics, Hayward (1995)
22. Marriott, P.: On the local geometry of mixture models. Biometrika **89**, 77–93 (2002)
23. Marriott, P.: Extending local mixture models. AISM **59**, 95–110 (2006)
24. Maroufy, V., Marriott, P.: Computing boundaries in local mixture models. Geometric Science of Information. Lecture Notes in Computer Science, vol. 9389, pp. 577–585 (2015)
25. Maroufy, V., Marriott, P.: Mixture models: building a parameter space. Stat. Comput. 1–7 (2016)
26. Maroufy, V., Marriott, P.: Local and global robustness in conjugate Bayesian analysis and sparsity models. Statistica Sinica. In press (2018)
27. Nielsen, G.G., Gill, R.D., Anderson, P.K., Sorensen, T.I.A.: A counting process approach to maximum likelihood estimation in frailty models. Scand. J. Stat. **19**(1), 25–43 (1992)
28. Vaupel, J.W., Manton, K.G., Stallard, E.: The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography **16**, 439–454 (1979)
29. Zucker, D., Gorfine, M., Hsu, L.: Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: asymptotic theory. J. Stat. Plan. Inference **138**(7), 1998–2016 (2008)

Some Universal Insights on Divergences for Statistics, Machine Learning and Artificial Intelligence

Michel Broniatowski and Wolfgang Stummer

Abstract Dissimilarity quantifiers such as divergences (e.g. Kullback–Leibler information, relative entropy) and distances between *probability distributions* are widely used in statistics, machine learning, information theory and adjacent artificial intelligence (AI). Within these fields, in contrast, some applications deal with divergences between *other-type* real-valued functions and vectors. For a broad readership, we present a correspondingly unifying framework which – by its nature as a "structure on structures" – also qualifies as a basis for similarity-based multistage AI and more humanlike (robustly generalizing) machine learning. Furthermore, we discuss some specificalities, subtleties as well as pitfalls when e.g. one "moves away" from the probability context. Several subcases and examples are given, including a new approach to obtain parameter estimators in continuous models which is based on noisy divergence minimization.

1 Outline

The goals formulated in the abstract are achieved in the following way and order: to address a wide audience, throughout the paper (with a few connection-indicative exceptions) we entirely formulate and investigate divergences and distances between

M. Broniatowski
Sorbonne Universite Pierre et Marie Curie, LPSM, 4 place Jussieu, 75252 Paris, France
e-mail: michel.broniatowski@sorbonne-universite.fr

W. Stummer (✉)
Department of Mathematics, University of Erlangen–Nürnberg, Cauerstrasse 11,
91058 Erlangen, Germany
e-mail: stummer@math.fau.de

W. Stummer
Affiliated Faculty Member of the School of Business and Economics, University of
Erlangen–Nürnberg, Lange Gasse 20, 90403 Nürnberg, Germany

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication
Technology, https://doi.org/10.1007/978-3-030-02520-5_8

functions, even for the probability context. In Sect. 2, we provide some non-technical background and overview of some of their principally possible usabilities for tasks in data analytics such as statistics, machine learning, and artificial intelligence (AI). Furthermore, we indicate some connections with geometry and information. Thereafter, in Sect. 3 we introduce a new *structured* framework (toolkit) of divergences between functions, and discuss their building-blocks, boundary behaviour, as well as their identifiability properties. Several subcases, running examples, technical subtleties of practical importance, etc. are illuminated, too. Finally, we study divergences between "entirely different functions" which e.g. appear in the frequent situation when for data-derived *discrete* functions one wants to find a closest possible continuous-function model (cf. Sect. 4); several corresponding noisy minimum-divergence procedures are compared – for the first time within a unifying framework – and new methods are derived too.

2 Some General Motivations and Uses of Divergences

2.1 Quantification of Proximity

As a starting motivation, it is basic knowledge that there are numerous ways of evaluating the proximity $d(p,q)$ of two real numbers p and q of primary interest. For instance, to quantify that p and q nearly coincide one could use the difference $d^{(1)}(p,q) := p - q \approx 0$ or the fraction $d^{(2)}(p,q) := \frac{p}{q} \approx 1$, scaled (e.g. magnifying, zooming-in) versions $d^{(3)}_m(p,q) := m \cdot (p-q) \approx 0$ or $d^{(4)}_m(p,q) := m \cdot \frac{p}{q} \approx 1$ with "scale" m of secondary (auxiliary) interest, as well as more flexible hybrids $d^{(5)}_{m_1,m_2,m_3}(p,q) := m_3 \cdot \big(\frac{p}{m_1} - \frac{q}{m_2}\big) \approx 0$ where m_i may also take one of the values p, q. All these "dissimilarities" $d^{(j)}(\cdot,\cdot)$ can principally take any sign and they are asymmetric which is consistent with the – in many applications required – desire that one of the two primary-interest numbers (say p) plays a distinct role; moreover, the involved divisions cause technical care if one principally allows for (convergence to) zero-valued numbers. A more sophisticated, nonlinear alternative to $d^{(1)}(\cdot,\cdot)$ is given by the dissimilarity $d^{(6)}_\phi(p,q) := \phi(p) - (\phi(q) + \phi'(q) \cdot (p-q))$ where $\phi(\cdot)$ is a strictly convex, differentiable function and thus $d^{(6)}_\phi(p,q)$ quantifies the difference between $\phi(p)$ and the value at p of the tangent line taken at $\phi(q)$. Notice that $d^{(6)}_\phi(\cdot,\cdot)$ is generally still asymmetric but always stays nonnegative independently of the possible signs of the "generator" ϕ and the signs of p,q. In contrast, as a nonlinear alternative to $d^{(4)}_m(\cdot,\cdot)$ one can construct from ϕ the dissimilarity $d^{(7)}_\phi(p,q) := q \cdot \phi\big(\frac{p}{q}\big)$ (where $m = q$) which is also asymmetric but can become negative depending on the signs of p, q, ϕ. More generally, one often wants to work with dissimilarities $d(\cdot,\cdot)$ having the properties

(D1) $d(p,q) \geqslant 0$ for all p, q (nonnegativity),
(D2) $d(p,q) = 0$ if and only if $p = q$ (reflexivity; identity of indiscernibles[1]),

and such $d(\cdot,\cdot)$ is then called a *divergence* (or disparity, contrast function). Loosely speaking, the divergence $d(p,q)$ of p and q can be interpreted as a kind of "directed distance from p to q".[2] As already indicated above, the underlying directness turns out to be especially useful in contexts where the first component (point), say p, is always/principally of "more importance" or of "higher attention" than the second component, say q; this is nothing unusual, since after all, one of our most fundamental daily-life constituents – namely time – is directed (and therefore also time-dependent quantities)! Moreover, as a further analogue consider the "way/path-length" $d(p,q)$ a taxi would travel from point p to point q in parts of a city with at least one one-way street. Along the latter, there automatically exist points $p \neq q$ such that $d(p,q) \neq d(q,p)$; this non-equality may even hold for all $p \neq q$ if the street pattern is irregular enough; the same holds on similar systems of connected "one-way loops", directed graphs, etc. However, sometimes the application context demands for the usage of a dissimilarity $d(\cdot,\cdot)$ satisfying (D1), (D2) and

(D3) $d(p,q) = d(q,p)$ for all p, q (symmetry),

and such $d(\cdot,\cdot)$ is denoted as a *distance*; notice that we don't assume that the triangle inequality holds. Hence, we regard a distance as a symmetric divergence. Moreover, a distance $d(\cdot,\cdot)$ can be constructed from a divergence $\widetilde{d}(\cdot,\cdot)$ e.g. by means of either the three "symmetrizing operations" $d(p,q) := \widetilde{d}(p,q) + \widetilde{d}(q,p)$, $d(p,q) := \min\{\widetilde{d}(p,q), \widetilde{d}(q,p)\}$, $d(p,q) := \max\{\widetilde{d}(p,q), \widetilde{d}(q,p)\}$ for all p and q.

In many real-life applications, the numbers p, q of primary interest as well as the scaling numbers m_i of secondary interest are typically replaced by real-valued functions $x \rightarrow p(x)$, $x \rightarrow q(x)$, $x \rightarrow m_i(x)$, where $x \in \mathscr{X}$ is taken from some underlying set $\mathscr{X}$. To address the entire functions as objects we use the abbreviations $P := \big\{p(x)\big\}_{x\in\mathscr{X}}$, $Q := \big\{q(x)\big\}_{x\in\mathscr{X}}$, $M_i := \big\{m_i(x)\big\}_{x\in\mathscr{X}}$, and alternatively sometimes also $p(\cdot)$, $q(\cdot)$, $m_i(\cdot)$. This is conform with the high-level data processing paradigms in "functional programming" and "symbolic computation", where functions are basically treated as whole entities, too.

Depending on the nature of the data-analytical task, the function P of primary interest may stem either from a hypothetical model, or its analogue derived from observed/measured data, or its analogue derived from artificial computer-generated (simulated) data; the same holds for Q where "cross-over constellations" (w.r.t. to the origin of P) are possible.

The basic underlying set (space) $\mathscr{X}$ respectively the function argument x can play different roles, depending on the application context. For instance, if $\mathscr{X} \subset \mathbb{N}$ is a subset of the integers $\mathbb{N}$ then $x \in \mathscr{X}$ may be an index and $p(x)$ may describe the xth real-valued data-point. Accordingly, P is then a s-dimensional vector where s is the total number of elements in $\mathscr{X}$ with eventually allowing for $s = \infty$. In other

[1] See e.g. Weller-Fahy et al. [93].

[2] Alternatively, one can think of $d(p,q)$ as degree of proximity from p to q.

situations, x itself may be a data point of arbitrary nature (i.e. $\mathscr{X}$ can be any set) and $p(x)$ a real value attributed to x; this $p(x)$ may be of direct or of indirect use. The latter holds for instance in cases where $p(\cdot)$ is a density function (on $\mathscr{X}$) which roughly serves as a "basis" for the operationalized calculation of the "local aggregations over all[3] $A \subset \mathscr{X}$" in the sense of $A \rightarrow \sum_{x \in A} p(x)$ or $A \rightarrow \int_A p(x) \,\mathrm{d}\widetilde{\lambda}(x)$ subject to some "integrator" $\widetilde{\lambda}(\cdot)$ (including classical Riemann integrals $\mathrm{d}\widetilde{\lambda}(x) = \mathrm{d}x$); as examples for nonnegative densities $p(\cdot) \geqslant 0$ one can take "classical" (volumetric, weights-concerning) inertial-mass densities, population densities, probability densities, whereas densities $p(\cdot)$ with possible negative values can occur in electromagnetism (charge densities, polarization densities), in other fields of contemporary physics (negative inertial-mass respectively gravitational-mass densities) as well as in the field of acoustic metamaterials (effective density), to name but a few.

Especially when used as a set of possible states/data configurations (rather than indices), $\mathscr{X}$ can be of arbitrary complexity. For instance, each x itself may be a real-valued continuous function on a time interval $[0, T]$ (i.e. $x : [0, T] \rightarrow] - \infty, \infty[$) which describes the scenario of the overall time-evolution of a quantity of interest (e.g. of a time-varying quantity in an deterministic production process of one machine, of the return on a stock, of a neural spike train). Accordingly, one can take e.g. $\mathscr{X} = C\big([0, T],] - \infty, \infty[\big)$ to be the set of all such continuous functions, and e.g. $p(\cdot)$ a density thereupon (which is then a function on functions). Other kinds of functional data analytics can be covered in an analogous fashion.

To proceed with the proximity-quantification of the primary-interest functions $P := \big\{p(x)\big\}_{x \in \mathscr{X}}$, $Q := \big\{q(x)\big\}_{x \in \mathscr{X}}$, in accordance with the above-mentioned investigations one can deal with the pointwise dissimilarities/divergences $d_{\phi}^{(j)}(p(x), q(x))$, $d_{m_1(x), m_2(x), m_3(x)}^{(5)}(p(x), q(x))$ for fixed $x \in \mathscr{X}$, but in many contexts it is crucial to take "summarizing" dissimilarities/divergences

$$D_{\phi}^{(j)}(P, Q) := \sum_{x \in \mathscr{X}} d_{\phi}^{(j)}(p(x), q(x)) \cdot \lambda(x) \quad \text{or} \quad D_{\phi}^{(j)}(P, Q) := \int\limits_{\mathscr{X}} d_{\phi}^{(j)}(p(x), q(x)) \,\mathrm{d}\lambda(x)$$

subject to some weight-type "summator"/"integrator" $\lambda(\cdot)$ (including classical Riemann integrals); analogously, one can deal with $D_{\phi, M_1, M_2, M_3}^{(5)}(P, Q) := \sum_{x \in \mathscr{X}} d_{m_1(x), m_2(x), m_3(x)}^{(5)}(p(x), q(x)) \cdot \lambda(x)$ or $D_{\phi, M_1, M_2, M_3}^{(5)}(P, Q) := \int_{\mathscr{X}} d_{m_1(x), m_2(x), m_3(x)}^{(5)}(p(x), q(x)) \,\mathrm{d}\lambda(x)$. Notice that the requirements (D1), (D2) respectively (D3) carry principally over in a straightforward manner also to these pointwise and aggregated dissimilarities between the functions (rather than real points), and accordingly one calls them (pointwise/aggregated) divergences respectively distances, too.

[3] Measurable.

2.2 Divergences and Geometry

There are several ways how pointwise dissimilarities $d(\cdot,\cdot)$ respectively aggregated dissimilarities $D(\cdot,\cdot)$ between two functions $P := \big\{p(x)\big\}_{x\in\mathscr{X}}$ and $Q := \big\{q(x)\big\}_{x\in\mathscr{X}}$ can be connected with geometric issues. To start with an "all-encompassing view", following the lines of e.g. Birkhoff [14] and Millmann and Parker [50], one can build from any set $\mathscr{S}$, whose elements can be interpreted as "points", together with a collection $\mathscr{L}$ of non-empty subsets of $\mathscr{S}$, interpreted as "lines" (as a manifestation of a principle sort of structural connectivity between points), and an arbitrary *distance* $\mathfrak{d}(\cdot,\cdot)$ on $\mathscr{S}\times\mathscr{S}$, an axiomatic constructive framework of geometry which can be of far-reaching nature; therein, $\mathfrak{d}(\cdot,\cdot)$ plays basically the role of a marked ruler. Accordingly, each triplet $(\mathscr{S},\mathscr{L},\mathfrak{d}(\cdot,\cdot))$ forms a distinct "quantitative geometric system"; the most prominent classical case is certainly $\mathscr{S}=\mathbb{R}^2$ with $\mathscr{L}$ as the collection of all vertical and non-vertical lines, equipped with the Euclidean distance $\mathfrak{d}(\cdot,\cdot)$, hence generating the usual Euclidean geometry in the two-dimensional space. In the case that $\mathfrak{d}(\cdot,\cdot)$ is only an *asymmetric divergence* but not a distance anymore, we propose that some of the outcoming geometric building blocks have to be interpreted in a direction-based way (e.g. the use of $\mathfrak{d}(\cdot,\cdot)$ as a marked directed ruler, the construction of points of equal divergence from a center viewed as distorted directed spheres, etc.). For $d(\cdot,\cdot)$ one takes $\mathscr{S}\subset\mathbb{R}$ whereas for $D(\cdot,\cdot)$ one has to work with $\mathscr{S}$ being a family of real-valued functions on $\mathscr{X}$.

Secondly, from any *distance* $\mathfrak{d}(\cdot,\cdot)$ on a "sufficiently rich" set $\mathscr{S}$ and a finite number of (fixed or adaptively flexible) distinct "reference points" s_i $(i=1,\ldots,n)$ one can construct the corresponding Voronoi cells $V(s_i)$ by

$$V(s_i) := \{z\in\mathscr{S}:\ \mathfrak{d}(z,s_i)\leqslant\mathfrak{d}(z,s_j)\ \text{for all}\ j=1,\ldots,n\}.$$

This produces a tesselation (tiling) of $\mathscr{S}$ which is very useful for classification purposes. Of course, the geometric shape of these tesselations is of fundamental importance. In the case that $\mathfrak{d}(\cdot,\cdot)$ is only an *asymmetric divergence* but not a distance anymore, then $V(s_i)$ has to be interpreted as a directed Voronoi cell and then there is also the "reversely directed" alternative

$$\widetilde{V}(s_i) := \{z\in\mathscr{S}:\ \mathfrak{d}(s_i,z)\leqslant\mathfrak{d}(s_j,z)\ \text{for all}\ j=1,\ldots,n\}.$$

Recent applications where $\mathscr{S}\subset\mathbb{R}^d$ and $\mathfrak{d}(\cdot,\cdot)$ is a Bregman divergence or a more general conformal divergence, can be found e.g. in Boissonnat et al. [15], Nock et al. [64] (and the references therein), where they also deal with the corresponding adaption of k-nearest neighbour classification methods.

Thirdly, consider a "specific framework" where the functions $P := \widetilde{P}_{\theta_1} := \big\{\widetilde{p}_{\theta_1}(x)\big\}_{x\in\mathscr{X}}$ and $Q := \widetilde{P}_{\theta_2} := \big\{\widetilde{p}_{\theta_2}(x)\big\}_{x\in\mathscr{X}}$ depend on some parameters $\theta_1\in\Theta$, $\theta_2\in\Theta$, which reflect the strive for a complexity-reducing representation of "otherwise intrinsically complicated" functions P, Q. The way of dependence of the function (say) $\widetilde{p}_\theta(\cdot)$ on the underlying parameter θ from an appropriate space Θ

of e.g. manifold type, may show up directly e.g. via its operation/functioning as a relevant system-indicator, or it may be manifested implicitly e.g. such that $\widetilde{p}_\theta(\cdot)$ is the solution of an optimization problem with θ-involving constraints. In such a framework, one can induce divergences $D(\widetilde{P}_{\theta_1}, \widetilde{P}_{\theta_2}) =: f(\theta_1, \theta_2)$ and – under sufficiently smooth dependence – study their corresponding differential-geometric behaviour of $f(\cdot, \cdot)$ on Θ. An example is provided by the Kullback–Leibler divergence between two distributions of the same exponential family of distributions, which defines a Bregman divergence on the parameter space. This and related issues are subsumed in the research field of "information geometry"; for comprehensive overviews see e.g. Amari [3], Amari [1], Ay et al. [8]. Moreover, for recent connections between divergence-based information geometry and optimal transport the reader is e.g. referred to Pal and Wong [66, 67], Karakida and Amari [34], Amari et al. [2], Peyre and Cuturi [71], and the literature therein.

Further relations of divergences with other approaches to geometry can be overviewed e.g. from the wide-range-covering research-article collections in Nielsen and Bhatia [58], Nielsen and Barbaresco [55–57]. Finally, geometry also enters as a tool for visualizing quantitative effects on divergences.

2.3 Divergences and Uncertainty in Data

In general, data-uncertainty (including "deficiencies" like data incompleteness, fakery, unreliability, faultiness, vagueness, etc.) can enter the framework in various different ways. For instance, in situations where $x \in \mathscr{X}$ plays the role of an index (e.g. $\mathscr{X} = \{1, 2, \ldots, s\}$) and $p(x)$ describes the xth real-valued data-point, the uncertainty is typically[4] incorporated by adding a random argument $\omega \in \Omega$ to end up with the "vectors" $P(\omega) := \big\{p(x, \omega)\big\}_{x \in \mathscr{X}}$, $Q(\omega) := \big\{q(x, \omega)\big\}_{x \in \mathscr{X}}$ of random data points. Accordingly, one ends up with random-variable-type pointwise divergences $\omega \rightarrow d_\phi^{(j)}(p(x, \omega), q(x, \omega))$, $\omega \rightarrow d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x, \omega), q(x, \omega))$ $(x \in \mathscr{X})$ as well as with the random-variable-type "summarizing" divergences $\omega \rightarrow D_\phi^{(j)}(P(\omega), Q(\omega)) := \sum_{x \in \mathscr{X}} d_\phi^{(j)}(p(x, \omega), q(x, \omega)) \cdot \lambda(x)$ respectively $\omega \rightarrow D_\phi^{(j)}(P(\omega), Q(\omega)) := \int_{\mathscr{X}} d_\phi^{(j)}(p(x, \omega), q(x, \omega)) \,\mathrm{d}\lambda(x)$, as well as with $\omega \rightarrow D^{(5)}_{\phi, M_1, M_2, M_3}(P(\omega), Q(\omega)) := \sum_{x \in \mathscr{X}} d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x, \omega), q(x, \omega)) \cdot \lambda(x)$, resp. $\omega \rightarrow D^{(5)}_{\phi, M_1, M_2, M_3}(P(\omega), Q(\omega)) := \int_{\mathscr{X}} d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x, \omega), q(x, \omega)) \mathrm{d}\lambda(x)$. More generally, one can allow for random scales $m_1(x, \omega), m_2(x, \omega), m_3(x, \omega)$.

In other situations with finitely-many-elements carrying $\mathscr{X}$, the state x may e.g. describe a possible outcome $Y(\omega)$ of an uncertainty-prone observation of a quantity Y of interest and $p(x)$, $q(x)$ represent the corresponding probability mass functions ("discrete density functions") at x under two alternative probability mechanisms Pr, $\widetilde{Pr}$ (i.e. $p(x) = Pr[\{\omega \in \Omega : Y(\omega) = x\}]$, $q(x) = \widetilde{Pr}[\{\omega \in \Omega : Y(\omega) = x\}]$); as

[4] In a probabilistic approach rather than a chaos-theoretic approach.

already indicated above, $P := \big\{p(x)\big\}_{x \in \mathscr{X}}$ respectively $Q := \big\{q(x)\big\}_{x \in \mathscr{X}}$ serve then as a kind of "basis" for the computation of the probabilities $\sum_{x \in A} p(x)$ respectively $\sum_{x \in A} q(x)$ that an arbitrary event $\{\omega \in \Omega : Y(\omega) \in A\}$ ($A \subset \mathscr{X}$) occurs. Accordingly, the pointwise divergences $d_{\phi}^{(j)}(p(x), q(x))$, $d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))$ ($x \in \mathscr{X}$), and the aggregated divergences $D_{\phi}^{(j)}(P, Q) := \sum_{x \in \mathscr{X}} d_{\phi}^{(j)}(p(x), q(x))$, $D^{(5)}_{\phi,M_1,M_2,M_3}(P, Q) := \sum_{x \in \mathscr{X}} d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))$, $D^{(5)}_{\phi,M_1,M_2,M_3}(P, Q) := \int_{\mathscr{X}} d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x)) \, \mathrm{d}\lambda(x)$ can then be regarded as (nonnegative, reflexive) dissimilarities between the two alternative uncertainty-quantification-bases P and Q. Analogously, when e.g. $\mathscr{X} = \mathbb{R}^n$ is the n-dimensional Euclidean space and P, Q are classical probability density functions interpreted roughly via $p(x)\mathrm{d}x = Pr[\{\omega \in \Omega : Y(\omega) \in [x, x + \mathrm{d}x[\}$, $q(x)\mathrm{d}x = \widetilde{Pr}[\{\omega \in \Omega : Y(\omega) \in [x, x + \mathrm{d}x[\}$, then $d_{\phi}^{(j)}(p(x), q(x))$, $d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))$ ($x \in \mathscr{X}$), $D_{\phi}^{(j)}(P, Q) := \int_{\mathscr{X}} d_{\phi}^{(j)}(p(x), q(x)) \, \mathrm{d}x$, $D^{(5)}_{\phi,M_1,M_2,M_3}(P, Q) := \int_{\mathscr{X}} d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x)) \, \mathrm{d}x$ serve as dissimilarities between the two alternative uncertainty-quantification-bases P, Q.

Let us finally mention that in concrete applications, the "degree" of intrinsic data-uncertainty may be zero (deterministic), low (e.g. small random data contamination and small random deviations from a "basically" deterministic system, slightly noisy data, measurement errors) or high (forecast of the price of a stock in one year from now). Furthermore, the data may contain "high unusualnesses" ("surprising observations") such as outliers and inliers. All this should be taken into account when choosing or even designing the right type of divergence which have different sensitivity to such issues (see e.g. Kißlinger and Stummer [37] and the references therein).

2.4 Divergences, Information and Model Uncertainty

In the main spirit of this book on geometric structures of information, let us also connect the latter with dissimilarities in a wide sense which is appropriate enough for our ambitions of universal modeling. In correspondingly adapting some conception e.g. of Buckland [20] to our above-mentioned investigations, in the following we regard a density function (say) $p(\cdot)$ as a fundamental basis of information understood as quantified real – respectively hypothetical – knowledge which can be communicated about some particular (family of) subjects or (family of) events; according to this information-as-knowledge point of view, pointwise dissimilarities/divergences/distances $d(p(x), q(x))$ ($x \in \mathscr{X}$) respectively aggregated dissimilarities/divergences/distances $D(P, Q)$ quantify the proximity between the two information-bases $P := \big\{p(x)\big\}_{x \in \mathscr{X}}$ and $Q := \big\{q(x)\big\}_{x \in \mathscr{X}}$ in a directed/nonnegative directed/nonnegative symmetric way. Hence, $d(\cdot, \cdot)$ respectively $D(\cdot, \cdot)$ themselves can be seen as a higher-level information on pairs of information bases.

Divergences can be used for the quantification of information-concerning issues for model uncertainty (model risk) and exploratory model search in various different ways. For instance, suppose that we search for (respectively learn to understand) a true unknown density function $Q^{true} := \left\{q^{true}(x)\right\}_{x\in\mathscr{X}}$ of an underlying data-generating mechanism of interest, which is often supposed to be a member of a prefixed class $\mathscr{P}$ of "hypothetical model-candidate density functions"; frequently, this task is (e.g. for the sake of fast tractability) simplified to a setup of finding the true unknown parameter $\theta = \theta_0$ – and hence $Q^{true} = Q_{\theta_0}$ – within a parametric family $\mathscr{P} := \{Q_\theta\}_{\theta\in\Theta}$. Let us first consider the case where the data-generating mechanism of interest Q^{true} is purely deterministic and hence also all the candidates $Q \in \mathscr{P}$ are (taken to be) *not* of probability-density-function type. Although one has no intrinsic data-uncertainty, one faces another type of knowledge-lack called model-uncertainty. Then, one standard goal is to "track down" (respectively learn to understand) this true unknown Q^{true} respectively Q_{θ_0} by collecting and purpose-appropriately postprocessing some corresponding data observations. Accordingly, one attempts to design a density-function-construction rule (mechanism, algorithm) $data \rightarrow P^{data} := \left\{p^{data}(x)\right\}_{x\in\mathscr{X}}$ to produce data-derived information-basis-type replica of a "comparable principal form" as the anticipated Q^{true}. This rule should theoretically guarantee that P^{data} converges – with reasonable "operational" speed – to Q^{true} as the number N^{data} of data grows, which particularly implies that (say) $D(P^{data}, Q^{true})$ for some prefixed aggregated divergence $D(\cdot, \cdot)$ becomes close to zero "fast enough". On these grounds, one reasonable strategy to down-narrow the true unknown data-generating mechanism Q^{true} is to take a prefixed class $\mathscr{P}^{hyp}$ of hypothetical density-function models and compute $infodeg := \inf_{Q\in\mathscr{P}^{hyp}} D(P^{data}, Q)$ which in the light of the previous discussions can be interpreted as an "unnormalized degree of informative evidence of Q^{true} being a member of $\mathscr{P}^{hyp}$", or from a reversed point of view, as an "unnormalized degree of goodness of approximation (respectively fit) of the data-derived density function P^{data} through/by means of $\mathscr{P}^{hyp}$". Within this current paradigm, if $infodeg$ is too large (to be specified in a context-dependent, appropriately quantified sense by taking into account the size of N^{data}), then one has to repeat the same procedure with a different class $\widetilde{\mathscr{P}^{hyp}}$; on the other hand, if (and roughly only if) $infodeg$ is small enough then $\widehat{Q^{data}} := \arg\inf_{Q\in\mathscr{P}^{hyp}} D(P^{data}, Q)$ (which may not be unique) is "the most reasonable" approximation. This procedure is repeated recursively as soon as new data points are observed.

In contrast to the last paragraph, let us now cope with the case where the true unknown data-generating mechanism of interest is prone to uncertainties (i.e. is random, noisy, risk-prone) and hence Q^{true} as well as all the candidates $Q \in \mathscr{P}$ *are* of probability-density-function type. Even more, the data-derived information-basis-type replica $\omega \rightarrow data(\omega) \rightarrow P^{data(\omega)} := \left\{p^{data(\omega)}(x)\right\}_{x\in\mathscr{X}}$ of Q^{true} is now a density-function-valued (!) random variable; notice that in an above-mentioned "full-scenario" time-evolutionary context, this becomes a density-function-on-functions-valued random variable. Correspondingly, the above-mentioned procedure for the deterministic case has to be adapted and the notions of convergence and smallness have to be stochastified, which leads to the need of considerably more advanced techniques.

Another field of applying divergences to a context of synchronous model and data uncertainty is Bayesian sequential updating. In such a "doubly uncertain" framework, one deals with a parametric context of probability density functions $Q^{true} = Q_{\theta_0}$, $\mathscr{P} := \{Q_\theta\}_{\theta\in\Theta}$ where the uncertain knowledge about the parameter θ (to be learnt) is operationalized by replacing it with a random variable ϑ on Θ. Based on both (i) an initial prior distribution $Prior_1[\cdot] := Pr[\vartheta \in \cdot\,]$ of ϑ (with probability density function pdf $\theta \rightarrow prior_1(\theta)$) and (ii) observed data $data_1(\omega), \ldots, data_{N^{data}}(\omega)$ of number N^{data}, a posterior distribution $Post_1[\cdot, \omega] := Pr[\vartheta \in \cdot \,|\, data_1(\omega), \ldots, data_{N^{data}}(\omega);\ prior[\cdot]\,]$ of ϑ (with pdf $\theta \rightarrow post_1(\theta, \omega)$) is determined with (amongst other things) the help of the well-known Bayes formula. This procedure is repeated recursively with new incoming data input (block) $data_{N^{data}+1}$, where the new prior distribution $Prior_2[\cdot, \omega] := Post_1[\cdot, \omega]$ is chosen as the old posterior and the new posterior distribution is $Post_2[\cdot, \omega] := Pr[\vartheta \in \cdot \,|\, data_1, \ldots, data_{N^{data}}, data_{N^{data}+1};\ Prior_2[\cdot, \omega]\,]$ (with pdf $\theta \rightarrow post_2(\theta, \omega)$), etc. The corresponding (say) aggregated divergence $D(P(\omega), Q(\omega))$ between the probability-density-valued random variables $\omega \rightarrow P(\omega) := \big\{prior_2(\theta, \omega)\big\}_{\theta\in\Theta}$, and $\omega \rightarrow Q(\omega) := \big\{post_2(\theta, \omega)\big\}_{\theta\in\Theta}$ serves as "degree of informativity of the new data-point observation on the learning of the true unknown θ_0".

As another application in a "doubly uncertain" framework, divergences $D(P, Q)$ appear also in a dichotomous Bayesian testing problem between the two alternative probability densities functions P and Q, where $D(P, Q)$ represents an appropriate average (over prior probabilities) of the corresponding difference between the prior Bayes risk (prior minimal mean decision loss) and the posterior Bayes risk (posterior minimal mean decision loss). This, together with non-averaging versions and an interpretation of $D(P, Q)$ as a (weighted-average) statistical information measure in the sense of De Groot [29] can be found e.g. in Österreicher and Vajda [65]; see also Stummer [78–80], Liese and Vajda [42], Reid and Williamson [73]. In contrast of this employment of $D(P, Q)$ as quantifier of "decision risk reduction" respectively "model risk reduction" respectively "information gain", a different use of divergences $D(P, Q)$ in a "double uncertain" general Bayesian context of dichotomous loss-dependent decisions between arbitrary probability density functions P and Q can be found in Stummer and Vajda [81], where they achieve $D_{\phi_\alpha}(P, Q)$ (for some power functions ϕ_α cf. (5)) as upper and lower bound of the Bayes risk (minimal mean decision loss) itself and also give applications to decision making of time-continuous, non-stationary financial stochastic processes.

Divergences can be also employed to detect distributional changes in streams (respectively clouds) $(data_j)_{j\in\tau}$ of uncertain (random, noisy, risk-prone) data indexed by j from an arbitrary countable set τ (e.g. the integers, an undirected graph); a survey together with some general framework can be found in Kißlinger and Stummer [38]: the basic idea is to pick out two[5] non-identical, purpose-appropriately chosen subcollections respectively sample patterns (e.g. windows)

[5] Where one of them may e.g. stem from training data.

$data_{one}(\omega) := (data_{s_1}(\omega), \ldots, data_{s_{N_1}}(\omega))$, $data_{two}(\omega) := (data_{t_1}(\omega), \ldots, data_{t_{N_2}}(\omega))$, and to build from them data-derived probability-density functions $\omega \to data_{one}(\omega) \to P^{data_{one}(\omega)} := \big\{p^{data_{one}(\omega)}(x)\big\}_{x \in \mathscr{X}}$, $\omega \to data_{two}(\omega) \to P^{data_{two}(\omega)} := \big\{p^{data_{two}(\omega)}(x)\big\}_{x \in \mathscr{X}}$. If a correspondingly chosen (say) aggregated divergence $D\big(P^{data_{one}(\omega)}, P^{data_{two}(\omega)}\big)$ – which plays the role of a condensed change-score – is "significantly large" in the sense that it is large enough – compared to some sound threshold which within the model reflects the desired "degree of confidential plausibility" – then there is strong indication of a distributional change which we then "believe in". Notice that both components of the divergence $D(\cdot,\cdot)$ are now probability-density-function-valued random variables. The sound threshold can e.g. be derived from advanced random asymptotic theory.

From the above discussion it is clear that divergence-based model-uncertainty methods are useful tools in concrete applications for machine learning and artificial intelligence, see e.g. Collins et al. [25], Murata et al. [54], Banerjee et al. [9], Tsuda et al. [87], Cesa-Bianchi and Lugosi [21], Nock and Nielsen [63], Sugiyama et al. [85], Wu et al. [94], Nock et al. [62], Nielsen et al. [60], respectively Minka [51], Cooper et al. [26], Lizier [46], Zhang et al. [96], Chhogyal [22], Cliff et al. [23, 24].

3 General Framework

For the rest of this paper, we shall use the following

Main (i.e. non-locally used) Notation and Symbols

$\mathbb{R}, \mathbb{N}, \mathbb{R}^d$	Set of real respectively integer numbers respectively d-dimensional vectors
Θ, θ	Set of parameters, see p. 188
$\mathbb{1}$	Function with constant value 1
$\mathbf{1}_A(z) = \delta_z[A]$	Indicator function on the set A evaluated at data point z, which is equal to Dirac's one-point distribution on z evaluated at A
$\#A$	Number of elements in set A
$\mathscr{X}$; $\mathscr{X}_\#$	Space/set where data can take values in; space/set of countable size
$\mathscr{F}$	System of admissible events/data-collections (σ-algebra) on $\mathscr{X}$
λ	Reference measure/integrator/summator, see p. 160 & Sect. 3.1 on p. 165
λ-a.a.	λ-almost all, see p. 160
λ_L	Lebesgue measure ("Riemann-type" integrator), see p. 160, & Sect. 3.1
$\lambda_\#$	Counting measure ("classical summator"), see p. 160 & Sect. 3.1 on p. 165
$P := \big\{p(x)\big\}_{x \in \mathscr{X}}$	Function from which the divergence/dissimilarity is measured from, see p. 160
$Q := \big\{q(x)\big\}_{x \in \mathscr{X}}$	Function to which the divergence/dissimilarity is measured to, see p. 160
$M_i := \big\{m_i(x)\big\}_{x \in \mathscr{X}}$	Scaling function ($i = 1, 2$) respec. aggregation function ($i = 3$), see p. 161, (1) and paragraph (I1) thereafter, as well as Sect. 3.3 on p. 170
$p(\cdot), q(\cdot), m_i(\cdot)$,	Alternative representations of P, Q, M_i
$R := \big\{r(x)\big\}_{x \in \mathscr{X}}$	Function used for the aggregation function $m_3(\cdot)$, see Sect. 3.3.1 on p. 171
W_i	Connector function of the form $W_i := \big\{w_i(x, y, z)\big\}_{x,y,z \in \ldots}$, for adaptive scaling and aggregation functions $m_i(x) = w_i(x, p(x), q(x))$ $(i = 1, 2, 3)$, see e.g. Assumption 2 on p. 163 and Sect. 3.3.1.3 on p. 181

$\mathbb{P}, \mathbb{Q}, \mathbb{M}_i, \mathbb{W}_i$	Functions with $\mathbb{p}(x) \geqslant 0$, $\mathbb{q}(x) \geqslant 0$, $\mathbb{m}_i(x) \geqslant 0$, $\mathbb{w}_i(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$
$\mathbb{Q}^{\chi} := \{\mathbb{q}^{\chi}(x)\}_{x \in \mathscr{X}}$	Function for the aggregation function $m_3(\cdot)$, see Sect. 4.2 on p. 184, (73)
$\vec{\mathbb{P}}, \vec{\mathbb{Q}}$	λ-probability density functions (incl. probability mass functions for $\lambda = \lambda_{\#}$), i.e. for which $\vec{\mathbb{p}}(x) \geqslant 0$, $\vec{\mathbb{q}}(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$ and $\int_{\mathscr{X}} \vec{\mathbb{p}}(x) \, \mathrm{d}\lambda(x) = 1$, see Remark 2 on p. 172
$\vec{\mathbb{Q}}_{\theta} := \{\vec{\mathbb{q}}_{\theta}(x)\}_{x \in \mathscr{X}}$	λ-probab. density function which depends on a parameter $\theta \in \Theta$, see p. 188
$\mathscr{R}(\frac{P}{M_1})$	Range (image) of the function $\{\frac{p(x)}{m_1(x)}\}_{x \in \mathscr{X}}$, see paragraph (I2) on p. 161
$\mathscr{R}(Y_1, \ldots, Y_N)$	Range (image) of the random variables $Y_1, \ldots, Y_N$, see p. 182
$\widetilde{\vec{\mathbb{Q}}}_{\theta} := \{\widetilde{\vec{\mathbb{q}}}_{\theta}(x)\}_{x \in \mathscr{X}}$	λ-probab. density function (modification of $\vec{\mathbb{Q}}_{\theta}$) defined by $\widetilde{\vec{\mathbb{q}}}_{\theta}(x) := \vec{\mathbb{q}}_{\theta}(x) \cdot (1 - \mathbf{1}_{\mathscr{R}(Y_1(\omega), \ldots, Y_N(\omega))}(x))$, see p. 191
$\phi := \{\phi(t)\}_{t \in]a,b[}$	Divergence generator, a convex real-valued function on $]a, b[$, see p. 161, (1) and paragraph (I2), as well as Sect. 3.2 on p. 165
$\Phi(]a, b[)$;	Class of all such ϕ, see paragraph (I2) on p. 161
$\overline{\phi} := \{\phi(t)\}_{t \in [a,b]}$	Continuous extension of ϕ on $[a, b]$, with $\overline{\phi}(t) = \phi(t)$ for all $t \in]a, b[$, see (I2)
$\phi'_{+,c}(t)$	c-weighted mixture of left-hand and right-hand derivative of ϕ at t, see (I2)
$\Phi_{C_1}(]a, b[)$	Subclass of everywhere continuously differentiable ϕ, with derivative $\phi'(t)$ (being equal to $\phi'_{+,c}(t)$ for all $c \in [0, 1]$), see (I2) on p. 161
ϕ_{α}	α-power-function type divergence generator, see (5) on p. 166, (14), (18), (19)
ϕ_{TV}	Generator of total variation distance, see (31) on p. 169
ϕ_{ie}	Divergence generator with interesting effects, see (35) on p. 170
$\psi_{\phi,c}$	Function given by $\psi_{\phi,c}(s, t) := \phi(s) - \phi(t) - \phi'_{+,c}(t) \cdot (s - t) \geqslant 0$, see (I2)
$\overline{\psi_{\phi,c}}$	Bivariate extension of $\psi_{\phi,c}$, see (I2) on p. 161
$\overline{\int}_{\mathscr{X}} \ldots, \overline{\sum}_{\mathscr{X}} \ldots$	Integral/sum over extension of integrand/summand . . ., see (I2) & (2) on p. 165
$D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q)$	Divergence between two functions P (scaled by M_1) and Q (scaled by M_2), generated by ϕ and weight c, and aggregated by $\mathbb{M}_3$ and λ, see (1) on p. 161
$D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q)$	As above, but with $\phi \in \Phi_{C_1}(]a, b[)$ and obsolete c, see Sect. 3.2 on p. 165
$D_{\lambda}(P, Q)$	General λ-aggregated divergence, see p. 189, respectively pseudo-divergence, see Definition 2 on p. 195
$D_{\mathbb{M},\lambda}(\vec{\mathbb{P}}, \vec{\mathbb{Q}})$	Pointwise decomposable pseudo-divergence, scaled by $\mathbb{M}$ and aggregated by $\mathbb{M}$ and λ, see Sect. 4.6 on p. 200
NN0, NN1	Nonnegativity setup 0 respectively 1, see p. 166 resp. p. 171
$\mathfrak{P}^{\mathbb{R}\cdot\lambda}, \mathfrak{Q}^{\mathbb{R}\cdot\lambda}, \mathfrak{M}^{\mathbb{R}\cdot\lambda}$	Measures with λ-densities $\mathbb{p}(\cdot) \cdot \mathbb{r}(\cdot)$, $\mathbb{q}(\cdot) \cdot \mathbb{r}(\cdot)$, $\mathbb{m}(\cdot) \cdot \mathbb{r}(\cdot)$, see Remark 2 on p. 171
$\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}, \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}$	Probability measures (distributions) with λ-densities $\vec{\mathbb{p}}(\cdot)$, $\vec{\mathbb{q}}(\cdot)$, see Remark 2
$\mathscr{Q}^{\lambda_2}_{\Theta}, \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_2}_{\theta}$	Class of probability measures with λ_2-densities $\vec{\mathbb{q}}_{\theta}(\cdot)$ with parameter $\theta \in \Theta$, see p. 188
$\vec{\mathfrak{P}}^{emp}_{N}, \vec{\mathbb{P}}^{emp}_{N}, \vec{\mathbb{p}}^{emp}_{N}(\cdot)$	Data-derived empirical (probability) distribution, and probability mass function ($\lambda_{\#}$-density) thereof, see Remark 2 on p. 172
$\vec{\mathfrak{P}}^{\overline{emp}(\omega)}_{N}, \vec{\mathbb{P}}^{\overline{emp}(\omega)}_{N}$	Data-derived "extended" empirical (probability) distribution, and probability mass function thereof, see (85) on p. 190 and thereafter
DPD, CASD	Density-power divergences (see p. 174), Csiszar–Ali–Silvey divergences (see p. 177)
$\ell i_1, \phi^*(0), \ell i_2, \ell i_3$	Certain limits, see (50), (71), (72)
$\mathbb{P} \perp \mathbb{Q}$	The functions $\mathbb{P}$, $\mathbb{Q}$ are "essentially different", see (64) to (66) and thereafter
$\mathbb{P} \not\perp \mathbb{Q}$	Negation of $\mathbb{P} \perp \mathbb{Q}$, see p. 192
$\mathbb{P} \sim \mathbb{Q}$	The functions $\mathbb{P}$, $\mathbb{Q}$ are "equivalent" (concerning zeros), see (80)
$\mathbb{P} \nsim \mathbb{Q}$	Negation of $\mathbb{P} \sim \mathbb{Q}$, see p. 195
$\widehat{\theta}_{N,D_{\lambda_2}}$	Minimum-divergence estimator ("approximator") of the true unknown parameter θ_0, based on N data observations, see (82) on p. 189

$\widehat{\theta}_{N,D_{\lambda_{\#}}}, \widehat{\theta}_{N,D_{\lambda}}$	Certain minimum-divergence estimators, see (83), (86)
$\widehat{\theta}_{N,decD_{\lambda}}, \ \widehat{\theta}_{N,decD_{\vec{\mathbb{Q}}_{\tau},\lambda}}$	Certain minimum-divergence estimators, see (107), (123)
$\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}$	Certain minimum-divergence estimator, see (135)
$\mathscr{P}^{\lambda}$	Certain class of nonnegative, mutually equivalent functions, see p. 194
$\mathscr{P}^{\lambda\not\sim}, \widetilde{\mathscr{P}}^{\lambda}$	Certain classes of nonnegative functions, see p. 194
$\mathscr{P}^{\lambda}_{\Theta}, \mathscr{P}^{\lambda\perp}_{emp}, \mathscr{P}^{\lambda}_{\Theta,emp}$	Certain classes of nonnegative functions, see p. 195
$\mathfrak{D}^0, \mathfrak{D}^1, \rho_{\mathbb{Q}}$	Functionals and mapping for decomposable pseudo-divergences, see Definition 3 on p. 195
$\psi^{dec}, \psi^0, \psi^1, \rho$	Mappings for pointwise decomposable pseudo-divergences, see Definition 3 on p. 196
h_0, h_1, h_2	Mappings for pointwise decomposable pseudo-divergences, see Definition 3 on p. 196
ψ^{dec}_{m}	Perspective function of ψ^{dec}, see (120)

New Divergence Toolkit

In the above Sect. 2, we have motivated that for many different tasks within a broad spectrum of situations, it is useful to employ divergences as "directed distances", including distances as their symmetric special case. For the rest of the paper, we shall only deal with aggregated forms of divergences, and thus drop the attribute "aggregated" from now on. In the following, we present a fairly universal, flexible, multi-component system of divergences by adapting and widening the concept of scaled Bregman divergences of Stummer [81] and Stummer and Vajda [84] to the current context of arbitrary (measurable) functions. To begin with, let us assume that the modeled respectively observed (random) data take values in a state space $\mathscr{X}$ (with at least two distinct values), equipped with a system $\mathscr{F}$ of admissible events (σ-algebra) and a σ-finite measure λ (e.g. the Lebesgue measure, the counting measure, etc.). Furthermore, we suppose that $x \rightarrow p(x) \in [-\infty, \infty]$ and $x \rightarrow q(x) \in [-\infty, \infty]$ are (correspondingly measurable) functions on $\mathscr{X}$ which satisfy $p(x) \in]-\infty, \infty[$, $q(x) \in]-\infty, \infty[$ for λ-almost all (abbreviated as λ-a.a.) $x \in \mathscr{X}$.[6] To address the entire functions as objects we write $P := \big\{p(x)\big\}_{x\in\mathscr{X}}$, $Q := \big\{q(x)\big\}_{x\in\mathscr{X}}$ and alternatively sometimes also $p(\cdot), q(\cdot)$. To better highlight the very important special case of *λ-probability density functions* – where $p(x) \geqslant 0, q(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$ and $\int_{\mathscr{X}} p(x)\,\mathrm{d}\lambda(x) = 1, \int_{\mathscr{X}} q(x)\,\mathrm{d}\lambda(x) = 1$ – we use the notation $\vec{\mathbb{P}}, \vec{\mathbb{p}}, \vec{\mathbb{Q}}, \vec{\mathbb{q}}$ instead of P, p, Q, q (where $\overrightarrow{\cdots}$ symbolizes a lying 1). For instance, if $\lambda = \lambda_L$ is the Lebesgue measure on the s-dimensional Euclidean space $\mathscr{X} = \mathbb{R}^s$, then $\vec{\mathbb{P}}, \vec{\mathbb{Q}}$ are "classical" (e.g. Gaussian) probability density functions. In contrast, in the *discrete setup* where the state space (i.e. the set of all possible data points) $\mathscr{X} = \mathscr{X}_{\#}$ has countably many elements and $\lambda := \lambda_{\#}$ is the counting measure (i.e., $\lambda_{\#}[\{x\}] = 1$ for all $x \in \mathscr{X}_{\#}$), then $\vec{\mathbb{P}}, \vec{\mathbb{Q}}$ are probability mass functions and (say) $\vec{\mathbb{p}}(x)$ can be interpreted as probability that the data point x is taken by the underlying random (uncertainty-prone) mechanism. If $p(x) \geqslant 0$, $q(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$ (but not necessarily with the restrictions $\int_{\mathscr{X}} p(x)\,\mathrm{d}\lambda(x) = 1 = \int_{\mathscr{X}} q(x)\,\mathrm{d}\lambda(x)$) then we write $\mathbb{P}, \mathbb{Q}, \mathbb{p}, \mathbb{q}$ instead of P, p, Q, q.

[6] This means that there exists a $N \in \mathscr{F}$ with $\lambda[N] = 0$ (where the empty set $N = \emptyset$ is allowed) such that for all $x \in \mathscr{X}\backslash\{N\}$ (say) $p(x) \in]-\infty, \infty[$ holds.

Back to generality, we quantify the dissimilarity between the two functions P,Q in terms of divergences $D^c_\beta(P, Q)$ with $\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$, defined by

$$\begin{aligned} 0 &\leqslant D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q) \\ &:= \int_{\mathscr{X}} \Big[\phi\Big(\tfrac{p(x)}{m_1(x)}\Big) - \phi\Big(\tfrac{q(x)}{m_2(x)}\Big) - \phi'_{+,c}\Big(\tfrac{q(x)}{m_2(x)}\Big) \cdot \Big(\tfrac{p(x)}{m_1(x)} - \tfrac{q(x)}{m_2(x)}\Big)\Big] \cdot \mathbb{m}_3(x)\,\mathrm{d}\lambda(x) \qquad (1) \end{aligned}$$

(see Stummer [81], Stummer and Vajda [84] for the case $c = 1, m_1(x) = m_2(x) = \mathbb{m}_3(x)$). Here, we use:

(I1) (measurable) *scaling functions* $m_1 : \mathscr{X} \to [-\infty, \infty]$ and $m_2 : \mathscr{X} \to [-\infty, \infty]$ as well as a nonnegative (measurable) *aggregating function* $\mathbb{m}_3 : \mathscr{X} \to [0, \infty]$ such that $m_1(x) \in]-\infty, \infty[$, $m_2(x) \in]-\infty, \infty[$, $\mathbb{m}_3(x) \in [0, \infty[$ for λ-a.a. $x \in \mathscr{X}$.[7] In accordance with the above notation, we use the symbols $M_i := \big\{m_i(x)\big\}_{x\in\mathscr{X}}$ respectively $m_i(\cdot)$ to refer to the entire functions, and $\mathbb{M}_{\mathbb{i}}$, $\mathbb{m}_{\mathbb{i}}(\cdot)$ when they are nonnegative as well as $\vec{\mathbb{M}}_i$, $\vec{\mathbb{m}}_i(\cdot)$ when they manifest λ-probability density functions. Furthermore, let us emphasize that we allow for / cover adaptive situations in the sense that all three functions $m_1(x)$, $m_2(x)$, $\mathbb{m}_3(x)$ (evaluated at x) may also depend on $p(x)$ and $q(x)$.

(I2) the so-called "divergence-generator" ϕ which is a continuous, convex (finite) function $\phi : E \to]-\infty, \infty[$ on some appropriately chosen open interval $E =]a, b[$ such that $[a, b]$ covers (at least) the union $\mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big)$ of both ranges $\mathscr{R}\big(\frac{P}{M_1}\big)$ of $\big\{\frac{p(x)}{m_1(x)}\big\}_{x\in\mathscr{X}}$ and $\mathscr{R}\big(\frac{Q}{M_2}\big)$ of $\big\{\frac{q(x)}{m_2(x)}\big\}_{x\in\mathscr{X}}$; for instance, $E =]0, 1[$, $E =]0, \infty[$ or $E =]-\infty, \infty[$; the class of all such functions will be denoted by $\Phi(]a, b[)$. Furthermore, we assume that ϕ is continuously extended to $\overline{\phi} : [a, b] \to [-\infty, \infty]$ by setting $\overline{\phi}(t) := \phi(t)$ for $t \in]a, b[$ as well as $\overline{\phi}(a) := \lim_{t\downarrow a} \phi(t)$, $\overline{\phi}(b) := \lim_{t\uparrow b} \phi(t)$ on the two boundary points $t = a$ and $t = b$. The latter two are the only points at which infinite values may appear. Moreover, for any fixed $c \in [0, 1]$ the (finite) function $\phi'_{+,c} :]a, b[\to]-\infty, \infty[$ is well-defined by $\phi'_{+,c}(t) := c \cdot \phi'_+(t) + (1 - c) \cdot \phi'_-(t)$, where $\phi'_+(t)$ denotes the (always finite) right-hand derivative of ϕ at the point $t \in]a, b[$ and $\phi'_-(t)$ the (always finite) left-hand derivative of ϕ at $t \in]a, b[$. If $\phi \in \Phi(]a, b[)$ is also continuously differentiable – which we denote by $\phi \in \Phi_{C_1}(]a, b[)$ – then for all $c \in [0, 1]$ one gets $\phi'_{+,c}(t) = \phi'(t)$ $(t \in]a, b[)$ and in such a situation we always suppress the obsolete indices c, $+$ in the corresponding expressions. We also employ the continuous continuation $\overline{\phi'_{+,c}} : [a, b] \to [-\infty, \infty]$ given by $\overline{\phi'_{+,c}}(t) := \phi'_{+,c}(t)$ $(t \in]a, b[)$, $\overline{\phi'_{+,c}}(a) := \lim_{t\downarrow a} \phi'_{+,c}(t)$, $\overline{\phi'_{+,c}}(b) := \lim_{t\uparrow b} \phi'_{+,c}(t)$. To explain the precise meaning of (1), we also make use of the (finite, nonnegative) function $\psi_{\phi,c} :]a, b[\times]a, b[\to [0, \infty[$ given by $\psi_{\phi,c}(s, t) := \phi(s) - \phi(t) - \phi'_{+,c}(t) \cdot (s - t) \geqslant 0$ $(s, t \in]a, b[)$. To extend this to a lower semi-continuous

[7] As an example, let $\mathscr{X} = \mathbb{R}$, $\lambda = \lambda_L$ be the Lebesgue measure (and hence, except for rare cases, the integral turns into a Riemann integral) and $\vec{\mathbb{m}}_1(x) := \frac{1}{2} \cdot x^{-1/2} \cdot \mathbf{1}_{[0,1]}(x) \geqslant 0$; since $\int_{\mathscr{X}} \vec{\mathbb{m}}_1(x)\,\mathrm{d}\lambda(x) = 1$ this qualifies as a probability density and thus is a possible candidate for $\vec{\mathbb{m}}_1(x) = \vec{\mathbb{q}}(x)$ in Sect. 3.3.1.2 below.

function $\overline{\psi_{\phi,c}} : [a,b] \times [a,b] \rightarrow [0,\infty]$ we proceed as follows: firstly, we set $\overline{\psi_{\phi,c}}(s,t) := \psi_{\phi,c}(s,t)$ for all $s,t \in]a,b[$. Moreover, since for fixed $t \in]a,b[$, the function $s \rightarrow \psi_{\phi,c}(s,t)$ is convex and continuous, the limit $\overline{\psi_{\phi,c}}(a,t) := \lim_{s \rightarrow a} \psi_{\phi,c}(s,t)$ always exists and (in order to avoid overlines in (1)) will be interpreted/abbreviated as $\phi(a) - \phi(t) - \phi'_{+,c}(t) \cdot (a-t)$. Analogously, for fixed $t \in]a,b[$ we set $\overline{\psi_{\phi,c}}(b,t) := \lim_{s \rightarrow b} \psi_{\phi,c}(s,t)$ with corresponding shorthand notation $\phi(b) - \phi(t) - \phi'_{+,c}(t) \cdot (b-t)$. Furthermore, for fixed $s \in]a,b[$ we interpret $\phi(s) - \phi(a) - \phi'_{+,c}(a) \cdot (s-a)$ as

$$\begin{aligned}\overline{\psi_{\phi,c}}(s,a) := & \big\{\phi(s) - \overline{\phi'_{+,c}}(a) \cdot s + \lim_{t \rightarrow a} \big(t \cdot \overline{\phi'_{+,c}}(a) - \phi(t)\big)\big\} \cdot \mathbf{1}_{]-\infty,\infty[}\big(\overline{\phi'_{+,c}}(a)\big) \\ & + \infty \cdot \mathbf{1}_{\{-\infty\}}\big(\overline{\phi'_{+,c}}(a)\big)\,,\end{aligned}$$

where the involved limit always exists but may be infinite. Analogously, for fixed $s \in]a,b[$ we interpret $\phi(s) - \phi(b) - \phi'_{+,c}(b) \cdot (s-b)$ as

$$\begin{aligned}\overline{\psi_{\phi,c}}(s,b) := & \big\{\phi(s) - \overline{\phi'_{+,c}}(b) \cdot s + \lim_{t \rightarrow b} \Big(t \cdot \overline{\phi'_{+,c}}(b) - \phi(t)\Big)\big\} \cdot \mathbf{1}_{]-\infty,\infty[}\big(\overline{\phi'_{+,c}}(b)\big) \\ & + \infty \cdot \mathbf{1}_{\{+\infty\}}\big(\overline{\phi'_{+,c}}(b)\big)\,,\end{aligned}$$

where again the involved limit always exists but may be infinite. Finally, we always set $\overline{\psi_{\phi,c}}(a,a) := 0$, $\overline{\psi_{\phi,c}}(b,b) := 0$, and $\overline{\psi_{\phi,c}}(a,b) := \lim_{s \rightarrow a} \overline{\psi_{\phi,c}}(s,b)$, $\overline{\psi_{\phi,c}}(b,a) := \lim_{s \rightarrow b} \overline{\psi_{\phi,c}}(s,a)$. Notice that $\overline{\psi_{\phi,c}}(\cdot,\cdot)$ is lower semi-continuous but not necessarily continuous. Since ratios are ultimately involved, we also consistently take $\overline{\psi_{\phi,c}}\big(\frac{0}{0},\frac{0}{0}\big) := 0$. Taking all this into account, we interpret $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ as $\int_{\mathcal{X}} \overline{\psi_{\phi,c}}\big(\frac{p(x)}{m_1(x)},\frac{q(x)}{m_2(x)}\big) \mathrm{m}_3(x)\,\mathrm{d}\lambda(x)$ at first glance (see further investigations in Assumption 2 below), and use the (in lengthy examples) less clumsy notation $\overline{\int}_{\mathcal{X}} \psi_{\phi,c}\big(\frac{p(x)}{m_1(x)},\frac{q(x)}{m_2(x)}\big) \mathrm{m}_3(x)\,\mathrm{d}\lambda(x)$ as a shortcut for the implicitly involved boundary behaviour. □

Notice that despite of the "difference-structure" in the integrand of (1), the splitting of the integral into differences of several "autonomous" integrals may not always be feasible due to the possible appearance of differences between infinite integral values. Furthermore, there is non-uniqueness in the construction (1); for instance, one (formally) gets $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) = D^c_{\tilde{\phi},M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ for any $\tilde{\phi}(t) := \phi(t) + c_1 + c_2 \cdot t \quad (t \in E)$ with $c_1, c_2 \in \mathbb{R}$. Moreover, there exist "essentially different" pairs $(\phi, \mathbb{M})$ and $(\breve{\phi}, \breve{\mathbb{M}})$ (where $\phi(t) - \breve{\phi}(t)$ is nonlinear in t) for which $D^c_{\phi,\mathbb{M},\mathbb{M},\mathbb{M},\lambda}(P,Q) = D^c_{\breve{\phi},\breve{\mathbb{M}},\breve{\mathbb{M}},\breve{\mathbb{M}},\lambda}(P,Q)$ (see e.g. [37]). Let us also mention that we could further generalize (1) by adapting the divergence concept of Stummer and Kißlinger [82] who also deal even with non-convex non-concave divergence generators ϕ; for the sake of brevity, this is omitted here.

Notice that by construction we obtain the following important assertion:

Theorem 1 *Let $\phi \in \Phi(]a,b[)$ and $c \in [0,1]$. Then there holds $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) \geqslant 0$ with equality if $\frac{p(x)}{m_1(x)} = \frac{q(x)}{m_2(x)}$ for λ-almost all $x \in \mathscr{X}$. Depending on the concrete situation, $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ may take infinite value.*

To get "sharp identifiability" (i.e. reflexivity) one needs further assumptions on $\phi \in \Phi(]a,b[)$, $c \in [0,1]$. As a motivation, consider the case where $\mathbb{m}_3(x) \equiv 1$ and $\phi \in \Phi(]a,b[)$ is affine linear on the whole interval $]a,b[$, and hence its extension $\overline{\phi}$ is affine-linear on $[a,b]$. Accordingly, one gets for the integrand-builder $\overline{\psi_{\phi,c}}(s,t) \equiv 0$ and hence $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) = \int_{\mathscr{X}} \overline{\psi_{\phi,c}}\big(\frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)\, \mathrm{d}\lambda(x) = 0$ even in cases where $\frac{p(x)}{m_1(x)} \neq \frac{q(x)}{m_2(x)}$ for λ-a.a. $x \in \mathscr{X}$. In order to avoid such and similar phenomena, we use the following set of requirements:

Assumption 2 Let $c \in [0,1]$, $\phi \in \Phi(]a,b[)$ and $\mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big) \subset [a,b]$. The aggregation function is supposed to be of the form $\mathbb{m}_3(x) = \mathbb{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)$ for some (measur.) function $\mathbb{w}_3 : \mathscr{X} \times [a,b] \times [a,b] \to [0,\infty]$. Moreover, for all $s \in \mathscr{R}\big(\frac{P}{M_1}\big)$, all $t \in \mathscr{R}\big(\frac{Q}{M_2}\big)$ and λ-a.a. $x \in \mathscr{X}$, let the following conditions hold:

(a) ϕ is strictly convex at t;
(b) if ϕ is differentiable at t and $s \neq t$, then ϕ is not affine-linear on the interval $[\min(s,t), \max(s,t)]$ (i.e. between t and s);
(c) if ϕ is not differentiable at t, $s > t$ and ϕ is affine linear on $[t,s]$, then we exclude $c = 1$ for the ("globally/universally chosen") subderivative $\phi'_{+,c}(\cdot) = c \cdot \phi'_{+}(\cdot) + (1-c) \cdot \phi'_{-}(\cdot)$;
(d) if ϕ is not differentiable at t, $s < t$ and ϕ is affine linear on $[s,t]$, then we exclude $c = 0$ for $\phi'_{+,c}(\cdot)$;
(e) $\mathbb{w}_3(x,s,t) < \infty$;
(f) $\mathbb{w}_3(x,s,t) > 0$ if $s \neq t$;
(g) $\mathbb{w}_3(x,a,a) \cdot \psi_{\phi,c}(a,a) := 0$ by convention (even in cases where the function $\mathbb{w}_3(x,\cdot,\cdot) \cdot \psi_{\phi,c}(\cdot,\cdot)$ is not continuous on the boundary point (a,a));
(h) $\mathbb{w}_3(x,b,b) \cdot \psi_{\phi,c}(b,b) := 0$ by convention (even in cases where the function $\mathbb{w}_3(x,\cdot,\cdot) \cdot \psi_{\phi,c}(\cdot,\cdot)$ is not continuous on the boundary point (b,b));
(i) $\mathbb{w}_3(x,a,t) \cdot \psi_{\phi,c}(a,t) > 0$, where $\mathbb{w}_3(x,a,t) \cdot \psi_{\phi,c}(a,t) := \lim_{s \to a} \mathbb{w}_3(x,s,t) \cdot \psi_{\phi,c}(s,t)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,a,t) \cdot \psi_{\phi,c}(a,t) := 1$ (or any other strictly positive constant);
(j) $\mathbb{w}_3(x,b,t) \cdot \psi_{\phi,c}(b,t) > 0$, where $\mathbb{w}_3(x,b,t) \cdot \psi_{\phi,c}(b,t)$ is analogous to (i);
(k) $\mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a) > 0$, where $\mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a) := \lim_{t \to a} \mathbb{w}_3(x,s,t) \cdot \psi_{\phi,c}(s,t)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a) := 1$ (or any other strictly positive constant);
(l) $\mathbb{w}_3(x,s,b) \cdot \psi_{\phi,c}(s,b) > 0$, where $\mathbb{w}_3(x,s,b) \cdot \psi_{\phi,c}(s,b)$ is analogous to (k);
(m) $\mathbb{w}_3(x,a,b) \cdot \psi_{\phi,c}(a,b) > 0$, where $\mathbb{w}_3(x,a,b) \cdot \psi_{\phi,c}(a,b) := \lim_{s \to a} \mathbb{w}_3(x,s,b) \cdot \psi_{\phi,c}(s,b)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,a,b) \cdot \psi_{\phi,c}(a,b) := 1$ (or any other strictly positive constant);
(n) $\mathbb{w}_3(x,b,a) \cdot \psi_{\phi,c}(b,a) > 0$, where $\mathbb{w}_3(x,b,a) \cdot \psi_{\phi,c}(b,a) := \lim_{s \to b} \mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,b,a) \cdot \psi_{\phi,c}(b,a) := 1$ (or any other strictly positive constant). □

Under Assumption 2, we always interpret the corresponding divergence

$$D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) := D^{c}_{\phi,M_1,M_2,\mathbb{W}_3,\lambda}(P,Q) :=$$
$$:= \overline{\int}_{\mathscr{X}} \mathbbm{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big) \cdot \big[\phi\big(\frac{p(x)}{m_1(x)}\big) - \phi\big(\frac{q(x)}{m_2(x)}\big)$$
$$-\phi'_{+,c}\big(\frac{q(x)}{m_2(x)}\big) \cdot \big(\frac{p(x)}{m_1(x)} - \frac{q(x)}{m_2(x)}\big)\big]\, \mathrm{d}\lambda(x)$$

as $\int_{\mathscr{X}} \overline{\mathbbm{w}_3 \cdot \psi_{\phi,c}}\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)\, \mathrm{d}\lambda(x)$, where $\overline{\mathbbm{w}_3 \cdot \psi_{\phi,c}}(x,s,t)$ denotes the extension of the function $\mathscr{X} \times]a,b[\times]a,b[\ni (x,s,t) \rightarrow \mathbbm{w}_3(x,s,t) \cdot \psi_{\phi,c}(s,t)$ on $\mathscr{X} \times [a,b] \times [a,b]$ according to the conditions (g) to (n) above.

Remark 1 (a) We could even work with a weaker assumption obtained by replacing s with $\frac{p(x)}{m_1(x)}$ as well as t with $\frac{q(x)}{m_2(x)}$ and by requiring that then the correspondingly plugged-in conditions (a) to (n) hold for λ-a.a. $x \in \mathscr{X}$.
(b) Notice that our above context subsumes aggregation functions of the form $\mathbbm{m}_3(x) = \tilde{\mathbbm{w}}_3(x, p(x), q(x), m_1(x), m_2(x))$ with $\tilde{\mathbbm{w}}_3(x, z_1, z_2, z_3, z_4)$ having appropriately imbeddable behaviour in its arguments x, z_1, z_2, z_3, z_4, the outcoming ratios $\frac{z_1}{z_3}, \frac{z_2}{z_4}$ and possible boundary values thereof. □

The following requirement is stronger than the "model-individual/dependent" Assumption 2 but is more "universally applicable" (amongst *all* models such that $\mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big) \subset [a,b]$, take e.g. $E =]a,b[$ as $E =]0,\infty[$ or $E =]-\infty,\infty[$):

Assumption 3 Let $c \in [0,1]$, $\phi \in \Phi(]a,b[)$ on some fixed $]a,b[\in]-\infty,+\infty[$ such that $]a,b[\supset \mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big)$. The aggregation function is of the form $\mathbbm{m}_3(x) = \mathbbm{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)$ for some (measurable) function $\mathbbm{w}_3 : \mathscr{X} \times [a,b] \times [a,b] \rightarrow [0,\infty]$. Furthermore, for all $s \in]a,b[$, $t \in]a,b[$ and λ-a.a. $x \in \mathscr{X}$, the conditions (a) to (n) of Assumption 2 hold.

Important examples in connection with the Assumptions 2, 3 will be given in Sect. 3.2 (for ϕ) and Sect. 3.3 (for m_1, m_2, $\mathbbm{w}_3$) below. With these assumptions at hand, we obtain the following non-negativity and reflexivity assertions:

Theorem 4 *Let the Assumption 2 be satisfied. Then there holds:*
(1) $D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) \geqslant 0$. *Depending on the concrete situation,* $D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ *may take infinite value.*
(2) $D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) = 0$ *if and only if* $\frac{p(x)}{m_1(x)} = \frac{q(x)}{m_2(x)}$ *for λ-a.a. $x \in \mathscr{X}$.*

Theorem 4 – whose proof will be given in the appendix – says that $D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ is indeed a "proper" divergence under the Assumption 2. Hence, the latter will be assumed for the rest of the paper, unless stated otherwise: for instance, we shall sometimes work with the stronger Assumption 3; thus, for more comfortable reference, we state explicitly

Corollary 1 *Under the more universally applicable Assumption 3, the Assertions (1) and (2) of Theorem 4 hold.*

Under some non-obvious additional constraints on the functions P, Q it may be possible to show the Assertions (1), (2) of Theorem 4 by even dropping the purely generator-concerning Assumptions 2(b) to (d); see e.g. Sect. 3.3.1.2 below. In the following, we discuss several important features and special cases of $\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$ in a well-structured way. Let us start with the latter.

3.1 The Reference Measure λ

In (1), λ can be interpreted as a "governer" upon the *principle* aggregation structure, whereas the "aggregation function" $\mathbb{m}_3$ tunes the *fine* aggregation details. For instance, if one chooses $\lambda = \lambda_L$ as the Lebesgue measure on $\mathscr{X} \subset \mathbb{R}$, then the integral in (1) turns out to be of Lebesgue-type and (with some rare exceptions) consequently of Riemann-type. In contrast, in the *discrete setup* where $\mathscr{X} := \mathscr{X}_\#$ has countably many elements and is equipped with the counting measure $\lambda := \lambda_\# := \sum_{z \in \mathscr{X}_\#} \delta_z$ (where δ_z is Dirac's one-point distribution $\delta_z[A] := \mathbf{1}_A(z)$, and thus $\lambda_\#[\{z\}] = 1$ for all $z \in \mathscr{X}_\#$) then (1) simplifies to

$$
\begin{aligned}
&0 \leqslant D^{c}_{\phi,M_1,M_2,\mathbb{M}_3,\lambda_\#}(P,Q)\\
&:= \overline{\sum}_{z\in\mathscr{X}} \Big[\phi\big(\tfrac{p(z)}{m_1(z)}\big) - \phi\big(\tfrac{q(z)}{m_2(z)}\big) - \phi'_{+,c}\big(\tfrac{q(z)}{m_2(z)}\big) \cdot \big(\tfrac{p(z)}{m_1(z)} - \tfrac{q(z)}{m_2(z)}\big)\Big] \cdot \mathbb{m}_3(z)\,, \qquad (2)
\end{aligned}
$$

which we interpret as $\sum_{z\in\mathscr{X}} \overline{\psi_{\phi,c}}\big(\tfrac{p(z)}{m_1(z)}, \tfrac{q(z)}{m_2(z)}\big) \cdot \mathbb{m}_3(z)$ with the same conventions and limits as in the paragraph right after (1); if $\mathscr{X}_\# = \{z_0\}$ for arbitrary $z_0 \in \widetilde{X}$, we obtain the corresponding one-point divergence over any space $\widetilde{X}$.

3.2 The Divergence Generator ϕ

We continue with the inspection of interesting special cases of
$\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$ by dealing with the first component. For this, let $\Phi_{C_1}(]a,b[)$ be the class of all functions $\phi \in \Phi(]a,b[)$ which are also continuously differentiable on $E =]a,b[$. For divergence generator $\phi \in \Phi_{C_1}(]a,b[)$, the formula (1) becomes (recall that we suppress the obsolete c and subderivative index $+$)

$$
\begin{aligned}
&0 \leqslant D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)\\
&:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m_1(x)}\big) - \phi\big(\tfrac{q(x)}{m_2(x)}\big) - \phi'\big(\tfrac{q(x)}{m_2(x)}\big) \cdot \big(\tfrac{p(x)}{m_1(x)} - \tfrac{q(x)}{m_2(x)}\big)\Big] \cdot \mathbb{m}_3(x)\,\mathrm{d}\lambda(x)\,, \qquad (3)
\end{aligned}
$$

whereas (2) turns into

$$
\begin{aligned}
&0 \leqslant D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda_{\#}}(P,Q) \\
&:= \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m_1(x)}\big) - \phi\big(\tfrac{q(x)}{m_2(x)}\big) - \phi'\big(\tfrac{q(x)}{m_2(x)}\big)\cdot\big(\tfrac{p(x)}{m_1(x)} - \tfrac{q(x)}{m_2(x)}\big)\Big]\cdot \mathrm{m}_3(x).
\end{aligned}
$$

Formally, by defining the integral functional $g_{\phi,\mathbb{M}_3,\lambda}(\xi) := \int_{\mathscr{X}} \phi(\xi(x)) \cdot \mathrm{m}_3(x)\, \mathrm{d}\lambda(x)$ and plugging in e.g. $g_{\phi,\mathbb{M}_3,\lambda}\big(\frac{P}{M_1}\big) = \int_{\mathscr{X}} \phi\big(\frac{p(x)}{m_1(x)}\big) \cdot \mathrm{m}_3(x)\, \mathrm{d}\lambda(x)$, the divergence in (3) can be interpreted as

$$
\begin{aligned}
&0 \leqslant D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) \\
&= g_{\phi,\mathbb{M}_3,\lambda}\big(\tfrac{P}{M_1}\big) - g_{\phi,\mathbb{M}_3,\lambda}\big(\tfrac{Q}{M_2}\big) - g'_{\phi,\mathbb{M}_3,\lambda}\big(\tfrac{Q}{M_2}, \tfrac{P}{M_1} - \tfrac{Q}{M_2}\big)
\end{aligned}
\tag{4}
$$

where $g'_{\phi,\mathbb{M}_3,\lambda}\big(\eta, \cdot\big)$ denotes the corresponding directional derivate at $\eta = \frac{Q}{M_2}$. If one has a "nonnegativity-setup" (NN0) in the sense that for all $x \in \mathscr{X}$ there holds $\frac{p(x)}{m_1(x)} \geqslant 0$ and $\frac{q(x)}{m_2(x)} \geqslant 0$ (but not necessarily $p(x) \geqslant 0$, $q(x) \geqslant 0$, $m_1(x) \geqslant 0$, $m_2(x) \geqslant 0$) then one can take $a = 0$, $b = \infty$, i.e. $E =]0,\infty[$, and employ the strictly convex power functions

$$
\begin{aligned}
&\tilde{\phi}(t) := \tilde{\phi}_\alpha(t) := \tfrac{t^\alpha - 1}{\alpha(\alpha-1)} \;\in]-\infty,\infty[, \qquad t \in]0,\infty[,\ \alpha \in \mathbb{R}\backslash\{0,1\}\,, \\
&\phi(t) := \phi_\alpha(t) := \tilde{\phi}_\alpha(t) - \tilde{\phi}'_\alpha(1)\cdot(t-1) = \tfrac{t^\alpha - 1}{\alpha(\alpha-1)} - \tfrac{t-1}{\alpha-1} \;\in [0,\infty[, \quad t \in]0,\infty[, \\
&\hfill \alpha \in \mathbb{R}\backslash\{0,1\}\,,
\end{aligned}
\tag{5}
$$

which satisfy (with the notations introduced in the paragraph right after (1))

$$
\phi_\alpha(1) = 0, \quad \phi'_\alpha(t) = \tfrac{t^{\alpha-1}-1}{\alpha-1}, \quad \phi'_\alpha(1) = 0, \quad \phi''_\alpha(t) = t^{\alpha-2} > 0, \quad t \in]0,\infty[, \tag{6}
$$

$$
\begin{aligned}
&\phi_\alpha(0) := \lim_{t\downarrow 0}\phi_\alpha(t) = \tfrac{1}{\alpha}\cdot\mathbf{1}_{]0,1]\cup]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha), \\
&\hfill \phi_\alpha(\infty) := \lim_{t\uparrow\infty}\phi_\alpha(t) = \infty,
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
&\phi'_\alpha(0) := \lim_{t\downarrow 0}\phi'_\alpha(t) = \tfrac{1}{1-\alpha}\cdot\mathbf{1}_{]1,\infty[}(\alpha) - \infty\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha), \\
&\phi'_\alpha(\infty) \;:=\; \lim_{t\uparrow\infty}\phi'_\alpha(t) \;=\; \infty\cdot\mathbf{1}_{]1,\infty[}(\alpha) + \tfrac{1}{1-\alpha}\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha) \;=\; \lim_{t\uparrow\infty}\tfrac{\phi_\alpha(t)}{t},
\end{aligned}
\tag{8}
$$

$$
\psi_{\phi_\alpha}(s,t) = \tfrac{1}{\alpha\cdot(\alpha-1)}\cdot\Big[s^\alpha + (\alpha-1)\cdot t^\alpha - \alpha\cdot s\cdot t^{\alpha-1}\Big], \quad s,t \in]0,\infty[, \tag{9}
$$

$$
\psi_{\phi_\alpha}(0,t) = \tfrac{t^\alpha}{\alpha}\cdot\mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha), \quad t \in]0,\infty[, \tag{10}
$$

$$
\begin{aligned}
&\psi_{\phi_\alpha}(\infty,t) = \infty, \quad t \in]0,\infty[, \\
&\lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_\alpha}(s,1) = \tfrac{1}{1-\alpha}\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha) + \infty\cdot\mathbf{1}_{]1,\infty[}(\alpha), \\
&\psi_{\phi_\alpha}(s,0) = \tfrac{s^\alpha}{\alpha\cdot(\alpha-1)}\cdot\mathbf{1}_{]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha), \quad s \in]0,\infty[, \\
&\psi_{\phi_\alpha}(s,\infty) = \tfrac{s^\alpha}{\alpha\cdot(\alpha-1)}\cdot\mathbf{1}_{]-\infty,0[}(\alpha) + \infty\cdot\mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha), \quad s \in]0,\infty[,
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
&\psi_{\phi_\alpha}(0,0) := 0 \text{ (which is unequal to } \lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha < 0 \\
&\qquad\qquad \text{and which is unequal to } \lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha > 1), \\
&\psi_{\phi_\alpha}(\infty,\infty) := 0 \text{ (which is unequal to } \lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha \in \mathbb{R}\backslash\{0,1\} \\
&\qquad\qquad \text{and which is unequal to } \lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha \in]0,1[\cup]1,\infty[), \\
&\psi_{\phi_\alpha}(0,\infty) := \lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_\alpha}(s,t) = \infty
\end{aligned}
\tag{12}
$$

$$\begin{aligned} &\qquad\qquad\qquad \text{(which coincides with } \lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha\in\mathbb{R}\backslash\{0,1\}), \\ &\psi_{\phi_\alpha}(\infty,0) := \lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_\alpha}(s,t) = \infty \qquad (13) \\ &\qquad\qquad\qquad \text{(which coincides with } \lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha\in\mathbb{R}\backslash\{0,1\}). \end{aligned}$$

The perhaps most important special case is $\alpha = 2$, for which (5) turns into

$$\phi_2(t) := \tfrac{(t-1)^2}{2}, \quad t \in]0,\infty[= E, \tag{14}$$

having for $s, t \in]0,\infty[$ the properties (cf. (7)–(13))

$$\begin{aligned} &\phi_2(1) = 0, \quad \phi_2'(1) = 0, \quad \phi_2(0) = \tfrac{1}{2}, \quad \phi_2(\infty) = \infty, \quad \phi_2'(0) = -\tfrac{1}{2}, \\ &\phi_2'(\infty) = \infty = \lim_{t\uparrow\infty}\tfrac{\phi_2(t)}{t}, \psi_{\phi_2}(s,t) = \tfrac{(s-t)^2}{2}, \qquad (15) \\ &\psi_{\phi_2}(0,t) = \tfrac{t^2}{2}, \quad \psi_{\phi_2}(\infty,t) = \infty, \quad \lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_2}(s,1) = \infty, \\ &\psi_{\phi_2}(s,0) = \tfrac{s^2}{2}, \quad \psi_{\phi_2}(s,\infty) = \infty, \quad \psi_{\phi_2}(0,0) := 0, \qquad (16) \\ &\psi_{\phi_2}(\infty,\infty) := 0, \quad \psi_{\phi_2}(0,\infty) = \infty, \quad \psi_{\phi_2}(\infty,0) = \infty. \end{aligned}$$

Also notice that the divergence-generator ϕ_2 of (14) can be trivially extended to

$$\bar{\phi}_2(t) := \tfrac{(t-1)^2}{2}, \quad t \in]-\infty,\infty[= \bar{E}, \tag{17}$$

which is useful in a general setup (GS) where for all $x \in \mathscr{X}$ one has $\frac{p(x)}{m_1(x)} \in [-\infty,\infty]$ and $\frac{q(x)}{m_2(x)} \in [-\infty,\infty]$. Convex extensions to $]a,\infty[$ with $a \in]-\infty,0[$ can be easily done by the shift $\bar{\phi}_\alpha(t) := \phi_\alpha(t-a)$.

Further examples of everywhere strictly convex differentiable divergence generators $\phi \in \Phi_{C_1}(]a,b[)$ for the "nonnegativity-setup" (NN0) (i.e. $a = 0$, $b = \infty$, $E =]0,\infty[$) can be obtained by taking the α-limits

$$\begin{aligned} &\tilde{\phi}_1(t) := \lim_{\alpha\to 1}\phi_\alpha(t) = t\cdot\log t \ \in [-e^{-1},\infty[, \qquad t\in]0,\infty[, \\ &\phi_1(t) := \lim_{\alpha\to 1}\phi_\alpha(t) = \tilde{\phi}_1(t) - \tilde{\phi}_1'(1)\cdot(t-1) = t\cdot\log t + 1 - t \in [0,\infty[,\ t\in]0,\infty[, \quad (18) \\ &\tilde{\phi}_0(t) := \lim_{\alpha\to 0}\phi_\alpha(t) = -\log t \ \in]-\infty,\infty[, \qquad t\in]0,\infty[, \\ &\phi_0(t) := \lim_{\alpha\to 0}\phi_\alpha(t) = \tilde{\phi}_0(t) - \tilde{\phi}_0'(1)\cdot(t-1) = -\log t + t - 1 \in [0,\infty[,\ t\in]0,\infty[, \quad (19) \end{aligned}$$

which satisfy

$$\phi_1(1) = 0, \quad \phi_1'(t) = \log t, \quad \phi_1'(1) = 0, \quad \phi_1''(t) = t^{-1} > 0, \quad t \in]0, \infty[,$$

$$\phi_1(0) := \lim_{t \downarrow 0} \phi_1(t) = 1, \quad \phi_1(\infty) := \lim_{t \uparrow \infty} \phi_1(t) = \infty, \tag{20}$$

$$\phi_1'(0) := \lim_{t \downarrow 0} \phi_1'(t) = -\infty, \quad \phi_1'(\infty) := \lim_{t \uparrow \infty} \phi_1'(t) = +\infty = \lim_{t \uparrow \infty} \tfrac{\phi_1(t)}{t}, \tag{21}$$

$$\psi_{\phi_1}(s,t) = s \cdot \log\left(\tfrac{s}{t}\right) + t - s, \quad s, t \in]0, \infty[, \tag{22}$$

$$\psi_{\phi_1}(0,t) = t, \quad \psi_{\phi_1}(\infty, t) = \infty, \quad \lim_{s \to \infty} \tfrac{1}{s} \cdot \psi_{\phi_1}(s, 1) = \infty, \quad t \in]0, \infty[, \tag{23}$$

$$\psi_{\phi_1}(s, 0) = \infty, \quad \psi_{\phi_1}(s, \infty) = \infty, \quad s \in]0, \infty[, \tag{24}$$

$$\psi_{\phi_1}(0,0) := 0 \text{ (which coincides with } \lim_{t \to 0} \lim_{s \to 0} \psi_{\phi_1}(s,t)$$
$$\text{but which does not coincide with } \lim_{s \to 0} \lim_{t \to 0} \psi_{\phi_1}(s,t) = \infty),$$

$$\psi_{\phi_1}(\infty, \infty) := 0 \text{ (which does not coincide with}$$
$$\lim_{t \to \infty} \lim_{s \to \infty} \psi_{\phi_1}(s,t) = \lim_{s \to \infty} \lim_{t \to \infty} \psi_{\phi_1}(s,t) = \infty,$$

$$\psi_{\phi_1}(0, \infty) := \lim_{s \to 0} \lim_{t \to \infty} \psi_{\phi_1}(s,t) = \infty$$
$$\text{(which coincides with } \lim_{t \to \infty} \lim_{s \to 0} \psi_{\phi_1}(s,t)),$$

$$\psi_{\phi_1}(\infty, 0) := \lim_{s \to \infty} \lim_{t \to 0} \psi_{\phi_1}(s,t) = \infty$$
$$\text{(which coincides with } \lim_{t \to 0} \lim_{s \to \infty} \psi_{\phi_1}(s,t)),$$

as well as

$$\phi_0(1) = 0, \quad \phi_0'(t) = 1 - \tfrac{1}{t}, \quad \phi_0'(1) = 0, \quad \phi_0''(t) = t^{-2} > 0, \quad t \in]0, \infty[, \tag{25}$$

$$\phi_0(0) := \lim_{t \downarrow 0} \phi_0(t) = \infty, \quad \phi_0(\infty) := \lim_{t \uparrow \infty} \phi_0(t) = \infty, \tag{26}$$

$$\phi_0'(0) := \lim_{t \downarrow 0} \phi_0'(t) = -\infty, \quad \phi_0'(\infty) := \lim_{t \uparrow \infty} \phi_0'(t) = 1 = \lim_{t \uparrow \infty} \tfrac{\phi_0(t)}{t}, \tag{27}$$

$$\psi_{\phi_0}(s,t) = -\log\left(\tfrac{s}{t}\right) + \tfrac{s}{t} - 1, \quad s, t \in]0, \infty[, \tag{28}$$

$$\psi_{\phi_0}(0,t) = \infty, \quad \psi_{\phi_0}(\infty, t) = \infty, \quad \lim_{s \to \infty} \tfrac{1}{s} \cdot \psi_{\phi_0}(s, 1) = 1, \quad t \in]0, \infty[, \tag{29}$$

$$\psi_{\phi_0}(s, 0) = \infty, \quad \psi_{\phi_0}(s, \infty) = \infty, \quad s \in]0, \infty[, \tag{30}$$

$$\psi_{\phi_0}(0,0) := 0 \text{ (which does not coincide with}$$
$$\lim_{t \to 0} \lim_{s \to 0} \psi_{\phi_0}(s,t) = \lim_{s \to 0} \lim_{t \to 0} \psi_{\phi_0}(s,t) = \infty),$$

$$\psi_{\phi_0}(\infty, \infty) := 0 \text{ (which does not coincide with}$$
$$\lim_{t \to \infty} \lim_{s \to \infty} \psi_{\phi_0}(s,t) = \lim_{s \to \infty} \lim_{t \to \infty} \psi_{\phi_0}(s,t) = \infty),$$

$$\psi_{\phi_0}(0, \infty) := \lim_{s \to 0} \lim_{t \to \infty} \psi_{\phi_0}(s,t) = \infty$$
$$\text{(which coincides with } \lim_{t \to \infty} \lim_{s \to 0} \psi_{\phi_0}(s,t)),$$

$$\psi_{\phi_0}(\infty, 0) := \lim_{s \to \infty} \lim_{t \to 0} \psi_{\phi_0}(s,t) = \infty$$
$$\text{(which coincides with } \lim_{t \to 0} \lim_{s \to \infty} \psi_{\phi_0}(s,t)).$$

An important, but (in our context) technically delicate, convex divergence generator is $\phi_{TV}(t) := |t - 1|$ which is non-differentiable at $t = 1$; the latter is also the only point of strict convexity. Further properties are for arbitrarily fixed $s, t \in]0, \infty[$, $c \in [0, 1]$ (if not stated otherwise)

$$\phi_{TV}(1)=0,\quad \phi_{TV}(0)=1,\quad \phi_{TV}(\infty)=\infty, \tag{31}$$

$$\phi'_{TV,+,c}(t) = \mathbf{1}_{]1,\infty[}(t) + (2c-1)\cdot \mathbf{1}_{\{1\}}(t) - \mathbf{1}_{]0,1[}(t),$$

$$\phi'_{TV,+,1}(t) = \mathbf{1}_{[1,\infty[}(t) - \mathbf{1}_{]0,1[}(t),$$

$$\phi'_{TV,+,\frac{1}{2}}(t) = \mathbf{1}_{]1,\infty[}(t) - \mathbf{1}_{]0,1[}(t) = \mathrm{sgn}(t-1)\cdot \mathbf{1}_{]0,\infty[}(t),$$

$$\phi'_{TV,+,c}(1) = 2c-1, \qquad \phi'_{TV,+,1}(1) = 1, \quad \phi'_{TV,+,\frac{1}{2}}(1) = 0, \tag{32}$$

$$\phi'_{TV,+,c}(0) = \lim_{t\to 0}\phi'_{TV,+,c}(t) = -1, \quad \phi'_{TV,+,c}(\infty) = \lim_{t\to\infty}\phi'_{TV,+,c}(t) = 1,$$

$$\begin{aligned}\psi_{\phi_{TV},c}(s,t) = {} & \mathbf{1}_{]0,1[}(t)\cdot 2(s-1)\cdot \mathbf{1}_{]1,\infty[}(s) + \mathbf{1}_{]1,\infty[}(t)\cdot 2(1-s)\cdot \mathbf{1}_{]0,1]}(s) \\ & + \mathbf{1}_{\{1\}}(t)\cdot \Big[2(1-c)\cdot(s-1)\cdot \mathbf{1}_{]1,\infty[}(s) + 2c\cdot(1-s)\cdot \mathbf{1}_{]0,1]}(s)\Big],\end{aligned}$$

$$\psi_{\phi_{TV},\frac{1}{2}}(s,1) = |s-1|, \tag{33}$$

$$\psi_{\phi_{TV},c}(0,t) = \lim_{s\to 0}\psi_{\phi_{TV},c}(s,t) = 2\cdot \mathbf{1}_{]1,\infty[}(t) + 2c\cdot \mathbf{1}_{\{1\}}(t),$$

$$\psi_{\phi_{TV},c}(\infty,t) = \lim_{s\to\infty}\psi_{\phi_{TV},c}(s,t) = \infty\cdot \mathbf{1}_{]0,1[}(t) + \ \infty\cdot \mathbf{1}_{\{1\}}(t)\cdot \mathbf{1}_{[0,1[}(c),$$

$$\lim_{s\to\infty}\tfrac{1}{s}\cdot \psi_{\phi_{TV},c}(s,1) = 2(1-c), \tag{34}$$

$$\psi_{\phi_{TV},c}(s,0) = \lim_{t\to 0}\psi_{\phi_{TV},c}(s,t) = 2(s-1)\cdot \mathbf{1}_{]1,\infty[}(s),$$

$$\psi_{\phi_{TV},c}(s,\infty) = \lim_{t\to\infty}\psi_{\phi_{TV},c}(s,t) = 2(1-s)\cdot \mathbf{1}_{]0,1]}(s),$$

$$\begin{aligned}\psi_{\phi_{TV},c}(0,0) := 0 \text{ (which coincides with both } & \lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_{TV},c}(s,t) \\ & \text{and } \lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_{TV},c}(s,t)),\end{aligned}$$

$$\begin{aligned}\psi_{\phi_{TV},c}(\infty,\infty) := 0 \text{ (which coincides with both } & \lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_{TV},c}(s,t) \\ & \text{and } \lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_{TV},c}(s,t)),\end{aligned}$$

$$\begin{aligned}\psi_{\phi_{TV},c}(0,\infty) := {} & \lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_{TV},c}(s,t) = 2 \\ & \text{(which coincides with } \lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_{TV},c}(s,t)),\end{aligned}$$

$$\begin{aligned}\psi_{\phi_{TV},c}(\infty,0) := {} & \lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_{TV},c}(s,t) = \infty \\ & \text{(which coincides with } \lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_{TV},c}(s,t)).\end{aligned}$$

In particular, one sees from Assumption 2(a) that – in our context – ϕ_{TV} can only be potentially applied if $\frac{q(x)}{m_2(x)} = 1$ for λ-a.a. $x \in \mathscr{X}$ and from Assumption 2(c), (d) that we *generally* have to exclude $c = 1$ and $c = 0$ for $\phi'_{+,c}(\cdot)$ (i.e. we choose $c \in]0,1[$); as already mentioned above, under some non-obvious additional constraints on the functions P, Q it may be possible to drop the Assumptions 2(c), (d), see for instance Sect. 3.3.1.2 below.

Another interesting and technically delicate example is the divergence generator $\phi_{ie}(t) := t - 1 + \frac{(1-t)^3}{3}\cdot \mathbf{1}_{[0,1]}(t)$ which is convex, twice continuously differentiable, strictly convex at any point $t \in]0,1]$ and affine-linear on $[1,\infty[$. More detailed, one obtains for arbitrarily fixed $s, t \in]0,\infty[$ (if not stated otherwise):

$$
\begin{aligned}
&\phi_{ie}(1) = 0, \quad \phi_{ie}(0) = -\tfrac{2}{3}, \quad \phi_{ie}(\infty) = \infty, \\
&\phi_{ie}'(t) = 1 - (1-t)^2 \cdot \mathbf{1}_{]0,1[}(t), \\
&\phi_{ie}'(1) = 1, \quad \phi_{ie}'(0) = \lim_{t \to 0} \phi_{ie}'(t) = 0, \quad \phi_{ie}'(\infty) = \lim_{t \to \infty} \phi_{ie}'(t) = 1, \\
&\phi_{ie}''(t) = 2(1-t) \cdot \mathbf{1}_{]0,1[}(t), \quad \phi_{ie}''(1) = 0, \\
&\psi_{\phi_{ie}}(s,t) = \tfrac{(1-s)^3}{3} \cdot \mathbf{1}_{]0,1[}(s) + \ (1-t)^2 \cdot \Big[\tfrac{2}{3} \cdot (1-t) + (s-1) \Big] \cdot \mathbf{1}_{]0,1[}(t), \\
&\psi_{\phi_{ie}}(s,1) = \tfrac{(1-s)^3}{3} \cdot \mathbf{1}_{]0,1[}(s), \\
&\psi_{\phi_{ie}}(0,t) \ = \ \lim_{s \to 0} \psi_{\phi_{ie}}(s,t) \ = \ \tfrac{1}{3} \cdot \mathbf{1}_{[1,\infty[}(t) + \tfrac{1}{3} \ \cdot \ \Big[1 - (1-t)^2 \cdot (1-2t) \Big] \cdot \mathbf{1}_{]0,1[}(t), \\
&\psi_{\phi_{ie}}(\infty,t) = \lim_{s \to \infty} \psi_{\phi_{ie}}(s,t) = \infty \cdot \mathbf{1}_{]0,1[}(t), \\
&\lim_{s \to \infty} \tfrac{1}{s} \cdot \psi_{\phi_{ie}}(s,1) = 0, \\
&\psi_{\phi_{ie}}(s,0) = \lim_{t \to 0} \psi_{\phi_{ie}}(s,t) = \big(s - \tfrac{1}{3} \big) \cdot \mathbf{1}_{[1,\infty[}(s) + s^2 \cdot \big(1 - \tfrac{s}{3} \big) \cdot \mathbf{1}_{]0,1[}(s), \\
&\psi_{\phi_{ie}}(s,\infty) = \lim_{t \to \infty} \psi_{\phi_{ie}}(s,t) = \tfrac{(1-s)^3}{3} \cdot \mathbf{1}_{]0,1[}(s), \\
&\psi_{\phi_{ie}}(0,0) := 0 \text{ (which coincides with both } \lim_{t \to 0} \lim_{s \to 0} \psi_{\phi_{ie}}(s,t) \\
&\qquad\qquad\qquad \text{and } \lim_{s \to 0} \lim_{t \to 0} \psi_{\phi_{ie}}(s,t)), \\
&\psi_{\phi_{ie}}(\infty,\infty) := 0 \text{ (which coincides with both } \lim_{t \to \infty} \lim_{s \to \infty} \psi_{\phi_{ie}}(s,t) \\
&\qquad\qquad\qquad \text{and } \lim_{s \to \infty} \lim_{t \to \infty} \psi_{\phi_{ie}}(s,t)), \\
&\psi_{\phi_{ie}}(0,\infty) := \lim_{s \to 0} \lim_{t \to \infty} \psi_{\phi_{ie}}(s,t) = \tfrac{1}{3} \\
&\qquad\qquad\qquad \text{(which coincides with } \lim_{t \to \infty} \lim_{s \to 0} \psi_{\phi_{ie}}(s,t)), \\
&\psi_{\phi_{ie}}(\infty,0) := \lim_{s \to \infty} \lim_{t \to 0} \psi_{\phi_{ie}}(s,t) = \infty \\
&\qquad\qquad\qquad \text{(which coincides with } \lim_{t \to 0} \lim_{s \to \infty} \psi_{\phi_{ie}}(s,t)).
\end{aligned}
\tag{35}
$$

In particular, one sees from the Assumptions 2(a), (b) that – in our context – ϕ_{ie} can only be potentially applied in the following two disjoint situations:

(i) $\frac{q(x)}{m_2(x)} < 1$ for λ-a.a. $x \in \mathscr{X}$;
(ii) $\frac{q(x)}{m_2(x)} = 1$ and $\frac{p(x)}{m_1(x)} \leqslant 1$ for λ-a.a. $x \in \mathscr{X}$.

As already mentioned above, under some non-obvious additional constraints on the functions P, Q it may be possible to drop Assumption 2(b) and consequently (ii) can then be replaced by
$\widetilde{(ii)}$ $\frac{q(x)}{m_2(x)} = 1$ for λ-a.a. $x \in \mathscr{X}$;
see for instance Sect. 3.3.1.2 below.

3.3 *The Scaling and the Aggregation Functions m_1, m_2, $\mathbb{m}_3$*

In the above two Sects. 3.1 and 3.2, we have illuminated details of the choices of the first and the last component of $\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$. Let us now discuss the *principal* roles as well as examples of m_1, m_2, $\mathbb{m}_3$, which widen considerably the divergence-modeling flexibility and thus bring in a broad spectrum of goal-oriented situation-based applicability. To start with, recall that in accordance with (1), the aggregation function $\mathbb{m}_3$ tunes the fine aggregation details (whereas λ can be interpreted as a "governer" upon the basic/principle aggregation structure); furthermore, the function $m_1(\cdot)$ scales the function $p(\cdot)$ and $m_2(\cdot)$ the function $q(\cdot)$. From a modeling perspective, these two scaling functions can e.g. be "purely direct" in the sense that $m_1(x)$, $m_2(x)$ are chosen to directly reflect some dependence on the data-reflecting state $x \in \mathscr{X}$ (independent of the choice of P,Q), or "purely adaptive" in the sense that $m_1(x) = w_1(p(x), q(x))$, $m_2(x) = w_2(p(x), q(x))$ for some appropriate (measurable) "connector functions" w_1, w_2 on the product $\mathscr{R}(P) \times \mathscr{R}(Q)$ of the ranges of $\big\{p(x)\big\}_{x\in\mathscr{X}}$ and $\big\{q(x)\big\}_{x\in\mathscr{X}}$, or "hybrids" $m_1(x) = w_1(x, p(x), q(x))$ $m_2(x) = w_2(x, p(x), q(x))$. Also recall that in consistency with Assumption 2 we always assume $\mathbb{m}_3(x) = \mathbb{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)$ for some (measurable) function $\mathbb{w}_3 : \mathscr{X} \times [a, b] \times [a, b] \to [0, \infty]$. Whenever applicable and insightfulness-enhancing, we use the notation $D^c_{\phi,W_1,W_2,\mathbb{W}_3,\lambda}(P, Q)$ instead of $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q)$.

Let us start with the following important sub-setup:

3.3.1 $\mathbf{m_1(x) = m_2(x) := m(x)}$, $\mathbf{\mathbb{m}_3(x) = r(x) \cdot m(x) \in [0, \infty]}$ for Some (meas.) Function $\mathbf{r : \mathscr{X} \to \mathbb{R}}$ Satisfying $\mathbf{r(x) \in]-\infty, 0[\cup]0, \infty[}$ for $\boldsymbol{\lambda}$−a.a. $\mathbf{x \in \mathscr{X}}$

As an interpretation, here the scaling functions are strongly coupled with the aggregation function; in order to avoid "case-overlapping", we assume that the function $r(\cdot)$ does not (explicitly) depend on the functions $m(\cdot)$, $p(\cdot)$ and $q(\cdot)$ (i.e. it is not of the form $r(\cdot) = h(\cdot, m(\cdot), p(\cdot), q(\cdot))$). From (1) one can deduce

$$0 \leqslant D^c_{\phi,M,M,R\cdot M,\lambda}(P, Q)$$
$$:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m(x)}\big) - \phi\big(\tfrac{q(x)}{m(x)}\big) - \phi'_{+,c}\big(\tfrac{q(x)}{m(x)}\big) \cdot \big(\tfrac{p(x)}{m(x)} - \tfrac{q(x)}{m(x)}\big)\Big] \cdot m(x) \cdot r(x)\, \mathrm{d}\lambda(x)\,, \qquad (36)$$

which for the discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_\#, \lambda_\#)$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}_\#$) simplifies to

$$0 \leqslant D^c_{\phi,M,M,R\cdot M,\lambda_\#}(P, Q)$$
$$= \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m(x)}\big) - \phi\big(\tfrac{q(x)}{m(x)}\big) - \phi'_{+,c}\big(\tfrac{q(x)}{m(x)}\big) \cdot \big(\tfrac{p(x)}{m(x)} - \tfrac{q(x)}{m(x)}\big)\Big] \cdot m(x) \cdot r(x)\,. \qquad (37)$$

Remark 2 (a) If one has a "nonnegativity-setup" (NN1) in the sense that for λ-almost all $x \in \mathscr{X}$ there holds $\mathbb{m}(x) \geqslant 0$, $\mathbb{r}(x) \geqslant 0$, $\mathbb{p}(x) \geqslant 0$, $\mathbb{q}(x) \geqslant 0$, then (36) (and hence also (37)) can be interpreted as scaled Bregman divergence $B_\phi\big(\mathfrak{P}, \mathfrak{Q}\,|\,\mathfrak{M}\big)$ between the two nonnegative measures $\mathfrak{P}$, $\mathfrak{Q}$ (on $(\mathscr{X}, \mathscr{F})$) defined by $\mathfrak{P}[\bullet] := \mathfrak{P}^{\mathbb{R}\cdot\lambda}[\bullet] := \int_\bullet \mathbb{p}(x) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x)$ and $\mathfrak{Q}[\bullet] := \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\bullet] := \int_\bullet \mathbb{q}(x) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x)$, with scaling by the nonnegative measure $\mathfrak{M}[\bullet] := \mathfrak{M}^{\mathbb{R}\cdot\lambda}[\bullet] := \int_\bullet \mathbb{m}(x) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x)$.

(b) In a context of $\mathbb{r}(x) \equiv 1$ and "λ-probability-densities" $\vec{\mathbb{p}}$, $\vec{\mathbb{q}}$ on general state space $\mathscr{X}$, then $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}[\bullet] := \int_\bullet \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda(x)$ and $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}[\bullet] := \int_\bullet \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda(x)$ are probability measures (where $\mathbb{1}$ stands for the function with constant value 1). Accordingly, (36) (and hence also (37)) can be interpreted as scaled Bregman divergence $B_\phi\big(\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}, \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda} \,|\, \mathfrak{M}^{\mathbb{1}\cdot\lambda}\big)$ which has been first defined in Stummer [81], Stummer and Vajda [84], see also Kisslinger and Stummer [35–37] for the "purely adaptive" case $\mathbb{m}(x) = \mathbb{w}\big(\vec{\mathbb{p}}(x), \vec{\mathbb{q}}(x)\big)$ and indications on non-probability measures. For instance, if Y is a random variable taking values in the discrete space $\mathscr{X}_\#$, then (with a slight abuse of notation[8]) $\vec{\mathbb{q}}(x) = \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_\#}[Y = x]$ may be its probability mass function under a hypothetical/candidate law $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_\#}$, and $\vec{\mathbb{p}}(x) = \frac{1}{N} \cdot \#\{i \in \{1, \dots, N\} : Y_i = x\} =: \vec{\mathbb{p}}_N^{emp}(x)$ is the probability mass function of the corresponding data-derived "empirical distribution" $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_\#}[\bullet] := \vec{\mathfrak{P}}_N^{emp}[\bullet] := \frac{1}{N} \cdot \sum_{i=1}^N \delta_{Y_i}[\bullet]$ of an N-size independent and identically distributed (i.i.d.) sample $Y_1, \dots, Y_N$ of Y which is nothing but the probability distribution reflecting the underlying (normalized) histogram. Typically, for small respectively medium sample size N one gets $\vec{\mathbb{p}}_N^{emp}(x) = 0$ for some states $x \in \mathscr{X}$ which are feasible but "not yet" observed; amongst other things, this explains why density-zeros play an important role especially in statistics and information theory. This concludes the current Remark 2. □

In the following, we illuminate two important special cases of the scaling (and aggregation-part) function $m(\cdot)$, namely $\mathbb{m}(x) := 1$ and $m(x) := q(x)$:

3.3.1.1 $\mathbb{m}_1(\mathbf{x}) = \mathbb{m}_2(\mathbf{x}) := \mathbf{1}$, $\mathbb{m}_3(\mathbf{x}) = \mathbb{r}(\mathbf{x})$ for Some (Measurable) Function $\mathbb{r} : \mathscr{X} \to [\mathbf{0}, \infty]$ Satisfying $\mathbb{r}(\mathbf{x}) \in]\mathbf{0}, \infty[$ for $\boldsymbol{\lambda}$-a.a. $\mathbf{x} \in \mathscr{X}$

Accordingly, (36) turns into

$$\begin{aligned} 0 &\leqslant D^c_{\phi,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(P, Q) \\ &:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(p(x)\big) - \phi\big(q(x)\big) - \phi'_{+,c}\big(q(x)\big) \cdot \big(p(x) - q(x)\big)\Big] \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x)\,, \end{aligned} \tag{38}$$

which for the discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_\#, \lambda_\#)$ becomes[9]

$$\begin{aligned} 0 &\leqslant D^c_{\phi,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda_\#}(P, Q) \\ &:= \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(p(x)\big) - \phi\big(q(x)\big) - \phi'_{+,c}\big(q(x)\big) \cdot \big(p(x) - q(x)\big)\Big] \cdot \mathbb{r}(x) \end{aligned} \tag{39}$$

[8] Respectively working with canonical space representation and $Y := id$.

[9] As a side remark, let us mention here that in the special case of continuously differentiable *strictly log-convex* divergence generator ϕ, one can construct divergences which are tighter than (38) respectively (39), see Stummer and Kißlinger [82]; in a finite discrete space and for differentiable *exponentially concave* divergence generator ϕ, a similar tightening (called L-divergence) can be found in Pal and Wong [66, 67].

Notice that for $\mathbbm{r}(x) \equiv 1$, the divergences (38) and (39) are "consistent extensions" of the motivating pointwise dissimilarity $d_{\phi}^{(6)}(\cdot,\cdot)$ from Sect. 2. A special case of (38) is e.g. the rho-tau divergence (cf. Lemma 1 of Zhang and Naudts [95]).
Let us exemplarily illuminate the special case $\phi = \phi_\alpha$ together with $\mathbbm{p}(x) \geqslant 0$, $\mathbbm{q}(x) \geqslant 0$, for λ-almost all $x \in \mathscr{X}$ which by means of (9), (22), (28) turns (38) into the "explicit-boundary" version (of the corresponding "implicit-boundary-describing" $\overline{\int} \ldots$)[10]

$$
\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \overline{\int}_{\mathscr{X}} \tfrac{\mathbbm{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \big[\mathbbm{p}(x)^\alpha + (\alpha-1)\cdot\mathbbm{q}(x)^\alpha - \alpha\cdot\mathbbm{p}(x)\cdot\mathbbm{q}(x)^{\alpha-1}\big]\,\mathrm{d}\lambda(x) &(40)\\
&= \int_{\mathscr{X}} \tfrac{\mathbbm{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \big[\mathbbm{p}(x)^\alpha + (\alpha-1)\cdot\mathbbm{q}(x)^\alpha - \alpha\cdot\mathbbm{p}(x)\cdot\mathbbm{q}(x)^{\alpha-1}\big]\cdot \mathbf{1}_{]0,\infty[}\big(\mathbbm{p}(x)\cdot\mathbbm{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+ \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\big[\tfrac{\mathbbm{p}(x)^\alpha}{\alpha\cdot(\alpha-1)}\cdot\mathbf{1}_{]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha)\big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbbm{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+ \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\big[\tfrac{\mathbbm{q}(x)^\alpha}{\alpha}\cdot\mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha)\big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbbm{p}(x)\big)\,\mathrm{d}\lambda(x)\,,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } \alpha \in \mathbb{R}\backslash\{0,1\}, &(41)\\
&0 \leqslant D_{\phi_1,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\big[\mathbbm{p}(x)\cdot\log\big(\tfrac{\mathbbm{p}(x)}{\mathbbm{q}(x)}\big) + \mathbbm{q}(x) - \mathbbm{p}(x)\big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{p}(x)\cdot\mathbbm{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+ \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbbm{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+ \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\mathbbm{q}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbbm{p}(x)\big)\,\mathrm{d}\lambda(x) &(42)\\
&0 \leqslant D_{\phi_0,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\Big[-\log\big(\tfrac{\mathbbm{p}(x)}{\mathbbm{q}(x)}\big) + \tfrac{\mathbbm{p}(x)}{\mathbbm{q}(x)} - 1\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{p}(x)\cdot\mathbbm{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+ \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbbm{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+ \int_{\mathscr{X}} \mathbbm{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbbm{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbbm{p}(x)\big)\,\mathrm{d}\lambda(x)\,, &(43)
\end{aligned}
$$

where we have employed (10), (11) (23), (24), (29), (30); notice that $D_{\phi_1,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q})$ is a generalized version of the Kullback–Leibler information divergence (resp. of the relative entropy). According to the above calculations, one should exclude $\alpha \leqslant 0$ whenever $\mathbbm{p}(x) = 0$ for all x in some A with $\lambda[A] > 0$, respectively $\alpha \leqslant 1$ whenever $\mathbbm{q}(x) = 0$ for all x in some $\tilde{A}$ with $\lambda[\tilde{A}] > 0$ (a refined alternative for $\alpha = 1$ is given in Sect. 3.3.1.2 below). As far as splitting of the first integral e.g. in (42) resp. (43) is concerned, notice that the integral $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] := \int_{\mathscr{X}} \big[\mathbbm{q}(x) - \mathbbm{p}(x)\big]\cdot\mathbbm{r}(x)\,\mathrm{d}\lambda(x)$ resp. $\int_{\mathscr{X}} \Big[\frac{\mathbbm{p}(x)}{\mathbbm{q}(x)} - 1\Big]\cdot\mathbbm{r}(x)\,\mathrm{d}\lambda(x)$ may be finite even in cases where $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbbm{p}(x)\cdot\mathbbm{r}(x)\,\mathrm{d}\lambda(x) = \infty$ and $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbbm{q}(x)\cdot\mathbbm{r}(x)\,\mathrm{d}\lambda(x) = \infty$ (especially in case of unbounded data space (e.g. $\mathscr{X} = \mathbb{R}$) when an additive constant is involved and $\mathbbm{r}(\cdot)$ is bounded from above); furthermore, there are situations where $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] < \infty$ and thus $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] = 0$ but $\int_{\mathscr{X}} \Big[\frac{\mathbbm{p}(x)}{\mathbbm{q}(x)} - 1\Big]\cdot\mathbbm{r}(x)\,\mathrm{d}\lambda(x) = \infty$. For $\alpha = 2$, we obtain from (41) and (15) to (16)

[10] The first resp. second resp. third integral in (41) can be interpreted as divergence-contribution of the function-(support-)overlap resp. of one part of the function-nonoverlap (e.g. describing "extreme outliers") resp. of the other part of the function-nonoverlap (e.g. describing "extreme inliers").

$$0 \leqslant D_{\phi_2,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{2} \cdot \big[\mathbb{p}(x) - \mathbb{q}(x)\big]^2 \, \mathrm{d}\lambda(x) \, , \tag{44}$$

where we can exceptionally drop the non-negativity constraints $\mathbb{p}(x) \geqslant 0, \mathbb{q}(x) \geqslant 0$. As for interpretation, (44) is nothing but half of the $\mathbb{r}(\cdot)$-weighted squared $L^2(\lambda)$-distance between $\mathbb{p}(\cdot)$ and $\mathbb{q}(\cdot)$.

In the special sub-setup of $\mathbb{r}(x) \equiv 1$ and "λ-probability-densities" $\vec{\mathbb{p}}$ $\vec{\mathbb{q}}$ on data space $\mathscr{X}$ (cf. Remark 2(b)), we can deduce from (41)–(43) the divergences

$$D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{1}\cdot\mathbb{1},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) \tag{45}$$

which for the choice $\alpha > 0$ can be interpreted as "order$-\alpha$" density-power divergences DPD of Basu et al. [10] between the two corresponding probability measures $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}$ and $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}$; for their statistical applications see e.g. Basu et al. [12], Ghosh and Basu [30, 31] and the references therein, and for general $\alpha \in \mathbb{R}$ see e.g. Stummer and Vajda [84]. In particular, the case $\alpha = 1$ corresponding divergence in (45) is called "Kullback–Leibler information divergence" between $\vec{\mathbb{P}}$ and $\vec{\mathbb{Q}}$, and is also known under the name "relative entropy". For $\alpha = 2$, we derive $D_{\phi_2,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ from (44) with $\mathbb{r}(x) = 1$ which is nothing but half of the squared L^2-distance between the two "λ-probability-densities" $\vec{\mathbb{p}}$ and $\vec{\mathbb{q}}$.

For the special discrete setup $(\mathscr{X},\lambda) = (\mathscr{X}_\#,\lambda_\#)$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}_\#$), the divergences (41)–(44) simplify to

$$\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \Big[\big(\mathbb{p}(x)\big)^\alpha + (\alpha-1)\cdot\big(\mathbb{q}(x)\big)^\alpha - \alpha\cdot\mathbb{p}(x)\cdot\big(\mathbb{q}(x)\big)^{\alpha-1}\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \Big[\tfrac{\mathbb{p}(x)^\alpha}{\alpha\cdot(\alpha-1)} \cdot \mathbf{1}_{]1,\infty[}(\alpha) + \infty \cdot \mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha)\Big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \Big[\tfrac{\mathbb{q}(x)^\alpha}{\alpha} \cdot \mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha) + \infty \cdot \mathbf{1}_{]-\infty,0[}(\alpha)\Big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \, , \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } \alpha \in \mathbb{R}\backslash\{0,1\}, \quad (46) \\
&0 \leqslant D_{\phi_1,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{p}(x)\cdot\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \mathbb{q}(x) - \mathbb{p}(x)\Big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \infty \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \, , \\
&0 \leqslant D_{\phi_0,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \Big[-\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1\Big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \infty \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \infty \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \, , \\
&0 \leqslant D_{\phi_2,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda_\#}(\mathbb{P},\mathbb{Q}) = \textstyle\sum_{x\in\mathscr{X}} \tfrac{\mathbb{r}(x)}{2} \cdot \big[\mathbb{p}(x) - \mathbb{q}(x)\big]^2 .
\end{aligned}$$

Hence, as above, one should exclude $\alpha \leqslant 0$ whenever $\mathbb{p}(x) = 0$ for all x in some A with $\lambda[A] > 0$, respectively $\alpha \leqslant 1$ whenever $\mathbb{q}(x) = 0$ for all x in some $\tilde{A}$ with $\lambda[\tilde{A}] > 0$ (a refined alternative for $\alpha = 1$ is given in Sect. 3.3.1.2 below).

In particular, take the probability context of Remark 2(b), with discrete random variable Y, hypothetical probability mass function $\mathbb{q}(x) := \vec{\mathbb{q}}(x) = \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_{\#}}[Y = x]$, and data-derived probability mass function (relative frequency) $\mathbb{p}(x) := \vec{\mathbb{p}}_N^{emp}(x) = \frac{1}{N} \cdot \#\{i \in \{1, \ldots, N\} : Y_i = x\}$ with sample size N. For $\mathbb{r}(x) \equiv 1$, the corresponding sample-size-weighted divergences $2N \cdot D_{\phi_\alpha, \mathbb{1}, \mathbb{1}, \mathbb{1}, \lambda_{\#}}(\vec{\mathbb{P}}_N^{emp}, \vec{\mathbb{Q}})$ (for $\alpha \in \mathbb{R}$) can be used as goodness-of-fit test statistics; see e.g. Kisslinger and Stummer [37] for their limit behaviour as the sample size N tends to infinity.

3.3.1.2 $\mathbf{m_1(x) = m_2(x) := q(x)}$, $\mathbb{m}_3\mathbf{(x) = r(x) \cdot q(x) \in [0, \infty]}$ for Some (meas.) Function $\mathbf{r} : \mathscr{X} \to \mathbb{R}$ Satisfying $\mathbf{r(x)} \in]-\infty, \mathbf{0}[\cup]\mathbf{0}, \infty[$ for $\boldsymbol{\lambda}$–a.a. $\mathbf{x} \in \mathscr{X}$

In such a set-up, the divergence (36) becomes

$$
\begin{aligned}
0 &\leqslant D^c_{\phi, Q, Q, R\cdot Q, \lambda}(P, Q) \\
&= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{q(x)}\big) - \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\tfrac{p(x)}{q(x)} - 1\big)\Big] \cdot q(x) \cdot r(x)\, \mathrm{d}\lambda(x) \qquad (47)\\
&= \overline{\int}_{\mathscr{X}} \Big[q(x) \cdot \phi\big(\tfrac{p(x)}{q(x)}\big) - q(x) \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(p(x) - q(x)\big)\Big] \cdot r(x)\, \mathrm{d}\lambda(x)\,, \qquad (48)
\end{aligned}
$$

where in accordance with the descriptions right after (1) we require that $\phi :]a, b[\to \mathbb{R}$ is convex and strictly convex at $1 \in]a, b[$ and incorporate the zeros of $p(\cdot), q(\cdot), r(\cdot)$ by the appropriate limits and conventions. In the following, we demonstrate this in a non-negativity set-up where for λ-almost all $x \in \mathscr{X}$ one has $\mathbb{r}(x) \in]0, \infty[$ as well as $\mathbb{p}(x) \in [0, \infty[$, $\mathbb{q}(x) \in [0, \infty[$, and hence $E =]a, b[=]0, \infty[$. In order to achieve a reflexivity result in the spirit of Theorem 4, we have to check for – respectively analogously adapt most of – the points in Assumption 2: to begin with, the weight $w(x, s, t)$ evaluated at $s := \mathbb{p}(x), t := \mathbb{q}(x)$ has to be substituted/replaced by $\widetilde{w}(x, \widetilde{t}) := \mathbb{r}(x) \cdot \widetilde{t}$ evaluated at $\widetilde{t} = \mathbb{q}(x)$, and the dissimilarity $\psi_{\phi,c}(s, t)$ has to be substituted/replaced by $\widetilde{\widetilde{\psi}}_{\phi,c}(\widetilde{s}, \widetilde{t}) := \psi_{\phi,c}\big(\frac{\widetilde{s}}{\widetilde{t}}, 1\big)$ with the plug-in $\widetilde{s} = \mathbb{p}(x)$. Putting things together, instead of the integrand-generating term $w(x, s, t) \cdot \psi_{\phi,c}(s, t)$ we have to inspect the boundary behaviour of $\widetilde{w}(x, \widetilde{t}) \cdot \widetilde{\widetilde{\psi}}_{\phi,c}(\widetilde{s}, \widetilde{t})$ being explicitly given (with a slight abuse of notation) by the function $\widetilde{\psi}_{\phi,c} :]0, \infty[^3 \to [0, \infty[$ in

$$
\begin{aligned}
&\widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big) := r \cdot \widetilde{t} \cdot \psi_{\phi,c}\big(\tfrac{\widetilde{s}}{\widetilde{t}}, 1\big) = r \cdot \widetilde{t} \cdot \big[\phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big) - \phi(1) - \phi'_{+,c}(1) \cdot \big(\tfrac{\widetilde{s}}{\widetilde{t}} - 1\big)\big] \\
&= r \cdot \widetilde{t} \cdot \big[\phi\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r}\big) - \phi(1) - \phi'_{+,c}(1) \cdot \big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r} - 1\big)\big] = r \cdot \widetilde{t} \cdot \psi_{\phi,c}\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r}, 1\big)\,. \qquad (49)
\end{aligned}
$$

Since the general right-hand-derivative concerning assumption $t \in \mathscr{R}\big(\frac{Q}{M_2}\big)$ has $\frac{\widetilde{s}}{\widetilde{t}} = 1$ as its analogue, we require that the convex function $\phi :]0, \infty[\to]-\infty, \infty[$ is strictly convex (only) at 1 in conformity with Assumption 2(a) (which is also employed in

Assumption 3); for the sake of brevity we use the short-hand notation $2(a)$ etc. in the following discussion. We shall not need 2(b) to 2(d) in the prevailing context, so that the above-mentioned generator $\phi_{TV}(t) := |t-1|$ is allowed for achieving reflexivity (for reasons which will become clear in the proof of Theorem 5 in the appendix). The analogue of 2(e) is $\mathbb{r}(x) \cdot \widetilde{t} < \infty$ which is always (almost surely) automatically satisfied (a.a.sat.), whereas 2(f) converts to "$\mathbb{r}(x) \cdot \widetilde{t} > 0$ for all $\widetilde{s} \neq \widetilde{t}$" which is also a.a.sat. except for the case $\widetilde{t} = 0$ which will be below incorporated in combination with $\psi_{\phi,c}$-multiplication (cf. (50)). For the derivation of the analogue of 2(k) we observe that for fixed $r > 0, \widetilde{s} > 0$ the function $\widetilde{t} \rightarrow \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big)$ is (the r-fold of) the perspective function (at $\widetilde{s}$) of the convex function $\psi_{\phi,c}\big(\cdot, 1\big)$ and thus convex with existing limit

$$\begin{aligned} &\ell i_1 := r \cdot 0 \cdot \psi_{\phi,c}\big(\tfrac{\widetilde{s}}{0}, 1\big) := \lim_{t \rightarrow 0} \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big) = \\ &= -r \cdot \widetilde{s} \cdot \phi'_{+,c}(1) + r \cdot \widetilde{s} \cdot \lim_{\widetilde{t} \rightarrow 0} \big[\tfrac{\widetilde{t}}{\widetilde{s}} \cdot \phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big)\big] = r \cdot \widetilde{s} \cdot (\phi^*(0) - \phi'_{+,c}(1)) \geqslant 0\,, \quad (50) \end{aligned}$$

where $\phi^*(0) := \lim_{u \rightarrow 0} u \cdot \phi\big(\frac{1}{u}\big) = \lim_{v \rightarrow \infty} \frac{\phi(v)}{v}$ exists but may be infinite (recall that $\phi'_{+,c}(1)$ is finite). Notice that in contrast to 2(k) we need not assume $\ell i_1 > 0$ (and thus do not exclude ϕ_{TV}). To convert 2(i), we employ the fact that for fixed $r > 0, \widetilde{t} > 0$ the function $\widetilde{s} \rightarrow \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big)$ is convex with existing limit

$$r \cdot \widetilde{t} \cdot \psi_{\phi,c}\Big(\frac{0}{\widetilde{t}}, 1\Big) := \lim_{s \rightarrow 0} \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big) = r \cdot \widetilde{t} \cdot (\phi(0) + \phi'_{+,c}(1) - \phi(1)) > 0\,,$$

where $\phi(0) := \lim_{u \rightarrow 0} \phi(u)$ exists but may be infinite. To achieve the analogue of 2(g), let us first remark that for fixed $r > 0$ the function $(\widetilde{s}, \widetilde{t}) \rightarrow \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big)$ may not be continuous at $(\widetilde{s}, \widetilde{t}) = (0, 0)$, but due to the very nature of a divergence we make the 2(g)-conform convention of setting

$$r \cdot 0 \cdot \psi_{\phi,c}\big(\tfrac{0}{0}, 1\big) := \widetilde{\psi}_{\phi,c}\big(r, 0, 0\big) := 0$$

(notice that e.g. the power function ϕ_{-1} of (5) with index $\alpha = -1$ obeys $\lim_{\widetilde{t} \rightarrow 0} \widetilde{\psi}_{\phi_{-1}}\big(r, \widetilde{t}, \widetilde{t}\big) = 0 \neq \frac{r}{2} = \lim_{\widetilde{t} \rightarrow 0} \widetilde{\psi}_{\phi_{-1}}\big(r, \widetilde{t}^2, \widetilde{t}\big)$). The analogues of the remaining Assumptions 2(h),(j),(ℓ),(m),(n) are (almost surely) obsolete because of our basic (almost surely) finiteness requirements. Summing up, with the above-mentioned limits and conventions we write (47) explicitly as

$$
\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\big[\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi^{*}(0) - \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi^{*}(0) - \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\,. \qquad (51)
\end{aligned}
$$

In case of $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] := \int_{\mathscr{X}} \mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, the divergence (51) becomes

$$
\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi^{*}(0) - \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0) + \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x) - \phi(1)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\,\mathrm{d}\lambda(x)\,.
\end{aligned}
\qquad (52)
$$

Moreover, in case of $\phi\big(1\big)=0$ and $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] = \int_{\mathscr{X}} \big(\mathbb{p}(x)-\mathbb{q}(x)\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) \in]-\infty,\infty[$ (but not necessarily $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{p}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$), the divergence (51) turns into

$$
\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) + \phi(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&-\phi'_{+,c}\big(1\big)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\,. \qquad (53)
\end{aligned}
$$

Let us remark that (53) can be interpreted as ϕ-divergence $D^{c}_{\phi}\big(\mu,\nu\big)$ between the two nonnegative measures μ,ν (on $(\mathscr{X},\mathscr{F})$) (cf. Stummer and Vajda [83]), where $\mu[\bullet] := \mathfrak{P}^{\mathbb{R}\cdot\lambda}[\bullet]$ and $\nu[\bullet] := \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\bullet]$. In the following, we briefly discuss two important sub-cases. First, in the "λ-probability-densities" context of Remark 2(b) one has for general $\mathscr{X}$ the manifestation $\mathbb{p}(x) := \overrightarrow{\mathbb{p}}(x) \geqslant 0$, $\mathbb{q}(x) := \overrightarrow{\mathbb{q}}(x) \geqslant 0$, and under the constraint $\phi(1)=0$ the corresponding divergence $D^{c}_{\phi,\overrightarrow{\mathbb{Q}},\overrightarrow{\mathbb{Q}},\mathbb{R}\cdot\overrightarrow{\mathbb{Q}},\lambda}(\overrightarrow{\mathbb{P}},\overrightarrow{\mathbb{Q}})$ turns out to be the ($\mathbb{r}$-)"local ϕ-divergence of Avlogiaris et al. [6, 7]; in case of $\mathbb{r}(x)\equiv 1$ this reduces – due to the fact $\int_{\mathscr{X}} \big(\overrightarrow{\mathbb{p}}(x) - \overrightarrow{\mathbb{q}}(x)\big)\,\mathrm{d}\lambda(x) = 0$ – to the classical Csiszar-Ali-Silvey ϕ-divergence CASD ([4, 27], see also e.g. Liese and Vajda [41], Vajda [89])

$$
\begin{aligned}
0 \leqslant D^{c}_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) &= \textstyle\int_{\mathscr{X}} \vec{\mathbb{q}}(x) \cdot \phi\big(\frac{\vec{\mathbb{p}}(x)}{\vec{\mathbb{q}}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}(x) \cdot \vec{\mathbb{q}}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\phi^{*}(0) \cdot \textstyle\int_{\mathscr{X}} \vec{\mathbb{p}}(x) \cdot \mathbf{1}_{\{0\}}\big(\vec{\mathbb{q}}(x)\big)\,\mathrm{d}\lambda(x) + \phi(0) \cdot \int_{\mathscr{X}} \vec{\mathbb{q}}(x) \cdot \mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}(x)\big)\,\mathrm{d}\lambda(x) \\
&-\phi'_{+,c}\big(1\big) \cdot \textstyle\int_{\mathscr{X}} \big(\vec{\mathbb{p}}(x) - \vec{\mathbb{q}}(x)\big)\,\mathrm{d}\lambda(x) \\
&= \textstyle\int_{\mathscr{X}} \vec{\mathbb{q}}(x) \cdot \phi\big(\frac{\vec{\mathbb{p}}(x)}{\vec{\mathbb{q}}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}(x) \cdot \vec{\mathbb{q}}(x)\big)\,\mathrm{d}\lambda(x) \\
&\quad +\phi^{*}(0) \cdot \mathfrak{P}^{\mathbb{1}\cdot\lambda}[\vec{\mathbb{q}}(x) = 0] + \phi(0) \cdot \mathfrak{Q}^{\mathbb{1}\cdot\lambda}[\vec{\mathbb{p}}(x) = 0]\,;
\end{aligned}
\tag{54}
$$

if $\phi(1) \neq 0$ then one has to additionally subtract $\phi(1)$ (cf. the corresponding special case of (52)). In particular, for the special sub-setup where for λ-almost all $x \in \mathscr{X}$ there holds $\mathbb{p}(x) := \vec{\mathbb{p}}(x) > 0$, $\mathbb{q}(x) := \vec{\mathbb{q}}(x) > 0$, $\mathbb{r}(x) \equiv 1$, $\phi(1) = 0$, one ends up with the reduced Csiszar-Ali-Silvey divergence

$$
0 \leqslant D^{c}_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) = \textstyle\int_{\mathscr{X}} \vec{\mathbb{q}}(x) \cdot \phi\big(\frac{\vec{\mathbb{p}}(x)}{\vec{\mathbb{q}}(x)}\big)\,\mathrm{d}\lambda(x)
$$

which can be interpreted as a "consistent extension" of the motivating pointwise dissimilarity $d_{\phi}^{(7)}(\cdot,\cdot)$ from the introductory Sect. 2; notice the fundamental structural difference to the divergence (38) which reflects $d_{\phi}^{(6)}(\cdot,\cdot)$. For comprehensive treatments of statistical applications of CASD, the reader is referred to Liese and Vajda [41], Read and Cressie [72], Vajda [89], Pardo [68], Liese and Miescke [40], Basu et al. [13].

Returning to the general divergence setup (51), we derive the reflexivity result (to be proved in the appendix):

Theorem 5 *Let* $c \in [0,1]$, $\mathbb{r}(x) \in]0,\infty[$ *for* λ*-a.a.* $x \in \mathscr{X}$, $\mathscr{R}\big(\frac{\mathbb{P}}{\mathbb{Q}}\big) \cup \{1\} \subset [a,b]$, *and* $\phi \in \Phi(]a,b[)$ *be strictly convex at* $t = 1$. *Moreover, suppose that*

$$
\textstyle\int_{\mathscr{X}} \big(\mathbb{p}(x) - \mathbb{q}(x)\big) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \;=\; 0 \tag{55}
$$

(but not necessarily $\int_{\mathscr{X}} \mathbb{p}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, $\int_{\mathscr{X}} \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$*). Then:*

(1) $D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) \geqslant 0$. *Depending on the concrete situation,* $D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})$ *may take infinite value.*

$$
\text{(2)}\;\; D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = 0 \;\; \textit{if and only if} \;\; \mathbb{p}(x) = \mathbb{q}(x)\, \textit{for}\, \lambda\textit{-a.a.}\; x \in \mathscr{X}. \tag{56}
$$

Remark 3 (a) In the context of non-negative measures, the special case $c = 1$ – together with $\int_{\mathscr{X}} \mathbb{p}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, $\int_{\mathscr{X}} \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$ – of Theorem 5 was first achieved by Stummer and Vajda [83].
(b) Assumption (55) is always automatically satisfied if one has coincidence of finite total masses in the sense of $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{p}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) = \int_{\mathscr{X}} \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) = \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] < \infty$. For $\mathbb{r}(x) \equiv 1$ this is always satisfied for λ-probability densities $\mathbb{p}(x) := \vec{\mathbb{p}}(x)$, $\mathbb{q}(x) := \vec{\mathbb{q}}(x)$, since $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}[\mathscr{X}] = \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}[\mathscr{X}] = 1$.

(c) Notice that in contrast to Theorem 4, the generator-concerning Assumptions 2(b)–(d) are replaced by the "model-concerning" constraint (55). This opens the gate for the use of the generators ϕ_{ie} and ϕ_{TV} for cases where (55) is satisfied. For the latter, we obtain with $c = \frac{1}{2}$ explicitly from (49) and (33)

$$\widetilde{\psi}_{\phi_{TV},\frac{1}{2}}\big(r,\widetilde{s},\widetilde{t}\big) := r\cdot\widetilde{t}\cdot\psi_{\phi_{TV},\frac{1}{2}}\big(\frac{\widetilde{s}}{\widetilde{t}},1\big) = r\cdot\widetilde{t}\cdot\big|\frac{\widetilde{s}}{\widetilde{t}}-1\big| = r\cdot\big|\widetilde{s}-\widetilde{t}\big|,$$

and hence from (51) together with $\phi_{TV}(1)=0$, $\phi_{TV}(0)=1$ (cf. (31)), $\phi'_{TV,+,\frac{1}{2}}(1) = 0$ (cf. (32)), $\phi^{*}_{TV}(0) = \lim_{s\to\infty}\frac{1}{s}\cdot\psi_{\phi_{TV},\frac{1}{2}}(s,1) = 1$ (cf. (34)) we get

$$0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \mathbb{r}(x)\cdot\big|\mathbb{p}(x)-\mathbb{q}(x)\big|\,\mathrm{d}\lambda(x) \tag{57}$$

which is nothing but the (possibly infinite) $\mathbb{r}(\cdot)$-weighted L_1-distance between the functions $x\to\mathbb{p}(x)$ and $x\to\mathbb{q}(x)$.

(d) In the light of (52), Theorem 4 (adapted to the current context) and Theorem 5, let us indicate that if one wants to use $\Xi := \int_{\mathscr{X}} \mathbb{q}(x)\cdot\phi\big(\frac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)$ (with appropriate zero-conventions) as a divergence, then one should either employ generators ϕ satisfying $\phi(1)=\phi'_{+,c}(1)=0$, or employ models fulfilling the assumption (56) together with generators ϕ satisfying $\phi(1)=0$. On the other hand, if this integral Ξ appears in your application context "naturally", then one should be aware that Ξ may become negative depending on the involved set-up; for a counter-example, see Stummer and Vajda [83]. This concludes Remark 3.

As an important example, we illuminate the special case $\phi=\phi_\alpha$ with $\alpha\in\mathbb{R}\backslash\{0,1\}$ (cf. (5)) under the constraint $(\mathfrak{P}^{\mathbb{R}\cdot\lambda}-\mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] = \int_{\mathscr{X}}\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)$ $\in]-\infty,\infty[$. Accordingly, the "implicit-boundary-describing" divergence (48) resp. the corresponding "explicit-boundary" version (53) turn into the generalized power divergences of order α (cf. Stummer and Vajda [83] for $\mathbb{r}(x)\equiv 1$)[11]

$$\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \overline{\int}_{\mathscr{X}} \tfrac{1}{\alpha\cdot(\alpha-1)}\cdot\Big[\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)^{\alpha}-\alpha\cdot\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}+\alpha-1\Big]\cdot\mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) \qquad (58)\\
&= \tfrac{1}{\alpha\cdot(\alpha-1)}\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\Big[\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)^{\alpha}-\alpha\cdot\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}+\alpha-1\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\alpha}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)+\phi_{\alpha}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&= \tfrac{1}{\alpha\cdot(\alpha-1)}\int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{p}(x)^{\alpha}\cdot\mathbb{q}(x)^{1-\alpha}-\mathbb{q}(x)\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\tfrac{1}{1-\alpha}\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot(\mathbb{p}(x)-\mathbb{q}(x))\,\mathrm{d}\lambda(x)+\infty\cdot\mathbf{1}_{]1,\infty[}(\alpha)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big(\tfrac{1}{\alpha\cdot(1-\alpha)}\cdot\mathbf{1}_{]0,1]\cup]1,\infty[}(\alpha)+\infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha)\big)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x),
\end{aligned}$$

[11] This can be interpreted analogously as in footnote 10.

where we have employed (8) and (7); especially, one gets for $\alpha = 2$

$$\begin{aligned}
&0 \leqslant D_{\phi_2,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \overline{\int}_{\mathscr{X}} \tfrac{1}{2} \cdot \tfrac{(\mathbb{p}(x)-\mathbb{q}(x))^2}{\mathbb{q}(x)} \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= \tfrac{1}{2} \int_{\mathscr{X}} \mathbb{r}(x) \cdot \tfrac{(\mathbb{p}(x)-\mathbb{q}(x))^2}{\mathbb{q}(x)} \cdot \mathbf{1}_{[0,\infty[}(\mathbb{p}(x)) \cdot \mathbf{1}_{]0,\infty[}(\mathbb{q}(x))\,\mathrm{d}\lambda(x) \\
&+\infty \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)
\end{aligned}$$

which is called Pearsons's chisquare divergence. Under the same constraint $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] \in]-\infty,\infty[$, the case $\alpha = 1$ leads by (18)–(22) to the generalized Kullback–Leibler divergence (generalized relative entropy)

$$\begin{aligned}
&0 \leqslant D_{\phi_1,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \overline{\int}_{\mathscr{X}} \Big[\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} \cdot \log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + 1 - \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} \Big] \cdot \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x) \cdot (\mathbb{q}(x) - \mathbb{p}(x))\,\mathrm{d}\lambda(x) + \infty \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)
\end{aligned}$$

(which equals (42)), and for $\alpha = 0$ one gets from (19), (25)–(27) the generalized reverse Kullback–Leibler divergence (generalized reverse relative entropy)

$$\begin{aligned}
&0 \leqslant D_{\phi_0,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \overline{\int}_{\mathscr{X}} \big[-\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1 \big] \cdot \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \log\big(\tfrac{\mathbb{q}(x)}{\mathbb{p}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x) \cdot (\mathbb{p}(x) - \mathbb{q}(x))\,\mathrm{d}\lambda(x) + \infty \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x).
\end{aligned}$$

Notice that instead of the limit in (50) one could also use the convention $r \cdot 0 \cdot \psi_\phi\big(\frac{s}{0}, 1\big) := \widetilde{\psi}_\phi\big(r, s, 0\big) := 0$; in the context of λ-probability densities, one then ends up with divergence by Rüschendorf [75].

For the discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_\#, \lambda_\#)$, the divergence in (51) simplifies to

$$\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda_\#}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x) \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x) \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) - \mathbb{q}(x)\big)\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big) \\
&+ \big[\phi^*(0) - \phi'_{+,c}(1)\big] \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)
\end{aligned} \tag{59}$$

which in case of $\phi(1) = \phi'_{+,c}(1) = 0$ – respectively $\phi(1) = 0$ and (55) – turns into

$$\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda_\#}(\mathbb{P},\mathbb{Q}) = \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big) \\
&+ \phi^*(0) \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) + \phi(0) \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big).
\end{aligned} \tag{60}$$

3.3.1.3 $\mathbf{m_1(x) = m_2(x) := w(p(x), q(x))}$, $\mathbb{m}_\mathbf{3}\mathbf{(x) = r(x) \cdot w(p(x), q(x)) \in [0, \infty[}$ for Some (Measurable) Functions $\mathbf{w} : \mathscr{R}(\mathbf{P}) \times \mathscr{R}(\mathbf{Q}) \rightarrow \mathbb{R}$ and $\mathbf{r} : \mathscr{X} \rightarrow \mathbb{R}$

Such a choice extends the context of the previous Sect. 3.3.1.2 where the "connector function" w took the simple form $w(u, v) = v$, as well as the setup of Sect. 3.3.1.1 dealing with constant $w(u, v) \equiv 1$. This introduces a wide flexibility with divergences of the form

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,W(P,Q),W(P,Q),R\cdot W(P,Q),\lambda}(P, Q) \\
&:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{w(p(x),q(x))}\big) - \phi\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \\
&-\phi'_{+,c}\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \cdot \big(\tfrac{p(x)}{w(p(x),q(x))} - \tfrac{q(x)}{w(p(x),q(x))}\big)\Big] \cdot w(p(x), q(x)) \cdot r(x) \,\mathrm{d}\lambda(x), \quad (61)
\end{aligned}$$

which for the discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_{\#}, \lambda_{\#})$ (recall $\lambda_{\#}[\{x\}] = 1$ for all $x \in \mathscr{X}_{\#}$) simplifies to

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,W(P,Q),W(P,Q),R\cdot W(P,Q),\lambda_{\#}}(P, Q) = \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{w(p(x),q(x))}\big) - \phi\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \\
&-\phi'_{+,c}\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \cdot \big(\tfrac{p(x)}{w(p(x),q(x))} - \tfrac{q(x)}{w(p(x),q(x))}\big)\Big] \cdot w(p(x), q(x)) \cdot r(x) \,. \qquad (62)
\end{aligned}$$

A detailed discussion of this wide class of divergences (61),(62) is beyond the scope of this paper. For the λ-probability density context (and an indication for more general functions), see the comprehensive paper of Kisslinger and Stummer [37] and the references therein. Finally, by appropriate choices of $w(\cdot, \cdot)$ we can even derive divergences of the form (60) but with non-convex non-concave ϕ: see e.g. the "perturbed" power divergences of Roensch and Stummer [74].

3.3.2 Global Scaling and Aggregation, and Other Paradigms

Our universal framework also contains, as special cases, scaling and aggregation functions of the form $m_i(x) := m_{\ell,i}(x) \cdot H_i\big((m_{g,i}(z))_{z\in\mathscr{X}}\big)$ for some (meas., possibly nonnegative) functions $m_{l,i} : \mathscr{X} \mapsto \mathbb{R}$, $m_{g,i} : \mathscr{X} \mapsto \mathbb{R}$ and some nonzero scalar functionals H_i thereupon ($i = 1, 2, 3,\ x \in \mathscr{X}$). Accordingly, the components $H_i\big(\ldots\big)$ can be viewed as "global tunings", and may depend adaptively on the primary-interest functions P and Q, i.e. $m_{g,i}(z) = w_{g,i}(x, p(x), q(x))$. For instance, in a finite discrete setup $(\mathscr{X}_{\#}, \lambda_{\#})$ with strictly convex and differentiable ϕ, $m_1(x) \equiv m_2(x) \equiv 1$, $m_3(x) = H_i\big((w_{g,3}(q(x)))_{z\in\mathscr{X}}\big)$ this reduces to the conformal divergences of Nock et al. [64] (they also indicate the extension to equal non-unity scaling $m_1(x) \equiv m_2(x)$), for which the subcase $w_{g,3}(q(x)) := \big(\phi'(q(x))\big)^2$, $H_3\left(\left(h(x)\right)_{x\in\mathscr{X}}\right) := \big(1 + \sum_{x\in\mathscr{X}} h(x)\big)^{-1/2}$ leads to the total Bregman divergences of Liu et al. [44, 45], Vemuri et al. [91]. In contrast, Nock et al. [62] use $m_1(x) \equiv m_1 = H_1\big((p(x))_{z\in\mathscr{X}}\big)$, $m_2(x) \equiv m_1 = H_1\big((q(x))_{z\in\mathscr{X}}\big)$, $m_3(x) \equiv 1$. A more detailed discussion can be found in Stummer and Kißlinger [82] and Roensch and Stummer [74], where they also give versions for nonconvex nonconcave divergence generators. Let us finally mention that for the construction of divergence families, there are other recent paradigms which are essentially different to (1), e.g. by means of measuring the tightness of inequalities (cf. Nielsen et al. [60, 61]), respectively of comparative convexity (cf. Nielsen et al. [59]).

4 Divergences for Essentially Different Functions

4.1 Motivation

Especially in divergence-based statistics, one is often faced with the situation where the functions $p(\cdot)$ and $q(\cdot)$ are of "essentially different nature". For instance, consider the situation where the uncertainty-prone data-generating mechanism is a random variable Y taking values in $\mathscr{X} = \mathbb{R}$ having a "classical" (e.g. Gaussian) probability density $\vec{\mathbb{q}}(\cdot)$ with respect to the one-dimensional Lebesque measure λ_L, i.e. $Pr[Y \in \bullet\,] := \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_L}[\bullet] := \int_{\bullet} \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda_L(x)$ where the latter is almost always a Riemann integral (i.e. $\mathrm{d}\lambda_L(x) = \mathrm{d}x$); notice that we have set $\mathbb{r}(x) \equiv 1$ $(x \in \mathbb{R})$. As already indicated above, under independent and identically distributed (i.i.d.) data observations $Y_1, \ldots, Y_N$ of Y one often builds the corresponding "empirical distribution" $\vec{\mathfrak{P}}_N^{emp}[\bullet] := \frac{1}{N} \cdot \sum_{i=1}^{N} \delta_{Y_i}[\bullet]$ which is nothing but the probability distribution reflecting the underlying (normalized) histogram. By rewriting $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_\#}[\bullet] := \vec{\mathfrak{P}}_N^{emp}[\bullet] = \int_{\bullet} \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda_\#(x)$ with empirical probability mass function $\vec{\mathbb{p}}(x) := \frac{1}{N} \cdot \#\{i \in \{1, \ldots, N\} : Y_i = x\} =: \vec{\mathbb{p}}_N^{emp}(x)$ one encounters some basic problems for a straightforward application of divergence concepts: the two aggregating measures λ_L and $\lambda_\#$ do not coincide and actually they are of "essentially different" nature; moreover, $\vec{\mathbb{p}}(\cdot)$ is nonzero only on the range $\mathscr{R}(Y_1, \ldots, Y_N) = \{z_1, \ldots, z_s\}$ of distinguishable points $z_1, \ldots, z_s$ $(s \leqslant N)$ occupied by $Y_1, \ldots, Y_N$. In particular, one has $\lambda_L[\{z_1, \ldots, z_s\}] = 0$. Accordingly, building a "non-coarsely discriminating" dissimilarity/divergence $D(\vec{\mathbb{P}}, \vec{\mathbb{Q}})$ between such type of functions $\vec{\mathbb{P}} := \big\{\vec{\mathbb{p}}(x)\big\}_{x \in \mathscr{X}}$ and $\vec{\mathbb{Q}} := \big\{\vec{\mathbb{q}}(x)\big\}_{x \in \mathscr{X}}$, is a task like "comparing apples with pears". There are several solutions to tackle this. To begin with, in the following we take the "encompassing" approach of quantifying their dissimilarity by means of their common superordinate characteristics as "fruits". Put in mathematical terms, we choose e.g. $\mathscr{X} = \mathbb{R}$, $\lambda = \lambda_L + \lambda_\#$ and work with the particular representations $\vec{\mathbb{p}}(x) := \widetilde{\vec{p}}(x) \cdot \mathbf{1}_{\{z_1,\ldots,z_s\}}(x)$ with $\widetilde{\vec{p}}(x) > 0$ for λ-almost all $x \in \{z_1, \ldots, z_s\}$ as well as $\vec{\mathbb{q}}(x) := \widetilde{\vec{q}}(x) \cdot \mathbf{1}_{\widetilde{A}\backslash\{z_1,\ldots,z_s\}}(x)$ with $\widetilde{\vec{q}}(x) > 0$ for λ-almost all $x \in \widetilde{A}\backslash\{z_1, \ldots, z_s\}$ with some large enough (measurable) subset $\widetilde{A}$ of $\mathscr{X} = \mathbb{R}$ such that

$$1 = \int_{\mathscr{X}} \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda_\#(x) = \int_{\mathscr{X}} \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda(x) \text{ and } 1 = \int_{\mathscr{X}} \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda_L(x) = \int_{\mathscr{X}} \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda(x) \tag{63}$$

hold. In fact, with these choices one gets $Pr[Y \in \bullet\,] = \int_{\bullet} \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda(x)$ and $\vec{\mathfrak{P}}_N^{emp}[\bullet] = \int_{\bullet} \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda(x)$, as well as

$$\mathbb{p}(x) \cdot \mathbb{q}(x) = 0 \quad \text{for } \lambda\text{-almost all } x \in \mathscr{X}, \tag{64}$$

$$\mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) = \mathbb{p}(x) \quad \text{for } \lambda\text{-almost all } x \in \mathscr{X}, \tag{65}$$

$$\mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) = \mathbb{q}(x) \quad \text{for } \lambda\text{-almost all } x \in \mathscr{X} \tag{66}$$

for the special choices $\mathbb{p}(x) = \vec{\mathbb{p}}(x)$ and $\mathbb{q}(x) = \vec{\mathbb{q}}(x)$. By means of these and (63), the divergence (51) simplifies to

$$D^{c}_{\phi,\overset{\rightharpoondown}{\mathbb{Q}},\overset{\rightharpoondown}{\mathbb{Q}},\mathbb{1}\cdot\overset{\rightharpoondown}{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) = \phi^{*}(0) + \phi(0) - \phi(1) \; > \; 0. \tag{67}$$

Since for arbitrary space $\mathscr{X}$ (and not only $\mathbb{R}$) and any aggregator λ thereupon, the formula (67) holds for all functions $\vec{\mathbb{P}} := \big\{\vec{\mathbb{p}}(x)\big\}_{x\in\mathscr{X}}$, $\vec{\mathbb{Q}} := \big\{\vec{\mathbb{q}}(x)\big\}_{x\in\mathscr{X}}$ which satisfy (63) as well as (64)–(66) for λ-almost all $x \in \mathscr{X}$, and since $\phi^{*}(0) + \phi(0) - \phi(1)$ is just a constant (which may be infinite), these divergences $D^{c}_{\phi,\overset{\rightharpoondown}{\mathbb{Q}},\overset{\rightharpoondown}{\mathbb{Q}},\mathbb{1}\cdot\overset{\rightharpoondown}{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ are not suitable for discriminating between such "essentially different" (basically orthogonal) λ-probability densities $\vec{\mathbb{P}}$ and $\vec{\mathbb{Q}}$. More generally, under the validity of (64)–(66) for λ-almost all $x \in \mathscr{X}$ – which we denote by $\mathbb{P} \perp \mathbb{Q}$ and which basically amounts to pair of functions of the type

$$\mathbb{p}(x) := \widetilde{p}(x) \cdot \mathbf{1}_{A}(x) \quad \text{with } \widetilde{p}(x) > 0 \text{ for } \lambda\text{-almost all } x \in A, \tag{68}$$

$$\mathbb{q}(x) := \widetilde{q}(x) \cdot \mathbf{1}_{B\backslash A}(x) \quad \text{with } \widetilde{q}(x) > 0 \text{ for } \lambda\text{-almost all } x \in B\backslash A, \tag{69}$$

with some (measurable) subsets $\widetilde{A} \subset B$ of $\mathscr{X}$ – the divergence (51) turns into

$$\begin{aligned} D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) &= \big[\phi^{*}(0) - \phi'_{+,c}(1)\big] \cdot \textstyle\int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)\,\mathrm{d}\lambda(x) \\ &\quad + \big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \textstyle\int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)\,\mathrm{d}\lambda(x) \; > \; 0 \end{aligned} \tag{70}$$

which now depends on $\mathbb{P}$ and $\mathbb{Q}$, in a rudimentary "weighted-total-mass" way. Inspired by this, we specify a statistically interesting divergence subclass:

Definition 1 We say that a divergence (respectively dissimilarity respectively distance)[12] $D(\cdot,\cdot)$ is encompassing for a class $\widetilde{\mathscr{P}}$ of functions if

- for arbitrarily fixed $Q := \big\{q(x)\big\}_{x\in\mathscr{X}} \in \widetilde{\mathscr{P}}$ the function $P := \big\{p(x)\big\}_{x\in\mathscr{X}} \to D(P, Q)$ is non-constant on the subfamily of all $P \in \widetilde{\mathscr{P}}$ with $P \perp Q$, and
- for arbitrarily fixed $P \in \widetilde{\mathscr{P}}$ the function $Q \to D(P, Q)$ is non-constant on the subfamily of all $Q \in \widetilde{\mathscr{P}}$ with $Q \perp P$.

Accordingly, due to (67) the prominently used divergences $D^{c}_{\phi,\overset{\rightharpoondown}{\mathbb{Q}},\overset{\rightharpoondown}{\mathbb{Q}},\mathbb{1}\cdot\overset{\rightharpoondown}{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ are not encompassing for the class of $\widetilde{\mathscr{P}}$ of all λ-probability densities; more gener-

[12] i.e. the properties (D1) and (D2) (respectively (D2) respectively (D1), (D2) and (D3)) are satisfied.

ally, because of (70) the divergences $D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})$ are in general encompassing for the class of $\widetilde{\mathscr{P}}$ of all λ-probability densities, but not for $\widetilde{\mathscr{P}} := \{\widetilde{P} := \big\{\widetilde{\mathbb{p}}(x)\big\}_{x\in\mathscr{X}} \mid \int_{\mathscr{X}} \mathbb{r}(x)\cdot\widetilde{\mathbb{p}}(x)\,\mathrm{d}\lambda(x) = \widetilde{c}\}$ for any fixed $\widetilde{c}$.

4.2 $\mathbb{m}_1(\mathbf{x}) = \mathbb{m}_2(\mathbf{x}) := \mathbb{q}(\mathbf{x})$, $\mathbb{m}_3(\mathbf{x}) = \mathbb{r}(\mathbf{x})\cdot\mathbb{q}(\mathbf{x})^{\chi} \in [0,\infty]$ *for Some* $\chi > 1$ *and Some (Measurable) Function* $\mathbb{r} : \mathscr{X} \to [0,\infty[$

In the following, we propose a new way of repairing the above-mentioned encompassing-concerning deficiency for λ-probability density functions, by introducing a new divergence in terms of choosing a generator $\phi :]0,\infty[\to \mathbb{R}$ which is convex and strictly convex at 1, the scaling function $\mathbb{m}_1(x) = \mathbb{m}_2(x) := \mathbb{q}(x)$ as in the non-negativity set-up of Sect. 3.3.1.2, but the more general aggregation function $\mathbb{m}_3(x) = \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi} \in [0,\infty[$ for some power $\chi > 1$ and some (measurable) function $\mathbb{r} : \mathscr{X} \to [0,\infty[$ which satisfies $\mathbb{r}(x) \in]0,\infty[$ for λ-almost all $x \in \mathscr{X}$. To incorporate the zeros of $\mathbb{p}(\cdot)$, $\mathbb{q}(\cdot)$, $\mathbb{r}(\cdot)$ by appropriate limits and conventions, we proceed analogously to Sect. 3.3.1.2. Accordingly, we inspect the boundary behaviour of the function $\widetilde{\psi}_{\phi,c} :]0,\infty[^3 \to [0,\infty[$ given by

$$\widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big) := r\cdot\widetilde{t}^{\chi}\cdot\psi_{\phi,c}\big(\tfrac{\widetilde{s}}{\widetilde{t}},1\big) = r\cdot\widetilde{t}^{\chi}\cdot\Big[\phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big)-\phi(1)-\phi'_{+,c}(1)\cdot\big(\tfrac{\widetilde{s}}{\widetilde{t}}-1\big)\Big]$$
$$= r\cdot\widetilde{t}^{\chi}\cdot\Big[\phi\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r}\big)-\phi(1)-\phi'_{+,c}(1)\cdot\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r}-1\big)\Big] = r\cdot\widetilde{t}^{\chi}\cdot\psi_{\phi,c}\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r},1\big).$$

As in Sect. 3.3.1.2, the Assumption 2(a) is conformly satisfied, for which we use the short-hand notation $2(a)$ etc. in the following discussion. Moreover, we require the validity of 2(b)–2(d) at the point $t = 1$. The analogue of 2(e) is $\mathbb{r}(x)\cdot\widetilde{t}^{\chi} < \infty$ which is always (almost surely) automatically satisfied (a.a.sat.), whereas 2(f) converts to "$\mathbb{r}(x)\cdot\widetilde{t}^{\chi} > 0$ for all $\widetilde{s} \neq \widetilde{t}$" which is also a.a.sat. except for the case $\widetilde{t} = 0$ which will be incorporated below. For the derivation of the analogue of 2(k) we observe that for fixed $r > 0, \widetilde{s} > 0$

$$\ell i_2 := r\cdot 0^{\chi}\cdot\psi_{\phi,c}\big(\tfrac{\widetilde{s}}{0},1\big) := \lim_{t\to 0}\widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big) =$$
$$= r\cdot\widetilde{s}^{\chi}\cdot\lim_{\widetilde{t}\to 0}\Big[\tfrac{\widetilde{t}^{\chi}}{\widetilde{s}^{\chi}}\cdot\phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big)\Big] = r\cdot\widetilde{s}^{\chi}\cdot\phi^{*}_{\chi}(0) \geqslant 0, \tag{71}$$

where $\phi^{*}_{\chi}(0) := \lim_{u\to 0} u^{\chi-1}\cdot u\cdot\phi\big(\frac{1}{u}\big) = \lim_{v\to\infty}\frac{\phi(v)}{v^{\chi}}$ exists but may be infinite. To convert 2(i), we employ the fact that for fixed $r > 0$, $\widetilde{t} > 0$ the function $\widetilde{s} \to \widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big)$ is convex with existing limit

$$\ell i_3 := r\cdot\widetilde{t}^{\chi}\cdot\psi_{\phi,c}\big(\frac{0}{\widetilde{t}},1\big) := \lim_{s\to 0}\widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big)$$
$$= r\cdot\widetilde{t}^{\chi}\cdot\big(\phi(0)+\phi'_{+,c}(1)-\phi(1)\big) > 0. \tag{72}$$

To achieve the analogue of 2(g), let us first remark that for fixed $r > 0$ the function $(\widetilde{s},\widetilde{t}) \rightarrow \widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big)$ may not be continuous at $(\widetilde{s},\widetilde{t}) = (0,0)$, but due to the very nature of a divergence we make the 2(g)-conform convention of setting

$$r \cdot 0^{\chi} \cdot \psi_{\phi,c}\big(\tfrac{0}{0},1\big) := \widetilde{\psi}_{\phi,c}\big(r,0,0\big) := 0\,.$$

The analogues of the Assumptions 2(h), (j), (ℓ), (m), (n) are obsolete because of our basic finiteness requirements. Putting together all the building-blocks, with the above-mentioned limits and conventions we obtain the divergence

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q})\\
&:= \overline{\int}_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)^{\chi}\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi}\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\,\mathrm{d}\lambda(x)\\
&:= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)^{\chi}\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi}\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)^{\chi}\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0)+\phi'_{+,c}(1)-\phi(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi}\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)^{\chi}\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi}\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)^{\chi}\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0)+\phi'_{+,c}(1)-\phi(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi}\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\,.
\end{aligned}\tag{73}$$

In case of $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}_{\chi}[\mathscr{X}] := \int_{\mathscr{X}} \mathbb{q}(x)^{\chi}\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, the divergence (73) becomes

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)^{\chi}\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)^{\chi}\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0)+\phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi}\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&-\phi(1)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi}\,\mathrm{d}\lambda(x)\,.
\end{aligned}\tag{74}$$

Moreover, in case of $\phi\big(1\big) = 0$ and $\int_{\mathscr{X}}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) \in [0,\infty[$ (but not necessarily $\int_{\mathscr{X}} \mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1}\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, $\int_{\mathscr{X}} \mathbb{q}(x)^{\chi}\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$), the divergence (73) turns into

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi}\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)^{\chi}\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) + \phi(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)^{\chi}\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&-\phi'_{+,c}\big(1\big)\cdot\int_{\mathscr{X}}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)\,.
\end{aligned}$$

In contrast to the case $\chi = 1$ where for λ-probability-density functions $\vec{\mathbb{p}}$, $\vec{\mathbb{q}}$, the divergence (53) was further simplified due to $\int_{\mathscr{X}} \big(\vec{\mathbb{p}}(x) - \vec{\mathbb{q}}(x)\big)\, \mathrm{d}\lambda(x) = 0$, for the current setup $\chi > 1$ the latter has no impact for further simplification. However, in general, for the new divergence defined by (73) one gets for any $\mathbb{P} \perp \mathbb{Q}$ from (68), (69), (64)–(66) the expression

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \phi^{*}_{\chi}(0) \cdot \textstyle\int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)^{\chi}\, \mathrm{d}\lambda(x) + \big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \textstyle\int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi}\, \mathrm{d}\lambda(x) \qquad (75)
\end{aligned}$$

which is encompassing for the class of λ-probability functions. By inspection of the above calculations, one can even relax the assumptions away from convexity:

Theorem 6 *Let $\chi > 1$, $c \in [0,1]$, $\phi :]0,\infty[\to \mathbb{R}$ such that both $\phi'_{+,c}(1)$ and $\phi(0) := \lim_{s\to 0}\phi(s)$ exist and $\psi_{\phi,c}(s,1) = \phi(s) - \phi(1) - \phi'_{+,c}(1)\cdot(s-1) \geqslant 0$ for all $s > 0$. Moreover, assume that $\psi_{\phi,c}(s,1) = 0$ if and only if $s = 1$. Furthermore, let the limits $\ell i_2 \geqslant 0$ defined by (71) and $\ell i_3 \geqslant 0$ defined by (72) exist and satisfy $\ell i_2 + \ell i_3 > 0$. Then one gets for the divergence defined by (73):*
(1) $D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q}) \geqslant 0$. Depending on the concrete situation, $D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q})$ may take infinite value.
(2) $D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{p}(x) = \mathbb{q}(x)$ for λ-a.a. $x \in \mathscr{X}$.
(3) For $\mathbb{P} \perp \mathbb{Q}$, the representation (75) holds.

Remark 4 (1) As seen above, if the generator ϕ is in $\Phi(]0,\infty[)$ and satisfies the Assumptions 2(a)–(d) for $t = 1$, then the requirements on ϕ in Theorem 6 are automatically satisfied. The case $\chi = 1$ has already been covered by Theorem 5.
(2) For practical purposes, it is sometimes useful to work with a sub-setup of choices $\chi > 1$, $c \in [0,1]$ and ϕ such that $\ell i_2 \in]0,\infty[$ and/or $\ell i_3 \in]0,\infty[$. □

Let us give some examples. To begin with, for $\alpha \in \mathbb{R}\backslash\{0,1\}$ take the power functions $\phi(t) := \phi_\alpha(t) := \frac{t^\alpha - 1}{\alpha(\alpha-1)} - \frac{t-1}{\alpha-1} \in [0,\infty[$, $\quad t \in]0,\infty[$, with the properties $\phi_\alpha(1) = 0$, $\phi'_\alpha(1) = 0$ (cf. (6)) and $\phi_\alpha(0) := \lim_{t\downarrow 0}\phi_\alpha(t) = \frac{1}{\alpha}\cdot \mathbf{1}_{]0,1]\cup]1,\infty[}(\alpha) + \infty \cdot \mathbf{1}_{]-\infty,0[}(\alpha)$. Then, for arbitrary $\chi \in \mathbb{R}$ one gets the representation

$$\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q}) \\
&:= \overline{\int}_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x)^{\chi} \cdot \phi_\alpha\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi} \cdot \phi_\alpha(1) - \phi'_\alpha(1) \cdot \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\, \mathrm{d}\lambda(x) \\
&\qquad (76) \\
&= \overline{\int}_{\mathscr{X}} \Big[\phi_\alpha\big(\tfrac{\mathbb{p}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big) - \phi_\alpha\big(\tfrac{\mathbb{q}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big) \\
&-\phi'_\alpha\big(\tfrac{\mathbb{q}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big) \cdot \big(\tfrac{\mathbb{p}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))} - \tfrac{\mathbb{q}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big)\Big] \cdot w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x)) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) \\
&= D_{\phi_\alpha,\mathbb{Q}^{\widetilde{\chi}},\mathbb{Q}^{\widetilde{\chi}},\mathbb{R}\cdot\mathbb{Q}^{\widetilde{\chi}},\lambda}(\mathbb{P},\mathbb{Q}) \qquad (77)
\end{aligned}$$

with the adaptive scaling/aggregation function $w_{\widetilde{\chi}}(u,v) = v^{\widetilde{\chi}}$ and $\widetilde{\chi} := 1 + \frac{\chi-1}{1-\alpha}$; in other words, the divergence (76) can be seen as a particularly adaptively scaled

Bregman divergence of non-negative functions in the sense of Kißlinger and Stummer [37], from which their robustness and non-singularity-asymptotical-statistics properties can be derived as a special case (for the probability setup $\vec{\mathbb{P}}$, $\vec{\mathbb{Q}}$, $\mathbb{r}(x) \equiv 1$, and beyond). From (77), it is immediate to see that the case $\chi = 1$ corresponds to the generalized power divergences (58) of order $\alpha \in \mathbb{R}\backslash\{0, 1\}$, whereas $\chi = \alpha$ corresponds to the unscaled divergences (40), i.e.

$$0 \leqslant D_{\phi_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\alpha,\lambda}(\mathbb{P}, \mathbb{Q}) = D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P}, \mathbb{Q}) \tag{78}$$
$$= \overline{\int}_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \Big[\mathbb{p}(x)^\alpha + (\alpha-1)\cdot\mathbb{q}(x)^\alpha - \alpha\cdot\mathbb{p}(x)\cdot\mathbb{q}(x)^{\alpha-1}\Big]\,\mathrm{d}\lambda(x) \ (cf.\ (40))$$

which for $\alpha > 1, \mathbb{r}(x) \equiv 1, \mathbb{p} = \vec{\mathbb{p}}, \mathbb{q} = \vec{\mathbb{q}}$ is a multiple of the α-order density-power divergences DPD used by Basu et al. [10]; as a side remark, in the latter setup our divergence (77) manifests a smooth interconnection between PD and DPD which differs from that of Patra et al. [70], Ghosh et al. [32].

For (76), let us shortly inspect the corresponding ℓi_2 from (71) as well as ℓi_3 from (72). Only for $\alpha \in]0, 1[\cup]1, \infty[$, one gets finite $\ell i_3 = \frac{r\widetilde{t}^\chi}{\alpha} \in]0, \infty[$ for all $\chi \in \mathbb{R}, r > 0, \widetilde{t} > 0$. Additionally, one obtains finite ℓi_2 only for $\chi = 1$, $\alpha \in]0, 1[$ where $\ell i_2 = \frac{r\widetilde{s}}{1-\alpha}$ (PD case), respectively for $\chi > 1$, $\alpha \in]0, 1[\cup]1, \chi[$ where $\ell i_2 = 0$, respectively for $\alpha = \chi > 1$ where $\ell i_2 = \frac{r\widetilde{s}^\alpha}{\alpha\cdot(\alpha-1)}$ (DPD case), for all $r > 0, \widetilde{s} > 0$.

Another interesting example for the divergence $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P}, \mathbb{Q})$ in (73) is given for $\alpha \in \mathbb{R}\backslash\{0, 1\}$ by the generators

$$\phi(t) := \overset{\approx}{\phi}_\alpha(t) := \tfrac{(\alpha-1)\cdot t^\alpha - \alpha\cdot t^{\alpha-1}+1}{\alpha\cdot(\alpha-1)}, \quad t > 0, \quad \overset{\approx}{\phi}_\alpha(1) = 0, \ \overset{\approx}{\phi}'_\alpha(1) = 0,$$

for which $t \to \overset{\approx}{\phi}_\alpha(t) = \overset{\approx}{\phi}_\alpha(t) - \overset{\approx}{\phi}_\alpha(0) - \overset{\approx}{\phi}'_\alpha(1)\cdot(t-1) = \psi_{\phi_\alpha}(t, 1)$ is strictly decreasing on $]0, 1[$ and strictly increasing on $]1, \infty[$. Hence, the corresponding assumptions of Theorem 6 are satisfied. Beyond this, notice that $\overset{\approx}{\phi}_\alpha(\cdot)$ is strictly convex on $]0, \infty[$ if $\alpha \in]1, 2]$, respectively strictly convex on $]1 - \frac{1}{\alpha-1}, \infty[$ and strictly concave on $]0, 1 - \frac{1}{\alpha-1}[$ if $\alpha > 2$, respectively strictly convex on $]0, 1 + \frac{1}{1-\alpha}[$ and strictly concave on $]1 + \frac{1}{1-\alpha}, \infty[$ if $\alpha \in]-\infty, 0[\cup]0, 1[$. Furthermore, the corresponding ℓi_3 is finite only for $\alpha > 1$, namely $\ell i_3 = \frac{r\widetilde{t}^\chi}{\alpha\cdot(\alpha-1)} \in]0, \infty[$ for all $\chi \in \mathbb{R}$, $r > 0, \widetilde{t} > 0$. Additionally, if $\alpha > 1$ one gets finite ℓi_2 only for $\chi > \alpha > 1$ where $\ell i_2 = 0$, respectively for $\alpha = \chi > 1$ where $\ell i_2 = \frac{r\widetilde{s}^\alpha}{\alpha}$ for all $r > 0, \widetilde{s} > 0$. Notice that for $\chi = \alpha > 1$, the limits ℓi_2, ℓi_3 for the cases ϕ_α and $\overset{\approx}{\phi}_\alpha$ are asymmetric. Indeed, by straightforward calculations one can easily see that

$$0 \leqslant D_{\overset{\approx}{\phi}_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\alpha,\lambda}(\mathbb{P}, \mathbb{Q}) = D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{Q}, \mathbb{P})$$
$$= \overline{\int}_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \Big[\big(\mathbb{q}(x)\big)^\alpha + (\alpha-1)\cdot\big(\mathbb{p}(x)\big)^\alpha - \alpha\cdot\mathbb{q}(x)\cdot\big(\mathbb{p}(x)\big)^{\alpha-1}\Big]\,\mathrm{d}\lambda(x) \tag{79}$$

which is the “reversion” of the divergence (40).

4.3 Minimum Divergences - The Encompassing Method

So far, we have almost entirely dealt with aggregated divergences between functions $P := \big\{p(x)\big\}_{x \in \mathscr{X}}$, $Q := \big\{q(x)\big\}_{x \in \mathscr{X}}$ under the *same* aggregator (measure) λ. On the other hand, in Sect. 4.1 we have already encountered an important statistical situation where *two* aggregators λ_1 and λ_2 come into play. Let us now investigate such a context in more detail. To achieve this, for the rest of this paper we confine ourselves to the following probabilistic setup: the modeled respectively observed (random) data take values in a state space $\mathscr{X}$ (with at least two distinct values), equipped with a system $\mathscr{F}$ of admissible events (σ-algebra) and two σ-finite measures λ_1 and λ_2. Furthermore, let $\vec{\mathbbm{P}} := \big\{\vec{\mathbbm{p}}\big\}_{x \in \mathscr{X}}$, $\vec{\mathbbm{Q}} := \big\{\vec{\mathbbm{q}}\big\}_{x \in \mathscr{X}}$ such that $\vec{\mathbbm{p}}(x) \geqslant 0$ for λ_1-a.a. $x \in \mathscr{X}$, $\vec{\mathbbm{q}}(x) \geqslant 0$ for λ_2-a.a. $x \in \mathscr{X}$, $\int_{\mathscr{X}} \vec{\mathbbm{p}}(x) \, \mathrm{d}\lambda_1(x) = 1$, and $\int_{\mathscr{X}} \vec{\mathbbm{q}}(x) \, \mathrm{d}\lambda_2(x) = 1$; in other words, $\vec{\mathbbm{P}}$ is a λ_1-probability density function and $\vec{\mathbbm{Q}}$ is a λ_2-probability density function; the two corresponding probability measures are denoted by $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_1}[\bullet] := \int_{\bullet} \vec{\mathbbm{p}}(x) \, \mathrm{d}\lambda_1(x)$ and $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_2}[\bullet] := \int_{\bullet} \vec{\mathbbm{q}}(x) \, \mathrm{d}\lambda_2(x)$. Notice that we henceforth assume $\mathbbm{r}(x) = 1$ for all $x \in \mathscr{X}$.

More specific, we deal with a parametric framework of double uncertainty in the data and in the model (cf. Sect. 2.4). The former is described by a random variable Y taking values in the space $\mathscr{X}$ and by its probability law $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}[\bullet]$ which (as far as model risk is concerned) is supposed to be unknown but belong to a class $\mathscr{Q}_{\Theta}^{\lambda_2} = \{\vec{\mathfrak{Q}}_{\theta}^{\mathbb{1}\cdot\lambda_2}[\bullet] : \theta \in \Theta\}$ of probability measures on $(\mathscr{X}, \mathscr{F})$ indexed by a set of parameters $\Theta \subset \mathbb{R}^d$ (the non-parametric case works basically in analogous way, with more sophisticated technicalities). Accordingly, all $Pr[Y \in \bullet \,|\, \theta] = \vec{\mathfrak{Q}}_{\theta}^{\mathbb{1}\cdot\lambda_2}[\bullet] = \int_{\bullet} \vec{\mathbbm{q}}_{\theta}(x) \, \mathrm{d}\lambda_2(x)$ ($\theta \in \Theta$) are principal model-candidate laws, with θ_0 to be found out (approximately and with high confidence) by N concrete data observations described by the independent and identically distributed random variables $Y_1, \ldots Y_N$. Furthermore, we assume that the true unknown parameter θ_0 (to be learnt) is identifiable and that the family $\mathscr{Q}_{\Theta}^{\lambda_2}$ is (measure-theoretically) equivalent in the sense

$$\vec{\mathfrak{Q}}_{\theta}^{\mathbb{1}\cdot\lambda_2} \neq \vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2} \quad \text{and} \quad \vec{\mathfrak{Q}}_{\theta}^{\mathbb{1}\cdot\lambda_2} \sim \vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2} \quad \text{for all } \theta, \; \theta_0 \in \Theta \quad \text{with } \; \theta \neq \theta_0. \qquad (80)$$

As usual, the equivalence $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_2} \sim \widehat{\vec{\mathfrak{Q}}}^{\mathbb{1}\cdot\lambda_2}$ means that for λ_2-a.a. $x \in \mathscr{X}$ there holds the density-function-relation: $\vec{\mathbbm{q}}(x) = 0$ if and only if $\widehat{\vec{\mathbbm{q}}}(x) = 0$; this implies in particular that $\vec{\mathbbm{q}}(x) \cdot \mathbf{1}_{\{0\}}\big(\widehat{\vec{\mathbbm{q}}}(x)\big) = 0$ and $\widehat{\vec{\mathbbm{q}}}(x) \cdot \mathbf{1}_{\{0\}}\big(\vec{\mathbbm{q}}(x)\big) = 0$ for λ_2-a.a. $x \in \mathscr{X}$, and by cutting off "datapoints/states of zero contributions" one can then even take $\mathscr{X}$ small enough such that $\vec{\mathbbm{q}}(x) \cdot \widehat{\vec{\mathbbm{q}}}(x) > 0$ (and hence, $\mathbf{1}_{]0,\infty[}\big(\vec{\mathbbm{q}}(x) \cdot \widehat{\vec{\mathbbm{q}}}(x)\big) = 1$) for λ_2-a.a. $x \in \mathscr{X}$. Clearly, since any λ_2-aggregated divergence $D_{\lambda_2}(\cdot, \cdot)$ satisfies (the aggregated version of) the axioms (D1) and (D2), and since θ_0 is identifiable, one gets immediately in terms of the corresponding λ_2-probability density functions $\vec{\mathbbm{Q}}_{\theta} := \big\{\vec{\mathbbm{q}}_{\theta}(x)\big\}_{x \in \mathscr{X}}$

$$\theta_0 = \mathrm{argmin}_{\theta \in \Theta} \, D_{\lambda_2}\big(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}\big) \quad \text{for every } \theta_0 \in \Theta. \tag{81}$$

Inspired by this, one major idea of tracking down (respectively, learning) the true unknown θ_0 is to replace $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}$ by a data-observation-derived – and thus noisy – probability law $\omega \to \vec{\mathfrak{P}}_N^{obs(\omega);\,\mathbb{1}\cdot\lambda_1}[\bullet] := \int_\bullet \vec{\mathbb{p}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\,\mathrm{d}\lambda_1(x)$ where the λ_1-probability density function $\vec{\mathbb{P}}_N^{obs(\omega)} := \big\{\vec{\mathbb{p}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\big\}_{x\in\mathscr{X}}$ depends, as indexed, on the outcome of the observations $Y_1(\omega),\ldots,Y_N(\omega)$. If $\vec{\mathfrak{P}}_N^{obs(\omega);\,\mathbb{1}\cdot\lambda_1}$ converges in distribution to $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}$ as N tends to infinity, then one *intuitively* expects to obtain the so-called *minimum-divergence estimator* ("approximator")

$$\widehat{\theta}_N(\omega) := \widehat{\theta}_{N,D_{\lambda_2}}(\omega) := \mathrm{arginf}_{\theta\in\Theta} \, D_{\lambda_2}\big(\vec{\mathbb{P}}_N^{obs(\omega)}, \vec{\mathbb{Q}}_\theta\big) \tag{82}$$

which estimates θ_0 consistently in the usual sense of the convergence $\theta_n \to \theta_0$ for $n \to \infty$. However, by the nature of our divergence construction, the method (82) makes principal sense only if the two aggregators λ_1 and λ_2 coincide (and if (82) is analytically respectively computationally solvable)! Remark that the minimum distance estimator (82) depends on the choice of the divergence $D_{\lambda_2}(\cdot,\cdot)$.

Subsetup 1. For instance, if by nature the set $\mathscr{X}$ of all possible data points has only countably many elements, say $\mathscr{X} = \mathscr{X}_\# = \{z_1, \ldots z_s\}$ (where s is an integer larger than one or infinity), then a natural model-concerning aggregator is the counting measure $\lambda_2 := \lambda_\#$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}$), and hence $\vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\bullet] = \sum_{x\in\bullet} \vec{\mathbb{q}}_\theta(x) = \sum_{x\in\mathscr{X}} \mathbf{1}_\bullet(x)\cdot\vec{\mathbb{q}}_\theta(x)$ (where $\bullet$ stands for any arbitrary subset of $\mathscr{X}$). In such a context, a popular choice for the data-observation-derived probability law is the so-called "empirical distribution" $\omega \to \vec{\mathfrak{P}}_N^{obs(\omega);\,\mathbb{1}\cdot\lambda_1}[\bullet] = \int_\bullet \vec{\mathbb{p}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\,\mathrm{d}\lambda_1(x) := \sum_{x\in\bullet} \vec{\mathbb{p}}_N^{emp(\omega)}(x) =: \vec{\mathfrak{P}}_N^{emp(\omega)}[\bullet]$, where $\lambda_1 := \lambda_\# = \lambda_2$ and $\vec{\mathbb{p}}_N^{emp(\omega)}(x) := \frac{1}{N}\cdot\#\{i \in \{1,\ldots,N\} : Y_i(\omega) = x\}$ is the total number of x-observations divided by the total number N of observations. In other words, $\vec{\mathfrak{P}}_N^{obs(\omega);\,\mathbb{1}\cdot\lambda_1}[\bullet] := \vec{\mathfrak{P}}_N^{emp(\omega)}[\bullet] := \frac{1}{N}\cdot\sum_{i=1}^N \delta_{Y_i(\omega)}[\bullet]$, where $\delta_z[\bullet]$ is the corresponding Dirac (resp. one-point) distribution given by $\delta_z[A] := \mathbf{1}_A(z)$. Hence, in such a set-up it makes sense to solve the noisy minimization problem

$$\widehat{\theta}_N(\omega) := \widehat{\theta}_{N,D_{\lambda_\#}}(\omega) := \mathrm{arginf}_{\theta\in\Theta} \, D_{\lambda_\#}\big(\vec{\mathbb{P}}_N^{emp(\omega)}, \vec{\mathbb{Q}}_\theta\big) \tag{83}$$

where $\vec{\mathbb{P}}_N^{emp(\omega)} := \big\{\vec{\mathbb{p}}_N^{emp}(x)\big\}_{x\in\mathscr{X}}$ and $D_{\lambda_\#}(\cdot,\cdot)$ is the discrete version of any of the divergences above. Notice that – at least for small enough number N of observations – for some $x \in \mathscr{X}$ with $\lambda_\#[\{x\}] > 0$ one has $\vec{\mathbb{p}}_N^{emp}(x) = 0$ but $\vec{\mathbb{q}}_\theta(x) > 0$ (i.e. an "extreme inlier"), and hence, $\vec{\mathbb{q}}_\theta(x)\cdot\mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}_N^{emp}(x)\big) > 0$; this must be taken into account in the calculation of the explicit forms of the corresponding divergences.[13] By the assumed convergence, this effect disappears as N becomes large enough. □

Subsetup 2. Consider the "crossover case" where $\mathscr{X}$ is uncountable (e.g. $\mathscr{X} = \mathbb{R}$) and the family $\mathscr{Q}_{\Theta}^{\lambda_2}$ is assumed to be *continuous (nonatomic)* in the sense

$$0 = \vec{\mathfrak{Q}}_{\theta}^{\mathbb{1}\cdot\lambda_2}[\{z\}] = Pr[Y \in \{z\} \,|\, \theta] = \textstyle\int_{\mathscr{X}} \mathbf{1}_{\{z\}}(x) \cdot \vec{\mathbb{q}}_{\theta}(x)\, \mathrm{d}\lambda_2(x) \;\text{ for all } z \in \mathscr{X}, \theta \in \Theta \tag{84}$$

(e.g. $\vec{\mathbb{q}}_{\theta}(\cdot)$ are Gaussian densities with mean θ and variance 1), and the data-observation-derived probability law is the "extended" empirical distribution

$$\begin{aligned}\omega \to \vec{\mathfrak{P}}_{N}^{obs(\omega);\mathbb{1}\cdot\lambda_1}[\bullet] &= \textstyle\int_{\bullet} \vec{\mathbb{p}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\, \mathrm{d}\lambda_1(x) \\ &:= \textstyle\sum_{x\in\bullet} \vec{\mathbb{p}}_{N}^{emp(\omega)}(x) \cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) =: \vec{\mathfrak{P}}_{N}^{\overline{emp}(\omega)}[\bullet],\end{aligned} \tag{85}$$

where the extension on $\mathscr{X}$ is accomplished by attributing zeros to all x outside the finite range $\mathscr{R}(Y_1(\omega), \ldots, Y_N(\omega)) = \{z_1(\omega), \ldots, z_s(\omega)\}$ of distinguishable points $z_1(\omega), \ldots, z_s(\omega)$ $(s \leqslant N)$ occupied by the observations $Y_1(\omega), \ldots, Y_N(\omega)$; notice that the involved counting measure given by $\lambda_1[\bullet] := \sum_{z\in\mathscr{X}} \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(z) \cdot \delta_z[\bullet]$ puts 1 to each data-point z which has been observed. Because λ_1 and λ_2 are now essentially different, the minimum-divergence method (82) can not be applied directly (by taking either $\lambda := \lambda_1$ or $\lambda := \lambda_2$), despite of $\vec{\mathfrak{P}}_{N}^{\overline{emp}(\omega)}$ converging in distribution to $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}$ as N tends to infinity. □

There are several ways to circumvent the problem in Subsetup 2. In the following, we discuss in more detail our abovementioned new encompassing approach:

(Enc1) take the encompassing aggregator $\lambda := \lambda_1 + \lambda_2$ and the imbedding $\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)} := \{\vec{\mathbb{p}}_{N}^{\overline{emp}(\omega)}(x)\}_{x\in\mathscr{X}}$ with $\vec{\mathbb{p}}_{N}^{\overline{emp}(\omega)}(x) := \vec{\mathbb{p}}_{N}^{emp(\omega)}(x) \cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)$;

(Enc2) choose a "sufficiently discriminating" (e.g. encompassing) divergence $D_{\lambda}(\cdot, \cdot)$ from above and evaluate them with the density-functions obtained in (Enc1);

(Enc3) solve the corresponding noisy minimization problem

$$\widehat{\theta}_N(\omega) := \widehat{\theta}_{N,D_{\lambda}}(\omega) := \text{arginf}_{\theta\in\Theta}\; D_{\lambda}\big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}, \check{\vec{\mathbb{Q}}}_{\theta}\big) \tag{86}$$

for $\check{\vec{\mathbb{Q}}}_{\theta} := \vec{\mathbb{Q}}_{\theta}$ respectively $\check{\vec{\mathbb{Q}}}_{\theta} := \tilde{\vec{\mathbb{Q}}}_{\theta}$ (to be defined right below);

(Enc4) compute the noisy minimal distance $D_{\lambda}\big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}, \check{\vec{\mathbb{Q}}}_{\theta}\big) > 0$ as an indicator of "goodness of fit" (goodness of noisy approximation");

[13] E.g. applying the divergence (46) for $\alpha \in \mathbb{R}\backslash\{0, 1\}$, the sum-entry $\mathbb{r}(x) \cdot \frac{\vec{\mathbb{q}}_{\theta}(x)^{\alpha}}{\alpha}$ appears, which can be viewed as penalty for the cell x being empty of data observations ("intrinsic empty-cell-penalty"); for divergence (60), the penalty is $\phi(0) \cdot \mathbb{r}(x) \cdot \vec{\mathbb{q}}_{\theta}(x)$.

(Enc5) investigate sound statistical properties of the outcoming estimator $\widehat{\theta}_N(\omega)$, e.g. show probabilistic convergence (as N tends to infinity) to the true unknown parameter θ_0, compute the corresponding convergence speed, analyze its robustness against data-contamination, etc.

Typically, for fixed N the step (Enc3) is not straightforward to solve, and consequently, the tasks described in the unavoidable step (Enc4) become even much more complicated; a detailed discussion of both is – for the sake of brevity – beyond the scope of this paper. As far as (Enc1) is concerned, things are non-trivial due to the generally well-known fact that "continuous" densities are only almost-surely unique. Indeed, consider e.g. the case where the θ-family of functions $\vec{\mathbb{Q}}_\theta := \big\{\vec{\mathbb{q}}_\theta(x)\big\}_{x\in\mathscr{X}}$ satisfies

$$\vec{\mathbb{q}}_\theta(x) > 0 \ \text{ for all } x \in \mathscr{X} \text{ and } \vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\mathscr{X}] = \textstyle\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)\,\mathrm{d}\lambda_2(x) = 1 \ \text{ for all } \theta \in \Theta \tag{87}$$

and the alternative θ-family of functions $\widetilde{\vec{\mathbb{Q}}}_\theta := \big\{\widetilde{\vec{\mathbb{q}}}_\theta(x)\big\}_{x\in\mathscr{X}}$ defined by $\widetilde{\vec{\mathbb{q}}}_\theta(x) := \vec{\mathbb{q}}_\theta(x) \cdot (1 - \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x))$; for the latter, one obtains

$$\widetilde{\vec{\mathfrak{Q}}}_\theta^{\mathbb{1}\cdot\lambda_2}[\mathscr{X}] = \textstyle\int_{\mathscr{X}} \widetilde{\vec{\mathbb{q}}}_\theta(x)\,\mathrm{d}\lambda_2(x) = \int_{\mathscr{X}} \widetilde{\vec{\mathbb{q}}}_\theta(x)\,\mathrm{d}(\lambda_1+\lambda_2)(x) = 1 \ \text{ for all } \theta \in \Theta. \tag{88}$$

Furthermore, due to (85) one has

$$1 = \vec{\mathfrak{P}}_N^{\overline{emp}(\omega)}[\mathscr{X}] = \textstyle\int_{\mathscr{X}} \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\,\mathrm{d}\lambda_1(x) = \int_{\mathscr{X}} \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\,\mathrm{d}(\lambda_1+\lambda_2)(x) \tag{89}$$

and the validity of (64)–(66) with $\mathbb{p}(x) := \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)$, $\mathbb{q}(x) := \widetilde{\vec{\mathbb{q}}}_\theta(x)$ and $\lambda = \lambda_1 + \lambda_2$; in other words, there holds the singularity (measure-theoretical orthogonality) $\vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \perp \widetilde{\vec{\mathbb{Q}}}_\theta$ for all $\theta \in \Theta$. Accordingly, for the step (Enc2) one can e.g. take directly the (family of) encompassing divergences $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q})$ of (73) for $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}$,$\mathbb{Q} := \widetilde{\vec{\mathbb{Q}}}_\theta$, $\lambda := \lambda_1 + \lambda_2$, $\mathbb{r}(x) \equiv 1$, and apply (75) to get

$$\begin{aligned}0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{1}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q}) &= \phi^*_\chi(0) \cdot \textstyle\sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))} \big(\vec{\mathbb{p}}_N^{emp(\omega)}(x)\big)^\chi \\ &+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \textstyle\int_{\mathscr{X}} \big(\vec{\mathbb{q}}_\theta(x)\big)^\chi\,\mathrm{d}\lambda_2(x)\,;\end{aligned} \tag{90}$$

hence, the corresponding solution of (Enc3) does not depend on the data-observations $Y_1(\omega), \ldots, Y_N(\omega)$, and thus is "statistically non-relevant". As an important remark for the rest of this paper, let us mention that – only – in situations where no observations are taken into account, then $\widetilde{\vec{\mathbb{Q}}}_\theta = \vec{\mathbb{Q}}_\theta$, $\mathscr{R}(Y_1, \ldots, Y_N) = \emptyset$, and λ_1 collapses to the "zero aggregator" (i.e. $\lambda_1[\bullet] \equiv 0$).

In contrast, let us replace the alternative θ-family $\widetilde{\vec{\mathbb{Q}}}_\theta$ by the original $\vec{\mathbb{Q}}_\theta$, on which λ_1 acts differently. In fact, instead of (88) there holds

$$1 = \vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\mathscr{X}] = \int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)\,\mathrm{d}\lambda_2(x) < \int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)\,\mathrm{d}(\lambda_1+\lambda_2)(x)$$
$$= 1 + \sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))} \vec{\mathbb{q}}_\theta(x) \qquad \text{for all } \theta \in \Theta\,; \tag{91}$$

moreover, one has for all $\theta \in \Theta$ the non-singularity $\vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \not\perp \vec{\mathbb{Q}}_\theta$ but

$$\mathbf{1}_{\{0\}}\big(\vec{\mathbb{q}}_\theta(x)\big) = 0 \quad \text{for all } x \in \mathscr{X}\,, \tag{92}$$
$$\mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\big) = 1 - \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) \quad \text{for all } x \in \mathscr{X}\,, \tag{93}$$
$$\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\cdot\vec{\mathbb{q}}_\theta(x)\big) = \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) \quad \text{for all } x \in \mathscr{X}\,. \tag{94}$$

Correspondingly, for the step (Enc2) one can e.g. take directly the (family of) encompassing divergences $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q})$ of (73) for $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}$, $\mathbb{Q} := \vec{\mathbb{Q}}_\theta$, $\lambda := \lambda_1 + \lambda_2$, $\mathbb{r}(x) \equiv 1$; the corresponding solution of the noisy minimization problem (Enc3) generally *does depend* on the data-observations $Y_1(\omega), \ldots, Y_N(\omega)$, as required. Let us demonstrate this exemplarily for the special subsetup where $\phi : [0,\infty[\to [0,\infty[$ is continuous (e.g. strictly convex on $]0,\infty[$), differentiable at 1, $\phi(1) = \phi'(1) = 0$, $\phi(t) \in]0,\infty[$ for all $t \in [0,1[\cup[1,\infty[$, $\chi > 1$, $\mathbb{r}(x) \equiv 1$, and $\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\chi\,\mathrm{d}\lambda_2(x) \in]0,\infty[$ for all $\theta \in \Theta$. Then, for each fixed $\theta \in \Theta$ we derive from (73) and (92)–(94) the divergence

$$0 < D_{\phi,\vec{\mathbb{Q}}_\theta,\vec{\mathbb{Q}}_\theta,\mathbb{1}\cdot\vec{\mathbb{Q}}_\theta^{\chi},\lambda_1+\lambda_2}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\vec{\mathbb{Q}}_\theta\big)$$
$$= \int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\chi\cdot\phi\Big(\frac{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{q}}_\theta(x)}\Big)\cdot\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\cdot\vec{\mathbb{q}}_\theta(x)\big)\,\mathrm{d}(\lambda_1+\lambda_2)(x)$$
$$+\phi(0)\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\chi\cdot\mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\big)\,\mathrm{d}(\lambda_1+\lambda_2)(x)$$
$$= \sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))} \vec{\mathbb{q}}_\theta(x)^\chi\cdot\phi\Big(\frac{\vec{\mathbb{p}}_N^{emp(\omega)}(x)}{\vec{\mathbb{q}}_\theta(x)}\Big) + \phi(0)\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\chi\,\mathrm{d}\lambda_2(x) < \infty\,. \tag{95}$$

When choosing this divergence (95) in step (Enc2), we call the solution $\widehat{\theta}_N(\omega)$ of the corresponding noisy minimization problem (86) of step (Enc3) a *minimum* (ϕ,χ)*-divergence estimator* of the true unknown parameter θ_0; in ML and AI contexts, the pair (ϕ,χ) may be regarded as "hyperparameter". Exemplarily, for the power functions $\phi := \phi_\alpha$ (cf. (5)) with $\alpha = \chi > 1$, we obtain from (95) (see also (78), (41)) the divergence

$$]0,\infty[\ni D_{\phi_\alpha,\vec{\mathbb{Q}}_\theta,\vec{\mathbb{Q}}_\theta,\mathbb{1}\cdot\vec{\mathbb{Q}}_\theta^{\alpha},\lambda_1+\lambda_2}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\vec{\mathbb{Q}}_\theta\big) = \tfrac{1}{\alpha}\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\alpha\,\mathrm{d}\lambda_2(x)$$
$$+\sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}\Big[\tfrac{(\vec{\mathbb{p}}_N^{emp(\omega)}(x))^\alpha}{\alpha\cdot(\alpha-1)} - \vec{\mathbb{p}}_N^{emp(\omega)}(x)\cdot\tfrac{\vec{\mathbb{q}}_\theta(x)^{\alpha-1}}{\alpha-1} + \tfrac{\vec{\mathbb{q}}_\theta(x)^\alpha}{\alpha}\Big]$$
$$= \tfrac{1}{\alpha}\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\alpha\,\mathrm{d}\lambda_2(x)$$
$$+\tfrac{1}{N}\sum_{i=1}^N\Big[\tfrac{(\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega)))^{\alpha-1}}{\alpha\cdot(\alpha-1)} - \tfrac{\vec{\mathbb{q}}_\theta(Y_i(\omega))^{\alpha-1}}{\alpha-1} + \tfrac{\vec{\mathbb{q}}_\theta(Y_i(\omega))^\alpha}{\alpha\cdot\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\Big], \tag{96}$$

where for the last equality we have used the representation

$$\sum_{x\in\mathscr{X}} \mathbb{p}_N^{emp(\omega)}(x)\cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)\cdot \delta_x[\bullet] \;=\; \tfrac{1}{N}\cdot \sum_{i=1}^{N} \delta_{Y_i(\omega)}[\bullet]\,; \qquad (97)$$

notice that $\mathbb{p}_N^{emp(\omega)}(Y_i(\omega)) = \#\{j \in \{1,\ldots,N\} : Y_j(\omega) = Y_i(\omega)\}/N$. Clearly, the outcoming minimum (ϕ,χ)-divergence estimator of (95) (and in particular, the minimum (ϕ_α,α)-divergence estimator of (96)) depends on the data observations $Y_1(\omega),\ldots,Y_N(\omega)$, where for technical reasons as e.g. existence and uniqueness – as well as for the tasks (Enc4), (Enc5) – some further assumptions are generally needed; for the sake of brevity, corresponding details will appear in a forthcoming paper.

4.4 Minimum Divergences - Grouping and Smoothing

Next, we briefly indicate two other ways to circumvent the problem described in Subsetup 2 of Sect. 4.3, with continuous (nonatomic) $\mathscr{Q}_\Theta^{\lambda_2}$ and λ_2 from (84):

(GR) grouping (partitioning, quantization) of data: convert[14] everything into a purely discrete context, by subdividing the data-point-set $\mathscr{X} = \bigcup_{j=1}^{s} A_j$ into countably many – (say) $s \in \mathbb{N} \cup \{\infty\}\backslash\{1\}$ – (measurable) disjoint classes $A_1,\ldots,A_s$ with the property $\lambda_2[A_j] > 0$ ("essential partition"); proceed as in Subsetup 1 of Sect. 4.3, with $\mathscr{X}^{new} := \{A_1,\ldots,A_s\}$ instead of $\{z_1,\ldots,z_s\}$, and thus the ith data observation $Y_i(\omega)$ and the corresponding running variable x) manifest (only) the corresponding class-membership. For the subcase of Csiszar-Ali-Slivey divergences and adjacently related divergences, thorough statistical investigations (such as efficiency, robustness, types of grouping, grouping-error sensitivity, etc.) of the corresponding minimum-divergence-estimation can be found e.g. in Victoria-Feser and Ronchetti [92], Menendez et al. [47–49], Morales et al. [52, 53], Lin and He [43].

(SM) smoothing of the empirical density function: convert everything to a purely continuous context, by keeping the original data-point-set $\mathscr{X}$ and by "continuously modifying" (e.g. with the help of kernels) the empirical density $\mathbb{p}_N^{emp}(\cdot)$ to a function $\mathbb{p}_N^{emp,smo}(\cdot) \geqslant 0$ such that $\int_{\mathscr{X}} \mathbb{p}_N^{emp,smo}(x)\,\mathrm{d}\lambda_2(x) = 1$ and that for all $\theta \in \Theta$ there holds: $\mathbb{p}_N^{emp,smo}(x) = 0$ if and only if $\mathbb{q}_\theta(x) = 0$ (in addition to (80)). For the subcase of Csiszar-Ali-Slivey divergences, thorough statistical investigations (such as efficiency, robustness, information loss, etc.) of the corresponding minimum-divergence-estimation can be found e.g. in Basu and Lindsay [11], Park and Basu [69], Chapter 3 of Basu et al. [13], Kuchibhotla and Basu [39], Al Mohamad [5], and the references therein. Due to the "curse of dimensionality", such a solution cannot be applied successfully in a large-dimension setting, as required in the

[14] In several situations, such a conversion can appear in a natural way; e.g. an institution may generate/collect data of "continuous value" but mask them for external data analysts to group-frequencies, for reasons of confidentiality (information asymmetry).

so called "big data" paradigm. For instance (in preparation for divergence valuation), take $\mathscr{X} = \mathbb{R}^d$, λ_2 to be the $d-$dimensional Lebesgue measure and $\vec{\mathbb{p}}_N^{\,emp,smo}(x) := \frac{1}{N}\sum_{i=1}^{N} K(x, Y_i, h_n) = \int_{\mathscr{X}} K(x, y, h_n)\, \mathrm{d}\vec{\mathfrak{P}}_N^{\,emp}(y)$ where $K(\cdot,\cdot,\cdot)$ is an appropriate smooth kernel function with "bandwidth" h_n, e.g. $K(x, y, h_n) := \frac{1}{h_n}\widehat{K}\left(\frac{x-y}{h_n}\right)$ with appropriate nonnegative function $\widehat{K}(\cdot)$ satisfying $\int_{\mathbb{R}^d} \widehat{K}(y)\, \mathrm{d}\lambda_2(y) = 1$. Since such kernel smoothers KS use local averaging, and for large d most neighborhoods tend to be empty of data observations (because data often "live" on lower-dimensional manifolds, sparsity of data), a typical KS technique (choosing concrete kernels and bandwidths, etc.) needs then a huge amount N of data to provide a reasonable accuracy; for $d = 8$ one may need N to be 1 million. For background details, the reader is e.g. referred to DasGupta [28], Scott and Wand [77], Chapter 7 of Scott [76] and the references therein.

For the sake of brevity, a detailed discussion of (GR) and (SM) is beyond the scope of this paper.

4.5 Minimum Divergences - The Decomposability Method

Let us discuss yet another strategy to circumvent the problem described in Subsetup 2 of Sect. 4.3. As a motivation, for a divergence of the form

$$\begin{aligned} 0 \leqslant D_\lambda(\mathbb{P},\mathbb{Q}) &= \int_{\mathscr{X}} f_1(x)\cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} f_2(x)\cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\cdot \mathbf{1}_{]0,\infty[}(\mathbb{q}(x))\,\mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} f_3(x)\cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\cdot \mathbf{1}_{]0,\infty[}(\mathbb{p}(x))\,\mathrm{d}\lambda(x) \end{aligned} \tag{98}$$

with $f_1(x) \geqslant 0$, $f_2(x) \geqslant 0$, $f_3(x) \geqslant 0$, and an "adjacent" dissimilarity

$$\begin{aligned} \widetilde{D_\lambda}(\mathbb{P},\mathbb{Q}) &= \int_{\mathscr{X}} f_1(x)\cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} g_2(x))\cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\cdot \mathbf{1}_{]0,\infty[}(\mathbb{q}(x))\,\mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} g_3(x))\cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\cdot \mathbf{1}_{]0,\infty[}(\mathbb{p}(x))\,\mathrm{d}\lambda(x), \end{aligned} \tag{99}$$

there holds $D_\lambda(\mathbb{P},\mathbb{Q}) = \widetilde{D_\lambda}(\mathbb{P},\mathbb{Q})$ for all equivalent $\mathbb{P} \sim \mathbb{Q}$ (where for both, the second and third integral become zero), but (in case that $g_2(\cdot)$, $g_3(\cdot)$ differ sufficiently enough from $f_2(\cdot)$, $f_3(\cdot)$) one gets $D_\lambda(\mathbb{P},\mathbb{Q}) \neq \widetilde{D_\lambda}(\mathbb{P},\mathbb{Q})$ for $\mathbb{P} \perp \mathbb{Q}$ and even for $\mathbb{P} \nsim \mathbb{Q}$; in the latter two cases, depending on the signs of $g_2(\cdot)$, $g_3(\cdot)$, $\widetilde{D_\lambda}(\mathbb{P},\mathbb{Q})$ may even become negative.

Such issues are of importance for our current problem where e.g. $\mathbb{P} := \vec{\mathbb{P}}_N^{\,\overline{emp}(\omega)} \perp \widetilde{\vec{\mathbb{Q}}}_\theta =: \mathbb{Q}$. For further illuminations, and for the sake of a compact presentation, we use henceforth the notations $\mathscr{P}^\lambda$ for an arbitrarily fixed class of nonnegative, mutually equivalent functions (i.e. $\mathbb{P}_1 \sim \mathbb{P}_2$ for all $\mathbb{P}_1 \in \mathscr{P}^\lambda$, $\mathbb{P}_2 \in \mathscr{P}^\lambda$), and $\mathscr{P}^{\lambda\nsim}$ for a corresponding class of nonnegative (not necessarily mutually equivalent) functions such that $\mathbb{P}_1 \not\sim \mathbb{P}_2$ for all $\mathbb{P}_1^\lambda \in \mathscr{P}^\lambda$, $\mathbb{P}_2 \in \mathscr{P}^{\lambda\not\sim}$. Furthermore, we employ $\widetilde{\mathscr{P}}^\lambda := \mathscr{P}^\lambda \cup \mathscr{P}^{\lambda\not\sim}$ and specify:

Definition 2 We say that a function $D_\lambda : \widetilde{\mathscr{P}}^\lambda \otimes \mathscr{P}^\lambda \rightarrow \mathbb{R}$ is a pseudo-divergence on $\widetilde{\mathscr{P}}^\lambda \times \mathscr{P}^\lambda$, if its restriction to $\mathscr{P}^\lambda \cup \mathscr{P}^\lambda$ is a divergence, i.e.

$$\begin{aligned} & D_\lambda(\mathbb{P}, \mathbb{Q}) \geqslant 0 \ \text{ for all } \ \mathbb{P} \in \mathscr{P}^\lambda, \mathbb{Q} \in \mathscr{P}^\lambda, \quad \text{and} \\ & D_\lambda(\mathbb{P}, \mathbb{Q}) = 0 \ \text{ if and only if } \ \mathbb{P} = \mathbb{Q} \in \mathscr{P}^\lambda \, . \end{aligned} \tag{100}$$

If also $D_\lambda(\mathbb{P}, \mathbb{Q}) > 0$ for all $\mathbb{P} \in \mathscr{P}^{\lambda\not\sim}$,$\mathbb{Q} \in \mathscr{P}^\lambda$, then $D_\lambda(\cdot, \cdot)$ is a divergence.

As for interpretation, a pseudo-divergence $D_\lambda(\cdot, \cdot)$ acts like a divergence if both arguments are from $\mathscr{P}^\lambda$, but only like a dissimilarity if the first argument is from $\mathscr{P}^{\lambda\not\sim}$ and thus is "quite different" from the second argument. In the following, we often use pseudo-divergences for our noisy minimum-distance-estimation problem – cf. (81), (82) – by taking $\lambda = \lambda_1 + \lambda_2$, $\mathscr{P}^\lambda := \mathscr{P}_\Theta^\lambda := \big(\widetilde{\vec{\mathbb{Q}}}_\theta\big)_{\theta \in \Theta} := \big(\big\{\widetilde{\vec{\mathbb{q}}}_\theta(x)\big\}_{x \in \mathscr{X}}\big)_{\theta \in \Theta}$ (cf. (87), (88)), and $\mathscr{P}^{\lambda\not\sim} := \mathscr{P}_{emp}^{\lambda\perp} := \big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\big)_{N \in \mathbb{N}} = \big(\big\{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\big\}_{x \in \mathscr{X}}\big)_{N \in \mathbb{N}}$ (cf. (85), (Enc1)) covering all numbers N of data observations (sample sizes), and the according $\widetilde{\mathscr{P}}^\lambda := \mathscr{P}_{\Theta,emp}^\lambda = \mathscr{P}_\Theta^\lambda \cup \mathscr{P}_{emp}^{\lambda\perp}$; notice that by construction we have even the function-class-relationship $\perp$ which is stronger than $\not\sim$. In such a setup, we have seen that for the choice $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}$, $\mathbb{Q} := \widetilde{\vec{\mathbb{Q}}}_\theta$ the divergence $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{1}\cdot\mathbb{Q}^\chi,\lambda}$ $(\mathbb{P}, \mathbb{Q}) > 0$ of (90) is unpleasant for (Enc3) since the solution does not depend on the data-observations $Y_1(\omega), \ldots, Y_N(\omega)$; also recall the special case of power functions $\phi := \phi_\alpha$ (cf. (5)) with $\alpha = \chi > 1$ which amounts to the unscaled divergences (78), (40) and thus to (41). In (95), for general ϕ we have repaired this deficiency by replacing $\mathbb{Q} := \widetilde{\vec{\mathbb{Q}}}_\theta$ with $\mathbb{Q} := \vec{\mathbb{Q}}_\theta$, at the cost of getting total mass larger than 1 but by keeping the strict positivity of the involved divergence; especially for $\phi := \phi_\alpha$, the divergence (41) has then amounted to (96).

In contrast, let us show another method to repair the (Enc3)-deficiency of (41), by sticking to $\mathbb{Q} := \widetilde{\vec{\mathbb{Q}}}_\theta$ but changing the basically underlying divergence. In fact, we deal with the even more general

Definition 3 (a) We say that a pseudo-divergence $D_\lambda : \widetilde{\mathscr{P}}^\lambda \otimes \mathscr{P}^\lambda \rightarrow \mathbb{R}$ is decomposable if there exist functionals $\mathfrak{D}^0 : \widetilde{\mathscr{P}}^\lambda \mapsto \mathbb{R}$, $\mathfrak{D}^1 : \mathscr{Q} \mapsto \mathbb{R}$ and a (measurable) mapping $\rho_\mathbb{Q} : \mathscr{X} \mapsto \mathbb{R}$ (for each $\mathbb{Q} \in \mathscr{P}^\lambda$) such that[15]

[15] In an encompassing way, the part (a) reflects a measure-theoretic "plug-in" version of decomposable pseudo-divergences $D : (\mathscr{P}^{meas,\lambda_1} \cup \mathscr{P}^{meas,\lambda_2}) \otimes \mathscr{P}^{meas,\lambda_1} \mapsto \mathbb{R}$, where $\mathscr{P}^{meas,\lambda_1}$ is a family of mutually equivalent nonnegative measures of the form $\mathfrak{P}[\bullet] := \mathfrak{P}^{\mathbb{1}\cdot\lambda_1}[\bullet] := \int_\bullet \mathbb{p}(x)\, \mathrm{d}\lambda_1(x)$, $\mathscr{P}^{meas,\lambda_2}$ is a family of nonnegative measures of the form $\overline{\mathfrak{P}}[\bullet] := \overline{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_2}[\bullet] := \int_\bullet \mathbb{q}(x)\, \mathrm{d}\lambda_2(x)$ such that any $\mathfrak{P} \in \mathscr{P}^{meas,\lambda_1}$ is not equivalent to any $\overline{\mathfrak{P}} \in \mathscr{P}^{meas,\lambda_2}$, and (101) is replaced with $D(\mathfrak{P}, \mathfrak{Q}) = \mathfrak{D}^0(\mathfrak{P}) + \mathfrak{D}^1(\mathfrak{Q}) + \int_\mathscr{X} \rho_\mathfrak{Q}(x)\, \mathrm{d}\mathfrak{P}(x)$ for all $\mathbb{P} \in \mathfrak{P} \in \mathscr{P}^{meas,\lambda_1} \cup \mathscr{P}^{meas,\lambda_2}$, $\mathfrak{Q} \in \mathscr{P}^{meas,\lambda_2}$; cf. Vajda [90], Broniatowski and Vajda [18], Broniatowski et al. [19]; part (b) is new.

$$D_{\lambda}(\mathbb{P}, \mathbb{Q}) \ = \ \mathfrak{D}^0(\mathbb{P}) + \mathfrak{D}^1(\mathbb{Q}) + \int_{\mathscr{X}} \ \rho_{\mathbb{Q}}(x) \cdot \mathbb{p}(x) \, \mathrm{d}\lambda(x) \ \ \text{for all } \mathbb{P} \in \widetilde{\mathscr{P}}^{\lambda}, \mathbb{Q} \in \mathscr{P}^{\lambda}. \tag{101}$$

(b) We say that a pseudo-divergence $D_{\lambda} : \widetilde{\mathscr{P}}^{\lambda} \otimes \mathscr{P}^{\lambda} \rightarrow \mathbb{R}$ is pointwise decomposable if it is of the form $D_{\lambda}(\mathbb{P}, \mathbb{Q}) = \int_{\mathscr{X}} \psi^{dec}(\mathbb{p}(x), \mathbb{q}(x)) \, \mathrm{d}\lambda(x)$ for some (measurable) mapping $\psi^{dec} : [0, \infty[\times[0, \infty[\mapsto \mathbb{R}$ with representation

$$\begin{aligned} &\psi^{dec}(s, t) := \psi^0\big(s + h_0(x, s) \cdot \mathbf{1}_{\{0\}}(t)\big) \cdot \mathbf{1}_{]\overline{c}_0, \infty[}(s) \cdot \mathbf{1}_{]c_0, \infty[}(t) \\ &+\psi^1\big(t + h_1(x) \cdot \mathbf{1}_{\{0\}}(t)\big) \cdot \mathbf{1}_{]c_1, \infty[}(t) \\ &+\rho\big(t + h_2(x) \cdot \mathbf{1}_{\{0\}}(t)\big) \cdot s \quad \text{for all } (s, t) \in [0, \infty[\times[0, \infty[\backslash\{(0, 0)\} \, , \\ &\psi^{dec}(0, 0) := 0, \end{aligned} \tag{102}$$

with constants $c_0, c_1, \overline{c}_0 \in \{0, 1\}$, and (measurable) mappings $\psi^0, \psi^1, \rho : [0, \infty[\mapsto \mathbb{R}$, $h_1, h_2 : \mathscr{X} \mapsto [0, \infty[$, $h_0 : \mathscr{X} \times [0, \infty[\mapsto \mathbb{R}$, such that

$$\psi^{dec}(s, t) = \psi^0(s) + \psi^1(t) + \rho(t) \cdot s \geqslant 0 \quad \text{for all } (s, t) \in]0, \infty[\times]0, \infty[\, , \tag{103}$$

$$\psi^{dec}(s, t) = 0 \quad \text{if and only if} \quad s = t \, , \tag{104}$$

$$s + h_0(x, s) \geqslant 0 \quad \text{for all } s \in [0, \infty[\text{ and } \lambda\text{-almost all } x \in \mathscr{X} \, .$$

Remark 5 (a) Any pointwise decomposable pseudo-divergence is decomposable, under the additional assumption that the integral $\int_{\mathscr{X}} \ldots \, \mathrm{d}\lambda(x)$ can be split into three appropriate parts.
(b) For use in (Enc3), $\mathfrak{D}^1(\cdot)$ and $\rho_{\mathbb{Q}}(\cdot)$ should be non-constant.
(c) In the Definitions 2 and 3 we have put the "extension-role" to the first component $\mathbb{P}$; of course, everything can be worked out analogously for the second component $\mathbb{Q}$ by using (pseudo-)divergences $D_{\lambda} : \mathscr{P}^{\lambda} \times \widetilde{\mathscr{P}}^{\lambda} \rightarrow \mathbb{R}$.
(d) We could even extend (102) for bivariate functions $h_1(x, s), h_2(x, s)$. □

Notice that from (102) one obtains the boundary behaviour

$$\mathbb{R} \ni \psi^{dec}(s, 0) \ = \ \psi^0(s + h_0(x, s)) \ \cdot \ \check{c_0} \ + \ \psi^1(h_1(x)) \cdot \check{c_1} \ + \ \rho(h_2(x)) \ \cdot \ s \text{ for all } s > 0, \tag{105}$$

$$\mathbb{R} \ni \psi^{dec}(0, t) = \psi^0(0) \cdot \check{\overline{c}_0} + \psi^1(t) \quad \text{for all } t > 0 \, , \tag{106}$$

with $\check{c_0} := \mathbf{1}_{]c_0, \infty[}(0)$, $\check{c_1} := \mathbf{1}_{]c_1, \infty[}(0)$, $\check{\overline{c}_0} := \mathbf{1}_{]\overline{c}_0, \infty[}(0)$. Notice that $\psi^{dec}(s, 0)$ of (105) does generally not coincide with the eventually existent "(103)-limit" $\lim_{t \rightarrow 0}[\psi^0(s) + \psi^1(t) + \rho(t) \cdot s]$ $(s > 0)$, which reflects a possibly "non-smooth boundary behaviour" (also recall (98), (99)). Moreover, when choosing a decomposable pseudo-divergence (101) in step (Enc2), we operationalize the solution $\widehat{\theta}_N(\omega)$ of the corresponding noisy minimization problem (86) of step (Enc3) as follows:

Definition 4 (a) We say that a functional $T_{D_\lambda} : \mathscr{P}^{\lambda}_{\Theta,emp} \mapsto \Theta$ generates a minimum decomposable pseudo-divergence estimator (briefly, min $-decD_\lambda$-estimator)

$$\widehat{\theta}_{N,decD_\lambda}(\omega) := T_{D_\lambda}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\big) \quad \text{for} \quad \vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \in \mathscr{P}^{\lambda\perp}_{emp} \tag{107}$$

of the true unknown parameter θ_0, if $D_\lambda(\cdot,\cdot) : \mathscr{P}^{\lambda}_{\Theta,emp} \otimes \mathscr{P}^{\lambda}_{\Theta} \mapsto \mathbb{R}$ is a decomposable pseudo-divergence and

$$T_{D_\lambda}\big(\vec{\mathbb{P}}\big) \;=\; \operatorname{arginf}_{\theta\in\Theta}\big[\mathfrak{D}^1(\vec{\mathbb{Q}}_\theta) \;+\; \textstyle\int_{\mathscr{X}} \rho_{\vec{\mathbb{Q}}_\theta}(x) \;\cdot\; \vec{\mathbb{p}}(x)\, \mathrm{d}\lambda(x)\big] \;\text{for all}\; \vec{\mathbb{P}} \in \mathscr{P}^{\lambda}_{\Theta,emp}. \tag{108}$$

(b) If $D_\lambda(\cdot,\cdot)$ is a pointwise decomposable pseudo-divergence we replace (108) by

$$T_{D_\lambda}\big(\vec{\mathbb{P}}\big) = \operatorname{arginf}_{\theta\in\Theta} \textstyle\int_{\mathscr{X}} \psi^{dec}(\vec{\mathbb{p}}(x), \widetilde{\vec{\mathbb{q}}}_\theta(x))\, \mathrm{d}\lambda(x) \quad \text{for all}\; \vec{\mathbb{P}} \in \mathscr{P}^{\lambda}_{\Theta,emp}\,,$$

but do not introduce a new notion (also recall that $\lambda = \lambda_2$ and $\widetilde{\vec{\mathbb{q}}}_\theta(\cdot) = \vec{\mathbb{q}}_\theta(\cdot)$ for the case of no observations, e.g. if $\vec{\mathbb{P}} \in \mathscr{P}^{\lambda_2}_{\Theta}$).

To proceed, let us point out that by (107) and (97) every min $-decD_\lambda$-estimator rewrites straightforwardly as

$$\widehat{\theta}_{N,decD_\lambda}(\omega) = \operatorname{arginf}_{\theta\in\Theta}\big[\mathfrak{D}^1(\vec{\mathbb{Q}}_\theta) + \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \rho_{\vec{\mathbb{Q}}_\theta}(Y_i(\omega))\big] \tag{109}$$

and is Fisher consistent in the sense that

$$T_{\mathfrak{D}}(\vec{\mathbb{Q}}_{\theta_0}) = \operatorname{arginf}_{\theta\in\Theta} \mathfrak{D}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta) = \theta_0 \quad \text{for all}\; \theta_0 \in \Theta\,. \tag{110}$$

Furthermore, the criterion to be minimized in (109) is of the form

$$\theta \;\;\mapsto\;\; \mathfrak{D}^1(\vec{\mathbb{Q}}_\theta) + \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \rho_{\vec{\mathbb{Q}}_\theta}(Y_i(\omega))$$

which e.g. for the task (Enc5) opens the possibility to apply the methods of the asymptotic theory of so-called M-estimators (cf. e.g. Hampel et al. [33], van der Vaart and Wellner [88], Liese and Mieske [40]). The concept of min $-decD_\lambda$-estimators (101) were introduced in Vajda [90], Broniatowski and Vajda [18] within the probability-law-restriction of the non-encompassing, "plug-in" context of footnote 15.

In the following, we demonstrate that our new concept of pointwise decomposability defined by (102) is very useful and flexible for creating new min $-decD_\lambda$-estimators and imbedding existing ones. In fact, since in our current statistics-ML-AI context we have chosen $\lambda[\bullet] := \lambda_1[\bullet] + \lambda_2[\bullet]$ with
$\lambda_1[\bullet] := \sum_{z\in\mathscr{X}} \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(z) \cdot \delta_z[\bullet]$ and $\lambda_2[\bullet]$ stemming from (87), we have seen that $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \perp \widetilde{\vec{\mathbb{Q}}}_\theta =: \vec{\mathbb{Q}}$ for all $\theta \in \Theta$. Hence, from (102), (105), (106) we obtain

$$\begin{aligned}
& D_{\lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) = \int_{\mathscr{X}} \psi^{dec}\Big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x), \widetilde{\vec{\mathbb{q}}}_{\theta}(x)\Big) \, \mathrm{d}\lambda(x) \\
& = \int_{\mathscr{X}} \Big[\psi^{0}(0) \cdot \check{\bar{c}}_0 + \psi^{1}\Big(\widetilde{\vec{\mathbb{q}}}_{\theta}(x)\Big)\Big] \cdot \mathbf{1}_{\{0\}}(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)) \, \mathrm{d}(\lambda_1 + \lambda_2)(x) \\
& + \int_{\mathscr{X}} \Big[\psi^{0}\Big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x) + h_0\Big(x, \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\Big)\Big) \cdot \check{c}_0 \\
& + \psi^{1}(h_1(x)) \cdot \check{c}_1 + \rho(h_2(x)) \cdot \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\Big] \cdot \mathbf{1}_{\{0\}}(\widetilde{\vec{\mathbb{q}}}_{\theta}(x)) \, \mathrm{d}(\lambda_1 + \lambda_2)(x) \\
& = \int_{\mathscr{X}} \Big[\psi^{0}(0) \cdot \check{\bar{c}}_0 + \psi^{1}\Big(\vec{\mathbb{q}}_{\theta}(x)\Big)\Big] \, \mathrm{d}\lambda_2(x) + \sum_{x \in \mathscr{X}} \Big[\psi^{0}\Big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x) \\
& + h_0\Big(x, \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\Big)\Big) \cdot \check{c}_0 \\
& + \psi^{1}(h_1(x)) \cdot \check{c}_1 + \rho(h_2(x)) \cdot \vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\Big] \cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega), \ldots, Y_N(\omega))}(x) \\
& = \int_{\mathscr{X}} \Big[\psi^{0}(0) \cdot \check{\bar{c}}_0 + \psi^{1}\Big(\vec{\mathbb{q}}_{\theta}(x)\Big)\Big] \, \mathrm{d}\lambda_2(x) + \tfrac{1}{N} \sum_{i=1}^{N} \rho(h_2(Y_i(\omega))) \\
& + \tfrac{1}{N} \sum_{i=1}^{N} \frac{\psi^{0}\Big(\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega)) + h_0\Big(Y_i(\omega), \vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))\Big)\Big) \cdot \check{c}_0 + \psi^{1}(h_1(Y_i(\omega))) \cdot \check{c}_1}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))} ,
\end{aligned} \tag{111}$$

where we have employed (97); recall that $\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega)) = \#\{j \in \{1, \ldots, N\} : Y_j(\omega) = Y_i(\omega)\}/N$. Hence, we always choose $\mathfrak{D}^1(\vec{\mathbb{Q}}) = \mathfrak{D}^1(\widetilde{\vec{\mathbb{Q}}}_{\theta}) = \int_{\mathscr{X}} \Big[\psi^{0}(0) + \psi^{1}\Big(\widetilde{\vec{\mathbb{q}}}_{\theta}(x)\Big)\Big] \, \mathrm{d}\lambda_2(x) = \int_{\mathscr{X}} \Big[\psi^{0}(0) + \psi^{1}\Big(\vec{\mathbb{q}}_{\theta}(x)\Big)\Big] \, \mathrm{d}\lambda_2(x) = \mathfrak{D}^1(\vec{\mathbb{Q}}_{\theta})$. Notice that the functions h_0, h_1, h_2 may depend on the parameter θ. Indeed, for $h_0(x, s) \equiv 0$, $h_1(x) \equiv 0$, $h_2(x) = \vec{\mathbb{q}}_{\theta}(x)$ $(\neq \widetilde{\vec{\mathbb{q}}}_{\theta}(x))$, the pseudo-divergence (111) turns into

$$\begin{aligned}
& D_{\lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) = \int_{\mathscr{X}} \Big[\psi^{0}(0) \cdot \check{\bar{c}}_0 + \psi^{1}\Big(\vec{\mathbb{q}}_{\theta}(x)\Big)\Big] \, \mathrm{d}\lambda_2(x) + \tfrac{1}{N} \sum_{i=1}^{N} \rho\Big(\vec{\mathbb{q}}_{\theta}(Y_i(\omega))\Big) \\
& + \tfrac{1}{N} \sum_{i=1}^{N} \frac{\psi^{0}\Big(\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))\Big) \cdot \check{c}_0 + \psi^{1}(0) \cdot \check{c}_1}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))} ,
\end{aligned} \tag{112}$$

whereas for $h_0(x, s) \equiv 0$, $h_1(x) = \vec{\mathbb{q}}_{\theta}(x)$, $h_2(x) = \vec{\mathbb{q}}_{\theta}(x)$, (111) becomes

$$\begin{aligned}
& D_{\lambda}(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) = \int_{\mathscr{X}} \big[\psi^{0}(0) \cdot \check{\bar{c}}_0 + \psi^{1}\big(\vec{\mathbb{q}}_{\theta}(x)\big)\big] \, \mathrm{d}\lambda_2(x) \\
& + \tfrac{1}{N} \sum_{i=1}^{N} \Big[\rho\Big(\vec{\mathbb{q}}_{\theta}(Y_i(\omega))\Big) + \tfrac{\check{c}_1 \cdot \psi^{1}(\vec{\mathbb{q}}_{\theta}(Y_i(\omega)))}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\Big] + \tfrac{1}{N} \sum_{i=1}^{N} \frac{\psi^{0}\Big(\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))\Big) \cdot \check{c}_0}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))} .
\end{aligned} \tag{113}$$

The last sum in (112) respectively (113) is the desired $\mathfrak{D}^0(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)})$. As an example, let us take $c_0 = c_1 = \bar{c}_0 = -1$ (and hence, $\check{c}_0 = \check{c}_1 = \check{\bar{c}}_0 = 1$) and for $\alpha > 1$ the power functions $\phi(t) := \phi_{\alpha}(t) := \frac{t^{\alpha} - \alpha \cdot t + \alpha - 1}{\alpha \cdot (\alpha - 1)}$ $(t \in]0, \infty[)$ of (6), for which by (9) and (103) one derives immediately the decomposition $\psi^0(t) := \psi^0_\alpha(t) := \frac{t^\alpha}{\alpha(\alpha-1)} > 0, \psi^1(t) := \psi^1_\alpha(t) := \frac{t^\alpha}{\alpha} > 0, \rho(t) := \rho_\alpha(t) := -\frac{t^{\alpha-1}}{\alpha-1} < 0\,(t \in]0,\infty[)$. Accordingly, (111) simplifies to

$$\begin{aligned} & D_\lambda\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) := D_{\lambda,\alpha}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) \\ & = \tfrac{1}{\alpha}\int_{\mathscr{X}} \vec{\mathrm{q}}_\theta(x)^\alpha \,\mathrm{d}\lambda_2(x) - \tfrac{1}{N\cdot(\alpha-1)} \textstyle\sum_{i=1}^N \Big(h_2(Y_i(\omega))\Big)^{\alpha-1} \\ & + \tfrac{1}{N} \textstyle\sum_{i=1}^N \frac{\Big(\vec{\mathrm{p}}_N^{emp(\omega)}(Y_i(\omega)) + h_0\Big(Y_i(\omega), \vec{\mathrm{p}}_N^{emp(\omega)}(Y_i(\omega))\Big)\Big)^\alpha + (\alpha-1)\cdot\Big(h_1(Y_i(\omega))\Big)^\alpha}{\alpha\cdot(\alpha-1)\cdot\vec{\mathrm{p}}_N^{emp(\omega)}(Y_i(\omega))}, \end{aligned} \tag{114}$$

and in particular the special case (112) turns into

$$\begin{aligned} & D_{\lambda,\alpha}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \tfrac{1}{\alpha}\int_{\mathscr{X}} \vec{\mathrm{q}}_\theta(x)^\alpha \,\mathrm{d}\lambda_2(x) - \tfrac{1}{N\cdot(\alpha-1)} \textstyle\sum_{i=1}^N \Big(\vec{\mathrm{q}}_\theta(Y_i(\omega))\Big)^{\alpha-1} \\ & + \tfrac{1}{N\cdot\alpha\cdot(\alpha-1)} \textstyle\sum_{i=1}^N \Big(\vec{\mathrm{p}}_N^{emp(\omega)}(Y_i(\omega))\Big)^{\alpha-1}, \end{aligned} \tag{115}$$

whereas the special case (113) simplifies to

$$\begin{aligned} & 0 < D_{\lambda,\alpha}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \tfrac{1}{\alpha}\int_{\mathscr{X}} \vec{\mathrm{q}}_\theta(x)^\alpha \,\mathrm{d}\lambda_2(x) \\ & + \tfrac{1}{N} \textstyle\sum_{i=1}^N \Big[\frac{\Big(\vec{\mathrm{q}}_\theta(Y_i(\omega))\Big)^\alpha}{\alpha\cdot\vec{\mathrm{p}}_N^{emp(\omega)}(Y_i(\omega))} - \frac{\Big(\vec{\mathrm{q}}_\theta(Y_i(\omega))\Big)^{\alpha-1}}{\alpha-1}\Big] + \tfrac{1}{N}\textstyle\sum_{i=1}^N \frac{\Big(\vec{\mathrm{p}}_N^{emp(\omega)}(Y_i(\omega))\Big)^{\alpha-1}}{\alpha\cdot(\alpha-1)}. \end{aligned} \tag{116}$$

Notice that (116) coincides with (96), but both were derived within quite different frameworks: to obtain (116) we have used the concept of decomposable pseudo-divergences (which may generally become negative at the boundary) together with $\vec{\mathbb{Q}} := \widetilde{\vec{\mathbb{Q}}}_\theta$ which leads to total mass of 1 (cf. (88)); on the other hand, for establishing (96) we have employed the concept of divergences (which are generally always strictly positive at the boundary) together with $\vec{\mathbb{Q}} := \vec{\mathbb{Q}}_\theta$ which amounts to total mass greater than 1 (cf. (91)). Moreover, choosing $h_0(x,s) \equiv 0$, $h_1(x) \equiv 0$, $h_2(x) \equiv 0$ in (114) gives exactly the divergence (90) for the current generator $\phi(t) := \phi_\alpha(t)$ with $\alpha > 1$; recall that the latter has been a starting motivation for the search of repairs. For $c_0 = c_1 = \overline{c}_0 = -1$ and the limit case $\alpha \to 1$ one gets $\phi(t) := \phi_1(t) := t \cdot \log t + 1 - t\; (t \in]0,\infty[)$ of (18), for which by (22) and (103) we obtain the decomposition $\psi^0(t) := \psi^0_1(t) := t \cdot \log t - t$, $\psi^1(t) := \psi^1_1(t) := t > 0$, $\rho(t) := \rho_1(t) := -\log t$. Accordingly, (111) simplifies to

$$
\begin{aligned}
& D_{\lambda}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) := D_{\lambda,1}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) \\
& = 1 - \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \log\Big(h_2(Y_i(\omega))\Big) \\
& + \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \frac{\psi_1^0\left(\vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega)) + h_0\left(Y_i(\omega), \vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega))\right)\right) + h_1(Y_i(\omega))}{\vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega))} \, , \qquad (117)
\end{aligned}
$$

and in particular the special case (112) turns into

$$
D_{\lambda,1}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) = \frac{1}{N}\sum_{i=1}^{N} \log\Big(\vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega))\Big) - \frac{1}{N}\sum_{i=1}^{N} \log\Big(\vec{\mathbb{q}}_{\theta}(Y_i(\omega))\Big), \qquad (118)
$$

whereas the special case (113) becomes

$$
\begin{aligned}
& 0 < D_{\lambda,1}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \log\Big(\vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega))\Big) - \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \log\Big(\vec{\mathbb{q}}_{\theta}(Y_i(\omega))\Big) \\
& + \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \frac{\vec{\mathbb{q}}_{\theta}(Y_i(\omega))}{\vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega))} \, . \qquad (119)
\end{aligned}
$$

To end up this subsection, let us briefly indicate that choosing in step (Enc2) a decomposable pseudo-divergence of the form (respectively) (111)–(119), and in the course of (Enc3) minimize this over $\theta \in \Theta$, we end up at the corresponding min $-decD_{\lambda}$-estimator (109). For the special case (118) (i.e. $\alpha = 1$) this leads to the omnipresent, celebrated *maximum-likelihood-estimator* (MLE) which is known to be efficient but not robust. The particular choice (115) for $\alpha > 1$ gives the density-power divergence estimator DPDE of Basu et al. [10], where $\alpha = 2$ amounts to the (squared) L_2-estimator which is robust but not efficient (see e.g. Hampel et al. [33]); accordingly, taking $\alpha \in]1, 2[$ builds a smooth bridge between the robustness and efficiency. The reversed version of the DPDE can be analogously imbedded in our context, by employing our new approach with $\phi(t) := \widetilde{\widetilde{\phi}}_{\alpha}(t)$ (cf. (79)).

4.6 Minimum Divergences - Generalized Subdivergence Method

One can flexibilize some of the methods of the previous Sect. 4.5, by employing an additional (a.s.) strictly positive density function $\mathbb{M}$ to define a pseudo-divergence $D_{\mathbb{M},\lambda} : \widetilde{\mathscr{P}}^{\lambda} \otimes \mathscr{P}^{\lambda} \to \mathbb{R}$ of the form $D_{\mathbb{M},\lambda}(\mathbb{P}, \mathbb{Q}) =$ $\int_{\mathscr{X}} \psi^{dec}\big(\frac{\mathbb{p}(x)}{\mathbb{m}(x)}, \frac{\mathbb{q}(x)}{\mathbb{m}(x)}\big) \cdot \mathbb{m}(x) \, \mathrm{d}\lambda(x)$ for some (measurable) mapping $\psi^{dec} : [0, \infty[\times [0, \infty[\mapsto \mathbb{R}$ with representation

$$
\begin{aligned}
&\psi^{dec}(s,t) := \psi^{0}\Big(s + h_0(x,s)\cdot \mathbf{1}_{\{0\}}(t)\Big)\cdot \mathbf{1}_{]\overline{c}_0,\infty[}(s)\cdot \mathbf{1}_{]c_0,\infty[}(t)\\
&\quad +\psi^{1}\Big(t + h_1(x)\cdot \mathbf{1}_{\{0\}}(t)\Big)\cdot \mathbf{1}_{]c_1,\infty[}(t)\\
&\quad +\rho\Big(t + h_2(x)\cdot \mathbf{1}_{\{0\}}(t)\Big)\cdot s \quad \text{for all } (s,t)\in [0,\infty[\times[0,\infty[\backslash\{(0,0)\}\,, \quad (cf.(102))\\
&\psi^{dec}(0,0) := 0.
\end{aligned}
$$

It is straightforward to see that $D_{\mathbb{M},\lambda}(\cdot,\cdot)$ is a pointwise decomposable pseudo-divergence in the sense of Definition 3(b), and one gets for fixed $m > 0$

$$
\psi^{dec}_m(s,t) := m\cdot\psi^{dec}\Big(\tfrac{s}{m},\tfrac{t}{m}\Big) = m\cdot\psi^{0}\Big(\tfrac{s}{m}\Big) + m\cdot\psi^{1}\Big(\tfrac{t}{m}\Big) + \rho\Big(\tfrac{t}{m}\Big)\cdot s \geqslant 0
$$
$$
\text{for all } (s,t) \in]0,\infty[\times]0,\infty[\,, \qquad (120)
$$

$$
\begin{aligned}
&\psi^{dec}_m(s,t) = 0 \quad \text{if and only if} \quad s = t\,,\\
&\tfrac{s}{m} + h_0\Big(x,\tfrac{s}{m}\Big) \geqslant 0 \qquad \text{for all } s\in[0,\infty[\text{ and } \lambda\text{-almost all } x\in\mathscr{X}\,,\\
&\mathbb{R}\ni\psi^{dec}_m(s,0) = m\cdot\psi^{0}\Big(\tfrac{s}{m} + h_0\Big(x,\tfrac{s}{m}\Big)\Big)\cdot\breve{c}_0 + m\cdot\psi^{1}(h_1(x))\cdot\breve{c}_1 + \rho(h_2(x))\cdot s\\
&\hspace{20em}\text{for all } s > 0\,, (121)
\end{aligned}
$$

$$
\mathbb{R}\ni\psi^{dec}_m(0,t) = m\cdot\psi^{0}(0)\cdot\breve{\overline{c}}_0 + m\cdot\psi^{1}\Big(\tfrac{t}{m}\Big) \quad \text{for all } t > 0\,. \qquad (122)
$$

For each class-family member $\mathbb{M} := \vec{\mathbb{Q}}_\tau$ with arbitrarily fixed $\tau\in\Theta$, we can apply Definition 4 to $D_\lambda(\cdot,\cdot) := D_{\vec{\mathbb{Q}}_\tau,\lambda}(\cdot,\cdot)$, and arrive at the corresponding $\min -decD_{\vec{\mathbb{Q}}_\tau,\lambda}$-estimators

$$
\widehat{\theta}_{N,decD_{\vec{\mathbb{Q}}_\tau,\lambda}}(\omega) := T_{D_{\vec{\mathbb{Q}}_\tau,\lambda}}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\big) \quad \text{for } \vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \in \mathscr{P}^{\lambda\perp}_{emp} \qquad (123)
$$

of the true unknown parameter θ_0. Hence, analogously to the derivation of (111), we obtain from (102), (121), (122) for each $\tau\in\Theta$

$$
\begin{aligned}
&D_{\vec{\mathbb{Q}}_\tau,\lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\tilde{\vec{\mathbb{Q}}}_\theta\Big) = \textstyle\int_{\mathscr{X}} \psi^{dec}\Big(\frac{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{q}}_\tau(x)},\frac{\tilde{\vec{\mathbb{q}}}_\theta(x)}{\vec{\mathbb{q}}_\tau(x)}\Big)\cdot\vec{\mathbb{q}}_\tau(x)\,\mathrm{d}\lambda(x)\\
&= \textstyle\int_{\mathscr{X}} \psi^{1}\Big(\frac{\vec{\mathbb{q}}_\theta(x)}{\vec{\mathbb{q}}_\tau(x)}\Big)\cdot\vec{\mathbb{q}}_\tau(x)\,\mathrm{d}\lambda_2(x) + \sum_{x\in\mathscr{X}}\Big[\vec{\mathbb{q}}_\tau(x)\cdot\psi^{0}\Big(\frac{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{q}}_\tau(x)} + h_0\Big(x,\frac{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{q}}_\tau(x)}\Big)\Big)\cdot\breve{c}_0\\
&+\vec{\mathbb{q}}_\tau(x)\cdot\psi^{1}(h_1(x))\cdot\breve{c}_1 + \rho(h_2(x))\cdot\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\Big]\cdot\mathbf{1}_{\mathscr{R}\,(Y_1(\omega),\ldots,Y_N(\omega))}(x) + \psi^{0}(0)\cdot\breve{\overline{c}}_0\\
&= \textstyle\int_{\mathscr{X}} \psi^{1}\Big(\frac{\vec{\mathbb{q}}_\theta(x)}{\vec{\mathbb{q}}_\tau(x)}\Big)\cdot\vec{\mathbb{q}}_\tau(x)\,\mathrm{d}\lambda_2(x) + \frac{1}{N}\sum_{i=1}^{N}\rho(h_2(Y_i(\omega))) + \psi^{0}(0)\cdot\breve{\overline{c}}_0\\
&+\frac{1}{N}\sum_{i=1}^{N}\frac{\psi^{0}\Big(\frac{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}{\vec{\mathbb{q}}_\tau(Y_i(\omega))} + h_0\Big(Y_i(\omega),\frac{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}{\vec{\mathbb{q}}_\tau(Y_i(\omega))}\Big)\Big)\cdot\breve{c}_0 + \psi^{1}(h_1(Y_i(\omega)))\cdot\breve{c}_1}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\cdot\vec{\mathbb{q}}_\tau(Y_i(\omega))\,. \qquad (124)
\end{aligned}
$$

Just as in the derivation of (112) respectively (113), reasonable choices for the "boundary-functions" in (124) are $h_0(x,s) \equiv 0$, $h_1(x) \equiv 0$, $h_2(x) = \frac{\overrightarrow{\mathbb{q}}_{\theta}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}$, respectively $h_0(x,s) \equiv 0$, $h_1(x) \equiv \frac{\overrightarrow{\mathbb{q}}_{\theta}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}$, $h_2(x) = \frac{\overrightarrow{\mathbb{q}}_{\theta}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}$. As for example, consider for all $\theta_0, \theta, \tau \in \Theta$ the scaled Bregman divergences in the sense of Stummer [81], Stummer and Vajda [84] (cf. Remark (2)(b)), for which we get from (36) with $r(x) \equiv 1$

$$
\begin{aligned}
& 0 \leqslant D^{c}_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\mathbb{1}\cdot\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta}) \\
& := \int_{\mathscr{X}} \Big[\phi\big(\tfrac{\overrightarrow{\mathbb{q}}_{\theta_0}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}\big) - \phi\big(\tfrac{\overrightarrow{\mathbb{q}}_{\theta}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}\big) - \phi'_{+,c}\big(\tfrac{\overrightarrow{\mathbb{q}}_{\theta}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}\big) \cdot \big(\tfrac{\overrightarrow{\mathbb{q}}_{\theta_0}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)} - \tfrac{\overrightarrow{\mathbb{q}}_{\theta}(x)}{\overrightarrow{\mathbb{q}}_{\tau}(x)}\big) \Big] \cdot \overrightarrow{\mathbb{q}}_{\tau}(x)\, \mathrm{d}\lambda_2(x)\,, \\
& =: D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})\,,
\end{aligned}
\tag{125}
$$

from which – together with (120) – one can identify immediately the pointwise decomposability with $\psi^0(s) := \psi^0_{\phi}(s) := \phi(s)$, $\psi^1(t) := \psi^1_{\phi}(t) := t \cdot \phi'_{+,c}(t) - \phi(t)$, $\rho(t) := \rho_{\phi}(t) := -\phi'_{+,c}(t)$; by plugging this into (124), one obtains the objective $D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}\Big(\overset{\rightharpoonup}{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_{\theta}\Big)$, which in the course of (Enc3) should be – for fixed $\tau \in \Theta$ – minimized over $\theta \in \Theta$ in order to obtain the corresponding τ-individual" $\min -decD_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda}$-estimator $\widehat{\theta}_{N,\tau}(\omega) := \operatorname{arginf}_{\theta\in\Theta}\ D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda}\Big(\overset{\rightharpoonup}{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_{\theta}\Big)$. Recall that this choice can be motivated by $0 = \min_{\theta\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})$ and $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})$. Furthermore, one gets even $0 = \min_{\theta\in\Theta} \min_{\tau\in\Theta} D_{\phi,\overrightarrow{\mathbb{Q}}_{\tau},\lambda_2}$ $(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})$, $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} \min_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})$, and in case of $\max_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta}) < \infty$ also $0 = \min_{\theta\in\Theta} \max_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})$, $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} \max_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda_2}(\overrightarrow{\mathbb{Q}}_{\theta_0},\overrightarrow{\mathbb{Q}}_{\theta})$. This suggests the alternative, "τ-uniform" estimators $\widehat{\theta}_N(\omega) := \operatorname{argmin}_{\theta\in\Theta}\ \min_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda}\Big(\overset{\rightharpoonup}{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_{\theta}\Big)$, respectively $\widehat{\theta}_N(\omega) := \operatorname{argmin}_{\theta\in\Theta}\ \max_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda}\Big(\overset{\rightharpoonup}{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_{\theta}\Big)$. As a side remark, let us mention that in general, (say) $\min_{\tau\in\Theta} D_{\phi,\overset{\rightharpoonup}{\mathbb{Q}}_{\tau},\lambda}\Big(\overset{\rightharpoonup}{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_{\theta}\Big)$ is not necessarily decomposable anymore, and therefore the standard theory of M-estimators is not applicable to this class of estimators.

With our approach, we can generate numerous further estimators of the true unknown parameter θ_0, by permuting the positions – but not the roles (!) – of the parameters (θ_0, θ, τ) in the (pseudo-)divergences of the above investigations. For the sake of brevity, we only sketch two further cases; the full variety will appear elsewhere. To start with, consider the adaptively scaled and aggregated divergence

$$
\begin{aligned}
0 \leqslant\ & D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\theta) := D^{c}_{\phi,\vec{\mathbb{Q}}_{\theta_0}^{2}/\vec{\mathbb{Q}}_\tau,\,\vec{\mathbb{Q}}_{\theta}^{2}/\vec{\mathbb{Q}}_\tau,\,\mathbb{1}\cdot\vec{\mathbb{Q}}_{\theta_0},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\theta) \\
&:= \int_{\mathscr{X}} \Bigg[\phi\Bigg(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\frac{\vec{\mathbb{q}}_{\theta_0}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\Bigg) - \phi\Bigg(\frac{\vec{\mathbb{q}}_{\theta}(x)}{\frac{\vec{\mathbb{q}}_{\theta}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\Bigg) - \phi'_{+,c}\Bigg(\frac{\vec{\mathbb{q}}_{\theta}(x)}{\frac{\vec{\mathbb{q}}_{\theta}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\Bigg)\cdot\Bigg(\Bigg(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\frac{\vec{\mathbb{q}}_{\theta_0}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\Bigg) - \Bigg(\frac{\vec{\mathbb{q}}_{\theta}(x)}{\frac{\vec{\mathbb{q}}_{\theta}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\Bigg)\Bigg)\Bigg] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot\vec{\mathbb{q}}_\tau(x)\,\mathrm{d}\lambda_2(x) \\
&= \int_{\mathscr{X}} \Big[\phi\big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta_0}(x)}\big) - \phi\big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta}(x)}\big) - \phi'_{+,c}\big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta}(x)}\big)\cdot\big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta_0}(x)} - \tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta}(x)}\big)\Big]\cdot\vec{\mathbb{q}}_{\theta_0}(x)\,\mathrm{d}\lambda_2(x) \\
&=: \int_{\mathscr{X}} \Big[\psi^{0,rev}_{\vec{\mathbb{q}}_\tau(x)}(\vec{\mathbb{q}}_{\theta_0}(x)) + \psi^{1,rev}_{\vec{\mathbb{q}}_\tau(x)}(\vec{\mathbb{q}}_{\theta}(x)) + \rho^{rev}_{\vec{\mathbb{q}}_\tau(x)}(\vec{\mathbb{q}}_{\theta}(x))\cdot\vec{\mathbb{q}}_{\theta_0}(x)\Big]\,\mathrm{d}\lambda_2(x)
\end{aligned}
$$

(indeed, by Theorem 4 and (80) this is zero if and only if $\theta=\theta_0$). By means of the involved mappings $\psi^0(s) := \psi^{0,rev}_m(s) := s\cdot\phi(\frac{m}{s})$, $\psi^1(t) := \psi^{1,rev}_m(t) := -m\cdot\phi'_{+,c}(\frac{m}{t})$, $\rho(t) := \rho^{rev}_m(t) := \frac{m}{t}\cdot\phi'_{+,c}(\frac{m}{t}) - \phi(\frac{m}{t}) =: \phi^{\circledcirc}(\frac{m}{t})$ $(s,t,m>0)$, the properties (103), (104) are applicable and thus $D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda_2}(\cdot,\cdot)$ can be extended to a pointwise decomposable pseudo-divergence on $\widetilde{\mathscr{P}}^{\lambda}\otimes\mathscr{P}^{\lambda}$ by using (102) with appropriate functions h_0,h_1,h_2 and constants c_0,c_1,$\overline{c}_0$. Furthermore, by minimizing over $\theta\in\Theta$ the objective (111) with these choices $\psi^{0,rev}_m(\cdot)$, $\psi^{1,rev}_m(\cdot)$, $\rho^{rev}_m(\cdot)$, in the course of (Enc3) we end up at the corresponding min $-decD^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda}$ -estimator. In particular, the corresponding special case $h_0(x,s)\equiv 0$, $h_1(x)\equiv 1$, $h_2(x)=\vec{\mathbb{q}}_\theta(x)$ $(\neq \widetilde{\vec{\mathbb{q}}}_\theta(x))$ leads to the objective (cf. (112) but with $\psi^1(1)$ instead of $\psi^1(0)$)

$$
\begin{aligned}
& D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda_2}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \phi^*(0)\cdot\check{\overline{c}}_0 - \int_{\mathscr{X}} \vec{\mathbb{q}}_\tau(x)\cdot\phi'_{+,c}\Big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\Big)\,\mathrm{d}\lambda_2(x) \\
& +\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\phi^{\circledcirc}\Big(\tfrac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{\vec{\mathbb{q}}_\theta(Y_i(\omega))}\Big) \\
& +\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\Big[\phi\Big(\tfrac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\Big)\cdot\check{c}_0 - \tfrac{\vec{\mathbb{q}}_\tau(Y_i(\omega))\cdot\phi'_{+,c}(\vec{\mathbb{q}}_\tau(Y_i(\omega)))}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\cdot\check{c}_1\Big]
\end{aligned}
$$

to be minimized over θ. As a second possibility to permutate the positions of the parameters (θ_0,θ,τ), let us consider

$$
\begin{aligned}
& 0 \leqslant D^{c}_{\phi,\vec{\mathbb{Q}}_\theta,\vec{\mathbb{Q}}_\theta,\mathbb{1}\cdot\vec{\mathbb{Q}}_\theta,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\tau) \\
& := \int_{\mathscr{X}} \Big[\phi\big(\tfrac{\vec{\mathbb{q}}_{\theta_0}(x)}{\vec{\mathbb{q}}_\theta(x)}\big) - \phi\big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\big) - \phi'_{+,c}\big(\tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\big)\cdot\big(\tfrac{\vec{\mathbb{q}}_{\theta_0}(x)}{\vec{\mathbb{q}}_\theta(x)} - \tfrac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\big)\Big]\cdot\vec{\mathbb{q}}_\theta(x)\,\mathrm{d}\lambda_2(x)\,; \quad (126)
\end{aligned}
$$

this is a pointwise decomposable divergence between $\vec{\mathbb{Q}}_{\theta_0}$ and $\vec{\mathbb{Q}}_\tau$, but it is *not* a divergence – yet still a nonnegative and obviously *not* pointwise decomposable functional – between $\vec{\mathbb{Q}}_{\theta_0}$ and $\vec{\mathbb{Q}}_\theta$. Indeed, for $\theta=\theta_0\neq\tau$ one obtains $D^{c}_{\phi,\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_{\theta_0},\mathbb{1}\cdot\vec{\mathbb{Q}}_{\theta_0},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\tau) > 0$. Notice that from (126) one gets

$$\int_{\mathscr{X}} \phi\Big(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big) \cdot \vec{\mathbb{q}}_{\theta}(x)\, \mathrm{d}\lambda_2(x) \geqslant \int_{\mathscr{X}} \Big\{\Big[\phi\Big(\frac{\vec{\mathbb{q}}_{\tau}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big) - \phi'_{+,c}\Big(\frac{\vec{\mathbb{q}}_{\tau}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big) \cdot \frac{\vec{\mathbb{q}}_{\tau}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big] \cdot \vec{\mathbb{q}}_{\theta}(x)$$
$$+\phi'_{+,c}\Big(\frac{\vec{\mathbb{q}}_{\tau}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big) \cdot \vec{\mathbb{q}}_{\theta_0}(x)\Big\}\, \mathrm{d}\lambda_2(x) =: \mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})\,, \tag{127}$$

provided that the integral on the right-hand side exists and is finite. If moreover $\phi(1) = 0$, then by (54) the inequality (127) rewrites as

$$D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) := D^c_{\phi,\vec{\mathbb{Q}}_{\theta},\vec{\mathbb{Q}}_{\theta},\mathbb{1}\cdot\vec{\mathbb{Q}}_{\theta},\lambda}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) \geqslant D^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) \tag{128}$$

with (for fixed θ) equality if and only if $\theta_0 = \tau$; this implies that

$$D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) = \max_{\tau\in\Theta} \mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) \tag{129}$$
$$= \max_{\tau\in\Theta} \int_{\mathscr{X}} \Big[\psi^{1,sub}_{\vec{\mathbb{q}}_{\tau}(x)}(\vec{\mathbb{q}}_{\theta}(x)) + \rho^{sub}_{\vec{\mathbb{q}}_{\tau}(x)}(\vec{\mathbb{q}}_{\theta}(x)) \cdot \vec{\mathbb{q}}_{\theta_0}(x)\Big]\, \mathrm{d}\lambda_2(x) \tag{130}$$

with $\psi^0(s) := \psi^{0,sub}_m(s) \equiv 0, \psi^1(t) := \psi^{1,sub}_m(t) := t \cdot \phi(\frac{m}{t}) - m \cdot \phi'_{+,c}(\frac{m}{t}), \rho(t) := \rho^{sub}_m(t) := \phi'_{+,c}(\frac{m}{t})\ \ (s, t, m > 0)$. In other words, this means that the Csiszar-Ali-Silvey divergence CASD $D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})$ can be represented as the τ-maximum over – not necessarily nonnegative – pointwise decomposable (in the sense of (103), (104)) functionals $\mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})$ between $\vec{\mathbb{Q}}_{\theta_0}$ and $\vec{\mathbb{Q}}_{\theta}$. Furthermore, from Theorem 5 and (130) we arrive at

$$0 = \min_{\theta\in\Theta} D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) = \min_{\theta\in\Theta}\ \max_{\tau\in\Theta} \mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})$$
$$= \min_{\theta\in\Theta}\ \max_{\tau\in\Theta} \int_{\mathscr{X}} \Big[\psi^{1,sub}_{\vec{\mathbb{q}}_{\tau}(x)}(\vec{\mathbb{q}}_{\theta}(x)) + \rho^{sub}_{\vec{\mathbb{q}}_{\tau}(x)}(\vec{\mathbb{q}}_{\theta}(x)) \cdot \vec{\mathbb{q}}_{\theta_0}(x)\Big]\, \mathrm{d}\lambda_2(x)\,,$$
$$\theta_0 = \mathrm{argmin}_{\theta\in\Theta}\ \max_{\tau\in\Theta} \mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}). \tag{131}$$

Accordingly, in analogy to the spirit of (81), (82), (86), respectively Definition 4 and (110), in order to achieve an estimator of the true unknown parameter θ_0 we first extend the "pure parametric case" $\mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda_2} : \mathscr{P}^{\lambda}_{\Theta} \otimes \mathscr{P}^{\lambda}_{\Theta} \mapsto \mathbb{R}$ to a singularity-covering functional $\mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda} : \mathscr{P}^{\lambda}_{\Theta,emp} \otimes \mathscr{P}^{\lambda}_{\Theta} \mapsto \mathbb{R}$, although it is not a pseudo-divergence anymore; indeed, by employing the reduced form of (102) we take

$$\mathscr{D}^c_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda}(\vec{\mathbb{P}}, \vec{\mathbb{Q}}) := \int_{\mathscr{X}} \Big[\psi^{1,sub}_{\vec{\mathbb{q}}_{\tau}(x)}\Big(\vec{\mathbb{q}}(x) + h_1(x) \cdot \mathbf{1}_{\{0\}}(\vec{\mathbb{q}}(x))\Big) \cdot \mathbf{1}_{]c_1,\infty[}(\vec{\mathbb{q}}(x))$$
$$+\rho^{sub}_{\vec{\mathbb{q}}_{\tau}(x)}\Big(\vec{\mathbb{q}}(x) + h_2(x) \cdot \mathbf{1}_{\{0\}}(\vec{\mathbb{q}}(x))\Big) \cdot \vec{\mathbb{p}}(x)\Big]\, \mathrm{d}\lambda(x) \quad \text{for all } \vec{\mathbb{P}} \in \mathscr{P}^{\lambda}_{\Theta,emp}, \vec{\mathbb{Q}} \in \mathscr{P}^{\lambda}_{\Theta}\,. \tag{132}$$

Hence, analogously to the derivation of (111), we obtain from (132)

$$\begin{aligned}
&\sup_{\tau\in\Theta}\ \mathscr{D}^{c}_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_{\theta}\Big) \;=\; \sup_{\tau\in\Theta}\ \int_{\mathscr{X}}\ \psi^{1,sub}_{\vec{\mathbb{q}}_{\tau}(x)}\Big(\vec{\mathbb{q}}_{\theta}(x)\Big)\,\mathrm{d}\lambda_2(x)\\
&+\sum_{x\in\mathscr{X}}\Big[\psi^{1,sub}_{\vec{\mathbb{q}}_{\tau}(x)}\Big(h_1(x)\Big)\cdot\check{c}_1+\rho^{sub}_{\vec{\mathbb{q}}_{\tau}(x)}(h_2(x))\cdot\vec{\mathbb{p}}_{N}^{\overline{emp}(\omega)}(x)\Big]\cdot\mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)\\
&=\sup_{\tau\in\Theta}\ \int_{\mathscr{X}}\Big[\vec{\mathbb{q}}_{\theta}(x)\cdot\phi\Big(\frac{\vec{\mathbb{q}}_{\tau}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big)-\vec{\mathbb{q}}_{\tau}(x)\cdot\phi'_{+,c}\Big(\frac{\vec{\mathbb{q}}_{\tau}(x)}{\vec{\mathbb{q}}_{\theta}(x)}\Big)\Big]\mathrm{d}\lambda_2(x) \qquad (133)\\
&+\frac{1}{N}\sum_{i=1}^{N}\phi'_{+,c}\Big(\frac{\vec{\mathbb{q}}_{\tau}(Y_i(\omega))}{h_2(Y_i(\omega))}\Big)\\
&+\frac{1}{N}\sum_{i=1}^{N}\frac{h_1(Y_i(\omega))\cdot\phi\Big(\frac{\vec{\mathbb{q}}_{\tau}(Y_i(\omega))}{h_1(Y_i(\omega))}\Big)-\vec{\mathbb{q}}_{\tau}(Y_i(\omega))\cdot\phi'_{+,c}\Big(\frac{\vec{\mathbb{q}}_{\tau}(Y_i(\omega))}{h_1(Y_i(\omega))}\Big)}{\vec{\mathbb{p}}_{N}^{emp(\omega)}(Y_i(\omega))}\cdot\check{c}_1 \qquad (134)
\end{aligned}$$

to be minimized over $\theta\in\Theta$. In the view of (131), we can estimate (respectively learn) the true unknown parameter θ_0 by the estimator

$$\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)\;:=\;\operatorname{arginf}_{\theta\in\Theta}\ \sup_{\tau\in\Theta}\ \mathscr{D}^{c}_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_{\theta}\Big)\ \text{ for } \vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)}\in\mathscr{P}^{\lambda\perp}_{emp}\,, \qquad (135)$$

which under appropriate technical assumptions (integrability, etc.) exists, is finite, unique, and Fisher consistent; moreover, this method can be straightforwardly extended to non-parametric setups. Similarly to the derivation of (112) respectively (113), reasonable choices for the "boundary-functions" in (134) are $h_2(x):=\vec{\mathbb{q}}_{\theta}(x)$ together with $h_1(x)\equiv 1$ respectively $h_1(x):=\vec{\mathbb{q}}_{\theta}(x)$ (where the nominator in the last sum becomes $-\vec{\mathbb{q}}_{\tau}(Y_i(\omega))\cdot\phi'_{+,c}(1))$). In the special case with $c_1=0=\check{c}_1$ – where the choice of $h_1(\cdot)$ is irrelevant – and $h_2(x):=\vec{\mathbb{q}}_{\theta}(x)$, the estimator $\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)$ was first proposed independently by Liese and Vajda [42] under the name *modified ϕ-divergence estimator* and Broniatowski and Keziou [16, 17] under the name *minimum dual ϕ-divergence estimator*; furthermore, within this special-case setup, Broniatowski and Keziou [17] also introduced for each fixed $\theta\in\Theta$ the related, so-called *dual ϕ-divergence estimator* $\widehat{\theta}_{N,\theta,\mathscr{D}_{\phi,\lambda}}(\omega):=\operatorname{argsup}_{\tau\in\Theta}\ \mathscr{D}^{c}_{\phi,\vec{\mathbb{Q}}_{\tau},\lambda}\Big(\vec{\mathbb{P}}_{N}^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_{\theta}\Big)$. The latter four references also work within a nonparametric framework. Let us also mention that by (128) and (129), $\widehat{\theta}_{N,\mathscr{D}_{\phi,\lambda}}(\omega)$ can be interpreted as *maximum sub-ϕ-divergence estimator*, whereas $\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)$ can be viewed as *minimum super-ϕ-divergence estimator* (cf. Vajda [90], Broniatowski and Vajda [18] for the probability-measure-theoretic context of footnote 15).

Remark 6 Making use of the escort parameter τ proves to be useful in statistical inference under the model; its use under misspecification has been considered in Toma and Broniatowski [86], Al Mohamad [5], for Csiszar-Ali-Silvey divergences.

As a final example, consider $c_1=0$, $h_2(x):=\vec{\mathbb{q}}_{\theta}(x)$, and $\phi(t):=t\log t+1-t$, for which we can deduce

$$\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)=\widehat{\theta}_{N,\theta,\mathscr{D}_{\phi,\lambda}}(\omega)=\operatorname{argsup}_{\xi\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\log\Big(\vec{\mathbb{q}}_{\xi}(Y_i(\omega))\Big)$$

for all $\theta \in \Theta$, i.e. in this case all maximum sub-ϕ-divergence estimators and the minimum super-ϕ-divergence estimator exceptionally coincide, and give the celebrated maximum-likelihood estimator.

5 Conclusions

Motivated by fields of applications from statistics, machine learning, artificial intelligence and information geometry, we presented for a wide audience a new unifying framework of divergences between functions. Within this, we illuminated several important subcases – such as scaled Bregman divergences and Csiszar-Ali-Silvey ϕ-divergences – as well as involved subtleties and pitfalls. For the often desired task of finding the "continuous" model with best divergence-proximity to the observed "discrete" data, we summarized existing and also derived new approaches. As far as potential future studies is concerned, the kind of universal nature of our introduced toolkit suggests quite a lot of possibilities for further adjacent developments and concrete applications.

Acknowledgements We are grateful to three anonymous referees for their very useful suggestions and comments. W. Stummer wants to thank very much the Sorbonne Universite Pierre et Marie Curie Paris for its partial financial support.

Appendix: Proofs

Proof of Theorem 4. Assertion (1) and the "if-part" of (2) follow immediately from Theorem 1 which uses less restrictive assumptions. In order to show the "only-if" part of (2) (and the "if-part" of (2) in an alternative way), one can use the straightforwardly provable fact that the Assumption 2 implies

$$\overline{\mathbb{W}_3 \cdot \psi_{\phi,c}}(x, s, t) \;=\; 0 \qquad \text{if and only if} \qquad s \;=\; t \tag{136}$$

for all $s \in \mathscr{R}\big(\frac{P}{M_1}\big)$, all $t \in \mathscr{R}\big(\frac{Q}{M_2}\big)$ and λ-a.a. $x \in \mathscr{X}$. To proceed, assume that $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q) = 0$, which by the non-negativity of $\overline{\mathbb{W}_3 \cdot \psi_{\phi,c}}(\cdot, \cdot)$ implies that $\overline{\mathbb{W}_3 \cdot \psi_{\phi,c}}\big(\frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big) = 0$ for λ-a.a. $x \in \mathscr{X}$. From this and the "only-if" part of (136), we obtain the identity $\frac{p(x)}{m_1(x)} = \frac{q(x)}{m_2(x)}$ for λ-a.a. $x \in \mathscr{X}$. □

Proof of Theorem 5. Consistently with Theorem 1 (and our adaptions) the "if-part" follows from (51). By our above investigations on the adaptions of the Assumptions 2 to the current context, it remains to investigate the "only-if" part (2) for the following four cases (recall that ϕ is strictly convex at $t = 1$):
(ia) ϕ is differentiable at $t = 1$ (hence, c is obsolete and $\phi'_{+,c}(1)$ collapses to $\phi'(1)$) and the function ϕ is affine linear on $[1, s]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[a, 1]$;

(ib) ϕ is differentiable at $t = 1$, and the function ϕ is affine linear on $[s, 1]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[1, b]$;
(ii) ϕ is not differentiable at $t = 1$, $c = 1$, and the function ϕ is affine linear on $[1, s]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[a, 1]$;
(iii) ϕ is not differentiable at $t = 1$, $c = 0$, and the function ϕ is affine linear on $[s, 1]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[1, b]$.
It is easy to see from the strict convexity at 1 that for (ii) one has $\phi(0) + \phi'_{+,1}(1) - \phi(1) > 0$, whereas for (iii) one gets $\phi^*(0) - \phi'_{+,0}(1) > 0$; furthermore, for (ia) there holds $\phi(0) + \phi'(1) - \phi(1) > 0$ and for (ib) $\phi^*(0) - \phi'(1) > 0$. Let us first examine the situations (ia) respectively (ii) under the assumptive constraint $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P}, \mathbb{Q}) = 0$ with $c = 1$ respectively (in case of differentiability) obsolete c, for which we can deduce from (51)

$$\begin{aligned}
&0 = D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P}, \mathbb{Q}) \\
&\geqslant \int_{\mathscr{X}} \mathbb{r}(x) \cdot \big[\mathbb{q}(x) \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x) \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) - \mathbb{q}(x)\big)\big] \\
&\qquad\qquad\qquad\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{]\mathbb{p}(x),\infty[}\big(\mathbb{q}(x)\big)\, \mathrm{d}\lambda(x) \\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{]\mathbb{p}(x),\infty[}\big(\mathbb{q}(x)\big)\, \mathrm{d}\lambda(x) \; \geqslant 0,
\end{aligned}$$

and hence $\int_{\mathscr{X}} \mathbf{1}_{]\mathbb{p}(x),\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) \; = 0$. From this and (55) we obtain

$$0 \;=\; \int_{\mathscr{X}} \; \big(\mathbb{p}(x) \;-\; \mathbb{q}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) \;=\; \int_{\mathscr{X}} \; \big(\mathbb{p}(x) - \mathbb{q}(x)\big) \;\cdot\; \mathbf{1}_{]\mathbb{q}(x),\infty[}\big(\mathbb{p}(x)\big) \;\cdot\; \mathbb{r}(x)\, \mathrm{d}\lambda(x)$$

and therefore $\int_{\mathscr{X}} \mathbf{1}_{]\mathbb{q}(x),\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) \; = 0$. Since for λ-a.a. $x \in \mathscr{X}$ we have $\mathbb{r}(x) > 0$, we arrive at $\mathbb{p}(x) = \mathbb{q}(x)$ for λ-a.a. $x \in \mathscr{X}$. The remaining cases (ib) respectively (iii) can be treated analogously. □

References

1. Amari, S.-I.: Information Geometry and Its Applications. Springer, Japan (2016)
2. Amari, S.-I., Karakida, R., Oizumi, M.: Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. Info. Geo. (2018). https://doi.org/10.1007/s41884-018-0002-8
3. Amari, S.-I., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Oxford (2000)
4. Ali, M.S., Silvey, D.: A general class of coefficients of divergence of one distribution from another. J. R. Stat. Soc. **B–28**, 131–140 (1966)
5. Al Mohamad, D.: Towards a better understanding of the dual representation of phi divergences. Stat. Papers (2016). https://doi.org/10.1007/s00362-016-0812-5
6. Avlogiaris, G., Micheas, A., Zografos, K.: On local divergences between two probability measures. Metrika **79**, 303–333 (2016)
7. Avlogiaris, G., Micheas, A., Zografos, K.: On testing local hypotheses via local divergence. Stat. Methodol. **31**, 20–42 (2016)
8. Ay, N., Jost, J., Le, H.V., Schwachhöfer, L.: Information Geometry. Springer, Berlin (2017)
9. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**, 1705–1749 (2005)

10. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimizing a density power divergence. Biometrika **85**(3), 549–559 (1998)
11. Basu, A., Lindsay, B.G.: Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Ann. Inst. Stat. Math. **46**(4), 683–705 (1994)
12. Basu, A., Mandal, A., Martin, N., Pardo, L.: Robust tests for the equality of two normal means based on the density power divergence. Metrika **78**, 611–634 (2015)
13. Basu, A., Shioya, H., Park, C.: Statistical Inference: The Minimum Distance Approach. CRC Press, Boca Raton (2011)
14. Birkhoff, G.D: A set of postulates for plane geometry, based on scale and protractor. Ann. Math. **33**(2) 329–345 (1932)
15. Boissonnat, J.-D., Nielsen, F., Nock, R.: Bregman Voronoi diagrams. Discret. Comput. Geom. **44**(2), 281–307 (2010)
16. Broniatowski, M., Keziou, A.: Minimization of ϕ-divergences on sets of signed measures. Stud. Sci. Math. Hungar. **43**, 403–442 (2006)
17. Broniatowski, M., Keziou, A.: Parametric estimation and tests through divergences and the duality technique. J. Multiv. Anal. **100**(1), 16–36 (2009)
18. Broniatowski, M., Vajda, I.: Several applications of divergence criteria in continuous families. Kybernetika **48**(4), 600–636 (2012)
19. Broniatowski, M., Toma, A., Vajda, I.: Decomposable pseudodistances in statistical estimation. J. Stat. Plan. Inf. **142**, 2574–2585 (2012)
20. Buckland, M.K.: Information as thing. J. Am. Soc. Inf. Sci. **42**(5), 351–360 (1991)
21. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning and Games. Cambridge University Press, Cambridge (2006)
22. Chhogyal, K., Nayak, A., Sattar, A.: On the KL divergence of probability mixtures for belief contraction. In: Hölldobler,S., et al. (eds.) KI 2015: Advances in Artificial Intelligence. Lecture Notes in Artificial Intelligence, vol. 9324, pp. 249–255. Springer International Publishing (2015)
23. Cliff, O.M., Prokopenko, M., Fitch, R.: An information criterion for inferring coupling in distributed dynamical systems. Front. Robot. AI **3**(71). https://doi.org/10.3389/frobt.2016.00071 (2016)
24. Cliff, O.M., Prokopenko, M., Fitch, R.: Minimising the Kullback-Leibler divergence for model selection in distributed nonlinear systems. Entropy **20**(51). https://doi.org/10.3390/e20020051 (2018)
25. Collins, M., Schapire, R.E., Singer, Y.: Logistic regression, AdaBoost and Bregman distances. Mach. Learn. **48**, 253–285 (2002)
26. Cooper, V.N., Haddad, H.M., Shahriar, H.: Android malware detection using Kullback-Leibler divergence. Adv. Distrib. Comp. Art. Int. J., Special Issue 3(2) (2014)
27. Csiszar, I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. **A-8**, 85–108 (1963)
28. DasGupta, A.: Some results on the curse of dimensionality and sample size recommendations. Calcutta Stat. Assoc. Bull. **50**(3–4), 157–178 (2000)
29. De Groot, M.H.: Uncertainty, information and sequential experiments. Ann. Math. Stat. **33**, 404–419 (1962)
30. Ghosh, A., Basu, A.: Robust Bayes estimation using the density power divergence. Ann. Inst. Stat. Math. **68**, 413–437 (2016)
31. Ghosh, A., Basu, A.: Robust estimation in generalized linear models: the density power divergence approach. TEST **25**, 269–290 (2016)
32. Ghosh, A., Harris, I.R., Maji, A., Basu, A., Pardo, L.: A generalized divergence for statistical inference. Bernoulli **23**(4A), 2746–2783 (2017)
33. Hampel, F.R., Ronchetti, E.M., Rousseuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)
34. Karakida, R., Amari, S.-I.: Information geometry of Wasserstein divergence. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 119–126. Springer International (2017)

35. Kißlinger, A.-L., Stummer, W.: Some decision procedures based on scaled Bregman distance surfaces. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2013. Lecture Notes in Computer Science, vol. 8085, pp. 479–486. Springer, Berlin (2013)
36. Kißlinger, A.-L., Stummer, W.: New model search for nonlinear recursive models, regressions and autoregressions. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2015. Lecture Notes in Computer Science, vol. 9389, pp. 693–701. Springer International (2015)
37. Kißlinger, A.-L., Stummer, W.: Robust statistical engineering by means of scaled Bregman distances. In: Agostinelli, C., Basu, A., Filzmoser, P., Mukherjee, D. (eds.) Recent Advances in Robust Statistics - Theory and Applications, pp. 81–113. Springer, India (2016)
38. Kißlinger, A.-L., Stummer, W.: A new toolkit for robust distributional change detection. Appl. Stochastic Models Bus. Ind. **34**, 682–699 (2018)
39. Kuchibhotla, A.K., Basu, A.: A general setup for minimum disparity estimation. Stat. Prob. Lett. **96**, 68–74 (2015)
40. Liese, F., Miescke, K.J.: Statistical Decision Theory: Estimation, Testing, and Selection. Springer, New York (2008)
41. Liese, F., Vajda, I.: Convex Statistical Distances. Teubner, Leipzig (1987)
42. Liese, F., Vajda, I.: On divergences and informations in statistics and information theory. IEEE Trans. Inf. Theory **52**(10), 4394–4412 (2006)
43. Lin, N., He, X.: Robust and efficient estimation under data grouping. Biometrika **93**(1), 99–112 (2006)
44. Liu, M., Vemuri, B.C., Amari, S.-I., Nielsen, F.: Total Bregman divergence and its applications to shape retrieval. In: Proceedings of 23rd IEEE CVPR, pp. 3463–3468 (2010)
45. Liu, M., Vemuri, B.C., Amari, S.-I., Nielsen, F.: Shape retrieval using hierarchical total Bregman soft clustering. IEEE Trans. Pattern Anal. Mach. Intell. **34**(12), 2407–2419 (2012)
46. Lizier, J.T.: JIDT: an information-theoretic toolkit for studying the dynamcis of complex systems. Front. Robot. AI **1**(11). https://doi.org/10.3389/frobt.2014.00011 (2014)
47. Menendez, M., Morales, D., Pardo, L., Vajda, I.: Two approaches to grouping of data and related disparity statistics. Comm. Stat. - Theory Methods **27**(3), 609–633 (1998)
48. Menendez, M., Morales, D., Pardo, L., Vajda, I.: Minimum divergence estimators based on grouped data. Ann. Inst. Stat. Math. **53**(2), 277–288 (2001)
49. Menendez, M., Morales, D., Pardo, L., Vajda, I.: Minimum disparity estimators for discrete and continuous models. Appl. Math. **46**(6), 439–466 (2001)
50. Millmann, R.S., Parker, G.D.: Geometry - A Metric Approach With Models, 2nd edn. Springer, New York (1991)
51. Minka, T.: Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd., Cambridge, UK (2005)
52. Morales, D., Pardo, L., Vajda, I.: Digitalization of observations permits efficient estimation in continuous models. In: Lopez-Diaz, M., et al. (eds.) Soft Methodology and Random Information Systems, pp. 315–322. Springer, Berlin (2004)
53. Morales, D., Pardo, L., Vajda, I.: On efficient estimation in continuous models based on finitely quantized observations. Comm. Stat. - Theory Methods **35**(9), 1629–1653 (2006)
54. Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S.: Information geometry of U-boost and Bregman divergence. Neural Comput. **16**(7), 1437–1481 (2004)
55. Nielsen, F., Barbaresco, F. (eds.): Geometric Science of Information GSI 2013. Lecture Notes in Computer Science, vol. 8085. Springer, Berlin (2013)
56. Nielsen, F., Barbaresco, F. (eds.): Geometric Science of Information GSI 2015. Lecture Notes in Computer Science, vol. 9389. Springer International (2015)
57. Nielsen, F., Barbaresco, F. (eds.): Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589. Springer International (2017)
58. Nielsen, F., Bhatia, R. (eds.): Matrix Information Geometry. Springer, Berlin (2013)
59. Nielsen, F., Nock, R.: Bregman divergences from comparative convexity. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 639–647. Springer International (2017)

60. Nielsen, F., Sun, K., Marchand-Maillet, S.: On Hölder projective divergences. Entropy **19**, 122 (2017)
61. Nielsen, F., Sun, K., Marchand-Maillet,S.: K-means clustering with Hölder divergences. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 856–863. Springer International (2017)
62. Nock, R., Menon, A.K., Ong, C.S.: A scaled Bregman theorem with applications. Advances in Neural Information Processing Systems 29 (NIPS 2016), pp. 19–27 (2016)
63. Nock, R., Nielsen, F.: Bregman divergences and surrogates for learning. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 2048–2059 (2009)
64. Nock, R., Nielsen, F., Amari, S.-I.: On conformal divergences and their population minimizers. IEEE Trans. Inf. Theory **62**(1), 527–538 (2016)
65. Österreicher, F., Vajda, I.: Statistical information and discrimination. IEEE Trans. Inf. Theory **39**, 1036–1039 (1993)
66. Pal, S., Wong, T.-K.L.: The geometry of relative arbitrage. Math. Financ. Econ. **10**, 263–293 (2016)
67. Pal, S., Wong, T.-K.L.: Exponentially concave functions and a new information geometry. Ann. Probab. **46**(2), 1070–1113 (2018)
68. Pardo, L.: Statistical Inference Based on Divergence Measures. Chapman & Hall/CRC, Boca Raton (2006)
69. Park, C., Basu, A.: Minimum disparity estimation: asymptotic normality and breakdown point results. Bull. Inf. Kybern. **36**, 19–33 (2004)
70. Patra, S., Maji, A., Basu, A., Pardo, L.: The power divergence and the density power divergence families: the mathematical connection. Sankhya 75-B Part 1, 16–28 (2013)
71. Peyre, G., Cuturi M.: Computational Optimal Transport (2018). arXiv:1803.00567v1
72. Read, T.R.C., Cressie, N.A.C.: Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York (1988)
73. Reid, M.D., Williamson, R.C.: Information, divergence and risk for binary experiments. J. Mach. Learn. Res. **12**, 731–817 (2011)
74. Roensch, B., Stummer, W.: 3D insights to some divergences for robust statistics and machine learning. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 460–469. Springer International (2017)
75. Rüschendorf, L.: On the minimum discrimination information system. Stat. Decis. Suppl. Issue **1**, 263–283 (1984)
76. Scott, D.W.: Multivariate Density Estimation - Theory, Practice and Visualization, 2nd edn. Wiley, Hoboken (2015)
77. Scott, D.W., Wand, M.P.: Feasibility of multivariate density estimates. Biometrika **78**(1), 197–205 (1991)
78. Stummer, W.: On a statistical information measure of diffusion processes. Stat. Decis. **17**, 359–376 (1999)
79. Stummer, W.: On a statistical information measure for a generalized Samuelson-Black-Scholes model. Stat. Decis. **19**, 289–314 (2001)
80. Stummer, W.: Exponentials, Diffusions, Finance. Entropy and Information. Shaker, Aachen (2004)
81. Stummer, W.: Some Bregman distances between financial diffusion processes. Proc. Appl. Math. Mech. **7**(1), 1050503–1050504 (2007)
82. Stummer, W., Kißlinger, A-L.: Some new flexibilizations of Bregman divergences and their asymptotics. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 514–522. Springer International (2017)
83. Stummer, W., Vajda, I.: On divergences of finite measures and their applicability in statistics and information theory. Statistics **44**, 169–187 (2010)
84. Stummer, W., Vajda, I.: On Bregman distances and divergences of probability measures. IEEE Trans. Inf. Theory **58**(3), 1277–1288 (2012)
85. Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. Ann. Inst. Stat. Math. **64**, 1009–1044 (2012)

86. Toma, A., Broniatowski, M.: Dual divergence estimators and tests: robustness results. J. Multiv. Anal. **102**, 20–36 (2011)
87. Tsuda, K., Rätsch, G., Warmuth, M.: Matrix exponentiated gradient updates for on-line learning and Bregman projection. J. Mach. Learn. Res. **6**, 995–1018 (2005)
88. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, Berlin (1996)
89. Vajda, I.: Theory of Statistical Inference and Information. Kluwer, Dordrecht (1989)
90. Vajda, I.: Modifications of divergence criteria for applications in continuous families. Research Report No. 2230, Institute of Information Theory and Automation, Prague (2008)
91. Vemuri, B.C., Liu, M., Amari, S.-I., Nielsen, F.: Total Bregman divergence and its applications to DTI analysis. IEEE Trans. Med. Imag. **30**(2), 475–483 (2011)
92. Victoria-Feser, M.-P., Ronchetti, E.: Robust estimation for grouped data. J. Am. Stat. Assoc. **92**(437), 333–340 (1997)
93. Weller-Fahy, D.J., Borghetti, B.J., Sodemann, A.A.: A survey of distance and similarity measures used within network intrusion anomaly detection. IEEE Commun. Surv. Tutor. **17**(1), 70–91 (2015)
94. Wu, L., Hoi, S.C.H., Jin, R., Zhu, J., Yu, N.: Learning Bregman distance functions for semi-supervised clustering. IEEE Trans. Knowl. Data Engin. **24**(3), 478–491 (2012)
95. Zhang, J., Naudts, J.: Information geometry under monotone embedding, part I: divergence functions. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 205–214. Springer International (2017)
96. Zhang, J., Wang, X., Yao, L., Li, J., Shen, X.: Using Kullback-Leibler divergence to model opponents in poker. Computer Poker and Imperfect Information: Papers from the AAAI-14 Workshop (2014)

Information-Theoretic Matrix Inequalities and Diffusion Processes on Unimodular Lie Groups

Gregory S. Chirikjian

Abstract Unimodular Lie groups admit natural generalizations of many of the core concepts on which classical information-theoretic inequalities are built. Specifically, they have the properties of shift-invariant integration, an associative convolution operator, well-defined diffusion processes, and concepts of Entropy, Fisher information, Gaussian distribution, and Fourier transform. Equipped with these definitions, it is shown that many inequalities from classical information theory generalize to this setting. Moreover, viewing the Fourier transform for noncommutative unimodular Lie groups as a matrix-valued function, relationships between trace inequalities, diffusion processes, and convolution are examined.

Keywords Haar measure · Convolution · Group theory · Harmonic analysis · Diffusion processes · Inequalities

1 Introduction

This paper explores aspects of information-theoretic inequalities that naturally extend from $\mathbb{R}^n$ to an arbitrary unimodular Lie group. The exposition is mostly a condensed summary of material that can be found in [1, 2], but also presents some new inequalities. The motivations for investigating such inequalities are twofold: (1) physical information-processing agents (such as mobile robots or microscopic organisms) often have configuration spaces with Lie-group structure, and their localization in the world is therefore inextricably connected to both information theory and geometry; (2) Due to the identical form of the entropy functional in continuous information theory and in statistical mechanics, results from one field carry over to the other, and so it becomes possible to make statements about the statistical mechanical entropy of passive objects such as DNA and loops in proteins using the results of Lie-theoretic information theory.

G. S. Chirikjian (✉)
Johns Hopkins University, Baltimore, MD 21218, USA
e-mail: gchirik1@jhu.edu
URL: http://www.rpk.lcsr.jhu.edu

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_9

1.1 Mathematical Preliminaries

A unimodular Lie group, $(G, \circ)$, is one which possesses a bi-invariant integration measure. That is, it is possible to construct a measure μ and associated volume element $dg \doteq d\mu(g)$ around each $g \in G$ such that given any function $f : G \to \mathbb{C}$ whose measure

$$\mu(f) = \int_G f(g)\, dg$$

exists, the following invariance properties will hold

$$\int_G f(g)\, dg = \int_G f(g^{-1})\, dg = \int_G f(h \circ g)\, dg = \int_G f(g \circ h)\, dg \tag{1}$$

for arbitrary $h \in G$. Here, of course, g^{-1} is the inverse of $g \in G$, which is the unique element such that

$$g \circ g^{-1} = g^{-1} \circ g = e$$

with e being the identity element, which for every $g \in G$ satisfies

$$g \circ e = e \circ g = g\,.$$

This is the unique element of G with such properties.

The equalities in (1) are analogous to the properties of the Lebesgue integral

$$\int_{\mathbb{R}^n} f(\mathbf{x})\, d\mathbf{x} = \int_{\mathbb{R}^n} f(-\mathbf{x})\, d\mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{y} + \mathbf{x})\, d\mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{x} + \mathbf{y})\, d\mathbf{x}\,.$$

All compact Lie groups are unimodular, as are all nilpotent and semisimple Lie groups. When referring to Lie groups in this paper, the discussion is restricted to matrix Lie groups with elements that are square matrices, and group operation, $\circ$, being matrix multiplication and the identity element is the identity matrix in the case of a matrix Lie group. In this context, the set of $n \times n$ unitary matrices, $U(n)$, and all of its subgroups are compact Lie groups. An example of a noncompact semisimple Lie group is the much-studied $SL(2, \mathbb{R})$ consisting of all 2×2 matrices with real entries and unit determinant [3–6]. And an example of a nilpotent group is the Heisenberg group, $\mathcal{H}(3)$, consisting of matrices of the form

$$H(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & \alpha & \beta \\ 0 & 1 & \gamma \\ 0 & 0 & 1 \end{pmatrix}$$

where $\alpha, \beta, \gamma \in \mathbb{R}$. Probability distributions, harmonic analysis, and diffusion processes on this group have been studied in detail [7, 8].

As a nontrivial example of a noncompact unimodular Lie group that arises frequently in engineering applications, and which is neither semisimple nor nilpotent, consider the Special Euclidean group, $SE(n)$, which consists of elements of the form $g = (R, \mathbf{t}) \in SO(n) \times \mathbb{R}^n$ with the semi-direct product group law

$$(R_1, \mathbf{t}_1) \circ (R_2, \mathbf{t}_2) = (R_1 R_2, R_1 \mathbf{t}_2 + \mathbf{t}_1).$$

Here $R_i \in SO(n)$, the special orthogonal group consisting of $n \times n$ rotation matrices, and the resulting semi-direct product group is denoted as

$$SE(n) = \mathbb{R}^n \rtimes SO(n)\,.$$

Building on the classic works of Miller [9, 10] and Vilenkin [11–13], the author has published extensively on harmonic analysis and diffusion processes on this group in the context of applications in robotics and polymer science [14–16]. Detailed treatment of these topics can be found in [1, 2], and a recent concise summary can be found in [17]. In order to avoid repetition, examples in the present paper are instead illustrated with $\mathcal{H}(3)$ and $SO(3)$, though the general formulation is kept abstract and general, in the spirit of [18–25].

In the context of unimodular Lie groups, it then makes sense to consider probability density functions, i.e., $f : G \to \mathbb{R}$ with the properties

$$f(g) \geq 0 \;\text{ and }\; \int_G f(g)\, dg = 1.$$

Moreover, the concept of entropy of a pdf is simply

$$S(f) \doteq - \int_G f(g) \log f(g)\, dg\,, \tag{2}$$

and an entropy power is

$$N(f) \doteq \frac{1}{2\pi e} \exp\left(\frac{2}{\dim(G)} S(f)\right)\,.$$

For unimodular Lie groups, a well-defined concept of convolution of functions in $(L^1 \cap L^2)(G)$ exists:

$$(f_1 * f_2)(g) \doteq \int_G f_1(h) f_2(h^{-1} \circ g)\, dh\,. \tag{3}$$

This inherits the associative property from G:

$$((f_1 * f_2) * f_3)(g) = (f_1 * (f_2 * f_3))(g)\,.$$

Moreover, for broad classes of unimodular Lie groups, including all compact Lie groups and group extensions such as $SE(n)$, it is possible to define a concept of Fourier transform. This is based on the concept of an irreducible unitary representation (IUR) of G. An IUR is a unitary operator (which can be thought of as a square matrix of either finite or infinite dimension) with the properties

$$U(g_1 \circ g_2, \lambda) = U(g_1, \lambda)\, U(g_2, \lambda) \quad \text{and} \quad U(g^{-1}, \lambda) = U^*(g, \lambda)$$

where λ can be thought of as a frequency parameter and $*$ denotes the Hermitian conjugate. The space of all Λ values is called the unitary dual of G. In the case when G is Abelian, the unitary dual is also a group. In the case of compact Lie groups, Λ is a discrete space with countable elements. In the case of some noncompact unimodular Lie groups the space Λ has been fully characterized, and its description can be a quite complicated requiring a combination of continuous and discrete parameters.

Within this context the Fourier transform is defined as

$$\hat{f}(\lambda) = \int_G f(g)\, U(g^{-1}, \lambda)\, dg\,.$$

As with classical Fourier analysis, there is a convolution theorem

$$(\widehat{f_1 * f_2})(\lambda) = \hat{f}_2(\lambda)\, \hat{f}_1(\lambda)$$

and a reconstruction formula

$$f(g) = \int_\Lambda \operatorname{tr}\left[\hat{f}(\lambda)\, U(g, \lambda)\right] d\lambda\,.$$

Combining the above gives

$$(f_1 * f_2)(e) = \int_\Lambda \operatorname{tr}\left[\hat{f}_2(\lambda)\, \hat{f}_1(\lambda)\right] d\lambda = (f_2 * f_1)(e)\,.$$

Moreover, the Plancherel equality gives

$$\int_G f_1(g)\overline{f_2(g)}\, dg = \int_\Lambda \operatorname{tr}\left[\hat{f}_1(\lambda)\, \hat{f}_2^*(\lambda)\right] d\lambda\,.$$

When $f_1 = f_2 = f$, this becomes the familiar form of Parseval's equality

$$\int_G |f(g)|^2\, dg = \int_\Lambda \left\|\hat{f}(\lambda)\right\|_{HS}^2 d\lambda\,,$$

where $\|A\|_{HS}^2 = \operatorname{tr}(AA^*)$ is the familiar Hilbert-Schmidt norm.

In the context of Lie groups there are also natural generalizations of the concept of partial derivatives. Namely, if $X \in \mathcal{G}$ (the Lie algebra corresponding to the Lie group G), then a (left-invariant) directional derivative of $f(g)$ is computed as

$$(\tilde{X}f)(g) \doteq \frac{d}{dt} f\left(g \circ e^{tX}\right)\Big|_{t=0} .$$

If $\{E_i\}$ is a basis for $\mathcal{G}$, then $(\tilde{E}_i f)(g)$ can be thought of as partial derivatives, and the derivative in the direction $X = \sum_i x_i E_i$ can be written as

$$(\tilde{X}f)(g) = \sum_i x_i (\tilde{E}_i f)(g) .$$

Making a choice $\{E_i\}$ and defining an inner product by imposing orthonormality conditions $(E_i, E_j) = \delta_{ij}$ in effect fixes a metric for $\mathcal{G}$, which can be transported by left or right action to define a metric on G.

Operational properties of the Fourier transform include

$$\widehat{(\tilde{X}f)}(\lambda) = u(X, \lambda)\, \hat{f}(\lambda)$$

where

$$u(X, \lambda) \doteq \frac{d}{dt} U(e^{tX}, \lambda)\Big|_{t=0} .$$

This matrix function is linear in X, and so

$$u\left(\sum_i x_i E_i, \lambda\right) = \sum_i x_i u(E_i, \lambda) .$$

The concept of a Fisher information matrix with elements

$$F_{ij}(f) \doteq \int_G \frac{(\tilde{E}_i f)(g)(\tilde{E}_j f)(g)}{f(g)}\, dg$$

and a diffusion process on G can be described with an equation of the form[1]

$$\frac{\partial f}{\partial t} = -\sum_{i=1}^{dim(G)} h_i \tilde{E}_i f + \frac{1}{2} \sum_{i,j=1}^{dim(G)} D_{ij} \tilde{E}_i \tilde{E}_j f . \qquad (4)$$

[1] The diffusion coefficients are D_{ij} and the drift coefficients are h_i. When $h_i = 0$ for all $i \in \{1, \ldots, dim(G)\}$ the diffusion process is called driftless.

With all of this in mind, it becomes possible to explore generalizations of inequalities from classical information theory. Two major hindrances to trivially extending these inequalities to the noncommutative case are: (1) for general functions on G,

$$(f_1 * f_2)(g) \neq (f_2 * f_1)(g) ;$$

and, (2) unlike in the case of $\mathbb{R}^n$ where the Gaussian distribution is simultaneously (a) the maximal entropy distribution subject to covariance constraints, (b) is closed under convolution and conditioning, and (c) solves the Euclidean-space version of (4), these attributes cannot in general be simultaneously satisfied for more general Lie groups. For example, the Entropy-Power Inequality (EPI) is not even true for the circle group $\mathbb{R}/\mathbb{Z} \cong SO(2)$.

This paper summarizes what is already known, introduces some new results, and poses some questions for future exploration. For additional background on convolution, entropy, and Fourier analysis (in the Euclidean on Lie group cases), see [26–38].

1.2 Structure of the Paper

Section 2 explains how to compute the integration measure for unimodular Lie groups. Section 3 reviews the concept of functions of positive type on unimodular Lie groups, and the resulting properties of Fourier matrices for such functions. Section 4 explains why trace inequalities are significant in the harmonic analysis of diffusion processes on noncommutative unimodular Lie groups, and reviews some of the most well-known trace inequalities and various conjectures put forth in the literature. Section 5 reviews how Fisher information arises in quantifying entropy increases under diffusion and reviews a generalization the de Bruijn identity, which is shown to hold for unimodular Lie groups in general. This too involves trace inequalities. Section 6 reviews definitions of mean of covariance of probability densities on unimodular Lie groups, and how the propagate under convolution. Section 7 illustrates the theory with specific examples ($SO(3)$ as an example of compact Lie groups and the Heisenberg group $\mathcal{H}(3)$ as an example of a noncommutative noncompact unimodular Lie group.)

2 Explicit Computation of the Bi-invariant Integration Measure

This section explains how to compute the bi-invariant integration measure for a unimodular Lie group, summarizing the discussion in [1, 2].

To begin, an inner product $(\cdot, \cdot)$ between arbitrary elements of the Lie algebra, $Y = \sum_i y_i E_i$ and $Z = \sum_j z_j E_j$, can be defined such that

$$(Y, Z) \doteq \sum_{i=1}^{n} y_i z_i \qquad \text{where} \qquad (E_i, E_j) = \delta_{ij}. \tag{5}$$

The basis $\{E_i\}$ is then orthonormal with respect to this inner product. The definition of the inner product together with the constraint of orthonormality of a particular choice of basis $\{E_i\}$ in (5) defines a metric tensor for the Lie group.

Let $\mathbf{q} = [q_1, \ldots, q_n]^T$ be a column vector of local coordinates. Then $g(t) = \tilde{g}(\mathbf{q}(t))$ is a curve in G where $\tilde{g} : \mathbb{R}^n \to G$ is the local parametrization of the Lie group G. Henceforth the tilde will be dropped since it will be clear from the argument whether the function $g(t)$ or $g(\mathbf{q})$ is being referred to. The right-Jacobian matrix[2] for an n-dimensional Lie group parameterized with local coordinates $q_1, \ldots, q_n$ is the matrix $J_r(\mathbf{q})$ that relates rates of change $\dot{\mathbf{q}}$ to $g^{-1}\dot{g}$, and likewise for $J_l(\mathbf{q})$ and $\dot{g}g^{-1}$, where a dot denotes d/dt. Specifically,

$$\dot{g}g^{-1} = \sum_j \omega_j^l E_j \qquad \text{and} \qquad \boldsymbol{\omega}^l = J_l(\mathbf{q})\dot{\mathbf{q}}$$

and

$$g^{-1}\dot{g} = \sum_j \omega_j^r E_j \qquad \text{and} \qquad \boldsymbol{\omega}^r = J_r(\mathbf{q})\dot{\mathbf{q}}.$$

In other words,

$$(\dot{g}g^{-1}, E_k) = \left(\sum_j \omega_j^l E_j, E_k\right) = \sum_j \omega_j^l \left(E_j, E_k\right) = \sum_j \omega_j^l \delta_{jk} = \omega_k^l.$$

The scalars ω_k^l can be stacked in an array to form the column vector $\boldsymbol{\omega}^l = [\omega_1^l, \omega_2^l, \ldots, \omega_n^l]^T$. Analogous calculations follow for the "r" case. This whole process is abbreviated with the "$\vee$" operation as

$$\left(\dot{g}g^{-1}\right)^{\vee} = \boldsymbol{\omega}^l \qquad \text{and} \qquad \left(g^{-1}\dot{g}\right)^{\vee} = \boldsymbol{\omega}^r. \tag{6}$$

Given an orthogonal basis $E_1, \ldots, E_n$ for the Lie algebra, projecting the left and right tangent operators onto this basis yields elements of the right- and left-Jacobian matrices[3]:

[2] Here 'right' and 'left' respectively refer to differentiation appearing on the right or left side in calculations. As such a 'right' quantity denoted with a subscript r is left invariant, and a 'left' quantity denoted with a subscript l is right invariant.

[3] The 'l' and 'r' convention used here for Jacobians and for vector fields is opposite that used in the mathematics literature. The reason for the choice made here is to emphasize the location of the "the most informative part" of the expression. In Jacobians, this is the location of the partial derivatives. In vector fields this is where the components defining the field appear.

$$(J_r)_{ij} = \left(g^{-1}\frac{\partial g}{\partial q_j}, E_i\right) \quad \text{and} \quad (J_l)_{ij} = \left(\frac{\partial g}{\partial q_j}g^{-1}, E_i\right). \tag{7}$$

In terms of the $\vee$ operation this is written as

$$\left(g^{-1}\frac{\partial g}{\partial q_j}\right)^{\vee} = J_r(\mathbf{q})\,\mathbf{e}_j \quad \text{and} \quad \left(\frac{\partial g}{\partial q_j}g^{-1}\right)^{\vee} = J_l(\mathbf{q})\,\mathbf{e}_j.$$

As another abuse of notation, the distinction between $J(\mathbf{q})$ and $J(g(\mathbf{q}))$ can be blurred in both the left and right cases. Again, it is clear which is being referred to from the argument of these matrix-valued functions.

Note that $J_r(h \circ g) = J_r(g)$ and $J_l(g \circ h) = J_l(g)$. For unimodular Lie groups,

$$|\det(J_r)(\mathbf{q})| = |\det(J_l)(\mathbf{q})| \quad \text{and} \quad dg = |\det(J_{r,l})(\mathbf{q})|\, d\mathbf{q}\,. \tag{8}$$

This dg has the bi-invariance property, and is called the *Haar measure*. Examples of how this looks in different coordinates are given for $\mathcal{H}(3)$ and $SO(3)$ in Sect. 7. In the compact case, it is always possible to find a constant c to normalize as $d'g \doteq c \cdot dg$ such that $\int_G d'g = 1$.

3 Functions of Positive Type

In harmonic analysis, a function $\varphi : G \to \mathbb{C}$ is called a *function of positive type* if for every $c_i \in \mathbb{C}$ and every $g_i, g_j \in G$ and any $n \in \mathbb{Z}_{>0}$ the inequality

$$\sum_{i,j=1}^{n} c_i\, \overline{c_j}\, \varphi(g_i \circ g_j^{-1}) \geq 0\,.$$

In some texts, such functions are also called *positive definite*, whereas in others that term is used only when the inequality above excludes equality except when all values of c_i are zero. Here a function of positive type will be taken to be one for which the matrix $M = [m_{ij}]$ with entries $m_{ij} \doteq \varphi(g_i \circ g_j^{-1})$ is Hermitian positive semi-definite (which can be shown to be equivalent to the above expression), and a positive definite function is one for which M is positive definite.

Some well-known properties of functions of positive type include [19, 21, 23]:

$$\begin{aligned} \varphi(e) &= \overline{\varphi(e)} \geq 0 \\ |\varphi(g)| &\leq \varphi(e) \\ \varphi(g^{-1}) &= \overline{\varphi(g)}\,. \end{aligned}$$

Moreover, if φ_1 and φ_2 are two such functions, then so are $\overline{\varphi_i}$, $\varphi_1 \cdot \varphi_2$, as are linear combinations of the form $a_1\varphi_1 + a_2\varphi_2$ where $a_i \in \mathbb{R}_{>0}$.

Clearly, if φ is a function constructed as

$$\varphi(g;\lambda) \doteq \mathrm{tr}\left[A^* U(g,\lambda) A\right]$$

when A is positive definite, then

$$\sum_{i,j=1}^{n} c_i\,\overline{c_j}\,\varphi(g_i \circ g_j^{-1}) = \sum_{i,j=1}^{n} c_i\,\overline{c_j}\,\mathrm{tr}\left[A^* U(g_i \circ g_j^{-1},\lambda) A\right]$$

$$= \left\| A^* \sum_{i=1}^{n} U(g_i,\lambda) \right\|_{HS}^{2} \geq 0$$

because

$$U(g_i \circ g_j^{-1},\lambda) = U(g_i,\lambda)\,U^*(g_j,\lambda)\,.$$

And hence $\varphi(g;\lambda)$ is a function of positive type. Moreover, by the same reasoning, if $f(g)$ is any functions for which $\hat{f}(\lambda)$ is a Hermitian positive definite matrix, then $f(g)$ will be a positive definite function. And, according to Hewitt and Ross [21, 23] (p. 683 Lemma D.12), if A and B are both positive definite matrices, then so is their product. This has implications regarding the positivity of the convolution of positive functions on a group.

In particular, if $\rho_t(g) = \rho(g;t)$ is the solution to a driftless diffusion equation with Dirac-delta initial conditions, then the Fourier-space solution is written as

$$\hat{\rho}(\lambda;t) = \exp\left[\frac{1}{2}\sum_{i,j=1}^{dim(G)} D_{ij} u(E_i,\lambda) u(E_j,\lambda)\right],$$

which is Hermitian positive definite, and hence $\rho_t(g)$ is a real-valued positive definite function for each value of $t \in \mathbb{R}_{\geq 0}$. Moreover,[4]

$$\rho_t(g) = \rho_t(g^{-1})\,.$$

It is not difficult to show that given two symmetric functions, $\rho_1(g) \doteq \rho_{t_1}(g;D_1)$ and $\rho_2(g) \doteq \rho_{t_2}(g;D_2)$, that

$$(\rho_1 * \rho_2)(g) = (\rho_2 * \rho_1)(g^{-1})\,.$$

Though this does not imply that $(\rho_1 * \rho_2)(g)$ is symmetric, it is easy to show that $(\rho_1 * \rho_2 * \rho_1)(g)$ is symmetric.

Moreover, if $f : G \to \mathbb{R}_{\geq 0}$ is a pdf which is not symmetric, it is not difficult to show that

[4] Here the dependence on $D = [D_{ij}]$ has been suppressed, but really $\rho_t(g) = \rho_t(g;D)$.

$$f'(g) \doteq \frac{f(g) + f(g^{-1})}{2}$$

and

$$f''(g) \doteq \frac{f(g)f(g^{-1})}{(f * f)(e)}$$

are symmetric pdfs.

For any positive definite symmetric pdf, the Fourier transform is a positive definite Hermitian matrix because

$$\begin{aligned}\hat{\rho}(\lambda) &= \int_G \rho(g)\, U(g^{-1}, \lambda)\, dg = \int_G \rho(g^{-1})\, U(g^{-1}, \lambda)\, dg \\ &= \int_G \rho(g)\, U(g, \lambda)\, dg = \int_G \rho(g)\, U^*(g^{-1}, \lambda)\, dg = \hat{\rho}^*(\lambda)\,.\end{aligned}$$

From positive definiteness, it is possible to write

$$\hat{\rho}(\lambda) = \exp H(\lambda)$$

where H is Hermitian, though not necessarily positive definite.

Moreover, every special unitary matrix can be expressed as the exponential of a skew-Hermitian matrix, and even more than that, if $g = \exp X$, then the IUR matrix $U(g, \lambda)$ can be computed as

$$U(g, \lambda) \;=\; \exp Z(\log(g), \lambda)$$

where

$$Z(X, \lambda) = \sum_{i=1}^{dim(G)} x_i u(E_i, \lambda) = -Z^*(X, \lambda)\,.$$

In analogy with the way that the exponential map for a matrix Lie group is simply the matrix exponential defined by the Taylor series, here and throughout this work the logarithm is the matrix logarithm defined by its Taylor series.

In this light, the Fourier inversion formula has in it the evaluation of

$$\mathrm{tr}\left[\exp H \exp Z\right]\,,$$

and the evaluation of $(\rho_1 * \rho_2)(e)$ has in it

$$\mathrm{tr}\left[\exp H_1 \exp H_2\right]\,.$$

Also, in the evaluation of probability densities for diffusion processes with drift, it is desirable to find approximations of the form

$$\mathrm{tr}\left[\exp(H+Z)\right] \approx \mathrm{tr}\left[\exp H' \exp Z'\right] .$$

For these reasons, there are connections between harmonic analysis on unimodular Lie groups and trace inequalities.

4 Trace Inequalities

In the Fourier reconstruction formula for a diffusion process on a Lie group, the trace of the product of exponentials of two matrices is computed. It is therefore relevant to consider: (1) when can the product of two exponentials be simplified; (2) even when the product cannot be simplified, it would be useful to determine when the trace operation has the effect of simplifying the result. For example, if A and B are bandlimited matrices and $\mathrm{tr}(e^A e^B) \approx \mathrm{tr}(\exp(A+B))$ then computing the eigenvalues of $A+B$, exponentiating each eigenvalue, and summing potentially could be much faster than directly exponentiating the matrices, and then taking the trace of the product. The statements that follow therefore may have some relevance to the rapid evaluation of the Fourier inversion formula for diffusion processes on Lie groups. Other matrix inequalities that may be applicable to extend the current analysis include [39–52].

4.1 Generalized Golden–Thompson Inequalities

For $n \times n$ Hermitian matrices A and B, the *Golden–Thompson inequality* [53–55] is

$$\varphi(e^A e^B) \geq \varphi(e^{A+B}) \tag{9}$$

where φ is one of a large number of so-called spectral functions. For the case when $\varphi(\cdot) = \mathrm{tr}(\cdot)$ (which is the case of primary interest in this chapter) this was proven in [54], and generalized in [56].

The Thompson Conjecture [57]: If H and K are Hermitian matrices, there exist unitary matrices U and V dependent on H and K such that

$$e^{iH} e^{iK} = e^{i(UHU^* + VKV^*)}. \tag{10}$$

The So-Thompson Conjecture [58]: If H and K are Hermitian matrices, there exist unitary matrices U and V dependent on H and K such that

$$e^{H/2} e^{K} e^{H/2} = e^{UHU^* + VKV^*}. \tag{11}$$

Interestingly, such conjectures have been proven [59] using techniques associated with random walks on symmetric space of Lie groups [60], thereby bringing the problem back to the domain of interest in this chapter.

In [61], Cohen et al. prove that *spectral matrix functions*,[5] $\varphi : \mathbb{C}^{n\times n} \to \mathbb{C}$ (including the trace) satisfy the inequality

$$\varphi(e^{(A+A^*)/2}e^{(B+B^*)/2}) \geq |\varphi(e^{A+B})| \tag{12}$$

when the following condition holds:

$$\varphi([XX^*]^s) \geq |\varphi(X^{2s})| \qquad \forall\, X \in \mathbb{C}^{n\times n} \qquad \text{for} \qquad s = 1, 2, \ldots$$

An interesting corollary to (12) is that if A is skew-Hermitian, and $B \in \mathbb{C}^{n\times n}$, then [61]:

$$\varphi(e^{(B+B^*)/2}) \geq |\varphi(e^{A+B})|. \tag{13}$$

Bernstein proved the following general statement for $A \in \mathbb{C}^{n\times n}$ [62]:

$$\mathrm{tr}(e^{A^*}e^{A}) \leq \mathrm{tr}(e^{A^*+A}) \tag{14}$$

Inequalities involving functions of products of exponentials have a long history (see e.g., [63, 64]) and remain an area of active investigation. A few recent papers include [65–67].

4.2 *Matrix Inequalities from Systems Theory*

Another sort of matrix inequality that may be useful would be extensions of results that comes from systems theory. For example, it is known that if $A, B \in \mathbb{R}^{n\times n}$ and $B = B^T > 0$ then [68]

$$\lambda_n(\hat{A})\mathrm{tr}(B) \leq \mathrm{tr}(AB) \leq \lambda_1(\hat{A})\mathrm{tr}(B) \tag{15}$$

where $\hat{A} = (A + A^T)/2$ and the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$. This has been tightened by Fang et al. [69]:

$$\lambda_n(\hat{A})\mathrm{tr}(B) - \lambda_n(B)[n\lambda_n(\hat{A}) - \mathrm{tr}(A)] \leq \mathrm{tr}(AB) \leq \lambda_1(\hat{A})\mathrm{tr}(B) - \lambda_n(B)[n\lambda_1(\hat{A}) - \mathrm{tr}(A)] \tag{16}$$

Additional modifications have been made by Park [70].

Under the same conditions on A and B, Komaroff and Lasserre independently derived the inequality [71, 72]:

$$\sum_{i=1}^{n} \lambda_i(\hat{A})\lambda_{n-i+1}(B) \leq \mathrm{tr}(AB) \leq \sum_{i=1}^{n} \lambda_i(\hat{A})\lambda_i(B) \tag{17}$$

[5] These are functions that depend only the eigenvalues of a matrix, and are therefore invariant under similarity transformations.

and Lasserre tightened this with the result [73]:

$$f(-\epsilon) \leq \text{tr}(AB) \leq f(\epsilon) \qquad \forall\, \epsilon > 0 \tag{18}$$

where

$$f(\epsilon) \doteq \frac{1}{\epsilon}\sum_{i=1}^{n}[\lambda_i(B+\epsilon\hat{A})\lambda_i(B) - \text{tr}(B^2)].$$

Apparently, when $A \in \mathbb{C}^{n\times n}$ and $B = B^* \in \mathbb{C}^{n\times n}$ (17) holds with the substitutions $A^T \to A^*$ and $\text{tr}(AB) \to \text{Re}[\text{tr}(AB)]$ [74]. Generalized formulas for products of arbitrary real matrices have been made recently [74–76].

4.3 Classical Matrix Inequalities

In a sense, inequalities of the form in the previous section can be traced back to work done as early as the 1930s. Mirsky [77] attributes the following result for arbitrary $A, B \in \mathbb{C}^{n\times n}$ to a 1937 paper by John von Neumann:

$$|\text{tr}(AB)| \leq \sum_{i=1}^{n}\mu_i(A)\mu_i(B) \tag{19}$$

where $\mu_1(A) \geq \mu_2(A) \geq \cdots \geq \mu_n(A)$ are the singular values of A.

Hoffman and Wieland [78] states that for $n \times n$ normal matrices A and B (i.e., $A^*A = AA^*$ and $B^*B = BB^*$) permutations $\pi, \sigma \in \Pi_n$ can be found such that

$$\sum_{i=1}^{n}|\lambda_i(A) - \lambda_{\pi(i)}|^2 \leq \|A - B\|^2 \leq \sum_{i=1}^{n}|\lambda_i(A) - \lambda_{\sigma(i)}|^2 \tag{20}$$

where $\|A\|^2 = \text{tr}(AA^*)$ is the Frobenius norm. For generalizations see [79]. In particular, Richter [80] and Mirsky [81] have shown that if A and B are both $n \times n$ Hermitian matrices,

$$\sum_{i=1}^{n}\lambda_i(A)\lambda_{n+1-i}(B) \leq \text{tr}(AB) \leq \sum_{i=1}^{n}\lambda_i(A)\lambda_i(B) \tag{21}$$

and Marcus [82] showed that for normal matrices A and B, there exist permutations π and σ for which

$$\sum_{i=1}^{n}\lambda_i(A)\lambda_{\pi(i)}(B) \leq \text{Re}[\text{tr}(AB)] \leq \sum_{i=1}^{n}\lambda_i(A)\lambda_{\sigma(i)}(B) \tag{22}$$

4.4 The Arithmetic-Mean-Geometric-Mean (AM-GM) Inequality

The *arithmetic-geometric-mean inequality* states that the arithmetic mean of a set of positive real numbers is always less than the geometric mean of the same set of numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\lambda_i \geq \prod_{i=1}^{n}\lambda_i^{\frac{1}{n}}. \tag{23}$$

This fact can be used to derive many useful inequalities. For example, Steele [83] uses the AM-GM inequality to derive a reverse Cauchy-Schwarz inequality of the form

$$\left(\sum_{k=1}^{n}a_k^2\right)^{\frac{1}{2}}\left(\sum_{k=1}^{n}b_k^2\right)^{\frac{1}{2}} \leq \frac{m+M}{2\sqrt{mM}}\sum_{k=1}^{n}a_k b_k \tag{24}$$

where $\{a_k\}, \{b_k\} \subset \mathbb{R}_{>0}$ and $0 < m \leq a_k/b_k \leq M < \infty$ for all $k \in \{1, \ldots, n\}$.

It is no coincidence that the numbers in (23) are denoted as λ_i because when they are interpreted as the eigenvalues of a positive definite Hermitian matrix, $A = A^* > 0$, and so

$$\frac{1}{n}\text{tr}(A) \geq |A|^{\frac{1}{n}}. \tag{25}$$

This is useful for bounding the trace of the matrix exponential of a not-necessarily-positive-definite Hermitian matrix, $H = H^*$, since $A = \exp H = A^* > 0$ can be substituted into (25). The determinant-trace equality for the matrix exponential $\det(\exp(H)) = e^{\text{tr}(H)}$ then gives $|\exp H|^{\frac{1}{n}} = |e^{\text{tr}(H)}|^{\frac{1}{n}}$ and so

$$\frac{1}{n}\text{tr}(\exp H) \geq e^{\text{tr}(H)/n}. \tag{26}$$

Though (25) is more fundamental than (26), the latter is directly useful in studying properties of diffusion processes on Lie groups.

It is also interesting to note that (25) generalizes in several ways. For example, if

$$\mu_p(\lambda_1, \ldots, \lambda_n) \doteq \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i^p\right)^{\frac{1}{p}}$$

then the AM-GM inequality can be stated as

$$\lim_{p\to 0}\mu_p(\lambda_1, \ldots, \lambda_n) \leq \mu_1(\lambda_1, \ldots, \lambda_n)$$

and more generally,

$$\mu_p(\lambda_1, \ldots, \lambda_n) \leq \mu_q(\lambda_1, \ldots, \lambda_n) \qquad p < q. \tag{27}$$

For each fixed choice of $\lambda_1, \ldots, \lambda_n$, the function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by $f(p) = \mu_p(\lambda_1, \ldots, \lambda_n)$ is an increasing function.

When $p = -1$, the *harmonic mean*

$$\mu_{-1}(\lambda_1, \ldots, \lambda_n) = n \cdot \left[\sum_{i=1}^{n} \frac{1}{\lambda_i}\right]^{-1}$$

results, and (27) for $p = -1$ and $q = 0$ implies that

$$n \cdot \left[\sum_{i=1}^{n} \frac{1}{\lambda_i}\right]^{-1} \leq \left[\prod_{i=1}^{n} \lambda_i\right]^{\frac{1}{n}}.$$

The implication of this for positive definite Hermitian matrices is that

$$n \cdot \mathrm{tr}[A^{-1}] \leq [\det A]^{\frac{1}{n}} \iff \mathrm{tr} A \leq \frac{1}{n}[\det A]^{-\frac{1}{n}}. \tag{28}$$

One generalization of (23) is the *weighted AM-GM inequality*

$$\frac{1}{\alpha}\sum_{i=1}^{n} \alpha_i \lambda_i \geq \left(\prod_{i=1}^{n} \lambda_i^{\alpha_i}\right)^{\frac{1}{\alpha}} \quad \text{where} \quad \alpha = \sum_{i=1}^{n} \alpha_i \quad , \quad \alpha_i \in \mathbb{R}_{>0}. \tag{29}$$

Another generalization is *Ky Fan's inequality* [84]:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} \lambda_i}{\frac{1}{n}\sum_{i=1}^{n}(1-\lambda_i)} \geq \frac{\prod_{i=1}^{n} \lambda_i^{\frac{1}{n}}}{\prod_{i=1}^{n}(1-\lambda_i)^{\frac{1}{n}}} \tag{30}$$

which holds for $0 \leq \lambda_i \leq \frac{1}{2}$. If the numbers λ_i are viewed as the eigenvalues of a matrix, A, then Ky Fan's inequality can be written as

$$\frac{\mathrm{tr} A}{\mathrm{tr}(\mathbb{I} - A)} \geq \frac{|A|^{\frac{1}{n}}}{|\mathbb{I} - A|^{\frac{1}{n}}} \quad \text{or} \quad \mathrm{tr} A \geq \frac{n}{1 + |A|^{-\frac{1}{n}}|\mathbb{I} - A|^{\frac{1}{n}}}. \tag{31}$$

Of course this statement should then be restricted to those matrices that have real eigenvalues that obey $0 \leq \lambda_i(A) \leq \frac{1}{2}$.

4.5 Consequences for Harmonic Analysis and Diffusion Processes

If $\rho_{D^{(k)}}(g,t)$ denotes the solution to the driftless diffusion equation on G with diffusion coefficients $D_{ij}^{(k)}$, subject to initial conditions $\rho_{D^{(k)}}(g,0)=\delta(g)$, then from the Golden–Thompson inequality

$$(\rho_{D^{(1)}} * \rho_{D^{(2)}})(e;t) \ \geq \ \rho_{D^{(1)}+D^{(2)}}(e;t)\,. \tag{32}$$

Alternatively, using the fact that $B^{\frac{1}{2}}AB^{\frac{1}{2}}$ is Hermitian positive definite whenever A and B are, the trace of the product $\mathrm{tr}[AB]=\mathrm{tr}[B^{\frac{1}{2}}AB]$ can be bounded using (25), resulting in

$$\frac{1}{n}\mathrm{tr}(AB) \geq |A|^{\frac{1}{n}}|B|^{\frac{1}{n}}. \tag{33}$$

This can then be used to bound $(\rho_{D^{(1)}} * \rho_{D^{(2)}})(e;t)$ from below as well. But it can also be used in a different way. Given a diffusion with drift, the Fourier matrices will be of the form $\exp(H+Z)$ where $H=H^*$ and $Z=-Z^*$. Then the convolution of two diffusions with drifts being the negative of each other will be $\exp(H+Z)\exp(H-Z)$, which is Hermitian, and hence

$$\frac{1}{n}\mathrm{tr}(\exp(H+Z)\exp(H-Z)) \geq |\exp(H+Z)|^{\frac{1}{n}}|\exp(H-Z)|^{\frac{1}{n}}. \tag{34}$$

Then from the determinant-trace equality, we can simplify

$$|\exp(H+Z)| = e^{\mathrm{tr}(H+Z)} \quad \text{and} \quad |\exp(H-Z)| = e^{\mathrm{tr}(H-Z)}\,,$$

thereby giving that

$$\frac{1}{n}\mathrm{tr}(\exp(H+Z)\exp(H-Z)) \geq e^{\frac{2}{n}\mathrm{tr}(H)}. \tag{35}$$

These will be demonstrated in Sect. 7.

In the case when $G=\mathbb{R}^n$, covariances add under convolution and for a diffusion $\Sigma^{(k)}=D^{(k)}t$, and so

$$(\rho_{D^{(1)}} * \rho_{D^{(2)}})(\mathbf{0};t) \ = \ \frac{1}{(2\pi t)^{n/2}\left|D^{(1)}+D^{(2)}\right|^{\frac{1}{2}}} \ = \ \rho_{D^{(1)}+D^{(2)}}(\mathbf{0};t)\,.$$

This begs the question of how to define and propagate covariances on a unimodular Lie group, and what relationships may exist with Fisher information. Inequalities relating Fisher information and entropy are reviewed in Sect. 5, followed by definitions of covariance in Sect. 6 and the relationship be Fisher information and covariance.

5 Inequalities Involving Fisher Information and Diffusion Processes

This section connects trace inequalities with Fisher information and the rate of entropy increase under a diffusion process. The results presented here are an abridged version of those presented in [2].

5.1 Rate of Increase of Entropy Under Diffusion

The entropy of a pdf on a Lie group is defined in (2) If $f(g,t)$ is a pdf that satisfies a diffusion equation (regardless of the details of the initial conditions) then some interesting properties of $S_f(t)$ can be studied. In particular, if $\dot{S}_f = dS_f/dt$, then differentiating under the integral sign gives

$$\dot{S}_f = -\int_G \left\{ \frac{\partial f}{\partial t} \log f + \frac{\partial f}{\partial t} \right\} dg.$$

But from the properties of a diffusion equation,

$$\int_G \frac{\partial f}{\partial t}\, dg = \frac{d}{dt} \int_G f(g,t)\, dg = 0,$$

and so the second term in the above braces integrates to zero.

Substitution of

$$\frac{\partial f}{\partial t} = \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \tilde{E}_i^r \tilde{E}_j^r f - \sum_{k=1}^{n} h_k \tilde{E}_k^r f$$

into the integral for $\dot{S}_f$ gives

$$\begin{aligned}
\dot{S}_f &= -\int_G \left\{ \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \tilde{E}_i^r \tilde{E}_j^r f - \sum_{k=1}^{n} h_k \tilde{E}_k^r f \right\} \log f \, dg \\
&= -\frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \int_G (\tilde{E}_i^r \tilde{E}_j^r f) \log f \, dg - \sum_{k=1}^{n} h_k \int_G (\tilde{E}_k^r f) \log f \, dg \\
&= \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \int_G (\tilde{E}_j^r f)(\tilde{E}_i^r \log f) \, dg + \sum_{k=1}^{n} h_k \int_G f \, (\tilde{E}_k^r \log f) \, dg \\
&= \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \int_G \frac{1}{f} (\tilde{E}_j^r f)(\tilde{E}_i^r f) \, dg + \sum_{k=1}^{n} h_k \int_G \tilde{E}_k^r f \, dg \\
&= \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \int_G \frac{1}{f} (\tilde{E}_j^r f)(\tilde{E}_i^r f) \, dg \\
&\geq 0
\end{aligned}$$

5.2 The Generalized de Briujn Identity

This section generalizes the de Bruijn identity, in which entropy rates are related to Fisher information.

Theorem 1 *Let $f_{D,\mathbf{h},t}(g) = f(g, t; D, \mathbf{h})$ denote the solution of the diffusion equation (4) with constant $\mathbf{h} = [h_1, \dots, h_n]^T$ subject to the initial condition $f(g, 0; D, \mathbf{h}) = \delta(g)$. Then for any well-behaved pdf $\alpha(g)$,*

$$\frac{d}{dt} S(\alpha * f_{D,\mathbf{h},t}) = \frac{1}{2} \mathrm{tr}[DF^r(\alpha * f_{D,\mathbf{h},t})]. \tag{36}$$

Proof It is easy to see that the solution of the diffusion equation

$$\frac{\partial \rho}{\partial t} = \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \tilde{E}_i^r \tilde{E}_j^r \rho - \sum_{k=1}^{n} h_k \tilde{E}_k^r \rho \tag{37}$$

subject to the initial conditions $\rho(g, 0) = \alpha(g)$ is simply $\rho(g, t) = (\alpha * f_{D,\mathbf{h},t)}(g)$. This follows because all derivatives "pass through" the convolution integral for $\rho(g, t)$ and act on $f_{D,\mathbf{h},t}(g)$.

Taking the time derivative of $S(\rho(g, t))$ gives

$$\frac{d}{dt} S(\rho) = -\frac{d}{dt} \int_G \rho(g, t) \log \rho(g, t)\, dg = -\int_G \left\{ \frac{\partial \rho}{\partial t} \log \rho + \frac{\partial \rho}{\partial t} \right\} dg. \tag{38}$$

Using (37), the partial with respect to time can be replaced with Lie derivatives. But

$$\int_G \tilde{E}_k^r \rho\, dg = \int_G \tilde{E}_i^r \tilde{E}_j^r \rho\, dg = 0,$$

so the second term on the right side of (38) completely disappears. Using the integration-by-parts formula[6]

$$\int_G f_1\, \tilde{E}_k^r f_2\, dg = -\int_G f_2\, \tilde{E}_k^r f_1\, dg,$$

with $f_1 = \log \rho$ and $f_2 = \rho$ then gives

[6] There are no surface terms because, like the circle and real line, each coordinate in the integral either wraps around or goes to infinity.

$$
\begin{aligned}
\frac{d}{dt} S(\alpha * f_{D,\mathbf{h},t}) &= \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} \int_G \frac{1}{\alpha * f_{D,\mathbf{h},t}} \tilde{E}_j^r(\alpha * f_{D,\mathbf{h},t}) \tilde{E}_i^r(\alpha * f_{D,\mathbf{h},t})\, dg \\
&= \frac{1}{2} \sum_{i,j=1}^{n} D_{ij} F_{ij}^r(\alpha * f_{D,\mathbf{h},t}) \;=\; \frac{1}{2} \mathrm{tr}\left[D\, F^r(\alpha * f_{D,\mathbf{h},t})\right].
\end{aligned}
$$

The implication of this is that

$$
S(\alpha * f_{D,\mathbf{h},t_2}) - S(\alpha * f_{D,\mathbf{h},t_1}) = \frac{1}{2} \int_{t_1}^{t_2} \mathrm{tr}\left[D F^r(\alpha * f_{D,\mathbf{h},t})\right] dt.
$$

6 Mean, Covariance, and their Propagation Under Convolution

This section reviews concepts of mean and covariance for unimodular matrix Lie groups, and how they propagate under convolution. In these definitions, the concepts of Lie-theoretic exponential and logarithm play central roles. For a matrix Lie group, G, with corresponding Lie algebra, $\mathcal{G}$, the exponential map

$$
\exp : \mathcal{G} \longrightarrow G
$$

simply can be viewed as the matrix exponential defined by the Taylor series. In general, this map is neither surjective nor injective. However, it is possible to characterize the largest path-connected subset $\mathcal{G}^\circ \subset \mathcal{G}$ for which the image $G^\circ \doteq \exp(\mathcal{G}^\circ) \subset G$ has a well-defined inverse map

$$
\log : G^\circ \longrightarrow \mathcal{G}\,.
$$

This is also simply the matrix logarithm defined by its Taylor series.

For $SO(3)$, $SE(2)$, and $SE(3)$ which are three of the most common unimodular matrix Lie groups encountered in applications, the exponential map is surjective and G and G° differ only by a set of measure zero.

In what follows, it is assumed that all probability density functions $f : G \longrightarrow \mathbb{R}_{\geq 0}$ are either supported in G°, or that

$$
\int_G f(g)\, dg = \epsilon + \int_{G^\circ} f(g)\, dg
$$

where ϵ is an inconsequential probability. With this in mind, it becomes possible to blur the difference between G and G°.

6.1 Defining Mean

At least three different definitions for the mean of a pdf on a unimodular Lie group exist in the literature. The definitions reviewed here are all in the context of matrix-Lie-theoretic language which grew out of the author's applied work [85, 86]. For similar definitions expressed in differential-geometric terms see [87–89].

Directly generalizing the definition

$$\mathbf{m} = \int_{\mathbb{R}^n} \mathbf{x}\, f(\mathbf{x})\, d\mathbf{x}$$

to a Lie group is problematic because $\int_G g\, f(g)\, dg$ is not an element of the group. However, it is possible to define $m_0 \in G$ such that

$$\log m_0 = \int_G \log g\, f(g)\, dg\,.$$

Alternatively,

$$\int_{\mathbb{R}^n} (\mathbf{x} - \mathbf{m})\, f(\mathbf{x})\, d\mathbf{x} = \mathbf{0}$$

generalizes to searching for $m_1 \in G$ such that

$$\int_G \log\left(m_1^{-1} \circ g\right)\, f(g)\, dg \;=\; \mathbb{O}.$$

Thirdly,

$$\mathbf{m} = \underset{\mathbf{y}\,\in\,\mathbb{R}^n}{\operatorname{argmin}} \int_{\mathbb{R}^n} (\mathbf{x} - \mathbf{y})^2\, f(\mathbf{x})\, d\mathbf{x}$$

generalizes as

$$m_2 = \underset{h\,\in\,G}{\operatorname{argmin}} \int_G \left\|\log\left(h^{-1} \circ g\right)\right\|^2\, f(g)\, dg\,.$$

In general, no two of m_0, m_1, and m_2 are equal. However, in practice for distributions that are concentrated, they are quite close to each other. That said, if $f(g) = \rho(g)$ is a symmetric function, then all three reduce to the identity element, and hence are equal in this special case.

Though m_0 seems simple and straight forward, it has the undesirable property that shifting a symmetric pdf as $\rho(\mu^{-1} \circ g)$ does not automatically shift the mean from e to μ. m_2 has the problem that the norm $\|\cdot\|$ requires a choice of metric, and for noncompact unimodular Lie groups, a bi-invariant metric generally does not exist. Therefore, conjugating by an arbitrary $a \in G$ a symmetric pdf as $\rho(a^{-1} \circ g \circ a)$, which in the Euclidean setting would leave the mean fixed at e, results in a change to the value of the mean m_2 which depends on a.

In contrast, the mean m_1 shifts naturally with shifts of the pdf because

$$\int_G \log\left(m_1^{-1} \circ g\right)\, \rho(\mu^{-1} \circ g)\, dg \;=\; \int_G \log\left(m_1^{-1} \circ \mu \circ h\right)\, \rho(h) dh.$$

hence $m_1^{-1} \circ \mu = e$, or $m_1 = \mu$. Under conjugation of the pdf, the appearance of log linearly in the definition of m_1 means that

$$\int_G \log\left(m_1^{-1} \circ g\right) \rho(a^{-1} \circ g \circ a)\, dg = \int_G \log\left(m_1^{-1} \circ a \circ h \circ a^{-1}\right) \rho(h)\, dh \;=\; \mathbb{O}$$

can be written as

$$\int_G a^{-1}\, \log\left(m_1^{-1} \circ a \circ h \circ a^{-1}\right)\, a\, \rho(h)\, dh \;=\; a^{-1}\, \mathbb{O}\, a = \mathbb{O}.$$

But since

$$a^{-1}\, \log(g)\, a = \log(a^{-1} \circ g \circ a)\,,$$

then the mean m_1 of the conjugated pdf will be the conjugated mean. The implication of this general result in the special case when ρ is symmetric is

$$a^{-1} \circ m_1^{-1} \circ a = e \;\Longrightarrow\; m_1 = e,$$

giving the desirable property of invariance of the mean of a symmetric function under conjugation.

For these reasons, m_1 is chosen here (and in the author's previous work) as the best definition of the mean, and this is what will be used henceforth, and denoted as μ. The value of μ can be obtained numerically with an iterative procedure using m_0 as the initial starting point.

6.2 Defining Covariance

Previously the concepts of $\log : G^\circ \to \mathcal{G}$ and $\vee : \mathcal{G} \;\to\; \mathbb{R}^n$ where defined. The composition of these maps is defined as

$$\log^\vee G^\circ \;\to\; \mathbb{R}^n\,.$$

That is, for any $g \in G^\circ$, $\log^\vee(g) \in \mathbb{R}^n$.

One way to define the covariance of pdf on a unimodular Lie group G is [1, 2, 85, 86]

$$\Sigma \doteq \int_G \log^\vee(\mu^{-1} \circ g)[\log^\vee(\mu^{-1} \circ g)]^T f(g)\, dg\,. \tag{39}$$

This definition is natural as a generalization of the concept of covariance in Euclidean space when the pdf of interest is relatively concentrated. Then, for example, a Gaussian distribution can be defined as

$$f(g; \mu, \Sigma) \doteq \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}[\log^\vee(\mu^{-1} \circ g)]^T \Sigma^{-1} \log^\vee(\mu^{-1} \circ g)\right)\,.$$

This definition makes sense when the tails decay to negligible values inside a ball around μ for which the exponential and logarithm maps form a bijective pair. Otherwise, the topological properties of G become relevant.

Alternative definitions of scalar variance can be found in [2, 20, 90]. If the covariance as defined in (39) has been computed for pdfs f_1 and f_2, a convenient and accurate approximation for the covariance of $f_1 * f_2$ is known [1, 86]. This is known as a covariance propagation formula. In contrast, the scalar definitions in [20, 90] have exact propagation formulas, but these quantities do not have the form or properties that are usually associated with covariance of pdfs on Euclidean space.

An altogether different way to define covariance that does not involve any approximation is to recognize that for a Gaussian distribution with the mean serving as the statistic, the Cramér-Rao Bound becomes the equality

$$\Sigma_{gaussian} = F^{-1}_{gaussian}\,, \tag{40}$$

and since a Gaussian distribution with $\mu = e$ solves a driftless diffusion equation subject to Dirac delta initial conditions, it is possible to define a kind of covariance for such processes by computing the Fisher information and using (40). By generalization, an alternative definition of covariance can be taken as

$$\Sigma' \doteq F^{-1}\,.$$

The exact properties of this definition under convolution are unknown.

The covariance propagation formula for (39) involves the concept of the adjoint matrix, $Ad(g)$. This concept is reviewed in the following section.

6.3 The Adjoint Operators ad and Ad

Given $X, Y \in \mathcal{G}$, "little ad" operator is defined as

$$ad_Y(X) \doteq [Y, X] = YX - XY\,,$$

and "big Ad" is

$$Ad_g(X) \doteq gXg^{-1}\,.$$

Here “ad” is short for “adjoint”. Both of these are linear in X. That is, for arbitrary $c_1, c_2 \in \mathbb{R}$ and $X_1, X_2 \in \mathbb{R}^{n\times n}$

$$ad_Y(c_1X_1 + c_2X_2) = c_1 ad_Y(X_1) + c_2 ad_Y(X_2)$$

and

$$Ad_g(c_1X_1 + c_2X_2) = c_1 Ad_g(X_1) + c_2 Ad_g(X_2)\,.$$

Sometimes Ad_g is written as $Ad(g)$ and ad_Y is written as $ad(Y)$. It turns out that these are related as

$$Ad(\exp(Y)) = \exp(ad(Y)).$$

By introducing a basis for the Lie algebra of a Lie group, it is possible to express the Ad and ad operators as square matrices of the same dimension as the group. The distinction between operators and matrices can sometimes be confusing, which is why, for example, the matrices of $ad(X)$ and $Ad(A)$ are written as $[ad(X)]$ and $[Ad(A)]$ in [1, 2] where

$$[ad(X)]_{ij} = (E_i, ad(X)E_j) \quad \text{and} \quad [Ad(A)]_{ij} = (E_i, Ad(A)E_j)$$

computed using the inner product $(\cdot,\cdot)$.

6.4 Covariance Propagation

Given two pdfs with mean and covariance specified, i.e., $f_{(\mu_i,\Sigma_i)}(g)$ for $i = 1, 2$, one would like to be able to write expressions for μ_3, Σ_3 such that

$$f_{(\mu_1,\Sigma_1)} * f_{(\mu_2,\Sigma_2)} = f_{(\mu_3,\Sigma_3)}.$$

In the case when $G = \mathbb{R}^n$, the result is simply $\mathbf{m}_3 = \mathbf{m}_1 + \mathbf{m}_2$, and $\Sigma_3 = \Sigma_1 + \Sigma_2$. This result is nonparametric. That is, it does not require the pdfs to have a specific form, such as a Gaussian.

For general unimodular Lie groups, there is no simple exact formula. However, when the pdfs are concentrated, i.e., both $\|\Sigma_i\|$ are small, then it is possible to write [85]

$$\mu_3 \approx \mu_1 \circ \mu_2 \;\; \text{and} \;\; \Sigma_3 \approx \Sigma_1^{\mu_2} + \Sigma_2.$$

where

$$\Sigma_1^{\mu_2} \doteq [Ad_{\mu_2^{-1}}]\Sigma_1[Ad_{\mu_2^{-1}}]^T.$$

A higher-order approximation of the form [86]

$$\Sigma_3 \approx \Sigma_1^{\mu_2} + \Sigma_2 + \Phi(\Sigma_1^{\mu_2}, \Sigma_2) \tag{41}$$

is also possible, but in many applications $\Phi(\Sigma_1^{\mu_2}, \Sigma_2)$ is negligible because it depends quadratically on the elements of $\Sigma_1^{\mu_2}$ and Σ_2.

If $f_{(\mu,\Sigma)}(g) = f(g; \mu, \Sigma)$ denotes a Gaussian distribution in exponential coordinates on a d-dimensional Lie group, it will be of the form

$$f(g; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}[\log^\vee(\mu^{-1} \circ g)]^T \Sigma^{-1} \log^\vee(\mu^{-1} \circ g)\right) .$$

Then

$$f_{(\mu_1,\Sigma_1)} * f_{(\mu_2,\Sigma_2)} \approx f_{(\mu_3,\Sigma_3)} . \tag{42}$$

The quality of these approximations, as well as those that are even more accurate, have been studied in [1]. This is also a nonparametric result.

In the case when the distributions are more spread out, it is possible to compute the covariance in a different way using the group Fourier transform using the convolution theorem. For example, if $\Sigma_i = D^{(i)}t$ and

$$f_{(\mu_i,\Sigma_i)}(g) \doteq \rho_{\Sigma_i}(\mu_i^{-1} \circ g; t) ,$$

and since

$$\hat{f}_i(\lambda) = U(\mu_i, \lambda) \exp\left(\sum_{j,k} \sigma_{jk}^{(i)} E_j E_k\right) ,$$

from the convolution theorem it is possible to write the Fourier version of (42) as

$$U(\mu_2, \lambda) \exp\left(\sum_{j,k} \sigma_{jk}^{(2)} E_j E_k\right) U(\mu_1, \lambda) \exp\left(\sum_{j,k} \sigma_{jk}^{(1)} E_j E_k\right) \approx U(\mu_3, \lambda) \exp\left(\sum_{j,k} \sigma_{jk}^{(3)} E_j E_k\right) .$$

which can then be substituted in the reconstruction formula to reproduce (42), which produces an approximate expression involving traces, which is of a different type than the trace inequalities studied previously in the literature.

7 Examples

This section illustrates ideas presented earlier in this paper on the Heisenberg and rotation groups.

7.1 The Heisenberg Group, $\mathcal{H}(3)$

The Heisenberg group, $H(3)$, is defined by elements of the form

$$g(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & \alpha & \beta \\ 0 & 1 & \gamma \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{where} \qquad \alpha, \beta, \gamma \in \mathbb{R} \tag{43}$$

and the operation of matrix multiplication. Therefore, the group law can be viewed in terms of parameters as

$$g(\alpha_1, \beta_1, \gamma_1) g(\alpha_2, \beta_2, \gamma_2) = g(\alpha_1 + \alpha_2, \beta_1 + \beta_2 + \alpha_1\gamma_2, \gamma_1 + \gamma_2).$$

The identity element is the identity matrix $g(0, 0, 0)$, and the inverse of an arbitrary element $g(\alpha, \beta, \gamma)$ is

$$g^{-1}(\alpha, \beta, \gamma) = g(-\alpha, \alpha\gamma - \beta, -\gamma).$$

7.1.1 Lie Algebra and Exponential Map

Basis elements for the Lie algebra are

$$E_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}; \qquad E_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}; \qquad E_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{44}$$

A linear mapping between the Lie algebra spanned by this basis with $\mathbb{R}^3$ is defined by $E_i^\vee = \mathbf{e}_i$.

The Lie bracket is defined as $[E_i, E_j] = E_iE_j - E_jE_i = -[E_j, E_i]$, and so, as is always the case, $[E_i, E_i] = \mathbb{O}$. For these particular basis elements,

$$[E_1, E_2] = [E_2, E_3] = \mathbb{O} \qquad \text{and} \qquad [E_1, E_3] = E_2.$$

In addition, all double brackets involving the first two listed above are also zero,

$$[E_i, [E_1, E_2]] = [E_i, [E_2, E_3]] = \mathbb{O} \text{ for } i = 1, 2, 3$$

From these, and the bilinearity of the Lie bracket, it follows that for arbitrary

$$X = \sum_i x_i E_i \text{ and } Y = \sum_j x_j E_j$$

that

$$[X,Y]=\sum_{i,j}x_iy_j[E_i,E_j]=(x_1y_3-x_3y_1)E_2\,. \tag{45}$$

If the inner product for the Lie algebra spanned by these basis elements is defined as $(X,Y)=\mathrm{tr}(XY^T)$, then this basis is orthonormal: $(E_i,E_j)=\delta_{ij}$.

The group $H(3)$ is nilpotent because $(x_1E_1+x_2E_2+x_3E_3)^n=0$ for all $n\geq 3$. As a result, the matrix exponential is a polynomial in the coordinates $\{x_i\}$:

$$\exp\begin{pmatrix}0 & x_1 & x_2\\ 0 & 0 & x_3\\ 0 & 0 & 0\end{pmatrix}=g(x_1,x_2+\frac{1}{2}x_1x_3,x_3). \tag{46}$$

The parametrization in (43) can be viewed as the following product of exponentials:

$$g(\alpha,\beta,\gamma)=g(0,\beta,0)g(0,0,\gamma)g(\alpha,0,0)=\exp(\beta E_2)\exp(\gamma E_3)\exp(\alpha E_1).$$

The logarithm is obtained by solving for each x_i as a function of α,β,γ. By inspection this is $x_1=\alpha$, $x_3=\gamma$ and $x_2=\beta-\alpha\gamma/2$. Therefore,

$$\log g(\alpha,\beta,\gamma)=\begin{pmatrix}0 & \alpha & \beta-\alpha\gamma/2\\ 0 & 0 & \gamma\\ 0 & 0 & 0\end{pmatrix}.$$

7.1.2 Adjoint Matrices for $\mathcal{H}(3)$

The adjoint matrix, defined by $[Ad(g)]\mathbf{x}=(gXg^{-1})^\vee$, is computed by evaluating

$$\begin{pmatrix}1 & \alpha & \beta\\ 0 & 1 & \gamma\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}0 & x_1 & x_2\\ 0 & 0 & x_3\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}1 & -\alpha & \alpha\gamma-\beta\\ 0 & 1 & -\gamma\\ 0 & 0 & 1\end{pmatrix}=\begin{pmatrix}0 & x_1 & -\gamma x_1+x_2+\alpha x_3\\ 0 & 0 & x_3\\ 0 & 0 & 0\end{pmatrix}.$$

Therefore,

$$(gXg^{-1})^\vee=\begin{pmatrix}x_1\\ -\gamma x_1+x_2+\alpha x_3\\ x_3\end{pmatrix}\quad\text{and}\quad[Ad(g(\alpha,\beta,\gamma))]=\begin{pmatrix}1 & 0 & 0\\ -\gamma & 1 & \alpha\\ 0 & 0 & 1\end{pmatrix}.$$

The fact that $\det[Ad(g)]=1$ for all $g\in G$ indicates that this group is unimodular. This fact is independent of the parametrization. It can also be shown that for $X=\sum_{i=1}^3 x_iE_i$ that

$$[Ad(\exp X)]=\begin{pmatrix}1 & 0 & 0\\ -x_3 & 1 & x_1\\ 0 & 0 & 1\end{pmatrix}. \tag{47}$$

7.2 Bi-invariant Integration Measure

The Jacobian matrices for this group can be computed in either parametrization. In terms of α, β, γ,

$$\frac{\partial g}{\partial \alpha} = E_1; \qquad \frac{\partial g}{\partial \beta} = E_2; \qquad \frac{\partial g}{\partial \gamma} = E_3.$$

A straightforward calculation then gives

$$g^{-1}\frac{\partial g}{\partial \alpha} = E_1; \qquad g^{-1}\frac{\partial g}{\partial \beta} = E_2; \qquad g^{-1}\frac{\partial g}{\partial \gamma} = E_3 - \alpha E_2.$$

Therefore

$$J_r(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\alpha \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad J_l(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & 0 & 0 \\ -\gamma & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{48}$$

Then $|J_l(\alpha, \beta, \gamma)| = |J_r(\alpha, \beta, \gamma)| = 1$ and bi-invariant integration measure expressed in these coordinates is simply

$$dg = d\alpha d\beta d\gamma\,.$$

In exponential coordinates

$$J_r(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 \\ x_3/2 & 1 & -x_1/2 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad J_l(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 \\ -x_3/2 & 1 & x_1/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{49}$$

and

$$dg = dx_1 dx_2 dx_3\,.$$

7.3 Covariance Propagation and the EPI for $\mathcal{H}(3)$

From (45)the only nonzero term in the second-order covariance propagation formula from [1, 86] is

$$\frac{1}{4}[X, Y]^{\vee}\left([X, Y]^{\vee}\right)^T = \frac{1}{4}(x_1 y_3 - x_3 y_1)^2 \mathbf{e}_2 \mathbf{e}_2^T.$$

Hence, for $\mathcal{H}(3)$, the second-order term in (41), which results from integrating the above as was done in [86], becomes

$$\Phi(\Sigma_1^{\mu_2}, \Sigma_2) = a\mathbf{e}_2\mathbf{e}_2^T$$

where

$$a = \frac{1}{4}\left(\sigma_{11}^{(1)}\sigma_{33}^{(2)} - \sigma_{13}^{(1)}\sigma_{31}^{(2)} - \sigma_{31}^{(1)}\sigma_{13}^{(2)} + +\sigma_{33}^{(1)}\sigma_{11}^{(2)}\right) \geq 0$$

with $\sigma_{ij}^{(1)} = \mathbf{e}_i^T \Sigma_1^{\mu_2}\mathbf{e}_j$ and $\sigma_{ij}^{(2)} = \mathbf{e}_i^T \Sigma_2\mathbf{e}_j$.

Moreover, from the matrix identity for $A = A^T > 0$,

$$\det(A + \mathbf{u}\mathbf{v}^T) = (1 + \mathbf{v}^T A^{-1}\mathbf{u}) \det A \,,$$

it follows that

$$\det(\Sigma_{(1*2)}) = \det(\Sigma_1^{\mu_2} + \Sigma_2 + a\mathbf{e}_2\mathbf{e}_2^T) = (1 + a\mathbf{e}_2^T(\Sigma_1^{\mu_2} + \Sigma_2)^{-1}\mathbf{e}_2)\det(\Sigma_1^{\mu_2} + \Sigma_2)\,,$$

where $\Sigma_1^{\mu_2} = Ad_{\mu_2^{-1}}\Sigma_1 Ad_{\mu_2^{-1}}^T$. Consequently, from the Euclidean EPI and the unimodularity of G, which implies that

$$\det(\Sigma_1^{\mu_2}) = \det(\Sigma_1)\,,$$

it is not difficult to see that

$$\det(\Sigma_{(1*2)})^{\frac{1}{\dim}(G)} \geq \det(\Sigma_1^{\mu_2} + \Sigma_2)^{\frac{1}{\dim}(G)} \geq \det(\Sigma_1)^{\frac{1}{\dim}(G)} + \det(\Sigma_2)^{\frac{1}{\dim}(G)}.$$

That is, the entropy power inequality for pdfs from diffusion processes on $\mathcal{H}(3)$ follows in the small time limit from the classical EPI, and it is less restrictive than in the Euclidean case.

7.4 The Case of $SO(3)$

The group of rotations of three-dimensional space has elements that are 3 × 3 special orthogonal matrices, i.e., those satisfying

$$RR^T = \mathbb{I} \quad \text{and} \quad \det(R) = +1\,.$$

That is, they satisfy $RR^T = \mathbb{I}$ and $\det R = +1$. It is easy to see that closure of these properties under multiplication is satisfied because

$$(R_1R_2)^T(R_1R_2) = R_2^T R_1^T R_1 R_2 = R_2^T R_2 = \mathbb{I}$$

and

$$\det(R_1R_2) = \det(R_1)\det(R_2) = 1 \cdot 1 = 1.$$

7.4.1 The Lie Algebra

The Lie algebra $so(3)$ consists of skew-symmetric matrices of the form

$$X = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix} = \sum_{i=1}^{3} x_i E_i \,. \tag{50}$$

Every such matrix can be associated with a vector $\mathbf{x}$ by making the identification

$$E_i^{\vee} = \mathbf{e}_i \iff E_i = \hat{\mathbf{e}}_i \,.$$

For $SO(3)$ the adjoint matrices are

$$[Ad(R)] = R \quad \text{and} \quad [ad(X)] = X.$$

Furthermore,

$$[X, Y]^{\vee} = \mathbf{x} \times \mathbf{y}.$$

7.4.2 Exponential and Logarithm

It is well known that the exponential map $\exp : so(3) \to SO(3)$ is related to Euler's Theorem as

$$R = \exp(\theta N) = \mathbb{I} + \sin\theta\, N + (1 - \cos\theta)\, N^2 \,,$$

where $\theta \in [0, \pi]$ is the angle of rotation around the axis $\mathbf{n} \in S^2$, with N being the associated skew-symmetric matrix. Then $X = \theta N$ and $\mathbf{x} = \theta\mathbf{n}$. It is convenient to limit $\theta \in [0, \pi]$ and to allow $\mathbf{n}$ to take any value in the unit sphere, S^2. Moreover,

$$\mathrm{tr}(R) = 1 + 2\cos\theta \quad \text{and} \quad N = \frac{R - R^T}{2\sin\theta} \,.$$

Then, since

$$\theta = \cos^{-1}\left[\frac{\mathrm{tr}(R) - 1}{2}\right] \quad \text{and} \quad \sin(\cos^{-1} a) = \sqrt{1 - a^2} \,,$$

it follows that $\sin\theta$ can be written explicitly in terms of R as

$$\sin\theta = \sqrt{1 - \frac{(\mathrm{tr}(R) - 1)^2}{4}} = \sqrt{\frac{3}{4} - \frac{(\mathrm{tr}(R))^2}{4} + \frac{2\mathrm{tr}(R)}{4}} \,.$$

Since $X = \theta N = \log R$, it follows that

$$\log(R) \; = \; \frac{\cos^{-1}\left[\frac{\mathrm{tr}(R)-1}{2}\right](R-R^T)}{\sqrt{3-(\mathrm{tr}(R))^2+2\mathrm{tr}(R)}} \, . \tag{51}$$

This expression breaks down when $\theta = \pi$, which defines a set of measure zero, and hence is inconsequential when evaluating the logarithm under an integral.

7.4.3 Invariant Integration Measure

Two common ways to parameterize rotations are using the matrix exponential $R = \exp X$ and using Euler angles such as $R = R_3(\alpha)R_1(\beta)R_3(\gamma)$ where $0 \leq \alpha, \gamma \leq 2\pi$ and $0 \leq \beta \leq \pi$.

Relatively simple analytical expressions were derived by Park [91] for the Jacobian J_l when $R = \exp X$ as

$$J_l(\mathbf{x}) = \mathbb{I} + \frac{1-\cos\|\mathbf{x}\|}{\|\mathbf{x}\|^2}X + \frac{\|\mathbf{x}\|-\sin\|\mathbf{x}\|}{\|\mathbf{x}\|^3}X^2 \tag{52}$$

The corresponding Jacobian J_r and its inverse are [1, 2]

$$J_r(\mathbf{x}) = \mathbb{I} - \frac{1-\cos\|\mathbf{x}\|}{\|\mathbf{x}\|^2}X + \frac{\|\mathbf{x}\|-\sin\|\mathbf{x}\|}{\|\mathbf{x}\|^3}X^2$$

In terms of ZXZ Euler angles,

$$J_l(\alpha,\beta,\gamma) = [\mathbf{e}_3,\, R_3(\alpha)\mathbf{e}_1,\, R_3(\alpha)R_1(\beta)\mathbf{e}_3] = \begin{pmatrix} 0 & \cos\alpha & \sin\alpha\sin\beta \\ 0 & \sin\alpha & -\cos\alpha\sin\beta \\ 1 & 0 & \cos\beta \end{pmatrix}. \tag{53}$$

and

$$J_r = R^T J_l = [R_3(-\gamma)R_1(-\beta)\mathbf{e}_3,\, R_3(-\gamma)\mathbf{e}_1,\, \mathbf{e}_3] = \begin{pmatrix} \sin\beta\sin\gamma & \cos\gamma & 0 \\ \sin\beta\cos\gamma & -\sin\gamma & 0 \\ \cos\beta & 0 & 1 \end{pmatrix}. \tag{54}$$

From this we see that

$$dR = \frac{2(1-\cos\|\mathbf{x}\|)}{\|\mathbf{x}\|^2}\, dx_1dx_2dx_3 = \sin\beta\, d\alpha d\beta d\gamma \, .$$

From these it can be shown that

$$\int_{SO(3)} dR = 8\pi^2 \, .$$

7.4.4 Fourier Series

For $SO(3)$, irreducible unitary representations (IURs) [1] are enumerated by $l \in \mathbb{Z}_{\geq 0}$, and for any $R, A \in SO(3)$ these $(2l+1) \times (2l+1)$ IUR matrices have the fundamental properties

$$U^l(RA) = U^l(R)\, U^l(A) \;\text{ and }\; U^l(R^T) = U^l(R)^*$$

where $*$ is the Hermitian conjugate of a matrix. The explicit forms of these matrices when R is expressed in Euler angles are well known in Physics as the Wigner-D functions [92–96]

For functions $f \in L^2(SO(3))$, the Fourier coefficients are computed as

$$\hat{f}^l_{mn} = \int_{SO(3)} f(A) U^l_{mn}(A^{-1})\, dA\,. \tag{55}$$

The following orthogonality relation holds

$$\int_{SO(3)} U^l_{mn}(A)\overline{U^s_{pq}(A)}\, dA = \frac{1}{2l+1}\delta_{ls}\delta_{mp}\delta_{nq} \tag{56}$$

where dA is scaled so that $\int_{SO(3)} dA = 1$. The Fourier series on $SO(3)$ has the form

$$f(A) = \sum_{l=0}^{\infty}(2l+1)\sum_{m=-l}^{l}\sum_{n=-l}^{l} \hat{f}^l_{mn} U^l_{nm}(A)\,, \tag{57}$$

which results from the completeness relation

$$\sum_{l=0}^{\infty}(2l+1)\sum_{m=-l}^{l}\sum_{n=-l}^{l} U^l_{mn}(R^{-1}) U^l_{nm}(A) = \delta(R^{-1}A)\,. \tag{58}$$

Another way to write (57) is

$$f(A) = \sum_{l=0}^{\infty}(2l+1)\text{trace}\left[\hat{f}^l\, U^l(A)\right]. \tag{59}$$

7.5 Diffusions on $SO(3)$

A diffusion process on $SO(3)$ commonly encountered in applications is of the form

$$\frac{\partial f}{\partial t} = \frac{1}{2}\sum_{i,j=1}^{3} D_{ij}\tilde{E}_i\tilde{E}_j f + \sum_{k=1}^{3} d_k \tilde{E}_k f\,. \tag{60}$$

By expanding the PDF in the PDE in (60) into a Fourier series on $SO(3)$, the solution can be obtained once we know how the differential operators X_i^R transform the matrix elements $U^l_{m,n}(A)$. Explicitly,

$$\tilde{E}_1 U^l_{mn} = \frac{1}{2}c^l_{-n}U^l_{m,n-1} - \frac{1}{2}c^l_n U^l_{m,n+1}; \tag{61}$$

$$\tilde{E}_2 U^l_{mn} = \frac{1}{2}ic^l_{-n}U^l_{m,n-1} + \frac{1}{2}ic^l_n U^l_{m,n+1}; \tag{62}$$

$$\tilde{E}_3 U^l_{mn} = -inU^l_{mn}; \tag{63}$$

where $c^l_n = \sqrt{(l-n)(l+n+1)}$ for $l \geq |n|$ and $c^l_n = 0$ otherwise. From this definition it is clear that $c^k_k = 0$, $c^l_{-(n+1)} = c^l_n$, $c^l_{n-1} = c^l_{-n}$, and $c^l_{n-2} = c^l_{-n+1}$.

By repeated application of these rules, it can be shown that [1]

$$\mathcal{F}\left(\frac{1}{2}\sum_{i,j=1}^{3} D_{ij}\tilde{E}_i\tilde{E}_j f + \sum_{i=1}^{3} d_i\tilde{E}_i f\right)^l_{mn} = \sum_{k=\max(-l,m-2)}^{\min(l,m+2)} \mathcal{A}^l_{m,k}\hat{f}^l_{k,n},$$

where

$$\mathcal{A}^l_{m,m+2} = \left[\frac{(D_{11}-D_{22})}{8} + \frac{i}{4}D_{12}\right]c^l_{m+1}c^l_{-m-1};$$

$$\mathcal{A}^l_{m,m+1} = \left[\frac{(2m+1)}{4}(D_{23}-iD_{13}) + \frac{1}{2}(d_1+id_2)\right]c^l_{-m-1};$$

$$\mathcal{A}^l_{m,m} = \left[-\frac{(D_{11}+D_{22})}{8}(c^l_{-m}c^l_{m-1} + c^l_m c^l_{-m-1}) - \frac{D_{33}m^2}{2} - id_3 m\right];$$

$$\mathcal{A}^l_{m,m-1} = \left[\frac{(2m-1)}{4}(D_{23}+iD_{13}) + \frac{1}{2}(-d_1+id_2)\right]c^l_{m-1};$$

$$\mathcal{A}^l_{m,m-2} = \left[\frac{(D_{11}-D_{22})}{8} - \frac{i}{4}D_{12}\right]c^l_{-m+1}c^l_{m-1};$$

Hence, application of the $SO(3)$-Fourier transform to (60) and corresponding initial conditions reduces (60) to a set of linear time-invariant ODEs of the form

$$\frac{d\hat{f}^l}{dL} = \mathcal{A}^l \hat{f}^l \qquad \text{with} \qquad \hat{f}^l(0) = \mathbb{I}_{2l+1}. \tag{64}$$

Here $\mathbb{I}_{2l+1}$ is the $(2l+1) \times (2l+1)$ identity matrix and the banded matrix $\mathcal{A}^l$ are of the following form for $l = 0, 1, 2, 3$:

$$\mathcal{A}^0 = \mathcal{A}^0_{0,0} = 0; \quad \mathcal{A}^1 = \begin{pmatrix} \mathcal{A}^1_{-1,-1} & \mathcal{A}^1_{-1,0} & \mathcal{A}^1_{-1,1} \\ \mathcal{A}^1_{0,-1} & \mathcal{A}^1_{0,0} & \mathcal{A}^1_{0,1} \\ \mathcal{A}^1_{1,-1} & \mathcal{A}^1_{1,0} & \mathcal{A}^1_{1,1} \end{pmatrix};$$

$$\mathcal{A}^2 = \begin{pmatrix} \mathcal{A}^2_{-2,-2} & \mathcal{A}^2_{-2,-1} & \mathcal{A}^2_{-2,0} & 0 & 0 \\ \mathcal{A}^2_{-1,-2} & \mathcal{A}^2_{-1,-1} & \mathcal{A}^2_{-1,0} & \mathcal{A}^2_{-1,1} & 0 \\ \mathcal{A}^2_{0,-2} & \mathcal{A}^2_{0,-1} & \mathcal{A}^2_{0,0} & \mathcal{A}^2_{0,1} & \mathcal{A}^2_{0,2} \\ 0 & \mathcal{A}^2_{1,-1} & \mathcal{A}^2_{1,0} & \mathcal{A}^2_{1,1} & \mathcal{A}^2_{1,2} \\ 0 & 0 & \mathcal{A}^2_{2,0} & \mathcal{A}^2_{2,1} & \mathcal{A}^2_{2,2} \end{pmatrix};$$

$$\mathcal{A}^3 = \begin{pmatrix} \mathcal{A}^3_{-3,-3} & \mathcal{A}^3_{-3,-2} & \mathcal{A}^3_{-3,-1} & 0 & 0 & 0 & 0 \\ \mathcal{A}^3_{-2,-3} & \mathcal{A}^3_{-2,-2} & \mathcal{A}^3_{-2,-1} & \mathcal{A}^3_{-2,0} & 0 & 0 & 0 \\ \mathcal{A}^3_{-1,-3} & \mathcal{A}^3_{-1,-2} & \mathcal{A}^3_{-1,-1} & \mathcal{A}^3_{-1,0} & \mathcal{A}^3_{-1,1} & 0 & 0 \\ 0 & \mathcal{A}^3_{0,-2} & \mathcal{A}^3_{0,-1} & \mathcal{A}^3_{0,0} & \mathcal{A}^3_{0,1} & \mathcal{A}^3_{0,2} & 0 \\ 0 & 0 & \mathcal{A}^3_{1,-1} & \mathcal{A}^3_{1,0} & \mathcal{A}^3_{1,1} & \mathcal{A}^3_{1,2} & \mathcal{A}^3_{1,3} \\ 0 & 0 & 0 & \mathcal{A}^3_{2,0} & \mathcal{A}^3_{2,1} & \mathcal{A}^3_{2,2} & \mathcal{A}^3_{2,3} \\ 0 & 0 & 0 & 0 & \mathcal{A}^3_{3,1} & \mathcal{A}^3_{3,2} & \mathcal{A}^3_{3,3} \end{pmatrix}.$$

The solution to (64) is then of the form of a matrix exponential:

$$\hat{f}^l(L) = e^{L\mathcal{A}^l}. \tag{65}$$

Since $\mathcal{A}^l$ is a band-diagonal matrix for $l > 1$, the matrix exponential can be calculated much more efficiently (either numerically or symbolically) for large values of l than for general matrices of dimension $(2l+1) \times (2l+1)$.

Given the explicit forms provided above, (32)–(35) can be verified.

7.5.1 Lack of an Entropy-Power Inequality

For all unimodular Lie groups, the EPI holds for concentrated Gaussian pdfs for which the first-order covariance propagation formula from [85] holds by application of the Euclidean EPI to Gaussians. However, for compact Lie groups (including the circle and n-torus) the EPI always breaks down. For example, the uniform distribution on the circle, $\rho(\theta) = 1/2\pi$, has entropy $S(\rho) = \log(1/2\pi)$. But since this distribution is stable under convolution, we have that $S(\rho * \rho) = S(\rho)$ and so the EPI cannot hold since $N(\rho * \rho) = N(\rho) < 2 \cdot N(\rho)$. Similarly, unlike for $\mathcal{H}(3)$, the EPI does not hold for $SO(3)$.

8 Conclusions

Many inequalities of information theory that are based on probability densities on Euclidean space extend to the case of probabilities on Lie groups. In addition to reviewing appropriate concepts of integration, convolution, partial derivative, Fourier transform, covariance, and diffusion processes on unimodular Lie groups, this paper also presents some new inequalities that extend to this setting those known in the classical Abelian case.

References

1. Chirikjian, G.S., Kyatkin, A.B.: Harmonic Analysis for Engineers and Applied Scientists. Dover Publications, Mineola, NY (2016)
2. Chirikjian, G.S.: Stochastic Models, Information Theory, and Lie Groups: Volume 2 - Analytic Methods and Modern Applications. Birkhäuser, Boston (2011)
3. Howe, R., Tan, E.C.: Non-abelian Harmonic Analysis. Springer, Berlin (1992)
4. Lang, S.: $SL_2(R)$. Addison-Wesley, Reading (1975)
5. Chandra, H.: Spherical functions on a semisimple Lie group II. Am. J. Math. **27**, 569–579 (1960)
6. Jorgenson, J., Lang, S.: Spherical Inversion on $SL_n(R)$. Springer, Berlin (2001)
7. Thangavelu, S.: Harmonic Analysis on the Heisenberg Group. Birkhäuser, Boston (1998)
8. Neuenschwander, D.: Probabilities on the Heisenberg Group: Limit Theorems and Brownian Motion. Lecture Notes in Mathematics, vol. 1630. Springer, Berlin (1996)
9. Miller Jr., W.: Lie Theory and Special Functions. Academic Press, New York (1968)
10. Miller Jr., W.: Some applications of the representation theory of the Euclidean group in three-space. Commun. Pure App. Math. **17**, 527–540 (1964)
11. Vilenkin, N.Ja., Klimyk, A.U.: Representation of Lie Groups and Special Functions, vol. 1–3. Kluwer Academic Publishers, Dordrecht, Holland (1991)
12. Vilenkin, N.J.: Special Functions and the Theory of Group Representations. American Mathematical Society, Providence (1968)
13. Vilenkin, N.J., Akim, E.L., Levin, A.A.: The matrix elements of irreducible unitary representations of the group of Euclidean three-dimensional space motions and their properties. Dokl. Akad. Nauk SSSR **112**, 987–989 (1957). (in Russian)
14. Wang, Y., Zhou, Y., Maslen, D.K., Chirikjian, G.S.: Solving the phase-noise Fokker-Planck equation using the motion-group Fourier transform. IEEE Trans. Commun. **54**(5), 868–877 (May, 2006)
15. Zhou, Y., Chirikjian, G.S.: Conformational statistics of semi-flexible macromolecular chains with internal joints. Macromolecules **39**(5), 1950–1960 (2006)
16. Chirikjian, G.S., Kyatkin, A.B.: An operational calculus for the Euclidean motion group with applications in robotics and polymer science. J. Fourier Anal. Appl. **6**(6), 583–606 (December, 2000)

17. Chirikjian, G.S.: Degenerate diffusions and harmonic analysis on SE(3): a tutorial. In: Albeverio, S., Cruzeiro, A., Holm, D. (eds.) Stochastic Geometric Mechanics, pp. 77–99. Springer, Berlin (2017)
18. Chirikjian, G.S.: Information-theoretic inequalities on unimodular Lie groups. J. Geom. Mech. **2**(2), 119–158 (June, 2010)
19. Folland, G.B.: A Course in Abstract Harmonic Analysis. CRC Press, Boca Raton, FL (1995)
20. Grenander, U.: Probabilities on Algebraic Structures. Dover Publications, Mineola (2008)
21. Gross, K.I.: Evolution of noncommutative harmonic analysis. Am. Math. Mon. **85**(7), 525–548 (1978)
22. Gurarie, D.: Symmetry and Laplacians. Introduction to Harmonic Analysis, Group Representations and Applications, Dover edn. Elsevier Science Publisher, The Netherlands, 1992 (2008)
23. Hewitt, E., Ross, K.A.: Abstract Harmonic Analysis I, and II. Springer, Berlin, 1963 and 1970. (Reprinted 1994)
24. Sugiura, M.: Unitary Representations and Harmonic Analysis, 2nd edn. North-Holland, Amsterdam (1990)
25. Taylor, M.E.: Noncommutative Harmonic Analysis. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI (1986)
26. Kunze, R.: L_p Fourier transforms on locally compact unimodular groups. Trans. Am. Math. Soc. **89**, 519–540 (1958)
27. Applebaum, D.: Probability on Compact Lie Groups. Springer, New York (2014)
28. Beckner, W.: Sharp inequalities and geometric manifolds. J. Fourier Anal. Appl. **3**, 825–836 (1997)
29. Beckner, W.: Geometric inequalities in Fourier analysis. Essays on Fourier Analysis in Honor of Elias M. Stein, pp. 36–68. Princeton University Press, Princeton (1995)
30. Blachman, N.M.: The convolution inequality for entropy powers. IEEE Trans. Inform. Theory **11**(2), 267–271 (1965)
31. Carlen, E.A.: Superadditivity of fishers information and logarithmic Sobolev inequalities. J. Funct. Anal. **101**, 194–211 (1991)
32. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley-Interscience, Hoboken, NJ (2006)
33. Dembo, A., Cover, T.M., Thomas, J.A.: Information theoretic inequalities. IEEE Trans. Inf. Theory **37**(6), 1501–1518 (1991)
34. Varopoulos, N.T., Saloff-Coste, L., Coulhon, T.: Analysis and Geometry on Groups. Cambridge University Press, Cambridge (1992)
35. Maslen, D.K.: Fast Transforms and Sampling for Compact Groups, Ph.D. Dissertation, Department of Mathematics, Harvard University (May 1993)
36. Maslen, D.K., Rockmore, D.N.: Generalized FFTSA survey of some recent results. DIMACS Ser. Discret. Math. Theor. Comput. Sci. **28**, 183–237 (1997)
37. Hardy, G.H., Littlewood, J.E., Pólya, G.: Inequalities. Cambridge University Press, Cambridge (1932)
38. Pólya, G.: Isoperimetric Inequalities in Mathematical Physics. Princeton University Press, Princeton (1951)
39. Simon, B.: Trace Ideals and their Applications. Mathematical Surveys and Monographs, 2nd edn. American Mathematical Society, Providence (2010)
40. Bhatia, R.: Positive Definite Matrices. Princeton University Press, Princeton (2007)
41. Bernstein, D.S.: Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory. Princeton University Press, Princeton (February 22, 2005)
42. Trotter, H.F.: On the product of semi-groups of operators. Proc. Amer. Math. Soc. 10545–10551 (1959)
43. So, W.: Equality cases in matrix exponential inequalities. SIAM J. Matrix Anal. Appl. **13**(4), 1154–1158 (October, 1992)
44. Mirsky, L.: A Note on normal matrices. Am. Math. Mon. **63**(7), 479 (August–September, 1956)

45. Reid, R.M.: Some eigenvalue properties of persymmetric matrices. SIAM Rev. **39**(2), 313–316 (June, 1997)
46. Schur, I.: Über die charakteristischen Wurzeln einer linearen Substitution mit einer Anwendung auf die Theorie der Integralgleichungen. Math. Annalen **66**, 488–510 (1909)
47. Thurston, H.S.: On the characteristic equations of products of square matrices. Am. Math. Mon. **38**, 322–324 (1931)
48. Scott, W.M.: On characteristic roots of matrix products. Am. Math. Mon. **48**(3), 201–203 (March, 1941)
49. Bhatia, R., Parthasarathy, K.R.: Positive definite functions and operator inequalities. Bull. Lond. Math. Soc. **32**, 214–228 (2000)
50. Andruchow, E., Corach, G., Stojanoff, D.: Geometric operator inequalities. Linear Algebra Appl. **258**, 295–310 (1997)
51. Bhatia, R., Kittaneh, F.: On singular values of a product of operators. SIAM J. Matrix Anal. Appl. **11**, 272–277 (1990)
52. Bhatia, R.: Matrix Analysis. Springer, Berlin (1996)
53. Thompson, C.J.: Inequalities and partial orders on matrix spaces. Indiana Univ. Math. J. **21**(5), 469–480 (1971)
54. Golden, S.: Lower bounds for the Helmholtz function. Phys. Rev. **137**, B1127–B1128 (1965)
55. Thompson, C.J.: Mathematical Statistical Mechanics. Macmillan, New York (1972); Reprint, Princeton University Press, Princeton (1979, 1992)
56. Lenard, A.: Generalization of the Golden-Thompson inequality $Tr(e^A e^B) \geq Tr(e^{A+B})$. Indiana Univ. Math. J. **21**, 457–467 (1971)
57. Thompson, R.C.: Special cases of a matrix exponential formula. Linear Algebra Appl. **107**, 283–292 (1988)
58. So, W., Thompson, R.C.: Products of exponentials of Hermitian and complex symmetric matrices. Linear and Multilinear Algebra **29**, 225–233 (1991)
59. So, W.: The high road to an exponential formula. Linear Algebra Appl. **379**, 69–75 (2004)
60. Klyachko, A.A.: Random walks on symmetric spaces and inequalities for matrix spectra. Linear Algebra Appl. **319**, 37–59 (2000)
61. Cohen, J.E., Friedland, S., Kato, T., Kelly, F.P.: Eigenvalue inequalities for products of matrix exponentials. Linear Algebra Appl. **45**, 55–95 (1982)
62. Bernstein, D.S.: Inequalities for the trace of matrix exponentials. SIAM J. Matrix Anal. Appl. **9**, 156–158 (1988)
63. Fan, K.: Maximum properties and inequalities for the eigenvalues of completely continuous operators. PNAS **37**, 760–766 (1951)
64. Fan, K.: On a theorem of Weyl concerning eigenvalues of linear transformations I. Proc. Nat. Acad. Sci. USA **35**, 652–655 (1949)
65. Bebiano, N., da Providência, J., Jr., Lemos, R.: Matrix inequalities in statistical mechanics. Linear Algebra Appl. **376**(1), 265–273 (January, 2004)
66. Friedland, S., So, W.: Product of matrix exponentials. Linear Algebra Appl. **196**, 193–205 (1994)
67. Friedland, S., Porta, B.: The limit of the product of the parameterized exponentials of two operators. J. Funct. Anal. **210**, 436–464 (2004)
68. Mori, T.: Comments on a matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equation. IEEE Trans. Autom. Control **AC-29**, 1088 (November, 1988)
69. Fang, Y., Loparo, K.A., Feng, X.: Inequalities for the trace of matrix product. IEEE Trans. Autom. Control **39**(12), 2489–2490 (December, 1994)
70. Park, P.-G.: On the trace bound of a matrix product. IEEE Trans. Autom. Control **41**(12), 1799–1802 (December, 1996)
71. Komaroff, N.: Bounds on eigenvalues of matrix products with an application to the algebraic Riccati equation. IEEE Trans. Autom. Control **35**(3), 348–350 (March, 1990)
72. Lasserre, J.B.: A trace inequality for the matrix product. IEEE Trans. Autom. Control **40**, 1500–1501 (1995)

73. Lasserre, J.B.: Tight bounds for the trace of a matrix product. IEEE Trans. Autom. Control **42**(4), 578–581 (April, 1997)
74. Zhang, F., Zhang, Q.: Eigenvalue inequalities for matrix product. IEEE Trans. Autom. Control **51**(9), 1506–1509 (September, 2006)
75. Xing, W., Zhang, Q., Wang, Q.: A trace bound for a general square matrix product. IEEE Trans. Autom. Control **45**(8), 1563–1565 (August, 2000)
76. Liu, J., He, L.: A new trace bound for a general square matrix product. IEEE Trans. Autom. Control **52**(2), 349–352 (February 2007)
77. Mirsky, L.: A trace inequality of John von Neumann. Monatshefte für Mathematik **79**, 303–306 (1975)
78. Hoffman, A.J., Wielandt, H.W.: The variation of the spectrum of a normal matrix. Duke Math. J. **20**, 37–40 (1953)
79. Cochran, J.A., Hinds, E.W.: Improved error bounds for the eigenvalues of certain normal operators. SIAM J. Numer. Anal. **9**(3), 446–453 (September, 1972)
80. Richter, H.: Zur Abschätzung von Matrizennormen. Mathematische Nachrichten **18**, 178–187 (1958)
81. Mirsky, L.: On the trace of matrix products. Mathematische Nachrichten **20**, 171–174 (1959)
82. Marcus, M.: An eigenvalue inequality for the product of normal matrices. Am. Math. Mon. **63**(3), 173–174 (March, 1956)
83. Steele, J.M.: The Cauchy-Schwarz Master Class : An Introduction to the Art of Mathematical Inequalities. Cambridge University Press, Cambridge, New York (2004)
84. Neuman, E., Sándor, J.: On the Ky Fan inequality and related inequalities I. Math. Inequalities Appl. **5**(1), 49–56 (2002)
85. Wang, Y., Chirikjian, G.S.: Error propagation on the Euclidean group with applications to manipulator kinematics. IEEE Trans. Robot. **22**(4), 591–602 (August, 2006)
86. Wang, Y., Chirikjian, G.S.: Nonparametric second-order theory of error propagation on the Euclidean group. Int. J. Robot. Res. **27**(11–12), 1258–1273 (2008)
87. Pennec, X.: *L'incertitude dans les problèmes de reconnaissance et de recalage–Applications en imagerie médicale et biologie moléculaire*, (Doctoral dissertation, Ecole Polytechnique X) (1996)
88. Pennec, X.: Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. J. Math Imaging Vis. **25**, 127 (July, 2006)
89. Pennec, X., Arsigny, V.: Exponential barycenters of the canonical Cartan connection and invariant means on Lie groups. In: Barbaresco, F., Mishra, A., Nielsen, F. (eds.) Matrix Information Geometry, pp. 123–166. Springer, Berlin (May, 2012)
90. Heyer, H.: Probability Measures on Locally Compact Groups. Springer, New York (1977)
91. Park, F.C.: The Optimal Kinematic Design of Mechanisms. Ph.D. thesis, Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA (1991)
92. Biedenharn, L.C., Louck, J.D.: Angular Momentum in Quantum Physics. Encyclopedia of Mathematics and Its Applications, vol. 8. Cambridge University Press, Cambridge (1985). (paperback version 2009)
93. Gelfand, I.M., Minlos, R.A., Shapiro, Z.Ya.: Representations of the Rotation and Lorentz Groups and their Applications. Macmillan, New York (1963)
94. Talman, J.: Special Functions. W. A. Benjamin Inc., Amsterdam (1968)
95. Varshalovich, D.A., Moskalev, A.N., Khersonskii, V.K.: Quantum Theory of Angular Momentum. World Scientific, Singapore (1988)
96. Wigner, E.P.: Group Theory and Its Applications to the Quantum Mechanics of Atomic Spectra. Academic Press, New York (1959)

Warped Riemannian Metrics for Location-Scale Models

Salem Said, Lionel Bombrun and Yannick Berthoumieu

Abstract The present contribution shows that warped Riemannian metrics, a class of Riemannian metrics which play a prominent role in Riemannian geometry, are also of fundamental importance in information geometry. Precisely, the starting point is a new theorem, which states that the Rao–Fisher information metric of any location-scale model, defined on a Riemannian manifold, is a warped Riemannian metric, whenever this model is invariant under the action of some Lie group. This theorem is a valuable tool in finding the expression of the Rao–Fisher information metric of location-scale models defined on high-dimensional Riemannian manifolds. Indeed, a warped Riemannian metric is fully determined by only two functions of a single variable, irrespective of the dimension of the underlying Riemannian manifold. Starting from this theorem, several original results are obtained. The expression of the Rao–Fisher information metric of the Riemannian Gaussian model is provided, for the first time in the literature. A generalised definition of the Mahalanobis distance is introduced, which is applicable to any location-scale model defined on a Riemannian manifold. The solution of the geodesic equation, as well as an explicit construction of Riemannian Brownian motion, are obtained, for any Rao–Fisher information metric defined in terms of warped Riemannian metrics. Finally, using a mixture of analytical and numerical computations, it is shown that the parameter space of the von Mises–Fisher model of n-dimensional directional data, when equipped with its Rao–Fisher information metric, becomes a Hadamard manifold, a simply-connected complete Riemannian manifold of negative sectional curvature, for $n = 2, \ldots, 8$. Hopefully, in upcoming work, this will be proved for any value of n.

S. Said (✉)
Laboratoire IMS, CNRS, Université de Bordeaux, Bordeaux, France
e-mail: salem.said@u-bordeaux.fr

L. Bombrun
Laboratoire IMS, Bordeaux Sciences Agro, Bordeaux, France
e-mail: lionel.bombrun@u-bordeaux.fr

Y. Berthoumieu
Laboratoire IMS, Bordeaux INP, Bordeaux, France
e-mail: yannick.berthoumieu@u-bordeaux.fr

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_10

1 Introduction

Warped Riemannian metrics are a class of Riemannian metrics which arise throughout Riemannian geometry [11, 40]. For example, the Riemannian metrics of surfaces of revolution, and of spaces of constant curvature (when restricted to polar coordinate charts), are warped Riemannian metrics. Closely similar to warped Riemannian metrics, warped semi-Riemannian metrics are very important in theoretical physics. Indeed, many gravitational models are given by warped semi-Riemannian metrics [38]. The present contribution shows that warped metrics, in addition to their well-known role in geometry and physics, play a fundamental role in information geometry, and have a strong potential for applications in statistical inference and statistical learning.

A unified definition of warped Riemannian metrics was first formulated in [11]. Here, only a special case of this definition is required. Precisely, let M be a complete Riemannian manifold, with length element ds^2_M , and consider the product manifold $\mathcal{M} = M \times (0\,,\infty)$, equipped with the length element $ds^2_{\mathcal{M}}$

$$ds^2_{\mathcal{M}}(z) \,=\, dr^2 \,+\, \beta^2(r)\, ds^2_M(x) \qquad \text{for } \; z = (x, r) \in \mathcal{M} \tag{1a}$$

where $\beta^2(r)$ is a strictly positive function. Then, the length element $ds^2_{\mathcal{M}}$ defines a warped Riemannian metric on $\mathcal{M}$. In Riemannian geometry, the coordinate r is a distance function, measuring the distance to some point or hypersurface [40]. In physics, r is replaced by the time t, and dr^2 is replaced by $-dt^2$ in formula (1a) (this is the meaning of "semi-Riemannian") [38]. In any case, the coordinate x can be thought of as a spatial coordinate which determines a position in M.

The intuition behind the present contribution is that warped Riemannian metrics are natural candidates for Riemannian metrics on location-scale models. Indeed, if $\mathcal{P}$ is a location-scale model on M, with location parameter $\bar{x} \in M$ and scale parameter $\sigma > 0$, then the parameter space of $\mathcal{P}$ is exactly $\mathcal{M} = M \times (0\,,\infty)$ with its points $z = (\bar{x}, \sigma)$. Thus, a warped Riemannian metric on $\mathcal{M}$ can be defined using (1a), after introducing a new scale parameter $r = r(\sigma)$ and setting $x = \bar{x}$.

As it turns out, this intuition is far from arbitrary. The main new result in the present contribution, Theorem 1 of Sect. 3, states that the Rao–Fisher information metric of any location-scale model is a warped Riemannian metric, whenever this model is invariant under the action of some Lie group. Roughly, Theorem 1 states that if M is a Riemannian symmetric space under the transitive action of a Lie group of isometries G, and if each probability density $p(x|\,\bar{x}, \sigma)$, belonging to the model $\mathcal{P}$, verifies the invariance condition

$$p(\,g\cdot x|\,g\cdot \bar{x}\,,\sigma) \,=\, p(x|\bar{x}, \sigma) \quad \text{for all } g \in G \tag{1b}$$

where $g \cdot x$ denotes the action of $g \in G$ on $x \in M$, then the Rao–Fisher information metric of the model $\mathcal{P}$ is a warped Riemannian metric.

A technical requirement for Theorem 1 is that the Riemannian symmetric space M should be irreducible. The meaning of this requirement, and the fact that it can be relaxed in certain cases, are discussed in Remarks 4 and 5 of Sect. 3. The proof of Theorem 1 is given in Appendix A.

A fundamental idea of information geometry is that the parameter space of a statistical model $\mathcal{P}$ should be considered as a Riemannian manifold [3, 15]. According to [7, 15], the unique way of doing so is by turning Fisher's information matrix into a Riemannian metric, the Rao–Fisher information metric. In this connection, Theorem 1 shows that, when the statistical model $\mathcal{P}$ is a location-scale model which is invariant under the action of a Lie group, information geometry inevitably leads to the study of warped Riemannian metrics.

In addition to stating and proving Theorem 1, the present contribution aims to explore its implications, with regard to the Riemannian geometry of location-scale models, and to lay the foundation for its applications in statistical inference and statistical learning.

To begin, Sect. 4 applies Theorem 1 to two location-scale models, the von Mises–Fisher model of directional data[16, 33], and the Riemannian Gaussian model of data in spaces of covariance matrices [14, 41, 42]. This leads to the analytic expression of the Rao–Fisher information metric of each one of these two models. Precisely, the Rao–Fisher information metric of the von Mises–Fisher model is given in Proposition 2, and that of the Riemannian Gaussian model is given in Proposition 3. The result of Proposition 2 is essentially already contained in [33], (see p. 199), but Proposition 3 is new in the literature.

Finding the analytic expression of the Rao–Fisher information metric, or equivalently of Fisher's information matrix, of a location-scale model $\mathcal{P}$ defined on a high-dimensional non-trivial manifold M, is a very difficult task when attempted by direct calculation. Propositions 2 and 3 show that this task is greatly simplified by Theorem 1. Precisely, if the dimension of M is d, then the dimension of the parameter space $\mathcal{M} = M \times (0, \infty)$ is $d + 1$. Therefore, a priori, the expression of the Rao–Fisher information metric involves $(d + 1)(d + 2)/2$ functions of both parameters $\bar{x}$ and σ of the model $\mathcal{P}$. Instead of so many functions of both $\bar{x}$ and σ, Theorem 1 reduces the expression of the Rao–Fisher information metric to only two functions of σ alone. In the notation of (1a), these two functions are $\alpha(\sigma) = dr/d\sigma$ and $\beta(\sigma) = \beta(r(\sigma))$.

Section 5 builds on Theorem 1 to introduce a general definition of the Mahalanobis distance, applicable to any location-scale model $\mathcal{P}$ defined on a manifold M. Precisely, assume that the model $\mathcal{P}$ verifies the conditions of Theorem 1, so its Rao–Fisher information metric is a warped Riemannian metric. Then, the *generalised Mahalanobis distance* is defined as the Riemannian distance on M which is induced by the restriction of the Rao–Fisher information metric to M. The expression of the generalised Mahalanobis distance is given in Propositions 4 and 5. It was recently applied to visual content classification in [10].

The generalised Mahalanobis distance includes the classical Mahalanobis distance as a special case. Precisely, assume $\mathcal{P}$ is the isotropic normal model defined on $M = \mathbb{R}^d$, so each density $p(x|\,\bar{x}, \sigma)$ is a d-variate normal density with mean $\bar{x}$ and covariance matrix σ^2 times the identity. Then, $\mathcal{P}$ verifies the invariance condition (1b) under the action of the group G of translations in $\mathbb{R}^d$. Therefore, by Theorem 1, its Rao–Fisher information metric is a warped Riemannian metric. This metric is already known in the literature, in terms of the length element [8, 9]

$$ds^2_{\mathcal{M}}(z) = \frac{2d}{\sigma^2}\,d\sigma^2 + \frac{1}{\sigma^2}\,\|d\bar{x}\|^2 \quad \text{for} \quad z = (\bar{x}, \sigma) \in \mathcal{M} \tag{1c}$$

where $\|d\bar{x}\|^2$ denotes the Euclidean length element on $\mathbb{R}^d$. Now, the restriction of this Rao–Fisher information metric to $M = \mathbb{R}^d$ is given by the second term in (1c), which is clearly the Euclidean length element divided by σ^2, and corresponds to the classical Mahalanobis distance [36] — note that (1c) is brought into the form (1a) by letting $r(\sigma) = (2d)^{1/2}\log(\sigma)$.

Section 6 illustrates the results of Sects. 4 and 5, by applying them to the special case of the Riemannian Gaussian model defined on $M = \mathcal{P}_n$, the space of $n \times n$ real covariance matrices. In particular, it gives directly applicable expressions of the Rao–Fisher information metric, and of the generalised Mahalanobis distance, corresponding to this model. As it turns out, the generalised Mahalanobis distance defines a whole new family of affine-invariant distances on $\mathcal{P}_n$, in addition to the usual affine-invariant distance, which was introduced to the information science community in [39].

Section 7 provides the solution of the geodesic equation of any of the Rao–Fisher information metrics arising from Theorem 1. The main result is Proposition 6, which states that the solution of this equation, for given initial conditions, reduces to the solution of a one-dimensional second-order differential equation. This implies that geodesics with given initial conditions can be constructed at a reasonable numerical cost, which opens the possibility, with regard to future work, of a practical implementation of Riemannian line-search optimisation algorithms, which find the extrema of cost functions by searching for them along geodesics [1, 32].

Section 8 is a follow up to Sect. 7, focusing on the construction of Riemannian Brownian motion, instead of the solution of the geodesic equation. The main result is Proposition 7, which states that the construction of Riemannian Brownian motion reduces to the solution of a one-dimensional stochastic differential equation. This implies that Brownian paths can be constructed at a reasonable computational cost, which opens the possibility of a practical implementation of Riemannian stochastic search algorithms, which have the ability to avoid local minima and saddle points [5].

Section 9 is motivated by the special case of the isotropic normal model, with its Rao–Fisher information metric given by (1c). It is well-known that, after a trivial change of coordinates, the length element (1c) coincides with the length element of the Poincaré half-space model of hyperbolic geometry [8, 9]. This means that the parameter space of the isotropic normal model, when equipped with its Rao–Fisher information metric, becomes a space of constant negative curvature, and in particular a Hadamard manifold, a simply-connected complete Riemannian manifold of negative sectional curvature [13, 40]. One cannot but wonder whether other location-scale models also give rise to Hadamard manifolds in this way. This is investigated using Propositions 2 and 6, for the case of the von Mises–Fisher model. A mixture of analytical and numerical computations leads to a surprising new observation: the parameter space of the von Mises–Fisher model of n-dimensional directional data, when equipped with its Rao–Fisher information metric, becomes a Hadamard manifold, for $n = 2, \ldots, 8$. Ongoing research warrants the conjecture that this is true for any value of n, but this is yet to be proved.

Theorem 1, the main new result in the present contribution, has many potential applications, which will be developed in future work. Indeed, this theorem can provide the expression of the Rao–Fisher information metric, or equivalently of Fisher's information matrix, of a location-scale model, even if this model is defined on a high-dimensional non-trivial manifold. By doing so, it unlocks access to the many applications which require this expression, both in statistical inference and in statistical learning. In statistical inference, the expression of the Rao–Fisher information metric allows the computation of the Cramér-Rao lower bound, and the construction of asymptotic chi-squared statistics [30, 48]. In statistical learning, it allows the practical implementation of the natural gradient algorithm, which has the advantages of full reparameterisation invariance, of asymptotic efficiency of its stochastic version, and of the linear rate of convergence of its deterministic version [2, 12, 35].

A first step, towards developing the applications of Theorem 1, was recently taken in [50]. Using the expression of the Rao–Fisher information metric of the Riemannian Gaussian model, this derived and successfully implemented the natural gradient algorithm, for the problem of on-line learning of an unknown probability density, on the space $\mathcal{P}_n$ of $n \times n$ real covariance matrices. Future work will focus on extending this algorithm to more problems of statistical learning, including problems of on-line classification and regression in the space $\mathcal{P}_n$, and in other spaces of covariance matrices. In addition to its use in deriving the natural gradient algorithm for problems of statistical learning, the expression of the Rao–Fisher information metric of the Riemannian Gaussian model can be used in deriving so-called natural evolutionary strategies, for black-box optimisation in spaces of covariance matrices. These would generalize currently existing natural evolutionary strategies, mostly restricted to black-box optimisation in Euclidean space [37, 47].

The following Sect. 2 provides background on warped Riemannian metrics. While the present introduction expressed Riemannian metrics using length elements, Sect. 2 will use scalar products on the tangent space.

2 Background on Warped Riemannian Metrics

Assume M is a complete Riemannian manifold with Riemannian metric Q, and consider the manifold $\mathcal{M} = M \times (0\,, \infty)$. A warped Riemannian metric I on $\mathcal{M}$ is given in the following way [11, 38, 40]. Let α and β be positive functions, defined on $(0\,, \infty)$. Then, for $z = (x, \sigma) \in \mathcal{M}$, let the scalar product I_z on the tangent space $T_z\mathcal{M}$ be defined by

$$I_z(U,U) \;=\; (\alpha(\sigma)\,u_\sigma)^2 \;+\; \beta^2(\sigma)\,Q_x(u,u) \quad U \in T_z\mathcal{M} \tag{2a}$$

where $U \;=\; u_\sigma\,\partial_\sigma \;+\; u$ with $u_\sigma \in \mathbb{R}$ and $u \in T_xM$. The functions α and β are part of the definition of the warped metric I. Once these functions are fixed, it is possible to introduce a change of coordinates $r = r(\sigma)$ which eliminates α from (2a). Precisely, if $dr/d\sigma = \alpha(\sigma)$ then

$$I_z(U,U) \;=\; u_r^2 \;+\; \beta^2(r)\,Q_x(u,u) \tag{2b}$$

where $U \;=\; u_r\,\partial_r \;+\; u$ and $\beta(r) = \beta(\sigma(r))$.

The coordinate r will be called *vertical distance*. This is not a standard terminology, but is suggested as part of the following geometric picture. For $z = (x,\sigma) \in \mathcal{M}$, think of x as a horizontal coordinate, and of σ as a vertical coordinate. Accordingly, the points $z_0 = (x,\sigma_0)$ and $z_1 = (x,\sigma_1)$ lie on the same vertical line. It can be shown from (2b) that the Riemannian distance between z_0 and z_1 is

$$d(z_0,z_1) \;=\; r(\sigma_1) - r(\sigma_0) \quad \text{where} \;\; \sigma_0 \;< \sigma_1 \tag{3a}$$

Precisely, $d(z_0,z_1)$ is the Riemannian distance induced by the warped Riemannian metric I.

The vertical distance r can be used to express a necessary and sufficient condition for completeness of the manifold $\mathcal{M}$, equipped with the warped Riemannian metric I. Namely, $\mathcal{M}$ is a complete Riemannian manifold, if and only if

$$\lim_{\sigma\to\infty} r(\sigma) - r(\sigma_0) = \infty \;\;\text{and}\;\; \lim_{\sigma\to 0} r(\sigma_1) - r(\sigma) = \infty \tag{3b}$$

where σ_0 and σ_1 are arbitrary. This condition is a special case of Lemma 7.2 in [11].

Let $K^{\mathcal{M}}$ and K^M denote the sectional curvatures of $\mathcal{M}$ and M, respectively. The relation between these two is given by the curvature equations of Riemannian geometry [18, 40]. These are,

$$\text{Gauss equation} \;:\; K_z^{\mathcal{M}}(u,v) \;= \beta^{-2}\,K_x^M(u,v) - \left(\beta^{-1}\partial_r\beta\right)^2 \tag{4a}$$

$$\text{Jacobi equation} \;:\; K_z^{\mathcal{M}}(u,\partial_r) \;= -\,\beta^{-1}\,\partial_r^2\beta \tag{4b}$$

for $u,v \in T_xM$. Here, the notations $K_z^{\mathcal{M}}$ and K_x^M mean that $K^{\mathcal{M}}$ is computed at z, and K^M is computed at x, where $z = (x,\sigma)$. Equation (4) are a special case of Lemma 7.4 in [11].

Note, as a corollary of these equations, that $\mathcal{M}$ has negative sectional curvature $K^{\mathcal{M}} < 0$, if M has negative sectional curvature $K^M < 0$ and β is a strictly convex function of r.

Remark 1 Equation (2) contain an abuse of notation. Namely, u denotes a tangent vector to M at x, and a tangent vector to $\mathcal{M}$ at z, at the same time. In the mathematical literature (for example, in [11, 38]), one writes $d\pi_z(U)$ instead of u, using the derivative $d\pi$ of the projection mapping $\pi(z) = x$, and this eliminates any ambiguity. In the present contribution, a deliberate choice is made to use a lighter, though not entirely correct, notation. ■

Remark 2 Consider the proof of Eq. (3). For (3a), let $\gamma(t)$ and $c(t)$ be curves connecting $z_0 = (x, \sigma_0)$ and $z_1 = (x, \sigma_1)$. Assume these are parameterised by $t \in [0, 1]$ as follows,

$$\gamma(t) \; : \; \begin{Bmatrix} x_\gamma(t) = x \quad \text{(constant)} \\ r_\gamma(t) = r(\sigma_0) + t\,(r(\sigma_1) - r(\sigma_0)) \end{Bmatrix} \quad ; \quad c(t) \; : \; \begin{Bmatrix} x_c(t) \\ r_c(t) \end{Bmatrix}$$

If $L(\gamma)$ and $L(c)$ denote the lengths of these curves, then from (2b),

$$L(c) = \int_0^1 \left((\dot{r}_c)^2 + \beta^2\, Q(\dot{x}_c, \dot{x}_c)\right)^{1/2} dt \geq \int_0^1 \dot{r}_c\, dt \; = \; r(\sigma_1) - r(\sigma_0) = L(\gamma)$$

where the dot denotes differentiation with respect to t, and the inequality is strict unless $\gamma = c$. This shows that $\gamma(t)$ is the unique length-minimising geodesic connecting z_0 and z_1. Thus, $d(z_0, z_1) = L(\gamma)$, and this gives (3a). For (3b), note that Lemma 7.2 in [11] states that $\mathcal{M}$ is complete, if and only if $(0\,, \infty)$ is complete when equipped with the distance $d_{(0,\infty)}(\sigma_0\,, \sigma_1) \; = |r(\sigma_1) - r(\sigma_0)|$. However, this is equivalent to (3b). ■

Remark 3 For each $\sigma \in (0\,, \infty)$, the manifold M can be embedded into the manifold $\mathcal{M}$, in the form of the hypersurface $M_\sigma = M \times \{\sigma\}$. Through this embedding, the warped Riemannian metric I of $\mathcal{M}$ induces a Riemannian metric Q^σ on M. By definition, this metric Q^σ is obtained by the restriction of I to the tangent vectors of M_σ [18, 40]. It follows from (2) that

$$Q^\sigma_x(u, u) \; = \; \beta^2(\sigma)\, Q_x(u, u) \tag{5}$$

The induced metric Q^σ will be called an extrinsic metric on M, since it comes from the ambient space $\mathcal{M}$. By (5), the extrinsic metric Q^σ is equal to a scaled version of the Riemannian metric Q of M, with scaling factor $\beta(\sigma)$. ■

3 Connection with Location-Scale Models

This section establishes the connection between warped Riemannian metrics and location-scale models. The main result is Theorem 1, which states that the Rao–Fisher information metric of any location-scale model is a warped Riemannian metric, whenever this model is invariant under the action of some Lie group.

To state this theorem, assume M is an irreducible Riemannian symmetric space, with invariant Riemannian metric Q, under the transitive action of a Lie group of isometries G [22]. Consider a location-scale model $\mathcal{P}$ defined on M,

$$\mathcal{P} = \{\, p(x|\bar{x}, \sigma)\, ;\ \bar{x} \in M\, ,\ \sigma \in (0\, , \infty)\} \tag{6}$$

To each point $z = (\bar{x}, \sigma)$ in the parameter space $\mathcal{M} = M \times (0\, , \infty)$, this model associates a probability density $p(x|\bar{x}, \sigma)$ on M, which has a location parameter $\bar{x}$ and a scale parameter σ. Precisely, $p(x|\bar{x}, \sigma)$ is a probability density with respect to the invariant Riemannian volume element of M.

The condition that the model $\mathcal{P}$ is invariant under the action of the Lie group G means that,

$$p(\, g \cdot x \,|\, g \cdot \bar{x}\, , \sigma) \ = \ p(x|\bar{x}, \sigma) \quad \text{for all } g \in G \tag{7}$$

where $g \cdot x$ denotes the action of $g \in G$ on $x \in M$.

The Rao–Fisher information metric of the location-scale model $\mathcal{P}$ is a Riemannian metric I on the parameter space $\mathcal{M}$ of this model [3]. It is defined as follows, for $z = (\bar{x}, \sigma) \in \mathcal{M}$ and $U \in T_z\mathcal{M}$,

$$I_z(U, U) \ = \ \mathbb{E}_z\left[(\, d\ell(z)\, U\,)^2\right] \tag{8}$$

where $\mathbb{E}_z$ denotes expectation with respect to the probability density $p(x|z) = p(x|\bar{x}, \sigma)$, and $\ell(z)$ is the log-likelihood function, given by $\ell(z)(x) = \log p(x|z)$.

In the following statement, $\nabla_{\bar{x}}\, \ell(z)$ denotes the Riemannian gradient of $\ell(z)$, taken with respect to $\bar{x} \in M$, while the value of σ is fixed.

Theorem 1 *If condition (7) is verified, then the Rao–Fisher information metric I of (8) is a warped Riemannian metric given by (2a), where*

$$\alpha^2(\sigma) \ = \mathbb{E}_z\, (\partial_\sigma \ell(z))^2 \quad \beta^2(\sigma) \ = \ \mathbb{E}_z\, Q\, (\nabla_{\bar{x}}\, \ell(z)\, ,\, \nabla_{\bar{x}}\, \ell(z)\,)\, /\dim M \tag{9}$$

The expectations appearing in (9) do not depend on $\bar{x}$, so $\alpha(\sigma)$ and $\beta(\sigma)$ are well-defined functions of σ.

Remark 4 Recall the definition of an irreducible Riemannian symmetric space [22]. A Riemannian manifold M, whose group of isometries is denoted G, is called a Riemannian symmetric space, if for each $\bar{x} \in M$ there exists an isometry $s_{\bar{x}} \in G$, whose effect is to fix $\bar{x}$ and reverse the geodesic curves passing through $\bar{x}$. Further, M is called irreducible if it verifies the following condition. Let $K_{\bar{x}}$ be the subgroup of G which consists of those elements k such that $k \cdot \bar{x} = \bar{x}$. For each $k \in K_{\bar{x}}$, its derivative $dk_{\bar{x}}$ is a linear mapping of $T_{\bar{x}}M$. The mapping $k \mapsto dk_{\bar{x}}$ is a representation of $K_{\bar{x}}$ in $T_{\bar{x}}M$, called the isotropy representation, and M is called an irreducible Riemannian symmetric space if the isotropy representation is irreducible. That is, if the isotropy representation has no invariant subspaces in $T_{\bar{x}}M$, except $\{0\}$ and $T_{\bar{x}}M$. Irreducible Riemannian symmetric spaces are classified in [22] (Table I, p. 346 and Table II, p. 354). They include spaces of constant curvature, such as spheres and hyperbolic spaces, as well as spaces of positive definite matrices which have determinant equal to 1, and whose entries are real or complex numbers, or quaternions. ■

Remark 5 It is sometimes possible to apply Theorem 1, even when the Riemannian symmetric space M is not irreducible. For example, in Sect. 4.2, Theorem 1 will be used to find the expression of the Rao–Fisher information metric of the Riemannian Gaussian model [14, 41, 42]. For this model, when M is not irreducible, the Rao–Fisher information metric turns out to be a so-called multiply-warped Riemannian metric, rather than a warped Riemannian metric. The concrete case of $M = \mathcal{P}_n$, the space of $n \times n$ real covariance matrices, is detailed in Sect. 6. ■

Proof of Theorem 1 Recall the expression $U = u_\sigma \, \partial_\sigma + u$ with $u_\sigma \in \mathbb{R}$ and $u \in T_{\bar{x}}M$. Since the Rao–Fisher information metric I is bilinear and symmetric,

$$I_z(U, U) = I_z(\partial_\sigma, \partial_\sigma)\, u_\sigma^2 + 2I_z(\partial_\sigma, u)\, u_\sigma + I_z(u, u)$$

It is possible to show the following,

$$I_z(\partial_\sigma, \partial_\sigma) = \alpha^2(\sigma) \tag{10a}$$

$$I_z(\partial_\sigma, u) = 0 \tag{10b}$$

$$I_z(u, u) = \beta^2(\sigma)\, Q_{\bar{x}}(u, u) \tag{10c}$$

where $\alpha^2(\sigma)$ and $\beta^2(\sigma)$ are given by (9).
Proof of (10a): this is immediate from (8). Indeed,

$$I_z(\partial_\sigma, \partial_\sigma) = \mathbb{E}_z \left[(\, d\ell(z)\, \partial_\sigma \,)^2 \right] = \mathbb{E}_z \left[(\partial_\sigma \ell(z))^2 \right]$$

Proof of (10b): this is carried out in Appendix A, using the fact that M is a Riemannian symmetric space.
Proof of (10c): this is carried out in Appendix A, using the fact that M is irreducible, by an application of Schur's lemma from the theory of group representations [27].

The fact that the expectations appearing in (9) do not depend on $\bar{x}$ is also proved in Appendix A. Throughout the proof of the theorem, the following identity is used, which is equivalent to condition (7). For any real-valued function f on M,

$$\mathbb{E}_{g \cdot z}\, f = \mathbb{E}_z\, (f \circ g) \tag{11}$$

Here, $g \cdot z = (g \cdot \bar{x}, \sigma)$, and $f \circ g$ is the function $(f \circ g)(x) = f(g \cdot x)$, for $g \in G$ and $z = (\bar{x}, \sigma)$. ■

4 Examples: von Mises–Fisher and Riemannian Gaussian

This section applies Theorem 1 to finding the expression of the Rao–Fisher information metric of two location-scale models. These are the von Mises–Fisher model, which is widely used in the study of directional data [16, 33], and the Riemannian

Gaussian model, recently introduced in the study of data with values in spaces of covariance matrices [14, 41, 42].

The application of Theorem 1 to these two models is encapsulated in the following Proposition 1. Precisely, both of these models are of a common exponential form, which can be described as follows. Let M be an irreducible Riemannian symmetric space, as in Sect. 3. In the notation of (6), consider a location-scale model $\mathcal{P}$ defined on M, by

$$p(x|\bar{x},\sigma) \;=\; \exp\left[\,\eta(\sigma)\,D(x,\bar{x}) \;-\; \psi(\eta(\sigma))\right] \tag{12a}$$

where $\eta(\sigma)$ is a certain parameter, to be called the natural parameter, and where $D : M \times M \to \mathbb{R}$ verifies the condition,

$$D(\,g\cdot x\,,\;\, g\cdot\bar{x}) \;=\; D(x\,,\;\bar{x}) \quad \text{for all } g \in G \tag{12b}$$

There is no need to assume that the function D is positive.

Proposition 1 *If the model $\mathcal{P}$ is given by Eq. (12), then the Rao–Fisher information metric I of this model is a warped Riemannian metric,*

$$I_z(U,U) \;=\; \psi''(\eta)\,u^2_\eta \;+\; \beta^2(\eta)\;Q_{\bar{x}}(u,u) \tag{13a}$$

where $U \;=\; u_\eta\,\partial_\eta \;+\; u$ with $u_\eta \in \mathbb{R}$ and $u \in T_{\bar{x}}M$, and where

$$\beta^2(\eta) \;=\; \eta^2\;\mathbb{E}_z\;Q\,(\nabla_{\bar{x}}\,D,\nabla_{\bar{x}}\,D\,)\,/\dim M \tag{13b}$$

Proof For a model $\mathcal{P}$ defined by (12a), condition (12b) is equivalent to condition (7). Therefore, by application of Theorem 1, it follows that I is a warped Riemannian metric, of the form (2a),

$$I_z(U,U) \;=\; (\alpha(\sigma)\,u_\sigma)^2 \;+\; \beta^2(\sigma)\;Q_{\bar{x}}(u,u) \tag{14a}$$

where $\alpha^2(\sigma)$ and $\beta^2(\sigma)$ are given by (9). Consider the first term in (14a). By the change of coordinates formula [29], $u_\sigma = \sigma'(\eta)\,u_\eta$, where the prime denotes differentiation with respect to η. It follows that

$$(\alpha(\sigma)\,u_\sigma)^2 \;=\; \alpha^2(\sigma)\,\left(\sigma'(\eta)\right)^2\;u^2_\eta \tag{14b}$$

However, by (9),

$$\alpha^2(\sigma)\,\left(\sigma'(\eta)\right)^2 \;=\; \mathbb{E}_z\,\left(\,\partial_\sigma\ell(z)\,\sigma'(\eta)\,\right)^2 \;=\; \mathbb{E}_z\,\left(\,\partial_\eta\ell(z)\,\right)^2 \tag{14c}$$

Here, the log-likelihood $\ell(z)$ is found from (12a),

$$\ell(z)(x) = \eta(\sigma)\,D(x,\bar{x}) \;-\; \psi(\eta(\sigma)) \tag{14d}$$

Therefore, the last expression in (14c) is

$$\mathbb{E}_z \left(\partial_\eta \ell(z) \right)^2 \;=\; -\mathbb{E}_z \, \partial^2_\eta \ell(z) \;=\; \psi''(\eta) \tag{14e}$$

where the first equality is the same as in [3], (see p. 28). Now, (14b) and (14c) imply

$$(\alpha(\sigma)\, u_\sigma)^2 \;=\; \psi''(\eta)\, u^2_\eta \tag{14f}$$

Replacing this in (14a), and writing $\beta(\eta) = \beta(\sigma(\eta))$, gives

$$I_z(U, U) \;=\; \psi''(\eta)\, u^2_\eta \;+\; \beta^2(\eta)\; Q_{\bar{x}}(u, u) \tag{14g}$$

which is the same as (13a). To prove the proposition, it remains to show that $\beta^2(\eta)$ is given by (13b). To do so, note that it follows from (14d),

$$\nabla_{\bar{x}}\, \ell(z) = \nabla_{\bar{x}} \left[\, \eta(\sigma)\, D(x, \bar{x}) \;-\; \psi(\eta(\sigma))\right] = \eta(\sigma)\, \nabla_{\bar{x}}\, D(x, \bar{x})$$

Replacing this in (9) gives,

$$\beta^2(\eta) \;=\; \mathbb{E}_z\, Q\left(\nabla_{\bar{x}}\, \ell(z)\,,\, \nabla_{\bar{x}}\, \ell(z)\,\right)/\dim M \;\; = \eta^2 \;\; \mathbb{E}_z\, Q\left(\nabla_{\bar{x}}\, D, \nabla_{\bar{x}}\, D\,\right)/\dim M$$

and this is the same as (13b). ■

4.1 The von Mises–Fisher Model

The von Mises–Fisher model is a mainstay of directional statistics [16, 33]. In the notation of (12), this model corresponds to $M = S^{n-1}$, the unit sphere in $\mathbb{R}^n$, and to $G = O(n)$, the group of $n \times n$ real orthogonal matrices, which acts on $\mathbb{R}^n$ by rotations. Then, the expressions appearing in (12a) are

$$D(x, \bar{x}) = \langle x, \bar{x} \rangle \quad \psi(\eta) = \nu \log(2\pi) \;+\; \log\left(\eta^{1-\nu} I_{\nu-1}(\eta)\right) \tag{15}$$

for $\eta \in [\,0\,, \infty)$. Here, $\langle x, \bar{x} \rangle$ denotes the Euclidean scalar product in $\mathbb{R}^n$, so that condition (12b) is clearly verified, and $I_{\nu-1}$ denotes the modified Bessel function of order $\nu - 1$, where $\nu = n/2$. The natural parameter η and the scale parameter σ should be considered identical, in the sense that $\eta(\sigma) = \sigma$, as long as $\sigma \in (0\,, \infty)$. However, η takes on the additional value $\eta = 0$, which requires a special treatment.

Remark 6 The parameter space of the von Mises–Fisher model will be identified with the space $\mathbb{R}^n$. This is done by mapping each couple $(\bar{x}, \eta)$ to the point $z = \eta\,\bar{x}$ in $\mathbb{R}^n$. This mapping defines a diffeomorphism from the set of couples $(\bar{x}, \eta)$ where $\eta \in (0\,, \infty)$, to the open subset $\mathbb{R}^n - \{0\} \subset \mathbb{R}^n$. On the other hand, it maps all couples $(\bar{x}, \eta = 0)$, to the same point $z = 0 \in \mathbb{R}^n$. Note that each couple $(\bar{x}, \eta)$ where

$\eta \in (0\,,\infty)$ defines a distinct von Mises–Fisher distribution, which is a unimodal distribution with its mode at $\bar{x}$. On the other hand, all couples $(\bar{x}, \eta = 0)$ define the same von Mises–Fisher distribution, which is the uniform distribution on S^{n-1}. Therefore, it is correct to map all of these couples to the same point $z = 0$. ■

Proposition 1 will only provide the Rao–Fisher information metric of the von Mises–Fisher model on the subset $\mathbb{R}^n - \{0\}$ of the parameter space $\mathbb{R}^n$. Therefore, it is necessary to verify that this metric has a well-defined limit at the point $z = 0$. This is carried out in Proposition 2 below. In the statement of this proposition, a tangent vector $U \in T_z\mathbb{R}^n$, at a point $z \in \mathbb{R}^n - \{0\}$, is written in the form

$$U = u_\eta\,\bar{x} + \eta\,u \tag{16a}$$

where $z = \eta\,\bar{x}$, and where $u_\eta \in \mathbb{R}$ and $u \in T_{\bar{x}}S^{n-1}$. Here, u_η and u are unique, for a given U. Precisely,

$$u_\eta \;=\; \langle U, \bar{x}\rangle \quad u \;=\; \frac{1}{\eta}\,[\,U \;-\; \langle U, \bar{x}\rangle\,\bar{x}\,] \tag{16b}$$

as follows since $\bar{x}$ and u are orthogonal.

Proposition 2 *The Rao–Fisher information metric I of the von Mises–Fisher model is a well-defined Riemannian metric on the parameter space $\mathbb{R}^n$. On $\mathbb{R}^n - \{0\}$, it is a warped Riemannian metric of the form (13a), where*

$$\psi''(\eta) = \frac{1}{n} + \frac{n-1}{n}\,\frac{I_{\nu+1}(\eta)}{I_{\nu-1}(\eta)} - \frac{I^2_{\nu}(\eta)}{I^2_{\nu-1}(\eta)} \tag{17a}$$

$$\beta^2(\eta) = \frac{\eta^2}{n}\left(1 + \frac{I_{\nu+1}(\eta)}{I_{\nu-1}(\eta)}\right) \tag{17b}$$

and it extends smoothly to the value,

$$I_0(U, U) \;=\; \frac{1}{n}\,\|U\|^2 \tag{17c}$$

at the point $z = 0$. Here, $\|\cdot\|$ denotes the Euclidean norm.

Proof The Rao–Fisher information metric I on $\mathbb{R}^n - \{0\}$ is given by Proposition 1. This proposition applies because $M = S^{n-1}$ is an irreducible Riemannian symmetric space [22] (Table II, p. 354). Accordingly, for any point $z \in \mathbb{R}^n - \{0\}$, the metric I_z is given by (13a). Formulae (17) are proved as follows.
Proof of (17a): this is carried out in Appendix B, using the derivative and recurrence relations of modified Bessel functions [46].
Proof of (17b): this follows from (13b) and (15). By (15),

$$\nabla_{\bar{x}}\, D(x, \bar{x}) \;=\; x - \langle x, \bar{x}\rangle\,\bar{x}$$

which is just the orthogonal projection of x onto the tangent space $T_{\bar{x}}S^{n-1}$. Replacing in (13b) gives

$$\beta^2(\eta) \;=\; \frac{\eta^2}{n-1}\,\mathbb{E}_z\,\|\nabla_{\bar{x}}\,D\|^2 \;=\; \frac{\eta^2}{n-1}\,\mathbb{E}_z\left(1-\langle x,\bar{x}\rangle^2\right) \tag{18a}$$

Here, in the first equality, $n-1$ appears because $\dim S^{n-1} = n-1$. The second equality follows by Pythagoras' theorem,

$$\|\,x-\langle x,\bar{x}\rangle\,\bar{x}\,\|^2 \;=\; \|\,x\,\|^2 - \|\,\langle x,\bar{x}\rangle\,\bar{x}\,\|^2 \;=\; 1-\langle x,\bar{x}\rangle^2$$

since x and $\bar{x}$ belong to the unit sphere S^{n-1}. Formula (17b) is derived from (18a) in Appendix B, using the derivative and recurrence relations of modified Bessel functions [46].
Proof of (17c): for any point $z \in \mathbb{R}^n - \{0\}$, the metric I_z is given by (13a). This reads

$$I_z(U,U) \;=\; \psi''(\eta)\,u_\eta^2 \;+\; \beta^2(\eta)\,\|u\|^2 \tag{18b}$$

Consider the limit of this expression at the point $z = 0$. In (17a) and (17b), this corresponds to the limit at $\eta = 0$. This can be evaluated using the power series development of modified Bessel functions [46]. When replaced in (17a) and (17b), this gives the following developments,

$$\psi''(\eta) = \frac{1}{n} - \quad \frac{12}{n^2(n+2)}\left(\frac{\eta}{2}\right)^2 + O\left(\eta^4\right)$$

$$\beta^2(\eta) = \frac{4}{n}\left(\frac{\eta}{2}\right)^2 + O\left(\eta^4\right)$$

which immediately imply that

$$\lim_{\eta\to 0}\;\psi''(\eta) = \frac{1}{n} \tag{18c}$$

$$\lim_{\eta\to 0}\;\beta^2(\eta) = 0 \tag{18d}$$

Replacing (18c) and (18d) in (18b) gives,

$$\lim_{z\to 0}\;I_z(U,U) = \frac{1}{n}\,u_\eta^2 \tag{18e}$$

Note that, from (16a),

$$\|U\|^2 = u_\eta^2 \;+\; \eta^2\,\|u\|^2$$

by Pythagoras' theorem, since $\bar{x}$ and u are orthogonal. At the point $z = 0$, one has $\eta = 0$, so that

$$\|U\|^2_{z=0} = u^2_\eta$$

This shows that (18e) is the same as

$$\lim_{z\to 0} \; I_z(U,U) = \frac{1}{n}\,\|U\|^2 \tag{18f}$$

This limit does not depend on the path along which z tends to $z=0$. Therefore, I_z extends smoothly to I_0, which is given by (17c), at the point $z=0$. This shows that I is a well-defined Riemannian metric throughout the parameter space $\mathbb{R}^n$. ■

4.2 *The Riemannian Gaussian Model*

The Riemannian Gaussian model was recently introduced as a means of describing unimodal populations of covariance matrices [14, 41, 42]. This model can be defined on any Riemannian symmetric space of non-positive sectional curvature. Let M be such a symmetric space and denote G its group of isometries. Then, the expressions appearing in (12a) are

$$D(x,\bar{x}) = d^2(x,\bar{x}) \quad \eta(\sigma) = -\frac{1}{2\sigma^2} \tag{19}$$

where $d(x,\bar{x})$ denotes the Riemannian distance in M, and condition (12b) is verified since each isometry $g \in G$ preserves this Riemannian distance. The function $\psi(\eta)$ is a strictly convex function of $\eta \in (-\infty\,, 0)$, which can be expressed by means of a multiple integral [42], (see Proposition 1 in this reference). Precisely, $\psi(\eta)$ is the cumulant generating function of the squared Riemannian distance $d^2(x,\bar{x})$.

Proposition 1 cannot be applied directly to the Riemannian Gaussian model (19). This is because, in most cases of interest, the Riemannian symmetric space M is not irreducible. In such cases, before applying Proposition 1, it is necessary to introduce the De Rham decomposition theorem [22, 40].

Remark 7 Assume the Riemannian symmetric space M is moreover simply-connected. Then, the De Rham decomposition theorem implies that M is a Riemannian product of irreducible Riemannian symmetric spaces [22] (Proposition 5.5, p. 310). Precisely, $M = M_1 \times \cdots \times M_r$ where each M_q is an irreducible Riemannian symmetric space, and the Riemannian metric and distance of M can be expressed as follows,

$$Q(u,u) \;=\; \sum_{q=1}^{r} Q(u_q\,, u_q) \tag{20a}$$

$$d^2(x,y) \;=\; \sum_{q=1}^{r} d^2(x_q\,, y_q) \tag{20b}$$

where $x, y \in M$ are written $x = (x_1, \ldots, x_r)$ and $y = (y_1, \ldots, y_r)$ with $x_q, y_q \in M_q$, and where $u \in T_x M$ is written $u = u_1 + \cdots + u_r$ with $u_q \in T_{x_q} M_q$ naturally identified with an element of $T_x M$. Since M has non-positive sectional curvature, each M_q is either a Euclidean space, or a so-called space of non-compact type, having negative sectional curvature [22]. A concrete example of the De Rham decomposition is treated in Sect. 6, where $M = \mathcal{P}_n$ is the space of $n \times n$ real covariance matrices. ■

The following Proposition 3 gives the Rao–Fisher information metric of the Riemannian Gaussian model. Since this model is, in general, defined on a Riemannian symmetric space M which is not irreducible, the Rao–Fisher information metric turns out to be a multiply-warped Riemannian metric, rather than a warped Riemannian metric.

Remark 8 In the notation of (20), a multiply-warped Riemannian metric I is a Riemannian metric defined on $\mathcal{M} = M \times (0, \infty)$, in the following way [17, 45]

$$I_z(U, U) = (\alpha(\sigma)\, u_\sigma)^2 + \sum_{q=1}^{r} \beta_q^2(\sigma)\, Q_{\bar{x}}(u_q, u_q) \tag{21}$$

for $z = (\bar{x}, \sigma)$ and $U \in T_z\mathcal{M}$, where $U = u_\sigma\, \partial_\sigma + u$ with $u_\sigma \in \mathbb{R}$ and $u \in T_{\bar{x}} M$. Here, the functions α and β_q are positive functions defined on $(0, \infty)$. ■

Proposition 3 *The Rao–Fisher information metric of the Riemannian Gaussian model is a multiply-warped Riemannian metric. In terms of $\bar{x} \in M$ and $\eta = -1/2\sigma^2$, this metric has the following expression*

$$I_z(U, U) = \psi''(\eta)\, u_\eta^2 + \sum_{q=1}^{r} \left(4\eta^2 \psi_q'(\eta)/\dim M_q\right) Q_{\bar{x}}(u_q, u_q) \tag{22}$$

where $U = u_\eta\, \partial_\eta + u$, and where $\psi_q(\eta)$ is the cumulant generating function of the squared Riemannian distance $d^2(x_q, \bar{x}_q)$.

Proof Assume first that the Riemannian symmetric space M is irreducible, so that Proposition 1 applies directly, and the Rao–Fisher information metric is given by (13a),

$$I_z(U, U) = \psi''(\eta)\, u_\eta^2 + \beta^2(\eta)\, Q_{\bar{x}}(u, u) \tag{23a}$$

To obtain $\beta^2(\eta)$, replace into (13b) the fact that

$$\nabla_{\bar{x}} D(x, \bar{x}) = -2 \exp_{\bar{x}}^{-1}(x) \quad Q\,(\nabla_{\bar{x}} D, \nabla_{\bar{x}} D) = 4\, d^2(x, \bar{x})$$

where exp denotes the Riemannian exponential mapping, corresponding to the Riemannian metric Q of M [13], (see p. 407). It then follows from (13b) that,

$$\beta^2(\eta) = 4\eta^2\, \mathbb{E}_z\, d^2(x, \bar{x})/\dim M = 4\eta^2\, \psi'(\eta)/\dim M \tag{23b}$$

where the second equality holds since $\psi(\eta)$ is the cumulant generating function (log-moment generating function) of $d^2(x, \bar{x})$. From (23a) and (23b),

$$I_z(U,U) \;=\; \psi''(\eta)\,u_\eta^2 \,+\, \left(4\eta^2\psi'(\eta)/\dim M\right)\; Q_{\bar{x}}(u,u) \tag{23c}$$

which is the same as (22) with $r = 1$. This proves the proposition in the special case where M is irreducible. For the general case where M is not irreducible, write $U = u_\eta\,\partial_\eta + u$, with $u = u_1 + \cdots + u_r$ as in Remark 7. It is possible to prove that,

$$q \neq p \;: I_z(u_p\,, u_q) = 0 \tag{24a}$$

$$u = u_p \;: I_z(U,U) \;=\; \psi''(\eta)\,u_\eta^2 + \left(4\eta^2\psi_p'(\eta)/\dim M_p\right) Q_{\bar{x}}(u_p\,, u_p) \tag{24b}$$

Then, since the Rao–Fisher information metric I is bilinear and symmetric, (22) follows immediately, and the proposition is proved in the general case.

Proof of identities (24): this is carried out using the following properties (25). Note first that the probability density function of the Riemannian Gaussian model is given by (12a) and (19),

$$p(x|\bar{x},\sigma) \;=\; \exp\left[\,\eta(\sigma)\,d^2(x,\bar{x}) \;-\; \psi(\eta(\sigma))\right] \tag{25a}$$

By substituting (20b) in this expression, it is seen that

$$p(x|\bar{x},\sigma) \;=\; \prod_{q=1}^{r} \exp\left[\,\eta(\sigma)\,d^2(x_q\,,\bar{x}_q) \;-\; \psi_q(\eta(\sigma))\right] \;=\; \prod_{q=1}^{r} p(x_q|\bar{x}_q\,,\sigma) \tag{25b}$$

where $\psi_q(\eta)$ is the cumulant generating function of $d^2(x_q\,, \bar{x}_q)$, as stated after (22). The last equality shows that $(x_q\,; q = 1, \ldots, r)$ are independent, and that each x_q has a Riemannian Gaussian density on the irreducible Riemannian symmetric space M_q, with parameters $z_q = (\bar{x}_q\,, \sigma)$. Now, identities (24) can be obtained from definition (8) of the Rao–Fisher information metric. To apply this definition, note from (25b), that the log-likelihood function $\ell(z)$ can be written,

$$\ell(z)(x) \;=\; \log p(x|z) \;=\; \sum_{q=1}^{r} \ell(z_q)(x_q) \quad \text{where} \;\; \ell(z_q)(x_q) = \log p(x_q|z_q) \tag{25c}$$

Proof of (24a): recall the polarisation identity, from elementary linear algebra [28], (see p. 29),

$$I_z(u_p\,, u_q) \;=\; \frac{1}{4}\,I_z(u_p + u_q\,, u_p + u_q) - \frac{1}{4}\,I_z(u_p - u_q\,, u_p - u_q)$$

By replacing (8) into this identity, it can be seen that,

$$I_z(u_p\,,u_q) \;=\; \mathbb{E}_z\left(\left(d\ell(z)\,u_p\right)\left(d\ell(z)\,u_q\right)\right) \tag{26}$$

Using (25c), it is then possible to write

$$I_z(u_p\,,u_q) \;=\; \mathbb{E}_z\left(\left(d\ell(z_p)\,u_p\right)\left(d\ell(z_q)\,u_q\right)\right) \;=\; \mathbb{E}_{z_p}\left(d\ell(z_p)\,u_p\right)\,\mathbb{E}_{z_q}\left(d\ell(z_q)\,u_q\right)$$

Here, the first equality follows from (26), since $u_p \in T_{\bar{x}_p}M_p$ and $u_q \in T_{\bar{x}_q}M_q$, and the second equality holds since x_p and x_q are independent. Now, each one of the two expectations appearing on the right-hand side is equal to zero, since the expectation of the derivative of the log-likelihood must be zero [3], (see p. 28). This shows that (24a) holds. ■

Proof of (24b): the condition $u = u_p$ implies $U = u_\eta\,\partial_\eta \,+\, u_p$. Replacing this in (8), it follows using (25c),

$$I_z(U,U) \;=\; \mathbb{E}_z\left(\sum_{q=1}^{r} d\ell(z_q)\,U\right)^2 = \mathbb{E}_z\left(\sum_{q=1}^{r} u_\eta\,\partial_\eta\ell(z_q) \,+\, d\ell(z_p)\,u_p\right)^2 \tag{27a}$$

where the second equality holds since $u_p \in T_{\bar{x}_p}M_p$. Since the x_q are independent, it is clear from (25c) that the $\ell(z_q)$ are independent. Accordingly, by expanding the right-hand side of (27a),

$$I_z(U,U) = \sum_{q\neq p} u_\eta^2\,\mathbb{E}_{z_q}\left(\partial_\eta\ell(z_q)\right)^2 + \mathbb{E}_{z_p}\left(d\ell(z_p)\,U\right)^2 \tag{27b}$$

Applying (14e) from the proof of Proposition 1 to each term in the sum over $q \neq p$, it follows that

$$I_z(U,U) = \sum_{q\neq p} \psi_q''(\eta)\,u_\eta^2 \,+\, \mathbb{E}_{z_p}\left(d\ell(z_p)\,U\,\right)^2 \tag{27c}$$

By (8), the expectation appearing in the second term is given by the Rao–Fisher information metric of the Riemannian Gaussian model on the irreducible Riemannian symmetric space M_p . This can be replaced from (23c), so that

$$\begin{aligned} I_z(U,U) &= \textstyle\sum_{q\neq p}\psi_q''(\eta)\,u_\eta^2 \,+\, \psi_p''(\eta)\,u_\eta^2 \,+\, \left(4\eta^2\psi_p'(\eta)/\dim M_p\right)\,Q_{\bar{x}}(u_p\,,u_p) \\ &= \textstyle\sum_q \psi_q''(\eta)\,u_\eta^2 \qquad\qquad\qquad\;\; +\, \left(4\eta^2\psi_p'(\eta)/\dim M_p\right)\,Q_{\bar{x}}(u_p\,,u_p) \end{aligned}$$

This immediately yields (24b), upon noting from (25a) and (25b) that $\psi(\eta) = \sum_q \psi_q(\eta)$. ■

Now, since identities (24) have been proved, (22) follows from the fact that the Rao–Fisher information metric I is bilinear and symmetric. ■

5 The Generalised Mahalanobis Distance

This section builds on Remark 3, made at the end of Sect. 2, in order to generalise the definition of the classical Mahalanobis distance, to the context of a location-scale model $\mathcal{P}$ defined on a Riemannian symmetric space M.

To begin, assume that, as in Theorem 1, the Riemannian symmetric space M is irreducible and the location-scale model $\mathcal{P}$ verifies condition (7). Then, according to Theorem 1, the Rao–Fisher information metric I of the model $\mathcal{P}$ is a warped Riemannian metric on the parameter space $\mathcal{M}$.

Recall from Remark 3 that this warped Riemannian metric I induces an extrinsic Riemannian metric Q^{σ} on M, for each $\sigma \in (0\,,\infty)$. The *generalised Mahalanobis distance* is defined to be the Riemannian distance on M which is induced by the extrinsic Riemannian metric Q^{σ}. The generalised Mahalanobis distance between $\bar{x}$ and $\bar{y}$ in M is denoted $d(\bar{x},\bar{y}\,|\sigma)$. It is given by the following Proposition 4. The proof of Proposition 4 is omitted, since it is elementary.

Proposition 4 *The generalised Mahalanobis distance $d(\bar{x},\bar{y}\,|\sigma)$ between $\bar{x}$ and $\bar{y}$ in M is given by*

$$d(\bar{x},\bar{y}\,|\sigma) \;=\; \beta(\sigma)\,d(\bar{x},\bar{y}) \tag{28}$$

where the function $\beta(\sigma)$ is given by (9), and where $d(\bar{x},\bar{y})$ denotes the Riemannian distance in M.

Remark 9 The generalised Mahalanobis distance (28) reduces to the classical Mahalanobis distance, when $\mathcal{P}$ is the isotropic normal model on $M = \mathbb{R}^d$. In this case, the Rao–Fisher metric I is given by (1c) in the introduction, so that $\beta(\sigma) = 1/\sigma$. Replacing this in (28) yields

$$d(\bar{x},\bar{y}\,|\sigma) \;=\; \frac{1}{\sigma}\,\|\bar{x}-\bar{y}\| \tag{29}$$

where $\|\bar{x}-\bar{y}\|$ is the Euclidean distance in $M = \mathbb{R}^d$. Now, (29) is the classical Mahalanobis distance [36]. ■

Expression (28) of the generalised Mahalanobis distance is valid only under the assumption that the Riemannian symmetric space M is irreducible. This assumption does not hold, when the model $\mathcal{P}$ is the Riemannian Gaussian model studied in Sect. 4.2. For this model, an alternative expression of the generalised Mahalanobis distance is given in Proposition 5 below.

As in Sect. 4.2, let $\mathcal{P}$ be the Riemannian Gaussian model on a Riemannian symmetric space M, where M is simply-connected and has non-positive sectional curvature. Proposition 3 states that the Rao–Fisher information metric I of the model $\mathcal{P}$ is a multiply-warped Riemannian metric on the parameter space $\mathcal{M}$. For each $\sigma \in (0\,,\infty)$, this multiply-warped Riemannian metric I induces an extrinsic Riemannian metric Q^{σ} on M. Precisely, Q^{σ} can be obtained from (22) of Proposition 3,

$$Q^{\sigma}_{\bar{x}}(u,u) \;=\; \sum_{q=1}^{r} \beta^2_q(\sigma)\, Q_{\bar{x}}(u_q\,,u_q) \quad \beta^2_q(\sigma) \;=\; 4\eta^2\psi^{\prime}_q(\eta)/\dim M_q \tag{30}$$

The generalised Mahalanobis distance $d(\bar{x},\,\bar{y}\,|\sigma)$ is the Riemannian distance between $\bar{x}$ and $\bar{y}$ in M, induced by this extrinsic Riemannian metric Q^{σ}.

Proposition 5 *When $\mathcal{P}$ is the Riemannian Gaussian model, the generalised Mahalanobis distance $d(\bar{x},\,\bar{y}\,|\sigma)$ between $\bar{x}$ and $\bar{y}$ in M is given by*

$$d^2(\bar{x},\bar{y}\,|\sigma) \;=\; \sum_{q=1}^{r} \beta^2_q(\sigma)\, d^{\,2}(\bar{x}_q\,,\bar{y}_q) \tag{31}$$

where the notation is that of (20b).

Proof The proof hinges on the fact that the extrinsic Riemannian metric Q^{σ} of (30) is an invariant Riemannian metric on M. In other words, if G is the group of isometries of M, then

$$Q^{\sigma}_{g\cdot\bar{x}}(dg_{\bar{x}}\,u, dg_{\bar{x}}\,u) = Q^{\sigma}_{\bar{x}}(u,u) \quad \text{for all } g\in G \tag{32a}$$

where $dg_{\bar{x}}$ is the derivative of the isometry g at the point $\bar{x}$. The proof of (32a) is not detailed here. It follows since the Riemannian metric Q is also an invariant Riemannian metric on M, so that Q also verifies (32a), and since Q^{σ} is related to Q by (30). A general result in [22] (Corollary 4.3, p. 182), states that all invariant Riemannian metrics on M have the same geodesics. In particular, the metrics Q^{σ} and Q have the same geodesics, and therefore the same Riemannian exponential mapping exp. To find the generalised Mahalanobis distance between $\bar{x}$ and $\bar{y}$ in M, let $u = \exp^{-1}_{\bar{x}}(\bar{y})$, and note that

$$d^2(\bar{x},\bar{y}\,|\sigma) \;=\; Q^{\sigma}_{\bar{x}}(u,u) \;=\; \sum_{q=1}^{r} \beta^2_q(\sigma)\, Q_{\bar{x}}(u_q\,,u_q) \tag{32b}$$

where the second equality follows from (30). Now, to prove (31) it is enough to prove that

$$Q_{\bar{x}}(u_q\,,u_q) \;=\; d^{\,2}(\bar{x}_q\,,\bar{y}_q) \tag{32c}$$

Indeed, (31) is then obtained by replacing (32c) into (32b). The proof of (32c) follows by writing, as in (32b),

$$d^2(\bar{x},\bar{y}) \;=\; Q_{\bar{x}}(u,u) \;=\; \sum_{q=1}^{r} Q_{\bar{x}}(u_q\,,u_q) \;=\; \sum_{q=1}^{r} d^{\,2}(\bar{x}_q\,,\bar{y}_q) \tag{32d}$$

where the second equality follows from (20a), and the third equality follows from (20b). Since (32d) is an identity which holds for arbitrary $\bar{x} = (\bar{x}_1\,,\ldots,\,\bar{x}_r)$ and $\bar{y} = (\bar{y}_1\,,\ldots,\,\bar{y}_r)$, it follows that (32c) must hold true, as required. ■

Remark 10 The generalised Mahalanobis distance, whether given by (28) or by (31), has interesting properties, both geometric and statistical. From the geometric viewpoint, it is an invariant Riemannian distance on M,

$$d(g \cdot \bar{x}, g \cdot \bar{y}\,|\sigma) \;=\; d(\bar{x}, \bar{y}\,|\sigma) \quad \text{for all } g \in G \tag{33}$$

while, from the statistical viewpoint, just like the classical Mahalanobis distance, it can be used to build asymptotic chi-squared statistics, for hypothesis testing or classification [10]. This statistical aspect of the generalised Mahalanobis distance will be developed in future work. ■

6 A Concrete Example for the Riemannian Gaussian Model

The aim of this section is to illustrate the geometric concepts involved in Propositions 3 and 5, by applying these concepts to the concrete example of the Riemannian Gaussian model defined on $M = \mathcal{P}_n$, the space of $n \times n$ real covariance matrices.

The space $\mathcal{P}_n$ is a Riemannian symmetric space, which is simply-connected and has non-positive sectional curvature [22, 44]. It is usually equipped with its affine-invariant Riemannian metric [6, 44],

$$Q_{\bar{x}}(u, u) \;=\; \text{tr}\left[\bar{x}^{-1}u\right]^2 \quad \bar{x} \in \mathcal{P}_n\,,\; u \in T_{\bar{x}}\mathcal{P}_n \tag{34a}$$

This metric is invariant under the action of the group of isometries $G = GL(n, \mathbb{R})$ on $\mathcal{P}_n$, which is given by affine transformations,

$$g \cdot \bar{x} = g\,\bar{x}\,g^t \tag{34b}$$

where ^{t} denotes the transpose. Moreover, this metric induces a Riemannian distance on $\mathcal{P}_n$, which is given by,

$$d^2(\bar{x}, \bar{y}) \;=\; \text{tr}\left[\log\left(\bar{x}^{-1/2}\,\bar{y}\,\bar{x}^{-1/2}\right)\right]^2 \tag{34c}$$

This distance is also invariant under the action of the group $GL(n, \mathbb{R})$ on $\mathcal{P}_n$. In other words: $d(g \cdot \bar{x}, g \cdot \bar{y}) \;=\; d(\bar{x}, \bar{y})$.

The Riemannian Gaussian model on $\mathcal{P}_n$ is given by the probability density function [14, 41]

$$p(x|\bar{x}, \sigma) \;=\; Z^{-1}(\sigma)\,\exp\left[-\frac{d^2(x, \bar{x})}{2\sigma^2}\right] \tag{35a}$$

which is a probability density function with respect to the invariant volume element associated to the Riemannian metric (34a). The normalising factor $Z(\sigma)$ can be expressed as a multiple integral [41], (see Proposition 4 in this reference),

$$Z(\sigma) \;=\; \mathrm{C}_n \int_{\mathbb{R}^n} e^{-\,\|r\|^2/2\sigma^2} \prod_{i<j} \sinh\big(|r_i - r_j|/2\big)\, dr_1 \,\ldots\, dr_n \tag{35b}$$

where C_n is a numerical constant which only depends on n, and the integration variable is denoted $r = (r_1\,, \ldots\,, r_n) \in \mathbb{R}^n$. If $\eta(\sigma) = -1/2\sigma^2$, then $\psi(\eta) = \log\, Z(\sigma)$ is a strictly convex function of $\eta \in (-\infty\,, 0)$.

With (34) and (35) in mind, consider the application of Proposition 3 to the Riemannian Gaussian model on $\mathcal{P}_n$. This will lead to the expression of the Rao–Fisher information metric I of this model.

De Rham decomposition of $\mathcal{P}_n$: recall first that Proposition 3 uses the De Rham decomposition, introduced in Remark 7. For the Riemannian symmetric space $\mathcal{P}_n$, the De Rham decomposition states that $\mathcal{P}_n$ is a Riemannian product of irreducible Riemannian symmetric spaces $\mathcal{P}_n = \mathbb{R} \times S\mathcal{P}_n$, where $S\mathcal{P}_n$ is the set of $\bar{s} \in \mathcal{P}_n$ such that $\det(\bar{s}) = 1$. The identification of $\mathcal{P}_n$ with $\mathbb{R} \times S\mathcal{P}_n$ is obtained by identifying each $\bar{x} \in \mathcal{P}_n$ with a couple $(\bar{\tau}, \bar{s})$, where $\bar{\tau} \in \mathbb{R}$ and $\bar{s} \in S\mathcal{P}_n$ are given by

$$\bar{\tau} = \log\det(\bar{x}) \quad \bar{s} = e^{-\,\bar{\tau}/n}\,\bar{x} \tag{36a}$$

Note that the spaces $\mathbb{R}$ and $S\mathcal{P}_n$ are indeed irreducible Riemannian symmetric spaces. This is clear for $\mathbb{R}$, which is one-dimensional and cannot be decomposed into a product of lower-dimensional spaces. The fact that $S\mathcal{P}_n$ is irreducible can be found in [22], (Table II, p. 354). It will be convenient to write $\bar{x} = (\bar{x}_1, \bar{x}_2)$ where $\bar{x}_1 = \bar{\tau}$ and $\bar{x}_2 = \bar{s}$. If $u \in T_{\bar{x}}\mathcal{P}_n$, then $u = u_1 + u_2$,

$$u_1 = \frac{1}{n}\,\mathrm{tr}(\bar{x}^{-1}u)\,\bar{x} \quad u_2 = u - \frac{1}{n}\,\mathrm{tr}(\bar{x}^{-1}u)\,\bar{x} \tag{36b}$$

Here, $u_1 \in T_{\bar{x}_1}\mathbb{R}$, where $T_{\bar{x}_1}\mathbb{R} \subset T_{\bar{x}}\mathcal{P}_n$ is the one-dimensional subspace consisting of symmetric matrices v of the form $v = t\,\bar{x}$ with t any real number. On the other hand, $u_2 \in T_{\bar{x}_2}S\mathcal{P}_n$, where $T_{\bar{x}_2}S\mathcal{P}_n \subset T_{\bar{x}}\mathcal{P}_n$ is the subspace consisting of symmetric matrices v which satisfy $\mathrm{tr}(\bar{x}^{-1}v) = 0$. Using (36a) and (36b), (20a) and (20b) of Remark 7 can be written down,

$$Q_{\bar{x}}(u, u) \;=\; Q_{\bar{x}}(u_1\,, u_1) + Q_{\bar{x}}(u_2\,, u_2) \tag{36c}$$

$$d^2(\bar{x}, \bar{y}) \;=\; \frac{1}{n}\;|\bar{x}_1 - \bar{y}_1|^2 + d^2(\bar{x}_2\,, \bar{y}_2) \tag{36d}$$

where $Q_{\bar{x}}$ is the affine-invariant metric (34a) and $d(\bar{x}, \bar{y})$ or $d(\bar{x}_2\,, \bar{y}_2)$ is the Riemannian distance (34c). The proof of formulae (36c) and (36d) is a direct calculation, and is not detailed here.

The Rao–Fisher metric I : according to (22) of Proposition 3, the Rao–Fisher information metric I of the Riemannian Gaussian model on $\mathcal{P}_n$ is given by,

$$I_z(U,U) = \psi''(\eta)\,u_\eta^2 + \frac{4\eta^2\psi_1'(\eta)}{\dim \mathbb{R}}\,Q_{\bar{x}}(u_1\,,u_1) + \frac{4\eta^2\psi_2'(\eta)}{\dim S\mathcal{P}_n}\,Q_{\bar{x}}(u_2\,,u_2) \tag{37a}$$

for $z = (\bar{x},\sigma)$ in the parameter space $\mathcal{M} = \mathcal{P}_n \times (0\,,\infty)$, and for $U = u_\eta\,\partial_\eta + u$ where $u = u_1 + u_2$ is given by (36b). Indeed, (37a) results from (22), by putting $r = 2$, as well as $M_1 = \mathbb{R}$ and $M_2 = S\mathcal{P}_n$. The functions appearing in (37a) are $\psi(\eta) = \log\,Z(\sigma)$ with $Z(\sigma)$ given by (35b), and, as shown in Remark 11 below,

$$\psi_1(\eta) = \frac{1}{2}\,\log(2\pi n) - \frac{1}{2}\,\log\,(-2\eta) \quad \psi_2(\eta) = \psi(\eta) - \psi_1(\eta) \tag{37b}$$

Moreover, $\dim \mathbb{R} = 1$ and $\dim S\mathcal{P}_n = \dim \mathcal{P}_n - 1 = n(n+1)/2 - 1$. Replacing into (37a) gives,

$$I_z(U,U) = \psi''(\eta)\,u_\eta^2 - 2\eta\,Q_{\bar{x}}(u_1\,,u_1) + \frac{8\eta^2\psi_2'(\eta)}{n^2+n-2}\,Q_{\bar{x}}(u_2\,,u_2) \tag{37c}$$

This expression of the Rao–Fisher information metric of the Riemannian Gaussian model on $\mathcal{P}_n$ can be computed directly from (34a), (36b) and (37b), once the function $\psi(\eta)$ is known. This function $\psi(\eta)$ has been tabulated for values of n up to $n = 50$, using a Monte Carlo method which was developed specifically for the evaluation of (35b) [49].

Remark 11 Assume x follows the Riemannian Gaussian probability density (35a) on $\mathcal{P}_n$. If $x = (x_1\,,x_2)$ where $x_1 \in \mathbb{R}$ and $x_2 \in S\mathcal{P}_n$, then the densities of x_1 and x_2 can be found by replacing (36d) into (35a). Precisely, this gives

$$p(x|\bar{x},\sigma) \,\propto\, \exp\left[-\frac{|x_1 - \bar{x}_1|^2}{2n\sigma^2}\right] \times \exp\left[-\frac{d^2(x_2\,,\bar{x}_2)}{2\sigma^2}\right]$$

It follows from this decomposition that x_1 and x_2 are independent, and that x_1 follows a univariate normal distribution of mean $\bar{x}_1$ and of variance $n\sigma^2$. In particular, the moment generating function $\psi_1(\eta)$ of the squared distance $|x_1 - \bar{x}_1|^2$ has the expression stated in (37b). ■

The generalised Mahalanobis distance on $\mathcal{P}_n$: applying Proposition 5 will yield the expression of the generalised Mahalanobis distance on $\mathcal{P}_n$. The Rao–Fisher information metric I as given by (37c) induces an extrinsic Riemannian metric Q^σ on $\mathcal{P}_n$, for each $\sigma \in (0\,,\infty)$,

$$Q^\sigma_{\bar{x}}(u,u) = -\,2\eta\,Q_{\bar{x}}(u_1\,,u_1) + \frac{8\eta^2\psi_2'(\eta)}{n^2+n-2}\,Q_{\bar{x}}(u_2\,,u_2) \quad \eta = -\frac{1}{2\sigma^2} \tag{38a}$$

The generalised Mahalanobis distance on $\mathcal{P}_n$ is the Riemannian distance induced on $\mathcal{P}_n$ by the extrinsic Riemannian metric Q^σ. If the generalised Mahalanobis distance between $\bar{x}$ and $\bar{y}$ in $\mathcal{P}_n$ is denoted $d(\bar{x}, \bar{y}\,|\sigma)$, then (31) of Proposition 5, along with (38a), imply

$$d^2(\bar{x}, \bar{y}\,|\sigma) = \frac{|\bar{x}_1 - \bar{y}_1|^2}{n\sigma^2} + \frac{4\psi_2'\left((-2\sigma^2)^{-1}\right)}{(n^2+n-2)\sigma^4}\, d^2(\bar{x}_2\,, \bar{y}_2) \tag{38b}$$

This distance can be computed directly from (34c), (36a) and (37b), once the function $\psi(\eta)$ has been tabulated using the Monte Carlo method of [49], or computed in any other way.

Affine invariance of the generalised Mahalanobis distance: the affine-invariant Riemannian metric Q of (34a) is well-known to the information science community, having been introduced in [39]. Besides the metric Q, a whole new family of affine-invariant Riemannian metrics Q^σ is provided by (38a). Indeed, to say that Q is affine-invariant means that it is invariant under affine transformations (34b). In other words

$$Q_{g\cdot\bar{x}}(dg_{\bar{x}}\,u, dg_{\bar{x}}\,u) \;=\; Q_{\bar{x}}(u, u) \quad \text{for all } g \in GL(n, \mathbb{R}) \tag{39a}$$

where $dg_{\bar{x}}$ denotes the derivative of the affine transformation (34b) at the point $\bar{x} \in \mathcal{P}_n$. On the other hand, a direct verification shows that each one of the metrics Q^σ also verifies (39a), so that

$$Q^\sigma_{g\cdot\bar{x}}(dg_{\bar{x}}\,u, dg_{\bar{x}}\,u) \;=\; Q^\sigma_{\bar{x}}(u, u) \quad \text{for all } g \in GL(n, \mathbb{R}) \tag{39b}$$

This means that each one of the metrics Q^σ is an affine-invariant Riemannian metric, as claimed. Furthermore, the fact that the metric Q^σ is invariant under affine transformations implies that the generalised Mahalanobis distance (38b) is also invariant under these transformations,

$$d(g\cdot\bar{x}, g\cdot\bar{y}\,|\sigma) \;=\; d(\bar{x}, \bar{y}\,|\sigma) \quad \text{for all } g \in GL(n, \mathbb{R}) \tag{39c}$$

This is because the generalised Mahalanobis distance (38b) is the Riemannian distance induced on $\mathcal{P}_n$ by Q^σ.

7 The Solution of the Geodesic Equation

The present section provides the solution of the geodesic equation of a multiply-warped Riemannian metric. The main result is the following Proposition 6. This proposition shows that the solution of the geodesic equation of a multiply-warped Riemannian metric, for given initial conditions, reduces to the solution of a one-dimensional second-order differential equation. As stated in Remark 8, warped Riemannian metrics are a special case of multiply-warped Riemannian metrics. Therefore, Proposition 6 also applies to the solution of the geodesic equation of a warped Riemannian metric. This special case of warped Riemannian metrics was treated separately in [38].

Let I be a multiply-warped Riemannian metric defined on $\mathcal{M} = M \times (0\,,\infty)$, in the notation of (21),

$$I_z(U, U) \;=\; (\alpha(\sigma)\, u_\sigma)^2 + \sum_{q=1}^{r} \beta_q^2(\sigma)\, Q_{\bar{x}}(u_q\,, u_q) \tag{40}$$

for $z = (\bar{x}, \sigma)$ and $U \in T_z\mathcal{M}$, with $u = u_\sigma\, \partial_\sigma + u$ and $u = u_1 + \cdots + u_r$. As in (2b) of Sect. 2, introduce the vertical distance coordinate r, which is defined by $dr/d\sigma = \alpha(\sigma)$.

Proposition 6 *Let $\gamma(t)$ be a geodesic of the multiply-warped Riemannian metric I, with initial conditions $\gamma(0) = z$ and $\dot{\gamma}(0) = U$, and let $\gamma(t) = (\bar{x}(t), \sigma(t))$ and $r(t) = r(\sigma(t))$. Then, $r(t)$ verifies the second-order differential equation*

$$\ddot{r} \;=\; -\frac{1}{2}\frac{d}{dr}V(r) \quad V(r) \;=\; \sum_{q=1}^{r} \frac{\beta_q^2(r(0))}{\beta_q^2(r)}\, I_z(u_q\,, u_q) \tag{41a}$$

and $\bar{x}(t)$ is given by

$$\bar{x}(t) \;=\; \exp_{\bar{x}}\left[\sum_{q=1}^{r}\left(\int_0^t \frac{\beta_q^2(r(0))}{\beta_q^2(r(s))} ds\right) u_q\right] \tag{41b}$$

where exp *denotes the Riemannian exponential mapping of the metric Q on M.*

Proof The proof is given in Appendix C. It is a generalisation of the proof dealing with the special case of warped Riemannian metrics, which can be found in [38] (Proposition 38, p. 208). ■

Proposition 6 shows that the main difficulty, involved in computing a geodesic $\gamma(t)$ of the multiply-warped Riemannian metric I, lies in the solution of the second order differential equation (41a). Indeed, once this equation is solved, computing $\gamma(t)$ essentially reduces to an application of exp, which is the Riemannian exponential mapping of the metric Q on M. In the context of the present contribution, Q is an invariant metric on M, where M is a Riemannian symmetric space. Therefore, exp has a straightforward expression [22] (Theorem 3.3, p. 173). In particular, for the examples treated in Sect. 4, the expression of exp is well-known in the literature. For the von Mises–Fisher model, this expression is elementary, since geodesics on a sphere in Euclidean space are the great circles on this sphere. For the Riemannian Gaussian model, when this model is defined on the space $M = \mathcal{P}_n$ of $n \times n$ real covariance matrices, the expression of exp is widely used in the literature, as found in [39].

Remark 12 The differential equation (41a) is the equation of motion of a one-dimensional conservative mechanical system. As such, its solution can be carried out by quadrature [21], (see p. 11). Precisely,

$$t = \pm \int_{r(0)}^{r(t)} \frac{dr}{\sqrt{E - V(r)}} \tag{42a}$$

where the total energy E is a conserved quantity, in the sense that $\dot{E} = 0$, and it can be shown that $E = I_z(U, U)$. Recalling that $dr/d\sigma = \alpha(\sigma)$, this integral can be written

$$t = \pm \int_{\sigma(0)}^{\sigma(t)} \frac{\alpha(\sigma)}{\sqrt{E - V(\sigma)}}\, d\sigma \quad V(\sigma) = V(r(\sigma)) \tag{42b}$$

Here, if t is interpreted as time, then the integral on the right-hand side gives the time necessary to go from $\sigma(0)$ to $\sigma(t)$. In particular, replacing $\sigma(t)$ by ∞ and by 0 gives the two quantities

$$t_\infty = \int_{\sigma(0)}^{\infty} \frac{\alpha(\sigma)}{\sqrt{E - V(\sigma)}}\, d\sigma \quad t_0 = \int_{0}^{\sigma(0)} \frac{\alpha(\sigma)}{\sqrt{E - V(\sigma)}}\, d\sigma \tag{42c}$$

where t_∞ is the time necessary for $\sigma(t)$ to reach the value $\sigma = \infty$, and t_0 is the time necessary for $\sigma(t)$ to reach the value $\sigma = 0$. Since $\mathcal{M} = M \times (0, \infty)$, these two values $\sigma = \infty$ and $\sigma = 0$ are excluded from $\mathcal{M}$. Therefore, the geodesic $\gamma(t)$ cannot be extended beyond the time $t = \min(t_\infty, t_0)$, as it would then escape from $\mathcal{M}$. ■

Remark 13 A vertical geodesic is a geodesic $\gamma(t)$ for which $\dot{\gamma}(0) = U$ with $U = u_\sigma\, \partial_\sigma$. This means that all the u_q are zero. In (41a), this implies that $V(r) = 0$, so that $\ddot{r} = 0$ and $r(t)$ is an affine function of t. In (41b), this implies that $\bar{x}(t) = \bar{x}$ is constant. In Remark 2 of Sect. 2, it was shown that a vertical geodesic is a unique length-minimising geodesic. For a vertical geodesic, (42c) reads

$$t_\infty = \frac{1}{\sqrt{E}} \int_{\sigma(0)}^{\infty} \alpha(\sigma)\, d\sigma \quad t_0 = \frac{1}{\sqrt{E}} \int_{0}^{\sigma(0)} \alpha(\sigma)\, d\sigma \tag{43a}$$

These formulae provide another way of understanding conditions (3b) from Sect. 2. Precisely, since $dr/d\sigma = \alpha(\sigma)$, it is clear that t_∞ and t_0 are given by

$$t_\infty = \frac{1}{\sqrt{E}} \lim_{\sigma \to \infty} r(\sigma) - r(\sigma(0)) \quad t_0 = \frac{1}{\sqrt{E}} \lim_{\sigma \to 0} r(\sigma(0)) - r(\sigma) \tag{43b}$$

Thus, the first condition in (3b) is equivalent to $t_\infty = \infty$, which means that $\sigma(t)$ cannot reach the value $\sigma = \infty$ within a finite time, and the second condition in (3b) is equivalent to $t_0 = \infty$, which means that $\sigma(t)$ cannot reach the value $\sigma = 0$ within a finite time. ■

8 The Construction of Riemannian Brownian Motion

The present section provides an explicit construction of the Riemannian Brownian motion associated to any multiply-warped Riemannian metric. The main result is the following Proposition 7, which shows that this construction reduces to the solution of a one-dimensional stochastic differential equation. This is in analogy with Proposition 6 of the previous section.

Let I be a multiply-warped Riemannian metric on $\mathcal{M} = M \times (0\,,\infty)$, given by (40). Recall that a Riemannian Brownian motion associated to I is a diffusion process z in $\mathcal{M}$ which satisfies the identity, see [23, 24],

$$df(z(t)) \,=\, \frac{1}{2}\,\Delta_{\mathcal{M}} f(z(t)) \,+\, dm^f(t) \tag{44a}$$

for each smooth function f on $\mathcal{M}$, where $\Delta_{\mathcal{M}}$ is the Laplace–Beltrami operator of I, and where m^f is a local martingale with respect to the augmented natural filtration of z. In other words, z is a Riemannian Brownian motion associated to I, if z solves the martingale problem associated to $(1/2)\Delta_{\mathcal{M}}$. The Laplace–Beltrami operator $\Delta_{\mathcal{M}}$ is given by the following formula, which will be proved in Appendix D,

$$\Delta_{\mathcal{M}} f \,=\, \frac{1}{G}\,\partial_r\,(G\,\partial_r f) \,+\, \sum_{q=1}^{r} \frac{1}{\beta_q^2(r)}\,\Delta_{M_q} f \tag{44b}$$

where $G = G(r)$ is given by $G = \prod_{q=1}^{r} \beta_q^{\dim M_q}$. Equation (44a) defines a Riemannian Brownian motion process z, associated to the multiply-warped Riemannian metric I, through its relationship to the Laplace–Beltrami operator $\Delta_{\mathcal{M}}$. On the other hand, the following Proposition 7 shows how such a Riemannian Brownian motion process may be constructed explicitly.

Proposition 7 *Let z be a stochastic process with values in $\mathcal{M}$, and write $z(t) = (\bar{x}(t), \sigma(t))$ where $\bar{x}(t) = (\bar{x}_1(t)\,, \ldots\,, \bar{x}_r(t))$ and $\bar{x}_q(t) \in M_q$. Let r and $(\theta_1\,, \ldots\,, \theta_r)$ be independent diffusion processes, where r is a one-dimensional diffusion process, which satisfies the stochastic differential equation*

$$dr(t) = \frac{1}{2}\,\frac{\partial_r G}{G}(r(t))dt + dw(t) \tag{45a}$$

with w a standard Brownian motion, and where θ_q is a Riemannian Brownian motion in M_q . If each $\bar{x}_q$ is a time-changed version of θ_q ,

$$\bar{x}_q(t) \,=\, \left(\theta_q \circ \tau_q\right)(t) \quad \tau_q(t) = \int_0^t \frac{ds}{\beta_q^2(r(s))} \tag{45b}$$

and if $\sigma(t) = \sigma(r(t))$, then the process z is a Riemannian Brownian motion associated to the multiply-warped Riemannian metric I given by (40).

Proof The proof of this proposition is given in Appendix D. ■

Remark 14 The above Proposition 7 can be used for numerical simulation of a Riemannian Brownian motion z associated to a multiply-warped Riemannian metric I. According to Proposition 7, to simulate z, it is enough to simulate the solution r of the one-dimensional stochastic differential equation (45a), and then to independently simulate each θ_q as a Riemannian Brownian motion in the Riemannian symmetric space M_q . To simulate the solution r of (45a), it is enough to use any of the numerical schemes described in [26]. On the other hand, Riemannian Brownian motion in a symmetric space can be simulated using Lie group stochastic exponentials [4, 20, 31]. A detailed description of these methods falls outside the present scope, and is reserved for future work. ■

9 Surprising Observation: Hadamard Manifolds

In Sect. 2, the completeness and curvature of a warped Riemannian metric I were characterised by Formulae (3b) and (4), respectively. Here, based on Sect. 4.1, these formulae will be applied to the case where I is the Rao–Fisher information metric of the von Mises–Fisher model defined on S^{n-1}. Precisely, this application is carried out using a mixture of analytical and numerical computations, for the von Mises–Fisher model defined on S^{n-1} where $n = 2, \ldots, 8$. The result is a surprising observation: the parameter space of the von Mises–Fisher model, when equipped with the Rao–Fisher information metric I, becomes a Hadamard manifold, a simply-connected complete Riemannian manifold of negative sectional curvature [13, 40]. Since this observation is true for several values of n, it gives rise to a family of Hadamard manifolds. Part of this claim can be proved for any value of $n = 2, \ldots,$ as in the following proposition.

Proposition 8 *For any value of $n = 2, \ldots,$ the parameter space of the von Mises–Fisher model defined on S^{n-1} is a simply-connected manifold, which moreover becomes a complete Riemannian manifold when equipped with the Rao–Fisher information metric I.*

Proof Recall from Remark 6 that the parameter space of the von Mises–Fisher model defined on S^{n-1} is identified with $\mathbb{R}^n$. Of course, $\mathbb{R}^n$ is a simply-connected manifold [43]. Thus, to prove the proposition, it remains to prove that the parameter space $\mathbb{R}^n$ becomes a complete Riemannian manifold when equipped with the Rao–Fisher information metric I. This will be done by proving that all geodesics of the metric I which pass through the point $z = 0$ in $\mathbb{R}^n$ can be extended indefinitely. Then, a corollary of the Hopf–Rinow theorem [13] (Corollary I.7.2, p. 29) implies the required completeness of the parameter space $\mathbb{R}^n$.

First, note that the geodesics of the metric I which pass through the point $z = 0$ are exactly the radial straight lines in $\mathbb{R}^n$. Indeed, according to Remark 6, if $\gamma(t)$ is a geodesic of I where $\gamma(t) = (\bar{x}(t), \eta(t))$, then $\gamma(t)$ is identified with the curve

$z(t) = \eta(t)\,\bar{x}(t)$ in $\mathbb{R}^n$. Moreover, by Proposition 2, the restriction of I to $\mathbb{R}^n - \{0\}$ is a warped Riemannian metric of the general form (13a). Then, by Remark 13, the vertical geodesics $\gamma(t)$ of this warped Riemannian metric are parameterised by $r(t) =$ affine function of t and $\bar{x}(t) = \bar{x} =$ constant, where r is the vertical distance coordinate. Therefore, each vertical geodesic $\gamma(t)$ can be parameterised by $\eta(t) = \eta(r(t))$ and $\bar{x}(t) = \bar{x} =$ constant. This is identified with the curve $z(t) = \eta(t)\,\bar{x}$, which is a radial straight line in $\mathbb{R}^n$, in the direction $\bar{x}$. It remains to note that the geodesics of the metric I are just the geodesics of its restriction to $\mathbb{R}^n - \{0\}$, extended by continuity whenever they reach the point $z = 0$.

Let $z(t) = \eta(t)\,\bar{x}$ describe a geodesic of the metric I, as just explained. To say that this geodesic can be extended indefinitely is equivalent to saying that $\eta(t)$ cannot reach the value $\eta = \infty$ within a finite time. For the von Mises–Fisher model, $\eta(\sigma) = \sigma$ as long as $\sigma \in (0\,,\infty)$. Therefore, according to Remark 13, by evaluating the two conditions (3b), it is possible to know whether $\eta(t)$ can reach the two values $\eta = \infty$ and $\eta = 0$ within a finite time. These conditions now read

$$\lim_{\eta\to\infty} r(\eta) - r(\eta_0) \overset{?}{=} \infty \quad\text{and}\quad \lim_{\eta\to 0} r(\eta_1) - r(\eta) \overset{?}{=} \infty$$

where η_0 and η_1 are arbitrary. By (13a), $r(\eta)$ is defined by $dr/d\eta = \left(\psi''(\eta)\right)^{1/2}$. Therefore, the two conditions in (3b) are identical to

$$\int_{\eta_0}^{\infty} \left(\psi''(\eta)\right)^{1/2} d\eta \overset{?}{=} \infty \quad\text{and}\quad \int_{0}^{\eta_1} \left(\psi''(\eta)\right)^{1/2} d\eta \overset{?}{=} \infty \tag{46a}$$

where $\psi''(\eta)$ is given by (17a). For the first integral, recall the asymptotic expansion of modified Bessel functions at $\eta = \infty$ [46], (Sect. 7.23, p. 203. This formula appears with the wrong sign for the second term in parentheses, in [33]),

$$I_\nu(\eta) \;=\; \frac{e^\eta}{\sqrt{2\pi\eta}}\left(1 - \frac{4\nu^2-1}{8\eta} + \frac{(4\nu^2-1)(4\nu^2-3^2)}{2!(8\eta)^2}\right) + O\left(\eta^{-3}\right)$$

Using this asymptotic expansion, it follows by performing some direct calculations, and recalling that $\nu = n/2$,

$$\begin{aligned} \tfrac{I_{\nu+1}(\eta)}{I_{\nu-1}(\eta)} &= 1 - \tfrac{n}{\eta} + \tfrac{n(n-1)}{2\eta^2} + O\left(\eta^{-3}\right) \\ \tfrac{I^2_{\nu}(\eta)}{I^2_{\nu-1}(\eta)} &= 1 - \tfrac{n-1}{\eta} + \tfrac{(n-1)(n-2)}{2\eta^2} + O\left(\eta^{-3}\right) \end{aligned}$$

Replacing these expressions into (17a) immediately gives

$$\psi''(\eta) \;=\; \frac{n-1}{2\eta^2} + O\left(\eta^{-3}\right) \tag{46b}$$

Since $n > 1$, this implies that the first integral in (46a) is divergent, as required. The second integral in (46a) is actually convergent. Indeed, $\psi''(\eta)$ is a continuous function in the neighborhood of $\eta = 0$, as seen in the proof of Proposition 2, for the limit (18c). Thus, the first condition in (3b) is verified, while the second condition is not verified. This means that $\eta(t)$ cannot reach the value $\eta = \infty$ within a finite time, but that it can reach the value $\eta = 0$ within a finite time. The first of these two statements shows that the geodesic described by $z(t)$ can indeed be extended indefinitely. Now, any geodesic of the metric I which passes through the point $z = 0$ is described by some $z(t)$ of this form. ■

The idea behind the proof of the completeness part of Proposition 8 can be summarised as follows. The restriction of the Rao–Fisher information metric I to $\mathbb{R}^n - \{0\}$ is a warped Riemannian metric. Thus, as stated in Sect. 2, $\mathbb{R}^n - \{0\}$ will be a complete Riemannian manifold, when equipped with this warped Riemannian metric, if and only if the two conditions in (3b) are verified. Once these conditions are evaluated, it turns out the first one is verified, but the second one is not. Thus, $\mathbb{R}^n - \{0\}$ is not a complete Riemannian manifold. However, this is only due to the fact that the point $z = 0$ is excluded. Once this point is included, the parameter space $\mathbb{R}^n$ is obtained, and this is a complete Riemannian manifold, when equipped with the Rao–Fisher information metric I. Precisely, a vertical geodesic in $\mathbb{R}^n - \{0\}$ can reach the point $z = 0$ within a finite time, but then all it does is pass through this point, and immediately return to $\mathbb{R}^n - \{0\}$. However, this vertical geodesic cannot escape to infinity within a finite time.

Proposition 8 established that the parameter space $\mathbb{R}^n$ of the von Mises–Fisher model is a simply-connected complete Riemannian manifold, for any value of n. To show that this parameter space is a Hadamard manifold, it remains to show that it has negative sectional curvature. This is done using numerical computation, for $n = 2, \ldots, 8$.

Precisely, let K_z denote the sectional curvature of $\mathbb{R}^n$ at a point $z \in \mathbb{R}^n$, with respect to the Rao–Fisher information metric I. Since, according to Proposition 2, I is a warped Riemannian metric, the sectional curvature K_z can be computed from formulae (4). To evaluate formula (4a), the Gauss equation, it is enough to note that for the von Mises–Fisher model, K^M is a constant equal to $+1$. Indeed, K^M is the sectional curvature of the unit sphere S^{n-1}. Then, formula (4a) reads

$$K_z(u, v) \;=\; \frac{1}{\beta^2} - \left(\frac{\partial_r \beta}{\beta}\right)^2 \quad u, v \in T_{\bar{x}} S^{n-1} \tag{47a}$$

where $z = \eta\,\bar{x}$. On the other hand, formula (4b), the Jacobi equation, can be copied directly,

$$K_z(u, \partial_r) \;=\; -\frac{\partial_r^2 \beta}{\beta} \tag{47b}$$

Here, $\beta(\eta)$ is given by (17b) of Proposition 2, and r is the vertical distance coordinate defined by $dr/d\eta = \big(\psi''(\eta)\big)^{1/2}$. In each one of formulae (47), the right-hand side

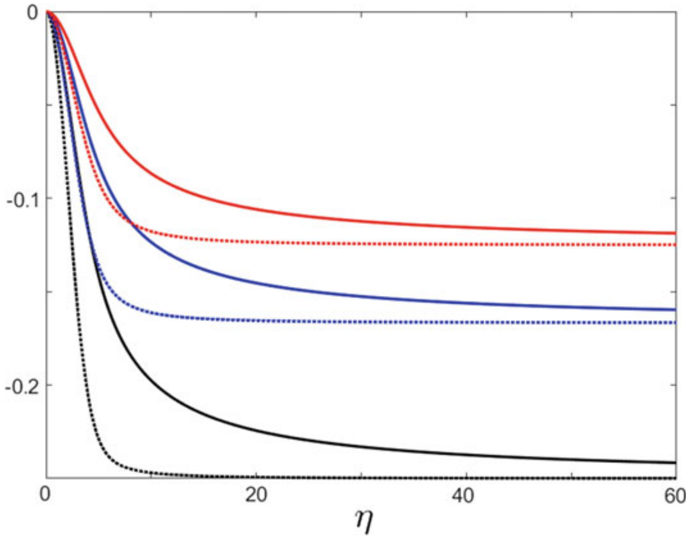


Fig. 1 Sectional curvature of the parameter space of the von Mises–Fisher model ($K^s(\eta)$ solid line, $K^r(\eta)$ dotted line, n = 3 black, n = 4 blue, n = 5 red)

is a real-valued function of η, independent of the vectors u, v. This will be denoted in the following way

$$K_z(u, v) = K^s(\eta) \quad K_z(u, \partial_r) = K^r(\eta) \tag{48}$$

Precisely, $K^s(\eta)$ is the sectional curvature of any section (u, v) tangent to the surface of a sphere centred at $z = 0$ and with Euclidean radius η. On the other hand, $K^r(\eta)$ is the sectional curvature of a radial section (u, ∂_r). In the following, $K^s(\eta)$ will be called the surface curvature, and $K^r(\eta)$ will be called the radial curvature.

Formulae (47) were computed numerically for $n = 2, \ldots, 8$. It was systematically found that the sectional curvatures $K^s(\eta)$ and $K^r(\eta)$ are negative for all values of η. From these numerical results, it can be concluded with certainty that the sectional curvature of $\mathbb{R}^n$, with respect to the Rao–Fisher information metric I, is negative, when n ranges from $n = 2$ to $n = 8$. Figure 1 gives a graphic representation for $n = 3, 4, 5$. The sectional curvatures $K^s(\eta)$ and $K^r(\eta)$ behave in the same way, for all considered values of n. Precisely, they are equal to zero at $\eta = 0$, and decrease to a limiting negative value, as η becomes large. This limiting value, denoted $K^s(\infty)$ and $K^r(\infty)$, for $K^s(\eta)$ and $K^r(\eta)$, respectively, is given in the following Table 1. Remarkably, it appears from this table that $K^s(\infty)$ and $K^r(\infty)$ have the same first two digits after the decimal point.

Remark 15 Based on Proposition 8, and on the numerical results reported here, it has been found that the parameter space $\mathbb{R}^n$ of the von Mises–Fisher model becomes a Hadamard manifold, when equipped with the Rao–Fisher information metric I,

Table 1 Limiting value of the surface and radial curvatures

	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$
$K^s(\infty)$	−0.50	−0.25	−0.16	−0.12	− 0.10	−0.08	−0.07
$K^r(\infty)$	−0.50	−0.25	−0.16	−0.12	−0.10	−0.08	−0.07

for $n = 2, \dots, 8$. Indeed, Proposition 8 shows that this parameter space is a simply-connected complete Riemannian manifold, for any value of $n = 2, \dots,$ while the numerical results of Fig. 1 and Table 1 show that it has negative sectional curvature. Hopefully, future work will provide a mathematical proof of the proposition that the sectional curvature of the parameter space $\mathbb{R}^n$ is negative for $n = 2, \dots,$ without restriction. ■

Remark 16 Preliminary results from ongoing research indicate that the Riemannian Gaussian model, which was studied in Sect. 4.2, when defined on $M = H^{n-1}$, the $(n-1)$-dimensional hyperbolic space, has similar properties to the von Mises–Fisher model, with regard to sectional curvature. Indeed, numerical computations show the sectional curvature of its parameter space $\mathcal{M} = H^{n-1} \times (0\,,\infty)$, equipped with the Rao–Fisher information metric I, is negative for $n = 3, 4, 5$. These numerical computations were carried out using formulae (4), for the sectional curvature of a warped Riemannian metric. This is justified because the hyperbolic space H^{n-1} is an irreducible Riemannian symmetric space, since it is a space of constant negative curvature, so the Rao–Fisher information metric (22) is a warped Riemannian metric. ■

Conclusion

The aim of the present contribution was to reveal a common geometric structure, shared by all location-scale models which are invariant under the action of some Lie group. Precisely, all of these location-scale models have a Rao–Fisher information metric which is a warped (eventually, multiply-warped) Riemannian metric. This provides a unified geometric framework for the study of a wide variety of location-scale models: von Mises–Fisher, and Riemannian Gaussian models, detailed in the above, or elliptically contoured distribution, generalised Wishart, and hyperboloid models, among many additional models. For such location-scale models, the rich yet tractable geometry of warped Riemannian metrics can be used to understand and solve important computational and theoretical problems. For example, future work will be able to address computational problems such as on-line estimation of mixture models on manifolds, regression between manifold-valued data sets, or black-box optimisation on manifolds. High-dimensional computations, with big data sets, as involved in these problems, are greatly simplified by the introduction of warped metrics, which afford exact computation of Riemannian gradients, Hessians, and geodesics, with a computational complexity largely independent of dimension. In particular, this means that exploiting the natural gradient algorithm, even with manifold-valued data and parameters, will be greatly streamlined. On the theoretical side, the connection between location-scale models and warped Riemannian metrics raises exciting new questions. Based on the theoretical and numerical results obtained above, it seems justified to make the conjecture that the parameter space of any invariant location-scale model, defined on a simply-connected symmetric space, turns into a Hadamard manifold, a simply-connected complete Riemannian manifold of negative curvature, when equipped with its Rao–Fisher information metric. Future work will focus on this conjecture, which seems to convey both geometric and statistical insight. Roughly, the fundamental example of a location-scale model is the isotropic normal model, whose parameter space turns into a space of constant negative curvature, indeed a Hadamard manifold, when equipped with its Rao–Fisher information metric. On the other hand, it is known to statisticians that any sufficiently regular location-scale model is locally normal, which means it is locally "similar" to an isotropic normal model. The question is then to know how this local statistical similarity translates into local geometric similarity. A glimpse of this statistical-geometric equivalence is seen from Table 1 in Sect. 9. For the von Mises–Fisher model, as η goes to infinity, the model converges to an isotropic normal model (see [33]). In this limit, it is seen in Table 1, that the two sectional curvatures $K^s(\eta)$ and $K^r(\eta)$ become equal, which means all sectional curvatures are equal, exactly as in a space of constant curvature. Intuitively, as the von Mises–Fisher model converges to an isotropic normal model, its geometry converges to that of a space of constant negative curvature.

Appendix A – Proof of Theorem 1

In order to complete the proof of Theorem 1, the following proposition is needed. The notation is that of Remark 4 and of (11).

Proposition 9 *Assume condition (7) holds. Then,*

$$\partial_\sigma \ell(z) \circ g \;=\; \partial_\sigma \ell(g^{-1} \cdot z) \quad \nabla_{\bar{x}}\, \ell(z) \circ g \;=\; dg_{\bar{x}}\, \nabla_{\bar{x}}\, \ell(g^{-1} \cdot z) \tag{49a}$$

In particular, if $g = s_{\bar{x}}$,

$$\partial_\sigma \ell(z) \circ s_{\bar{x}} \;=\; \partial_\sigma \ell(z) \quad \nabla_{\bar{x}}\, \ell(z) \circ s_{\bar{x}} \;=\; -\,\nabla_{\bar{x}}\, \ell(z) \tag{49b}$$

Proof Note that (49b) follows from (49a), by the definition of the geodesic-reversing isometry $s_{\bar{x}}$ [22]. Indeed, $s_{\bar{x}} \cdot \bar{x} = \bar{x}$ so $s_{\bar{x}}^{-1} \cdot z = z$. Moreover, $ds_{\bar{x}} = -\mathrm{Id}$, as a linear mapping of $T_{\bar{x}}M$, where Id denotes the identity. To prove (49a), note that

$$(\partial_\sigma \ell(z) \circ g)\,(x) \;=\; \partial_\sigma \log p(g \cdot x|z) \;=\; \partial_\sigma \log p(x|g^{-1} \cdot z) \tag{50a}$$

where the second equality follows from condition (7). However,

$$\partial_\sigma \log p(x|g^{-1}\cdot z) \;=\; \partial_\sigma \ell(g^{-1}\cdot z)(x) \tag{50b}$$

Replacing (50b) in (50a) gives,

$$\left(\partial_\sigma \ell(z)\circ g\right)(x) \;=\; \partial_\sigma \ell(g^{-1}\cdot z)(x)$$

which is the first part of (49a). For the second part, a similar reasoning can be applied. Precisely, using condition (7), it follows,

$$\left(d\ell(z)\circ g\right)(x) \;=\; d\log p(g\cdot x|z) = d\log p(x|g^{-1}\cdot z) \;=\; d\ell^{(g)}(z)(x) \tag{51a}$$

where $d\ell(z)$ denotes the derivative of $\ell(z)$ with respect to $\bar{x}$, and $\ell^{(g)}(z) = \ell(g^{-1}\cdot z)$, so (51a) implies that,

$$d\ell(z)\circ g \;=\; d\ell^{(g)}(z) \tag{51b}$$

By the chain rule [29], for $u \in T_{\bar{x}}M$,

$$d\ell^{(g)}(z)\big|_{\bar{x}}\; u \;=\; d\ell(g^{-1}\cdot z)\, dg^{-1}_{\bar{x}}\, u$$

Replacing in (51b),

$$d\ell(z)\circ g|_{\bar{x}} \;=\; d\ell(g^{-1}\cdot z)\, dg^{-1}_{\bar{x}} \tag{51c}$$

The second part of (49a) can now be obtained as follows. By the definition of the Riemannian gradient [40],

$$Q\left(\nabla_{\bar{x}}\, \ell(z)\circ g\,, u\right) \;=\; d\ell(z)\circ g|_{\bar{x}}\; u \;=\; d\ell(g^{-1}\cdot z)\, dg^{-1}_{\bar{x}}\, u \tag{52a}$$

where the second equality follows from (51c). However,

$$d\ell(g^{-1}\cdot z)\, dg^{-1}_{\bar{x}}\, u \;=\; Q\left(\nabla_{\bar{x}}\, \ell(g^{-1}\cdot z), dg^{-1}_{\bar{x}}\, u\right)$$

Since g is an isometry of M, its derivative $dg_{\bar{x}}$ preserves the Riemannian metric Q. Therefore,

$$Q\left(\nabla_{\bar{x}}\, \ell(g^{-1}\cdot z), dg^{-1}_{\bar{x}}\, u\right) \;=\; Q\left(dg_{\bar{x}}\, \nabla_{\bar{x}}\, \ell(g^{-1}\cdot z),\; u\right) \tag{52b}$$

Replacing (52b) in (52a) gives,

$$Q\left(\nabla_{\bar{x}}\, \ell(z)\circ g\,, u\right) \;=\; Q\left(dg_{\bar{x}}\, \nabla_{\bar{x}}\, \ell(g^{-1}\cdot z),\; u\right)$$

To finish the proof, it is enough to note that the vector u is arbitrary. ■

Proof of (10b): recall the polarisation identity, from elementary linear algebra [28], (see p. 29),

$$I_z(\partial_\sigma, u) = \frac{1}{4}\, I_z(\partial_\sigma + u, \partial_\sigma + u) - \frac{1}{4}\, I_z(\partial_\sigma - u, \partial_\sigma - u)$$

by replacing (8) into this identity, it can be seen that,

$$I_z(\partial_\sigma, u) \;=\; \mathbb{E}_z\,((\partial_\sigma \ell(z))\,(d\ell(z)\,u))$$

Then, by recalling the definition of the Riemannian gradient [40],

$$I_z(\partial_\sigma, u) \;=\; \mathbb{E}_z\,((\partial_\sigma \ell(z))\;Q\,(\nabla_{\bar{x}}\,\ell(z), u)) \tag{53a}$$

Denote the function under the expectation by f, and apply (11) with $g = s_{\bar{x}}$. Then,

$$\mathbb{E}_z\, f \;=\; \mathbb{E}_{s_{\bar{x}}\cdot z}\, f \;=\; \mathbb{E}_z\,(f \circ s_{\bar{x}}) \tag{53b}$$

since $s_{\bar{x}} \cdot z = z$. Note that (10b) amounts to saying that $\mathbb{E}_z\, f = 0$. To prove this, note that

$$f \circ s_{\bar{x}} = (\partial_\sigma \ell(z) \circ s_{\bar{x}})\;Q\,(\nabla_{\bar{x}}\ell(z) \circ s_{\bar{x}}\,, u) = -\partial_\sigma \ell(z) Q\,(\nabla_{\bar{x}}\ell(z), u) = -f \tag{53c}$$

where the second equality follows from (49b). Replacing in (53b) shows that $\mathbb{E}_z\, f = 0$. ■

Proof of (10c): the idea is to apply Schur's lemma to $I_z(u, u)$, considered as a symmetric bilinear form on $T_{\bar{x}}M$. First, it is shown that this symmetric bilinear form is invariant under the isotropy representation. That is,

$$I_z(u, u) \;=\; I_z\,(dk_{\bar{x}}\,u\,, dk_{\bar{x}}\,u) \quad \text{for all } k \in K_{\bar{x}} \tag{54a}$$

This is done using (11). Note from (8),

$$I_z(u, u) = \mathbb{E}_z\left[(\,Q\,(\nabla_{\bar{x}}\,\ell(z), u))^2\right] \tag{54b}$$

Denote the function under the expectation by f. By (11),

$$\mathbb{E}_z\, f \;=\; \mathbb{E}_{k^{-1}\cdot z}\, f \;=\; \mathbb{E}_z\left(f \circ k^{-1}\right) \tag{54c}$$

since $k^{-1} \cdot z = z$ for $k \in K_{\bar{x}}$. To find $f \circ k^{-1}$, note that,

$$Q\left(\nabla_{\bar{x}}\,\ell(z) \circ k^{-1}, u\right) \;=\; Q\left(dk_{\bar{x}}^{-1}\nabla_{\bar{x}}\,\ell(z), u\right) \;=\; Q\,(\nabla_{\bar{x}}\,\ell(z), dk_{\bar{x}}\,u)$$

where the first equality follows from (49a) and the fact that $k \cdot \bar{x} = \bar{x}$, and the second equality from the fact that $dk_{\bar{x}}$ preserves the Riemannian metric Q. Now, by (54b) and (54c),

$$I_z(u, u) = \mathbb{E}_z\, f = \mathbb{E}_z\left(f \circ k^{-1}\right) = \mathbb{E}_z\,(Q\,(\nabla_{\bar{x}}\ell(z), dk_{\bar{x}}\,u))^2 = I_z\,(dk_{\bar{x}}\,u\,, dk_{\bar{x}}\,u)$$

and this proves (54a).

Recall Schur's lemma, ([27], p. 240). Applied to (54a), this lemma implies that there exists some multiplicative factor β^2, such that

$$I_z(u,u) \;=\; \beta^2\, Q_{\bar{x}}(u,u) \tag{55a}$$

It remains to show that β^2 is given by (9). Taking the trace of (55a),

$$\operatorname{tr} I_z = \beta^2 \operatorname{tr} Q_{\bar{x}} = \beta^2 \dim M \tag{55b}$$

If $e_1\,, \ldots,\, e_d$ is an orthonormal basis of $T_{\bar{x}}M$, then by (54b),

$$\operatorname{tr} I_z = \mathbb{E}_z \sum_{i=1}^{d} \left(Q\left(\nabla_{\bar{x}}\, \ell(z), e_i\right)\right)^2 = \mathbb{E}_z\, Q\left(\nabla_{\bar{x}}\ell(z), \nabla_{\bar{x}}\ell(z)\right) \tag{55c}$$

Thus, (55b) and (55c) show that β^2 is given by (9). ■

To complete the proof of Theorem 1, it remains to show that the expectations appearing in (9) do not depend on $\bar{x}$.

For the first expectation, giving $\alpha^2(\sigma)$, note that,

$$\mathbb{E}_{g\cdot z}\left(\partial_\sigma \ell(g\cdot z)\right)^2 \;=\; \mathbb{E}_z\left(\partial_\sigma \ell(g\cdot z)\circ g\right)^2 \;=\; \mathbb{E}_z\left(\partial_\sigma \ell(z)\right)^2 \tag{56}$$

where the first equality follows from (11) and the second equality follows from (49a). Thus, this expectation has the same value, whether computed at $g\cdot z = (g\cdot\bar{x},\sigma)$, or at $z = (\bar{x},\sigma)$. Therefore, it does not depend on $\bar{x}$, since the action of G on M is transitive.

For the second expectation, giving $\beta^2(\sigma)$, note that by (11),

$$\mathbb{E}_{g\cdot z}\, Q\left(\nabla_{\bar{x}}\, \ell(g\cdot z)\,,\, \nabla_{\bar{x}}\, \ell(g\cdot z)\,\right) = \mathbb{E}_z\, Q\left(\nabla_{\bar{x}}\, \ell(g\cdot z)\circ g\,,\, \nabla_{\bar{x}}\, \ell(g\cdot z)\circ g\,\right) \tag{57a}$$

On the other hand, by (49a),

$$\nabla_{\bar{x}}\, \ell(g\cdot z)\circ g \;=\; dg_{\bar{x}}\, \nabla_{\bar{x}}\, \ell(z)$$

Moreover, since $dg_{\bar{x}}$ preserves the Riemannian metric Q,

$$\begin{aligned} Q(\nabla_{\bar{x}}\, \ell(g\cdot z)\circ g\,,\, \nabla_{\bar{x}}\, \ell(g\cdot z)\circ g) &= Q(dg_{\bar{x}}\, \nabla_{\bar{x}}\, \ell(z)\,,\, dg_{\bar{x}}\nabla_{\bar{x}}\, \ell(z)) \\ &= Q(\nabla_{\bar{x}}\, \ell(z)\,,\, \nabla_{\bar{x}}\, \ell(z)) \end{aligned}$$

Replacing in (57a) gives

$$\mathbb{E}_{g\cdot z}\, Q\left(\nabla_{\bar{x}}\, \ell(g\cdot z)\,,\, \nabla_{\bar{x}}\, \ell(g\cdot z)\,\right) \;=\; \mathbb{E}_z\, Q(\nabla_{\bar{x}}\, \ell(z)\,,\, \nabla_{\bar{x}}\, \ell(z)) \tag{57b}$$

so this expectation has the same value, at $g\cdot z$ and at z. By the same argument made after (56), it does not depend on $\bar{x}$. ■

Proof of (11): let dv denote the invariant Riemannian volume element of M, and note that,

$$\mathbb{E}_{g\cdot z}\, f \;=\; \int_M f(x)\, p(x|g\cdot z)\, dv(x) \;=\; \int_M f(x)\, p(g^{-1}\cdot x|z)\, dv(x) \tag{58a}$$

where the second equality follows from (7). Introduce the variable $y = g^{-1}\cdot x$. Since the volume element dv is invariant,

$$\int_M f(x)\, p(g^{-1}\cdot x|z)\, dv(x) \;=\; \int_M f(g\cdot y)\, p(y|z)\, dv(y) \tag{58b}$$

The last integral is the same as $\mathbb{E}_z\,(f\circ g)$. Therefore, (11) follows from (58a) and (58b). ■

Appendix B – Proof of Proposition 2

It remains to prove (17a) and (17b). To do so, introduce the following notation, using (15),

$$t = \langle x, \bar{x}\rangle \quad Z(\eta) = e^{\psi(\eta)} = (2\pi)^{\nu}\, \eta^{1-\nu} I_{\nu-1}(\eta) \tag{59a}$$

Then, $Z(\eta)$ is the moment generating function of t, so

$$\mathbb{E}_z(t) = \frac{Z'(\eta)}{Z(\eta)} \quad \text{and} \quad \mathbb{E}_z\left(t^2\right) = \frac{Z''(\eta)}{Z(\eta)} \tag{59b}$$

where the prime denotes differentiation with respect to η. Recall the derivative and recurrence relations of modified Bessel functions [46],

$$\left(\eta^{-a} I_a(\eta)\right)' = \eta^{-a} I_{a+1}(\eta) \quad I_{a-1}(\eta) - I_{a+1}(\eta) \;=\; \frac{2a}{\eta}\, I_a(\eta) \tag{60}$$

where a is any complex number. By applying these relations to (59b), it is possible to show, through a direct calculation,

$$\mathbb{E}_z(t) = \frac{I_\nu(\eta)}{I_{\nu-1}(\eta)} \tag{61a}$$

$$\mathbb{E}_z\left(t^2\right) = \frac{1}{n} + \frac{n-1}{n}\,\frac{I_{\nu+1}(\eta)}{I_{\nu-1}(\eta)} \tag{61b}$$

Formulae (61) will provide the proof of (17a) and (17b).

Proof of (17a): since $\psi(\eta)$ is the cumulant generating function of t,

$$\psi''(\eta) \;=\; \text{Var}_z(t) \;=\; \mathbb{E}_z\left(t^2\right) - \mathbb{E}_z\,(t)^2 \tag{62a}$$

where Var denotes the variance. Now, (17a) follows immediately by replacing from (61) into the right-hand side. ■

Proof of (17b): recall from (18a),

$$\beta^2(\eta) \;=\; \frac{\eta^2}{n-1}\,\mathbb{E}_z\left(1 - t^2\right) \tag{62b}$$

However, from (61b),

$$\mathbb{E}_z\left(1 - t^2\right) = \frac{n-1}{n}\left(1 + \frac{I_{\nu+1}(\eta)}{I_{\nu-1}(\eta)}\right) \tag{62c}$$

Now, (17b) follows by replacing (62c) into (62b). ■

Proof of (61a): using the derivative relation of modified Bessel functions, which is the first relation in (60), with $a = \nu - 1$, it follows that

$$Z'(\eta) \;=\; (2\pi)^{\nu}\,\eta^{1-\nu} I_{\nu}(\eta) \tag{63a}$$

Now, Formula (61a) follows by replacing this into (59b) and using (59a). ■

Proof of (61b): write (63a) in the form

$$Z'(\eta) \;=\; (2\pi)^{\nu}\,\eta\,\left(\eta^{-\nu} I_{\nu}(\eta)\right)$$

By the product rule

$$Z''(\eta) \;=\; (2\pi)^{\nu}\,\eta^{-\nu} I_{\nu}(\eta) + (2\pi)^{\nu}\,\eta\,\left(\eta^{-\nu} I_{\nu}(\eta)\right)'$$

The derivative in the second term can be evaluated from the derivative relation of modified Bessel functions, with $a = \nu$. Then,

$$Z''(\eta) \;=\; (2\pi)^{\nu}\,\eta^{-\nu} I_{\nu}(\eta) + (2\pi)^{\nu}\,\eta^{1-\nu} I_{\nu+1}(\eta)$$

Rearrange this formula as

$$Z''(\eta) \;=\; (2\pi)^{\nu}\,\eta^{1-\nu}\,\left(\eta^{-1} I_{\nu}(\eta) + I_{\nu+1}(\eta)\right)$$

By the recurrence relation of modified Bessel functions, which is the second relation in (60), with $a = \nu$, it then follows

$$Z''(\eta) \;=\; (2\pi)^{\nu}\,\eta^{1-\nu}\,\left(\frac{1}{2\nu}\,I_{\nu-1}(\eta) - \frac{1}{2\nu}\,I_{\nu+1}(\eta) + I_{\nu+1}(\eta)\right)$$

Recalling that $2\nu = n$, this can be written,

$$Z''(\eta) \;=\; (2\pi)^{\nu}\,\eta^{1-\nu}\,\left(\frac{1}{n}\,I_{\nu-1}(\eta) + \frac{n-1}{n}\,I_{\nu+1}(\eta)\right) \tag{63b}$$

Now, Formula (61b) follows by replacing this into (59b) and using (59a). ■

Appendix C – Proof of Proposition 6

The setting and notations are the same as in Sect. 7, except for the fact that $\bar{x}$ is written as x, without the bar, in order to avoid notations such as $\dot{\bar{x}}$ or $\ddot{\bar{x}}$. This being said, let $\tilde{\nabla}$ and ∇ denote the Levi-Civita connections of the Riemannian metrics I and Q, respectively. Thus, $\tilde{\nabla}$ is a connection on the tangent bundle of the manifold $\mathcal{M}$, and ∇ is a connection on the tangent bundle of the manifold M [13, 40]. Introduce the shape operator $S : T_xM \rightarrow T_xM$, which is given as in [40],

$$S(u) = \tilde{\nabla}_u\,\partial_r \quad u \in T_xM \tag{64}$$

for any $x \in M$. The following identities can be found in [40] (Sect. 2.4, p. 41),

$$\tilde{\nabla}_{\partial_r}\,\partial_r = 0 \tag{65a}$$

$$\tilde{\nabla}_{\partial_r}\,X = S(X) \tag{65b}$$

$$\tilde{\nabla}_X\,Y = \nabla_X\,Y - I(S(X), Y)\,\partial_r \tag{65c}$$

for any vector fields X and Y on M. Using these identities, it is possible to write the geodesic equation of the Riemannian metric I, in terms of the shape operator S. This is given in the following proposition.

Proposition 10 *Let $\gamma(t)$ be a curve in $\mathcal{M}$, with $\gamma(t) = (x(t), \sigma(t))$ and let $r(t) = r(\sigma(t))$. The curve $\gamma(t)$ is a geodesic of the Riemannian metric I if and only if it satisfies the geodesic equation*

$$\ddot{r} = I(S(\dot{x}), \dot{x}) \tag{66a}$$

$$\ddot{x} = -2\,\dot{r}\,S(\dot{x}) \tag{66b}$$

where $\ddot{x} = \nabla_{\dot{x}}\,\dot{x}$ is the acceleration of the curve $x(t)$ in M.

The shape operator S moreover admits a simple expression, which can be derived from expression (40) of the Riemannian metric I, using the fact that $\tilde{\nabla}$ is a metric connection [13] (Theorem I.5.1, p. 16).

Proposition 11 *In the notation of (40), the shape operator S is given by*

$$S(u) = \sum_{q=1}^{r} \frac{\partial_r\, \beta_q(r)}{\beta_q(r)}\, u_q \tag{67}$$

In other words, the decomposition $u = u_1 + \cdots + u_r$ provides a block-diagonalisation of S, where each block is a multiple of identity.

Combining Propositions 10 and 11, the geodesic equation (66) takes on a new form. Precisely, replacing (67) into (66) gives the following equations

$$\ddot{r} = \sum_{q=1}^{r} \beta_q(r)\partial_r\beta_q(r)\; Q(\dot{x}_q, \dot{x}_q) \tag{68a}$$

$$\ddot{x}_q = -2\,\dot{r}\, \frac{\partial_r\, \beta_q(r)}{\beta_q(r)}\, \dot{x}_q \tag{68b}$$

where $\dot{x} = \dot{x}_1 + \cdots + \dot{x}_r$ and $\ddot{x} = \ddot{x}_1 + \cdots + \ddot{x}_r$. The proof of Proposition 6 can be obtained directly from Eq. (68), using the following conservation laws.

Proposition 12 *Each one of the following quantities C_q is a conserved quantity,*

$$C_q \;=\; \beta^4_q(r)\, Q(\dot{x}_q\,,\dot{x}_q) \;\; \textit{for} \;\; q = 1\,,\ldots,\, r \tag{69}$$

In other words, C_q remains constant when evaluated along any geodesic $\gamma(t)$ of the Riemannian metric I.

For now, assume that Propositions 10–12 are true. To prove Proposition 6, note the following.

Proof of (41a): it is enough to show that the right-hand side of (41a) is the same as the right-hand side of (68a). To do so, note from (41a) and (69) that

$$V(r) \;=\; \sum_{q=1}^{r} \frac{\beta^2_q(r(0))}{\beta^2_q(r)}\, I_z(u_q\,,u_q) \;=\; \sum_{q=1}^{r} \frac{C_q}{\beta^2_q(r)} \tag{70a}$$

Indeed, since $\dot{x}_q(0) = u_q$ and since C_q is a conserved quantity

$$\beta^2_q(r(0))\, I_z(u_q\,,u_q) \;=\; \beta^4_q(r)\, Q(\dot{x}_q\,,\dot{x}_q)\big|_{t=0} \;=\; C_q$$

Now, replacing the derivative of (70a) into the right-hand side of (41a) directly leads to the right-hand side of (68a). ■

Proof of (41b): recall from Remark 7 that M is the Riemannian product of the M_q. Therefore, the Riemannian exponential mapping of M is also the product of the Riemannian exponential mappings of the M_q. Precisely, (41b) is equivalent to

$$x_q(t) \;=\; \exp_{x_q(0)}\left[\left(\int_0^t \frac{\beta_q^2(r(0))}{\beta_q^2(r(s))}ds\right)\, u_q\right] \quad \text{for } q = 1\,,\ldots,\, r \tag{70b}$$

This means that the curve $x_q(t)$ in M_q is a reparameterised geodesic $\big(\delta_q \circ F\big)\,(t)$ where $\delta_q(t)$ is the geodesic given by $\delta_q(t) = \exp(t\,u_q)$ and $F(t)$ is the integral inside the parentheses in (70b). To prove (41b), it is sufficient to prove that (70b) solves Eq. (68b). Using the chain rule, (70b) implies that

$$\ddot{x}_q \;=\; \dot{F}^{\,2}\left(\ddot{\delta}_q \circ F\right) \,+\, F''\left(\dot{\delta}_q \circ F\right) \;=\; \dot{F}^{\,2}\left(\ddot{\delta}_q \circ F\right) \,+\, \frac{F''}{F'}\,\dot{x}_q \;=\; \frac{F''}{F'}\,\dot{x}_q \tag{70c}$$

where the third equality follows because δ_q is a geodesic, and therefore its acceleration $\ddot{\delta}_q$ is zero. By replacing the definition of the function $F(t)$, it is seen that (70c) is the same as (68b). It follows that (70b) solves (68b), as required. ■

Proof of Proposition 10: recall the geodesic equation is $\tilde{\nabla}_{\dot{\gamma}}\,\dot{\gamma} = 0$, which means that the velocity $\dot{\gamma}(t)$ is self-parallel [13, 40]. Here, the velocity $\dot{\gamma}(t)$ is given by $\dot{\gamma}(t) \,=\, \dot{r}\,\partial_r \,+\, \dot{x}$. Accordingly, the left-hand side of the geodesic equation is

$$\tilde{\nabla}_{\dot{\gamma}}\,\dot{\gamma} \;=\; \tilde{\nabla}_{\dot{\gamma}}\,\dot{r}\,\partial_r \,+\, \tilde{\nabla}_{\dot{\gamma}}\,\dot{x} \;=\; \ddot{r}\,\partial_r \,+\, \dot{r}\,\tilde{\nabla}_{\dot{\gamma}}\,\partial_r \,+\, \tilde{\nabla}_{\dot{\gamma}}\,\dot{x} \tag{71a}$$

where the second equality follows by the product rule for the covariant derivative [13, 40]. The second and third terms on the right-hand side of (71a) can be written in terms of the shape operator S. Precisely, for the second term,

$$\tilde{\nabla}_{\dot{\gamma}}\,\partial_r \;=\; \dot{r}\,\tilde{\nabla}_{\partial_r}\,\partial_r \,+\, \tilde{\nabla}_{\dot{x}}\,\partial_r \;=\; S(\dot{x}) \tag{71b}$$

where the second equality follows from (64) and (65a). Moreover, for the third term,

$$\tilde{\nabla}_{\dot{\gamma}}\,\dot{x} \;=\; \dot{r}\,\tilde{\nabla}_{\partial_r}\,\dot{x} \,+\, \tilde{\nabla}_{\dot{x}}\,\dot{x} \;=\; \dot{r}\,S(\dot{x}) \,+\, \ddot{x} \,-\, I(S(\dot{x}),\dot{x})\,\partial_r \tag{71c}$$

where the second equality follows from (65b) and (65c). Replacing (71b) and (71c) into (71a), the left-hand side of the geodesic equation becomes

$$\tilde{\nabla}_{\dot{\gamma}}\,\dot{\gamma} \;=\; (\ddot{r} - I(S(\dot{x}),\dot{x}))\;\partial_r \,+\, (\ddot{x} + 2\dot{r}S(\dot{x}))$$

Setting this equal to zero immediately gives Eq. (66). ■

Proof of Proposition 11 : recall the shape operator S is symmetric, since it is essentially the Riemannian Hessian of r [40] (Sect. 2.4, p. 41). Therefore, it is enough to evaluate $I(S(u),u)$ for $u \in T_xM$. Let X by a vector field on M, with $X(x) = u$. Then,

$$I(S(u),u) \;=\; I\,(S(X),X) \;=\; I\left(\tilde{\nabla}_{\partial_r}\,X\,,\,X\right) \tag{72a}$$

where the second equality follows from (65b). Using the fact that $\tilde{\nabla}$ is a metric connection [13] (Theorem I.5.1, p. 16), the right-hand side can be written as

$$I\left(\tilde{\nabla}_{\partial_r} X\,,X\right) \,=\, \frac{1}{2}\,\partial_r\, I(X,X) \,=\, \frac{1}{2}\,\partial_r \sum_{q=1}^{r} \beta_q^2(r)\, Q_x(u_q\,,u_q) \tag{72b}$$

where the second equality follows from (40). It remains to note that

$$\frac{1}{2}\,\partial_r\, \beta_q^2(r)\, Q_x(u_q\,,u_q) \,=\, \frac{\partial_r\, \beta_q(r)}{\beta_q(r)}\, I_z(u_q\,,u_q)$$

Accordingly, (72a) and (72b) imply

$$I(S(u),u) \,=\, I\left(\sum_{q=1}^{r} \frac{\partial_r\, \beta_q(r)}{\beta_q(r)}\, u_q\,,u_q\right) \,=\, I\left(\sum_{q=1}^{r} \frac{\partial_r\, \beta_q(r)}{\beta_q(r)}\, u_q\,,u\right) \tag{72c}$$

and (67) follows from the fact that S is symmetric. ■

Proof of Proposition 12: to say that C_q is a conserved quantity means that $\dot{C}_q = 0$. From (69),

$$\dot{C}_q \,=\, 4\dot{r}\, \beta_q^3(r)\, \partial_r \beta_q(r)\, Q(\dot{x}_q\; \dot{x}_q) \,+\, \beta_q^4(r)\, \frac{d}{dt}\, Q(\dot{x}_q\; \dot{x}_q) \tag{73a}$$

The last derivative can be expressed as

$$\frac{d}{dt}\, Q(\dot{x}_q\; \dot{x}_q) \,=\, 2\, Q(\ddot{x}_q\,,\dot{x}_q) \,=\, -4\dot{r}\, \frac{\partial_r \beta_q(r)}{\beta_q(r)}\, Q(\dot{x}_q\,,\dot{x}_q) \tag{73b}$$

where the second equality follows from (68b). By replacing (73b) into (73a), it follows immediately that $\dot{C}_q = 0$. ■

Appendix D – Proof of Proposition 7

The proof of Formula (44b), for the Laplace–Beltrami operator $\Delta_{\mathcal{M}}$, will introduce some useful notation.

Proof of Formula (44b): for any smooth function f on $\mathcal{M}$, it follows from (40) that the Riemannian gradient $\tilde{\nabla} f$ of f, with respect to the multiply-warped Riemannian metric I, is given by

$$\tilde{\nabla} f \,=\, (\partial_r f)\; \partial_r \,+\, \sum_{q=1}^{r} \beta_q^{-2}(r)\, \nabla_{\bar{x}_q}\, f \tag{74a}$$

where $\nabla_{\bar{x}_q} f$ is the Riemannian gradient of f with respect to $\bar{x}_q \in M_q$, computed with all other arguments of f being fixed. Expression (74a) can be verified by checking that

$$df\, U = I(\tilde{\nabla} f, U)$$

for any tangent vector U to $\mathcal{M}$. This follows directly from (40) and (74a). Now, by definition of the Laplace–Beltrami operator [34], (see p. 443),

$$\Delta_{\mathcal{M}} f = \operatorname{div} \tilde{\nabla} f \tag{74b}$$

where the divergence div V of a vector field V on $\mathcal{M}$ is found from

$$\mathcal{L}_V \operatorname{vol} = (\operatorname{div} V) \operatorname{vol} \tag{74c}$$

with the notation $\mathcal{L}$ for the Lie derivative, and vol for the Riemannian volume element of the metric I. This last formula can be applied along with the following expression of vol, which follows from (40),

$$\operatorname{vol} = G(r)\, dr \bigwedge_q \operatorname{vol}_q(\bar{x}_q) \tag{74d}$$

where the function $G(r)$ was defined after (44b), and where $\wedge$ denotes the exterior product, and vol_q is the Riemannian volume of M_q. From (74c) and (74d), applying the product formula of the Lie derivative [34] (Theorem 7.4.8, p. 414),

$$\operatorname{div} V = \frac{1}{G}\, \partial_r\, (G\, V_r) + \sum_{q=1}^{r} \operatorname{div}_{M_q} V_q \tag{74e}$$

where $V = V_r\, \partial_r + \sum_q V_q$ with each V_q tangent to M_q, and where $\operatorname{div}_{M_q} V_q$ denotes the divergence of V_q with respect to $\bar{x}_q \in M_q$. Formula (44b) follows directly from (74a), (74b) and (74e). Indeed, for the vector field $V = \tilde{\nabla} f$,

$$V_r = \partial_r f \quad V_q = \beta_q^{-2}(r)\, \nabla_{\bar{x}_q} f$$

as can be seen from (74a). ■

For the proof of Proposition 7, assume the process z is a Riemannian Brownian motion associated to I. Write $z(t) = (\bar{x}(t), \sigma(t))$ where $\bar{x}(t) = (\bar{x}_1(t), \ldots, \bar{x}_r(t))$ and each $\bar{x}_q(t)$ belongs to M_q. The proof consists in showing that the joint distribution of the processes $r(t) = r(\sigma(t))$ and $\bar{x}_q(t)$ is the same as described in Proposition 7. This is done through the following steps.

Step 1 – $r(t)$ **verifies** (45a): recall that, for any smooth function f on $\mathcal{M}$, the process z verifies (44a). If $f = r$, then by (44a) and (44b)

$$dr(t) = \frac{1}{2}\,\Delta_{\mathcal{M}} r(t)dt + dm^r(t) = \frac{1}{2}\,\frac{\partial_r G}{G}(r(t))dt + dm^r(t) \tag{75}$$

To prove that $r(t)$ verifies (45a), it is enough to prove that $dm^r(t) = dw(t)$ where w is a standard Brownian motion. Note that $dr^2(t)$ can be computed in two ways. The first way, by Itô's formula and (75)

$$dr^2 = r\Delta_{\mathcal{M}} r(t)dt + 2r(t)dm^r(t) + d[m^r](t)$$

where $[m^r](t)$ is the quadratic variation process of the local martingale $m^r(t)$ [25] (Theorem 17.16, p. 339). The second way, by (44a) with $f = r^2$,

$$\begin{aligned} dr^2 &= \left(I(\tilde{\nabla} r, \tilde{\nabla} r) + r\Delta_{\mathcal{M}} r(t)\right) dt + dm^{r^2}(t)dt \\ &= (1 + r\Delta_{\mathcal{M}} r(t))\; dt + dm^{r^2}(t)dt \end{aligned}$$

where the second line follows from (40) because $\tilde{\nabla} r = \partial_r$. By equating these two expressions of $dr^2(t)$, it follows that $dt - d[m^r](t)$ is the differential of a continuous local martingale of finite variation, and therefore identically zero [25] (Proposition 17.2, p. 330). In other words, $d[m^r](t) = dt$ and Lévy's characterisation implies that $dm^r(t) = dw(t)$, where w is a standard Brownian motion [25] (Theorem 18.3, p. 352). ■

Step 2 – $\bar{x}_q(t)$ **verifies** (45b): if f is a smooth function on $\mathcal{M}$, such that $f(z) = f(\bar{x}_q)$, then by (44a) and (44b)

$$df(\bar{x}_q(t)) = \frac{1}{2}\beta_q^{-2}(r(t))\,\Delta_{M_q}\, f(\bar{x}_q(t))\,dt + dm^f(t) \tag{76}$$

Define $l_q(t)$ to be the inverse of the time change process $\tau_q(t)$ defined in (45b), and let $\theta_q(t) = \left(\bar{x}_q \circ l_q\right)(t)$. By applying the time change $l_q(t)$ to (76)

$$df(\theta_q(t)) = \frac{1}{2}\,\Delta_{M_q}\, f(\theta_q(t))\,dt + d(m^f \circ l_q)(t)$$

where the first term on the right-hand side is obtained by replacing from the definition of $\tau_q(t)$ in (45b). Recall that a time change of a local martingale is a local martingale [25] (Theorem 17.24, p. 344). Therefore, $m^f \circ l_q$ is a local martingale, so θ_q solves the martingale problem associated to $(1/2)\Delta_{M_q}$. This means that θ_q is a Riemannian Brownian motion in M_q . ■

Step 3 – $r(t)$ **and** $\theta_q(t)$ **are independent**: it is required to prove that the processes $r(t)$ and $\theta_1(t)\,, \ldots,\, \theta_r(t)$ are jointly independent. A detailed proof is given only of the fact that $r(t)$ and $\theta_q(t)$ are independent, for any fixed q. The complete proof is obtained by repeating similar arguments.

The following proof is modeled on [23] (Example 3.3.3, p. 84). Let $\left(\xi^i(t)\,;\, i = 1\,,\ldots,\,\dim M_q\right)$ be the stochastic anti-development of $\bar{x}_q(t)$. Precisely [19] (Definition 8.23, p. 119)

$$d\xi^i(t) \,=\, Q\left(e^i\,,d\bar{x}_q(t)\right) \,=\, \beta_q^{-2}(r(t))\,I\left(e^i\,,dz(t)\right) \tag{77a}$$

where the e^i form a parallel orthonormal moving frame above the stochastic process $\bar{x}_q(t)$ in M_q. On the other hand, it is possible to write, using Itô's formula [19] (Proposition 7.34, p. 109)

$$dr(t) \,=\, I\left(\partial_r\,, dz(t)\right) \,+\, \frac{1}{2}\,\Delta_{\mathcal{M}} r(t)\,dt \tag{77b}$$

Let $[\xi^i, r]$ denote the quadratic covariation of ξ^i and r. From [19] (Proposition 5.18, p. 63), since z is a Riemannian Brownian motion, it follows from (77a) and (77b) that

$$d[\xi^i, r](t) \,=\, \beta_q^{-2}(r(t))\,I(e^i, \partial_r)(z(t))\,dt \,=\, 0 \tag{77c}$$

as e^i and ∂_r are orthogonal with respect to I. This means that $(\xi^i(t))$ and $r(t)$ have zero quadratic covariation, and therefore $(\xi^i(t))$ and $w(t)$ have zero quadratic covariation. It follows as in [23] (Lemma 3.3.4, p. 85), that $r(t)$ and $\theta_q(t)$ are independent. ■

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
2. Amari, S.I.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
3. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence (2000)
4. Arnaudon, M.: Semi-martingales dans les espaces homogènes. Annales de l'I.H.P **29**(3), 269–288 (1993)
5. Arnaudon, M., Miclo, L.: A stochastic algorithm finding generalized means on compact manifolds. Stochastic. Process. Appl. **124**(10), 3463–3479 (2014)
6. Atkinson, C., Mitchell, A.: Rao's distance measure. Sankhya Ser. A **43**, 345–365 (1981)
7. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information geometry and sufficient statistics. Probab. Theory Relat. Fields **162**(1), 327–364 (2015)
8. Bensadon, J.: Black-box optimization using geodesics in statistical manifolds. Entropy **17**(1), 304–345 (2015)
9. Bensadon, J.: Applications of information theory to machine learning. Ph.D. thesis, Université Paris-Saclay (2016). https://tel.archives-ouvertes.fr/tel-01297163
10. Berthoumieu, Y., Bombrun, L., Said, S.: Classification approach based on the product of Riemannian manifolds from Gaussian parameterization space. In: International Conference on Image Processing (ICIP) (2017)
11. Bishop, R.L., O'NEILL, B.: Manifolds of negative curvature. Trans. Amer. Math. Soc. **145**, 1–49 (1969)

12. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning (2017). arXiv:1606.04838v2
13. Chavel, I.: Riemannian Geometry, A Modern Introduction. Cambridge University Press, Cambridge (2006)
14. Cheng, G., Vemuri, B.C.: A novel dynamic system in the space of SPD matrices with applications to appearance tracking. SIAM J. Imaging Sci. **6**(1), 592–615 (2013)
15. Chentsov, N.N.: Statistical Decision Rules and Optimal Inference. American Mathematical Society, Providence (1982)
16. Chikuse, Y.: Statistics on Special Manifolds. Lecture Notes in Statistics, 174. Springer Science+Business Media, LLC (2003)
17. Dobarro, F., Unal, B.: Curvature of multiply warped products. J. Geom. Phys. **55**, 75–106 (2005)
18. Do Carmo, M.P.: Riemannian Geometry, 2nd edn. Birkhauser, Boston (1992)
19. Emery, M.: Stochastic Calculus in Manifolds. Springer, Berlin (1980)
20. Estrade, A.: Exponentielle stochastique et intégrale multiplicative discontinues. Annales de l'I.H.P. **28**(1), 107–129 (1992)
21. Gallavotti, G.: The Elements of Mechanics. Springer Science+Business Media, LLC, New York (1983)
22. Helgason, S.: Differential Geometry and Symmetric Spaces. Academic Press, New York (1962)
23. Hsu, E.P.: Stochastic Analysis on Manifolds. American Mathematical Society, Providence (2002)
24. Ikeda, N., Watanabe, S.: Stochastic Differential Equations and Diffusion Processes. North-Holland Publishing Company, Amsterdam (1981)
25. Kallenberg, O.: Foundations of Modern Probability, 2nd edn. Springer, New York (2001)
26. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, New York (2011)
27. Knapp, A.W.: Lie Groups Beyond an Introduction, 2nd edn. Birkhauser, Boston (2002)
28. Lang, S.: Introduction to Linear Algebra, 2nd edn. Springer, New York (1986)
29. Lee, J.M.: Introduction to Smooth Manifolds, 2nd edn. Springer Science+Business Media, New York (2012)
30. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer Science+Business Media Inc, New York (2005)
31. Liao, E.: Lévy Processes on Lie Groups. Cambridge University Press, Cambridge (2004)
32. Luenbeger, D.G.: Linear and Nonlinear Programming, 2nd edn. Addison-Wesley publishing company, Reading (1973)
33. Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley, Chichester (2002)
34. Marsden, J.E., Ratiu, T.: Manifolds, Tensor Analysis, and Applications. Springer Publishing Company Inc., New York (2001)
35. Martens, J.: New insights and perspectives on the natural gradient method (2017). arXiv:1412.1193v8
36. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley, Hoboken (2004)
37. Ollivier, Y., Arnold, L., Auger, A., Hansen, N.: Information geometric optimization algorithms: a unifying picture via invariance principles. J. Mach. Learn. Res. **18**(18), 1–65 (2017)
38. O'Neill, B.: Semi-Riemannian Geometry with Applications to Relativity. Academic Press, San Diego (1983)
39. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. Int. J. Comput. Vision **66**(1), 41–66 (2006)
40. Petersen, P.: Riemannian Geometry, 2nd edn. Springer Science+Business Media, LLC, New York (2006)
41. Said, S., Bombrun, L., Berthoumieu, Y., Manton, J.H.: Riemannian Gaussian distributions on the space of symmetric positive definite matrices. IEEE. Trans. Inf. Theory **63**(4), 2153–2170 (2017)

42. Said, S., Hajri, H., Bombrun, L., Vemuri, B.C.: Gaussian distributions on Riemannian symmetric spaces: statistical learning with structured covariance matrices. IEEE. Trans. Inf. Theory **64**(2), 752–772 (2018)
43. Spanier, E.H.: Algebraic Topology. Springer, New York (1966)
44. Terras, A.: Harmonic Analysis on Symmetric Spaces and Applications, vol. II. Springer, New York (1988)
45. Unal, B.: Doubly warped products. Differ. Geom. Appl. **15**, 253–263 (2001)
46. Watson, G.N.: A Treatise on the Theory of Bessel Functions. Cambridge University Press, Cambridge (1922)
47. Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., Schmidhuber, J.: Natural evolution strategies. J. Mach. Learn. Res. **15**(1), 949–980 (2014)
48. Young, G.A., Smith, R.L.: Essentials of Statistical Inference. Cambridge University Press, Cambridge (2005)
49. Zanini, P., Said, S., Congedo, M., Berthoumieu, Y., Jutten, C.: Parameter estimates of Riemannian Gaussian distributions in the manifold of covariance matrices. In: Sensor Array and Multichannel Signal Processsing Workshop (SAM) (2016)
50. Zanini, P., Said, S., Berthoumieu, Y., Congedo, M., Jutten, C.: Riemannian online algorithm for estimating mixture model parameters. In: Geometric Science of Information (GSI) (2017)

Clustering in Hilbert's Projective Geometry: The Case Studies of the Probability Simplex and the Elliptope of Correlation Matrices

Frank Nielsen and Ke Sun

Abstract Clustering categorical distributions in the probability simplex is a fundamental task met in many applications dealing with normalized histograms. Traditionally, differential-geometric structures of the probability simplex have been used either by (i) setting the Riemannian metric tensor to the Fisher information matrix of the categorical distributions, or (ii) defining the dualistic information-geometric structure induced by a smooth dissimilarity measure, the Kullback–Leibler divergence. In this work, we introduce for this clustering task a novel computationally-friendly framework for modeling the probability simplex termed *Hilbert simplex geometry*. In the Hilbert simplex geometry, the distance function is described by a polytope. We discuss the pros and cons of those different statistical modelings, and benchmark experimentally these geometries for center-based k-means and k-center clusterings. Furthermore, since a canonical Hilbert metric distance can be defined on any bounded convex subset of the Euclidean space, we also consider Hilbert's projective geometry of the elliptope of correlation matrices and study its clustering performances.

Keywords Fisher-Riemannian geometry · Information geometry
Multinoulli distribution · Hilbert simplex geometry · Polytope distance
Elliptope · Finsler geometry · k-means clustering · k-center clustering

F. Nielsen (✉)
Sony Computer Science Laboratories, Tokyo, Japan
e-mail: Frank.Nielsen@acm.org

K. Sun
CSIRO Data61, Sydney, Australia
e-mail: Ke.Sun@data61.csiro.au

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication
Technology, https://doi.org/10.1007/978-3-030-02520-5_11

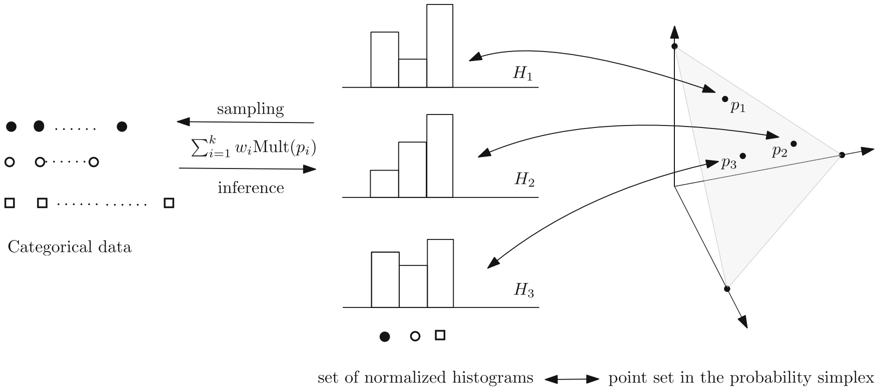

sampling
$\sum_{i=1}^{k} w_i \mathrm{Mult}(p_i)$
inference
Categorical data
H_1
H_2
H_3
p_1
p_2
p_3
set of normalized histograms ←→ point set in the probability simplex

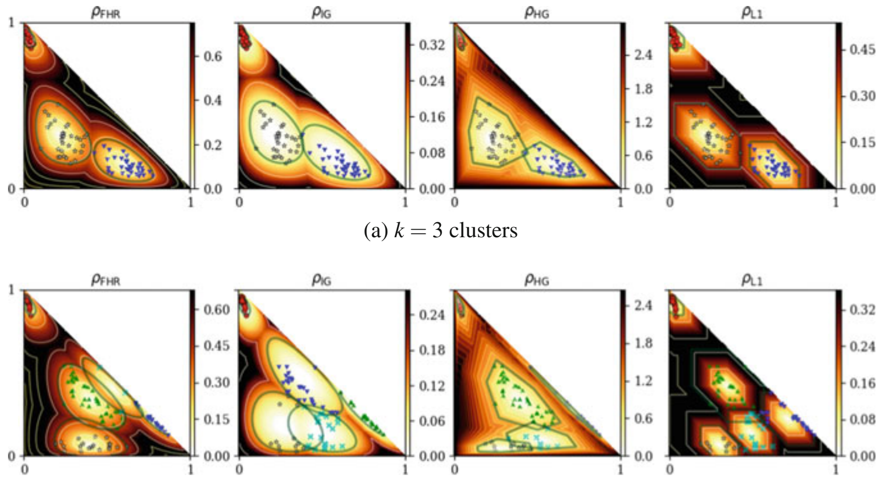
Fig. 1 Categorical datasets modeled by a generative statistical mixture model of multinoulli distributions can be visualized as a weighted set of normalized histograms or equivalently by a weighted point set encoding multinoulli distributions in the probability simplex Δ^d (here, $d = 2$ for trinoulli distributions — trinomial distributions with a single trial)

1 Introduction and Motivation

The categorical distributions and multinomial distributions are important probability distributions often met in data analysis [1], text mining [2], computer vision [3] and machine learning [4]. A multinomial distribution over a set $\mathscr{X} = \{e_0, \ldots, e_d\}$ of outcomes (e.g., the $d + 1$ distinct colored faces of a die) is defined as follows: Let $\lambda_p^i > 0$ denote the probability that outcome e_i occurs for $i \in \{0, \ldots, d\}$ (with $\sum_{i=0}^{d} \lambda_p^i = 1$). Denote by m the total number of events, with m_i reporting the number of outcome e_i. Then the probability $\Pr(X_0 = m_0, \ldots, X_d = m_d)$ that a multinomial random variable $X = (X_0, \ldots, X_d) \sim \mathrm{Mult}(p = (\lambda_p^0, \ldots, \lambda_p^d), m)$ (where X_i count the number of events e_i, and $\sum_{i=0}^{d} m_i = m$) is given by the following probability mass function (pmf):

$$\Pr(X_0 = m_0, \ldots, X_d = m_d) = \frac{m!}{\prod_{i=0}^{d} m_i!} \prod_{i=0}^{d} \left(\lambda_p^i\right)^{m_i}.$$

The multinomial distribution is called a binomial distribution when $d = 1$ (e.g., coin tossing), a Bernoulli distribution when $m = 1$, and a "multinoulli distribution" (or categorical distribution) when $m = 1$ and $d > 1$. The multinomial distribution is also called a generalized Bernoulli distribution. A random variable X following a multinoulli distribution is denoted by $X = (X_0, \ldots, X_d) \sim \mathrm{Mult}(p = (\lambda_p^0, \ldots, \lambda_p^d))$. The multinomial/multinoulli distribution provides an important *feature representation* in machine learning that is often met in applications [5–7] as normalized histograms (with non-empty bins) as illustrated in Fig. 1.



(a) $k = 3$ clusters

(b) $k = 5$ clusters

Fig. 2 Visualizing some k-center clustering results on a toy dataset in the space of trinomials Δ^2 for the considered four types of distances (and underlying geometries): Fisher-Hotelling-Rao metric distance (Riemannian geometry), Kullback–Leibler non-metric divergence (information geometry), Hilbert metric distance (Hilbert projective geometry), and total variation/L_1 metric distance (norm geometry). Observe that the L_1 balls have hexagonal shapes on the probability simplex (intersection of a rotated cube with the plane H_{Δ^d}). The color density maps indicate the distance from any point to its nearest cluster center

A multinomial distribution $p \in \Delta^d$ can be thought as a point lying in the probability simplex Δ^d (standard simplex) with coordinates $p = (\lambda_p^0, \ldots, \lambda_p^d)$ such that $\lambda_p^i = \Pr(X = e_i) > 0$ and $\sum_{i=0}^d \lambda_p^i = 1$. The open probability simplex Δ^d can be embedded in $\mathbb{R}^{d+1}$ on the hyperplane $H_{\Delta^d} : \sum_{i=0}^d x^i = 1$. Notice that observations with D categorical attributes can be clustered using k-mode [8] with respect to the Hamming distance. Here, we consider the different task of clustering a set $\Lambda = \{p_1, \ldots, p_n\}$ of n categorical/multinomial distributions in Δ^d [5] using center-based k-means++ or k-center clustering algorithms [9, 10], which rely on a dissimilarity measure (loosely called distance or divergence when smooth) between any two categorical distributions. In this work, we mainly consider four distances with their underlying geometries: (1) Fisher-Hotelling-Rao distance ρ_{FHR} (spherical geometry), (2) Kullback–Leibler divergence ρ_{IG} (dually flat geometry), (3) Hilbert distance ρ_{HG} (generalize Klein's hyperbolic geometry), and (4) the total variation/L1 distance (norm geometry). The geometric structures of spaces are necessary in algorithms, for example, to define midpoint distributions. Figure 2 displays the k-center clustering results obtained with these four geometries as well as the L^1 distance ρ_{L1} normed geometry on toy synthetic datasets in Δ^2. We shall now explain the Hilbert simplex geometry applied to the probability simplex, describe how to perform k-center clustering in Hilbert geometry, and report experimental results that demonstrate the superiority of the Hilbert geometry when clustering multinomials and correlation matrices.

The rest of this paper is organized as follows: Sect. 2 formally introduces the distance measures in Δ^d. Section 3 introduces how to efficiently compute the Hilbert distance. Section 4 presents algorithms for Hilbert minimax centers and Hilbert center-based clustering. Section 5 performs an empirical study of clustering multinomial distributions, comparing Riemannian geometry, information geometry, and Hilbert geometry. Section 6 presents a second use case of Hilbert geometry in machine learning: clustering correlation matrices in the elliptope [11]. Finally, Sect. 7 concludes this work by summarizing the pros and cons of each geometry. Although some contents require prior knowledge on geometric structures, we will present the detailed algorithms so that the general audience can still benefit from this work.

2 Four Distances with their Underlying Geometries

2.1 Fisher-Hotelling-Rao Riemannian Geometry

The Rao distance between two multinomial distributions is [6, 12]:

$$\rho_{\mathrm{FHR}}(p,q) = 2 \arccos\left(\sum_{i=0}^{d} \sqrt{\lambda_p^i \lambda_q^i}\right). \tag{1}$$

It is a Riemannian metric length distance (satisfying the symmetric and triangular inequality axioms) obtained by setting the metric tensor g to the *Fisher information matrix* (FIM) $\mathscr{I}(p) = (g_{ij}(p))_{d\times d}$ with respect to the coordinate system $(\lambda_p^1, \ldots, \lambda_p^d)$, where

$$g_{ij}(p) = \frac{\delta_{ij}}{\lambda_p^i} + \frac{1}{\lambda_p^0}.$$

We term this geometry the *Fisher-Hotelling-Rao (FHR) geometry* [13–16]. The metric tensor g allows one to define an inner product on each tangent plane T_p of the probability simplex manifold: $\langle u,\ v\rangle_p = u^\top g(p) v$. When g is everywhere the identity matrix, we recover the Euclidean (Riemannian) geometry with the inner product being the scalar product: $\langle u, v\rangle = u^\top v$. The geodesics $\gamma(p, q; \alpha)$ are defined by the Levi-Civita metric connection [17, 18] that is derived from the metric tensor. The FHR manifold can be embedded in the positive orthant of an Euclidean d-sphere in $\mathbb{R}^{d+1}$ by using the *square root representation* $\lambda \mapsto \sqrt{\lambda}$ [12]. Therefore the FHR manifold modeling of Δ^d has constant *positive* curvature: It is a spherical geometry restricted to the positive orthant with the metric distance measuring the arc length on a great circle.

2.2 Information Geometry

A divergence D is a smooth C^3 differentiable dissimilarity measure [19] that allows to define a dual structure in Information Geometry (IG), see [17, 18, 20]. A f-divergence is defined for a strictly convex function f with $f(1) = 0$ by:

$$I_f(p : q) = \sum_{i=0}^{d} \lambda_p^i f\left(\frac{\lambda_q^i}{\lambda_p^i}\right) \geq f(1) = 0.$$

It is a *separable* divergence since the d-variate divergence can be written as a sum of d univariate (scalar) divergences: $I_f(p : q) = \sum_{i=0}^{d} I_f(\lambda_p^i : \lambda_q^i)$. The class of f-divergences plays an essential role in information theory since they are provably the *only* separable divergences that satisfy the *information monotonicity* property [17, 21] (for $d \geq 2$). That is, by coarse-graining the histograms, we obtain lower-dimensional multinomials, say p' and q', such that $0 \leq I_f(p' : q') \leq I_f(p : q)$ [17]. The Kullback–Leibler (KL) divergence ρ_{IG} is a f-divergence obtained for the functional generator $f(u) = -\log u$:

$$\rho_{\mathrm{IG}}(p, q) = \sum_{i=0}^{d} \lambda_p^i \log \frac{\lambda_p^i}{\lambda_q^i}. \tag{2}$$

It is an asymmetric non-metric distance: $\rho_{\mathrm{IG}}(p, q) \neq \rho_{\mathrm{IG}}(q, p)$. In differential geometry, the structure of a manifold is defined by two independent components:

1. A *metric tensor* g that allows to define an inner product $\langle \cdot, \cdot \rangle_p$ at each tangent space (for measuring vector lengths and angles between vectors);
2. A *connection* ∇ that defines *parallel transport* $\prod_c^{\nabla}$, i.e., a way to move a tangent vector from one tangent plane T_p to any other one T_q along a smooth curve c, with $c(0) = p$ and $c(1) = q$.

In FHR geometry, the implicitly-used connection is called the Levi-Civita connection that is induced by the metric g: $\nabla^{\mathrm{LC}} = \nabla(g)$. It is a metric connection since it ensures that $\langle u, v \rangle_p = \langle \prod_{c(t)}^{\nabla^{\mathrm{LC}}} u, \prod_{c(t)}^{\nabla^{\mathrm{LC}}} v \rangle_{c(t)}$ for $t \in [0, 1]$. The underlying information-geometric structure of KL is characterized by a pair of *dual* connections [17] $\nabla = \nabla^{(-1)}$ (mixture connection) and $\nabla^* = \nabla^{(1)}$ (exponential connection) that induces a corresponding pair of dual geodesics (technically, ±1-autoparallel curves, [18]). Those connections are said *flat* as they define two dual global affine coordinate systems θ and η on which the θ- and η-geodesics are (Euclidean) straight line segments, respectively. For multinomials, the *expectation parameters* are: $\eta = (\lambda^1, \ldots, \lambda^d)$ and they one-to-one correspond to the *natural parameters*: $\theta = \left(\log \frac{\lambda^1}{\lambda^0}, \ldots, \log \frac{\lambda^d}{\lambda^0}\right) \in \mathbb{R}^d$. Thus in IG, we have two kinds of midpoint multinomials of p and q, depending on whether we perform the (linear) interpolation on the θ- or the η-geodesics. Informally speaking, the dual connections $\nabla^{(\pm 1)}$ are said coupled to the FIM since we have $\frac{\nabla + \nabla^*}{2} = \nabla(g) = \nabla^{\mathrm{LC}}$. Those dual (torsion-free affine) connections are not metric connections but enjoy the following metric-compatibility property when used together as follows: $\langle u, v\rangle_p = \langle \prod_{c(t)} u, \prod^*{}_{c(t)} v\rangle_{c(t)}$ (for $t \in [0, 1]$), where $\prod := \prod^{\nabla}$ and $\prod^* := \prod^{\nabla^*}$ are the corresponding induced dual parallel transports. The geometry of f-divergences [19] is the α-geometry (for $\alpha = 3 + 2f'''(1)$) with the dual $\pm\alpha$-connections, where $\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^* + \frac{1-\alpha}{2}\nabla$. The Levi-Civita metric connection is $\nabla^{\mathrm{LC}} = \nabla^{(0)}$. More generally, it was shown how to build a dual information-geometric structure for *any* divergence [19]. For example, we can build a dual structure from the symmetric Cauchy–Schwarz divergence [22]:

$$\rho_{\mathrm{CS}}(p, q) = -\log \frac{\langle \lambda_p, \lambda_q\rangle}{\sqrt{\langle \lambda_p, \lambda_p\rangle \langle \lambda_q, \lambda_q\rangle}}. \tag{3}$$

2.3 Hilbert Simplex Geometry

In Hilbert geometry (HG), we are given a bounded convex domain $\mathscr{C}$ (here, $\mathscr{C} = \Delta^d$), and the distance between any two points M, M' of $\mathscr{C}$ is defined [23] as follows: Consider the two intersection points AA' of the line (MM') with $\mathscr{C}$, and order them on the line so that we have A, M, M', A'. Then the Hilbert metric distance [24] is defined by:

$$\rho_{\mathrm{HG}}(M, M') = \begin{cases} \left|\log \frac{|A'M||AM'|}{|A'M'||AM|}\right|, & M \neq M', \\ 0 & M = M'. \end{cases} \tag{4}$$

It is also called the Hilbert cross-ratio metric distance [25, 26]. Notice that we take the absolute value of the logarithm since the Hilbert distance is a *signed distance* [27]. When $\mathscr{C}$ is the unit ball, HG lets us recover the Klein hyperbolic geometry [26]. When $\mathscr{C}$ is a quadric bounded convex domain, we obtain the Cayley–Klein hyperbolic geometry [28] which can be studied with the Riemannian structure and the corresponding metric distance called the curved Mahalanobis distances [29, 30]. Cayley–Klein hyperbolic geometries have negative curvature. Elements on the boundary are called ideal elements [31].

In Hilbert geometry, the geodesics are *straight* Euclidean lines making them convenient for computation. Furthermore, the domain boundary $\partial\mathscr{C}$ needs not to be smooth: One may also consider bounded polytopes [32]. This is particularly interesting for modeling Δ^d, the d-dimensional open standard simplex. We call this geometry the *Hilbert simplex geometry* [33]. In Fig. 3, we show that the Hilbert distance between two multinomial distributions p (M) and q (M') can be computed by finding the two intersection points of the line $(1 - t)p + tq$ with $\partial\Delta^d$, denoted as $t_0 \leq 0$ and $t_1 \geq 1$. Then

$$\rho_{\mathrm{HG}}(p, q) = \left|\log \frac{(1 - t_0)t_1}{(-t_0)(t_1 - 1)}\right| = \log\left(1 - \frac{1}{t_0}\right) - \log\left(1 - \frac{1}{t_1}\right).$$

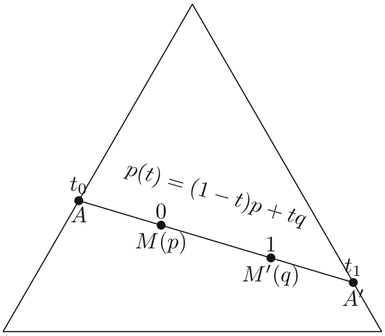


Fig. 3 Computing the Hilbert metric distance for trinomials on the 2D probability simplex as the logarithm of the cross ratio $(M, M'; A, A')$ of the four collinear points A, M, M' and A'

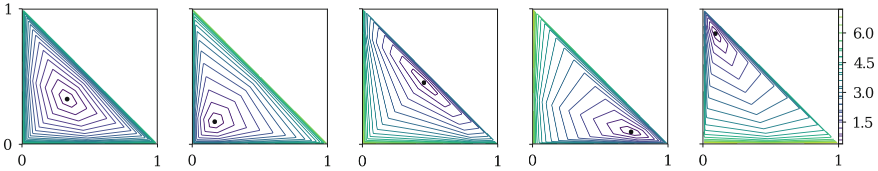


Fig. 4 Balls in the Hilbert simplex geometry Δ^2 have polygonal Euclidean shapes of constant combinatorial complexity. At infinitesimal scale, the balls have hexagonal shapes, showing that the Hilbert geometry is not Riemannian

The shape of balls in polytope-domain HG is Euclidean polytopes[1] [26], as depicted in Fig. 4. Furthermore, the Euclidean shape of the balls does not change with the radius. Hilbert balls have hexagons shapes in 2D [34], rhombic dodecahedra shapes in 3D, and are polytopes [26] with $d(d+1)$ facets in dimension d. When the polytope domain is not a simplex, the combinatorial complexity of balls depends on the center location [34], see Fig. 5. The HG of the probability simplex yields a non-Riemannian geometry, because, at an infinitesimal radius, the balls are polytopes and not ellipsoids (corresponding to squared Mahalanobis distance balls used to visualize metric tensors [35]). The isometries in Hilbert polyhedral geometries are studied in [36]. In Appendix 9, we recall that any Hilbert geometry induces a Finslerian structure that becomes Riemannian iff the boundary is an ellipsoid (yielding the hyperbolic Cayley–Klein geometries [27]). Notice that in Hilbert simplex/polytope geometry, the geodesics are not unique (see Figure 2 of [25]).

[1] To contrast with this result, let us mention that infinitesimal small balls in Riemannian geometry have Euclidean ellipsoidal shapes (visualized as Tissot's indicatrix in cartography).

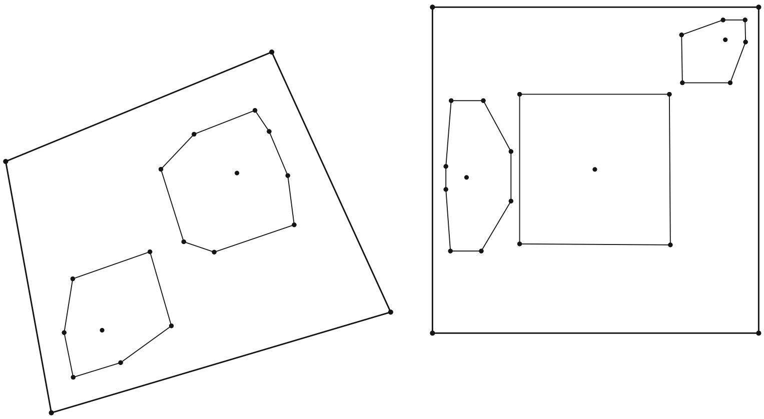


Fig. 5 Hilbert balls in quadrangle domains have combinatorial complexity depending on the center location

2.4 L_1-Norm Geometry

The Total Variation (TV) metric distance between two multinomials p and q is defined by:

$$\mathrm{TV}(p,q) = \frac{1}{2}\sum_{i=0}^{d} |\lambda_p^i - \lambda_q^i|.$$

It is a statistical f-divergence obtained for the generator $f(u) = \frac{1}{2}|u-1|$. The L_1-norm induced distance ρ_{L1} (L1) is defined by:

$$\rho_{L1}(p,q) = \|\lambda_p - \lambda_q\|_1 = \sum_{i=0}^{d} |\lambda_p^i - \lambda_q^i| = 2\mathrm{TV}(p,q).$$

Therefore the distance ρ_{L1} satisfies information monotonicity (for coarse-grained histograms p' and q' of $\Delta^{D'}$ with $D' < D$):

$$0 \leq \rho_{L1}(p', q') \leq \rho_{L1}(p,q).$$

For trinomials, the ρ_{L1} distance is given by:

$$\rho_{L1}(p,q) = |\lambda_p^0 - \lambda_q^0| + |\lambda_p^1 - \lambda_q^1| + |\lambda_q^0 - \lambda_p^0 + \lambda_q^1 - \lambda_p^1|.$$

The L_1 distance function is a polytopal distance function described by the dual polytope $\mathscr{Z}$ of the d-dimensional cube called the standard (or regular) d-cross-polytope [37], the orthoplex [38] or the d-cocube [39]: The cross-polytope $\mathscr{Z}$ can be obtained as the convex hull of the $2d$ unit standard base vectors $\pm e_i$ for

Table 1 Comparing the geometric modelings of the probability simplex Δ^d

	Riemannian Geometry	Information Rie. Geo.	Non-Rie. Hilbert Geo.
Structure	$(\Delta^d, g, \nabla^{\mathrm{LC}} = \nabla(g))$	$(\Delta^d, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$	(Δ^d, ρ)
	Levi-Civita $\nabla^{\mathrm{LC}} = \nabla^{(0)}$	Dual connections $\nabla^{(\pm\alpha)}$ so that $\frac{\nabla^{(\alpha)}+\nabla^{(-\alpha)}}{2} = \nabla^{(0)}$	Connection of $\mathbb{R}^d$
Distance	Rao distance (metric)	α-divergence (non-metric) KL or reverse KL for $\alpha = \pm 1$	Hilbert distance (metric)
Property	Invariant to reparameterization	Information monotonicity	Isometric to a normed space
Calculation	Closed-form	Closed-form	Easy (Algorithm 1)
Geodesic	Shortest path	Straight either in θ/η	Straight
Smoothness	Manifold	Manifold	Non-manifold
Curvature	Positive	Dually flat	Negative

$i \in \{0, \ldots, d-1\}$. The cross-polytope is one of the three regular polytopes in dimension $d \geq 4$ (with the hypercubes and simplices): It has $2d$ vertices and 2^d facets. Therefore the L_1 balls on the hyperplane H_{Δ^d} supporting the probability simplex is the intersection of a $(d+1)$-cross-polytope with d-dimensional hyperplane H_{Δ^d}. Thus the "multinomial ball" $\mathrm{Ball}_{L_1}(p, r)$ of center p and radius r is defined by $\mathrm{Ball}_{L_1}(p, r) = (\lambda_p \oplus r\mathscr{Z}) \cap H_{\Delta^d}$. In 2D, the shape of L_1 trinomial balls is that of a regular octahedron (twelve edges and eight faces) cut by the 2D plane H_{Δ^2}: Trinomial balls have hexagonal shapes as illustrated in Fig. 2 (for ρ_{L1}). In 3D, trinomial balls are Archimedean solid cuboctahedra, and in arbitrary dimension, the shapes are polytopes with $d(d+1)$ vertices [40]. Let us note in passing, that in 3D, the L_1 multinomial cuboctahedron ball has the dual shape of the Hilbert rhombic dodecahedron ball.

Table 1 summarizes the characteristics of the three main geometries: FHR, IG, and HG. Let us conclude this introduction by mentioning the Cramér–Rao lower bound and its relationship with information geometry [41]: Consider an unbiased estimator $\hat{\theta} = T(X)$ of a parameter θ estimated from measurements distributed according to a smooth density $p(x;\, \theta)$ (i.e., $X \sim p(x;\, \theta)$). The Cramér–Rao Lower Bound (CRLB) states that the variance of $T(X)$ is greater or equal to the inverse of the FIM $\mathscr{I}(\theta)$: $V_\theta[T(X)] \succ \mathscr{I}^{-1}(\theta)$. For regular parametric families $\{p(x;\, \theta)\}_\theta$, the FIM is a positive-definite matrix and defines a metric tensor, called the Fisher metric in Riemannian geometry. The FIM is the cornerstone of information geometry [17] but requires the differentiability of the probability density function (pdf).

A better lower bound that does not require the pdf differentiability is the Hammersley–Chapman–Robbins Lower Bound [42, 43] (HCRLB):

$$V_\theta[T(X)] \geq \sup_{\Delta} \frac{\Delta^2}{E_\theta\left[\left(\frac{p(x;\theta+\Delta)-p(x;\theta)}{p(x;\theta)}\right)^2\right]}. \tag{5}$$

By introducing the χ^2-divergence, $\chi^2(P:Q)=\int\left(\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)^2\mathrm{d}Q$, we rewrite the HCRLB using the χ^2-divergence in the denominator as follows:

$$V_\theta[T(X)]\geq\sup_{\Delta}\frac{\Delta^2}{\chi^2(P(x;\theta+\Delta):P(x;\theta))}. \tag{6}$$

Note that the FIM is not defined for non-differentiable pdfs, and therefore the Cramér–Rao lower bound does not exist in that case.

3 Computing Hilbert Distance in Δ^d

Let us start with the simplest case: The 1D probability simplex Δ^1, the space of Bernoulli distributions. Any Bernoulli distribution can be represented by the activation probability of the random bit x: $\lambda=p(x=1)\in\Delta^1$, corresponding to a point in the interval $\Delta^1=(0,1)$. We write the Bernoulli manifold as an exponential family as

$$p(x)=\exp\left(x\theta-F(\theta)\right),\quad x\in\{0,\ 1\},$$

where $F(\theta)=\log(1+\exp(\theta))$. Therefore $\lambda=\frac{\exp(\theta)}{1+\exp(\theta)}$ and $\theta=\log\frac{\lambda}{1-\lambda}$.

3.1 1D Probability Simplex of Bernoulli Distributions

By definition, the Hilbert distance has the closed form:

$$\rho_{\mathrm{HG}}(p,q)=\left|\log\frac{\lambda_q(1-\lambda_p)}{\lambda_p(1-\lambda_q)}\right|=\left|\log\frac{\lambda_p}{1-\lambda_p}-\log\frac{\lambda_q}{1-\lambda_q}\right|.$$

Note that $\theta_p=\log\frac{\lambda_p}{1-\lambda_p}$ is the canonical parameter of the Bernoulli distribution.

The FIM of the Bernoulli manifold in the λ-coordinates is given by: $g=\frac{1}{\lambda}+\frac{1}{1-\lambda}=\frac{1}{\lambda(1-\lambda)}$. The FHR distance is obtained by integration as:

$$\rho_{\mathrm{FHR}}(p,q)=2\arccos\left(\sqrt{\lambda_p\lambda_q}+\sqrt{(1-\lambda_p)(1-\lambda_q)}\right).$$

Notice that $\rho_{\mathrm{FHR}}(p,q)$ has finite values on $\partial\Delta^1$.

The KL divergence of the ±1-geometry is:

$$\rho_{\mathrm{IG}}(p,q)=\lambda_p\log\frac{\lambda_p}{\lambda_q}+(1-\lambda_p)\log\frac{1-\lambda_p}{1-\lambda_q}.$$

The KL divergence belongs to the family of α-divergences [17].

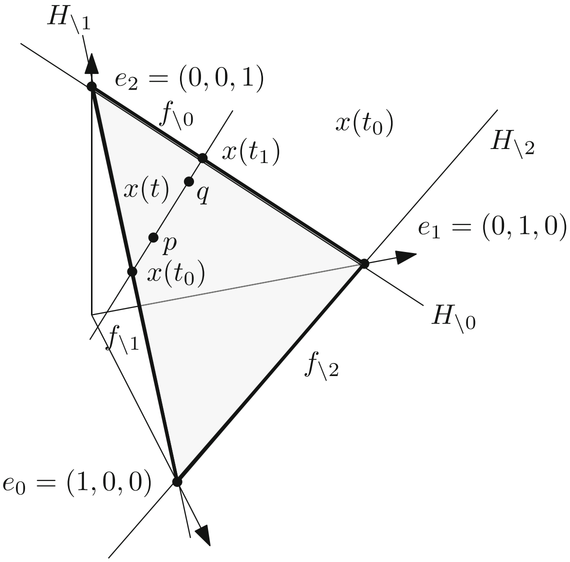


Fig. 6 Calculating the two intersection points $x(t_0)$ and $x(t_1)$ of the line (pq) with the boundary of the probability simplex Δ_d: For each facet $f_{\setminus i}$, we calculate the intersection point of line $x(t) = (1-t)p + tq$ with the d-dimensional hyperplane $H_{\setminus i}$ supporting the facet $f_{\setminus i}$

3.2 Arbitrary Dimension Case

Given $p, q \in \Delta^d$, we first need to compute the intersection of line (pq) with the border of the d-dimensional probability simplex to get the two intersection points p' and q' so that p', p, q, q' are ordered on (pq). Once this is done, we simply apply the formula in Eq. 4 to get the Hilbert distance.

A d-dimensional simplex consists of $d + 1$ vertices with their corresponding $(d - 1)$-dimensional facets. For the probability simplex Δ^d, let $e_i = (\underbrace{0, \ldots, 0}_{i}, 1, 0, \ldots, 0)$ denote the $d + 1$ vertices of the standard simplex embedded in the hyperplane $H_\Delta : \sum_{i=0}^{d} \lambda^i = 1$ in $\mathbb{R}^{d+1}$. Let $f_{\setminus j}$ denote the simplex facets that is the convex hull of all vertices except e_j: $f_{\setminus j} = \text{hull}(e_0, \ldots, e_{j-1}, e_{j+1}, \ldots, e_d)$. Let $H_{\setminus j}$ denote the hyperplane supporting this facet, which is the affine hull $f_{\setminus j} = \text{affine}(e_0, \ldots, e_{j-1}, e_{j+1}, \ldots, e_d)$.

To compute the two intersection points of (pq) with Δ^d, a naive algorithm consists in computing the unique intersection point r_j of the line (pq) with each hyperplane $H_{\setminus j}$ $(j = 0, \ldots, d)$ and checking whether r_j belongs to $f_{\setminus j}$.

A much more efficient implementation given by Algorithm (1) calculates the intersection point of the line $x(t) = (1-t)p + tq$ with each $H_{\setminus j}$ $(j = 0, \ldots, d)$. These intersection points are represented using the coordinate t. For example, $x(0) = p$ and $x(1) = q$. Due to convexity, any intersection point with $H_{\setminus j}$ must satisfy either $t \leq 0$ or $t \geq 1$. Then, the two intersection points with $\partial \Delta^d$ are obtained by $t_0 = \max\{t : \exists j,\ x(t) \in H_{\setminus j} \text{ and } t \leq 0\}$ and $t_1 = \min\{t : \exists j,\ x(t) \in H_{\setminus j} \text{ and } t \geq 1\}$. Figure 6 illustrates this calculation method. This algorithm only requires $O(d)$ time and $O(1)$ memory.

Lemma 1 *The Hilbert distance in the probability simplex can be computed in optimal $\Theta(d)$ time.*

Algorithm 1: Computing the Hilbert distance

```
Data: Two points $p = (\lambda_p^0, \ldots, \lambda_p^d)$, $q = (\lambda_q^0, \ldots, \lambda_q^d)$ in the $d$-dimensional simplex $\Delta^d$
Result: Their Hilbert distance $\rho_{\mathrm{HG}}(p, q)$
begin
    $t_0 \leftarrow -\infty$; $t_1 \leftarrow +\infty$;
    for $i = 0 \ldots d$ do
        if $\lambda_p^i \neq \lambda_q^i$ then
            $t \leftarrow \lambda_p^i / (\lambda_p^i - \lambda_q^i)$;
            if $t_0 < t \leq 0$ then
                $t_0 \leftarrow t$;
            else if $1 \leq t < t_1$ then
                $t_1 \leftarrow t$;
    if $t_0 = -\infty$ or $t_1 = +\infty$ then
        Output $\rho_{\mathrm{HG}}(p, q) = 0$;
    else if $t_0 = 0$ or $t_1 = 1$ then
        Output $\rho_{\mathrm{HG}}(p, q) = \infty$;
    else
        Output $\rho_{\mathrm{HG}}(p, q) = \left| \log(1 - \frac{1}{t_0}) - \log(1 - \frac{1}{t_1}) \right|$;
```

Once an arbitrary distance ρ is chosen, we can define a ball centered at c and of radius r as $B_\rho(c, r) = \{x \ : \ \rho(c, x) \leq r\}$. Figure 4 displays the hexagonal shapes of the Hilbert balls for various center locations in Δ^2.

Theorem 1 (Balls in a simplicial Hilbert geometry [26]) *A ball in the Hilbert simplex geometry has a Euclidean polytope shape with $d(d+1)$ facets.*

Note that when the domain is not simplicial, the Hilbert balls can have varying combinatorial complexity depending on the center location. In 2D, the Hilbert ball can have $s \sim 2s$ edges inclusively, where s is the number of edges of the boundary of the Hilbert domain $\partial\mathscr{C}$.

Since a Riemannian geometry is locally defined by a metric tensor, at infinitesimal scales, Riemannian balls have Mahalanobis smooth ellipsoidal shapes: $B_\rho(c, r) = \{x \ : \ (x - c)^\top g(c)(x - c) \leq r^2\}$. This property allows one to visualize Riemannian metric tensors [35]. Thus we conclude that:

Lemma 2 ([26]) *Hilbert simplex geometry is a non-manifold metric length space.*

As a remark, let us notice that slicing a simplex with a hyperplane does not always produce a lower-dimensional simplex. For example, slicing a tetrahedron by a plane yields either a triangle or a quadrilateral. Thus the restriction of a d-dimensional ball B in a Hilbert simplex geometry Δ^d to a hyperplane H is a $(d-1)$-dimensional ball $B' = B \cap H$ of varying combinatorial complexity, corresponding to a ball in the induced Hilbert sub-geometry in the convex sub-domain $H \cap \Delta^d$.

3.3 Visualizing Distance Profiles

Figure 7 displays the distance profile from any point in the probability simplex to a fixed reference point (trinomial) based on the following common distance measures [18]: Euclidean (metric) distance, Cauchy–Schwarz (CS) divergence, Hellinger (metric) distance, Fisher-Rao (metric) distance, KL divergence and Hilbert simplicial (metric) distance. The Euclidean and Cauchy–Schwarz divergence are clipped to Δ^2. The Cauchy–Schwarz distance is projective so that $\rho_{\mathrm{CS}}(\lambda p, \lambda' q) = \rho_{\mathrm{CS}}(p, q)$ for any $\lambda, \lambda' > 0$ [44].

4 Center-Based Clustering

We concentrate on comparing the efficiency of Hilbert simplex geometry for clustering multinomials. We shall compare the experimental results of k-means++ and k-center multinomial clustering for the three distances: Rao and Hilbert metric distances, and KL divergence. We describe how to adapt those clustering algorithms to the Hilbert distance.

4.1 k-means++ Clustering

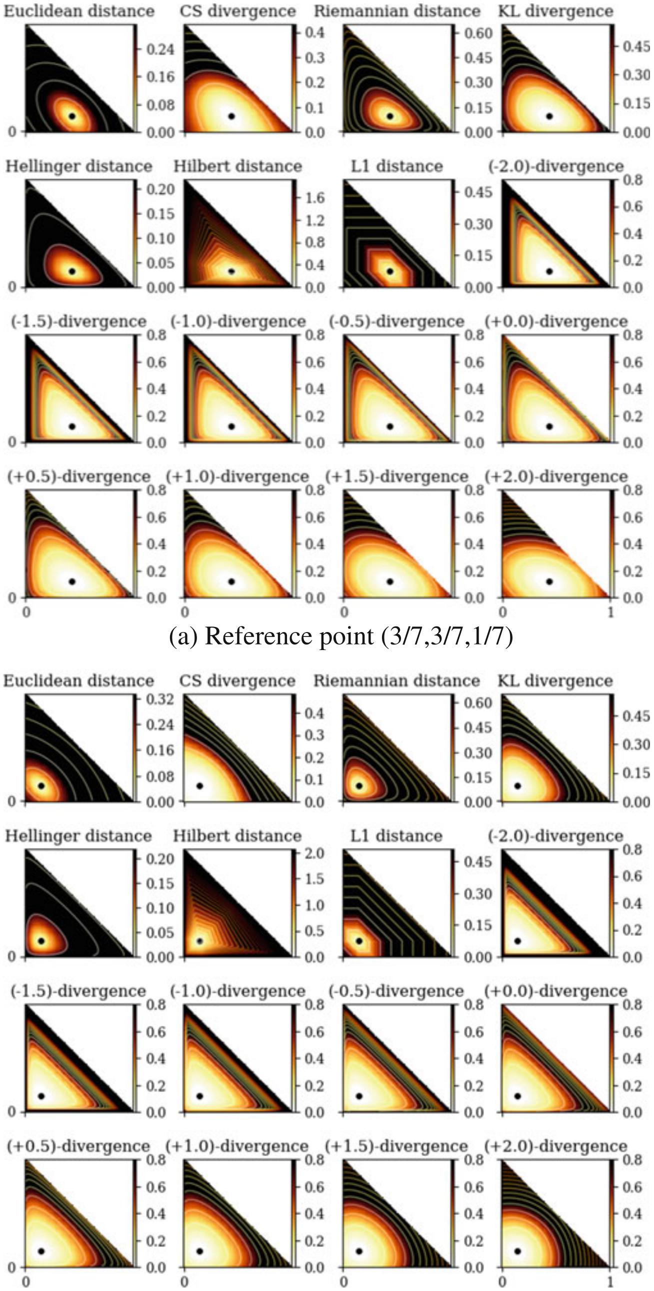
The celebrated k-means clustering [45] minimizes the sum of within-cluster variances, where each cluster has a center representative element. When dealing with $k = 1$ cluster, the center (also called centroid or cluster prototype) is the center of mass defined as the minimizer of

$$E_D(\Lambda, c) = \frac{1}{n} \sum_{i=1}^{n} D(p_i : c),$$

where $D(\cdot : \cdot)$ is a dissimilarity measure. For an arbitrary D, the centroid c may not be available in closed form. Nevertheless, using a generalization of the k-means++ initialization [9] (picking randomly seeds), one can bypass the centroid computation, and yet guarantee probabilistically a good clustering.

Let $C = \{c_1, \ldots, c_k\}$ denote the set of k cluster centers. Then the generalized k-means energy to be minimized is defined by:

$$E_D(\Lambda, C) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} D(p_i : c_j).$$



(a) Reference point (3/7,3/7,1/7)

(b) Reference point (5/7,1/7,1/7)

Fig. 7 A comparison of different distance measures on Δ^2. The distance is measured from $\forall p \in \Delta^2$ to a fixed reference point (the black dot). Lighter color means shorter distance. Darker color means longer distance. The contours show equal distance curves with a precision step of 0.2

By defining the distance $D(p, C) = \min_{j\in\{1,\ldots,k\}} D(p : c_j)$ of a point to a set, we can rewrite the objective function as $E_D(\Lambda, C) = \frac{1}{n}\sum_{i=1}^n D(p_i, C)$. Let $E_D^*(\Lambda, k) = \min_{C\,:\,|C|=k} E_D(\Lambda, C)$ denote the global minimum of $E_D(\Lambda, C)$ wrt some given Λ and k.

The k-means++ seeding proceeds for an arbitrary divergence D as follows: Pick uniformly at random at first seed c_1, and then iteratively choose the $(k-1)$ remaining seeds according to the following probability distribution:

$$\Pr(c_j = p_i) = \frac{D(p_i, \{c_1, \ldots, c_{j-1}\})}{\sum_{i=1}^n D(p_i, \{c_1, \ldots, c_{j-1}\})} \quad (2 \le j \le k).$$

Since its inception (2007), this k-means++ seeding has been extensively studied [46]. We state the general theorem established by [47]:

Theorem 2 (Generalized k-means++ performance, [47]) *Let κ_1 and κ_2 be two constants such that κ_1 defines the quasi-triangular inequality property:*

$$D(x : z) \le \kappa_1 \left(D(x : y) + D(y : z)\right), \quad \forall x, y, z \in \Delta^d,$$

and κ_2 handles the symmetry inequality:

$$D(x : y) \le \kappa_2 D(y : x), \quad \forall x, y \in \Delta^d.$$

Then the generalized k-means++ seeding guarantees with high probability a configuration C of cluster centers such that:

$$E_D(\Lambda, C) \le 2\kappa_1^2(1+\kappa_2)(2+\log k)E_D^*(\Lambda, k). \tag{7}$$

The ratio $\frac{E_D(\Lambda,C)}{E_D^*(\Lambda,k)}$ is called the *competitive factor*. The seminal result of ordinary k-means++ was shown [9] to be $8(2+\log k)$-competitive. When evaluating κ_1, one has to note that squared metric distances are not metric because they do not satisfy the triangular inequality. For example, the squared Euclidean distance is not a metric but it satisfies the 2-quasi-triangular inequality with $\kappa_1 = 2$.

We state the following general performance theorem:

Theorem 3 (k-means++ performance in a metric space) *In any metric space $(\mathscr{X}, d)$, the k-means++ wrt the squared metric distance d^2 is* $16(2+\log k)$*-competitive.*

Proof Since a metric distance is symmetric, it follows that $\kappa_2 = 1$. Consider the quasi-triangular inequality property for the squared non-metric dissimilarity d^2:

$$\begin{aligned} d(p,q) &\le d(p,q) + d(q,r),\\ d^2(p,q) &\le (d(p,q)+d(q,r))^2,\\ d^2(p,q) &\le d^2(p,q) + d^2(q,r) + 2d(p,q)d(q,r). \end{aligned}$$

Let us apply the inequality of arithmetic and geometric means[2]:

$$\sqrt{d^2(p,q)d^2(q,r)} \leq \frac{d^2(p,q)+d^2(q,r)}{2}.$$

Thus we have

$$d^2(p,q) \leq d^2(p,q)+d^2(q,r)+2d(p,q)d(q,r) \leq 2(d^2(p,q)+d^2(q,r)).$$

That is, the squared metric distance satisfies the 2-approximate triangle inequality, and $\kappa_1 = 2$. The result is straightforward from Theorem 2.

Theorem 4 (k-means++ performance in a normed space) *In any normed space* $(\mathscr{X}, \|\cdot\|)$*, the k-means++ with* $D(x:y) = \|x-y\|^2$ *is* $16(2+\log k)$*-competitive.*

Proof In any normed space $(\mathscr{X}, \|\cdot\|)$, we have both $\|x-y\| = \|y-x\|$ and the triangle inequality:

$$\|x-z\| \leq \|x-y\| + \|y-z\|.$$

The proof is very similar to the proof of Theorem 3 and is omitted.

Since any inner product space $(\mathscr{X}, \langle\cdot,\cdot\rangle)$ has an induced norm $\|x\| = \sqrt{\langle x,x\rangle}$, we have the following corollary.

Corollary 1 *In any inner product space* $(\mathscr{X}, \langle\cdot,\cdot\rangle)$*, the k-means++ with* $D(x:y) = \langle x-y, x-y\rangle$ *is* $16(2+\log k)$*-competitive.*

We need to report a bound for the squared Hilbert symmetric distance ($\kappa_2 = 1$). In [26] (Theorem 3.3), it was shown that Hilbert geometry of a bounded convex domain $\mathscr{C}$ is isometric to a normed vector space iff $\mathscr{C}$ is an open simplex: $(\Delta^d, \rho_{\mathrm{HG}}) \simeq (V^d, \|\cdot\|_{\mathrm{NH}})$, where $\|\cdot\|_{\mathrm{NH}}$ is the corresponding norm. Therefore $\kappa_1 = 2$. We write "NH" for short for this equivalent normed Hilbert geometry. Appendix 8 recalls the construction due to [25], and shows the squared Hilbert distance fails the triangle inequality and it is not a distance induced by an inner product.

As an empirical study, we randomly generate $n = 10^6$ tuples (x, y, z) based on the uniform distribution in Δ^d. For each tuple (x, y, z), we evaluate the ratio

$$\kappa_1 = \frac{D(x:z)}{D(x:y)+D(y:z)}.$$

Figure 8 shows the statistics for four different choices of D: (1) $D(x:y) = \rho^2_{\mathrm{FHR}}(x,y)$; (2) $D(x:y) = \frac{1}{2}\mathrm{KL}(x:y) + \frac{1}{2}\mathrm{KL}(y:x)$; (3) $D(x:y) = \rho^2_{\mathrm{HG}}(x,y)$; (4) $D(x:y) = \rho^2_{\mathrm{L1}}(x,y)$. We find experimentally that κ_1 is upper bounded by 2 for ρ^2_{FHR}, ρ^2_{HG} and ρ^2_{L1}, while the average κ_1 value is smaller than 0.5. For all the compared distances, $\kappa_2 = 1$. Therefore ρ_{FHR} and ρ_{HG} have better k-means++ performance guarantee as compared to ρ_{IG}.

[2] For positive values a and b, the arithmetic-geometric mean inequality states that $\sqrt{ab} \leq \frac{a+b}{2}$.

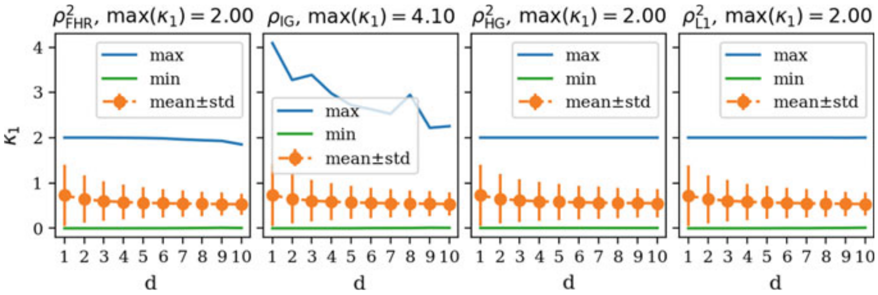
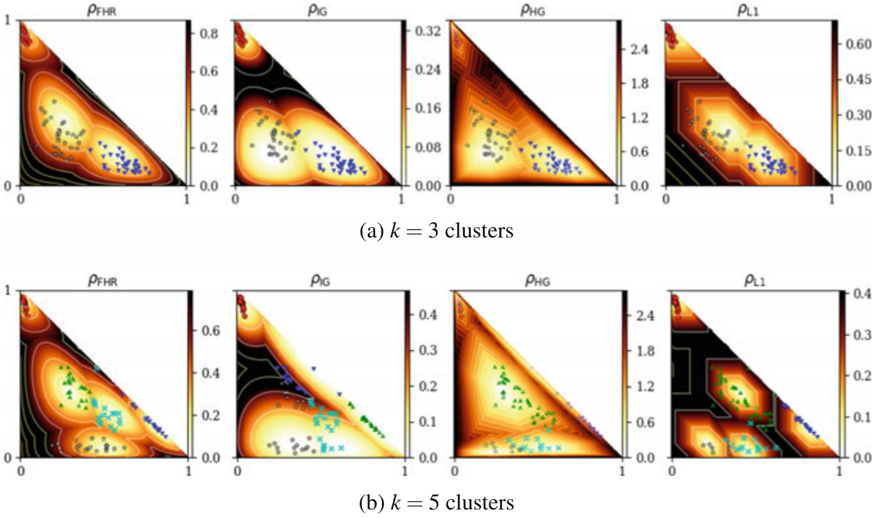


Fig. 8 The maximum, mean, standard deviation, and minimum of κ_1 on 10^6 randomly generated tuples (x, y, z) in Δ^d for $d = 1, \ldots, 10$



(a) $k = 3$ clusters

(b) $k = 5$ clusters

Fig. 9 k-Means++ clustering results on a toy dataset in the space of trinomials Δ^2. The color density maps indicate the distance from any point to its nearest cluster center

We get by applying Theorem 4:

Corollary 2 (k-means++ in Hilbert simplex geometry) *The k-means++ seeding in a Hilbert simplex geometry in fixed dimension is* $16(2 + \log k)$*-competitive.*

Figure 9 displays the clustering results of k-means++ in Hilbert simplex geometry as compared to the other geometries for $k \in \{3, 5\}$.

The KL divergence can be interpreted as a separable Bregman divergence [48]. The Bregman k-means++ performance has been studied in [48, 49], and a competitive factor of $O(\frac{1}{\mu})$ is reported using the notion of Bregman μ-similarity (that is suited for data-sets on a compact domain).

In [50], spherical k-means++ is studied wrt the distance $d_S(x, y) = 1 - \langle x, y\rangle$ for any pair of points x, y on the unit sphere. Since $\langle x, y\rangle = \|x\|_2\|y\|_2 \cos(\theta_{x,y}) =$

$\cos(\theta_{x,y})$, we have $d_S(x, y) = 1 - \cos(\theta_{x,y})$, where $\theta_{x,y}$ denotes the angle between a pair of unit vectors x and y. This distance is called the cosine distance since it amounts to one minus the cosine similarity. Notice that the cosine distance is related to the squared Euclidean distance via the identity: $d_S(x, y) = \frac{1}{2}\|x - y\|^2$. The cosine distance is different from the spherical distance that relies on the arccos function.

Since divergences may be asymmetric, one can further consider mixed divergence $M(p : q : r) = \lambda D(p : q) + (1 - \lambda)D(q : r)$ for $\lambda \in [0, 1]$, and extend the k-means++ seeding procedure and analysis [51].

For a given data set, we can compute κ_1 or κ_2 by inspecting triples and pairs of points, and get data-dependent competitive factor improving the bounds mentioned above.

4.2 k-Center Clustering

Let Λ be a finite point set. The cost function for a k-center clustering with centers C ($|C| = k$) is:

$$f_D(\Lambda, C) = \max_{p_i \in \Lambda} \min_{c_j \in C} D(p_i : c_j).$$

The farthest first traversal heuristic [10] has a guaranteed approximation factor of 2 for any metric distance (see Algorithm 3).

In order to use the k-center clustering algorithm described in Algorithm 2, we need to be able to compute the 1-center (or minimax center) for the Hilbert simplex geometry, that is the Minimum Enclosing Ball (MEB, also called the Smallest Enclosing Ball, SEB).

Algorithm 2: k-Center clustering

```
Data: A set of points $p_1, \ldots, p_n \in \Delta^d$. A distance measure $\rho$ on $\Delta^d$. The maximum number
      $k$ of clusters. The maximum number $T$ of iterations.
Result: A clustering scheme assigning each $p_i$ a label $l_i \in \{1, \ldots, k\}$
begin
    Randomly pick $k$ cluster centers $c_1, \ldots, c_k$ using the kmeans++ heuristic;
    for $t = 1, \ldots, T$ do
        for $i = 1, \ldots, n$ do
            $l_i \leftarrow \arg\min_{l=1}^{k} \rho(p_i, c_l)$;
        for $l = 1, \ldots, k$ do
            $c_l \leftarrow \arg\min_c \max_{i:l_i=l} \rho(p_i, c)$;
    Output $\{l_i\}_{i=1}^{n}$;
```

Algorithm 3: A 2-approximation of the k-center clustering for any metric distance ρ.

```
Data: A set $\Lambda$; a number $k$ of clusters; a metric distance $\rho$.
Result: A 2-approximation of the $k$-center clustering
begin
    $c_1 \leftarrow$ ARandomPointOf($\Lambda$);
    $C \leftarrow \{c_1\}$;
    for $i = 2, \ldots, k$ do
        $c_i \leftarrow \arg\max_{p \in \Lambda} \rho(p, C)$;
        $C \leftarrow C \cup \{c_i\}$;
Output $C$;
```

We may consider the SEB equivalently either in Δ^d or in the normed space V^d. In both spaces, the shapes of the balls are convex. Let $\Lambda = \{p_1, \ldots, p_n\}$ denote the point set in Δ^d, and $\mathcal{V} = \{v_1, \ldots, v_n\}$ the equivalent point set in the normed vector space (following the mapping explained in Appendix 8). Then the SEBs $B_{\mathrm{HG}}(\Lambda)$ in Δ^d and $B_{\mathrm{NH}}(\mathcal{V})$ in V^d have respectively radii r^*_{HG} and r^*_{NH} defined by:

$$r^*_{\mathrm{HG}} = \min_{c \in \Delta^d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(p_i, c),$$

$$r^*_{\mathrm{NH}} = \min_{v \in V^d} \max_{i \in \{1,\ldots,n\}} \|v_i - v\|_{\mathrm{NH}}.$$

The SEB in the normed vector space $(V^d, \|\cdot\|_{\mathrm{NH}})$ amounts to find the minimum covering norm polytope of a finite point set. This problem has been well-studied in computational geometry [52–54]. By considering the equivalent Hilbert norm polytope with $d(d+1)$ facets, we state the result of [54]:

Theorem 5 (SEB in Hilbert polytope normed space, [54]) *A $(1+\varepsilon)$-approximation of the SEB in V^d can be computed in $O(d^3 \frac{n}{\varepsilon})$ time.*

We shall now report two algorithms for computing the SEBs: One exact algorithm in V^d that does not scale well in high dimensions, and one approximation in Δ^d that works well for large dimensions.

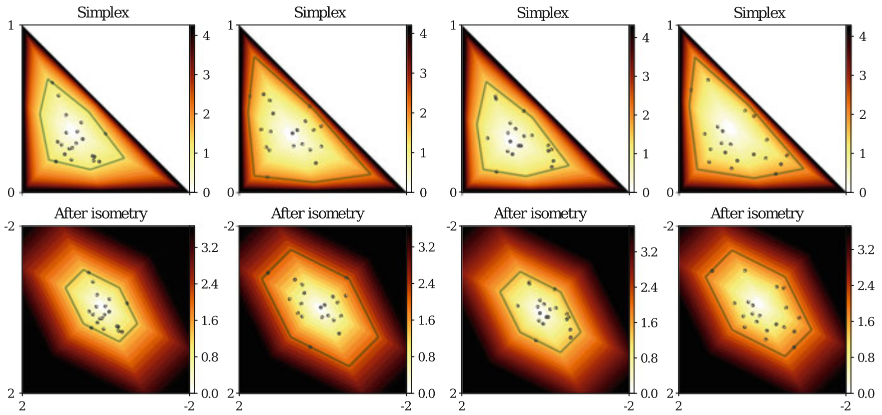
4.2.1 Exact Smallest Enclosing Ball in a Hilbert Simplex Geometry

Given a finite point set $\{p_1, \ldots, p_n\} \in \Delta^d$, the SEB in Hilbert simplex geometry is centered at

$$c^* = \arg\min_{c \in \Delta^d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(c, x_i),$$

with radius

$$r^* = \min_{c \in \Delta^d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(c, x_i).$$



(a) 3 points on the border (b) 2 points on the border

Fig. 10 Computing the SEB in Hilbert simplex geometry amounts to compute the SEB in the corresponding normed vector space

An equivalent problem is to find the SEB in the isometric normed vector space V^d via the mapping reported in Appendix 8. Each simplex point p_i corresponds to a point v_i in the V^d.

Figure 10 displays some examples of the exact smallest enclosing balls in the Hilbert simplex geometry and in the corresponding normed vector space.

To compute the SEB, one may also consider the generic LP-type randomized algorithm [55]. We notice that an enclosing ball for a point set in general has a number k of points on the border of the ball, with $2 \leq k \leq \frac{d(d+1)}{2}$. Let $D = \frac{d(d+1)}{2}$ denote the varying size of the combinatorial basis, then we can apply the LP-type framework (we check the axioms of locality and monotonicity, [56]) to solve efficiently the SEBs.

Theorem 6 (Smallest Enclosing Hilbert Ball is LP-type, [56, 57]) *The smallest enclosing Hilbert ball amounts to find the smallest enclosing ball in a vector space with respect to a polytope norm that can be solved using a LP-type randomized algorithm.*

The Enclosing Ball Decision Problem (EBDP, [58]) asks for a given value r, whether $r \geq r^*$ or not. The decision problem amounts to find whether a set $\{rB_V + v_i\}$ of translates can be stabbed by a point [58]: That is, whether $\cap_{i=1}^n (rB_V + v_i)$ is empty or not. Since these translates are polytopes with $d(d+1)$ facets, this can be solved in linear time using *Linear Programming*.

Theorem 7 (Enclosing Hilbert Ball Decision Problem) *The decision problem to test whether $r \geq r^*$ or not can be solved by Linear Programming.*

This yields a simple scheme to approximate the optimal value r^*: Let $r_0 = \max_{i\in\{2,\dots,n\}} \|v_i - v_1\|_{\mathrm{NH}}$. Then $r^* \in [\frac{r_0}{2}, r_0] = [a_0, b_0]$. At stage i, perform a dichotomic search on $[a_i, b_i]$ by answering the decision problem for $r_{i+1} = \frac{a_i+b_i}{2}$, and update the radius range accordingly [58].

However, the LP-type randomized algorithm or the decision problem-based algorithm do not scale well in high dimensions. Next, we introduce a simple approximation algorithm that relies on the fact that the line segment $[pq]$ is a geodesic in Hilbert simplex geometry. (Geodesics are not unique. See Figure 2 of [25].)

4.2.2 Geodesic Bisection Approximation Heuristic

In Riemannian geometry, the 1-center can be arbitrarily finely approximated by a simple geodesic bisection algorithm [59, 60]. This algorithm can be extended to HG straightforwardly as detailed in Algorithm 4.

Algorithm 4: Geodesic walk for approximating the Hilbert minimax center, generalizing [60]

```
Data: A set of points $p_1, \dots, p_n \in \Delta^d$. The maximum number $T$ of iterations.
Result: $c \approx \arg\min_c \max_i \rho_{\mathrm{HG}}(p_i, c)$
begin
    $c_0 \leftarrow$ ARandomPointOf$(\{p_1, \dots, p_n\})$;
    for $t = 1, \dots, T$ do
        $p \leftarrow \arg\max_{p_i} \rho_{\mathrm{HG}}(p_i, c_{t-1})$;
        $c_t \leftarrow c_{t-1} \#^{\rho}_{1/(t+1)} p$;
    Output $c_T$;
```

The algorithm first picks up a point c_0 at random from Λ as the initial center, then computes the farthest point p (with respect to the distance ρ), and then walk on the geodesic from c_0 to p by a certain amount to define c_1, etc. For an arbitrary distance ρ, we define the operator $\#^{\rho}_{\alpha}$ as follows:

$$p \#^{\rho}_{\alpha} q = v = \gamma(p, q, \alpha), \quad \rho(p : v) = \alpha\rho(p : q),$$

where $\gamma(p, q, \alpha)$ is the geodesic passing through p and q, and parameterized by α ($0 \leq \alpha \leq 1$). When the equations of the geodesics are explicitly known, we can either get a closed form solution for $\#^{\rho}_{\alpha}$ or perform a bisection search to find v' such that $\rho(p : v') \approx \alpha\rho(p : q)$. See [61] for an extension and analysis in hyperbolic geometry. See Fig. 11 to get an intuitive idea on the *experimental* convergence rate of Algorithm 4. See Fig. 12 for visualizations of centers wrt different geometries.

Furthermore, this iterative algorithm implies a core-set [62] (namely, the set of farthest points visited during the geodesic walks) that is useful for clustering large

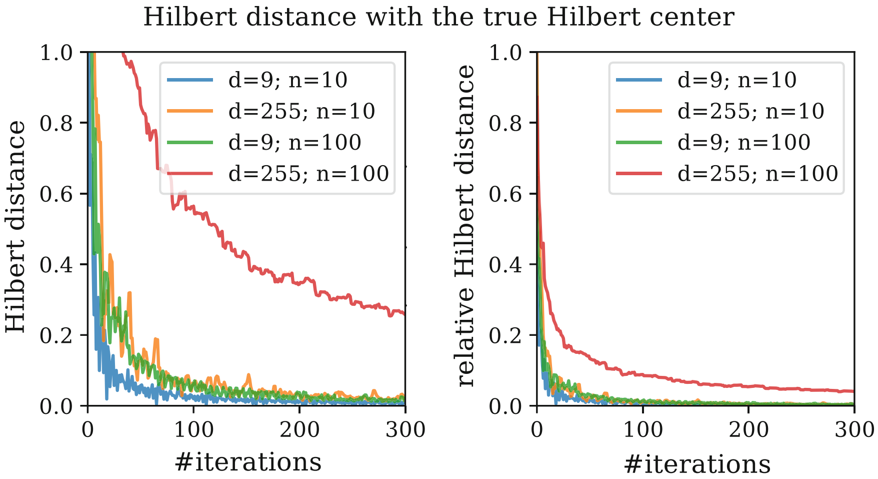


Fig. 11 Convergence rate of Algorithm 4 measured by the Hilbert distance between the current minimax center and the true center (left) or their Hilbert distance divided by the Hilbert radius of the dataset (right). The plot is based on 100 random points in Δ^9/Δ^{255}

data-sets [63]. See [52] for core-set results on containment problems wrt a convex homothetic object (the equivalent Hilbert polytope norm in our case).

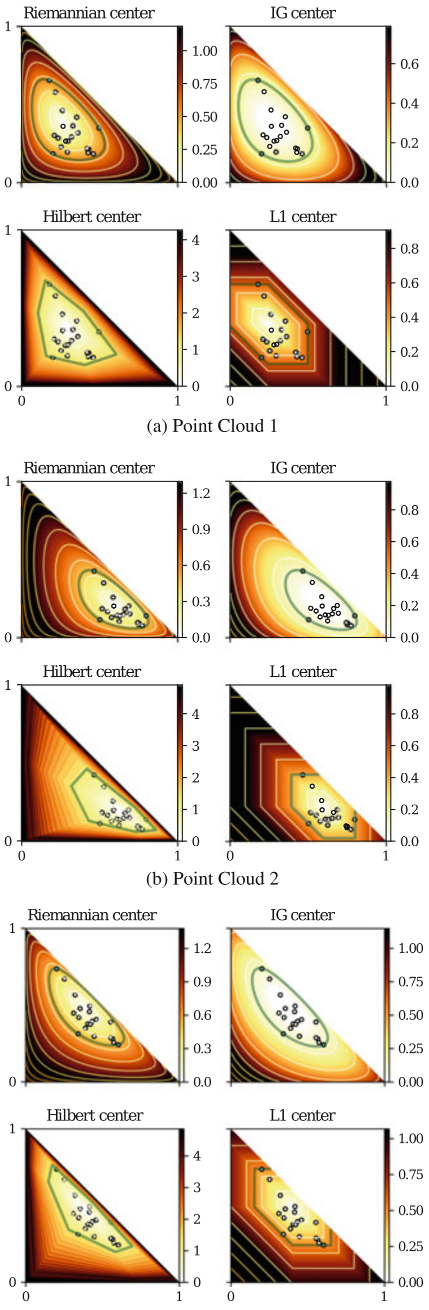
A simple algorithm dubbed MinCon [53] can find an approximation of the Minimum Enclosing Polytope. The algorithm induces a core-set of size $O(\frac{1}{\varepsilon^2})$ although the theorem is challenged in [52].

Thus by combining the k-center seeding [10] with the Lloyd-like batched iterations, we get an efficient k-center clustering algorithm for the FHR and Hilbert metric geometries. When dealing with the Kullback–Leibler divergence, we use the fact that KL is a Bregman divergence, and use the 1-center algorithm ([64, 65] for approximation in any dimension, or [55] which is exact but limited to small dimensions).

Since Hilbert simplex geometry is isomorphic to a normed vector space [26] with a polytope norm with $d(d+1)$ facets, the Voronoi diagram in Hilbert geometry of Δ^d amounts to compute a Voronoi diagram wrt a polytope norm [66–68].

5 Experiments

We generate a dataset consisting of a set of clusters in a high dimensional statistical simplex Δ^d. Each cluster is generated independently as follows. We first pick a random center $c = (\lambda_c^0, \ldots, \lambda_c^d)$ based on the uniform distribution on Δ^d. Then any random sample $p = (\lambda^0, \ldots, \lambda^d)$ associated with c is independently generated by



(a) Point Cloud 1

(b) Point Cloud 2

(c) Point Cloud 3

Fig. 12 The Riemannian/IG/Hilbert/L_1 minimax centers of three point clouds in Δ^2 based on Algorithm 4. The color maps show the distance from $\forall p \in \Delta^2$ to the corresponding center

$$\lambda^i = \frac{\exp(\log \lambda_c^i + \sigma\varepsilon^i)}{\sum_{i=0}^d \exp(\log \lambda_c^i + \sigma\varepsilon^i)},$$

where $\sigma > 0$ is a noise level parameter, and each ε^i follows independently a standard Gaussian distribution (generator 1) or the Student's t-distribution with five degrees of freedom (generator 2). Let $\sigma = 0$, we get $\lambda^i = \lambda_c^i$. Therefore p is randomly distributed around c. We repeat generating random samples for each cluster center, and make sure that different clusters have almost the same number of samples. Then we perform clustering based on the configurations $n \in \{50, 100\}$, $d \in \{9, 255\}$, $\sigma \in \{0.5, 0.9\}$, $\rho \in \{\rho_{\mathrm{FHR}}, \rho_{\mathrm{IG}}, \rho_{\mathrm{HG}}, \rho_{\mathrm{EUC}}, \rho_{\mathrm{L1}}\}$. For simplicity, the number of clusters k is set to the ground truth. For each configuration, we repeat the clustering experiment based on 300 different random datasets. The performance is measured by the normalized mutual information (NMI), which is a scalar indicator in the range $[0, 1]$ (the larger the better).

The results of k-means++ and k-centers are shown in Tables 2 and 3, respectively. The large variance of NMI is because that each experiment is performed on random datasets wrt different random seeds. Generally, the performance deteriorates as we increase the number of clusters, increase the noise level or decrease the dimensionality, which have the same effect to reduce the inter-cluster gap.

The key comparison is the three columns ρ_{FHR}, ρ_{HG} and ρ_{IG}, as they are based on exactly the same algorithm with the only difference being the underlying geometry. We see clearly that in general, their clustering performance presents the order HG > FHR > IG. The performance of HG is superior to the other two geometries, especially when the noise level is large. Intuitively, the Hilbert balls are more compact in size and therefore can better capture the clustering structure (see Fig. 2).

The column ρ_{EUC} is based on the Euclidean enclosing ball. It shows the worst scores because the intrinsic geometry of the probability simplex is far from the Euclidean geometry.

6 Hilbert's Projective Geometry of the Space of Correlation Matrices

In this section, we present the Hilbert's projective geometry to the space of correlation matrices

$$\mathscr{C}^d = \{C_{d\times d} \,:\, C \succ 0; C_{ii} = 1, \forall i\}.$$

If $C_1, C_2 \in \mathscr{C}$, then $(1-\lambda)C_1 + \lambda C_2 \in \mathscr{C}$ for $0 < \lambda < 1$. Therefore $\mathscr{C}$ is a convex set, known as an *elliptope* [11] embedded in the p.s.d. cone. See Fig. 13 for an intuitive view of $\mathscr{C}_3$, where the coordinate system (x, y, z) is the off-diagonal entries of $C \in \mathscr{C}_3$.

In order to compute the Hilbert distance $\rho_{\mathrm{HG}}(C_1, C_2)$, we need to compute the intersection of the line (C_1, C_2) with $\partial\mathscr{C}$, denoted as C_1' and C_2', then we have

Table 2 k-means++ clustering accuracy in NMI on randomly generated datasets based on different geometries. The table shows the mean and standard deviation after 300 independent runs for each configuration. ρ is the distance measure. n is the sample size. d is the dimensionality of Δ^d. σ is noise level

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.76 ± 0.22	0.76 ± 0.24	**0.81 ± 0.22**	0.64 ± 0.23	0.70 ± 0.22
			0.9	0.44 ± 0.20	0.44 ± 0.20	**0.57 ± 0.22**	0.31 ± 0.17	0.38 ± 0.18
		255	0.5	0.80 ± 0.24	0.81 ± 0.24	**0.88 ± 0.21**	0.74 ± 0.25	0.79 ± 0.24
			0.9	0.65 ± 0.27	0.66 ± 0.28	**0.72 ± 0.27**	0.46 ± 0.24	0.63 ± 0.27
	100	9	0.5	0.76 ± 0.22	0.76 ± 0.21	**0.82 ± 0.22**	0.60 ± 0.21	0.69 ± 0.23
			0.9	0.42 ± 0.19	0.41 ± 0.18	**0.54 ± 0.22**	0.27 ± 0.14	0.34 ± 0.16
		255	0.5	0.82 ± 0.23	0.82 ± 0.24	**0.89 ± 0.20**	0.74 ± 0.24	0.80 ± 0.25
			0.9	0.66 ± 0.26	0.66 ± 0.28	**0.72 ± 0.26**	0.45 ± 0.25	0.64 ± 0.27
5	50	9	0.5	0.75 ± 0.14	0.74 ± 0.15	**0.81 ± 0.13**	0.61 ± 0.13	0.68 ± 0.13
			0.9	0.44 ± 0.13	0.42 ± 0.13	**0.55 ± 0.15**	0.31 ± 0.11	0.36 ± 0.12
		255	0.5	0.83 ± 0.15	0.83 ± 0.15	**0.88 ± 0.14**	0.77 ± 0.16	0.82 ± 0.15
			0.9	0.71 ± 0.17	0.70 ± 0.19	**0.75 ± 0.17**	0.50 ± 0.17	0.68 ± 0.18
	100	9	0.5	0.74 ± 0.13	0.74 ± 0.14	**0.80 ± 0.14**	0.60 ± 0.13	0.67 ± 0.13
			0.9	0.42 ± 0.11	0.40 ± 0.12	**0.55 ± 0.15**	0.29 ± 0.09	0.35 ± 0.11
		255	0.5	0.83 ± 0.14	0.83 ± 0.15	**0.88 ± 0.13**	0.77 ± 0.15	0.81 ± 0.15
			0.9	0.69 ± 0.18	0.69 ± 0.18	**0.73 ± 0.17**	0.48 ± 0.17	0.67 ± 0.18

(a) Generator 1

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.62 ± 0.22	0.60 ± 0.22	**0.71 ± 0.23**	0.45 ± 0.20	0.54 ± 0.22
			0.9	0.29 ± 0.17	0.27 ± 0.16	**0.39 ± 0.19**	0.17 ± 0.13	0.25 ± 0.15
		255	0.5	0.70 ± 0.25	0.69 ± 0.26	**0.74 ± 0.25**	0.37 ± 0.29	0.70 ± 0.26
			0.9	**0.42 ± 0.25**	0.35 ± 0.20	0.40 ± 0.19	0.03 ± 0.08	**0.44 ± 0.26**
	100	9	0.5	0.63 ± 0.22	0.61 ± 0.22	**0.71 ± 0.22**	0.46 ± 0.19	0.56 ± 0.20
			0.9	0.29 ± 0.15	0.26 ± 0.14	**0.38 ± 0.20**	0.18 ± 0.12	0.24 ± 0.14
		255	0.5	0.71 ± 0.26	0.69 ± 0.27	**0.75 ± 0.25**	0.31 ± 0.28	0.70 ± 0.27
			0.9	0.41 ± 0.26	0.33 ± 0.20	0.38 ± 0.18	0.02 ± 0.06	**0.43 ± 0.26**
5	50	9	0.5	0.64 ± 0.15	0.61 ± 0.14	**0.70 ± 0.14**	0.48 ± 0.14	0.57 ± 0.15
			0.9	0.31 ± 0.12	0.29 ± 0.12	**0.41 ± 0.15**	0.20 ± 0.09	0.26 ± 0.10
		255	0.5	0.74 ± 0.17	0.72 ± 0.17	**0.77 ± 0.16**	0.41 ± 0.20	0.74 ± 0.17
			0.9	0.44 ± 0.17	0.37 ± 0.16	0.44 ± 0.15	0.04 ± 0.06	**0.47 ± 0.17**
	100	9	0.5	0.62 ± 0.14	0.61 ± 0.14	**0.71 ± 0.14**	0.46 ± 0.13	0.54 ± 0.14
			0.9	0.30 ± 0.10	0.27 ± 0.11	**0.40 ± 0.13**	0.19 ± 0.08	0.25 ± 0.09
		255	0.5	0.73 ± 0.18	0.70 ± 0.18	**0.75 ± 0.16**	0.37 ± 0.20	0.73 ± 0.17
			0.9	0.43 ± 0.16	0.35 ± 0.14	0.41 ± 0.12	0.03 ± 0.06	**0.46 ± 0.18**

(b) Generator 2

Table 3 k-center clustering accuracy in NMI on randomly generated datasets based on different geometries. The table shows the mean and standard deviation after 300 independent runs for each configuration. ρ is the distance measure. n is the sample size. d is the dimensionality of the statistical simplex. σ is noise level

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.87 ± 0.19	0.85 ± 0.19	**0.92 ± 0.16**	0.72 ± 0.22	0.80 ± 0.20
			0.9	0.54 ± 0.21	0.51 ± 0.21	**0.70 ± 0.23**	0.36 ± 0.17	0.44 ± 0.19
		255	0.5	0.93 ± 0.16	0.92 ± 0.18	**0.95 ± 0.14**	0.89 ± 0.18	0.90 ± 0.19
			0.9	0.76 ± 0.24	0.72 ± 0.26	**0.82 ± 0.24**	0.50 ± 0.28	0.76 ± 0.25
	100	9	0.5	0.88 ± 0.17	0.86 ± 0.18	**0.93 ± 0.14**	0.70 ± 0.20	0.80 ± 0.20
			0.9	0.53 ± 0.20	0.49 ± 0.19	**0.70 ± 0.22**	0.33 ± 0.14	0.41 ± 0.18
		255	0.5	0.93 ± 0.16	0.92 ± 0.17	**0.95 ± 0.13**	0.88 ± 0.19	0.93 ± 0.16
			0.9	0.81 ± 0.22	0.75 ± 0.24	**0.83 ± 0.22**	0.47 ± 0.28	0.79 ± 0.22
5	50	9	0.5	0.82 ± 0.13	0.81 ± 0.13	**0.89 ± 0.12**	0.67 ± 0.13	0.75 ± 0.13
			0.9	0.50 ± 0.13	0.47 ± 0.13	**0.66 ± 0.15**	0.34 ± 0.11	0.40 ± 0.12
		255	0.5	**0.92 ± 0.11**	**0.91 ± 0.12**	**0.93 ± 0.11**	0.87 ± 0.13	**0.92 ± 0.12**
			0.9	0.77 ± 0.15	0.71 ± 0.17	**0.85 ± 0.17**	0.54 ± 0.19	0.74 ± 0.16
	100	9	0.5	0.83 ± 0.12	0.81 ± 0.13	**0.89 ± 0.11**	0.67 ± 0.11	0.76 ± 0.13
			0.9	0.48 ± 0.12	0.46 ± 0.12	**0.66 ± 0.15**	0.33 ± 0.09	0.39 ± 0.10
		255	0.5	**0.93 ± 0.10**	**0.92 ± 0.11**	**0.94 ± 0.09**	0.89 ± 0.11	0.92 ± 0.11
			0.9	0.81 ± 0.14	0.74 ± 0.15	**0.84 ± 0.16**	0.52 ± 0.19	0.79 ± 0.14

(a) Generator 1

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.68 ± 0.22	0.67 ± 0.22	**0.80 ± 0.20**	0.48 ± 0.22	0.60 ± 0.22
			0.9	0.32 ± 0.18	0.29 ± 0.17	**0.45 ± 0.21**	0.20 ± 0.14	0.26 ± 0.15
		255	0.5	0.79 ± 0.24	0.75 ± 0.24	**0.82 ± 0.22**	0.13 ± 0.23	**0.81 ± 0.24**
			0.9	0.35 ± 0.27	0.35 ± 0.21	**0.42 ± 0.19**	0.00 ± 0.02	0.32 ± 0.30
	100	9	0.5	0.66 ± 0.22	0.65 ± 0.22	**0.79 ± 0.21**	0.45 ± 0.19	0.59 ± 0.20
			0.9	0.30 ± 0.16	0.28 ± 0.14	**0.42 ± 0.19**	0.20 ± 0.12	0.26 ± 0.14
		255	0.5	0.78 ± 0.25	0.76 ± 0.24	**0.82 ± 0.21**	0.05 ± 0.14	0.77 ± 0.27
			0.9	0.29 ± 0.28	0.29 ± 0.20	**0.39 ± 0.20**	0.00 ± 0.02	0.22 ± 0.25
5	50	9	0.5	0.69 ± 0.14	0.66 ± 0.14	**0.77 ± 0.13**	0.50 ± 0.13	0.61 ± 0.14
			0.9	0.34 ± 0.12	0.30 ± 0.12	**0.46 ± 0.15**	0.22 ± 0.09	0.28 ± 0.10
		255	0.5	**0.80 ± 0.15**	0.76 ± 0.15	**0.82 ± 0.14**	0.24 ± 0.23	**0.81 ± 0.14**
			0.9	0.42 ± 0.21	0.38 ± 0.16	**0.46 ± 0.15**	0.00 ± 0.02	0.39 ± 0.22
	100	9	0.5	0.66 ± 0.13	0.64 ± 0.14	**0.77 ± 0.14**	0.47 ± 0.13	0.57 ± 0.13
			0.9	0.31 ± 0.11	0.28 ± 0.10	**0.44 ± 0.13**	0.21 ± 0.08	0.25 ± 0.09
		255	0.5	**0.80 ± 0.16**	0.76 ± 0.15	**0.82 ± 0.13**	0.12 ± 0.17	**0.81 ± 0.16**
			0.9	0.32 ± 0.19	0.30 ± 0.15	**0.41 ± 0.13**	0.00 ± 0.01	0.26 ± 0.18

(b) Generator 2

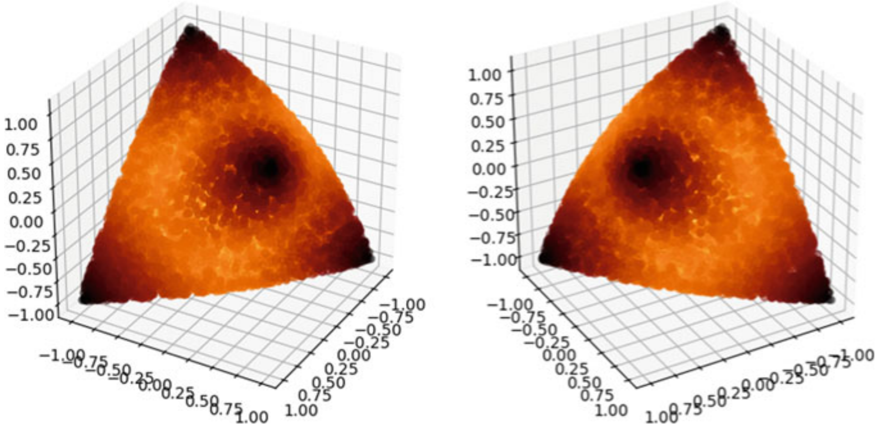


Fig. 13 The elliptope $\mathscr{C}_3$ (two different perspectives)

$$\rho_{\mathrm{HG}}(C_1, C_2) = \left| \log \frac{\|C_1 - C_2'\| \|C_1' - C_2\|}{\|C_1 - C_1'\| \|C_2 - C_2'\|} \right|.$$

Unfortunately there is no closed form solution of C_1' and C_2'. Instead, we apply a binary searching algorithm. Note a necessary condition for $C \in \mathscr{C}$ is that C has a positive spectrum. If C has at least one non-positive eigenvalue, then $C \notin \mathscr{C}$. To determine whether a given C is inside the elliptope requires a spectral decomposition of C. Therefore the computation of C_1' and C_2' is in general expensive.

We compare the Hilbert elliptope geometry with commonly used distance measures including the L_2 distance ρ_{EUC}, L_1 distance ρ_{L1}, and the square root of the log-det divergence

$$\rho_{\mathrm{LD}}(C_1, C_2) = \mathrm{tr}(C_1 C_2^{-1}) - \log |C_1 C_2^{-1}| - d.$$

Due to the high computational complexity, we only investigate k-means++ clustering. The investigated dataset consists of 100 matrices forming 3 clusters in $\mathscr{C}_3$ with almost identical size. Each cluster is independently generated according to

$$P \sim \mathscr{W}^{-1}(I_{3\times 3}, \nu_1),$$
$$C_i \sim \mathscr{W}^{-1}(P, \nu_2),$$

where $\mathscr{W}^{-1}(A, \nu)$ denotes the inverse Wishart distribution with scale matrix A and ν degrees of freedom, and C_i is a point in the cluster associated with P. Table 4 shows the k-means++ clustering performance in terms of NMI. Again Hilbert geometry is favorable as compared to alternatives, showing that the good performance of Hilbert clustering is generalizable.

Table 4 NMI (mean±std) of k-means++ clustering based on different distance measures in the elliptope (500 independent runs)

ν_1	ν_2	ρ_{HG}	ρ_{EUC}	ρ_{L1}	ρ_{LD}
4	10	**0.62 ± 0.22**	0.57±0.21	0.56±0.22	0.58±0.22
4	30	**0.85 ± 0.18**	0.80±0.20	0.81±0.19	0.82±0.20
4	50	**0.89 ± 0.17**	0.87±0.17	0.86±0.18	0.88±0.18
5	10	**0.50 ± 0.21**	0.49±0.21	0.48±0.20	0.47±0.21
5	30	**0.77 ± 0.20**	0.75±0.21	0.75±0.21	0.75±0.21
5	50	**0.84 ± 0.19**	0.82±0.19	0.82±0.20	**0.84 ± 0.18**

7 Conclusion

We introduced the Hilbert projective metric distance and its underlying non-Riemannian geometry for modeling the space of multinomials or the open probability simplex. We compared experimentally in simulated clustering tasks this geometry with the traditional differential geometric modelings (either the Fisher-Hotelling-Rao metric connection or the dually coupled non-metric affine connections of information geometry [17]).

The main feature of Hilbert geometry (HG) is that it is a metric non-manifold geometry, where geodesics are straight (Euclidean) line segments. This makes this geometry computationally attractive. In simplex domains, the Hilbert balls have fixed combinatorial (Euclidean) polytope structures, and HG is known to be isometric to a normed space [25, 69]. This latter isometry allows one to generalize easily the standard proofs of clustering (e.g., k-means or k-center). We demonstrated it for the k-means++ competitive performance analysis and for the convergence of the 1-center heuristic [60] (smallest enclosing Hilbert ball allows one to implement efficiently the k-center clustering). Our experimental k-means++ or k-center comparisons of HG algorithms with the manifold modeling approach yield superior performance. This may be intuitively explained by the sharpness of Hilbert balls as compared to the FHR/IG ball profiles.

Chentsov [70] defined statistical invariance on a probability manifold under Markov morphisms and proved that the Fisher Information Metric is the unique Riemannian metric (up to rescaling) for multinomials. However, this does not rule out that other distances (with underlying geometric structures) may be used to model statistical manifolds (e.g., Finsler statistical manifolds [71, 72], or the total variation distance — the only metric f-divergence [73]). Defining statistical invariance related to geometry is the cornerstone problem of information geometry that can be tackled from many directions (see [74] and references therein for a short review).

In this paper, we introduced Hilbert geometries in machine learning by considering clustering tasks in the probability simplex and in the correlation elliptope. A canonical Hilbert metric distance can be defined on any bounded convex subset of the Euclidean space with the key property that geodesics are straight Euclidean line segments thus making this geometry well-suited for fast and exact computations. Thus we may consider clustering in other bounded convex subsets like the simplotopes [75].

One future direction is to consider the Hilbert metric for regularization and sparsity in machine learning (due to its equivalence with a polytope normed distance).

Our Python codes are freely available online for reproducible research:
https://www.lix.polytechnique.fr/~nielsen/HSG/

8 Isometry of Hilbert Simplex Geometry to a Normed Vector Space

Consider the Hilbert simplex metric space $(\Delta^d, \rho_{\mathrm{HG}})$ where Δ^d denotes the d-dimensional open probability simplex and ρ_{HG} the Hilbert cross-ratio metric. Let us recall the isometry ([25], 1991) of the open standard simplex to a normed vector space $(V^d, \|\cdot\|_{\mathrm{NH}})$. Let $V^d = \{v \in \mathbb{R}^{d+1} \ : \ \sum_i v^i = 0\}$ denote the d-dimensional vector space sitting in $\mathbb{R}^{d+1}$. Map a point $p = (\lambda^0, \ldots, \lambda^d) \in \Delta^d$ to a point $v(x) = (v^0, \ldots, v^d) \in V^d$ as follows:

$$v^i = \frac{1}{d+1}\left(d \log \lambda^i - \sum_{j \neq i} \log \lambda^j\right) = \log \lambda^i - \frac{1}{d+1}\sum_j \log \lambda^j.$$

We define the corresponding norm $\|\cdot\|_{\mathrm{NH}}$ in V^d by considering the shape of its unit ball $B_V = \{v \in V^d \ : \ |v^i - v^j| \leq 1, \forall i \neq j\}$. The unit ball B_V is a symmetric convex set containing the origin in its interior, and thus yields a *polytope norm* $\|\cdot\|_{\mathrm{NH}}$ (Hilbert norm) with $2\binom{d+1}{2} = d(d+1)$ facets. Reciprocally, let us notice that a norm induces a unit ball centered at the origin that is convex and symmetric around the origin.

The distance in the normed vector space between $v \in V^d$ and $v' \in V^d$ is defined by:

$$\rho_V(v, v') = \|v - v'\|_{\mathrm{NH}} = \inf\left\{\tau \ : \ v' \in \tau(B_V \oplus \{v\})\right\},$$

where $A \oplus B = \{a + b \ : \ a \in A, b \in B\}$ is the Minkowski sum.

The reverse map from the normed space V^d to the probability simplex Δ^d is given by:

$$\lambda^i = \frac{\exp(v^i)}{\sum_j \exp(v^j)}.$$

Thus we have $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$. In 1D, $(V^1, \|\cdot\|_{\mathrm{NH}})$ is isometric to the Euclidean line.

Note that computing the distance in the normed vector space requires naively $O(d^2)$ time.

Unfortunately, the norm $\|\cdot\|_{\mathrm{NH}}$ does not satisfy the parallelogram law.[3] Notice that a norm satisfying the parallelogram law can be associated with an inner product via the polarization identity. Thus the isometry of the Hilbert geometry to a normed vector space is not equipped with an inner product. However, all norms in a finite dimensional space are equivalent. This implies that in finite dimension, $(\Delta^d, \rho_{\mathrm{HG}})$ is *quasi-isometric* to the Euclidean space $\mathbb{R}^d$. An example of Hilbert geometry in infinite dimension is reported in [25]. Hilbert spaces are not CAT spaces except when $\mathscr{C}$ is an ellipsoid [76].

9 Hilbert Geometry with Finslerian/Riemannian Structures

In a Riemannian geometry, each tangent plane T_pM of a d-dimensional manifold M is equivalent to $\mathbb{R}^d$: $T_pM \simeq \mathbb{R}^d$. The inner product at each tangent plane T_pM can be visualized by an ellipsoid shape, a convex symmetric object centered at point p. In a *Finslerian geometry*, a norm $\|\cdot\|_p$ is defined in each tangent plane T_pM, and this norm is visualized as a symmetric convex object with non-empty interior. Finslerian geometry thus generalizes Riemannian geometry by taking into account generic symmetric convex objects instead of ellipsoids for inducing norms at each tangent plane. Any Hilbert geometry induced by a compact convex domain $\mathscr{C}$ can be expressed by an equivalent Finslerian geometry by defining the norm in T_p at p as follows [76]:

$$\|v\|_p = F_{\mathscr{C}}(p, v) = \frac{\|v\|}{2}\left(\frac{1}{pp^+} + \frac{1}{pp^-}\right),$$

where $F_{\mathscr{C}}$ is the *Finsler metric*, $\|\cdot\|$ is an *arbitrary norm* on $\mathbb{R}^d$, and p^+ and p^- are the intersection points of the line passing through p with direction v:

$$p^+ = p + t^+ v, \quad p^- = p + t^- v.$$

A geodesic γ in a Finslerian geometry satisfies:

$$d_{\mathscr{C}}(\gamma(t_1), \gamma(t_2)) = \int_{t_1}^{t_2} F_{\mathscr{C}}(\gamma(t), \dot{\gamma}(t))\mathrm{d}t.$$

In T_pM, a ball of center c and radius r is defined by:

$$B(c, r) = \{v \ : \ F_{\mathscr{C}}(c, v) \leq r\}.$$

[3] Consider $A = (1/3, 1/3, 1/3)$, $B = (1/6, 1/2, 1/3)$, $C = (1/6, 2/3, 1/6)$ and $D = (1/3, 1/2, 1/6)$. Then $2AB^2 + 2BC^2 = 4.34$ but $AC^2 + BD^2 = 3.84362411135$.

Thus any Hilbert geometry induces an equivalent Finslerian geometry, and since Finslerian geometries include Riemannian geometries, one may wonder which Hilbert geometries induce Riemannian structures? The only Riemannian geometries induced by Hilbert geometries are the *hyperbolic Cayley–Klein geometries* [27, 29, 30] with the domain $\mathscr{C}$ being an ellipsoid. The Finslerian modeling of information geometry has been studied in [71, 72].

There is not a canonical way of defining measures in a Hilbert geometry since Hilbert geometries are Finslerian but not necessary Riemannian geometries [76]. The Busemann measure is defined according to the Lebesgue measure λ of $\mathbb{R}^d$: Let B_p denote the unit ball wrt. to the Finsler norm at point $p \in \mathscr{C}$, and B_e the Euclidean unit ball. Then the Busemann measure for a Borel set $\mathscr{B}$ is defined by [76]:

$$\mu_{\mathscr{C}}(\mathscr{B}) = \int_{\mathscr{B}} \frac{\lambda(B_e)}{\lambda(B_p)} \mathrm{d}\lambda(p).$$

The existence and uniqueness of center points of a probability measure in Finsler geometry have been investigated in [77].

10 Bounding Hilbert Norm with Other Norms

Let us show that $\|v\|_{\mathrm{NH}} \leq \beta_{d,c}\|v\|_c$, where $\|\cdot\|_c$ is any norm. Let $v = \sum_{i=0}^d e_i x_i$, where $\{e_i\}$ is a basis of $\mathbb{R}^{d+1}$. We have:

$$\|v\|_c \leq \sum_{i=0}^{d} |x_i| \|e_i\|_c \leq \|x\|_2 \underbrace{\sqrt{\sum_{i=0}^{d} \|e_i\|_c^2}}_{\beta_d},$$

where the first inequality comes from the triangle inequality, and the second inequality is from the Cauchy–Schwarz inequality. Thus we have:

$$\|v\|_{\mathrm{NH}} \leq \beta_d \|x\|_2,$$

with $\beta_d = \sqrt{d+1}$ since $\|e_i\|_{\mathrm{NH}} \leq 1$.

Let $\alpha_{d,c} = \min_{\{v\,:\,\|v\|_c=1\}} \|v\|_{\mathrm{NH}}$. Consider $u = \frac{v}{\|v\|_c}$. Then $\|u\|_c = 1$ so that $\|v\|_{\mathrm{NH}} \geq \alpha_{d,c}\|v\|_c$. To find α_d, we consider the unit ℓ_2 ball in V^d, and find the smallest $\lambda > 0$ so that λB_V fully contains the Euclidean ball (Fig. 14).

Therefore, we have overall:

$$\alpha_d \|x\|_2 \leq \|v\|_{\mathrm{NH}} \leq \sqrt{d+1}\|x\|_2$$

In general, note that we may consider two arbitrary norms $\|\cdot\|_l$ and $\|\cdot\|_u$ so that:

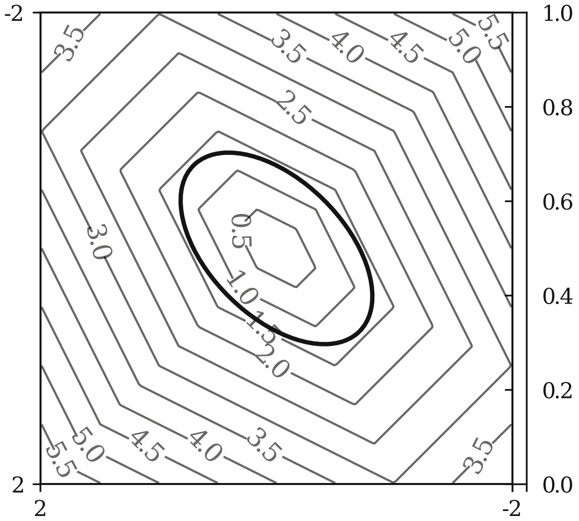


Fig. 14 Polytope balls B_V and the Euclidean unit ball B_E. From the figure the smallest polytope ball has radius ≈ 1.5

$$\alpha_{d,l}\|x\|_l \leq \|v\|_{\mathrm{NH}} \leq \beta_{d,u}\|x\|_u.$$

11 Funk Directed Metrics and Funk Balls

The Funk metric [78] wrt a convex domain $\mathscr{C}$ is defined by

$$F_{\mathscr{C}}(x, y) = \log\left(\frac{\|x-a\|}{\|y-a\|}\right),$$

where a is the intersection of the domain boundary and the affine ray $R(x, y)$ starting from x and passing through y. Correspondingly, the reverse Funk metric is

$$F_{\mathscr{C}}(y, x) = \log\left(\frac{\|y-b\|}{\|x-b\|}\right),$$

where b is the intersection of $R(y, x)$ with the boundary. The Funk metric is *not* a metric distance.

The Hilbert metric is simply the arithmetic symmetrization:

$$H_{\mathscr{C}}(x, y) = \frac{F_{\mathscr{C}}(x, y) + F_{\mathscr{C}}(y, x)}{2}.$$

It is interesting to explore clustering based on the Funk geometry, which we leave as a future work.

References

1. Agresti, A.: Categorical Data Analysis, vol. 482. Wiley, New Jercy (2003)
2. Aggarwal, C.C., Zhai, C.X.: Mining Text Data. Springer Publishing Company, Berlin (2012)
3. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: International Conference on Computer Vision, pp. 104–111. IEEE (2009)
4. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge (2012)
5. Chaudhuri, K., McGregor, A.: Finding metric structure in information theoretic clustering. In: Conference on Learning Theory (COLT), pp. 391–402 (2008)
6. Lebanon, G.: Learning Riemannian metrics. In: Conference on Uncertainty in Artificial Intelligence (UAI), pp. 362–369 (2002)
7. Rigouste, L., Cappé, O., Yvon, F.: Inference and evaluation of the multinomial mixture model for text clustering. Inf. Process. Manag. **43**(5), 1260–1280 (2007)
8. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. **2**(3), 283–304 (1998)
9. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1027–1035 (2007)
10. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theor. Comput. Sci. **38**, 293–306 (1985)
11. Tropp, J.A.: Simplicial faces of the set of correlation matrices. Discret. Comput. Geom. **60**(2), 512–529 (2018)
12. Kass, R.E., Vos, P.W.: Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics. Wiley-Interscience, New Jercy (1997)
13. Hotelling, H.: Spaces of statistical parameters. Bull. Amer. Math. Soc. **36**, 191 (1930)
14. Rao, C.R.: Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**(3), 81–91 (1945)
15. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. Breakthroughs in Statistics, pp. 235–247. Springer, New York (1992)
16. Stigler, S.M.: The epic story of maximum likelihood. Stat. Sci. **22**(4), 598–620 (2007)
17. Amari, Si: Information Geometry and Its Applications. Applied Mathematical Sciences, vol. 194. Springer, Japan (2016)
18. Calin, O., Udriste, C.: Geometric Modeling in Probability and Statistics. Mathematics and Statistics. Springer International Publishing, New York (2014)
19. Amari, Si, Cichocki, A.: Information geometry of divergence functions. Bull. Pol. Acad. Sci.: Tech. Sci. **58**(1), 183–195 (2010)
20. Shima, H.: The Geometry of Hessian Structures. World Scientific, Singapore (2007)
21. Liang, X.: A note on divergences. Neural Comput. **28**(10), 2045–2062 (2016)
22. Jenssen, R., Principe, J.C., Erdogmus, D., Eltoft, T.: The Cauchy–Schwarz divergence and Parzen windowing: connections to graph theory and mercer kernels. J. Frankl. Inst. **343**(6), 614–629 (2006)
23. Hilbert, D.: Über die gerade linie als kürzeste verbindung zweier punkte. Mathematische Annalen **46**(1), 91–96 (1895)
24. Busemann, H.: The Geometry of Geodesics. Pure and Applied Mathematics, vol. 6. Elsevier Science, Amsterdam (1955)
25. de la Harpe, P.: On Hilbert's metric for simplices. Geometric Group Theory, vol. 1, pp. 97–118. Cambridge University Press, Cambridge (1991)
26. Lemmens, B., Nussbaum, R.: Birkhoff's version of Hilbert's metric and its applications in analysis. Handbook of Hilbert Geometry, pp. 275–303 (2014)
27. Richter-Gebert, J.: Perspectives on Projective Geometry: A Guided Tour Through Real and Complex Geometry. Springer, Berlin (2011)
28. Bi, Y., Fan, B., Wu, F.: Beyond Mahalanobis metric: Cayley–Klein metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2339–2347 (2015)

29. Nielsen, F., Muzellec, B., Nock, R.: Classification with mixtures of curved Mahalanobis metrics. In: IEEE International Conference on Image Processing (ICIP), pp. 241–245 (2016)
30. Nielsen, F., Muzellec, B., Nock, R.: Large margin nearest neighbor classification using curved Mahalanobis distances (2016). arXiv:1609.07082 [cs.LG]
31. Stillwell, J.: Ideal elements in Hilbert's geometry. Perspect. Sci. **22**(1), 35–55 (2014)
32. Bernig, A.: Hilbert geometry of polytopes. Archiv der Mathematik **92**(4), 314–324 (2009)
33. Nielsen, F., Sun, K.: Clustering in Hilbert simplex geometry. CoRR arXiv: abs/1704.00454 (2017)
34. Nielsen, F., Shao, L.: On balls in a polygonal Hilbert geometry. In: 33st International Symposium on Computational Geometry (SoCG 2017). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik (2017)
35. Laidlaw, D.H., Weickert, J.: Visualization and Processing of Tensor Fields: Advances and Perspectives. Mathematics and Visualization. Springer, Berlin (2009)
36. Lemmens, B., Walsh, C.: Isometries of polyhedral Hilbert geometries. J. Topol. Anal. **3**(02), 213–241 (2011)
37. Condat, L.: Fast projection onto the simplex and the ℓ_1 ball. Math. Program. **158**(1–2), 575–585 (2016)
38. Park, P.S.: Regular polytopic distances. Forum Geom. **16**, 227–232 (2016)
39. Boissonnat, J.D., Sharir, M., Tagansky, B., Yvinec, M.: Voronoi diagrams in higher dimensions under certain polyhedral distance functions. Discret. Comput. Geom. **19**(4), 485–519 (1998)
40. Bengtsson, I., Zyczkowski, K.: Geometry of Quantum States: An Introduction to Quantum Entanglement. Cambridge University Press, Cambridge (2017)
41. Nielsen, F.: Cramér–Rao lower bound and information geometry. Connected at Infinity II, pp. 18–37. Springer, Berlin (2013)
42. Chapman, D.G.: Minimum variance estimation without regularity assumptions. Ann. Math. Stat. **22**(4), 581–586 (1951)
43. Hammersley, H.: On estimating restricted parameters. J. R. Stat. Society. Ser. B (Methodol.) **12**(2), 192–240 (1950)
44. Nielsen, F., Sun, K.: On Hölder projective divergences. Entropy **19**(3), 122 (2017)
45. Nielsen, F., Nock, R.: Further heuristics for k-means: the merge-and-split heuristic and the (k, l)-means. arXiv:1406.6314 (2014)
46. Bachem, O., Lucic, M., Hassani, S.H., Krause, A.: Approximate k-means++ in sublinear time. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 1459–1467 (2016)
47. Nielsen, F., Nock, R.: Total Jensen divergences: definition, properties and k-means++ clustering (2013). arXiv:1309.7109 [cs.IT]
48. Ackermann, M.R., Blömer, J.: Bregman clustering for separable instances. Scandinavian Workshop on Algorithm Theory, pp. 212–223. Springer, Berlin (2010)
49. Manthey, B., Röglin, H.: Worst-case and smoothed analysis of k-means clustering with Bregman divergences. J. Comput. Geom. **4**(1), 94–132 (2013)
50. Endo, Y., Miyamoto, S.: Spherical k-means++ clustering. Modeling Decisions for Artificial Intelligence, pp. 103–114. Springer, Berlin (2015)
51. Nielsen, F., Nock, R., Amari, Si: On clustering histograms with k-means by using mixed α-divergences. Entropy **16**(6), 3273–3301 (2014)
52. Brandenberg, R., König, S.: No dimension-independent core-sets for containment under homothetics. Discret. Comput. Geom. **49**(1), 3–21 (2013)
53. Panigrahy, R.: Minimum enclosing polytope in high dimensions (2004). arXiv:cs/0407020 [cs.CG]
54. Saha, A., Vishwanathan, S., Zhang, X.: New approximation algorithms for minimum enclosing convex shapes. In: ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1146–1160 (2011)
55. Nielsen, F., Nock, R.: On the smallest enclosing information disk. Inf. Process. Lett. **105**(3), 93–97 (2008)

56. Sharir, M., Welzl, E.: A combinatorial bound for linear programming and related problems. STACS **92**, 567–579 (1992)
57. Welzl, E.: Smallest enclosing disks (balls and ellipsoids). New Results and New trends in Computer Science, pp. 359–370. Springer, Berlin (1991)
58. Nielsen, F., Nock, R.: Approximating smallest enclosing balls with applications to machine learning. Int. J. Comput. Geom. Appl. **19**(05), 389–414 (2009)
59. Arnaudon, M., Nielsen, F.: On approximating the Riemannian 1-center. Comput. Geom. **46**(1), 93–104 (2013)
60. Bâdoiu, M., Clarkson, K.L.: Smaller core-sets for balls. In: ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 801–802 (2003)
61. Nielsen, F., Hadjeres, G.: Approximating covering and minimum enclosing balls in hyperbolic geometry. International Conference on Networked Geometric Science of Information, pp. 586–594. Springer, Cham (2015)
62. Bădoiu, M., Clarkson, K.L.: Optimal core-sets for balls. Comput. Geom. **40**(1), 14–22 (2008)
63. Bachem, O., Lucic, M., Krause, A.: Scalable and distributed clustering via lightweight coresets (2017). arXiv:1702.08248 [stat.ML]
64. Nielsen, F., Nock, R.: On approximating the smallest enclosing Bregman balls. In: Proceedings of the Twenty-Second Annual Symposium on Computational Geometry, pp. 485–486. ACM (2006)
65. Nock, R., Nielsen, F.: Fitting the smallest enclosing Bregman ball. ECML, pp. 649–656. Springer, Berlin (2005)
66. Deza, M., Sikirić, M.D.: Voronoi polytopes for polyhedral norms on lattices. Discret. Appl. Math. **197**, 42–52 (2015)
67. Körner, M.C.: Minisum hyperspheres, Springer Optimization and Its Applications, vol. 51. Springer, New York (2011)
68. Reem, D.: The geometric stability of Voronoi diagrams in normed spaces which are not uniformly convex (2012). arXiv:1212.1094 [cs.CG]
69. Foertsch, T., Karlsson, A.: Hilbert metrics and Minkowski norms. J. Geom. **83**(1–2), 22–31 (2005)
70. Cencov, N.N.: Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs, vol. 53. American Mathematical Society, Providence (2000)
71. Cena, A.: Geometric structures on the non-parametric statistical manifold. Ph.D. thesis, University of Milano (2002)
72. Shen, Z.: Riemann-Finsler geometry with applications to information geometry. Chin. Ann. Math. Ser. B **27**(1), 73–94 (2006)
73. Khosravifard, M., Fooladivanda, D., Gulliver, T.A.: Confliction of the convexity and metric properties in f-divergences. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **90**(9), 1848–1853 (2007)
74. Dowty, J.G.: Chentsov's theorem for exponential families (2017). arXiv:1701.08895 [math.ST]
75. Doup, T.M.: Simplicial Algorithms on the Simplotope, vol. 318. Springer Science & Business Media, Berlin (2012)
76. Vernicos, C.: Introduction aux géométries de Hilbert. Séminaire de théorie spectrale et géométrie **23**, 145–168 (2004)
77. Arnaudon, M., Nielsen, F.: Medians and means in Finsler geometry. LMS J. Comput. Math. **15**, 23–37 (2012)
78. Papadopoulos, A., Troyanov, M.: From Funk to Hilbert geometry (2014). arXiv:1406.6983 [math.MG]

Jean-Louis Koszul and the Elementary Structures of Information Geometry

Frédéric Barbaresco

Abstract This paper is a scientific exegesis and admiration of Jean-Louis Koszul's works on homogeneous bounded domains that have appeared over time as elementary structures of Information Geometry. Koszul has introduced fundamental tools to characterize the geometry of sharp convex cones, as Koszul-Vinberg characteristic Function, Koszul Forms, and affine representation of Lie Algebra and Lie Group. The 2nd Koszul form is an extension of classical Fisher metric. Koszul theory of hessian structures and Koszul forms could be considered as main foundation and pillars of Information Geometry.

Keywords Koszul-Vinberg characteristic function · Koszul forms
Affine representation of lie algebra and lie group
Homogeneous bounded domains

1 Preamble

> *«La Physique mathématique, en incorporant à sa base la notion de groupe, marque la suprématie rationnelle…Chaque géométrie – et sans doute plus généralement chaque organisation mathématique de l'expérience – est caractérisée par un groupe spécial de transformations… Le groupe apporte la preuve d'une mathématique fermée sur elle-même. Sa découverte clôt l'ère des conventions, plus ou moins indépendantes, plus ou moins cohérentes»* - **Gaston Bachelard, Le nouvel esprit scientifique, 1934**

In this article, I will pay tribute to a part of Professor Jean-Louis Koszul's work and fundamental and deep contributions of this great algebraist and geometer in the field of Information Geometry, which have many applications in the domain of applied mathematics, and in the emerging applications of Artificial Intelligence where the most efficient and robust algorithms are based on the natural gradient of the information geometry deduced from the Fisher matrix, as Yann Ollivier recently showed [1, 2].

F. Barbaresco (✉)
Thales Land & Air Systems, Voie Pierre-Gilles de Gennes, F91470 Limours, France
e-mail: frederic.barbaresco@thalesgroup.com

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_12

After the seminal papers of Fréchet [3], Rao [4] and Chentsov [5], many mathematicians and physicists have studied Information Geometry. One can quote in mathematics, the works of Amari [6, 7] in the 80 s, which does not refer to the Koszul publications of the 50s and 60s where Koszul introduced the elementary structures of the Hessian geometries, and generalized the Fisher metric for homogeneous convex domains. In the physical field, many physicists have also addressed Information Geometry, without references to Koszul. Weinhold [8] in 1976 and Ruppeiner [9] in 1979 empirically introduced the inverse dual metric defined by the Hessian of Entropy, or Ingarden [10, 11] in 1981 in Statistical Physics. Mrugala [12, 13] in 1978, and Janyszek [14] in 1989, tried to geometrize Thermodynamics by jointly addressing Information Geometry and Contact Geometry. All these authors were not familiar with Representations Theory introduced by Kirillov, and more particularly the affine representation of Lie groups and Lie algebras, used and developed by Koszul in mathematics and by Souriau in statistical mechanics [79–84]. It thus appears that the first foundations of the information geometry goes back to Fréchet's paper of 1943 [3] (and his Lecture given during the winter of 1939 at the Institut Henri Poincaré), who first introduced the Clairaut(-Legendre) equation (fundamental equation in Information Geometry between dual potentials) and Fisher metric as the Hessian of a convex function. This Fréchet's seminal work was followed by Koszul's 50's papers [15, 16] which introduced new forms that generalize Fisher metric for sharp convex cones. It was not until 1969 that Souriau completed this extension in the framework of the Lie Group Thermodynamics with a cohomological definition of Fisher metric [17]. This last extension was developed by Koszul at the beginning of 80's in his Lecture "Introduction to Symplectic Geometry" [18]. I will conclude this survey by making reference to Balian [19], who has developed during 80's Information Geometry in Quantum Physics with a Quantum Fisher metric given by Von Neumann Entropy hessian [20].

Inspired by the French mathematical tradition, and the teachings of his master Elie Cartan (Koszul was PhD student of Henri Cartan but was greatly influenced by Elie Cartan), Jean-Louis Koszul was a real "avant-garde", if we take the definition given by Clausewitz « *An avant-garde is a group of units intended to move in front of the army to: explore the terrain to avoid surprises, quickly occupy the strong positions of the battlefield (high points), screen and contain the enemy the time the army can deploy*". Indeed, Jean-Louis Koszul was a pioneer, who explored and cleared many areas of mathematics, detailed in the book "Selected papers of JL Koszul" [21]. What I will expose, in this paper, is therefore only one part of his work which concerns homogeneous bounded domains geometry, from seminal Elie Cartan's earlier work on symmetric bounded domains. In a letter from André Weil to Henri Cartan, cited in the proceedings of the conference "*Elie Cartan and today's mathematics*" in 1984, it says "*As to the symmetrical spaces, and more particularly to the symmetric bounded domains at the birth of which you contributed, I have kept alive the memory of the satisfaction I felt in finding some incarnations in Siegel from his first works on quadratic forms, and later to convince Siegel of the value of your father's ideas on the subject*". At this 1984 conference, two disciples of Elie Cartan gave a conference, Jean-Louis Koszul [22] and Jean-Marie Souriau (Fig. 1).

Fig. 1 (on the left) Jean-Louis Koszul student at ENS ULM in 1940, (on the right) Jean-Louis Koszul at GSI'13 "Geometric Science of Information" conference at the École des Mines de Paris August 2013

In the book "*Selected papers of JL Koszul*" [21], Koszul summarizes the work, I will detail in the following: "*It is with the problem of the determination of the homogeneous bounded domains posed by E. Cartan around 1935 that are related [my papers]. The idea of approaching the question through invariant Hermitian forms already appears explicitly in Cartan. This leads to an algebraic approach which constitutes the essence of Cartan's work and which, with the Lie J-algebras, was pushed much further by the Russian School* [23–36]. *It is the work of Piatetski Shapiro on the Siegel domains, then those of E.B. Vinberg on the homogeneous cones that led me to the study of the affine transformation groups of the locally flat manifolds and in particular to the convexity criteria related to invariant forms"*. In particular, J.L. Koszul source of inspiration is given in this last sentence of Elie Cartan's 1935 article [37]:

"*It is clear that if one could demonstrate that all homogeneous domains whose form* $\Phi = \sum_{i,j} \frac{\partial^2 \log K(z,z^*)}{\partial z_i \partial z_j^*} dz_i dz_j^*$ *is positive definite are symmetric, the whole theory of homogeneous bounded domains would be elucidated. This is a problem of Hermitian geometry certainly very interesting*". It was not until 1953 that the classification of non-Riemannian symmetric spaces has been achieved by Marcel Berger [38]. The work of Koszul has also been extended and deepened by one of his student Jacques Vey in [39, 40]. Jacques Vey has transposed the notion of hyperbolicity, developed

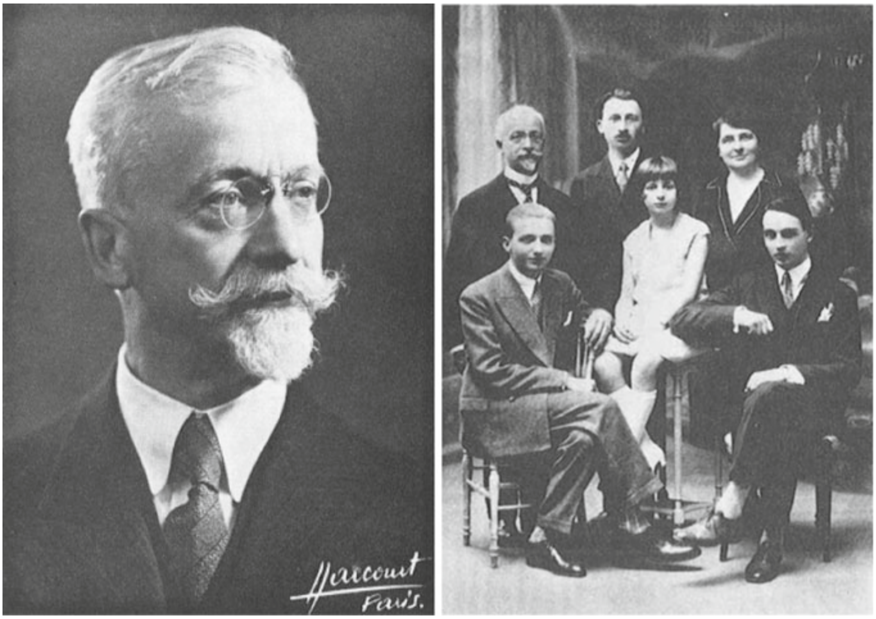


Fig. 2 (on the left) Professor Elie Cartan, (on the right) the Cartan family

by W. Kaup for Riemann surfaces, into the category of differentiable manifolds with flat linear connection (locally flat manifolds), which makes it possible to completely characterize the locally flat manifolds admitting as universal covering a convex open sharp cone of R^n, which had been studied by Koszul in [41]. The links between Koszul's work and those of Ernest B. Vinberg [23–30] were recently developed at the conference "*Transformation groups 2017*" in Moscow dedicated to the 80th anniversary of Professor EB Vinberg, in Dmitri Alekseevsky's talk on "*Vinberg's theory of homogeneous convex cones: developments and applications*" [42]. Koszul and Vinberg are actually associated with the concept of Koszul-Vinberg's characteristic function on convex cones, which I will develop later in the paper. Koszul introduced the so-called "*Koszul forms*" and a canonical metric given by the Hessian of the opposite of the logarithm of this Koszul-Vinberg characteristic function, from which I will show the links with Fisher's metric in Information Geometry, and its extension (Fig. 2).

Professor Koszul's main papers, which form the elementary structures of information geometry, are as follows:

- ***«Sur la forme hermitienne canonique des espaces homogènes complexes»*** [15] of 1955: Koszul considers the Hermitian structure of a homogeneous *G/B* manifold (*G* related Lie group and *B* a closed subgroup of *G*, associated, up to a constant factor, to the single invariant *G*, and to the invariant complex structure by the operations of *G*). Koszul says "*The interest of this form for the determination of* *homogeneous bounded domains has been emphasized by Elie Cartan: a necessary condition for G/B to be a bounded domain is indeed that this form is positive definite*". Koszul calculated this canonical form from infinitesimal data Lie algebra of *G,* the sub-algebra corresponding to *B* and an endomorphism algebra defining the invariant complex structure of *G/B.* The results obtained by Koszul proved that the homogeneous bounded domains whose group of automorphisms is semi-simple are bounded symmetric domains in the sense of Elie Cartan. Koszul also refers to André Lichnerowicz's work on Kählerian homogeneous spaces [43]. In this seminal paper, Koszul also introduced a left invariant form of degree 1 on *G*: $\Psi(X) = Tr_{g/b}[ad(JX) - J.ad(X)] \quad \forall X \in g$ with *J* an endomorphism of the Lie algebra space and the trace $Tr_{g/b}[.]$ corresponding to that of the endomorphism *g/b.* The Kähler form of the canonical Hermitian form is given by the differential of $-1/4\Psi(X)$ of this form of degree 1.

- ***«Exposés sur les espaces homogènes symétriques»*** [16] of 1959 is a Lecture written as part of a seminar held in September and October 1958 at the University of Sao Paulo, which details the determination of homogeneous bounded domains. He returned to [15] and showed that any symmetric bounded domain is a direct product of irreducible symmetric bounded domains, determined by Elie Cartan (4 classes corresponding to classical groups and 2 exceptional domains). For the study of irreducible symmetric bounded domains, Koszul refered to Elie Cartan, Carl-Ludwig Siegel and Loo-Keng Hua. Koszul illustrated the subject with two particular cases, the half-plane of Poincaré and the half-space of Siegel, and showed that with its trace formula of endomorphism *g/b,* he found that the canonical Kähler hermitian form and the associated metrics are the same as those introduced by Henri Poincaré and Carl-Ludwig Siegel [44] (who introduced them as invariant metric under action of the automorphisms of these spaces).
- ***«Domaines bornées homogènes et orbites de groupes de transformations affines»*** [45] of 1961 is written by Koszul at the Institute for Advanced Study at Princeton during a stay funded by the National Science Foundation. On a complex homogeneous space, an invariant volume defines with the complex structure the canonical invariant Hermitian form introduced in [15]. If the homogeneous space is holomorphically isomorphic to a bounded domain of a space C^n, this Hermitian form is positive definite because it coincides with the Bergmann metric of the domain. Koszul demonstrated in this article the reciprocal of this proposition for a class of complex homogeneous spaces. This class consists of some open orbits of complex affine transformation groups and contains all homogeneous bounded domains. Koszul addressed again the problem of knowing if a complex homogeneous space, whose canonical Hermitian form is positive definite is isomorphic to a bounded domain, but via the study of the invariant bilinear form defined on a real homogeneous space by an invariant volume and an invariant flat connection. Koszul demonstrated that if this bilinear form is positive definite then the homogeneous space with its flat connection is isomorphic to a convex open domain containing no straight line in a real vector space and extended it to the initial problem for the complex homogeneous spaces obtained in defining a complex structure in the variety of vectors of a real homogeneous space provided with an invariant

flat connection. It is in this article that Koszul used the affine representation of Lie groups and algebras. By studying the open orbits of the affine representations, he introduced an affine representation of G, written (f, q), and the following equation setting f the linear representation of the Lie algebra g of G, defined by f and q the restriction to g and the differential of q (f and q are differential respectively of f and q):

$$f(X)q(Y) - f(Y)q(X) = q([X, Y]) \quad \forall X, Y \in \mathsf{g}$$
$$\text{with} \quad f : \mathsf{g} \to gl(E) \quad \text{and} \quad q : \mathsf{g} \mapsto E$$

- ***«Ouverts convexes homogènes des espaces affines»*** [46] of 1962. Koszul is interested in this paper by the structure of the convex open non-degenerate Ω (with no straight line) and homogeneous (the group of affine transformations of E leaving stable Ω operates transitively in Ω) in a real affine space of finite dimension. Koszul demonstrated that they can be all deduced from non-degenerate and homogeneous convex open cones built in [45]. He used for this the properties of the group of affine transformations leaving stable a non-degenerate convex open domain and an homogeneous domain.
- ***«Variétés localement plates et convexité»*** [41] of 1965. Koszul established the following theorem: let *M be* a locally related differentiable manifold. If the universal covering of *M* is isomorphic as a flat manifold with a convex open domain containing no straight line in a real affine space, then there exists on *M* a closed differential form α such that $D\alpha$ *(D* linear covariant derivative of zero torsion) is positive definite in all respects and which is invariant under every automorphism of *M*. If *G* is a group of automorphisms of *M* such that *G\M* is quasi-compact and if there exists on *M* a closed 1-differential form α invariant by *G* and such that $D\alpha$ is positive definite at any point, then the universal covering of *M* is isomorphic as a flat manifold with a convex open domain that does not contain a straight line in a real affine space.
- ***«Lectures on Groups of Transformations»*** [47] of 1965. This is lecture notes given by Koszul at Bombay "Tata Institute of Fundamental Research " on transformation groups. In particular in Chap. 6, Koszul studied discrete linear groups acting on convex open cones in vector spaces based on the work of C.L. Siegel (work on quadratic forms [48]). Koszul used what I will call in the following Koszul-Vinberg characteristic function on convex sharp cone.
- ***«Déformations des variétés localement plates»*** [49] of 1968. Koszul provided other proofs of theorems introduced in [41]. Koszul considered related differentiable manifolds of dimension n and *TM* the fibered space of M. The linear connections on *M* constitute a subspace of the space of the differentiable applications of the *TM*x*TM* fiber *product* in the space *T(TM)* of the *TM* vectors. Any locally flat connection *D* (the curvature and the torsion are zero) defines a locally flat connection on the covering of *M,* and is hyperbolic when universal covering of *M,* with this connection, is isomorphic to a sharp convex open domain (without straight lines) in R^n. Koszul showed that, if *M* is a compact manifold, for a locally

flat connection on M to be hyperbolic, it is necessary and sufficient that there exists a closed differential form of degree 1 on M whose covariant differential is positive definite.

- ***«Trajectoires Convexes de Groupes Affines Unimodulaires»*** [50] in 1970 . Koszul demonstrated that a convex sharp open domain in R^n that admits a unimodular transitive group of affine automorphisms is an auto-dual cone. This is a more geometric demonstration of the results shown by Ernest Vinberg [29] on the automorphisms of convex cones.

The elementary geometric structures discovered by Jean-Louis Koszul are the foundations of Information Geometry. These links were first established by Professor Hirohiko Shima [51–56]. These links were particularly crystallized in Shima book 2007 "*The Geometry of Hessian Structures*" [57], which is dedicated to Professor Koszul. The origin of this work followed the visit of Koszul in Japan in 1964, for a mission coordinated with the French government. Koszul taught lectures on the theory of flat manifolds at Osaka University. Hirohiko Shima was then a student and attended these lectures with the teachers Matsushima and Murakami. This lecture was at the origin of the notion of Hessian structures and the beginning of the works of Hirohiko Shima. Henri Cartan noted concerning Koszul's ties with Japan, "*Koszul has attracted eminent mathematicians from abroad to Strasbourg and Grenoble. I would like to mention in particular the links he has established with representatives of the Japanese School of Differential Geometry*". Shima's book [57] is a systematic introduction to the theory of Hessian structures (provided by a pair of a flat connection D and an Hessian metric *g).* Koszul studied flat manifolds with a closed 1-form α, such that $D\alpha$ be positive definite, where $D\alpha$ is a hessian metric. However, not all Hessian metrics are globally of the form $g = D\alpha$. Shima introduces the notion of Codazzi structure for a pair (D,g), with D a torsion-free connection, which verifies the Codazzi equation $(D_X g)(Y, Z) = (D_Y g)(X, Z)$. A Hessian structure is a Codazzi structure for which connection D is flat. This is an extension of Riemannian geometry. It is then possible to define a connection *D'* and a dual Codazzi structure *(D',g)* with $D' = \nabla - D$ where ∇ is the Levi-Civita connection. For a hessian structure (D, g) with $g = Dd\varphi$, the dual Codazzi structure $\left(D', g\right)$ is also a Hessian structure and $g = D'd\varphi'$, where φ' is the Legendre transform of $\varphi : \varphi' = \sum_i x^i \frac{\partial \varphi}{\partial x^i} - \varphi$. Shima observed that Information Geometry framework could be introduced by dual connections, and not only founded on Fréchet, Rao and Chentsov works [5]. A hessian structure *(D, g)* is of Koszul type, if there is a closed 1-form ω as $g = D\omega$. Using D and the volume element of g, Koszul introduced a 2nd form, which plays a similar role to the Ricci tensor for a Kählerian metric. Let υ be the volume element of g, we define a closed 1-form α such that $D_X \upsilon = \alpha(X)\upsilon$ and a symmetric bilinear form $\gamma = D\alpha$. In the following, α and γ forms are called 1st and 2nd form of Koszul for Hessian structure (D, g). We can consider the forms associated with the Hessian dual structure (D', g) by $\alpha' = -\alpha$ and $\gamma' = \gamma - 2\nabla\alpha$. In the case of a homogeneous regular convex cone Ω, with D the canonical flat connection of the ambient vector space, the Koszul forms α and γ for the canonical Hessian structure $(D, g = Dd\psi)$

Fig. 3 From left to right, Jean-Louis Koszul, Hirohiko Shima and Michel Nguiffo Boyom at GSI'13 (Geometric Science of Information) conference at the École des Mines of Paris in August 2013

are given by $\alpha = d \log \psi$ and $\gamma = g$. The volume element υ determined by g is invariant under the action of the group of automorphisms G of Ω.

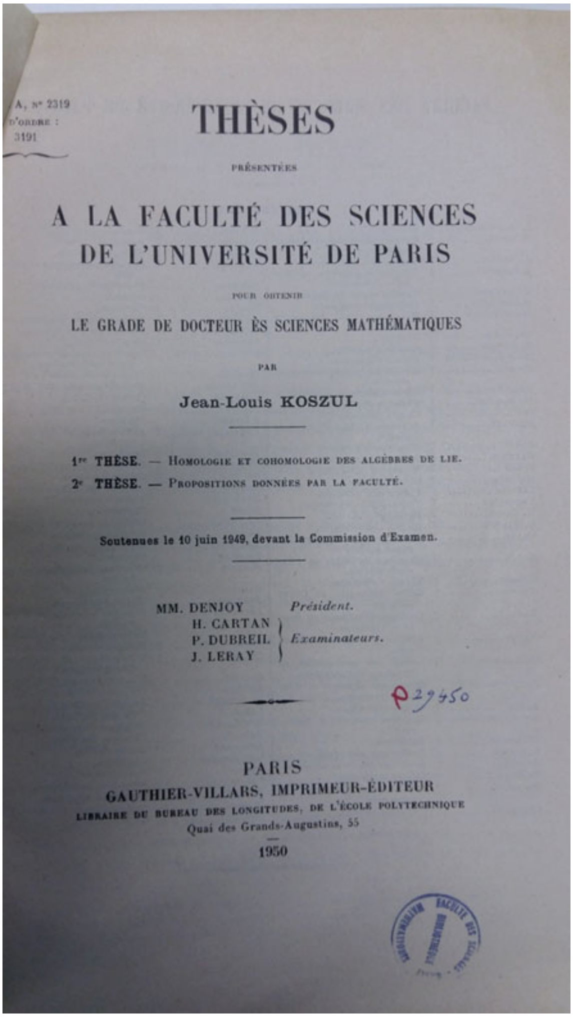
Jean-Louis Koszul attended the 1st GSI "*Geometric Science of Information*" conference in August 2013 at the Ecole des Mines in Paris, where he attended the presentation of Hirohiko Shima, given for his honor on the topic "*Geometry of Hessian Structures* " [58]. In the photo below, we can see from left to right, Jean-Louis Koszul, Hirohiko Shima and Michel Nguiffo Boyom. Professor Michel Boyom has extensively studied and developed, at the University of Montpellier, Koszul models [59–66] in relation to symplectic flat affine manifolds and to the cohomology of Koszul-Vinberg algebras (KV Cohomology). Professor Boyom with his PhD student Byande [67, 68] have explored other links with Information Geometry. André Lichnerowicz worked in parallel on a closed topic about homogeneous Kähler manifolds [69] (Fig. 3).

2 Biographical Reminder of Jean-Louis Koszul Scientific Life

Jean Louis André Stanislas Koszul born in Strasbourg in 1921, is the child of a family of four (with three older sisters, Marie Andrée, Antoinette and Jeanne). He is the son of André Koszul (born in Roubaix on November 19th 1878, professor at the

Strasbourg university), and Marie Fontaine (born in Lyon on June 19th 1887), who was a friend of Henri Cartan's mother. Henri Cartan writes on this friendship "*My mother in her youth, had been a close friend of the one who was to become Jean-Louis Koszul's mother*" [70]. His paternal grandparents were Julien Stanislas Koszul and Hélène Ludivine Rosalie Marie Salomé. He attended high school in Fustel-de-Coulanges in Strasbourg and the Faculty of Science in Strasbourg and in Paris. He entered ENS Ulm in the class of 1940 and defended his thesis with Henri Cartan. Henri Cartan noted "*This promotion included other mathematicians like Belgodère or Godement, and also physicists and some chemists, like Marc Julia and Raimond Castaing*" [70] (for the anecdote, the maiden name of my wife Anne, is *Belgodère*, with a filial link with Paul Belgodère of the Koszul ENS promotion). Jean-Louis Koszul married on July 17th 1948 with Denise Reyss-Brion, student of ENS Sèvres, entered in 1941. They have three children, Michel (married to Christine Duchemin), Anne (wife of Stanislas Crouzier) and Bertrand. He then taught in Strasbourg and was appointed Associate Professor at the University of Strasbourg in 1949, and had for colleagues R. Thom, M. Berger and B. Malgrange. He was promoted to professor status in 1956. He became a member of Bourbaki with the 2nd generation, J. Dixmier, R. Godement, S. Eilenberg, P. Samuel, J. P. Serre and L. Schwartz. Henri Cartan remarked in [70] "*In the vehement discussions within Bourbaki, Koszul was not one of those who spoke loudly; but we learned to listen to him because we knew that if he opened his mouth he had something to say*". About this Koszul's period at Strasbourg University, Pierre Cartier [71] said "*When I arrived in Strasbourg, Koszul was returning from a year spent in Institute for Advanced Studies in Princeton, and he was after the departure of Ehresman and Lichnerowicz to Paris the paternal figure of the Department of Mathematics (despite his young age). I am not sure of his intimate convictions, but he represented for me a typical figure of this Alsatian Protestantism, which I frequented at the time. He shared the seriousness, the honesty, the common sense and the balance. In particular, he knew how to resist the academic attraction of Paris. He left us after 2 years to go to Grenoble, in a maneuver uncommon at the time of exchange of positions with Georges Reeb*". He became Senior Lecturer at the University of Grenoble in 1963, and then an honorary professor at the Joseph Fourier University [72] and integrated in Fourier Institute led by C. Chabauty. During this period, B. Malgrange [73] remembered Koszul seminar on "algebra and geometry" with his three students J. Vey, D. Luna and J. Helmstetter. In Grenoble, he practiced mountaineering and was a member of the French Alpine Club. Koszul was awarded by Jaffré Prize in 1975 and was elected correspondent at the Academy of Sciences on January 28th 1980. Koszul was one of the CIRM conference center founder at Luminy. The following year, he was elected to the Academy of São Paulo. Jean-Louis Koszul died on January 12th 2018, at the age of 97.

As early as 1947, Jean-Louis Koszul published three articles in CRAS of the Academy of Sciences, on the Betti number of a simple compact Lie group, on cohomology rings, generalizing ideas of Jean-Leray, and finally on the homology of homogeneous spaces. Koszul's thesis, defended in June 10th 1949 under the direction of Henri Cartan, dealt with the homology and cohomology of Lie algebras [74]. The jury was composed of M. Denjoy (President), J. Leray, P. Dubreil and H. Cartan.



A, n° 2319
D'ORDRE :
3191

THÈSES

PRÉSENTÉES

A LA FACULTÉ DES SCIENCES
DE L'UNIVERSITÉ DE PARIS

POUR OBTENIR

LE GRADE DE DOCTEUR ÈS SCIENCES MATHÉMATIQUES

PAR

Jean-Louis KOSZUL

1re THÈSE. — Homologie et cohomologie des algèbres de Lie.
2e THÈSE. — Propositions données par la Faculté.

Soutenues le 10 juin 1949, devant la Commission d'Examen.

MM. DENJOY *Président.*
H. CARTAN
P. DUBREIL *Examinateurs.*
J. LERAY

P 29450

PARIS
GAUTHIER-VILLARS, IMPRIMEUR-ÉDITEUR
LIBRAIRE DU BUREAU DES LONGITUDES, DE L'ÉCOLE POLYTECHNIQUE
Quai des Grands-Augustins, 55

1950

Fig. 4 Cover page of Koszul's PhD report defended June 10th 1949 with a Jury composed of Professors Arnaud Denjoy, Henri Cartan, Paul Dubreil and Jean Leray, published in [74]

Under the title "Works of Koszul I, II and III", Henri Cartan reported Koszul's PhD results to Bourbaki seminar [75–77]. See also, André Haefliger paper [78] (Fig. 4).

In 1987, an International Symposium on Geometry was held in Grenoble in honor of Jean-Louis Koszul, whose proceedings were published in "*les Annales de l'Institut Fourier*", Volume 37, No. 4. This conference began with a presentation by Henri Cartan, who remembered the mention given to Koszul for his aggregation [70]: "*Distinguished Spirit; he is successful in his problems. Should beware, orally, of overly systematic trends. A little less subtle complications, baroque ideas, a little more common sense and balance would be desirable*". About his supervision of Koszul's PhD, Henri Cartan writed "*Why did he turn to guide him (so-called)? Is it because he found inspiration in Elie Cartan's work on the topology of Lie groups?*

Perhaps he was surprised to note that mathematical knowledge is not necessarily transmitted by descent. In any case, he helped me to better know what my father had brought to the theory" [70]. On the work of Koszul algebrisation, Henri Cartan notes " *Koszul was the first to give a precise algebraic formalization of the situation studied by Leray in his 1946 publication, which became the theory of the spectral sequence. It took a good deal of insight to unravel what lay behind Leray's study. In this respect, Koszul's Note in the July 1947 CRAS is of historical significance.*" [70]. From June 26th to July 2nd 1947, CNRS, received an International conference in Paris, on "*Algebraic Topology*". This was the first postwar international diffusion of Leray's ideas. Koszul writes about this lecture "*I can still see Leray putting his chalk at the end of his talk by saying (modestly?) that he definitely did not understand anything about Algebraic Topology*". In writing his lectures at the Collège de France, Leray adopted the algebraic presentation of the spectral suite elaborated by Koszul. As early as 1950, J.P. Serre used the term "Leray-Koszul suite". Speaking of Leray, Koszul wrote "*around 1955 I remember asking him what had put him on the path of what he called the ring of homology of a representation in his Notes to the CRAS of 1946. His answer was Künneth's theorem; I could not find out more*". The sheaf theory, introduced by Jean-Leray, followed in 1947, at the same time as the spectral sequences.

In 1950, Koszul published an important book of 62 pages entitled "*Homology and Cohomology of Lie Algebras*" [74] based on his PhD work, in which he studied the links between homology and cohomology (with real coefficients) of a compact connected Lie group and purely algebraic problems of Lie algebra. Koszul then gave a lecture in São Paulo on the topic "*sheaves and cohomology*". The superb lecture notes were published in 1957 and dealt with the cohomology of Čech with coefficients in a sheaf. In the autumn of 1958, he again organized a series of seminars in São Paulo, this time on symmetric spaces [16]. R. Bott commented on these seminars "*very pleasant. The pace is fast, and the considerable material is covered elegantly. In addition to the more or less standard theorems on symmetric spaces, the author discusses the geometry of geodesics, Bergmann's metrics, and finally studies the bounded domains with many details*". In the mid-1960s, Koszul taught at the Tata Institute in Bombay on transformation groups [47] and on fiber bundles and differential geometry. The second lecture dealt with the theory of connections and the lecture notes were published in 1965. In 1986 he published "*Introduction to symplectic geometry*" [18] following a Chinese course in China (with the agreement of Jean-Louis Koszul given in 2017, this lecture given at the University of Nanjing will be translated into English by Springer and will be published in 2018). This book takes up and develops works of Jean-Marie Souriau [17, 79] on homogeneous symplectic manifolds and the affine representation of Lie algebras and Lie groups in geometric mechanics (another fundamental source of Information Geometry structures extended on homogeneous varieties [80–84]). Chuan Yu Ma writes in a review, on this latest book in Chinese, that "*This work coincided with developments in the field of analytical mechanics. Many new ideas have also been derived using a wide variety of notions of modern algebra, differential geometry, Lie groups, functional analysis, differentiable manifolds, and representation theory. [Koszul's book] emphasizes the* *differential-geometric and topological properties of symplectic manifolds. It gives a modern treatment of the subject that is useful for beginners as well as for experts*".

In 1994, in [21], a comment by Koszul explains the problems he was preoccupied with when he invented what is now called the "*Koszul complex*". This was introduced to define a theory of cohomology for Lie algebras and proved to be a general structure useful in homological algebra.

3 Koszul-Vinberg Characteristic Function, Koszul Forms and Maximum Entropy Density

Through the study of the geometry of bounded homogeneous domains initiated by Elie Cartan [37, 85], Jean-Louis Koszul discovered that the elementary structures are associated with Hessian manifolds on sharp convex cones [15, 16, 41, 45–47, 49, 50]. In 1935, Elie Cartan proved in [37] that the symmetric homogeneous irreducible bounded domains could be reduced to 6 classes, 4 canonical models and 2 exceptional cases. Ilya Piatetski-Shapiro [31–35], after Luogeng Hua [86], extended Siegel's description [44, 48] to other symmetric spaces, and showed by a counterexample that Elie Cartan's conjecture, that all transitive domains are symmetrical, was false. At the same time, Ernest B. Vinberg [23–30] worked on the theory of homogeneous convex cones and the construction of Siegel domains [44, 48]. More recently, the classical complex symmetric spaces were studied by F. Berezin [87, 88] in the context of quantification. In parallel, O.S. Rothaus [89] and Piatetski-Shapiro [31–35] with Karpelevitch, explored the underlying geometry of these complexes homogeneous fields, and more particularly the fibration areas on the components of the shilov boundary. In Italy, I note the work of E. Vessentini [90] and U. Sampieri [91, 92]. The Siegel domains, which fit into these classes of structures, nowadays play an important role in the processing of radar spatio-temporal signals and, more broadly, in learning from structured covariance matrices.

Jean-Louis Koszul and Ernest B. Vinberg have introduced a hessian metric invariant by the group of linear automorphisms on a sharp convex cone Ω through a function, called characteristic function ψ. In the following Ω is a sharp convex cone in a vector space E of finite size on R (a convex cone is sharp if there is no straight lines). In dual space E^* of E, Ω^* is the set of linear strictly positive forms on $\overline{\Omega}-\{0\}$. Ω^*, dual cone of Ω, is also a sharp convex cone. If $\xi \in \Omega^*$, then intersection $\Omega \cap \{x \in E / \langle x, \xi \rangle = 1\}$ is bounded. $G = Aut(\Omega)$ is the group of linear transformation from E *that preserves* Ω (group of automorphisms). $G = Aut(\Omega)$ acts on Ω^* such that, $\forall g \in G = Aut(\Omega), \forall \xi \in E^*$ then $\bar{g}.\xi = \xi \circ g^{-1}$. Koszul introduce an integral, of Laplace kind, on sharp dual convex cone, as:

Koszul-Vinberg Characteristic definition:

Let $d\xi$ Lebesque measure on E^*, following integral:

$$\psi_\Omega(x) = \int\limits_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \quad \forall x \in \Omega \tag{1}$$

with Ω^* the dual cone, is analytical function on Ω, with $\psi_\Omega(x) \in]0, +\infty[$, called Koszul-Vinberg characteristic function of cone Ω.

Nota: the logarithm of the characteristic function is called « barrier function » for convex optimization algorithms. Yurii Nesterov and Arkadii Nemirovskii [93] have proved in modern theory of « *interior point* », using function $\Theta_\Omega(x) = \log(vol_n\{s \in \Omega^*/\langle s, x \rangle \leq 1\})$, that all convex cones in R^n have a self-dual barrier, linked with Koszul characteristic function.

Koszul-Vinberg Characteristic function has the following properties:

- Bergman kernel of $\Omega + i R^{n+1}$ is written $K_\Omega(\mathrm{Re}(z))$ up to a constant. K_Ω is defined by integral:

$$K_\Omega(x) = \int\limits_{\Omega^*} e^{-\langle \xi, x \rangle} \psi_{\Omega^*}(\xi)^{-1} d\xi \tag{2}$$

- ψ_Ω is an analytical function defined in the interior of Ω and $\psi_\Omega(x) \to +\infty$ when $x \to \partial\Omega$. If $g \in Aut(\Omega)$ then $\psi_\Omega(gx) = |\det g|^{-1}\psi_\Omega(x)$ and as $tI \in G = Aut(\Omega)$ for all $t > 0$, we have:

$$\psi_\Omega(tx) = \psi_\Omega(x)/t^n \tag{3}$$

- ψ_Ω is strictly log convex, such that $\phi_\Omega(x) = \log(\psi_\Omega(x))$ is strictly convex.

From this characteristic function, Koszul introduced two forms:

1st Koszul form $\boldsymbol{\alpha}$: Differential 1-form

$$\alpha = d\phi_\Omega = d\log\psi_\Omega = d\psi_\Omega/\psi_\Omega \tag{4}$$

is invariant with respect to all automorphisms $G = Aut(\Omega)$ of Ω. If $x \in \Omega$ and $u \in E$ then:

$$\langle \alpha_x, u \rangle = -\int\limits_{\Omega^*} \langle \xi, u \rangle . e^{-\langle \xi, x \rangle} d\xi \text{ and } \alpha_x \in -\Omega^* \tag{5}$$

and

2nd Koszul form $\boldsymbol{\gamma}$: Differential symmetric 2-form

$$\gamma = D\alpha = Dd \; \log\psi_\Omega \tag{6}$$

is a bilinear symmetric positive definite form invariant with respect to the action of $G = Aut(\Omega)$ and $D\alpha > 0$

Positivity is given by Schwarz inequality and:

$$Dd\ \log\psi_\Omega(u,v) = \int\limits_{\Omega^*} \langle \xi, u\rangle \langle \xi, v\rangle e^{-\langle \xi, u\rangle} d\xi \tag{7}$$

Koszul has proved that from this 2nd form, we can introduce an invariant Riemannian metric with respect to the action of cone automorphisms:
Koszul Metric: $D\alpha$ defines a Riemannian invariant structure by $Aut(\Omega)$, and the Riemannian metric is given by:

$$g = Dd\ \log\psi_\Omega \tag{8}$$

$$(Dd\ \log\psi(x))(u) = \frac{1}{\psi(u)^2}\left[\int\limits_{\Omega^*} F(\xi)^2 d\xi . \int\limits_{\Omega^*} G(\xi)^2 d\xi - \left(\int\limits_{\Omega^*} F(\xi).G(\xi) d\xi\right)^2\right] > 0$$

$$\text{with } F(\xi) = e^{-\frac{1}{2}\langle x,y\rangle} \quad \text{and} \quad G(\xi) = e^{-\frac{1}{2}\langle x,\xi\rangle}\langle u,\xi\rangle \tag{9}$$

The positivity could be proved by using Schwarz inequality, and the following properties for the derivative given by $d\ \log\psi = \frac{d\psi}{\psi}$ and $Dd\ \log\psi = \frac{Dd\psi}{\psi} - \left(\frac{d\psi}{\psi}\right)^2$ where $(d\psi(x))(u) = -\int\limits_{\Omega^*} e^{-\langle x,\xi\rangle}\langle u,\xi\rangle d\xi$ and $(Dd\psi(x))(u) = -\int\limits_{\Omega^*} e^{-\langle x,\xi\rangle}\langle u,\xi\rangle^2 d\xi$.

Koszul uses this diffeomorphism to define dual coordinates:

$$x^* = -\alpha_x = -d\ \log\psi_\Omega(x) \tag{10}$$

with $\langle df(x), u\rangle = D_u f(x) = \frac{d}{dt}\Big|_{t=0} f(x+tu)$. When the cone Ω is symmetric, the map $x \mapsto x^* = -\alpha_x$ is a bijection and an isometry with only one fixed point (the manifold is a symmetric Riemannian space given by its isometry):

$$(x^*)^* = x, \left\langle x, x^*\right\rangle - n\ \text{et}\ \psi_\Omega(x)\psi_{\Omega^*}(x^*) = cste \tag{11}$$

x^* is characterized by $x^* = \arg\ \min\{\psi(y)/y \in \Omega^*, \langle x,y\rangle = n\}$ and x^* is the gravity center of the transverse cut $\{y \in \Omega^*, \langle x,y\rangle = n\}$ of Ω^*:

$$x^* = \int\limits_{\Omega^*} \xi.e^{-\langle \xi,x\rangle} d\xi / \int\limits_{\Omega^*} e^{-\langle \xi,x\rangle} d\xi$$

$$\text{and} \quad \left\langle -x^*, h\right\rangle = d_h \log\psi_\Omega(x) = -\int\limits_{\Omega^*} \langle \xi, h\rangle e^{-\langle \xi,x\rangle} d\xi / \int\limits_{\Omega^*} e^{-\langle \xi,x\rangle} d\xi \tag{12}$$

In [94–97], Misha Gromov was interested by these structures. If we set $\Phi(x) = -\log\psi_\Omega(x)$, Gromov has observed that $x^* - d\Phi(x)$ is an injection where the image

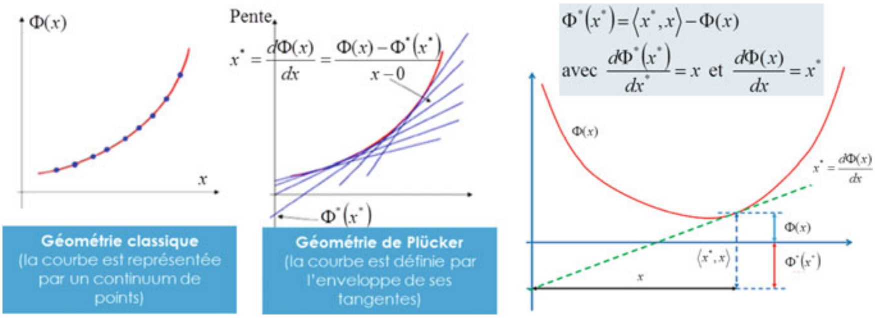

$\Phi(x)$
x
Géométrie classique
(la courbe est représentée par un continuum de points)
Pente
$x^* = \frac{d\Phi(x)}{dx} = \frac{\Phi(x) - \Phi^*(x^*)}{x-0}$
$\Phi^*(x^*)$
Géométrie de Plücker
(la courbe est définie par l'enveloppe de ses tangentes)
$\Phi^*(x^*) = \langle x^*, x \rangle - \Phi(x)$
avec $\frac{d\Phi^*(x^*)}{dx^*} = x$ et $\frac{d\Phi(x)}{dx} = x^*$
$\Phi(x)$
$x^* = \frac{d\Phi(x)}{dx}$
$\Phi(x)$
$\langle x^*, x \rangle$
$\Phi^*(x^*)$
x

Fig. 5 Legendre transform and Plücker geometry

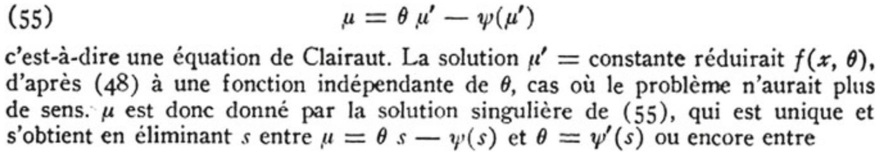
closure is equal to the convex envelop of the support and the volume of this envelop is the n-dimensionnel volume defined by the integral of hessian determinant of this function, $\Phi(x)$, where the map $\Phi \mapsto M(\Phi) = \int\limits_{\Omega} \det(Hess(\Phi(x))).dx$ obeys a non-trivial inequality given by Brunn-Minkowsky:

$$[M(\Phi_1 + \Phi_2)]^{1/2} \geq [M(\Phi_1)]^{1/n} + [M(\Phi_2)]^{1/n} \tag{13}$$

These relations appear also in statistical physics. As the physicist Jean-Marie Souriau [17, 80–84, 98] did, it is indeed possible to define the concept of Shannon's Entropy via the Lengendre transform associated with the opposite of the logarithm of this Koszul-Vinberg characteristic function. Taking up the seminal ideas of François Massieu [99–102] in Thermodynamics (classmate of the Corps des Mines, it is François Massieu who influenced Henri Poincaré [103] who introduced the characteristic function in Probability, with a Laplace transform, and not a Fourier transform as did then Paul Levy), which were recently developed by Roger Balian in Quantum Physics [19, 20, 104–111], replacing Shannon Entropy by von Neumann Entropy. I will also note the work of Jean-Leray on the extensions of the Laplace transform in [112]. Starting from the characteristic function of Koszul-Vinberg, it is thus possible to introduce an entropy of Koszul defined as the Legendre transform of this function, which is the opposite of the logarithm of the characteristic function of Koszul-Vinberg (a logarithm lies the characteristic function of Massieu and the characteristic function of Koszul or Poincaré). Starting from the Koszul function, its Legendre transform gives a dual potential function in the dual coordinate system.x^* (Fig. 5):

$$\Phi^*(x^*) = \langle x, x^* \rangle - \Phi(x) \text{ with } x^* = D_x \Phi \text{ and } x = D_{x^*} \Phi^* \text{ where } \Phi(x) = -\log \psi_\Omega(x) \tag{14}$$

Concerning the Legendre transform [113], Darboux gives in his book an interpretation of Chasles: "*What comes back according to a remark of M. Chasles, to replace the surface with its polar reciprocal with respect to a paraboloid*". We have the same reference to polar reciprocal in "*Lessons on the calculus of variations*" by Jacques



(55) $$\mu = \theta\,\mu' - \psi(\mu')$$

c'est-à-dire une équation de Clairaut. La solution $\mu' =$ constante réduirait $f(x, \theta)$, d'après (48) à une fonction indépendante de θ, cas où le problème n'aurait plus de sens. μ est donc donné par la solution singulière de (55), qui est unique et s'obtient en éliminant s entre $\mu = \theta\, s - \psi(s)$ et $\theta = \psi'(s)$ ou encore entre

Fig. 6 Legendre-Clairaut equation in 1943 Fréchet's paper

Hadamard, written by Maurice Fréchet (student of Hadamard), with references to M.E. Vessiot, which uses the "*figuratrice*", as polar reciprocal of the "*figurative*".

It is possible to express this Legendre transform only from the dual coordinate system x^*, using that $x = D_{x^*}\Phi^*$. We then obtain the Clairaut equation:

$$\Phi^*(x^*) - \left\langle (D_x\Phi)^{-1}(x^*), x^* \right\rangle - \Phi\left[(D_x\Phi)^{-1}(x^*)\right] \forall x^* \in \{D_x\Phi(x)/x \in \Omega\} \quad (15)$$

This equation was discovered by Maurice Fréchet in his 1943 paper [3] (see also in the appendix), in which he introduced for the first time the bound on the variance of any statistical estimator via the Fisher matrix, wrongly attributed to Cramer and Rao [4]. Fréchet was looking for "*distinguished densities*" [98], densities whose covariance matrix of the estimator of these parameters reaches this bound. Fréchet there showed that these densities were expressed while using this characteristic function $\Phi(x)$, and that these densities belong to the exponential densities family (Fig. 6).

Apparently, this discovery by Fréchet dates from winter of 1939, because Fréchet writes at the bottom of the page [3] "*The content of this dissertation formed part of our mathematical statistics Lecture at the Institut Henri Poincaré during the winter of 1939–1940. It is one of the chapters of the second edition (in preparation) of our ' Lessons in Mathematical Statistics', the first of which is 'Introduction: Preliminary Lecture on the Probability Calculation' (119 pages in quarto, typed in) has just been published at the University Documentation Center, Tournaments and Constans. Paris*". More details are given in appendix.

More recently Muriel Casalis [114, 115], the PhD student of Gérard Letac [116], has studied in her PhD, invariance of probability densities with respect to the affine group, and the links with densities of exponential families.

To make the link between the characteristic function of Koszul-Vinberg and Entropy of Shannon, we will detail the formulas of Koszul in the following developments. Using the fact that $-\langle \xi, x \rangle = \log e^{-\langle \xi, x \rangle}$, we can write:

$$-\left\langle x^*, x \right\rangle = \int\limits_{\Omega^*} \log\, e^{-\langle \xi,x \rangle}.e^{-\langle \xi,x \rangle} d\xi \Big/ \int\limits_{\Omega^*} e^{-\langle \xi,x \rangle} d\xi \quad (16)$$

and then developing the Legendre transform to make appear the density of maximum entropy in $\Phi^*(x^*)$, and also the Shannon entropy:

$$\Phi^*(x^*) = \langle x, x^*\rangle - \Phi(x) = -\int_{\Omega^*} \log e^{-\langle \xi,x\rangle}.e^{-\langle\xi,x\rangle}d\xi \Big/ \int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi + \log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi$$

$$\Phi^*(x^*) = \left[\left(\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi\right).\log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi - \int_{\Omega^*}\log e^{-\langle\xi,x\rangle}.e^{-\langle\xi,x\rangle}d\xi\right] \Big/ \int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi$$

$$\Phi^*(x^*) = \left[\log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi - \int_{\Omega^*}\log e^{-\langle\xi,x\rangle}.\frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}d\xi\right]$$

$$\Phi^*(x^*) = \left[\log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi.\left(\int_{\Omega^*}\frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}d\xi\right) - \int_{\Omega^*}\log e^{-\langle\xi,x\rangle}.\frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}d\xi\right]$$

with $\displaystyle\int_{\Omega^*}\frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}d\xi = 1$

$$\Phi^*(x^*) = \left[-\int_{\Omega^*}\frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}.\log\left(\frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}\right)d\xi\right] \tag{17}$$

In this last equation, $p_x(\xi) = e^{-\langle\xi,x\rangle}/\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi$ plays the role of maximum entropy density as introduced by Jaynes [117–119] (also called, Gibbs density in Thermodynamics). I call the associated entropy, Koszul Entropy:

$$\Phi^* = -\int_{\Omega^*} p_x(\xi)\log\, p_x(\xi)d\xi \tag{18}$$

with

$$p_x(\xi) = e^{-\langle\xi,x\rangle}\Big/\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi = e^{-\langle x,\xi\rangle-\log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi} = e^{-\langle x,\xi\rangle+\Phi(x)}\text{and}\, x^* = \int_{\Omega^*}\xi.p_x(\xi)d\xi \tag{19}$$

This Koszul density $p_x(\xi) = \frac{e^{-\langle\xi,x\rangle}}{\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi}$ help us to develop the log likelihood:

$$\log\, p_x(\xi) = -\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle\xi,x\rangle}d\xi = -\langle x,\xi\rangle + \Phi(x) \tag{20}$$

and deduce from the expectation:

$$E_\xi\left[-\log p_x(\xi)\right] = \langle x, x^*\rangle - \Phi(x) \tag{21}$$

We also obtain the equation about normalization:

$$\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi = -\log \int_{\Omega^*} e^{-[\Phi^*(\xi)+\Phi(x)]} d\xi = \Phi(x) - \log \int_{\Omega^*} e^{-\Phi^*(\xi)} d\xi$$
$$\Rightarrow \int_{\Omega^*} e^{-\Phi^*(\xi)} d\xi = 1 \tag{22}$$

But we have to make appear the variable x^* in $\Phi^*(x^*)$. We have then to write:

$$\log\ p_x(\xi) = \log\ e^{-\langle x,\xi\rangle+\Phi(x)} = \log e^{-\Phi^*(\xi)} = -\Phi^*(\xi)$$
$$\Rightarrow \Phi^* = -\int_{\Omega^*} p_x(\xi)\log p_x(\xi) d\xi = \int_{\Omega^*} \Omega^*(\xi) p_x(\xi) d\xi = \Phi^*(x^*) \tag{23}$$

Last equality is true, if we have:

$$\int_{\Omega^*} \Phi^*(\xi) p_x(\xi) d\xi - \Phi^*\left(\int_{\Omega^*} \xi . p_x(\xi) d\xi\right) \text{ with } x^* = \int_{\Omega^*} \xi . p_x(\xi) d\xi \tag{24}$$

This last relation is associated to classical Jensen inequality. Equality is obtained for Maximum Entropy density for $x^* = D_x \Phi$ [120]:

$$\text{Legendre - Moreau Transform:} \quad \Phi^*(x^*) = \underset{x}{Sup}\left[\langle x, x^*\rangle - \Phi(x)\right]$$
$$\Rightarrow \begin{cases} \Phi^*(x^*) \geq \langle x, x^*\rangle - \Phi(x) \\ \Phi^*(x^*) \geq \int_{\Omega^*} \Phi^*(\xi) p_x(\xi) d\xi \end{cases} \Rightarrow \begin{cases} \Phi^*(x^*) \geq E\left[\Phi^*(\xi)\right] \\ \text{equality if } x^* = \frac{d\Phi}{dx} \end{cases} \tag{25}$$

We obtain for the maximum entropy density, the equality:

$$E\left[\Phi^*(\xi)\right] = \Phi^*(E[\xi]), \xi \in \Omega^* \tag{26}$$

To make the link between this Koszul model and maximum entropy density [121–123] introduced by Jaynes [117–119], I use previous notation and I look for the density $p_x(\xi)$ that is the solution to this maximum entropy variational problem. Find the density that maximizes the Shannon entropy with constraint on normalization and on the knowledge of first moment:

$$\underset{p_x(.)}{Max}\left[-\int_{\Omega^*} p_x(\xi)\log p_x(\xi) d\xi\right] \text{ such that } \begin{cases} \int_{\Omega^*} p_x(\xi) d\xi = 1 \\ \int_{\Omega^*} \xi . p_x(\xi) d\xi = x^* \end{cases} \tag{27}$$

If we consider the density $q_x(\xi) = e^{-\langle \xi, x\rangle} / \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi = e^{-\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi}$ such that:

$$\begin{cases} \int\limits_{\Omega^*} q_x(\xi).d\xi = \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi / \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi = 1 \\ \log q_x(\xi) = \log e^{-\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi} = -\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle x, \xi\rangle} d\xi \end{cases} \tag{28}$$

By using the inequality $\log x \geq \left(1 - x^{-1}\right)$ with equality if $x = 1$, we can then write that:

$$-\int\limits_{\Omega^*} p_x(\xi) \log \frac{p_x(\xi)}{q_x(\xi)} d\xi \leq -\int\limits_{\Omega^*} p_x(\xi) \left(1 - \frac{q_x(\xi)}{p_x(\xi)}\right) d\xi \tag{29}$$

We develop the right term of the equation:

$$\int\limits_{\Omega^*} p_x(\xi) \left(1 - \frac{q_x(\xi)}{p_x(\xi)}\right) d\xi = \int\limits_{\Omega^*} p_x(\xi) d\xi - \int\limits_{\Omega^*} q_x(\xi) d\xi = 0 \tag{30}$$

knowing that $\int\limits_{\Omega^*} p_x(\xi) d\xi = \int\limits_{\Omega^*} q_x(\xi) d\xi = 1$, we can deduce that:

$$-\int\limits_{\Omega^*} p_x(\xi) \log \frac{p_x(\xi)}{q_x(\xi)} d\xi \leq 0 \Rightarrow -\int\limits_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \leq -\int\limits_{\Omega^*} p_x(\xi) \log q_x(\xi) d\xi \tag{31}$$

We have then to develop the right term by using previous expression of $q_x(\xi)$:

$$-\int\limits_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \leq -\int\limits_{\Omega^*} p_x(\xi) \left[-\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle x, \xi\rangle} d\xi \right] d\xi \tag{32}$$

$$-\int\limits_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \leq \left\langle x, \int\limits_{\Omega^*} \xi . p_x(\xi) d\xi \right\rangle + \log \int\limits_{\Omega^*} e^{-\langle x, \xi\rangle} d\xi \tag{33}$$

If we use that $x^* = \int\limits_{\Omega^*} \xi . p_x(\xi) d\xi$ and $\Phi(x) = -\log \int\limits_{\Omega^*} e^{-\langle x, \xi\rangle} d\xi$, then we obtain that the density $q_x(\xi) = e^{-\langle \xi, x\rangle} / \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi = e^{-\langle x, \xi\rangle - \log \int\limits_{\Omega^*} e^{-\langle \xi, x\rangle} d\xi}$ is the maximum entropy density constrained by $\int\limits_{\Omega^*} p_x(\xi) d\xi$ and $\int\limits_{\Omega^*} \xi . p_x(\xi) d\xi = x^*$:

$$-\int\limits_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \leq \left\langle x, x^* \right\rangle - \Phi(x) \tag{34}$$

$$-\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \le \Phi^*(x^*) \tag{35}$$

In the following, we will write $x^* = \hat{\xi}$, to give to this variable the link with momentum $\hat{\xi} = \int\limits_{\Omega^*} \xi.p_{\hat{\xi}}(\xi)d\xi$. To express the density with respect to the 1st moment as variable, we have to inverse $\hat{\xi} = \Theta(x) = \frac{d\Phi(x)}{dx}$, by writting $x = \Theta^{-1}(\hat{\xi})$ the inverse function (given by Legendre transform):

$$p_{\hat{\xi}}(\xi) = \frac{e^{-\left\langle \xi,\Theta^{-1}(\hat{\xi})\right\rangle}}{\int\limits_{\Omega^*} e^{-\left\langle \xi,\Theta^{-1}(\hat{\xi})\right\rangle}d\xi} \text{ with } \hat{\xi} = \int_{\Omega^*} \xi.p_{\hat{\xi}}(\xi)d\xi \text{ and } \Phi(x) = -\log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{36}$$

We find finally the Maximum entropy density parametrized by 1st moment $\hat{\xi}$.

4 Links Between Koszul-Vinberg Characteristic Function, Koszul Forms and Information Geometry

Koszul Hessian Geometry Structure is the key tool to define elementary structures of Information Geometry, that appears as one particular case of more general framework studied by Koszul. In the Koszul-Vinberg Characteristic function $\psi_\Omega(x) = \int\limits_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi$, $\forall x \in \Omega$ where Ω is a sharp convex cone and Ω^* its dual cone, the duality bracket < .,. > has to be defined. I will introduce it by using Cartan-Killing form $\langle x,y\rangle = -B(x,\theta(y))$ with $B(.,.)$ killing form and $\theta(.)$ Cartan involution. The inner product is then invariant with respect to automorphims of cone Ω. Koszul-Vinberg characteristic function could be developed as [124]:

$$\psi_\Omega(x+\lambda u) = \psi_\Omega(x) - \lambda\left\langle x^* + u\right\rangle + \frac{\lambda^2}{2}\langle K(x)u,u\rangle + \ldots \tag{37}$$

with $x^* = \frac{d\Phi(x)}{dx}$, $\Phi(x) = -\log\psi_\Omega(x)$ and $K(\mathrm{x}) = \frac{d^2\Phi(x)}{dx^2}$

In the following developments, I will write β, previous variable written x, because in thermodynamics, this variable corresponds to the Planck temperature, classically $\beta = \frac{1}{T}$. The variable β will be the dual variable of $\hat{\xi}$.

$$
\begin{aligned}
& p_{\hat{\xi}}(\xi) = \frac{e^{-\left\langle \Theta^{-1}(\hat{\xi}),\xi \right\rangle}}{\int\limits_{\Omega^*} e^{-\left\langle \Theta^{-1}(\hat{\xi}),\xi \right\rangle}.d\xi} \hat{\xi} = \Theta(\beta) = \frac{\partial \Phi(\beta)}{\partial \beta} \text{ with } \Phi(\beta) = -\log \psi_\Omega(\beta) \\
& \psi_\Omega(\beta) = \int\limits_{\Omega^*} e^{-\langle \beta,\xi \rangle} d\xi, \quad S(\hat{\xi}) = -\int\limits_{\Omega^*} p_{\hat{\xi}}(\xi) \log p_{\hat{\xi}}(\xi).d\xi \text{ and } \beta = \Theta^{-1}(\hat{\xi}) \\
& S(\hat{\xi}) = \left\langle \hat{\xi}, \beta \right\rangle - \Phi(\beta)
\end{aligned}
\tag{38}
$$

Inversion of the function $\Theta(.)$ is given by $\beta = \Theta^{-1}(\hat{\xi})$ is achieved by Legendre transform using relation between Entropy $S(\hat{\xi})$ and the function $\Phi(\beta)$ (opposite of the logarithm of the Koszul-Vinberg characteristic function):

$$
\begin{aligned}
& S(\hat{\xi}) = \left\langle \beta, \hat{\xi} \right\rangle - \Phi(\beta) \\
& \text{with} \quad \Phi(\beta) = -\log \int\limits_{\Omega^*} e^{-\langle \xi,\beta \rangle} d\xi \quad \forall \beta \in \Omega \quad \text{and} \quad \forall \xi, \hat{\xi} \in \Omega^*
\end{aligned}
\tag{39}
$$

We will prove that the 2nd Koszul form $-\frac{\partial^2 \Phi(\beta)}{\partial \beta^2}$ is linked with Fisher Metric of Information Geometry:

$$
I(\beta) = -E\left[\frac{\partial^2 \log p_\beta(\xi)}{\partial \beta^2} \right]
\tag{40}
$$

To compute the Fisher metric $I(\beta)$, we use the following relations between variable

$$
\begin{aligned}
& \begin{cases}
\log p_{\hat{\xi}}(\xi) = -\langle \xi, \beta \rangle + \Phi(\beta) \\
S(\widehat{\xi}) = -\int\limits_{\Omega^*} p_{\hat{\xi}}(\xi). \log p_{\hat{\xi}}(\xi).d\xi = -E\left[\log p_{\hat{\xi}}(\xi) \right]
\end{cases} \\
& \Rightarrow S(\widehat{\xi}) = \langle E[\xi], \beta \rangle - \Phi(\beta) = \left\langle \hat{\xi}, \beta \right\rangle - \Phi(\beta)
\end{aligned}
\tag{41}
$$

We can observe that the logarithm of the density is affine with respect to the variable β, and that the Fisher matrix is given by the hessian. We can then deduce that the Fisher Metric is given by the hessian.

$$
I(\beta) = -E\left[\frac{\partial^2 \log p_\beta(\xi)}{\partial \beta^2} \right] = -E\left[\frac{\partial^2 (-\langle \xi, \beta \rangle + \Phi(\beta))}{\partial \beta^2} \right] = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2} = \frac{\partial^2 \log \psi_\Omega(\beta)}{\partial \beta^2}
\tag{42}
$$

We can also identify the Fisher metric as a variance:

$$
\log \Psi_\Omega(\beta) = \log \int\limits_{\Omega^*} e^{-\langle \xi,\beta \rangle} d\xi \Rightarrow \frac{\partial \log \Psi_\Omega(\beta)}{\partial \beta} = -\frac{1}{\int\limits_{\Omega^*} e^{-\langle \xi,\beta \rangle} d\xi} \int\limits_{\Omega^*} \xi.e^{-\langle \xi,\beta \rangle} d\xi
\tag{43}
$$

$$\frac{\partial^2 \log \Psi_\Omega(\beta)}{\partial \beta^2} = -\frac{1}{\left(\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle} d\xi\right)^2}\left[-\int\limits_{\Omega^*} \xi^2.e^{-\langle \xi,\beta\rangle} d\xi.\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle} d\xi + \left(\int\limits_{\Omega^*} \xi^2.e^{-\langle \xi,\beta\rangle} d\xi\right)^2\right] \quad (44)$$

$$\frac{\partial^2 \log \Psi_\Omega(\beta)}{\partial \beta^2} = \int\limits_{\Omega^*} \xi^2.\frac{e^{-\langle \xi,\beta\rangle}}{\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle} d\xi} d\xi - \left(\int\limits_{\Omega^*} \xi.\frac{e^{-\langle \xi,\beta\rangle}}{\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle} d\xi} d\xi\right)^2$$

$$= \int\limits_{\Omega^*} \xi^2.p_\beta(\xi) d\xi - \left(\int\limits_{\Omega^*} \xi.p_\beta(\xi) d\xi\right)^2 \quad (45)$$

$$I(\beta) = -E_\xi\left[\frac{\partial^2 \log p_\beta(\xi)}{\partial \beta^2}\right] = \frac{\partial^2 \log \psi_\Omega(\beta)}{\partial \beta^2} = E_\xi\left[\xi^2\right] - E_\xi\left[\xi^2\right] = Var(\xi) \quad (46)$$

In 1977, Crouzeix [125, 126] has identified the following relation between both hessian of entropy and characteristic function $\frac{\partial^2 \Phi}{\partial \beta^2} = \left[\frac{\partial^2 S}{\partial \hat{\xi}^2}\right]^{-1}$ giving a relation between the dual metrics with respect to their dual coordinate systems. The metric could be given by Fisher metric or given by the hessian of Entropy *S:*

$$ds_g^2 = d\beta^T I(\beta) d\beta = \sum_{ij} g_{ij} d\beta_i d\beta_j \quad \text{with} \quad g_{ij} = [I(\beta)]_{ij} \quad (47)$$

Thanks to Crouzeix relation [125] [126], we observe that 2 geodesic distances given by hessian of dual potential functions in dual coordinates systems, are equal:

$$ds_h^2 = d\hat{\xi}^T \left[\frac{\partial^2 S(\hat{\xi})}{\partial \hat{\xi}^2}\right] d\hat{\xi} = \sum_{ij} h_{ij} d\hat{\xi}_i d\hat{\xi}_j \quad \text{with} \quad h_{ij} = \left[\frac{\partial^2 S(\hat{\xi})}{\partial \hat{\xi}^2}\right]_{ij} \quad (48)$$

$$ds_h^2 = ds_g^2 \quad (49)$$

One can ask oneself the question of what is the most natural product of duality. This question has been treated by Elie Cartan in his thesis in 1894, by introducing a form called Cartan-Killing form, a symmetric bilinear form naturally associated with any Lie algebra. This form of Cartan-Killing is defined via the endomorphism ad_x of Lie algebra g via the Lie bracket:

$$ad_x(y) = [x, y] \quad (50)$$

The trace of the composition of these 2 endomorphisms defines this bilinear form by:

$$B(x, y) = Tr\left(ad_x ad_y\right) \quad (51)$$

The Cartan-Killing form is symmetric:

$$B(x, y) = B(y, x) \quad (52)$$

and verify associativity property:

$$B([x, y], z) = B(x, [y, z]) \tag{53}$$

given by:

$$\begin{aligned} B([x, y], z) &= Tr\left(ad_{[x,y]}ad_z\right) = Tr\left(\left[ad_x, ad_y\right]ad_z\right) \\ &= Tr\left(ad_x\left[ad_y, ad_z\right]\right) = B(x, [y, z]) \end{aligned} \tag{54}$$

Elie Cartan proved that if g is a semi-simple Lie algebra (the form of Killing is non-degenerate) then any symmetric bilinear form is a scalar multiple of the Cartan-Killing form. The Cartan-Killing form is invariant under the action of automorphisms $\sigma \in Aut(g)$ of the algebra g:

$$B(\sigma(x), \sigma(y)) = B(x, y) \tag{55}$$

This invariance is deduced from:

$$\begin{gathered} \begin{cases} \sigma[x, y] = [\sigma(x), \sigma(y)] \\ z = \sigma(y) \end{cases} \Rightarrow \sigma\left[x, \sigma^{-1}(z)\right] = [\sigma(x), z] \\ \text{by writting } \; ad_{\sigma(x)} = \sigma \circ ad_x \circ \sigma^{-1} \end{gathered} \tag{56}$$

Then, we can write:

$$B(\sigma(x), \sigma(y)) = Tr\left(ad_{\sigma(x)}ad_{\sigma(y)}\right) = Tr\left(\sigma \circ ad_x ad_y \circ \sigma^{-1}\right) = Tr\left(ad_x ad_y\right) = B(a, y) \tag{57}$$

Cartan has introduced this natural inner product that is invariant by the automorphisms of the Lie algebra, from this Cartan-Killing form:

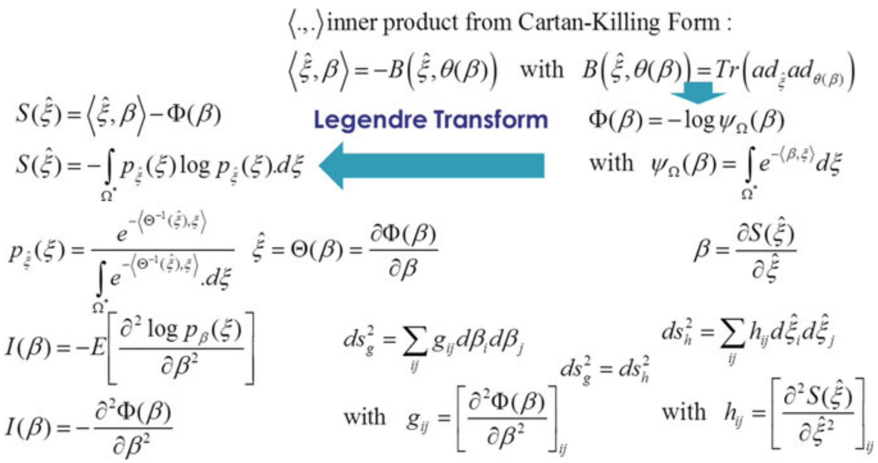
$$\langle x, y \rangle = -B(x, \theta(y)) \tag{58}$$

with $\theta \in g$ the Cartan involution (an involution on the Lie algebra g is an automorphism θ such that the square is equal to identity).

I summarize all these relations of information geometry from the characteristic function of Koszul-Vinberg, and the duality given via the Cartan-Killing form, as described in the figure below (Fig. 7):

Thanks to the expression of the characteristic function of Koszul-Vinberg and the Cartan-Killing form, one can express the maximum Entropy density in a very general way. For example, by applying these formulas to the cone Ω (self-dual: $\Omega^* = \Omega$) symmetric positive definite matrices $Sym^+(n)$, Cartan-Killing form gives us the product of duality:

$$\langle \eta, \xi \rangle = Tr(\eta^T \xi). \quad \forall \eta, \xi \in Sym^+(n) = \left\{\xi / \xi^T = \xi, \xi > 0\right\} \tag{59}$$



$\langle .,. \rangle$ inner product from Cartan-Killing Form :

$\langle \hat{\xi}, \beta \rangle = -B\left(\hat{\xi}, \theta(\beta)\right)$ with $B\left(\hat{\xi}, \theta(\beta)\right) = Tr\left(ad_{\hat{\xi}} ad_{\theta(\beta)}\right)$

$S(\hat{\xi}) = \langle \hat{\xi}, \beta \rangle - \Phi(\beta)$

Legendre Transform

$\Phi(\beta) = -\log \psi_\Omega(\beta)$

$S(\hat{\xi}) = -\int_{\Omega^*} p_{\hat{\xi}}(\xi) \log p_{\hat{\xi}}(\xi).d\xi$

with $\psi_\Omega(\beta) = \int_{\Omega^*} e^{-\langle \beta, \xi \rangle} d\xi$

$p_{\hat{\xi}}(\xi) = \dfrac{e^{-\langle \Theta^{-1}(\hat{\xi}), \xi \rangle}}{\int_{\Omega^*} e^{-\langle \Theta^{-1}(\hat{\xi}), \xi \rangle}.d\xi}$ $\quad \hat{\xi} = \Theta(\beta) = \dfrac{\partial \Phi(\beta)}{\partial \beta}$ $\quad \beta = \dfrac{\partial S(\hat{\xi})}{\partial \hat{\xi}}$

$I(\beta) = -E\left[\dfrac{\partial^2 \log p_\beta(\xi)}{\partial \beta^2}\right]$ $\quad ds_g^2 = \sum_{ij} g_{ij} d\beta_i d\beta_j$ $\quad ds_g^2 = ds_h^2$ $\quad ds_h^2 = \sum_{ij} h_{ij} d\hat{\xi}_i d\hat{\xi}_j$

$I(\beta) = -\dfrac{\partial^2 \Phi(\beta)}{\partial \beta^2}$ $\quad$ with $g_{ij} = \left[\dfrac{\partial^2 \Phi(\beta)}{\partial \beta^2}\right]_{ij}$ $\quad$ with $h_{ij} = \left[\dfrac{\partial^2 S(\hat{\xi})}{\partial \hat{\xi}^2}\right]_{ij}$

Fig. 7 Relations between cartan-killing form, koszul-vinberg characteristic function, potentials and dual coordinates, and metrics of information geometry

The maximum entropy density is given by:

$$\psi_\Omega(\beta) = \int_{\Omega^*} e^{-\langle \beta, \xi \rangle} d\xi = \det(\beta)^{-\frac{n+1}{2}} \psi_\Omega(I_d)$$
$$\text{and} \quad \hat{\xi} = \frac{\partial \Phi(\beta)}{\partial \beta} = \frac{\partial(-\log \psi_\Omega(\beta))}{\partial \beta} = \frac{n+1}{2}\beta^{-1} \tag{60}$$

From which, I can deduce the final expression:

$$p_{\hat{\xi}}(\xi) = e^{-\left\langle \Theta^{-1}(\hat{\xi}), \xi \right\rangle + \Phi\left(\Theta^{-1}(\hat{\xi})\right)} = \psi_\Omega(I_d).\left[\det\left(\alpha \hat{\xi}^{-1}\right)\right].e^{-Tr\left(\alpha \hat{\xi}^{-1} \xi\right)}$$
$$\text{with} \quad \alpha = \frac{n+1}{2} \tag{61}$$

We can apply this approach for multivariate Gaussian densities. In the case of multivariate Gaussian densities, as noted by Souriau [17, 79], the classical Gibbs expression can be rewritten by modifying the coordinate system and defining a new duality product [80–84, 98]. The multivariate Gaussian density is classically written with the following coordinate system (m, R), with m the mean vector, and R the covariance matrix of the vector z:

$$p_{\hat{\xi}}(\xi) = \frac{1}{(2\pi)^{n/2} \det(R)^{1/2}} e^{-\frac{1}{2}(z-m)^T R^{-1}(z-m)} \text{ with } \begin{cases} m = E(z) \\ R = E\left[(z-m)(z-m)^T\right] \end{cases} \tag{62}$$

By developing the term in the exponential:

$$\begin{aligned}\frac{1}{2}(z-m)^TR^{-1}(z-m)&=\frac{1}{2}\left[z^TR^{-1}z-m^TR^{-1}z-z^TR^{-1}m+m^TR^{-1}m\right]\\&=\frac{1}{2}z^TR^{-1}z-m^TR^{-1}z+\frac{1}{2}m^TR^{-1}m\end{aligned}\tag{63}$$

I can write this density as a Gibbs density by introducing a new duality bracket between $\left(z,zz^T\right)$ and $\left(-R^{-1}m,\frac{1}{2}R^{-1}\right)$:

$$\begin{aligned}&p_{\hat{\xi}}(\xi)=\frac{1}{(2\pi)^{n/2}\det(R)^{1/2}e^{\frac{1}{2}m^TR^{-1}m}}e^{-\left[-m^TR^{-1}z+\frac{1}{2}z^TR^{-1}z\right]}=\frac{1}{Z}e^{-\langle\xi,\beta\rangle}\\&\xi=\begin{bmatrix}z\\zz^T\end{bmatrix}\text{ and }\beta=\begin{bmatrix}-R^{-1}m\\\frac{1}{2}R^{-1}\end{bmatrix}=\begin{bmatrix}a\\H\end{bmatrix}\\&\text{with }\ \langle\xi,\beta\rangle=a^Tz+z^THz=Tr\left[za^T+H^Tzz^T\right]\end{aligned}\tag{64}$$

We can then write the density in Koszul form:

$$\begin{aligned}&p_{\hat{\xi}}(\xi)=\frac{1}{\int\limits_{\Omega^*}e^{-\langle\xi,\beta\rangle}.d\xi}e^{-\langle\xi,\beta\rangle}=\frac{1}{Z}e^{-\langle\xi,\beta\rangle}\\&\text{with }\ \log(Z)=n\log(2\pi)+\frac{1}{2}\log\det(R)+\frac{1}{2}m^TR^{-1}m\\&\xi=\begin{bmatrix}z\\zz^T\end{bmatrix},\hat{\xi}=E[\xi]=\begin{bmatrix}E[z]\\E\left[zz^T\right]\end{bmatrix}=\begin{bmatrix}m\\R+mm^T\end{bmatrix},\beta=\begin{bmatrix}a\\H\end{bmatrix}=\begin{bmatrix}-R^{-1}m\\\frac{1}{2}R^{-1}\end{bmatrix}\\&\text{with }\langle\xi,\beta\rangle=Tr\left[za^T+H^Tzz^T\right]\\&R=E\left[(z-m)(z-m)^T\right]=E\left[zz^T-mz^T-zm^T+mm^T\right]=E\left[zz^T\right]-mm^T\end{aligned}\tag{65}$$

We are then able to compute the Koszul-Vinberg characteristic function whose opposite of the logarithm provides the potential function:

$$\begin{aligned}&\psi_\Omega(\beta)=\int\limits_{\Omega^*}e^{-\langle\xi,\beta\rangle}.d\xi\\&\text{and }\ \Phi(\beta)=-\log\psi_\Omega(\beta)=\frac{1}{2}\left[-Tr\left[H^{-1}aa^T\right]+\log\left[(2)^n\det H\right]-n\log(2\pi)\right]\end{aligned}\tag{66}$$

that verifies the following relation given by Koszul and linked with 1st Koszul form:

$$\frac{\partial \Phi(\beta)}{\partial \beta} = \frac{\partial\left[-\log \psi_\Omega(\beta)\right]}{\partial \beta} = \int\limits_{\Omega^*} \xi \frac{e^{-\langle \xi,\beta\rangle}}{\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle}.d\xi} = \int\limits_{\Omega^*} \xi.p_{\hat{\xi}}(\xi).d\xi = \hat{\xi}$$

$$\frac{\partial \Phi(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial \Phi(\beta)}{\partial \alpha} \\ \frac{\partial \Phi(\beta)}{\partial H} \end{bmatrix} = \begin{bmatrix} m \\ R + mm^T \end{bmatrix} = \hat{\xi} \tag{67}$$

The 2nd dual potential is given by the Legendre transform of $\Phi(\beta)$:

$$S(\hat{\xi}) = \left\langle \hat{\xi}, \beta \right\rangle - \Phi(\beta) \quad \text{with} \quad \frac{\partial \Phi(\beta)}{\partial \beta} = \hat{\xi} \quad \text{and} \quad \frac{\partial S(\hat{\xi})}{\partial \hat{\xi}} = \beta$$

$$S\left(\hat{\xi}\right) = -\int\limits_{\Omega^*} \frac{e^{-\langle \xi,\beta\rangle}}{\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle}.d\xi} \log \frac{e^{-\langle \xi,\beta\rangle}}{\int\limits_{\Omega^*} e^{-\langle \xi,\beta\rangle}.d\xi}.d\xi = -\int\limits_{\Omega^*} p_{\hat{\xi}}(\xi) \log p_{\hat{\xi}}(\xi).d\xi \tag{68}$$

that is explicitly identified with the classical Shannon Entropy:

$$\begin{aligned} S(\hat{\xi}) &= -\int\limits_{\Omega^*} p_{\hat{\xi}}(\xi) \log p_{\hat{\xi}}(\xi).d\xi \\ &= \frac{1}{2}\left[\log(2)^n \det\left[H^{-1}\right] + n \log(2\pi.e)\right] = \frac{1}{2}\left[\log \det[R] + n \log(2\pi.e)\right] \end{aligned} \tag{69}$$

The Fisher metric of Information Geometry is given by the hessian of the opposite of the logarithm of the Koszul-Vinberg characteristic function:

$$\begin{aligned} & ds_g^2 = d\beta^T I(\beta) d\beta = \sum_{ij} g_{ij} d\beta_i d\beta_j \\ & \text{with} \quad g_{ij} = [I(\beta)]_{ij} \text{ and } I(\beta) = -E_\xi\left[\frac{\partial^2 \log p_\beta(\xi)}{\partial \beta^2}\right] = \frac{\partial^2 \log \psi_\Omega(\beta)}{\partial \beta^2} \end{aligned} \tag{70}$$

Then, for the multivariate Gaussian density, we have the following Fisher metric:

$$ds^2 = \sum_{ij} g_{ij} d\theta_i d\theta_j = dm^T R^{-1} dm + \frac{1}{2} Tr\left[\left(R^{-1} dR\right)^2\right] \tag{71}$$

Geodesic equations are given by Euler-Lagrange equations:

$$\begin{aligned} & \sum_{i=1}^n g_{ik} \ddot{\theta}_i + \sum_{i,j=1}^n \Gamma_{ijk} \dot{\theta}_i \dot{\theta}_j = 0, \ k = 1, \ldots, n \\ & \text{with} \quad \mathit{\Gamma}_{ijk} = \frac{1}{2}\left[\frac{\partial g_{jk}}{\partial \theta_i} + \frac{\partial g_{jk}}{\partial \theta_j} + \frac{\partial g_{ij}}{\partial \theta_k}\right] \end{aligned} \tag{72}$$

that can be reduced to the equations:

$$\begin{cases} \ddot{R} + \dot{m}\dot{m}^T - \dot{R}R^{-1}\dot{R} = 0 \\ \ddot{m} - \dot{R}R^{-1}\dot{m} = 0 \end{cases} \tag{73}$$

I use a result of Souriau [17] that the component of « moment map » are constants (geometrization of Emmy Noether theorem), to identify the following constants [83]:

$$\frac{d\Pi_R}{dt} = \begin{bmatrix} \frac{d\left(R^{-1}\dot{R}+R^{-1}\dot{m}m^T\right)}{dt} & \frac{d\left(R^{-1}\dot{m}\right)}{dt} \\ 0 & 0 \end{bmatrix} = 0$$

$$\Rightarrow \begin{cases} R^{-1}\dot{R} + R^{-1}\dot{m}m^T = B = cste \\ R^{-1}\dot{m} = b = cste \end{cases} \tag{74}$$

with Π_R the moment map introduced by Souriau [17]. This moment map could be computed if we consider the following Lie group acting in case of Gaussian densities:

$$\begin{bmatrix} Y \\ 1 \end{bmatrix} = \begin{bmatrix} R^{1/2} & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ 1 \end{bmatrix} = \begin{bmatrix} R^{1/2}X + m \\ 1 \end{bmatrix}, \begin{cases} (m, R) \in R^n \times Sym^+(n) \\ M = \begin{bmatrix} R^{1/2} & m \\ 0 & 1 \end{bmatrix} \in G_{aff} \end{cases}$$

$$X \approx \aleph(0, I) \rightarrow Y \approx \aleph(m, R) \tag{75}$$

$R^{1/2}$, square root of R, is given by Cholesky decomposition of R. $R^{1/2}$ is the Lie group of triangular matrix with positive elements on the diagonal. Euler-Poincaré equations, reduced equations from Euler-Lagrange equations, are then given by:

$$\begin{cases} \dot{m} = Rb \\ \dot{R} = R(B - bm^T) \end{cases} \tag{76}$$

Geodesic distance between multivariate Gaussian density is then obtained by "*geodesic shooting*" method that will provide iteratively the final solution from the tangent vector at the initial point:

$$\left(R^{-1}(0)\dot{m}(0), R^{-1}(0)\left(\dot{R}(0) + \dot{m}(0)m(0)^T\right)\right) = (b, B) \in R^n \times Sym^+(n) \tag{77}$$

From which, we then deduce the distance:

$$d = \sqrt{\dot{m}(0)^T R^{-1}(0)\dot{m}(0) + \frac{1}{2}Tr\left[\left(R^{-1}(0)\dot{R}(0)\right)^2\right]} \tag{78}$$

Geodesic shooting is obtained by using equations established by Eriksen [127, 128] for "*exponential map*" using the following change of variables:

$$\begin{cases} \Delta(t) = R^{-1}(t) \\ \delta(t) = R^{-1}(t)m(t) \end{cases} \Rightarrow \begin{cases} \dot{\Delta} = -B\Delta + bm^T \\ \dot{\delta} = -B\delta + \left(1 + \delta^T \Delta^{-1}\delta\right)b \\ \Delta(0) = I_p, \delta(0) = 0 \end{cases} \quad \text{with} \quad \begin{cases} \dot{\Delta}(0) = -B \\ \dot{\delta}(0) = b \end{cases} \tag{79}$$

The method based on geodesic shooting consists in iteratively approaching the solution by geodesic shooting in direction $\left(\dot{\delta}(0), \dot{\Delta}(0)\right)$, using the following exponential map (Fig. 8):

$$\Lambda(t) = \exp(tA) = \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} = \begin{pmatrix} \Delta & \delta & \Phi \\ \delta^T & \varepsilon & \gamma^T \\ \Phi^T & \gamma & \Gamma \end{pmatrix}$$

$$\text{with} \quad A = \begin{pmatrix} -B & b & 0 \\ b^T & 0 & -b^T \\ 0 & -b & B \end{pmatrix} \tag{80}$$

The principle of geodesic shooting is the following. We consider one geodesic χ between θ_0 and θ_1 with an initial tangent vector V from the origin, and assume that V is modified by W, with respect to $V + W$. Variation of final point θ_1 could be obtained by Jacobi vector field $J(0) = 0$ and $\dot{J}(0) = W$:

$$J(t) = \frac{d}{d\alpha} \exp_{\theta_0}(t(V + \alpha W))|_{\alpha=0} \tag{81}$$

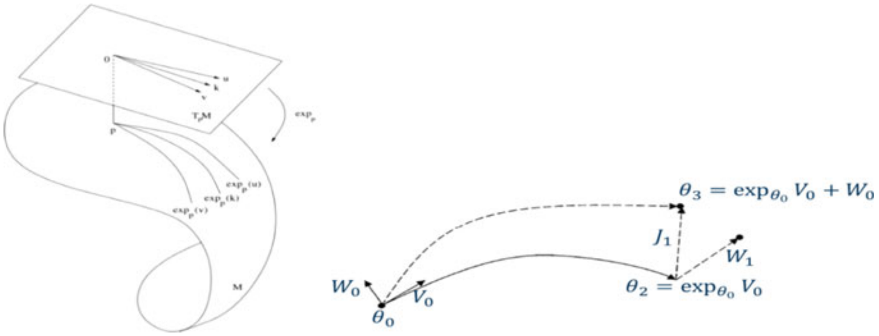

0
u
k
v
T_pM
exp_p
P
exp_p(u)
exp_p(k)
exp_p(v)
M
$\theta_3 = \exp_{\theta_0} V_0 + W_0$
J_1
W_1
W_0
V_0
$\theta_2 = \exp_{\theta_0} V_0$
θ_0

Fig. 8 Principle of geodesic shooting in the direction of the initial vector V_0 at the origin and correction by W_0

5 Koszul's Study of Homogeneous Bounded Domains and Affine Representations of Lie Groups and Lie Algebras

Jean-Louis Koszul [15, 16, 41, 45–47, 49, 50] and his student Jacques Vey [39, 40] introduced new theorems with more general extension than previous results:

Koszul theorem [50]: Let *Ω* be a sharp convex open in an affine space of E of finite dimension on R. If a unimodular Lie group of affine transformations operates transitively on Ω, Ω is a cone.

Koszul-Vey Theorem [40]: Let M a hessian connected manifold associated with the hessian metric g. Assume that M has a closed 1-form α such that $D\alpha = g$ and that there is a group G of affine automorphisms of M preserving α, then:

- If M/G is almost compact, then the manifold, universal covering of M, is affinely isomorphic to a convex domain of an affine space containing no straight line.
- If M/G is compact, then Ω is a sharp convex cone.

Jean-Louis Koszul developed his theory, studying the homogeneous domains, in particular the homogeneous symmetric bounded domains of Siegel, which we note *DS* [44, 48]. He has proved that there is a subgroup G in the group of complex affine automorphisms of these domains (Iwasawa subgroup), so that G acts on *DS* in a merely transitive way. The Lie algebra g of G has a structure which is an algebraic translation of the Kähler structure *DS*.

Koszul considered on *G/B* an invariant complex structure tensor I. All the invariant volumes on *G/B*, equal up to a constant factor, define with the complex structure the same invariant Hermitian form on *G/B*, called Hermitian canonical form, denoted h. Let E be a differentiable fiber space of base M and let p be the projection of E on M, such that $p^*((pX).f) = X.(p^*f)$. The projection $p : E \to M$ defines an injective homomorphism p^* of the space of differential forms of M in the space of the differential forms of E such that for any form α of degree n on M and any sequence of n projectable vectors fields, we have $p^*(\alpha(pX_1, pX_2, \ldots, pX_n)) = (p^*\alpha)(X_1, X_2, \ldots, X_n)$. Let I be the tensor of an almost complex structure on the basis M, there exists on E a tensor J of type *(1,1)* and only one which possesses the following properties $p(JX) = I(pX)$ and $J^2X = -X \mod h, X \in g$ for any vector field X on E. Let G be a connected Lie group and B a closed subgroup of G, we note g the Lie algebra left invariant vector fields on G and b sub-algebra of g corresponding to B. The canonical mapping of G on *G/B* is denoted p (defining E as before). We assume that there exists on *G/B* an invariant volume by G, which consist in assuming that, for all $s \in B$, the automorphism $X \to Xs$ of g defines by passing to the quotient an automorphism of determinant *1* in *g/b*. Let r be the dimension of *G/B* and $(X_i)_{1 \le i \le m}$ a base of g such that $X_i \in b$, for $r \le i \le m$. Let $(\xi_i)_{1 \le i \le m}$ the base of the space of differential forms of degree *1* left invariant on G such that $\xi_i\left(X_j\right) = \delta_{ij}$. If ω is an invariant volume on *G/B*, then $\Omega = p^*\omega$ is equal, up to a constant factor, to $\xi_1 \wedge \xi_2 \wedge \ldots \wedge \xi_r$. We will assume the base $\left(X_j\right)$ chosen so that this factor is equal to *1*, let $\Omega = \xi_1 \wedge \xi_2 \wedge \ldots \wedge \xi_r$. For any vector field that can be projected X on G, we have:

$$p^*(div(pX))\Omega = p^*((div(pX))\omega) = p^*((pX)\omega) = X\Omega = \sum_{j=1}^{r} \xi_j([X_j, X])\Omega \tag{82}$$

$$p^*(div(pX)) = \sum_{j=1}^{r} \xi_j([X_j, X]) \tag{83}$$

These elements being defined, Koszul calculates the Hermitian canonical form of *G/B*, denoted h, more particularly $\eta = p^*h$ on G. Let X and Y both right invariant vector fields on G. They are projectable and the fields pX and pY are conformal vector fields on *G/B* such that $div(pX) = div(pY) = 0$, because the volume and the complex structure of *G/B* are invariant under G. As a result, if κ is the Kähler form of h and if $\alpha = p^*\kappa$, then:

$$4\alpha(X, Y) = 4p^*(\kappa(pX, pY)) = p^*div(I[pX, pY]) \tag{84}$$

and as $p(J[X, Y]) = I[pX, pY]$, we obtain:

$$4\alpha(X, Y) = p^*div(J[X, Y]) = \sum_{i=1}^{2n} \xi_i([X_i, J[X, Y]]) \tag{85}$$

X and *Y* are two left invariant vectors fields on *G*. *X '* and *Y'* right invariant vectors fields coinciding with *X* and *Y* at the point *e*, neutral element of *G*. If $T = \left[X', Y'\right]$ is tight invariant vectors fields which coincide with $-[X, Y]$ on *e*, then:

$$[X, JT] = J[X, [X, Y]] - [X, J[X, Y]] \,\text{at point e} \tag{86}$$

At point *e*, we have the equality:

$$4\alpha(X, Y) = \sum_{i=1}^{2n} \xi_i([J[X, Y], X_i] - J[[X, Y], X_i]) \tag{87}$$

As the form α is invariant on the left by G, this equality is verified for all points. For any endomorphism Θ of the space g such that $\Theta b \subset b$, we denote by $Tr_b\Theta$ the trace of the restriction of Θ to b and by $Tr_{g/b}\Theta$ the trace of the endomorphism of *g/b* deduced from Θ by passage to the quotient, with $Tr\Theta = Tr_b\Theta + Tr_{g/b}\Theta$. We have:

$$T_{r_{g/b}\Theta} = \sum_{i=1}^{2n} \xi_i(\Theta X_i) \tag{88}$$

Whatever $X \in g$ and $s \in B$, we have $J(Xs) - (JX)s \in b$. If $ad(Y)$ is the endomorphism of g defined by $ad(Y).Z = [Y, Z]$, we have $(Jad(Y) - ad(Y)J)g \subset b$ for all $Y \in b$. We can deduce, for all $X \in g$, the endomorphism $ad(JX) - Jad(X)$ leaves steady the subspace b. Koszul defines a linear form Ψ on the space g by defining:

$$\Psi(X) = Tr_{g/b}(ad(JX) - Jad(X)) \text{ },\forall X \in g \tag{89}$$

Koszul has finally obtained the following fundamental theorem:

Theorem of Koszul [15]:

The Kähler form of the Hermitian canonical form has for image by p^* the differential of the form $-\frac{1}{4}\Psi(X) = -\frac{1}{4}Tr_{g/b}(ad(JX) - Jad(X)),\forall X \in g$

Koszul note that the form Ψ is independent of the choice of the tensor J. It is determined by the invariant complex structure of *G/B*. The form Ψ is right invariant by B. For all $s \in B$, note the endomorphism $r(s) : X \to Xs$ of g. Since $J(Xs) = (JX)s \mod b$ and that $Tr_{g/b}ad(Y) = 0$, we have:

$$\Psi(Xs) = Tr_{g/b}(ad((JX)s) - Jad(Xs)), \quad \forall X \in g \text{ }, \forall Y \in b \tag{90}$$

$$\Psi(Xs) = Tr_{g/b}(r(s)ad(JX)r(s)^{-1} - Jr(s)ad(X)r(s)^{-1}) \tag{91}$$

$$\Psi(Xs) = \Psi(X) + Tr_{g/b}((J - r(s)^{-1}Jr(s))ad(X)), \quad \forall X \in g, s \in B \tag{92}$$

As $\left(J - r(s)^{-1}Jr(s)\right)$ maps g in b, we get $\Psi(Xs) = \Psi(X)$. The form Ψ is not zero on b. This is not the image by p^* of a differential form of *G/B*. However, the right invariance of Ψ on B is translated, infinitesimally by the relation:

$$\Psi([b, g]) = (0) \tag{93}$$

Koszul proved that the canonical hermitian form h of a homogeneous Kähler manifold *G/B* has the following expression:

$$\begin{aligned} &\eta(X, Y) = \frac{1}{2}\Psi([JX, Y]) \\ &\text{with } \begin{cases} \Psi([X, Y]) = \Psi([JX, JY]) \\ \eta([JX, JY]) = \eta(X, Y) \end{cases} \forall X, Y \in g \end{aligned} \tag{94}$$

To do, the link with the first chapters, I can summarize the main result of Koszul that there is an integrable structure almost complex J on g, and for $l \in g^*$ defined by a positive J -invariant inner product on g:

$$\langle X, Y\rangle_l = \langle [JX, Y], l\rangle \tag{95}$$

Koszul has proposed as admissible form, $l \in g^*$, the form ξ:

$$\Psi(X) = \langle X, \xi \rangle = T_r[ad(JX) - J.ad(X)] \ \forall X \in g \tag{96}$$

Koszul proved that $\langle X, Y \rangle_\xi$ coincides, up to a positive multiplicative constant; with the real part of the Hermitian inner product obtained by the Bergman metric of symmetric homogeneous bounded domains DS by identifying g with the tangent space of DS. $\Psi(X)$ is the restriction to g of a differential form Ψ of degree 1, with left invariance on G. This form is fully defined by the invariant complex structure of G/B. This form is invariant to the choice of J. This form is invariant on the right by B. We have $\Psi([X, Y]) = 0$ with $X \in g, Y \in b$. The exterior differential $d\Psi$ of Ψ is the inverse image by the projection $G \to G/B$ of degree 2 form Ω. This form Ω is, up to a constant, the Kähler form h, defined by the canonical Hermitian form of G/B: $h(\pi.X, \pi.Y) = \frac{1}{2}(d\Psi)(X, J.Y), \forall X, Y \in G$ as it is proved in Bourbaki seminar by Koszul in [129].

The 1st Koszul form is then given by:

$$\alpha = -\frac{1}{4} d\Psi(X) \tag{97}$$

We can illustrate this structure for the simplest example of DS, the Poincaré upper half-plane $V = \{z = x + iy/y > 0\}$ which is isomorphic to the open $zz^* < 1$, which is a bounded domain. The group G of transformations $z \to az + b$ with a and b real values with $a > 0$ is simply transitive in V. We identify G and V by the application passing from $s \in G$ an element to the image $i = \sqrt{-1}$ by s.

Let's define vector fields $X = y\frac{d}{dx}$ and $Y = y\frac{d}{dy}$ which generate the vector space of left invariant vectors fields on G, and J an almost complex structure on V defined by $JX = Y$. As $[X, Y] = -Y$ and $ad(Y).Z = [Y, Z]$ then:

$$\begin{cases} Tr[ad(JX) - Jad(X)] = 2 \\ Tr[ad(JY) - Jad(Y)] = 0 \end{cases} \tag{98}$$

The Koszul forms and the Koszul metric are respectively given by:

$$\Psi(X) = 2\frac{dx}{y} \Rightarrow \alpha = -\frac{1}{4} d\Psi = -\frac{1}{2}\frac{dx \wedge dy}{y^2} \Rightarrow ds^2 = \frac{dx^2 + dy^2}{2y^2} \tag{99}$$

I note that $\alpha = -\frac{1}{4} d\Psi(X)$ is indeed the Kähler form of Poincaré's metric, which is invariant by the automorphisms of the upper half-plane.

The following example concerns $V = \{Z = X + iY/X, Y \in Sym(p), Y > 0\}$ the upper half-space of Siegel (which is the most natural extension of the Poincaré half-plane) with:

$$\begin{cases} SZ = (AZ + B)D^{-1} \\ A^T D = I, B^T D = D^T B \end{cases} \text{with } S = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \text{ and } J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{100}$$

We can then compute Koszul forms and the metric:

$$\Psi(dX+idY)=\frac{3p+1}{2}Tr(Y^{-1}dX)$$
$$\Rightarrow\begin{cases}\alpha=-\frac{1}{4}d\Psi=\frac{3p+1}{8}Tr\left(Y^{-1}dZ\wedge Y^{-1}d\bar{Z}\right)\\ ds^2=\frac{(3p+1)}{8}Tr\left(Y^{-1}dZY^{-1}d\bar{Z}\right)\end{cases}\qquad(101)$$

We recover Carl-Ludwig Siegel metric for the upper half space.

More recent development on Kähler manifolds are described in [130] et [131].

Koszul studied symmetric homogeneous spaces and defines the relation between invariant flat affine connections and the affine representations of Lie algebras and invariant Hessian metrics characterized by affine representations of Lie algebras. Koszul provides a correspondence between symmetric homogeneous spaces with invariant Hessian structures using affine representations of Lie algebras, and proves that a symmetric homogeneous space simply connected with an invariant Hessian structure is a direct product of a Euclidean space and of a homogeneous dual-cone. Let *G* be a connected Lie group and *G/K* a homogeneous space over which *G* acts effectively. Koszul gives a bijective correspondence between all planar *G*-invariantes connections on *G/K* and all of a certain class of affine representations of the Lie algebra of *G*. The main theorem of Koszul is:

Koszul's theorem: Let *G/K* be a homogeneous space of a connected Lie group *G* and be g and k the Lie algebras of *G* and *K*, assuming that *G/K* has G-invariant connection, then admits an affine representation *(f, q)* on the vector space *E*. Conversely, assume that *G* is simply connected and has an affine representation, then *G/K* admits a flat *G*-invariant connection.

In the foregoing, the basic tool studied by Koszul is the affine representation of Lie algebra and Lie group. To study these structures, Koszul introduced the following developments.

Let Ω a convex domain on R^n without any straight lines, and an associated convex cone $V(\Omega)=\{(\lambda x,x)\in R^n\times R/x\in\Omega,\lambda\in R^+\}$, then there exist an affine embedding:

$$\ell:x\in\Omega\mapsto\begin{bmatrix}x\\1\end{bmatrix}\in V(\Omega)\qquad(102)$$

If we consider η the group of homomorphism of $A(n,R)$ in $GL(n+1,R)$ given by:

$$s\in A(n,R)\mapsto\begin{bmatrix}\mathsf{f}(s)&\mathsf{q}(s)\\0&1\end{bmatrix}\in GL(n+1,R)\qquad(103)$$

and the affine representation of Lie algebra:

$$\begin{bmatrix}f&q\\0&0\end{bmatrix}\qquad(104)$$

with $A(n, R)$ the group of all affine representations of R^n. We have $\eta(G(\Omega)) \subset G(V(\Omega))$ and the pair (η, ℓ) of homomorphism $\eta : (G(\Omega) \to G(V(\Omega))$ and the application $\ell : \Omega \to V(\Omega)$ is equivariant.

If we observe Koszul affine representations of Lie algebra and Lie group, we have to consider G a convex Lie group and E a real or complex vector space of finite size, Koszul has introduced an affine representation of G in E such that:

$$\begin{aligned} &E \to E \\ &a \mapsto sa \, \forall s \in G \end{aligned} \tag{105}$$

is an affine representation. We set $A(E)$ the set of all affine transformation of a real vector space E, a Lie group called affine representation group of E. The set $GL(E)$ of all regular linear representation of E, a sub-group of $A(E)$.

We define a linear representation of G in $GL(E)$:

$$\begin{aligned} &\mathsf{f} : G \to GL(E) \\ &s \mapsto \mathsf{f}(s)a = sa - so \quad \forall a \in E \end{aligned} \tag{106}$$

and a map from G to E:

$$\begin{aligned} &\mathsf{q} : G \to E \\ &s \mapsto \mathsf{q}(s)so \quad \forall s \in G \end{aligned} \tag{107}$$

then, we have $\forall s, t \in G$:

$$\mathsf{f}(s)\mathsf{q}(t) + \mathsf{q}(s) = \mathsf{q}(st) \tag{108}$$

deduced from $\mathsf{f}(s)\mathsf{q}(t) + \mathsf{q}(s) = s\mathsf{q}(t) - s\mathsf{q} + so = s\mathsf{q}(t) = sto = \mathsf{q}(st)$.

Inversely, if a map q from G to E and a linear representation f from G to $GL(E)$ verifying previous equation, then we can define an affine representation from G in E, written by (f, q):

$$\mathrm{A}ff(s) : a \mapsto sa = \mathsf{f}(s)a + \mathsf{q}(s) \, \forall s \in G, \forall a \in E \tag{109}$$

The condition $\mathsf{f}(s)\mathsf{q}(t) + \mathsf{q}(s) = \mathsf{q}(st)$ is equal to the request that the following mapping is an homomorphism:

$$\mathrm{A}ff : s \in G \mapsto \mathrm{A}ff(s) \in A(E) \tag{110}$$

We write f the affine representation of Lie algebra g of G, defined by f and q the restriction to g to the differential of q (f and q differential of f and q respectively), Koszul proved the following equation:

$$\begin{aligned} &f(X)q(Y) - f(Y)q(X) = q([X, Y]) \forall X, Y \in \mathsf{g} \\ &\text{with} \quad f : \mathsf{g} \to gl(E) \quad \text{and}\, q : \mathsf{g} \mapsto E \end{aligned} \tag{111}$$

where $gl(E)$ the set of all linear endomorphisms of E, Lie algebra of $GL(E)$.

We use the assumption that:

$$q(Ad_s Y) = \left.\frac{d\mathsf{q}\left(s.e^{tY}.s^{-1}\right)}{dt}\right|_{t=0} = \mathsf{f}(s)f(Y)\mathsf{q}(s^{-1}) + \mathsf{f}(s)q(Y) \tag{112}$$

We then obtain:

$$q([X,Y]) = \left.\frac{d\mathsf{q}\left(Ad_{e^{tX}} Y\right)}{dt}\right|_{t=0} = f(X)q(Y)\mathsf{q}(e) + \mathsf{f}(e)f(Y)(-q(X)) + f(X)q(Y) \tag{113}$$

where e is neutral element of G. Since $\mathsf{f}(e)$ is identity map and $\mathsf{q}(e) = 0$, we have the equality:

$$f(X)q(Y) - f(Y)q(X) = q([X,Y]) \tag{114}$$

A pair (f,q) of linear representation of f of a Lie algebra g on E and a linear map q from g in E is an affine representation of g in E, if it satisfy:

$$f(X)q(Y) - f(Y)q(X) = q([X,Y]) \tag{115}$$

Inversely, if we assume that g has an affine representation (f,q) on E, by using the coordinate systems $\{x^1,\dots,x^n\}$ on E, we can express the affine map $v \mapsto f(X)v + q(Y)$ by a matrix representation of size $(n+1)\times(n+1)$:

$$aff(X) = \begin{bmatrix} f(X) & q(X) \\ 0 & 0 \end{bmatrix} \tag{116}$$

where $f(X)$ is a matrix of size $n \times n$ and $q(X)$ a vector of size n.

$X \mapsto aff(X)$ is an injective homomorphism of Lie algebra g in Lie algebra of matrices $(n+1)\times(n+1)$, $gl(n+1,R)$:

$$\left| \begin{array}{l} \mathsf{g} \to gl(n+1,R) \\ X \mapsto aff(X) \end{array} \right. \tag{117}$$

If we note $\mathsf{g}_{aff} = aff(\mathsf{g})$, we write G_{aff} linear Lie sub-group of $GL(n+1,R)$ generated by g_{aff} . One element of $s \in G_{aff}$ could be expressed by:

$$Aff(s) = \begin{bmatrix} \mathsf{f}(s) & \mathsf{q}(s) \\ 0 & 1 \end{bmatrix} \tag{118}$$

Let M_{aff} the orbit of G_{aff} from the origin o, then $M_{aff} = \mathsf{q}(G_{aff}) = G_{aff}/K_{aff}$ where $K_{aff} = \{s \in G_{aff}/\mathsf{q}(s) = 0\} = Ker(\mathsf{q})$.

We can give as example the following case. Let Ω a convex domain in R^n without any straight line, we define the cone $V(\Omega)$ in $R^{n+1} = R^n \times R$ by $V(\Omega) = \{(\lambda x, x) \in R^n \times R/x \in \Omega, \lambda \in R^+\}$. Then, there is an affine embedding:

$$\ell : x \in \Omega \mapsto \begin{bmatrix} x \\ 1 \end{bmatrix} \in V(\Omega) \tag{119}$$

If we consider η the group of homomorphisms of $A(n, R)$ in $GL(n + 1, R)$ given by:

$$s \in A(n, R) \mapsto \begin{bmatrix} \mathrm{f}(s) & \mathrm{q}(s) \\ 0 & 1 \end{bmatrix} \in GL(n + 1, R) \tag{120}$$

with $A(n, R)$ the group of all affine transformations in R^n. We have $\eta(G(\Omega)) \subset G(V(\Omega))$ and the pair (η, ℓ) of homomorphism $\eta : G(\Omega) \to G(V(\Omega))$ and the map $\ell : \Omega \to V(\Omega)$ are equivariant:

$$\ell \circ s = \eta(s) \circ \ell \,\textbf{and}\, d\ell \circ s = \eta(s) \circ d\ell \tag{121}$$

6 Koszul Lecture on Geometric and Analytics Mechanics, Related to Geometric Theory of Heat (Souriau's Lie Group Thermodynamics) and Theory of Information (Information Geometry)

Before that Professor Koszul passed away in January 2018, he gave his agreement to his book "Introduction to Symplectic Geometry" translation from Chinese to English by SPRINGER [18]. This Koszul's book translation genesis dates back to 2013. We had contacted Professor Jean-Louis Koszul, to deeper understand his work in the field of homogeneous bounded domains within the framework of Information Geometry. Professor Michel Boyom succeeded to convince Jean-Louis Koszul to answer positively to our invitation to attend the 1st GSI "*Geometric Science of Information*" conference in August 2013 at Ecole des Mines ParisTech in Paris, and more especialy to attend the talk of Hirohiko Shima, given for his honor on the topic "*Geometry of Hessian Structures*" (Fig. 9).

I was more particularly interested by Koszul's work developed in the paper « ***Domaines bornées homogènes et orbites de groupes de transformations affines*** » [45] of 1961, written by Koszul at the Institute for Advanced Studies at Princeton during a stay funded by the National Science Foundation. Koszul proved in this paper that on a complex homogeneous space, an invariant volume defines with the complex structure the canonical invariant Hermitian form introduced in [15]. It is in this article that Koszul uses the ***affine representation of Lie groups and Lie algebras***.

Fig. 9 Jean-Louis Koszul and Hirihiko Shima at GSI'13 "*Geometric Science of Information*" conference in Ecole des Mines ParisTech in Paris, October 2013

The use by Koszul of the affine representation of Lie groups and Lie algebras drew our attention, especially on the links of his approach with the similar one used by Jean-Marie Souriau in geometric mechanics in the framework of homogeneous symplectic manifolds. I have then looked for links between Koszul and Souriau works. I finally discovered, that in 1986, Koszul published this book "Introduction to symplectic geometry" following a Chinese course in China. I also observed that this book takes up and develops works of Jean-Marie Souriau on homogeneous symplectic manifolds and ***the affine representation of Lie algebras and Lie groups in geometric mechanics***.

I have then exchanged e-mails with Professor Koszul on Souriau works and on genesis of this Book. In May 2015, questioning Koszul on Souriau work on Geometric Mechanics and on Lie Group Thermodynamics, Koszul answered me "[*A l'époque où Souriau développait sa théorie, l'establishment avait tendance à ne pas y voir des avancées importantes. Je l'ai entendu exposer ses idées sur la thermodynamique mais je n'ai pas du tout réalisé à l'époque que la géométrie hessienne était en jeu.*] *At the time when Souriau was developing his theory, the establishment tended not to see significant progress. I heard him explaining his ideas on thermodynamics but I did not realize at the time that Hessian geometry was at stake*". In September 2016, I asked him the origins of Lie Group and Lie Algebra Affine representation. Koszul informed me that he attended Elie Cartan Lecture, where he presented seminal work on this topic: "*[Il y a là bien des choses que je voudrais comprendre (trop peut-être !), ne serait-ce que la relation entre ce que j'ai fait et les travaux de Souriau. Détecter l'origine d'une notion ou la première apparition d'un résultat est souvent difficile. Je ne suis certainement pas le premier à avoir utilisé des représentations affines de groupes ou d'algèbres de Lie. On peut effectivement imaginer que cela se trouve chez Elie Cartan, mais je ne puis rien dire de précis. A propos d'Elie Cartan: je n'ai pas été son élève. C'est Henri Cartan qui a été mon maître pendant mes années de thèse. En 1941 ou 42 j'ai entendu une brève série de conférences données par Elie à l'Ecole Normale et ce sont des travaux d'Elie qui ont été le point de départ de mon travail de thèse.] There are many things that I would like to understand (too much perhaps!), If only the relationship between what I did and the work of Souriau. Detecting the origin of a notion or the first appearance of a result is often difficult. I am certainly not the first to have used affine representations of Lie groups or Lie algebras. We can imagine that it is at Elie Cartan, but I cannot say anything specific. About Elie Cartan: I was not his student. It was Henri Cartan who was my master during my years of thesis. In 1941 or 42, I heard a brief series of lectures given by Elie at the Ecole Normale and it was Elie's work that was the starting point of my thesis work*".

After discovering the existence of this Koszul's book, written in Chinese based on a course given at Nankin, on "Introduction to Symplectic Geometry", where he made reference to Souriau's book and developed his main tools, I started to discuss its content. In January 2017, Koszul wrote me, with the usual humility "*[Ce petit fascicule d'introduction à la géométrie symplectique a été rédigé par un assistant de Nankin qui avait suivi mon cours. Il n'y a pas eu de version initiale en français.] This small introductory booklet on symplectic geometry was written by a Nanjing assistant who had taken my course. There was no initial version in French* ". I asked him if he had personal archive of this course, he answered "*[Je n'ai pas conservé de notes préparatoires à ce cours. Dites-moi à quelle adresse je puis vous envoyer un exemplaire du texte chinois.] I have not kept any preparatory notes for this course. Tell me where I can send you a copy of the Chinese text.* ". Professor Koszul then sent me his last copy of this book in Chinese, a small green book (Fig. 10).

I was not able to read the Chinese text, but I have observed in Chap. 4 "*Symplectic G-spaces*" and in Chap. 5 "*Poisson Manifolds*", that their equations content new original developments of Souriau work on moment map and affine representation of Lie Group and Lie Algebra. More especially, Koszul considered equivariance of moment map, where I recover Souriau theorem. Koszul shows that when *(M; ω)* is a connected Hamiltonian *G*-space and μ a moment map of the action of *G*, there exists an affine action of *G* on g^* (dual Lie algebra), whose linear part is the coadjoint action, for which the moment μ is equivariant. Koszul developed Souriau idea that this affine action is obtained by modifying the coadjoint action by means of a closed cochain (called cocycle by Souriau), and that *(M; ω)* is a *G*-Poisson space making reference to Souriau's book for more details.

About collaboration between Koszul and Souriau and another potential Lecture on Symplectic Geometry in Toulouse, Koszul informed me in February 2017 that: "*[J'ai plus d'une fois rencontré Souriau lors de colloques, mais nous n'avons jamais*

Fig. 10 Original small green Koszul’s book “Introduction to symplectic geometry” in Chinese

collaboré. Pour ce qui est de cette allusion à un “cours” donné à Toulouse, il y erreur. J’y ai peut être fait un exposé en 81, mais rien d’autre.] I have met Souriau more than once at conferences, but we have never collaborated. As for this allusion to a “course” given in Toulouse, there is error. I could have made a presentation in 81, but nothing else. “. Koszul admitted that he had no direct collaboration with Souriau: “*[Je ne crois pas avoir jamais parlé de ses travaux avec Souriau. Du reste j’avoue ne pas en avoir bien mesuré l’importance à l’époque] I do not think I ever talked about his work with Souriau. For the rest, I admit that I did not have a good idea of the importance at the time*”.

Considering the importance of this book for different communities, I tried to find an editor for its translation in English. By chance, I met Catriona Byrne from SPRINGER, when I gave a talk at IHES, invited by Pierre Cartier, on Koszul and Souriau works application in Radar. With help of Michel Boyom, we have convinced Professor Koszul to translate this book, proposing to contextualize this book with regard to the contemporary research trends in Geometric Mechanics, Lie Groups Thermodynamics and Geometric Science of Information. Professors Marle and Boyom accepted to check the translation and help me to write the forewords.

In the historical Foreword of this book, Koszul write "*The development of analytical mechanics provided the basic concepts of symplectic structures. The term symplectic structure is due largely to analytical mechanics. But in this book, the applications of symplectic structure theory to mechanics is not discussed in any detail*". Koszul considers in this book purely algebraic and geometric developments of Geometric/Analytic Mechanics developed during the 60th, more especially Jean-Marie Souriau works detailed in Chaps. 4 and 5. ***The originality of this book lies in the fact that Koszul develops new points of view, and demonstrations not considered initially by Souriau and Geometrical Mechanics community***.

Jean-Marie Souriau was the Creator of a new discipline called "*Mécanique Géométrique (Geometric Mechanics)*". Souriau observed that the collection of motions of a dynamical system is a manifold with an antisymmetric flat tensor that is a symplectic form where the structure contains all the pertinent information on the state of the system (positions, velocities, forces, etc.). Souriau said: "*[Ce que Lagrange a vu, que n'a pas vu Laplace, c'était la structure symplectique] What Lagrange saw, that Laplace didn't see, was the symplectic structure*". Using the symmetries of a symplectic manifold, Souriau introduced a mapping which he called the "*moment map*", which takes its values in a space attached to the group of symmetries (in the dual space of its Lie algebra). Souriau associated to this moment map, the notion of symplectic cohomology, linked to the fact that such a moment is defined up to an additive constant that brings into play an algebraic mechanism (called cohomology). Souriau proved that the moment map is a constant of the motion, and provided geometric generalization of Emmy Noether invariant theorem (invariants of E. Noether theorem are the components of the moment map). Souriau has defined in a geometrically way the Noetherian symmetries using the Lagrange-Souriau 2 form with the application map. Influenced by François Gallissot (Souriau and Galissot both attended ICM'54 in Moscow, and should have exchanged during this conference), Souriau has introduced in Mechanics the Lagrange 2-form, recovering seminal Lagrange ideas. Motivated by variational principles in a coordinate free formulation, inspired by Henri Poincaré and Elie Cartan who introduced a differential 1-form instead of the Lagrangian, Souriau introduced the Lagrange 2-form as the exterior differential of the Poincaré-Cartan 1-form, and obtained the phase space as a symplectic manifold. Souriau proposed to consider this Lagrange 2-form as the fundamental structure for Lagrangian system and not the classical Lagrangian function or the Poincaré-Cartan 1-form. This 2-form is called Lagrange-Souriau 2 form, and is the exterior derivative of the Lepage form (the Poincaré-Cartan form is a first order Lepage form). This structure is developed in Koszul book, where the authors shows that when $(M;\omega)$ is an exact symplectic manifold (when there exists a 1-form α on M such that $\omega = -d\alpha$), and that a symplectic action leaves not only ω, but α invariant, this action is strongly Hamiltonian ($(M;\omega)$ is a g-Poisson space). Koszul shows that a symplectic action of a Lie algebra g on an exact symplectic manifold $(M;\omega = -d\alpha)$ that leaves invariant not only ω, but also α, is strongly Hamiltionian.

In this Book in Chap. 4, Koszul calls symplectic G-space a symplectic manifold $(M;\omega)$ on which a Lie group G acts by a symplectic action (an action which leaves unchanged the symplectic form ω). Koszul then introduces and develop properties of the moment map μ (Souriau's invention) of a Hamiltonian action of the Lie algebra g. Koszul also defines the Souriau 2-cocycle, considering that the difference of two moments of the same Hamiltonian action is a locally constant application on M, showing that when μ is a moment map, for every pair *(a;b)* of elements of g, the function $c_{\mu}(a, b) = \{\langle \mu, a\rangle, \langle \mu, b\rangle\} - \langle \mu, \{a, b\}\rangle$ is locally constant on M, defining an antisymmetric bilinear application of *gxg* in *H⁰(M; R)* which verifies Jacobi's identity. ***This is the 2-cocycle introduced by Jean-Marie Souriau in Geometric Mechanics, that will play a fundamental role in Souriau Lie Groups Thermodynamics to define an extension of the Fisher Metric from Information Geometry (what I will call Fisher-Souriau metric in the following).***

To highlight the importance of this Koszul book, we will illustrate the links of the detailed tools, including demonstrations or original Koszul extensions, with Souriau's Lie Groups Thermodynamics, whose applications range from statistical physics to machine learning in Artificial Intelligence. In 1970, Souriau introduced the concept of co-adjoint action of a group on its momentum space, based on the orbit method works, that allows to define physical observables like energy, heat and momentum or moment as pure geometrical objects. In a first step to establish new foundations of thermodynamics, Souriau has defined a Gibbs canonical ensemble on a symplectic manifold M for a Lie group action on M. In classical statistical mechanics, a state is given by the solution of Liouville equation on the phase space, the partition function. As symplectic manifolds have a completely continuous measure, invariant by diffeomorphisms (the Liouville measure λ*)*, Souriau has proved that when statistical states are Gibbs states (as generalized by Souriau), they are the product of the Liouville measure by the scalar function given by the generalized partition function $e^{\Phi(\beta)-\langle\beta,U(\xi)\rangle}$ defined by the energy U (defined in the dual of the Lie algebra of this dynamical group) and the geometric temperature β, where Φ is a normalizing constant such the mass of probability is equal to 1, $\Phi(\beta) = -\log \int\limits_M e^{-\langle\beta,U(\xi)\rangle} d\lambda$. Jean-Marie Souriau then generalizes the Gibbs equilibrium state to all symplectic manifolds that have a dynamical group. Souriau has observed that if we apply this theory for Galileo, the symmetry will be broken. For each temperature β, element of the Lie algebra $\mathfrak{g}$, Souriau has introduced a tensor $\tilde{\Theta}_{\beta}$, equal to the sum of the cocycle $\tilde{\Theta}$ and the heat coboundary (with [.,.] Lie bracket):

$$\tilde{\Theta}_{\beta}(Z_1, Z_2) = \tilde{\Theta}(Z_1, Z_2) + \langle Q, ad_{Z1}(Z_2)\rangle \tag{122}$$

This tensor $\tilde{\Theta}_{\beta}$ has the following properties: $\tilde{\Theta}(X, Y) = \langle\Theta(X), Y\rangle$ where the map Θ is the symplectic one-cocycle of the Lie algebra $\mathfrak{g}$ with values in $\mathfrak{g}^*$, with $\Theta(X) = T_e\theta(X(e))$ where θ the one-cocycle of the Lie group G. $\tilde{\Theta}(X, Y)$ is constant on M and the map $\tilde{\Theta}(X, Y) : \mathfrak{g}\times\mathfrak{g} \to \Re$ is a skew-symmetric bilinear form, and is called the ***symplectic two-cocycle of Lie algebra*** $\mathfrak{g}$ associated to the *moment map J*, with the following properties:

$$\tilde{\Theta}(X, Y) = J_{[X,Y]} - \{J_X, J_Y\}\text{with } J \text{ the Moment Map} \tag{123}$$

$$\tilde{\Theta}([X, Y], Z) + \tilde{\Theta}([Y, Z], X) + \tilde{\Theta}([Z, X], Y) = 0 \tag{124}$$

where J_X linear application from $\mathfrak{g}$ to differential function on $M : \mathfrak{g} \to C^{\infty}(M, R), X \to J_X$ and the associated differentiable application J, called moment map:

$$J : M \to \mathfrak{g}^*, x \mapsto J(x) \quad \text{such that } J_X(x) = \langle J(x), X\rangle, X \in \mathfrak{g} \tag{125}$$

The geometric temperature, element of the algebra $\mathfrak{g}$, is in the kernel of the tensor $\tilde{\Theta}_\beta$:

$$\beta \in Ker\, \tilde{\Theta}_\beta \text{ such that } \tilde{\Theta}_\beta(\beta, \beta) = 0, \quad \forall\beta \in \mathfrak{g} \tag{126}$$

The following symmetric tensor $g_\beta([\beta, Z_1], [\beta, Z_2]) = \tilde{\Theta}_\beta(Z_1, [\beta, Z_2])$, defined on all values of $ad_\beta(.) = [\beta, .]$ is positive definite, and defines extension of classical Fisher metric in Information Geometry (as hessian of the logarithm of partition function):

$$g_\beta([\beta, Z_1], Z_2) = \tilde{\Theta}_\beta(Z_1, Z_2), \quad \forall Z_1 \in \mathfrak{g}, \forall Z_2 \in \mathrm{Im}\big(ad_\beta(.)\big) \tag{127}$$

$$\text{With } g_\beta(Z_1, Z_2) \geq 0, \quad \forall Z_1, Z_2 \in \mathrm{Im}\big(ad_\beta(.)\big) \tag{128}$$

***These equations are universal*, because they are not dependent on the symplectic manifold but only on the dynamical group *G*, the symplectic two-cocycle Θ, the temperature β and the heat Q. Souriau called it "*Lie groups thermodynamics*".**

This antisymmetric bilinear map (127) and (128), with definition (122) and (123) is exactly equal to the mathematical object introduced in Chap. 4 of Koszul's book by:

$$c_\mu(a, b) = \{\langle\mu, a\rangle, \langle\mu, b\rangle\} - \{\mu, \langle a, b\rangle\} \tag{129}$$

In this book, Koszul has studied this antisymmetric bilinear map considering the following developments. For any moment map μ, Koszul defines the skew symmetric bilinear form $\mathrm{c}_\mu(\mathrm{a, b})$ on Lie algebra by:

$$\mathrm{c}_\mu(\mathrm{a, b}) = \big\langle d\theta_\mu(a), b\big\rangle, a, b \in \mathfrak{g} \tag{130}$$

Koszul observes that if we use:

$$\theta_\mu(st) = \mu(stx) - Ad^*_{st}\mu(x) = \theta_\mu(s) + Ad^*_s\mu(tx) - Ad^*_s Ad^*_t\mu(x) = \theta_\mu(s) + Ad^*_s\theta_\mu(t)$$

by developing $d\mu(ax) =^t ad_a\mu(x) + d\theta_\mu(a), \; x \in M, a \in \mathfrak{g}$, he obtains:

$$\langle d\mu(ax), b\rangle = \langle\mu(x), [a, b]\rangle + \big\langle \mathrm{d}\theta_\mu(a), b\big\rangle = \{\langle\mu, a\rangle, \langle\mu, b\rangle\}(x), \; x \in M, a, b \in \mathfrak{g} \tag{131}$$

We have then:

$$c_\mu(\mathrm{a, b}) = \{\langle\mu, a\rangle, \langle\mu, b\rangle\} - \langle\mu, [a, b]\rangle = \left\langle \mathrm{d}\theta_\mu(a), b\right\rangle, a, b \in \mathfrak{g} \tag{132}$$

and the property:

$$c_\mu(\mathrm{[a, b], c}) + c_\mu([b, c], a) + c_\mu([c, a], b) = 0,\ a, b, c \in \mathfrak{g} \tag{133}$$

Koszul concludes by observing that if the moment map is transform as $\mu' = \mu + \phi$ then we have:

$$c_{\mu'}(a, b) = c_\mu(a, b) - \langle\phi, [a, b]\rangle \tag{134}$$

Finally using $c_\mu(\mathrm{a, b}) = \{\langle\mu, a\rangle, \langle\mu, b\rangle\} - \langle\mu, [a, b]\rangle = \left\langle \mathrm{d}\theta_\mu(a), b\right\rangle, a, b \in \mathfrak{g}$, koszul highlights the property that:

$$\left\{\mu^*(a), \mu^*(b)\right\} = \{\langle\mu, a\rangle, \langle\mu, b\rangle\} = \mu^*\big([a, b] + c_\mu(a, b)\big) = \mu^*\{a, b\}_{c_\mu} \tag{135}$$

In Chap. 4, Koszul introduces the equivariance of the moment map μ. Based on the definitions of the adjoint and coadjoint representations of a Lie group or a Lie algebra, Koszul proves that when *(M; ω)* is a connected Hamiltonian *G*-space and $\mu : M \to \mathfrak{g}^*$ a moment of the action of *G*, there exists an affine action of *G* on $g*$, whose linear part is the coadjoint action, for which the moment μ is equivariant. This affine action is obtained by modifying the coadjoint action by means of a cocycle. This notion is also developed in Chap. 5 for Poisson manifolds. Defining classical operation $Ad_s a = sas^{-1}, s \in G, a \in \mathfrak{g}$, $ad_a b = [a, b], a \in \mathfrak{g}, b \in \mathfrak{g}$ and coadjoint action given by $Ad_s^* =^t Ad_{s^{-1}}, s \in G$ with classical properties:

$$Ad_{\exp a} = \exp(-ad_a), a \in \mathfrak{g} \text{ or } Ad^*_{\exp a} = \exp^t(ad_a), a \in \mathfrak{g} \tag{136}$$

Koszul considers:

$$x \mapsto sx, x \in M, \mu : M \to \mathfrak{g}^* \tag{137}$$

From which, he obtains:

$$\langle d\mu(v), a\rangle = \omega(ax, v) \tag{138}$$

Koszul then study $\mu \circ s_M - Ad_s^* \circ \mu : M \to \mathfrak{g}^*$, and develops:

$$d\left\langle Ad_s^* \circ \mu, a\right\rangle = \left\langle Ad_s^* d\mu, a\right\rangle = \langle d\mu, Ad_{s^{-1}} a\rangle \tag{139}$$

$$\langle d\mu(v), Ad_{s^{-1}} a\rangle = \omega(s^{-1}asx, v) = \omega(asx, sv) = \langle d\mu(sv), a\rangle = (d\langle\mu \circ s_M, a\rangle)(v) \tag{140}$$

$$d\langle Ad_s^* \circ \mu, a\rangle = d\langle \mu \circ s_M, a\rangle \text{ and then proves that } d\langle \mu \circ s_M - Ad_s^* \circ \mu, a\rangle = 0 \tag{141}$$

Koszul considers the cocycle given by $\theta_\mu(s) = \mu(sx) - Ad_s^* \mu(x), s \in G$, and observes that:

$$\theta_\mu(st) = \theta_\mu(s) - Ad_s^* \theta_\mu(t), s, t \in G \tag{142}$$

From this action of the group on dual Lie algebra:

$$G \times \mathfrak{g}^* \to \mathfrak{g}^*, (s, \xi) \mapsto s\xi = Ad_s^* \xi + \theta_\mu(s) \tag{143}$$

Koszul introduces the following properties:

$$\mu(sx) = s\mu(x) = Ad_s^* \mu(x) + \theta_\mu(s), \forall s \in G, x \in M \tag{144}$$

$$G \times \mathfrak{g}^* \to \mathfrak{g}^*, (e, \xi) \mapsto e\xi = Ad_e^* \xi + \theta_\mu(e) = \xi + \mu(x) - \mu(x) = \xi \tag{145}$$

$$(s_1 s_2)\xi = Ad_{s_1 s_2}^* \xi + \theta_\mu(s_1 s_2) = Ad_{s_1}^* Ad_{s_2}^* \xi + \theta_\mu(s_1) + Ad_{s_1}^* \theta_\mu(s_2)$$

$$(s_1 s_2)\xi = Ad_{s_1}^* (Ad_{s_2}^* \xi + \theta_\mu(s_2)) + \theta_\mu(s_1) = s_1(s_2 \xi), \forall s_1, s_2 \in G, \xi \in \mathfrak{g}^* \tag{146}$$

This Koszul study of the moment map μ equivariance, and the existence of an affine action of G on g^*, whose linear part is the coadjoint action, for which the moment μ is equivariant, is at the cornerstone of Souriau Theory of Geometric Mechanics and Lie Groups Thermodynamics. I illustrate this importance by giving Souriau theorem for Lie Groups Thermodynamics, and the link with, what I call, Souriau-Fisher metric (a covariant definition of Fisher metric):

Theorem (Souriau Theorem of Lie Group Thermodynamics). *Let Ω be the largest open proper subset of $\mathfrak{g}$, Lie algebra of G, such that $\int_M e^{-\langle \beta, U(\xi)\rangle} d\lambda$ and $\int_M \xi . e^{-\langle \beta, U(\xi)\rangle} d\lambda$ are convergent integrals, this set Ω is convex and is invariant under every transformation $Ad_g(.)$. Then, the fundamental equations of Lie group thermodynamics are given by the action of the group:*

$$Action\, of\, Lie\, group\, on\, Lie\, algebra : \beta \to Ad_g(\beta) \tag{147}$$

$$Characteristic\, function\, after\, Lie\, group\, action : \Phi \to \Phi - \langle \theta(g^{-1}), \beta \rangle \tag{148}$$

$$Invariance\, of\, entropy\, with\, respect\, to\, action\, of\, Lie\, group : s \to s \tag{149}$$

$$Action\, of\, Lie\, group\, on\, geometric\, heat : Q \to a(g, Q) = Ad_g^*(Q) + \theta(g) \tag{150}$$

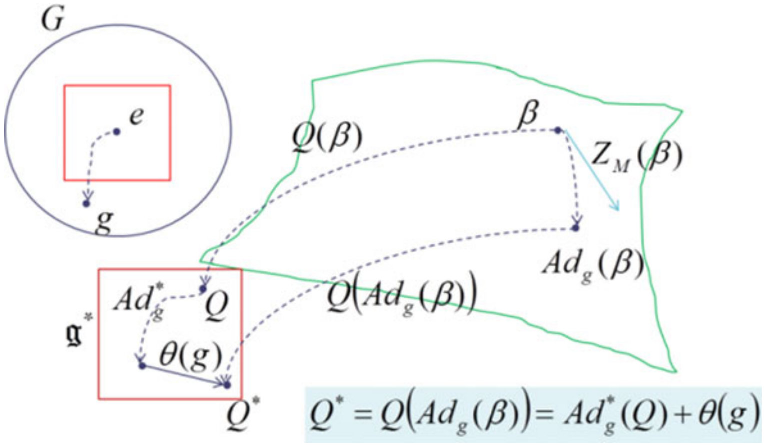


Fig. 11 Broken symmetry on geometric heat Q due to adjoint action of the group on temperature β as an element of the Lie algebra

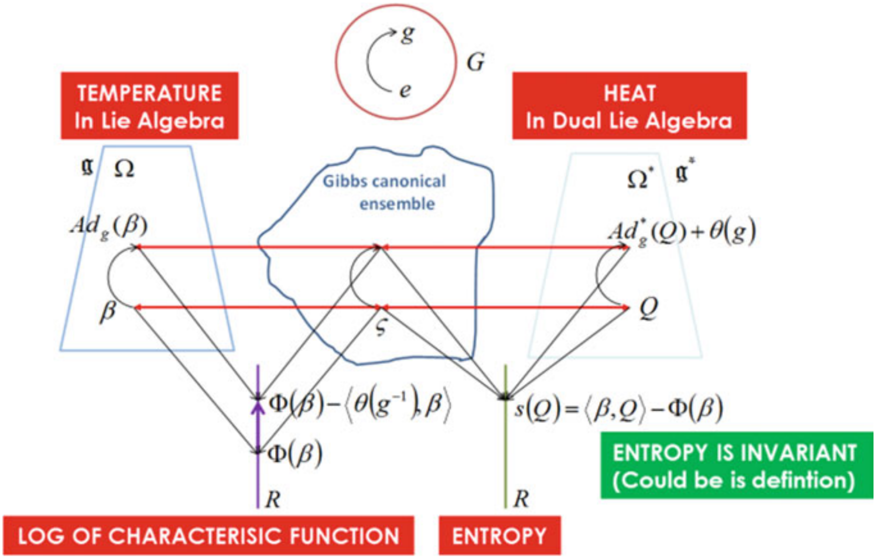


Fig. 12 Global Souriau scheme of Lie group thermodynamics, with entropy *s(Q)*, geometric heat Q element of dual Lie algebra and geometric temperature β element of Lie algebra

Souriau equations of Lie group thermodynamics, related to the moment map μ equivariance, and the existence of an affine action of G on g^*, whose linear part is the coadjoint action, for which the moment μ is equivariant, are summarized in the following figures (Figs. 11 and 12).

I finally observe that ***the Koszul antisymmetric bilinear map*** $c_\mu(a,b) = \{\langle \mu, a\rangle, \langle \mu, b\rangle\} - \langle \mu, \{a,b\}\rangle$ ***is equal to Souriau Riemannian metric***, introduced by mean of symplectic cocycle. I have observed that this metric is a generalization of the Fisher metric from Information Geometry, that I call the Souriau-Fisher metric, defined as a hessian of the partition function logarithm $g_\beta = -\frac{\partial^2 \Phi}{\partial \beta^2} = \frac{\partial^2 \log \psi_\Omega}{\partial \beta^2}$ as in classical information geometry. This new definition of Fisher metric has the property to be covariant under the action of the group G. I have established the equality of two terms, between Souriau definition based on Lie group cocycle Θ and parameterized by "*geometric heat*" Q (element of dual Lie algebra) and "*geometric temperature*" β (element of Lie algebra) and hessian of characteristic function $\Phi(\beta) = -\log \Psi_\Omega(\beta)$ with respect to the variable β:

$$g_\beta([\beta, Z_1], [\beta, Z_2]) = \langle \Theta(Z_1), [\beta, Z_2]\rangle + \langle Q, [Z_1, [\beta, Z_2]]\rangle = \frac{\partial^2 \log \psi_\Omega}{\partial \beta^2} \quad (151)$$

If we differentiate this relation of Souriau theorem $Q(Ad_g(\beta)) = Ad_g^*(Q) + \theta(g)$, this relation occurs:

$$\frac{\partial Q}{\partial \beta}(-[Z_1, \beta], .) = \tilde{\Theta}(Z_1[\beta, .]) + \left\langle Q, Ad_{z_1}([\beta, .])\right\rangle = \tilde{\Theta}_\beta(Z_1, [\beta, .]) \quad (152)$$

$$-\frac{\partial Q}{\partial \beta}([Z_1, \beta], Z_2.) = \tilde{\Theta}(Z_1, [\beta, Z_2]) + \left\langle Q, Ad_{.z_1}([\beta, Z_2])\right\rangle = \tilde{\Theta}_\beta(Z_1, [\beta, Z_2]) \quad (153)$$

$$\Rightarrow -\frac{\partial Q}{\partial \beta} = g_\beta([\beta, Z_1], [\beta, Z_2]) \quad (154)$$

The Souriau Fisher metric $I(\beta) = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2} = -\frac{\partial Q}{\partial \beta}$ has been considered by Souriau as a *generalization of* "*heat capacity*". Souriau called it the "*geometric capacity*" and is also equal to "*geometric susceptibility*".

7 Conclusion

The community of "*Geometric Science of Information*" (GSI) has lost a mathematician of great value, who informed his views by the depth of his knowledge of the elementary structures of hessian geometry and bounded homogeneous domains. His modesty was inversely proportional to his talent. Professor Koszul built in over 60 years of mathematical career, in the silence of his passions, an immense work, which makes him one of the great mathematicians of the XX's century, whose importance will only affirm with the time. In this troubled time and rapid transformation of society and science, the example of Professor Koszul must be regarded as a model for future generations, to avoid them the trap of fleeting glories and recognitions too fast acquired. The work of Professor Koszul is also a proof of fidelity to his masters

Fig. 13 (on the left) Jean-Louis Koszul at Grenoble in December 1993, (on the right) last interview of Jean-Louis Koszul in 2016 for 50th birthday of Institut Joseph Fourier in Grenoble

and in the first place to Prof. Elie Cartan, who inspired him throughout his life. Henri Cartan writes on this subject "*I do not forget the homage he paid to Elie Cartan's work in Differential Geometry during the celebration, in Bucharest, in 1969, of the centenary of his birth. It is not a coincidence that this centenary was also celebrated in Grenoble the same year. As always, Koszul spoke with the discretion and tact that we know him, and that we love so much at home*". I will conclude by quoting Jorge Luis Borges "*Forgetfulness and memory are also inventive*" (Brodie's report). Our generation and previous one have forgotten or misunderstood the depth of the work of Jean-Louis Koszul and Elie Cartan on the study of bounded homogeneous domains. It is our responsibility to correct this omission, and to make it the new inspiration for the Geometric Science of Information. I will conclude by requesting you to listen to the last interview of Jean-Louis Koszul for 50th birthday of Joseph Fourier Institute [72], especially when Koszul he is passionate by "*conifers and cedars trees planted by Claude Chabauty*", or by the "*pretty catalpa tree*" which was at the Fourier Institute and destroyed by wind, "*the tree with parentheses*" he says, to which he seemed to be sentimentally attached. He also regrets that the Institute did not use the 1% artistic fund for the art mosaic project in the library. In this Koszul family of mathematicians, musicians, and Scientifics, there was a constant recollection of "beauty" and "truth". Our society no longer cares about timeless "beauty". We have then to extase ourself with Jean-Louis Koszul by observing beautiful "*Catalpa tree*" with "*Parenthese Mushroom*", before there is no longer people to contemplate them (Fig. 13).

> "*Seul la nuit avec un livre éclairé par une chandelle – livre et chandelle, double îlot de lumière, contre les doubles ténèbres de l'esprit et de la nuit. J'étudie ! Je ne suis que le sujet du verbe étudier. Penser je n'ose. Avant de penser, il faut étudier. Seuls les philosophes pensent avant d'étudier.* » - **Gaston Bachelard, La flamme d'une chandelle, 1961**

Appendix

Clairaut(-Legendre) Equation of Maurice Fréchet associated to "distinguished functions" as fundamental equation of Information geometry

Before Rao [4, 124], in 1943, Maurice Fréchet [3] wrote a seminal paper introducing what was then called the Cramer-Rao bound. This paper contains in fact much more that this important discovery. In particular, Maurice Fréchet introduces more general notions relative to "*distinguished functions*", densities with estimator reaching the bound, defined with a function, solution of Clairaut's equation. The solutions "envelope of the Clairaut's equation" are equivalents to standard Legendre transform without convexity constraints but only smoothness assumption. This Fréchet's analysis can be revisited on the basis of Jean-Louis Koszul's works as seminal foundation of "*Information Geometry*".

I will use Maurice Fréchet notations, to consider the estimator:

$$T = H(X_1, \ldots, X_n) \tag{155}$$

and the random variable

$$A(X) = \frac{\partial \log p_\theta(X)}{\partial \theta} \tag{156}$$

that are associated to:

$$U = \sum_i A(X_i) \tag{157}$$

The normalizing constraint $\int_{-\infty}^{+\infty} p_\theta(x)dx = 1$ implies that: $\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \prod_i p_\theta(x_i)dx_i = 1$

If we consider the derivative if this last expression with respect to θ, then $\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \left[\sum_i A(x_i)\right] \prod_i p_\theta(x_i)dx_i = 0$ gives:

$$E_\theta[U] = 0 \tag{158}$$

Similarly, if we assume that $E_\theta[T] = \theta$, then $\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} H(x_1, \ldots, x_n) \prod_i p_\theta(x_i)dx_i = \theta$, and we obtain by derivation with respect to θ:

$$E[(T - \theta)U] = 1 \tag{159}$$

But as $E[T] = \theta$ and $E[U] = 0$, we immediatly deduce that:

$$E[(T - E[T])(U - E[U])] = 1 \tag{160}$$

From Schwarz inequality, we can develop the following relations:

$$\begin{aligned} &[E(ZT)]^2 \leq E\left[Z^2\right]E\left[T^2\right] \\ &1 \leq E\left[(T - E[T])^2\right]E\left[(U - E[U])^2\right] = (\sigma_T \sigma_U)^2 \end{aligned} \tag{161}$$

U being the summation of independent variables, Bienaymé equality could be applied:

$$(\sigma_U)^2 = \sum_i \left[\sigma_{A(X_i)}\right]^2 = n(\sigma_A)^2 \tag{162}$$

From which, Fréchet deduced the bound, rediscoved by Cramer and Rao 2 years later:

$$(\sigma_T)^2 \geq \frac{1}{n(\sigma_A)^2} \tag{163}$$

Fréchet observed that it is a remarkable inequality where the second member is independent of the choice of the function H defining the "*empirical value*" T, where the first member can be taken to any empirical value $T = H(X_1, \ldots, X_n)$ subject to the unique condition $E_\theta[T] = \theta$ regardless is θ.

The classic condition that the Schwarz inequality becomes an equality helps us to determine when σ_T reaches its lower bound $\frac{1}{\sqrt{n}\sigma_n}$.

The previous inequality becomes an equality if there are two numbers α and β (not random and not both zero) such that $\alpha\left(H' - \theta\right) + \beta U = 0$, with H' particular function among eligible H as we have the equality. This equality is rewritten $H' = \theta + \lambda' U$ with λ' a non-random number.

If we use the previous equation, then:

$$E[(T - E[T])(U - E[U])] = 1 \Rightarrow E\left[\left(H' - \theta\right)U\right] = \lambda' E_\theta\left[U^2\right] = 1 \tag{164}$$

We obtain:

$$U = \sum_i A(X_i) \Rightarrow \lambda' n E_\theta\left[A^2\right] = 1 \tag{165}$$

From which we obtain λ' and the form of the associated estimator H':

$$\lambda' = \frac{1}{nE\left[A^2\right]} \Rightarrow H' = \theta + \frac{1}{nE\left[A^2\right]} \sum_i \frac{\partial \log p_\theta(X_i)}{\partial \theta} \tag{166}$$

It is therefore deduced that the estimator that reaches the terminal is of the form:

$$H' = \theta + \frac{\sum\limits_{i} \frac{\partial \log p_\theta(X_i)}{\partial \theta}}{n \int\limits_{-\infty}^{+\infty} \left[\frac{\partial p_\theta(x)}{\partial \theta}\right]^2 \frac{dx}{p_\theta(x)}} \tag{167}$$

with $E\left[H'\right] = \theta + \lambda' E[U] = \theta$.

H' **would be one of the eligible functions, if** H' **would be independent of** θ. Indeed, if we consider:

$$E_{\theta_0}\left[H'\right] = \theta_0,\ E\left[\left(H' - \theta_0\right)^2\right] \le E_{\theta_0}\left[(H - \theta_0)^2\right] \forall H \text{ such that } E_{\theta_0}[H] = \theta_0 \tag{168}$$

$H = \theta_0$ satisfies the equation and inequality shows that it is almost certainly equal to θ_0. So to look for θ_0, we should know beforehand θ_0.

At this stage, Fréchet looked for "*distinguished functions*" ("*densités distinguées*" in French), as any probability density $p_\theta(x)$ such that the function:

$$h(x) = \theta + \frac{\frac{\partial \log p_\theta(x)}{\partial \theta}}{\int\limits_{-\infty}^{+\infty} \left[\frac{\partial p_\theta(x)}{\partial \theta}\right]^2 \frac{dx}{p_\theta(x)}} \tag{169}$$

is independant of θ. The objective of Fréchet is then to determine the minimizing function $T = H'(X_1, \ldots, X_n)$ that reaches the bound. By deduction from previous relations, we have:

$$\lambda(\theta) \frac{\partial \log p_\theta(x)}{\partial \theta} = h(x) - \theta \tag{170}$$

But as $\lambda(\theta) > 0$, ***we can consider*** $\frac{1}{\lambda(\theta)}$ ***as the second derivative of a function*** $\Phi(\theta)$ such that:

$$\frac{\partial \log p_\theta(x)}{\partial \theta} = \frac{\partial^2 \Phi(\theta)}{\partial \theta^2} [h(x) - \theta] \tag{171}$$

from which we deduce that:

$$\ell(x) = \log p_\theta(x) - \frac{\partial \Phi(\theta)}{\partial \theta} [h(x) - \theta] - \Phi(\theta) \tag{172}$$

Is an independant quantity of θ. ***A distinguished function*** will be then given by:

$$p_\theta(x) = e^{\frac{\partial \Phi(\theta)}{\partial \theta} [h(x) - \theta] + \Phi(\theta) + \ell(x)} \tag{173}$$

with the normalizing constraint $\int\limits_{-\infty}^{+\infty} p_\theta(x) dx = 1$.

These two conditions are sufficient. Indeed, reciprocally, let three functions $\Phi(\theta)$, $h(x)$ et $\ell(x)$ that we have, for any θ:

$$\int_{-\infty}^{+\infty} e^{\frac{\partial\Phi(\theta)}{\partial\theta}[h(x)-\theta]+\Phi(\theta)+\ell(x)}dx = 1 \tag{174}$$

Then the function is distinguished:

$$\theta + \frac{\frac{\partial \log p_\theta(x)}{\partial\theta}}{\int_{-\infty}^{+\infty}\left[\frac{\partial p_\theta(x)}{\partial\theta}\right]^2\frac{dx}{p_\theta(x)}} = \theta + \lambda(x)\frac{\partial^2\Phi(\theta)}{\partial\theta^2}[h(x)-\theta] \tag{175}$$

If $\lambda(x)\frac{\partial^2\Phi(\theta)}{\partial\theta^2} = 1$, when

$$\frac{1}{\lambda(x)} = \int_{-\infty}^{+\infty}\left[\frac{\partial \log p_\theta(x)}{\partial\theta}\right]^2 p_\theta(x)dx = (\sigma_A)^2 \tag{176}$$

The function is reduced to $h(x)$ and then is not dependent of θ.
We have then the following relation:

$$\frac{1}{\lambda(x)} = \int_{-\infty}^{+\infty}\left(\frac{\partial^2\Phi(\theta)}{\partial\theta^2}\right)^2[h(x)-\theta]^2 e^{\frac{\partial\Phi(\theta)}{\partial\theta}(h(x)-\theta)+\Phi(\theta)+\ell(x)}dx \tag{177}$$

The relation is valid for any θ, we can derive the previous expression from θ:

$$\int_{-\infty}^{+\infty} e^{\frac{\partial\Phi(\theta)}{\partial\theta}(h(x)-\theta)+\Phi(\theta)+\ell(x)}\left(\frac{\partial^2\Phi(\theta)}{\partial\theta^2}\right)[h(x)-\theta]dx = 0 \tag{178}$$

We can divide by $\frac{\partial^2\Phi(\theta)}{\partial\theta^2}$ because it doesn't depend on x.
If we derive again with respect to θ, we will have:

$$\int_{-\infty}^{+\infty} e^{\frac{\partial\Phi(\theta)}{\partial\theta}(h(x)-\theta)+\Phi(\theta)+\ell(x)}\left(\frac{\partial^2\Phi(\theta)}{\partial\theta^2}\right)[h(x)-\theta]^2dx = \int_{-\infty}^{+\infty} e^{\frac{\partial\Phi(\theta)}{\partial\theta}(h(x)-\theta)+\Phi(\theta)+\ell(x)}dx = 1 \tag{179}$$

Combining this relation with that of $\frac{1}{\lambda(x)}$, we can deduce that $\lambda(x)\frac{\partial^2\Phi(\theta)}{\partial\theta^2} = 1$ and as $\lambda(x) > 0$ then $\frac{\partial^2\Phi(\theta)}{\partial\theta^2} > 0$.

Fréchet emphasizes at this step, another way to approach the problem. We can select arbitrarily $h(x)$ and $l(x)$ and then $\Phi(\theta)$ is determined by:

$$\int_{-\infty}^{+\infty} e^{\frac{\partial\Phi(\theta)}{\partial\theta}[h(x)-\theta]+\Phi(\theta)+\ell(x)}dx = 1 \tag{180}$$

That could be rewritten:

$$e^{\theta.\frac{\partial\Phi(\theta)}{\partial\theta}-\Phi(\theta)} = \int_{-\infty}^{+\infty} e^{\frac{\partial\Phi(\theta)}{\partial\theta}h(x)+\ell(x)}dx \tag{181}$$

If we then fixed arbitrarily $h(x)$ and $l(x)$ and let s an arbitrary variable, the following function will be an explicit positive function given by $e^{\Psi(s)}$:

$$\int_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}dx = e^{\Psi(s)} \tag{182}$$

Fréchet obtained finally the function $\Phi(\theta)$ as solution of the equation:

$$\Phi(\theta) = \theta.\frac{\partial\Phi(\theta)}{\partial\theta} - \Psi\left(\frac{\partial\Phi(\theta)}{\partial\theta}\right) \tag{183}$$

Fréchet noted that this is the Alexis Clairaut Equation.

The case $\frac{\partial\Phi(\theta)}{\partial\theta} = cste$ would reduce the density to a function that would be independent of θ, and so $\Phi(\theta)$ is given by a singular solution of this Clairaut equation, that is unique and could be computed by eliminating the variable s between:

$$\Phi = \theta.s - \Psi(s) \,\text{and}\, \theta = \frac{\partial\Psi(s)}{\partial s} \tag{184}$$

Or between:

$$e^{\theta.s-\Phi(\theta)} = \int_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}dx \text{ and } \int_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}[h(x)-\theta]dx = 0 \tag{185}$$

$\Phi(\theta) = -\log \int_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}dx + \theta.s$ where s is induced implicitly through the constraint given by the integral $\int_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}[h(x)-\theta]dx = 0$.

When we known the distinguished function, H' is among functions $H(X_1, \ldots, X_n)$ verifying $E_\theta[H] = \theta$ and such that σ_H reaches for each value of θ, an absolute minimum, equal to $\frac{1}{\sqrt{n}\sigma_A}$. For the previous equation:

$$h(x) = \theta + \frac{\frac{\partial \log p_\theta(x)}{\partial \theta}}{\int\limits_{-\infty}^{+\infty} \left[\frac{\partial p_\theta(x)}{\partial \theta} \right]^2 \frac{dx}{p_\theta(x)}} \tag{186}$$

We can rewrite the estimator as:

$$H'(X_1, \ldots, X_n) = \frac{1}{n}[h(X_1) + \ldots + h(X_n)] \tag{187}$$

and compute the associated empirical value:

$$t = H'(x_1, \ldots, x_n) = \frac{1}{n} \sum_i h(x_i) = \theta + \lambda(\theta) \sum_i \frac{\partial \log p_\theta(x_i)}{\partial \theta} \tag{188}$$

If we take $\theta = t$, we have as $\lambda(\theta) > 0$:

$$\sum_i \frac{\partial \log p_t(x_i)}{\partial t} = 0 \tag{189}$$

When $p_\theta(x)$ is a distinguished function, the empirical value t of θ corresponding to a sample $x_1, \ldots, x_n$ is a root of previous equation in t. This equation has a root and only one when X is a distinguished variable. Indeed, as we have:

$$p_\theta(x) = e^{\frac{\partial \Phi(\theta)}{\partial \theta}[h(x) - \theta] + \Phi(\theta) + \ell(x)} \tag{190}$$

$$\sum_i \frac{\partial \log p_t(x_i)}{\partial t} = \frac{\partial^2 \Phi(t)}{\partial t^2} \left[\frac{\sum\limits_i h(x_i)}{n} - t \right] \text{with } \frac{\partial^2 \Phi(t)}{\partial t^2} > 0 \tag{191}$$

We can then recover the unique root: $t = \frac{\sum\limits_i h(x_i)}{n}$.

This function $T \equiv H'(X_1, \ldots, X_n) = \frac{1}{n} \sum\limits_i h(X_i)$ can have an arbitrary form, that is a sum of functions of each only one of the quantities and it is even the arithmetic average of N values of a same auxiliary random variable $Y = h(X)$. The dispersion is given by:

$$\left(\sigma_{T_n}\right)^2 = \frac{1}{n(\sigma_A)^2} = \frac{1}{n \int\limits_{-\infty}^{+\infty} \left[\frac{\partial p_\theta(x)}{\partial \theta} \right]^2 \frac{dx}{p_\theta(x)}} = \frac{1}{n \frac{\partial^2 \Phi(\theta)}{\partial \theta^2}} \tag{192}$$

and T_n follows the probability density:

$$p_\theta(t) = \sqrt{n} \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{n(t-\theta)^2}{2.\sigma_A^2}} \quad \text{with} \quad (\sigma_A)^2 = \frac{\partial^2 \Phi(\theta)}{\partial \theta^2} \tag{193}$$

- **Clairaut Equation and Legendre Transform**

To summarize, Fréchet paper novelty, I have just observed that Fréchet introduced distinguished functions depending on a function $\Phi(\theta)$, solution of the Clairaut equation:

$$\Phi(\theta) = \theta.\frac{\partial\Phi(\theta)}{\partial\theta} - \Psi\left(\frac{\partial\Phi(\theta)}{\partial\theta}\right) \tag{194}$$

Or given by the Legendre Transform:

$$\Phi = \theta.s - \Psi(s)\,\text{and}\,\theta = \frac{\partial\Psi(s)}{\partial s} \tag{195}$$

Fréchet also observed that this function $\Phi(\theta)$ could be rewritten: $\Phi(\theta) = -\log \int\limits_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}dx + \theta.s$ where s is induced implicitly by the constraints given by integral $\int\limits_{-\infty}^{+\infty} e^{s.h(x)+\ell(x)}[h(x) - \theta]dx = 0$.

This equation is the fundamental equation of Information Geometry.

The "Legendre" transform was introduced by Adrien-Marie Legendre in 1787 to solve a minimal surface problem Gaspard Monge in 1784. Using a result of Jean Baptiste Meusnier, a student of Monge, it solves the problem by a change of variable corresponding to the transform which now entitled with his name. Legendre wrote: "*I have just arrived by a change of variables that can be useful in other occasions.*" About this transformation, Darboux in his book gives an interpretation of Chasles: "*This comes after a comment by Mr. Chasles, to substitute its polar reciprocal on the surface compared to a paraboloïd.*" The equation of Clairaut was introduced 40 years earlier in 1734 by Alexis Clairaut. Solutions "envelope of the Clairaut equation" are equivalent to the Legendre transform with unconditional convexity, but only under differentiability constraint. Indeed, for a non-convex function, Legendre transformation is not defined where the Hessian of the function is canceled, so that the equation of Clairaut only make the hypothesis of differentiability. The portion of the strictly convex function g in Clairaut equation $y = px - g\ (p)$ to the function f giving the envelope solutions by the formula $y = f\ (x)$ is precisely the Legendre transformation. The approach of Fréchet may be reconsidered in a more general context on the basis of the work of Jean-Louis Koszul.

References

1. Ollivier, Y., Marceau-Caron, G.: Natural Langevin dynamics for neural networks. In: Nielsen, F., Barbaresco, F. (eds.) Proceedings of the Conference on Geometric Science of Information (GSI 2017). Lecture Notes in Computer Science, vol. 10589, pp. 451–459. Springer, Berlin (2017) (Best paper award.)
2. Ollivier, Y.: True asymptotic natural gradient optimization manuscript disponible sur. http://arxiv.org/abs/1712.08449
3. Fréchet, M.: Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. Rev. l'Institut Int. Stat. **11**(3/4), 182–205 (1943)
4. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–89 (1945)
5. Chentsov, N.N.: Statistical Decision Rules and Optimal Inferences, Transactions of Mathematics Monograph, vol. 53. American Mathematical Society, Providence, RI, USA, (1982)
6. Amari, S.I.: Differential Geometry of Statistical Models. SPRINGER Series Lecture Notes in Statistics, vol. 28 (1985)
7. Amari, S.I.: Information Geometry and Its Applications. SPRINGER Series Applied Mathematical Sciences, vol. 194 (2016)
8. Weinhold, F.: Thermodynamics and geometry. Phys. Today (1976)
9. Ruppeiner, G.: Thermodynamics: a Riemannian geometric model. Phys. Rev. A **20**, 1608 (1979)
10. Ingarden, R.S.: Information geometry in functional spaces of classical and quantum finite statistical systems. Int. J. Eng. Sci. **19**(12), 1609–1633 (1981)
11. Ingarden, R.S., Kossakowski, A., Ohya, M.: Information Dynamics and Open Systems: Classical and Quantum Approach. Springer, Berlin (1997). Accessed 31 March 1997
12. Mrugala, R.: Geometrical formulation of equilibrium phenomenological thermodynamics. Rep. Math. Phys. **14**, 419 (1978)
13. Mrugala, R., Nulton, J.D., Schön, J.C., Salamon, P.: Statistical approach to the geometric structure of thermodynamics. Phys. Rev. A **41**, 6, 3156 (1990)
14. Janyszek, H., Mrugala, R.: Riemannian geometry and the thermodynamics of model magnetic systems. Phys. Rev. A **39**(12), :6515–6523 (1989)
15. Koszul, J.L.: Sur la forme hermitienne canonique des espaces homogènes complexes. Can. J. Math. **7**, 562–576 (1955)
16. Koszul, J.L.: Exposés sur les Espaces Homogènes Symétriques. Publicação da Sociedade de Matematica de São Paulo, São Paulo, Brazil (1959)
17. Souriau, J.-M.: Structures des Systèmes Dynamiques. Dunod, Paris (1969)
18. Koszul, J.L.: Introduction to Symplectic Geometry. Science Press, Beijing (1986) (in chinese); translated by SPRINGER, with F.Barbaresco, C.M. Marle and M. Boyom forewords, SPRINGER, 2018
19. Balian, R., Alhassid, Y., Reinhardt, H.: Dissipation in many-body systems: a geometric approach based on information theory. Phys. Rep. **131**, 1–146 (1986)
20. Balian, R.: The entropy-based quantum metric. Entropy **16**, 3878–3888 (2014)
21. Selected Papers of J L Koszul, Series in Pure Mathematics, Volume 17, World Scientific Publishing, 1994
22. Koszul, J.L.: L'œuvre d'Élie Cartan en géométrie différentielle, in Élie Cartan, 1869–1951. Hommage de l'Académie de la République Socialiste de Roumanie à l'occasion du centenaire de sa naissance. Comprenant les communications faites aux séances du 4e Congrès du Groupement des Mathématiciens d'Expression Latine, tenu à Bucarest en 1969 (Editura Academiei Republicii Socialiste Romania, Bucharest, 1975) pp. 39–45
23. Vinberg, E.B.: Homogeneous cones. Dokl. Akad. Nauk SSSR. **133**, 9–12 (1960); Sov. Math. Dokl. **1**, 787–790 (1961)
24. Vinberg, E.B.: The Morozov-Borel theorem for real Lie groups. Dokl. Akad. Nauk SSSR. **141**, 270–273 (1961); Sov. Math. Dokl. **2**, 1416–1419 (1962)

25. Vinberg E.B., Convex homogeneous domains, Dokl. Akad. Nauk SSSR., 141 1961, 521–524; Soviet Math. Dokl., n°2, pp. 1470–1473, 1962
26. Vinberg, E.B.: Automorphisms of homogeneous convex cones. Dokl. Akad. Nauk SSSR. **143**, 265–268 (1962); Sov. Math. Dokl. **3**, 371–374 (1963)
27. Vinberg, E.B.: The Theory of Homogeneous Convex Cones, Trudy Moskovskogo Matematicheskogo Obshchestva, vol. 12, pp. 303–358 (1963); Trans. Moscow Math. Soc. **12**, 340–403 (1963)
28. Vinberg, E.B., Gindikin, S.G., Pyatetskii-Shapiro, I.I.: On the classification and canonical realization of bounded homogeneous domains. Trudy Moskov. Mat. Obshch. **12**, 359–388 (1963); Trans. Moscow Math. Soc. **12**, 404–437 (1963)
29. Vinberg, E.B.: The structure of the group of automorphisms of a convex cone. Trudy Moscov. Mat. Obshch. **13**, 56–83 (1964); Trans. Moscow Math. Soc. **13** (1964)
30. Vinberg, E.B.: Structure of the group of automorphisms of a homogeneous convex cone. Tr. Mosk. Mat. Obshchestva **13**, 56–83 (1965)
31. Pyatetskii-Shapiro, I.I.: Certain problems of harmonic analysis in homogeneous cones. Dokl. Akad. Nauk SSSR. 181–184 (1957)
32. Pyatetskii-Shapiro, I.I.: On a problem of E. Cartan Dokl. Akad. Nauk SSSR. **124**, 272–273 (1959)
33. Pyatetskii-Shapiro, I.I.: The geometry of homogeneous domains and the theory of automorphic functions, the solution of a problem of E. Cartan. Uspekhi Mat. Nauk. **14**(3), 190–192 (1959)
34. Pyatetskii-Shapiro, I.I.: On the classification of bounded homogeneous domains in n-dimensional complex space. Dokl. Akad. Nauk SSSR. **141**, 316–319 (1961); Sov. Math. Dokl. **141**, 1460–1463 (1962)
35. Pyatetskii-Shapiro, I.I.: On bounded homogeneous domains in n-dimensional complex space. Izv. Akad. Nauk SSSR. Ser. Mat. **26**, 107–124 (1962)
36. Gindikin, S.G.: Analysis in homogeneous domains. Uspekhi Mat. Nauk. **19**(4), 3–92 (1964); Russ. Math. Surv. **19**(4), 1–89 (1964)
37. Cartan, E.: Sur les domaines bornés de l'espace de n variables complexes. Abh. Math. Semin. Hambg. **1**, 116–162 (1935)
38. Berger, M.: Les espaces symétriques noncompacts. Ann. Sci. l'École Norm. Supérieure Sér. 3 Tome **74**(2), 85–177 (1957)
39. Vey, J.: Sur une notion d'hyperbolicité des variétés localement plates, Thèse de troisième cycle de mathématiques pures, Faculté des sciences de l'université de Grenoble (1969)
40. Vey, J.: Sur les automorphismes affines des ouverts convexes saillants, Annali della Scuola Normale Superiore di Pisa, Classe di Science, 3e série, tome **24**(4), 641–665 (1970)
41. Koszul, J.L.: Variétés localement plates et convexité. Osaka. J. Math. **2**, 285–290 (1965)
42. Alekseevsky, D.: Vinberg's Theory of Homogeneous Convex Cones: Developments and Applications, Transformation Groups 2017. Conference Dedicated to Prof. Ernest B. Vinberg on the occasion of his 80th birthday, Moscou, December (2017). https://www.mccme.ru/tg2017/slides/alexeevsky.pdf, http://www.mathnet.ru/present19121
43. Lichnerowicz, A.: Espaces homogènes Kähleriens. In: Colloque de Géométrie Différentielle. Publication du CNRSP, Paris, France, pp. 171–184 (1953)
44. Siegel, C.L.: Symplectic geometry. Am. J. Math. **65**, 1–86 (1943)
45. Koszul, J.L.: Domaines bornées homogènes et orbites de groupes de transformations affines. Bull. Soc. Math. France **89**, 515–533 (1961)
46. Koszul, J.L.: Ouverts convexes homogènes des espaces affines. Math. Z. **79**, 254–259 (1962)
47. Koszul, J.L.: Lectures on Groups of Transformations. Tata Institute of Fundamental Research, Bombay (1965)
48. Siegel, C.L.: Über der analytische Theorie der quadratischen Formen. Ann. Math. **36**, 527–606 (1935)
49. Koszul, J.L.: Déformations des variétés localement plates. Ann Inst Fourier **18**, 103–114 (1968)

50. Koszul, J.L.: Trajectoires convexes de groupes affines unimodulaires. In: Essays on Topology and Related Topics, pp. 105–110. Springer, Berlin (1970)
51. Shima, H.: Symmetric spaces with invariant locally Hessian structures. J. Math. Soc. Jpn. 581–589 (1977)
52. Shima, H.: Homogeneous Hessian manifolds. Ann. Inst. Fourier 91–128 (1980)
53. Shima, H.: Vanishing theorems for compact Hessian manifolds. Ann. Inst. Fourier 183–205 (1986)
54. Shima, H.: Harmonicity of gradient mappings of level surfaces in a real affine space. Geom. Dedicata 177–184 (1995)
55. Shima, H.: Hessian manifolds of constant Hessian sectional curvature. J. Math. Soc. Jpn. 735–753 (1995)
56. Shima, H.: Homogeneous spaces with invariant projectively flat affine connections. Trans. Am. Math. Soc. 4713–4726 (1999)
57. Shima, H.: The Geometry of Hessian Structures. World Scientific, Singapore (2007)
58. Shima, H.: In: Nielsen, F., Frederic, B. (eds.) Geometry of Hessian Structures. Springer Lecture Notes in Computer Science, vol. 8085, pp. 37–55 (2013) (planches: https://www.see.asso.fr/file/5104/download/25050), (vidéos GSI'13: https://www.youtube.com/watch?time_continue=139&v=6pyXxdIzDNQ, https://www.youtube.com/watch?time_continue=182&v=jG2tUjniOUs, https://www.youtube.com/watch?time_continue=6&v=I5kdMJvuNHA)
59. Boyom, M.N.: Sur les structures affines homotopes à zéro des groupes de Lie. J. Differ. Geom. **31**, 859–911 (1990)
60. Boyom, M.N.: Structures localement plates dans certaines variétés symplectiques. Math. Scand. **76**, 61–84 (1995)
61. Boyom, M.N.: Métriques kählériennes affinement plates de certaines variétés symplectiques. I Proc. Lond. Math. Soc. 3 **66**(2), 358–380 (1993)
62. Boyom, M.N.: The cohomology of Koszul-Vinberg algebras. Pac. J. Math. **225**, 119–153 (2006)
63. Boyom, M.N.: Some Lagrangian Invariants of Symplectic Manifolds, Geometry and Topology of Manifolds; Banach Center Institute of Mathematics, Polish Academy of Sciences, Warsaw, vol. 76, pp. 515–525 (2007)
64. Boyom, M.N., Byande, P.M.: KV Cohomology in Information Geometry Matrix Information Geometry, pp. 69–92. Springer, Heidelberg, Germany (2013)
65. Boyom, M.N.: Transversally Hessian foliations and information geometry I. Am. Inst. Phys. Proc. **1641**, 82–89 (2014)
66. Boyom, M.N., Wolak, R.: Transverse Hessian metrics information geometry MaxEnt 2014. In: AIP. Conference Proceedings American Institute of Physics (2015)
67. Byande, P.M., Ngakeu, F., Boyom, M.N., Wolak, R.: KV-cohomology and differential geometry of affinely flat manifolds. Information geometry. Afr. Diaspora J. Math. **14**, 197–226 (2012)
68. Byande, P.M.: Des Structures Affines à la Géométrie de L'information; European University Editions (2012)
69. Lichnerowicz, A.: Groupes de Lie à structures symplectiques ou Kähleriennes invariantes. In: Albert, C. (ed.) Géométrie Symplectique et Mécanique. Lecture Notes in Mathematics, vol 1416. Springer, Berlin (1990)
70. Cartan, H.: Allocution de Monsieur Henri Cartan, colloques Jean-Louis Koszul. Ann. l'Institut Fourier **37**(4), 1–4 (1987)
71. Malgrange, B.: Quelques souvenirs de Jean-Louis KOSZUL. Gazette des Mathématiciens **156**, 63–64 (2018)
72. Koszul, J.L.: Interview for "Institut Joseph Fourier" 50th birthday in 2016. https://www.youtube.com/watch?v=AzK5K7Q05sw
73. Barbaresco, F.: Jean-Louis Koszul et les structures élémentaires de la géométrie de l'information, revue MATAPLI 2018, SMAI (2018)
74. Koszul, J.M.: Homologie et cohomologie des algèbres de Lie. Bull. Soc. Math. Fr. Tome **78**, 65–127 (1950)

75. Cartan, H.: Les travaux de Koszul, I, Séminaire Bourbaki, Tome 1, Exposé **1**, 7–12 (1948–1951)
76. Cartan, H.: Les travaux de Koszul, II, Séminaire Bourbaki, Tome 1, Exposé **8**, 45–52 (1948–1951)
77. Cartan, H.: Les travaux de Koszul, III, Séminaire Bourbaki, Tome 1, Exposé **12**, 71–74 (1948–1951)
78. Haefliger, A.: Des espaces homogènes à la résolution de Koszul. Ann. l'inst. Fourier Tome **37**(4), 5–13 (1987)
79. Souriau, J.-M.: Mécanique statistique, groupes de Lie et cosmologie, Colloques int. du CNRS numéro 237, Géométrie symplectique et physique mathématique, pp. 59–113 (1974)
80. Barbaresco F., Koszul information geometry and Souriau Lie group thermodynamics. AIP Conference Proceedings 1641, 74, 2015, Proceedings of MaxEnt'14 Conference, Amboise, Septembre 2014
81. Barbaresco, F.: Koszul Information Geometry and Souriau Geometric Temperature/Capacity of Lie Group Thermodynamics, Entropy, vol. 16, pp. 4521–4565 (2014); Published in the book Information, Entropy and Their Geometric Structures, MDPI Publisher, September 2015
82. Barbaresco, F.: Symplectic structure of information geometry: fisher metric and Euler-Poincaré equation of Souriau Lie group thermodynamics. In: Geometric Science of Information, Second International Conference GSI 2015 Proceedings. Lecture Notes in Computer Science vol. 9389, pp. 529–540. Springer, Berlin (2015)
83. Barbaresco, F.: Geometric theory of heat from souriau lie groups thermodynamics and koszul hessian geometry: applications in information geometry for exponential families. Entropy **18**, 386 (2016)
84. Barbaresco, F.: Poly-symplectic Model of Higher Order Souriau Lie Groups Thermodynamics for Small Data Analytics, GSI 2017. Springer LNCS, vol. 10589, pp. 432–441 (2017)
85. Cartan, E.: Sur les invariants intégraux de certains espaces homogènes clos et les propriétés topologiques de ces espaces. Ann. Soc. Pol. De Math. **8**, 181–225 (1929)
86. Hua, L.K.: Harmonic Analysis of Functions of Several Complex Variables in the Classical Domains. American Mathematical Society, Providence, USA (1963)
87. Berezin, F.: Quantization in complex symmetric spaces. Izv. Akad. Nauk SSSR Ser. Math. **9**, 363–402 (1975)
88. Maliavin, P.: Invariant or quasi-invariant probability measures for infinite dimensional groups, Part II: unitarizing measures or Berezinian measures. Jpn. J. Math. **3**, 19–47 (2008)
89. Rothaus, O.S.: The Construction of Homogeneous Convex Cones. Annals of Mathematics, Series 2, vol. 83, pp. 358–376 (1966)
90. Vesentini, E.: Geometry of Homogeneous Bounded Domains. Springer, Berlin (2011); reprint of the 1st edn. C.IM.E., Ed. Cremonese, Roma 1968
91. Sampieri, U.: A generalized exponential map for an affinely homogeneous cone. Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni, Serie 8. **75(**6), 320–330 (1983)
92. Sampieri, U.: Lie group structures and reproducing kernels on homogeneous siegel domains. Ann. Mat. **152**(1), 1–19 (1988)
93. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM Series: Studies in Applied and Numerical Mathematics, pp. ix + 396 (1994)
94. Gromov, M.: Convex sets and kähler manifolds. In: Tricerri, F. (ed.) Advances in Differential Geometry and Topology, pp. 1–38. World Scientific, Singapore (1990)
95. Gromov, M.: In a Search for a Structure, Part 1: On Entropy. http://www.ihes.fr/~gromov/PDF/structre-serch-entropy-july5-2012.pdf (2012). Accessed 23 June 2012
96. Gromov, M.: Gromov Six Lectures on Probability, Symmetry, Linearity. October 2014, Jussieu, November 6th, 2014; Lecture Slides & video of Gromov lectures on youtube: http://www.ihes.fr/~gromov/PDF/probability-huge-Lecture-Nov-2014.pdf, https://www.youtube.com/watch?v=hb4D8yMdov4
97. Gromov, M.: Gromov Four Lectures on Mathematical Structures Arising from Genetics and Molecular Biology, IHES, October 2013; video of Lectures on youtube: https://www.youtube.com/watch?v=v7QuYuoyLQc&t=5935s

98. Barbaresco, F.: Les densités de probabilité « distinguées » et l'équation d'Alexis Clairaut: regards croisés de Maurice Fréchet et de Jean-Louis Koszul, GRETSI'17, Juan-Les-Pins (2017)
99. Massieu, F.: Sur les Fonctions caractéristiques des divers fluides. C. R. Acad. Sci. **69**, 858–862 (1869)
100. Massieu, F.: Addition au précédent Mémoire sur les Fonctions caractéristiques. C. R. Acad. Sci. **69**, 1057–1061 (1869)
101. Massieu, F.: Exposé des Principes Fondamentaux de la Théorie Mécanique de la Chaleur (note Destinée à Servir D'introduction au Mémoire de L'auteur sur les Fonctions Caractéristiques des Divers Fluides et la Théorie des Vapeurs); Académie des Sciences, Paris, 31 p. (1873)
102. Massieu, F.: Thermodynamique: Mémoire sur les Fonctions Caractéristiques des Divers Fluides et sur la Théorie des Vapeurs. Académie des Sciences, Paris, France, p. 92 (1876)
103. Poincaré, H.: Calcul des Probabilités. Gauthier-Villars, Paris, France (1896)
104. Balian, R., Balazs, N.: Equiprobability, inference and entropy in quantum theory. Ann. Phys. **179**, 97–144 (1987)
105. Balian, R.: On the principles of quantum mechanics. Am. J. Phys. **57**, 1019–1027 (1989)
106. Balian, R.: From Microphysics to Macrophysics: Methods and Applications of Statistical Physics, Vols. I and II. Springer, Heidelberg, Germany (1991, 1992)
107. Balian, R.: Incomplete descriptions and relevant entropies. Am. J. Phys. **67**, 1078–1090 (1999)
108. Balian, R., Valentin, P.: Hamiltonian structure of thermodynamics with gauge. Eur. Phys. J. B **21**, 269–282 (2001)
109. Balian, R.: Entropy, a protean concept. In: Dalibard, J., Duplantier, B., Rivasseau, V. (eds.) Poincaré Seminar 2003, pp. 119–144. Birkhauser, Basel, Switzerland (2004)
110. Balian, R.: Information in statistical physics. In: Studies in History and Philosophy of Modern Physics, Part B. Elsevier, Amsterdam (2005)
111. Balian, R.: François Massieu et les Potentiels Thermodynamiques, Évolution des Disciplines et Histoire des Découvertes. Académie des Sciences, Avril, France (2015)
112. Leray, J.: Un prolongement de la transformation de Laplace qui transforme la solution unitaire d'un opérateur hyperbolique en sa solution élémentaire. (Problème de Cauchy. IV.). Bulletin de la Société Mathématique de France **90**, 39–156 (1962)
113. Legendre, A.M.: Mémoire Sur L'intégration de Quelques Equations aux Différences Partielles; Mémoires de l'Académie des Sciences: Paris, France, pp. 309–351 (1787)
114. Casalis, M.: Familles exponentielles naturelles invariantes par un groupe de translations. C. R. Acad. Sci. Ser. I Math. **307**, 621–623 (1988)
115. Casalis, M.: Familles exponentielles naturelles invariantes par un groupe. Ph.D. Thesis, Thèse de l'Université Paul Sabatier, Toulouse, France (1990)
116. Letac, G.: Lectures on Natural Exponential Families and Their Variance Functions, Volume 50 of Monografias de Matematica (Mathematical Monographs); Instituto de Matematica Pura e Aplicada (IMPA). Rio de Janeiro, Brazil (1992)
117. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. Ser. II **106**, 620–630 (1957)
118. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. Ser. **108**, 171–190 (1957)
119. Jaynes, E.T.: Prior probabilities. IEEE Trans. Syst. Sci. Cybern. **4**, 227–241 (1968)
120. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. Acad. Sci. **255**, 2897–2899 (1962)
121. Dacunha-Castelle, D., Gamboa, F.: Maximum d'entropie et problèmes des moments. Ann. Inst. H. Poincaré Prob. Stat. **26**, 567–596 (1990)
122. Gamboa, F., Gassiat, E.: Maximum d'entropie et problème des moments: Cas multidimensionnel. Probab. Math. Stat. **12**, 67–83 (1991)
123. Dacunha-Castelle, D., Gamboa, F.: Maximum de l'entropie sous contraintes non linéaires. Ann. Inst H. Poincaré Prob. Stat. **4**, 567–596 (1990)
124. Burbea, J., Rao, C.R.: Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. J. Multivar. Anal. **12**, 575–596 (1982)

125. Crouzeix, J.P.: A relationship between the second derivatives of a convex function and of its conjugate. Math. Progr. **3**, 364–365 (1977)
126. Hiriart-Urruty, J.B.: A new set-valued second-order derivative for convex functions. In: Mathematics for Optimization; Mathematical Studies Series, vol. 129. Elsevier: Amsterdam, The Netherlands (1986)
127. Eriksen, P.: Geodesics Connected with the Fisher Metric on the Multivariate Normal Manifold; Technical Report, pp. 86–13; Institute of Electronic Systems, Aalborg University, Aalborg, Denmark (1986)
128. Eriksen, P.S.: Geodesics connected with the Fisher metric on the multivariate normal manifold. In: Proceedings of the GST Workshop, Lancaster, UK (1987). Accessed 28–31 Oct 1987
129. Koszul, J.L.: Formes hermitiennes canoniques des espaces homogènes complexes. Sémin. Bourbaki Tome 3, Exposé **108**, 69–75 (1954–1956)
130. Bourguignon, J.P.: La géométrie kählérienne, domaines de la géométrie différentielle, séminaire Histoires de géométries, FMSH (2005). https://youtu.be/SDmMo4a1vbk
131. Gauduchon, P.: Calabi's extremal Kähler metrics: an elementary introduction. https://germanio.math.unifi.it/wp-content/uploads/2015/03/dercalabi.pdf
132. Cartier, P.: In memoriam Jean-Louis KOSZUL. Gazette des Mathématiciens **156**, 64–66 (2018)
133. Kosmanek, E.-E.: Hommage à mon directeur de thèse, Gazette des Mathématiciens **156**, 67, (2018)