axioms

MDPI

Article

A Unified Approach to Aitchison's, Dually Affine, and Transport Geometries of the Probability Simplex

Giovanni Pistone [1] **and Muhammad Shoaib** [2,*]

[1] De Castro Statistics Initiative, Collegio Carlo Alberto, 10122 Torino, Italy; giovanni.pistone@carloalberto.org
[2] Department of Mathematics—DIMA, University of Genova, 16146 Genova, Italy
* Correspondence: muhammad.shoaib@dima.unige.it

Abstract: A critical processing step for AI algorithms is mapping the raw data to a landscape where the similarity of two data points is conveniently defined. Frequently, when the data points are compositions of probability functions, the similarity is reduced to affine geometric concepts; the basic notion is that of the straight line connecting two data points, defined as a zero-acceleration line segment. This paper provides an axiomatic presentation of the probability simplex's most commonly used affine geometries. One result is a coherent presentation of gradient flow in Aichinson's compositional data, Amari's information geometry, the Kantorivich distance, and the Lagrangian optimization of the probability simplex.

Keywords: affine geometry; Bayes space; information geometry

MSC: 53B12; 62B11; 94A16

check for updates

Citation: Pistone, G.; Shoaib, M. A Unified Approach to Aitchison's, Dually Affine, and Transport Geometries of the Probability Simplex. *Axioms* **2024**, *13*, 823. https://doi.org/10.3390/axioms13120823

Academic Editors: Anna Maria Fino and Sidney A. Morris

Received: 5 September 2024
Revised: 1 November 2024
Accepted: 21 November 2024
Published: 25 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

This paper discusses the geometry of the set of all probability measures whose support is a given finite set. The *sample space* Ω is the set of possible outcomes of the experiment of interest. While the use of a geometric language is well established in statistics, the formalization of a specific *affine* structure is due to the work of Amari, particularly in [1,2]. Amari defines a dually affine atlas of charts in the precise sense to be described below. This structure is fascinating because it gives an intrinsically geometric meaning to classical Fischerian objects such as log-likelihood, Fisher's score, and Fisher's information. This paper is based on a non-parametric representation of Amari's seminal work, which was introduced in a series of papers, starting with [3].

1.1. Probability Simplex

On the finite sample space Ω, a probability is characterized by its *probability function* $q \in \mathbb{R}^\Omega$, that is, the mapping $\Omega \to \mathbb{R}_{\geq 0}$ (the set of non-negative real numbers) in which the sum of the probabilities over all outcomes equals 1. Hence, the set of all probability functions, the *probability simplex*, is

$$\mathcal{P}(\Omega) = \left\{ q \colon \Omega \to \mathbb{R}_{\geq 0} \,\middle|\, \sum_{x \in \Omega} q(x) = 1 \right\}. \tag{1}$$

The probability simplex is a convex set, which means it contains all mixtures. Given two probability functions, $p, q \in \mathcal{P}(\Omega)$, the line segment joining p and q,

$$[0,1] \ni t \mapsto p(t) = (1-t)p + tq, \tag{2}$$

lies entirely within $\mathcal{P}(\Omega)$.

1.2. Compositions

According to [4,5], a *composition*, f, in the sample space Ω is a positive function of Ω, $f\colon \Omega \to \mathbb{R}_{\geq 0}$. Two compositions, f and g, are equivalent if there exists a constant, $k \in \mathbb{R}_{>0}$, such that $g = kf$. Each composition, f, is normalized to a probability function via

$$\mathbb{R}^{\Omega}_{\geq 0} \ni f \mapsto \frac{f}{\sum_{x\in\Omega} f(x)} \in \mathcal{P}(\Omega). \tag{3}$$

The above equation holds under the condition that $f \neq 0$ on Ω, as normalization requires a non-zero sum of $f's$ values. If $f = 0$ on Ω, normalization would not be possible. Two compositions are equivalent if and only if the normalized compositions are equal so that the set of equivalent compositions is the same as the probability simplex. However, the notion of the composition suggests a geometric structure other than the mixture of Equation (2). Given compositions f and g, the product $h = f \circ g$ is a composition, and the proportionality relation holds. If $f_1 = k_1 f$ and $g_1 = k_2 g$, then $f_1 g_1 = (k_1 k_2) fg$. Similarly, for the exponentiation $(\alpha, f) \mapsto f^{\alpha}$, $\alpha \in \mathbb{R}$.

These two operations provide a way of connecting two probability functions with the so-called *exponential arc*,

$$[0,1] \ni t \mapsto \frac{p^{1-t}q^{t}}{\sum_{x\in\Omega} p(x)q(x)} \in \mathcal{P}(\Omega). \tag{4}$$

The set of all compositions with the operations we have just described is sometimes called *Bayes space* [6] because the Bayes formula is possibly the most important example of a normalized product of compositions.

The mixture (2) and the exponential arc (4) will be central objects in the geometric presentation of the probability simplex.

1.3. Polytopes and Simplexes

We need to review some basic facts of convex analysis; see, for example, [7,8]. A set, S, in a Euclidean space $\mathbb{R}^n$ is *convex* if, for any two points, $p, q \in S$, the line segment joining p and q of Equation (2) lies entirely within S. The *convex hull* of a set, $P \in \mathbb{R}^n$, is the set of all convex combinations of points in P,

$$\operatorname{conv}(P) = \left\{ \sum_{i=1}^{k} \lambda_i p_i \,\middle|\, k \geq 1, p_i \in P, \lambda_i \geq 0, \sum_{i=1}^{k} \lambda_i = 1 \right\},$$

where the non-negative real numbers λ_i are the weights or coefficients of the convex combination, and p_i are points in P. In particular, a *polytope* is the convex hull of a finite set of points. A minimal set of such points is the set of vertices. The probability simplex on Ω (also called the standard simplex in other fields) is the polytope whose vertices are the delta probability functions $\delta_a(x) = (x = a)$ for all $a \in \Omega$; that is,

$$\delta_a(x) = \begin{cases} 1 & \text{if } x = a, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Each δ_a represents a probability for which the entire probability mass is concentrated at a.

An *affine combination* of the vectors $p_1, \ldots, p_k \in \mathbb{R}^{\Omega}$ is a vector of the form $\sum_{j=1}^{k} t_j p_j$, with $\sum_{j=1}^{k} t_j = 1$. The given vectors are *affinely independent* if no affine combination is zero. A *simplex* is a polytope generated via *affinely independent* points $a_1, \ldots, a_n$. All simplexes with the same number of vertices can be transformed into one another through a suitable affine function. Thus, any n-vertex simplex in any n-dimensional space is essentially the same shape, differing only by scaling, rotation, and translation.

The probability simplex is a closed set in its ambient Euclidean space. The topological border is a union of simplexes of smaller dimensions, each one corresponding to a reduced support of the probability functions. In statistical modeling, one usually assumes that the

support of each probability function in the model of interest is the same. For this reason, we will mainly consider the open probability simplex $\mathcal{P}_>(\Omega)$, which is the set of strictly positive probability functions.

1.4. Overview

This piece of research presents, in a coordinated and principled way, some basic topics in the geometrical calculus of the probability simplex. The specific topics are Aitchison's theory of compositional data and Bayes spaces, Amari's dually affine information geometry, and transport theory in product sample spaces. We define a differential geometric structure with an explicitly defined expression of the tangent bundle and discuss the relevant calculus operations, especially the notions of a natural gradient and gradient flow.

Section 2 is devoted to the definition of *affine space* as a set, M, endowed with a *displacement* function that maps every couple of points to a vector in such a way that the rule of addition of vectors holds. An affine space is a differentiable manifold for which the displacement function with the first point fixed provides a chart from $\mathcal{M}$ to a vector space. The Aitchison geometry of compositional data is an instance of such a geometry.

Section 3 presents a generalization of the notion of affine space to a more complex object, which allows us to set information geometry within this geometric framework. In affine spaces, the tangent space is the same at each point. However, the classical Fisher's methodology, together with Amari's methodology, suggests taking the Fisher's score as a definition of the tangent vector. As a consequence, it is convenient to assume that the primary expression of the tangent bundle is the set of weighted contrasts. This natural setting requires an extra structure, that is, an identification between tangent spaces, called a *parallel transport*. This section contains the main computational results, particularly the notions of natural gradient, duality of transports, and gradient flow.

Section 4 presents the analysis of the affine statistical bundle when the sample space is of product type. In this case, we introduce the notions of ANOVA decomposition and an optimal transport plan.

Section 5 deals with the second-order geometric analysis of the probability simplex, that is, the first-order geometric analysis of the statistical bundle. It is well known from differential geometry that second-order objects, such as accelerations, are not uniquely defined from the given atlas of charts on the base space. In our case, we have charts defined on a vector bundle expressing the probability functions and Fisher's scores, which allows for statistically natural definitions of accelerations. The availability of second-order objects prompts the development of results inspired by notions of mechanics, such as the calculus of Lagrangian functions. It should be noted that the use of a language inspired by physics here does not refer to the push-up of the probability simplex of a dynamic on the sample space; rather, it refers to an irreducible dynamic of the probability simplex.

2. Affine Geometry of the Probability Simplex

The geometries we are going to discuss are supported on the probability simplex $\mathcal{P}(\Omega)$ or on the open probability simplex $\mathcal{P}_>(\Omega)$ of a finite sample space Ω. The latter open convex set is the interior of the former full probability simplex (1). The term "geometry" is ambiguous and should be qualified by stating which type of geometry we are talking about. Actually, the field of statistics uses many types of geometries, such as linear, affine, metric, geodesic, Riemannian, and differential geometries. We discuss here affine geometry, to be precisely defined below. Our general references are the monographs [2,9,10] (Ch. I–II). The tutorials [11,12] present similar material as the present one but in a more restricted scope.

Affine geometry is a special instance of differential geometry. The former inherits the language of the latter: chart, atlas, expression, tangent space, and bundle…We refer to the non-parametric version of differential geometry as presented in [13,14]. Below, we list the basic concepts.

Given a set, M, and a real vector space, V, a *chart* is a one-to-one mapping, $s : U \to V$, where $U \subseteq M$, and $s(U)$ is open in V. Two charts, $s_1 \colon U_1 \to V_1$ and $s_2 \colon U_2 \to V_2$, are C^k-compatible if the change in chart mapping

$$s_2 \circ s_1^{-1} : s_1(U_1 \cap U_2) \to s_2(U_1 \cap U_2)$$

is either empty or a C^k-diffeomorphism of open sets. (As usual, $\circ$ stands for the composition of functions.) A family of two by two C^k-compatible charts is a C^k-*atlas*. Two atlases are compatible if their union is an atlas. A maximal C^k-atlas is a C^k-*manifold*.

A *curve* is a map from a real interval to the manifold. We assume, unless otherwise stated, that the interval is an open neighborhood of 0. Given a curve, $\gamma \colon t \mapsto \gamma(t) \in M$, and a chart, s, in the atlas, the curve $s \circ \gamma \colon t \mapsto s(\gamma(t)) \in V$ is the *expression* of the curve in the chart s. If one expression (hence, all) is differentiable, then $t \mapsto \frac{d}{dt}\, s \circ \gamma(t) \in V$ is the expression of the velocity at t of the curve γ in the chart s. The vector space of all possible expressions of velocities for curves passing through a given $p \in M$ is the expression of the *tangent space* at p.

2.1. Affine Space

The open probability simplex $\mathcal{P}_>(\Omega)$ and the probability simplex are, respectively, an open convex subset and a closed convex subset of the generated affine plane,

$$A(\Omega) = \left\{ v \in \mathbb{R}^\Omega \,\middle|\, \sum_{x \in \Omega} v(x) = 1 \right\}, \tag{6}$$

where $A(\Omega)$ consists of all vectors in the vector space $\mathbb{R}^\Omega$ whose components sum to 1. Probability functions are non-negative and sum to 1. As $A(\Omega)$ is an Euclidean space, the open probability simplex inherits Euclidean geometry. In some contexts, it is useful to relax the non-negativity constraint and allow the "probabilities" to take on negative values while maintaining the sum-to-one condition. By allowing $v(x)$ to be negative, we extend our consideration from the standard probability simplex to a larger space.

Example 1. *Consider a simple three-dimensional space with outcomes* $\Omega = \{1, 2, 3\}$. *A probability vector* $v \in \mathbb{R}^3$ *in this space might look like* $v = (v_1, v_2, v_3)$, *where* $v_1 + v_2 + v_3 = 1$. *For standard probabilities, we require* $v_1, v_2, v_3 \geq 0$. *However, if we relax this constraint,* v_1, v_2, v_3 *can assume negative values as long as they sum to 1. In the plot of Figure 1, the blue triangle represents the standard probability simplex where* $v_1, v_2, v_3 \geq 0$. *The red points represent example vectors in the generalized affine space, including one where* v_3 *is negative and another where* v_1 *exceeds 1.*

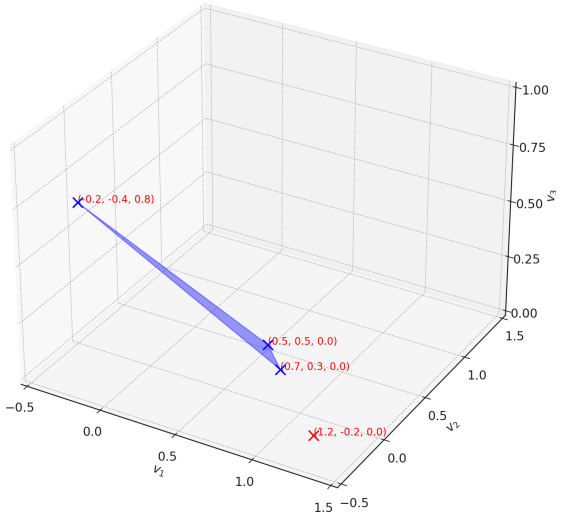


Figure 1. The blue triangle represents the probability simplex. The red points represent vectors in the generalized affine space, including points with negative values or values exceeding 1.

Example 2. *A* score function, *S, can be defined on the space of these generalized probability vectors. For a given vector, v, and an outcome, x, the score* $S(v, x)$ *quantifies how well the vector v predicts the outcome x. A common example of a score function is the Brier score [15], defined as follows:*

$$S(v, x) = \sum_{y \in \Omega} (v(y) - \delta_x(y))^2$$

where δ_x *is the indicator function of x, that is, 1 if* $y = x$ *and 0 otherwise.*

The following definition, due to [16], extends the notion of geometry on a hyperplane to a more general situation.

Definition 1 (Weyl's affine space)**.** *Given a set, M, and a real finite-dimensional vector space, V, an affine space,* $(M, V, \overrightarrow{\cdot\cdot})$*, consists of* displacement mapping,

$$M \times M \ni (p, q) \mapsto \overrightarrow{pq} \in V \,,$$

such that the following two assumptions hold.

1. *For each point* $p \in M$*, the partial mapping* $s_p : M \ni q \mapsto s_p(q) = \overrightarrow{pq}$ *is injective (one-to-one) with the open image* $s_p(M)$*.*
2. *Given points p, q, and r, the displacement vectors satisfy the parallelogram law,*

$$\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr} \,.$$

Property 1 ensures that, for a fixed point, p, the mapping from points in M to the vector space, V, via the displacement vector $\overrightarrow{pq}$ does not map different points to the same vector. In the language of differential geometry, the mapping s_p is a *chart*, and $v = s_p(q)$ is the coordinate of q in the p-chart. As $s_p(p) = 0$, we will say that p is the *origin* of the chart. The inverse chart, also called a *patch*, is $s_p^{-1} \colon s_p(M) \to M$.

Property 2 means that the vector displacement from p to r is the same as the sum of the vector displacement from p to q and from q to r; see Figure 2. This property ensures that adding displacements is consistent with the geometric intuition of moving along a path from one point to another via an intermediate point. In terms of the charts,

$$s_p(q) + s_q(r) = s_p(r) \,. \tag{7}$$

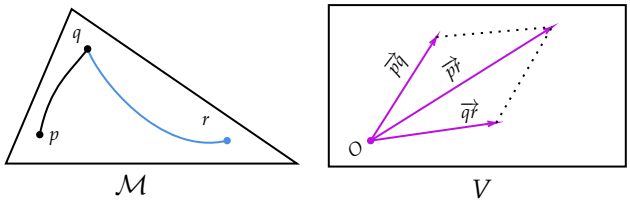


Figure 2. Parallelogram law in an affine space.

In particular, we have $s_p(p) = \overrightarrow{pp} = 0$; hence s_p is the chart centered at p. Also, $s_p(q) + s_q(p) = 0$. The change-of-chart transformation from origin p to origin q follows from Equation (7).

$$s_q \circ s_p^{-1}(v) = s_q(s_p^{-1}(v)) = s_q(p) + s_p(s_p^{-1}(v)) = s_q(p) + v \,. \tag{8}$$

The change-of-chart from origin p to origin q is the translation via $s_q(p) = -s_p(q)$.

2.2. Affine Manifold

In the affine space $(M, V, \overrightarrow{\cdot\cdot})$, the displacement $(p, q) \mapsto \overrightarrow{pq}$ defines a system of charts, $s_p \colon M \ni q \mapsto \overrightarrow{pq}$, as p varies in $\mathcal{M}$. Each s_p is one-to-one, and the change-of-charts (8) are translations; hence, they are highly smooth. The *affine manifold* is defined by the atlas

of charts $s_p : M \to V$, $p \in M$. The affine manifold differs from the C^k-manifold that it generates because the former is not maximal, while the latter is. Any vector basis of V provides a version of the affine space, where the vector space is a real vector space of parameters, producing a faithful parametrization of M.

Consider a curve, $t \mapsto \rho(t) \in M$, in the affine manifold. For any reference point $p \in M$, the curve $t \mapsto s_p(\rho(t))$ is the s_p-expression of the curve. The derivative $\frac{d}{dt} s_p(\rho(t))$ does not depend on the choice of the point p:

$$\frac{d}{dt} s_p(\rho(t)) = \frac{d}{dt}\big(s_p(q) + s_q(\rho(t))\big) = \frac{d}{dt} s_q(\rho(t))$$

Thus, the derivative $\dfrac{d}{dt} s_p(\rho(t))$ is independent of the choice of p and is called the *affine velocity of the curve*. It is denoted as

$$\frac{d}{dt}\rho(t) \quad \text{or} \quad \dot{\rho}(t)\,.$$

The vector space of all velocities evaluated at $t = 0$ of all curves $t \mapsto \rho(t) \in M$, such that $\rho(0) = p$, is the *tangent space* at p of the affine manifold, namely

$$T_p = \{\dot{q}(0) \,|\, t \mapsto q(t) \in M, q(0) = p\} \tag{9}$$

Conversely, given any $v \in V$, the curve $t \mapsto s_p^{-1}(tv)$ has affine velocity v, so that $T_p M = V$.

Example 3. *Let us discuss a basic elementary example that is relevant in the statistical literature on the design of experiments and regression modeling. If we take as base set M the hyperplane $A(\Omega)$ of Equation (6) and as vector space the vector space parallel to the probability simplex, that is, the space of* contrasts,

$$\mathrm{CO}(\Omega) = \left\{ v \in \mathbb{R}^\Omega \,\middle|\, \sum_{x\in\Omega} v(x) = 0 \right\}, \tag{10}$$

the displacement is $\overrightarrow{pq} = q - p \in V$. In fact, $\sum_x (q(x) - p(x)) = 1 - 1 = 0$. And the mapping $s_p(q) = q - p$ is one-to-one. The parallelogram law is $(q - p) + (r - q) = (r - p)$. The tangent space is defined in Equation (10). In the case with $\Omega = \{1, 2, 3\}$, a vector basis of $\mathrm{CO}(\Omega)$ is

$$v_1 = (1, -1, 0), \quad v_2 = (1, 0, -1)\,,$$

so that

$$(p, q) \mapsto s_p(q) = -(q(2) - p(2)) \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} - (q(3) - p(3)) \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

The chart at p in parametric coordinates is

$$q \mapsto \begin{pmatrix} -(q(2) - p(2)) \\ -(q(3) - p(3)) \end{pmatrix} \in \mathbb{R}^2\,.$$

Often, the interpretability of regression models depends on the suitable choice of vector space bases of $\mathrm{CO}(\Omega)$ and charts.

Interestingly, many types of affine manifolds on the open probability simplex are of statistical interest. They all use the space of contrasts (10) as vector space. Various vector bases of $\mathrm{CO}(\Omega)$ are commonly used in applications. One option is to choose a component $a \in \Omega$ and use the *reduced basis* $(\delta_x)_{x\neq a}$. With this basis, the components of the projection of $v \in \mathrm{CO}(\Omega)$ are $(v(x) - v(a))_{x\neq a}$. A second option is the *preferred point basis*. Fix $a \in \Omega$. If v is a contrast, it holds that $v(a) = -\sum_{x\neq a} v(x)$; hence $v = \sum_{x\neq a} v(x)(\delta_x - \delta_a)$ and $(\delta_x - \delta_a)_{x\neq a}$ is a basis.

The space of contrasts is a Euclidean space, for example, with the restriction of the inner product in $\mathbb{R}^\Omega$. Other inner products can be used. While not part of the affine geometry, these metrics, applied to the vector space of contrasts, provide various useful instances of metric on the affine space; see Example 2. It is important to realize that the choice of an affine displacement and the choice of an inner product are independent modeling choices.

Example 4 (Open probability simplex)**.** *The construction in Example 3 is not feasible for the probability simplex because the image of the restricted mapping*

$$\mathcal{P}(\Omega)\colon q \mapsto q - p \in \mathrm{CO}(\Omega)$$

is closed. We can use the same construction as before if we restrict to the open probability simplex,

$$\mathcal{P}_>(\Omega)\colon q \mapsto q - p \in \mathrm{CO}(\Omega). \tag{11}$$

Aitchison's Centered Log Ratio [5]

In this section, a key notion for statistical usage of Aitchison's geometry, namely the centered log ratio (CLR), is interpreted as a displacement. This allows for the interpretation of the moments of a random variable of the cumulant function in this framework.

If f and $f_1 = k_1 f$ are equivalent positive compositions, then $\log f_1 = \log f + \log k_1$. The logarithms of equivalent compositions differ by an additive constant. This remark suggests using a logarithmic scale by considering the set of positive probability functions $\mathcal{P}_>(\Omega)$, the vector space of contrasts $\mathrm{CO}(\Omega)$, $n = \#\Omega$, and the *centered log ratio* displacement

$$\mathrm{CLR}\colon \mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto \log \frac{q}{p} - \frac{1}{n} \sum_{x \in \Omega} \log \frac{q(x)}{p(x)}\,. \tag{12}$$

If f is a composition and ϕ is a random variable of Ω, then, if we define the $\mathbb{E}$-notation to include normalization,

$$\mathbb{E}_f[\phi] = \sum_{x \in \Omega} \frac{f(x)}{\sum_x f(x)}\, \phi(x)\,. \tag{13}$$

Let us check that all conditions for an affine manifold hold. The partial mapping at p is a chart with

$$s_p\colon \mathcal{P}_>(\Omega) \ni q \mapsto \mathrm{CLR}_p(q) = \log \frac{q}{p} - \mathbb{E}_1\left[\log \frac{q}{p}\right] \in \mathrm{CO}(\Omega)\,, \tag{14}$$

$$s_p^{-1}\colon \mathrm{CO}(\Omega) \ni v \mapsto \mathrm{e}^{v - \Psi_p(v)} \cdot p \quad , \quad \Psi_p(v) = \mathbb{E}_1\left[\log \frac{p}{q}\right] = \log \mathbb{E}_p[\mathrm{e}^v]\,. \tag{15}$$

The axioms for affine geometry are satisfied, as the equality $s_p(q) + s_q(r) = s_p(r)$ is easily checked,

$$\begin{aligned}\mathrm{CLR}_p(q) + \mathrm{CLR}_q(r) = \log \frac{q}{p} - \mathbb{E}_1\left[\log \frac{q}{p}\right] + \log \frac{r}{q} - \mathbb{E}_1\left[\log \frac{r}{q}\right] \\ = \log \frac{r}{p} - \mathbb{E}_1\left[\log \frac{r}{p}\right] = \mathrm{CLR}_p(r)\,.\end{aligned}$$

The patch mapping is a non-parametric exponential model [17,18] (§ 5.5) in which the sufficient-statistics v is a contrast. The convex mapping Ψ_p is sometimes called a cumulant function. The first and second derivatives of Ψ_p in the direction h and the directions h and k, respectively, are (as computed in the references above),

$$d\Psi_p(v)[h] = \mathbb{E}_q[h]\,, \quad d^2\Psi_p(v)[h,k] = \mathrm{Cov}_q(h,k)\,. \tag{16}$$

The velocity of the curve $t \mapsto q(t)$ is

$$\frac{d}{dt}\operatorname{CLR}_p(q(t)) = \frac{d}{dt}\log q(t) - \mathbb{E}_1\left[\frac{d}{dt}\log q(t)\right] = \frac{\dot{q}(t)}{q(t)} - \mathbb{E}_1\left[\frac{\dot{q}(t)}{q(t)}\right].$$

where $\dot{q}(t)$ is the derivative computed in $A(\Omega)$. The random variable $\dot{q}(t)/q(t)$ is called *Fisher's score* [18], and it has the remarkable property

$$\mathbb{E}_{q(t)}\left[\frac{\dot{q}(t)}{q(t)}\right] = \sum_{x\in\Omega}\dot{q}(x;t) = 0\,.$$

In the reduced basis, with $\Omega = \{1,\dots,n\}$, each $j \neq n$ component of the CLR displacement becomes

$$\begin{aligned}\operatorname{CLR}_p(q)(j) = \left(\log\frac{q(j)}{p(j)} - \mathbb{E}_1\left[\log\frac{q}{p}\right]\right) - \left(\log\frac{q(j)}{p(j)} - \mathbb{E}_1\left[\log\frac{q}{p}\right]\right)\\ = \log\frac{q(j)}{p(j)} - \log\frac{q(n)}{p(n)} = \log\frac{p(n)q(j)}{p(j)q(n)}\,.\end{aligned}$$

See [4] for a detailed treatment.

Within Aitchison's paradigm, we presented the idea of the centered log-ratio (CLR) transformation (12). Determining displacement mappings between probability distributions on the simplex depends critically on this transformation. By mapping compositions into a Euclidean space, the CLR makes the comparison and analysis of probability distributions easier in a way that makes sense with respect to the inherent geometry of the simplex (referring to the unique geometric properties of the probability simplex—a space that consists of probability distributions where all components are positive and sum to one). Aitchison geometry's mathematical foundations are explored in more detail in Section 3.6 with special attention to how it relates to information geometry (IG). We discuss how Aitchison geometry may be considered a specific example of IG, simplifying the displacement mappings with a fixed reference chart and deriving important conclusions, including the parallelogram law that guarantees these mappings to be consistent over a wide range of distributions. A further explanation and detailed connections with the centered log-ratio transformation are provided in Section 3.6 below.

2.3. Tangent Bundle

The motivation for defining a manifold structure, $\mathcal{M}$, through an atlas, $\mathcal{A}$, is the availability of a coherent setup for performing calculus operations. A typical problem in our interest is minimizing a smooth scalar field $G\colon \mathcal{M} \to \mathbb{R}$ via a gradient-flow algorithm. To this end, we must consider points on the manifold and derivatives, which exist in a vector space attached to each point.

The *tangent bundle* $T\mathcal{M}$ is defined as a set of pairs,

$$T\mathcal{M} = \left\{(q,v)\,\middle|\, q \in \mathcal{M}, v \in T_q\mathcal{M}\right\},$$

where $T_q\mathcal{M}$ is the tangent space at q, that is, the vector space of all possible velocities of curves through q. The situation is straightforward in all the examples of affine manifolds we have seen up to now, as the tangent space identifies with the space of contrasts $\mathrm{CO}(\Omega)$.

Example 5 (Exponential arc)**.** *Let $t \mapsto p(t)$ be the exponential arc in Equation (4) connecting the points q and $r \in \mathcal{P}_>(\Omega)$. In the CLR affine manifold, we represent the two equivalent probability densities with the CLR patch centered at p, as in Equation (15),*

$$q = \mathrm{e}^{v-\Psi_p(v)}\cdot p\,, \quad r = \mathrm{e}^{w-\Psi_p(w)}\cdot p\,, \quad \textit{with } v,w \in \mathrm{CO}(\Omega)\,,$$

and hence, the arc $t \mapsto p(t)$ is expressed in terms of a mixture of two random variables, v and w, and the convex mapping Ψ_p, that is, two canonical statistics and the cumulant function,

$$p(t) = \mathrm{e}^{(1-t)v+tw-\Psi_p((1-t)v+tw)} \cdot p \quad \textit{and} \quad \mathrm{CLR}_p(p(t)) = (1-t)v + tw\,.$$

The affine velocity is constant,

$$\dot{q}(t) = \frac{d}{dt}\,\mathrm{CLR}_p(p(t)) = \frac{d}{dt}((1-t)v - tw) = w - v\,.$$

The exponential arc is the line segment with constant velocity connecting q and r in the CLR affine geometry. The set of all possible velocities at a given point is an expression of the tangent space.

Example 6 (Gradient of the KL-divergence)**.** *For any $p \in \mathcal{P}_>(\Omega)$, the Kullback–Leibler divergence, in short KL-divergence (see [19] or § 3.3–4 in [9]), of q and p is the scalar field*

$$\mathcal{P}_>(\Omega) \ni q \mapsto \mathrm{D}(p\,\|q) = \mathbb{E}_p\left[\log\frac{p}{q}\right] \in \mathbb{R}\,.$$

The patch at p for the CLR displacement is Equation (15); hence, the expression at p of the KL-divergence is

$$G\colon v \mapsto \mathrm{D}\Big(p\,\|\mathrm{e}^{v-\Psi_p(v)} \cdot p\Big) = \mathbb{E}_p\big[-v + \Psi_p(v)\big] = \Psi_p(v) - \mathbb{E}_p[v]\,.$$

The expression is defined on the vector space $\mathrm{CO}(\Omega)$, and hence, we can compute the ordinary derivative in the direction $h \in \mathrm{CO}(\Omega)$ using Equation (16) and $q = \mathrm{e}^{v-\Psi_p(v)} \cdot p$,

$$\begin{aligned} dG(v)[h] = \frac{d}{dt}G(v+th)\Big|_{t=0} &= \frac{d}{dt}\Psi_p(v+th)\Big|_{t=0} - \mathbb{E}_p[h] \\ &= d\Psi_p(v)[h] - \mathbb{E}_p[h] = \mathbb{E}_q[h] - \mathbb{E}_p[h] = \mathbb{E}_p\Big[(\mathrm{e}^{v-\Psi_p(v)} - 1))h\Big] \\ &= \mathbb{E}_1\Big[np(\mathrm{e}^{v-\Psi_p(v)} - 1)h\Big]\,. \end{aligned}$$

As $(np(\mathrm{e}^{v-\Psi_1(v)} - 1) \in \mathrm{CO}(\Omega)$, if we define the inner product of $\mathrm{CO}(\Omega)$ to be $(v,w) \mapsto \langle v, w\rangle = \mathbb{E}_1[vw]$, then $np(\mathrm{e}^{v-\Psi_1(v)} - 1)$ is the gradient *of G,*

$$dG(v)[h] = \Big\langle pn(\mathrm{e}^{v-\Psi_p(v)}, h\Big\rangle\,.$$

The gradient is also defined on the affine manifold as follows. If $t \mapsto q(t)$ is a smooth curve with expression in the CLR, the expression at p is $t \mapsto \mathrm{CLR}_p(q(t))$, and the affine velocity is $\dot{q}(t) = \frac{d}{dt}\,\mathrm{CLR}_p(q(t)) = \frac{\dot{q}(t)}{q(t)} - \mathbb{E}_1\left[\frac{\dot{q}(t)}{q(t)}\right]$. Hence,

$$\frac{d}{dt}\,\mathrm{D}(p\,\|q(t)) = \frac{d}{dt}G(\mathrm{CLR}_p(q(t)) = \big\langle n(q(t) - p), \dot{q}(t)\big\rangle\,.$$

Example 7 (Entropy)**.** *The expression of the entropy [19]*

$$\mathcal{H}(q)\colon \mathcal{P}_>(\Omega) \ni q \mapsto -\mathbb{E}_q[\log q] = -\mathbb{E}_p\left[\frac{q}{p}\log q\right]$$

in the chart at p becomes, using the first Equation (16),

$$\begin{aligned} H_p\colon\ \mathrm{CO}(\Omega) \ni v \mapsto H_p(v) &= -\mathbb{E}_p\Big[\mathrm{e}^{v-\Psi_p(v)}(v - \Psi_p(v) + \log p)\Big] \\ &= -d\Psi_p(v)[v] + \Psi_p(v) + \mathcal{H}(p)\,, \end{aligned}$$

and hence, using the second Equation (16),

$$dH_p(v)[h] = -d^2\Psi_p(v)[v,h] - d\Psi_p(v)[h] + d\Psi_p(v)[h] = \mathrm{Cov}_{s_p^{-1}(v)}(v,h)\,.$$

If $t \mapsto q(t)$ *is a curve with expression* $v(t) = \log\frac{q(t)}{p} - \mathbb{E}_1\left[\log\frac{q(t)}{p}\right]$ *and velocity* $\overset{\star}{q}(t) = \frac{\dot{q}(t)}{q(t)} - \mathbb{E}_1\left[\frac{\dot{q}(t)}{q(t)}\right]$,

$$\frac{d}{dt}\mathcal{H}(q(t)) = \frac{d}{dt}H_p(\mathrm{CLR}_p(q(t))) = \mathrm{Cov}_{q(t)}\left(v(t), \overset{\star}{q}(t)\right) = \mathrm{Cov}_{q(t)}\left(\log\frac{q(t)}{p}, \frac{\dot{q}(t)}{q(t)}\right).$$

3. Affine Statistical Bundle

We define other affine geometries on the open probability simplex, using a richer atlas of charts and introducing an essential notion of duality. This approach highlights the relationship between our topic and Fisher's statistics and Boltzmann and Gibbs's statistical physics [20].

Our geometry definitions are kinematic, based on calculating the velocity and acceleration of curves. A one-dimensional statistical model is a curve, that is, a mapping, $I \ni t \mapsto q(t) \in \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the probability simplex on the finite sample space Ω, and I is an open interval of $\mathbb{R}$. Higher dimensional statistical models are treated according to the usual conventions of multivariate analysis.

3.1. Fisher's Score

As $\mathcal{P}(\Omega) \subset \mathbb{R}^\Omega$, the affine velocity $t \mapsto \dot{q}(t)$ is well defined and takes values in the space of contrasts $\mathrm{CO}(\Omega)$, as shown in the previous section.

Smooth curves in $\mathbb{R}^\Omega$, whose values stay in the probability simplex, have a notable property related to their support. If the curve stays in the open simplex, then the support is Ω; that is, $q(x;t) \neq 0$ for all $x \in \Omega$ and all $t \in I$ [21].

Otherwise, the curve hits the boundary of the probability simplex; that is, there exists at least a pair, $(a, \bar{t})) \in \Omega \times I$, such that $q(a;\bar{t}) = 0$. This implies $q(a;t) \geq q(a,\bar{t})$; hence, $\dot{q}(a;\bar{t}) = 0$ [3]. In conclusion,

$$q(x;t) = 0 \Rightarrow \dot{q}(x;t) = 0\,.$$

In the language of measure theory, $\dot{q}(t)$ is absolutely continuous with respect to $q(t)$, $\dot{q}(t) \ll q(t)$ [22].

Hence, given any $\mathbb{R}^\Omega$-smooth curve in the probability complex, there exists a curve, $t \mapsto \overset{\star}{q}(t) \in \mathbb{R}^\Omega$, called *Fisher's score*, such that the following applies:

$$\dot{q}(t) = \overset{\star}{q}(t)\; q(t)$$

Fisher's score has two remarkable properties [23]. First, if $q(x;t) > 0$, $t \in I$, then

$$\overset{\star}{q}(x;t) = \frac{\dot{q}(x;t)}{q(x;t)} = \frac{d}{dt}\log q(x;t)$$

and second,

$$\mathbb{E}_{q(t)}\left[\overset{\star}{q}(t)\right] = \sum_x \overset{\star}{q}(x;t)q(x;t) = \sum_x \dot{q}(x;t) = \frac{d}{dt}1 = 0\,. \tag{17}$$

Specifically, if $t \mapsto q(t) \in \mathcal{P}_>(\Omega)$ (the open probability simplex), then Fisher's score is given by

$$\overset{\star}{q}(t) = \frac{d}{dt}\log q(t)\,.$$

There are many statistical reasons to take the Fisher score as a proper notion of the parameter rate of change in a statistical model. See the textbook [18] (Ch. 4) and the monograph [22].

3.2. Statistical Bundle and Parallel Transports

For the general notion of a bundle, see, for example, [14]. The previous discussion leads us to consider, for each positive probability function $q \in \mathcal{P}_>(\Omega)$, the vector space of *q-contrasts*, that is, the random variables that are centered for q,

$$B_q = \left\{ u \in \mathbb{R}^\Omega \;\middle|\; \sum_{x\in\Omega} u(x)q(x) = 0 \right\}. \tag{18}$$

The *statistical bundle* $S\mathcal{P}_>(\Omega)$ is the bundle $S\mathcal{P}_>(\Omega) \to \mathcal{P}_>(\Omega)$ of q-contrasts,

$$S\mathcal{P}_>(\Omega) = \left\{ (q,u) \;\middle|\; q \in \mathcal{P}_>(\Omega), u \in \mathbb{R}^\Omega, \mathbb{E}_q[u] = 0 \right\}. \tag{19}$$

Each fiber $S_q\mathcal{P}_>(\Omega) = B_q$ will represent all Fisher's scores at the point q. In turn, the Fisher's score will be interpreted as velocities.

Example 8. *The plot in Figure 3 illustrates the probability simplex for two outcomes, the point $q = \{q(1), q(2)\}$ within the simplex, and the vectors u_1 and u_2 representing q-contrasts.*

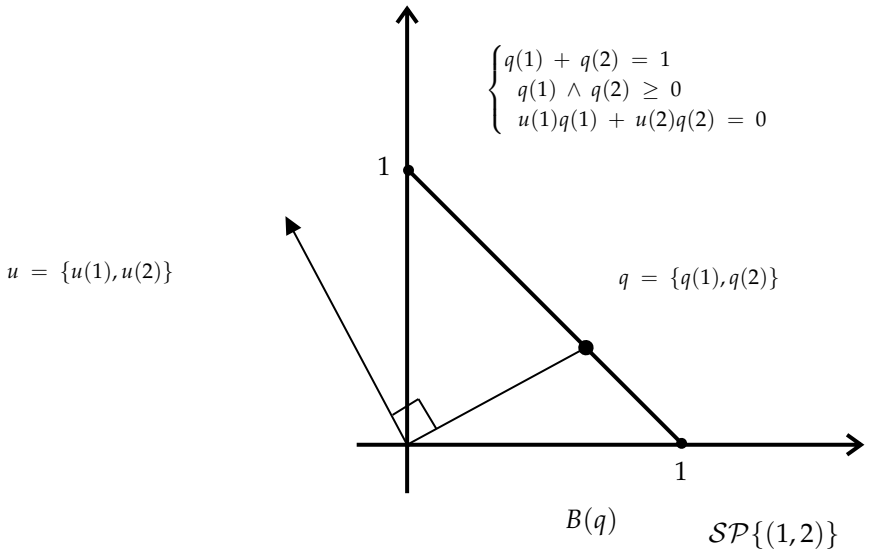


Figure 3. Example 8: the vector u is an element of the fiber at point q of the probability simplex.

Each fiber B_q is equipped with a natural *inner product*, the covariance at q,

$$\langle u, v \rangle_q = \mathrm{Cov}_q(u,v) = \mathbb{E}_q[uv]\,.$$

This inner product measures the angles and lengths of vectors within each fiber, contributing to the metric structure of the statistical bundle. The mapping $(q,u,v) \mapsto \langle u,v\rangle_q$ is the *metric* of the statistical bundle.

Example 9 (Fisher's metric)**.** *Given a smooth curve, $t \mapsto q(t) \in \mathcal{P}_>(\Omega)$, the Fisher's score $\overset{\star}{q}(t) = \frac{d}{dt}\log q(t) = \dot{q}(t)/q(t)$ provides a curve in the statistical bundle [24],*

$$t \mapsto (q(t), \overset{\star}{q}(t)) \in S\mathcal{P}_>(\Omega)$$

The variance in the Fisher's score is called Fisher's information *[25,26],*

$$t \mapsto \langle \overset{\star}{q}(t), \overset{\star}{q}(t) \rangle_{q(t)} = \mathbb{E}_{q(t)}\left[\left(\frac{\dot{q}(t)}{q(t)} \right)^2 \right] = \sum_x \frac{(\dot{q}(x;t))^2}{q(x;t)}\,.$$

Each fiber is a different vector space in the statistical bundle. Because of that, the mapping $t \mapsto (q(t), \overset{\star}{q}(t)) \in S\mathcal{P}_>(\Omega)$ requires a special handling of differentiation with a construction called *connection*. We define this notion as follows.

Definition 2 (Parallel transport)**.** *A* parallel transport *of the statistical bundle $S\mathcal{P}_>(\Omega)$ is a collection of bijective linear mappings between the fibers*

$$\mathbb{U}_p^q \colon S_p\mathcal{P}_>(\Omega) \to S_q\mathcal{P}_>(\Omega) \tag{20}$$

such that, for all $p, q, r \in \mathcal{P}_>(\Omega)$, the cocycle relation *holds,*

$$\mathbb{U}_q^r \mathbb{U}_p^q = \mathbb{U}_p^r \,. \tag{21}$$

Given a parallel transport, if $v \in S_q\mathcal{P}_>(\Omega)$ is proportional to $\mathbb{U}_p^q u$, $u \in S_p\mathcal{P}_>(\Omega)$, we say that u and v are *parallel*.

3.3. Affine Bundle

Here, we generalize Weyl's definition of affine space to adapt it to the statistical bundle. The novelty is that there is a different vector space at each point of the space.

Definition 3. *Consider the statistical bundle $S\mathcal{P}_>(\Omega)$ and a cocycle of parallel transports, $\mathbb{U}_p^q$, $p, q \in \mathcal{P}_>(\Omega)$. An* affine displacement *is a mapping,*

$$s \colon \mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto s_p(q) \in S_p\mathcal{P}_>(\Omega) \,,$$

such that the following applies:

1. *For each fixed p, the mapping $s_p : q \mapsto s_p(q)$ is 1-to-1, and it has an open image;*
2. *The* parallelogram law *holds,*

$$s_p(q) + \mathbb{U}_q^p s_q(r) = s_p(r) \,.$$

The affine displacement on the statistical bundle provides an atlas of charts, s_p, $p \in \mathcal{P}_>(\Omega)$. Each chart is a local coordinate system around the origin point p. The *change-of-chart map* is $s_p \circ s_q^{-1}$. At $\rho = s_q^{-1}(w)$, $w \in S_q\mathcal{P}_>(\Omega)$, it holds that

$$s_p \circ s_q^{-1}(w) = s_p(\rho) = s_p(q) + \mathbb{U}_q^p s_q(\rho) = s_p(q) + \mathbb{U}_q^p w \,. \tag{22}$$

The change-of-origin map restricts an affine map to $\mathbb{R}^\Omega$, whose linear part is the parallel transport (Figure 4).

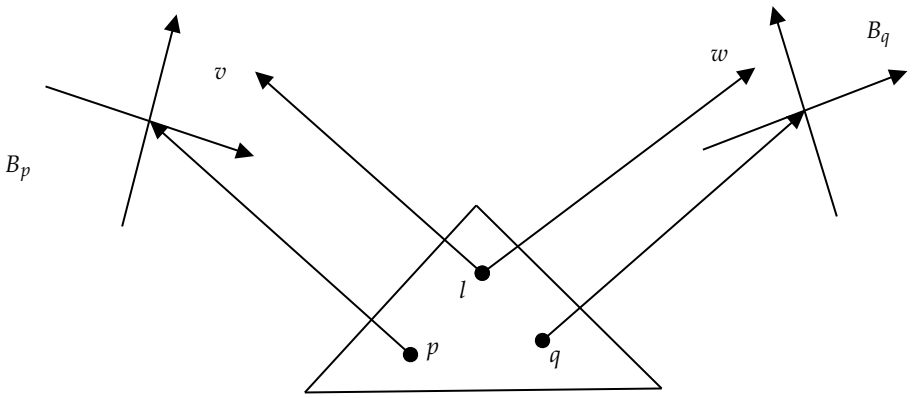


Figure 4. In the affine statistical manifold (the triangle), each point, for example, p, corresponds to a reference space, B_p. Given two reference spaces, B_p and B_q, the same point, l, has two different coordinates, $v \in B_p$ and $w \in B_q$.

Example 10 (Barycenter and convex combination). *Given points $q_1, \dots, q_n \in \mathcal{P}_>(\Omega)$, coefficients $\lambda_1, \dots, \lambda_n$ with $\sum_i \lambda_i = 1$, and a reference, p, the convex combination of p-coordinates, $\sum_i \lambda_i s_p(q_i)$, is the p-coordinate of a point, called the* barycenter,

$$\bar{q} = s_p^{-1}\left(\sum_i \lambda_i s_p(q_i)\right), \tag{23}$$

which does not depend on p. In fact, for $q \neq p$, we can use the change-of-chart of Equation (22) to obtain

$$s_q^{-1}\left(\sum_i \lambda_i s_q(q_i)\right) = s_q^{-1}\left(\sum_i \lambda_i \Big(s_q(p) + \mathbb{U}_p^q s_p(q_i)\Big)\right) = s_q^{-1}\left(s_q(p) + \mathbb{U}_p^q \sum_i \lambda_i s_p(q_i)\right) = s_q^{-1}\Big(s_q(p) + \mathbb{U}_p^q s_p(\bar{q})\Big) = \bar{q}\,.$$

Example 11 (Single fiber). *If s is an affine displacement with parallel transports $\mathbb{U}_p^q$, and (1) is the uniform probability function, then*

$$\bar{s}(p,q) = \mathbb{U}_p^{(1)} s(p,q)$$

defines an affine displacement with vector space $\mathrm{CO}(\Omega)$. In fact,

$$\bar{s}(p,q) + \bar{s}(q,r) = \mathbb{U}_p^{(1)} s(p,q) + \mathbb{U}_q^{(1)} s(q,r) = \mathbb{U}_p^{(1)}\Big(s(p,q) + \mathbb{U}_q^p s(q,r)\Big) = \mathbb{U}_p^{(1)} s(p,r) = \bar{s}(p,r)\,.$$

3.4. Velocity and Natural Gradient

If $t \mapsto q(t)$ is a curve in the open probability simplex, and s_p is a chart according to Definition 3, the expression of the velocity in the chart is, then,

$$t \mapsto \frac{d}{dt} s_p(q(t)) \in S_p\mathcal{P}_>(\Omega)$$

If we change the chart and use the second equations in Definition 3, the change in the expression of the velocity follows the parallel transport

$$\mathbb{U}_q^p \frac{d}{dt} s_q(q(t)) = \frac{d}{dt} s_p(q(t))\,.$$

Let us define an intrinsic notion of velocity using the so-called *moving frame*.

Definition 4 (Velocity). *The* velocity *of the curve $t \mapsto q(t)$ is*

$$\overset{\star}{q}(t)\colon t \mapsto \mathbb{U}_p^{q(t)} \frac{d}{dt} s_p(q(t))\,,$$

which does not depend on p. The velocity provides a lifting of the curve to the statistical bundle,

$$t \mapsto (q(t), \overset{\star}{q}(t)) \in S\mathcal{P}_>(\Omega)\,.$$

The function $F\colon \mathcal{P}_>(\Omega) \to \mathbb{R}$ is smooth if its expression in the chart at p, $F_p = F \circ s_p^{-1}\colon S_p\mathcal{P}_>(\Omega) \to \mathbb{R}$ is differentiable. Let us recall that the gradient ∇f of a smooth function, $f\colon \mathbb{R}^n \to \mathbb{R}$, is defined by the equation $\frac{d}{dt} f(x(t)) = \nabla f(x(t)) \cdot \dot{x}(t)$ for all smooth curves in $\mathbb{R}^n$. The same definition applies to any other inner product. After Amari, we call the *natural gradient* the gradient associated with the inner product on the statistical bundle.

Definition 5 (Natural gradient)**.** *If* $t \mapsto q(t) \in \mathcal{P}_>(\Omega)$, $q = q(0)$ *and* $\overset{\star}{q} = \overset{\star}{q}(0)$, *and* $F\mathcal{P}_>(\Omega) \to \mathbb{R}$ *is smooth, then*

$$\frac{d}{dt} F \circ q(t)\Big|_{t=0} = \frac{d}{dt} F_q \circ s_q(q(t))\Big|_{t=0} = dF_q(q(t))[\overset{\star}{q}(t)]\big|_{t=0} = dF_q(0)[\overset{\star}{q}] = \left\langle \operatorname{grad} F(q), \overset{\star}{q} \right\rangle_q , \quad (24)$$

where $q \mapsto \operatorname{grad} F(q)$ *is the* natural gradient *of* F*; that is,* $\operatorname{grad} F(q)$ *is the element of* $S_q\mathcal{P}_>(\Omega)$ *identified with the linear operator* $dF(0)$ *in the covariance inner product.*

Definition 6 (Exponential of a parallel transport)**.** *Assume that, for each given* $(q, v) \in S\mathcal{P}_>(\Omega)$, *the differential equation*

$$\overset{\star}{q}(t) = \mathbb{U}_q^{q(t)} v \,, \quad q(0) = q$$

has a unique global solution. Such a curve is auto-parallel, $\overset{\star}{q}(t) = \alpha(t)\mathbb{U}_{q(s)}^{q(t)} \overset{\star}{q}(s)$, *with* relative velocity $\alpha(t) = 1$. *The mapping*

$$\operatorname{Exp}\colon S\mathcal{P}_>(\Omega) \ni (q, v) \mapsto q(1) = \operatorname{Exp}_q(v) \in \mathcal{P}_>(\Omega) \,.$$

is called the exponential *of the affine bundle.*

The previous construction, leading from the affine bundle to the moving frame velocity and the natural gradient, is essential to the non-parametric version of Amari's information geometry. This specific construction is qualified below as the dually affine geometry of the probability simplex. It admits a number of immediate generalizations. In particular, it can be used to define a second-order calculus, as sketched in the example below.

Example 12 (Second-order affine geometry)**.** *As a basic set, let us assume the statistical bundle with the cocycle of parallel transports and, as a vector space, the product of the fibers. The following equation defines a displacement:*

$$((p, u), (q, v)) \mapsto (s_p(q), \mathbb{U}_q^p v - u) \in S_p\mathcal{P}_>(\Omega) \times S_p\mathcal{P}_>(\Omega) \,.$$

If $t \mapsto (q(t), v(t)) \in S\mathcal{P}_>(\Omega)$ *is a curve in the statistical bundle, then the derivative at* (p, u) *is*

$$t \mapsto \left(\frac{d}{dt}(q(t)), \frac{d}{dt} \mathbb{U}_{q(t)}^p v(t) \right)$$

In particular, if $v(t) = \overset{\star}{q}(t)$, *the second-order derivative is defined in the chart at* p *by*

$$t \mapsto \left(\frac{d}{dt} s_p(q(t)), \frac{d}{dt} \mathbb{U}_{q(t)}^p \overset{\star}{q}(t) \right) .$$

In the moving frame, the *acceleration* is defined by

$$\overset{\star\star}{q}(t) = \frac{d}{dt} \mathbb{U}_{q(t)}^p \overset{\star}{q}(t)\Big|_{p=q(t)} \quad (25)$$

3.5. Dually Affine Atlases on the Open Probability Simplex

According to the works of Amari [1] and Amari and Nagaoka [2], there are two main types of affine geometry on the probability simplex, the exponential one and the mixture one. We axiomatically describe such geometries via the notion of the statistical bundle endowed with parallel transport. This approach was initially presented in [27]. Each fiber of the statistical bundle is endowed with the covariance inner product to identify each fiber with its dual. In this identification, the two parallel transports are dual to each other.

Theorem 1 (Exponential transport and mixture transport)**.**

1. *The equation*
$$ {}^{e}\mathbb{U}_p^q \colon S_p\mathcal{P}_>(\Omega) \ni u \mapsto u - \mathbb{E}_q[u] \in S_q\mathcal{P}_>(\Omega) $$
defines a cocycle of parallel transports. It is the exponential transport*.*
2. *The equation*
$$ {}^{m}\mathbb{U}_p^q \colon S_p\mathcal{P}_>(\Omega) \ni u \mapsto \frac{p}{q}u \in S_q\mathcal{P}_>(\Omega) $$
defines a cocycle of parallel transports. It is the mixture transport*.*
3. *The exponential transport and the mixture transports are dual to each other in the covariance pairing. Namely, if* $u \in S_p\mathcal{P}_>(\Omega)$ *and* $v \in S_q\mathcal{P}_>(\Omega)$*,*
$$ \left\langle v, {}^{e}\mathbb{U}_p^q u \right\rangle_q = \left\langle {}^{m}\mathbb{U}_q^p v, u \right\rangle_p . $$

Proof. All checks for Definition 2 are immediate. The duality follows from

$$ \left\langle v, {}^{\mathrm{e}}\mathbb{U}_p^q u \right\rangle_q = \langle v, u - \mathbb{E}_q[u] \rangle_q = \langle v, u \rangle_q = \mathbb{E}_q[vu] = \mathbb{E}_p\left[\frac{q}{p}vu\right] = \left\langle {}^{\mathrm{m}}\mathbb{U}_q^p v, u \right\rangle_p . $$

□

For each parallel transport, we define an affine displacement. Each affine displacement maps the open probability simplex onto a vector space of contrasts. Conversely, the inverse maps are statistical models parametrized by a vector space of contrasts. The normalizing constants are interpreted in terms of statistical information.

Theorem 2 (Exponential chart and mixture chart)**.**

1. *The mapping*
$$ \mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto s_p(q) = \log\frac{q}{p} - \mathbb{E}_p\left[\log\frac{q}{p}\right] \in S_p\mathcal{P}_>(\Omega), $$
is an affine displacement for the exponential parallel transport Theorem 1 item 1. The inverse of the exponential chart s_p *is the exponential model*
$$ s_p^{-1} \colon S_p\mathcal{P}_>(\Omega) \ni v \mapsto \mathrm{e}^{v - K_p(v)} \cdot p\,, \quad K_p(v) = \log \mathbb{E}_p[\mathrm{e}^v]\,. $$
2. *The mapping*
$$ \mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto \eta_p(q) = \frac{q}{p} - 1 \in S_p\mathcal{P}_>(\Omega), $$
is an affine displacement for the mixture displacement Theorem 1 item 2. The inverse of the mixture chart η_p *is the mixture model*
$$ \eta_p^{-1} \colon S_p\mathcal{P}_>(\Omega) \supset \{v > -1\} \ni v \mapsto (1 + v) \cdot p\,. $$

Proof.

1. The generalized parallelogram law for the exponential displacement is
$$ \begin{aligned} &\left(\log\frac{q}{p} - \mathbb{E}_p\left[\log\frac{q}{p}\right]\right) + {}^{\mathrm{e}}\mathbb{U}_q^p\left(\log\frac{r}{q} - \mathbb{E}_q\left[\log\frac{r}{q}\right]\right) \\ &\qquad = \left(\log\frac{q}{p} - \mathbb{E}_p\left[\log\frac{q}{p}\right]\right) + \left(\log\frac{r}{q} - \mathbb{E}_p\left[\log\frac{r}{q}\right]\right) \\ &\qquad\qquad = \log\frac{r}{p} - \mathbb{E}_p\left[\log\frac{r}{p}\right]. \end{aligned} $$

2. The generalized parallelogram law for mixture displacement is

$$\left(\frac{q}{p}-1\right)+{}^{\mathrm{m}}\mathbb{U}_q^p\left(\frac{r}{q}-1\right)=\left(\frac{q}{p}-1\right)+\frac{q}{p}\left(\frac{r}{q}-1\right)=\frac{r}{p}-1\,.$$

□

The inverse exponential chart is defined on all p-contrasts by

$$s_p^{-1}(v)=\exp(v-K_p(v))\cdot p=e_p(v)$$

where

$$K_p(v)=\log\mathbb{E}_p[\mathrm{e}^v]$$

K is usually called the cumulant generating functional of v with respect to p. The inverse chart $s_p^{-1}(v)$ maps a vector p-contrast v back to a probability function. The expression $\exp(v-K_p(v))\cdot p$ involves an exponential map adjusted by a normalization factor $K_p(v)$. The quantity $\exp(v-K_p(v))$ ensures that the resulting vector after exponentiation is properly normalized to be a valid probability distribution. The term $e_p(v)=s_p^{-1}$ denotes this normalized exponential family.

Kullback–Leibler Divergence

The *Kullback–Leibler divergence* between two positive probability functions, p and q, is $\mathrm{D}(p\,\|q)=\mathbb{E}_p\left[\log\frac{p}{q}\right]$. Both partial functions

$$q\mapsto\mathrm{D}(p\,\|q)\quad\text{and}\quad p\mapsto\mathrm{D}(p\,\|q)$$

are commonly used in statistical optimization procedures. The primary reference is [28].

Example 13 (Expression of KL divergence)**.** *Let us find an expression for the Kullback–Leibler divergence by using a mixture chart on the first variable and the exponential chart on the second variable. Namely, if $p=(1+u)\cdot m$ and $q=\mathrm{e}^{v-K_m(v)}\cdot m$, the expression in the charts at m is*

$$\begin{aligned}\kappa_m\colon\{u\in S_m\mathcal{P}_>(\Omega)\,|\,1+u>0\}\times S_m\mathcal{P}_>(\Omega)\ni(u,v)\mapsto\\ \mathbb{E}_m\left[(1+u)\log\frac{1+u}{\mathrm{e}^{v-K_m(v)}}\right]=\\ \mathbb{E}_m[(1+u)\log(1+u)]-\mathbb{E}_m[(1+u)(v-K_m(v))]=\\ \mathbb{E}_m[(1+u)\log(1+u)]-\mathbb{E}_m[(1+u)v]-\mathbb{E}_m[(v-K_m(v))]=\\ H_m(u)+K_m(v)-\mathbb{E}_m[uv]\,,\end{aligned}\tag{26}$$

where $H_m(u)=\mathbb{E}_m[(1+u)\log(1+u)]$.

If we let $p=q$ in Equation (26), that is, $1+u=\mathrm{e}^{(v-K-m)(v)}$ and $\mathbb{E}_m[u]=\mathbb{E}_m[v]=0$, then $H_m(u)+K_m(v)=\langle u,v\rangle_m$.

Example 14 (Divergence and displacement)**.** *In Equation (26), we can turn back to the probability densities to obtain the* generalized parallelogram theorem *[29]*

$$\mathrm{D}(p\,\|q)=\mathrm{D}(p\,\|m)+\mathrm{D}(m\,\|q)-\langle u,v\rangle_m=\mathrm{D}(p\,\|m)+\mathrm{D}(m\,\|q)-\langle\eta_m(p),s_m(q)\rangle_m\,.$$

In particular, the generalized Pythagorean *theorem holds,*

$$\langle\eta_m(p),s_m(q)\rangle_m=0\quad\Rightarrow\quad\mathrm{D}(p\,\|q)=\mathrm{D}(p\,\|m)+\mathrm{D}(m\,\|q)\,.\tag{27}$$

The cumulant function $K_p(u)$ is central to the theory. It is the exponential normalizing constant of the exponential model $e_p(u)=\mathrm{e}^{u-K_p(u)}\cdot p$, and it equals $\mathrm{D}\big(p\,\|e_p(u)\big)$. Moreover, it has remarkable computational properties, summarized in the following theorem.

Notice that it is a formalism well known in statistical physics, where it is used to discuss the Gibbs–Boltzmann distribution. See, for example, [30].

Theorem 3 (Calculus of the cumulant functional)**.**

1. *The derivative of K_p at v in the direction h is*

$$dK_p(v)[h] = \mathbb{E}_{e_p(v)}[h] = \left\langle \frac{e_p(v)}{p} - 1, h \right\rangle_p . \tag{28}$$

2. *The second derivative of K_p at v in the directions h and k is*

$$d^2K_p(v)[h,k] = \mathrm{Cov}_{e_p(v)}(h,=)\left\langle {}^e\mathbb{U}_p^{e_p(v)}h, {}^e\mathbb{U}_p^{e_p(v)}k \right\rangle_{e_p(v)} = \left\langle {}^m\mathbb{U}_{e_p(v)}^{p}\, {}^e\mathbb{U}_p^{e_p(v)}h, k \right\rangle_p . \tag{29}$$

3. *The third derivative of K_p at v in the directions h_1, h_2, and h_3 is*

$$d^3K_p(v)[h_1,h_2,h_3] = \mathbb{E}_{e_p(v)}\left[\left({}^e\mathbb{U}_p^{e_p(v)}h_1 \right)\left({}^e\mathbb{U}_p^{e_p(v)}h_2 \right)\left({}^e\mathbb{U}_p^{e_p(v)}h_3 \right) \right].$$

4. *The cumulant functional K_p is convex, and its gradient is the covariance at p given by $\nabla K_p(v) = \frac{e_p(v)}{p} - 1$. The inverse gradient mapping is*

$$(\nabla K_p)^{-1} \colon S_p\mathcal{P}_>(\Omega) \ni w \mapsto \log(1+w) - \mathbb{E}_p[\log(1+w)],$$

and the convex conjugate is

$$K_p^*(w) \;=\; \left\langle w, (\nabla K)^{-1}w \right\rangle_p - K_p \circ (\nabla^{-1}K_p)(w) \;=\; \mathbb{E}_p[(1+w)\log(1+w)] . \tag{30}$$

Proof. The proof of the first three items follows through direct computation.

1.

$$\begin{aligned} dK_p(v)[h] &= \frac{d}{dt}K_p(v+th)\Big|_{t=0} = \frac{d}{dt}\log \mathbb{E}_p\left[e^{v+th}\right]\Big|_{t=0} \\ &= \frac{d}{dt}\left(\mathbb{E}_p\left[e^{v+th}\right]\right)^{-1}\mathbb{E}_p\left[he^{v+th}\right]\Big|_{t=0} = \left(\mathbb{E}_p[e^v]\right)^{-1}\mathbb{E}_p[he^v] \\ &= \mathbb{E}_{e_p(v)}[h] = \mathbb{E}_p\left[\frac{e_p(v)}{p}h\right] = \left\langle \frac{e_p(v)}{p} - 1, h \right\rangle_p \end{aligned}$$

2.

$$\begin{aligned} d^2K_p(v)[h,k] &= \frac{d}{dt}dK_p(v+tk)[h]\Big|_{t=0} = \frac{d}{dt}\left\langle \frac{e_p(v+tk)}{p} - 1, h \right\rangle_p\Big|_{t=0} \\ &= \left\langle \frac{d}{dt}e^{v+tk-K_p(v+tk)}\Big|_{t=0}, h \right\rangle_p = \left\langle \frac{e_p(v)}{p}\left(k - \frac{d}{dt}K_p(v+tk)\Big|_{t=0}\right), h \right\rangle_p \\ &= \left\langle \frac{e_p(v)}{p}\left(k - \mathbb{E}_{e_p(v)}[k]\right), h \right\rangle_p = \mathrm{Cov}_{e_p(v)}(h,k) \end{aligned}$$

3.

$$
\begin{aligned}
d^3K_p(v)[h_1,h_2,h_3] &= \frac{d}{dt}d^2K_p(v+th_3)[h_1,h_2]\Big|_{t=0} \\
&= \frac{d}{dt}\left\langle h_1, \frac{d}{dt}\left(\frac{e_p(v+th_3)}{p}\right)\Big(h_2-\mathbb{E}_{e_p(v+th_3)}[h_2]\Big)\right|_{t=0}\Bigg\rangle_p \\
= \left\langle h_1, \frac{d}{dt}\left(\frac{e_p(v+th_3)}{p}\right)\Big(h_2-\mathbb{E}_{e_p(v+th_3)}[h_2]\Big)\right|_{t=0}\Bigg\rangle_p &= \left\langle h_1, \frac{e_p(v)}{p}\Big(k-\mathbb{E}_{e_p(v)}[k]\Big), h\right\rangle_p \\
&= \left\langle \frac{e_p(v)}{p}h_1, \frac{e_p(v)}{p}\Big(h_2-\mathbb{E}_{e_p(v)}[h_2]\Big), \frac{e_p(v)}{p}h_3\right\rangle_p \\
&= \mathbb{E}_{e_p(v)}\Big[\Big(h_1-\mathbb{E}_{e_p(v)}[h_1]\Big)\Big(h_2-\mathbb{E}_{e_p(v)}[h_2]\Big)\Big(h_3-\mathbb{E}_{e_p(v)}[h_3]\Big)\Big]
\end{aligned}
$$

4. The convexity follows from items 1 and 2. The last statement is, in essence, a summary of the theorem itself.

□

3.6. Aitchison Geometry and Information Geometry

In Section Aitchison's Centered Log Ratio [5], we have shown that the Aitchison geometry of the probability simplex is an instance of affine geometry. The present section shows how to interpret the same formalism as using a particular atlas on the statistical bundle. For clarity, we admit some overlapping between the two sections.

Let $\mathcal{P}_>(\Omega)$ be the positive probability simplex over a finite set Ω, and let $T\mathcal{P}_>(\Omega)$ be its tangent space. Define the displacement mapping, $s_p : \mathcal{P}_>(\Omega) \to T\mathcal{P}_>(\Omega)$, for a reference point, $p \in \mathcal{P}_>(\Omega)$, and any $q \in \mathcal{P}_>(\Omega)$ as follows:

$$s_p(q) = \log\frac{q}{p} - \sum_{y\in\Omega}\frac{1}{\#\Omega}\log\frac{q(y)}{p(y)}\,,$$

where $\log\frac{q}{p}$ denotes the element-wise logarithm of the ratio of the probability distributions q and p, and $\sum_{y\in\Omega}\frac{1}{\#\Omega}\log\frac{q(y)}{p(y)}$ is the average of the log ratios over all elements $y \in \Omega$. This can also be written as follows:

$$s_p(q) = \log\frac{q}{p} - \mathbb{E}\left(\log\frac{q}{p}\right), \tag{31}$$

where $\mathbb{E}\left(\log\frac{q}{p}\right)$ is the uniform expectation of the log ratio.

The critical difference with exponential charts is the centering term, which is an expectation with respect to the uniform measure instead of the initial point, p.

The displacement mapping s_p of Equation (31) is surjective. For every $u \in T\mathcal{P}_>(\Omega)$, there exists a $q \in \mathcal{P}_>(\Omega)$, such that $s_p(q) = u$. The inverse mapping s_p^{-1} is given as follows:

$$s_p^{-1}(u) = \exp(u - K(u))\cdot p,$$

where the normalizing function $K(u)$ is defined as follows:

$$K(u) = \log\sum_{y\in\Omega} e^{u(y)}p(y) = \log\mathbb{E}_p[e^u]\,.$$

Notice the relationship between the two forms of the normalizing constant:

$$K(u) = \mathbb{E}\left(\log\frac{q}{p}\right).$$

This means that the normalizing constant $K(u)$ can be expressed as the p-expectation of the log ratio of the probability distributions q and p.

The parallelogram law holds true:

$$\left(\log\frac{q}{p}-\mathbb{E}\left(\log\frac{q}{p}\right)\right)+\left(\log\frac{r}{q}-\mathbb{E}\left(\log\frac{r}{q}\right)\right)=\log\frac{r}{p}-\mathbb{E}\left(\log\frac{r}{p}\right).$$

This law states that the sum of the centered log ratios between distributions q and p, and r and q, is equal to the centered log ratio between r and p. It ensures consistency in the displacement mappings across different distributions.

Example 15 (Uniform Distribution Case)**.** *If p is uniform, that is, $p(x)=D^{-1}$, where $D=\#\Omega$ (the number of elements in the set Ω), then the displacement mapping is simplified as follows:*

$$s_{D^{-1}}(q)(x)=\log q(x)-\sum_{y\in\Omega}\frac{1}{D}\log q(y)=\log\frac{q(x)}{\left(\prod_{y\in\Omega}q(y)\right)^{1/D}}.$$

The displacement mapping $s_{D^{-1}}(q)(x)$ involves taking the log of $q(x)$ and subtracting the average log of all $q(y)$. This is simplified to the log *of $q(x)$, divided by the geometric mean of $q(y)$ over all $y\in\Omega$.*

A few comments are relevant at this point. The particular chart, derived from the displacement mapping when p is uniform, matches a chart in the exponential atlas used in information geometry. Exponential affine geometry, which deals with exponential families of probability distributions, and Aitchison geometry, which deals with compositional data, are fundamentally the same, but they are represented differently. The tangent space is viewed as a statistical bundle in the exponential affine geometry. In Aitchison geometry, it is viewed as the tangent space of the simplex. This observation highlights the equivalence of these geometric frameworks when analyzing probability distributions or compositional data.

3.6.1. Mapping of n-Sequences

In Aitchison's geometry, the coordinates given in the chart above are called centered log-ratio CLR coefficients. In the present section, the CLR map transforms a set of (data) points in the (interior of the) simplex into vectors with $\#\Omega-1$ real-valued components. Statistical models for compositional data are often built and interpreted onto this set of transformed vectors. See, for example, [4,5].

Any n-sequence (n-sample) $q_1,\ldots,q_n\in\mathcal{P}_>(\Omega)$ is mapped to an n-sequence in the space of coordinates (contrasts) via the CLR operation as follows:

$$u_1=\mathrm{CLR}(q_1),\ldots,u_n=\mathrm{CLR}(q_n)\in T\mathcal{P}_>(\Omega).$$

The mapping translates the original probability distributions in the simplex $\mathcal{P}_>(\Omega)$ into coordinates in the tangent space $T\mathcal{P}_>(\Omega)$.

The *space of contrasts* $T\mathcal{P}_>(\Omega)$ is a vector space with an *inner product* defined as follows:

$$\langle u,v\rangle=\mathbb{E}(uv)=\mathrm{Cov}(u,v),$$

In the space of contrasts, $\langle u,v\rangle$ denotes the inner product of vectors u and v, $\mathbb{E}(uv)$ represents the expected value of their product, and $\mathrm{Cov}(u,v)$ represents their covariance. This inner product allows the computations required by us to perform all computations that require a vector-space structure: (1) vector addition and scalar multiplication are well defined in $T\mathcal{P}_>(\Omega)$; (2) the inner product induces a norm, making $T\mathcal{P}_>(\Omega)$ a normed

vector space; and (3) it can compute angles between vectors and project one vector onto another using this inner product.

In the compositional data literature, the formalism described above is called a *Bayes space*. See, for example, [6].

3.6.2. Mean, Variance, Velocity, and Velocity Curve

When data are analyzed, providing statistics (data summaries) such as the mean and variances is essential. In Aitchison geometry, these are computed into the tangent space. To fully appreciate their meaning, we highlight that the mean in $T\mathcal{P}_>(\Omega)$ is a function of geometric means in the original simplex, $\mathcal{P}_>(\Omega)$.

The *sample mean* of n sequences in centered log-ratio (CLR) coordinates is given as follows:

$$\frac{1}{n}\sum_{j=1}^{n} u_j = \frac{1}{n}\sum_{j=1}^{n} \mathrm{CLR}(q_j) = \frac{1}{n}\sum_{j=1}^{n}\left(\log q_j - \sum_{y\in\Omega}\frac{1}{\#\Omega}\log q_j(y)\right).$$

This is simplified to the following:

$$\frac{1}{n}\sum_{j=1}^{n}\left(\log q_j - \sum_{y\in\Omega}\frac{1}{\#\Omega}\log q_j(y)\right) = \log\left(\prod_{j=1}^{n} q_j\right)^{1/n} - \sum_{y\in\Omega}\frac{1}{\#\Omega}\log\left(\prod_{j=1}^{n} q_j(y)\right)^{1/n}.$$

It can also be written as follows:

$$\mathrm{CLR}\left(\left(\prod_{j=1}^{n} q_j\right)^{1/n}\right) = \mathrm{CLR}\left(\frac{\left(\prod_{j=1}^{n} q_j\right)^{1/n}}{\mathbb{E}\left(\left(\prod_{j=1}^{n} q_j\right)^{1/n}\right)}\right).$$

The *sample variance* is based on the squared norm:

$$\langle \mathrm{CLR}(q), \mathrm{CLR}(p)\rangle = \mathbb{E}(\mathrm{CLR}(q)\mathrm{CLR}(r)) = \mathbb{E}(\log q \log r) - \mathbb{E}(\log q)\mathbb{E}(\log r) = \mathrm{Cov}(\log q, \log r).$$

For CLR-transformed distributions, $\langle \mathrm{clr}(q), \mathrm{clr}(p)\rangle$ denotes their inner product, $\mathbb{E}(\mathrm{clr}(q)\mathrm{clr}(r))$ is the expectation of their product, $\mathbb{E}(\log q \log r)$ is the expectation of the product of log-transformed distributions, $\mathbb{E}(\log q)\mathbb{E}(\log r)$ is the product of their expectations, and $\mathrm{Cov}(\log q, \log r)$ represents their covariance.

A statistical model is a curve into $\mathcal{P}_>(\Omega)$, often inverse-mapped from a statistical bundle in $T\mathcal{P}_>(\Omega)$. In turn, the statistical bundle is estimated from some CLR-transformed data points.

Most differential algorithms for model estimation and optimization are based on the computation of the parametric rate of change of statistics, which, in turn, are based on the computation of velocities, accelerations, and gradients.

The *velocity* is as follows:

$$\frac{D}{dt}q(t) = \frac{d}{dt}\mathrm{CLR}(q(t)) = \frac{\dot{q}(t)}{q(t)} - \mathbb{E}\left(\frac{\dot{q}(t)}{q(t)}\right),$$

where $\frac{D}{dt}q(t)$ represents the velocity of the distribution $q(t)$, $\frac{d}{dt}\mathrm{clr}(q(t))$ is the time derivative of the clr-transformed distribution $q(t)$, $\frac{\dot{q}(t)}{q(t)}$ denotes the element-wise derivative of $q(t)$ divided by $q(t)$, and $\mathbb{E}\left(\frac{\dot{q}(t)}{q(t)}\right)$ is the expectation of $\frac{\dot{q}(t)}{q(t)}$, ensuring that the result has null expectation.

The *acceleration* is defined in the usual way. The velocity curve $t \mapsto \frac{D}{dt}q(t)$ is a curve in a vector space, and its derivative is the acceleration,

$$\frac{d}{dt}\frac{D}{dt}q(t) = \frac{\ddot{q}(t)}{q(t)} - \left(\frac{\dot{q}(t)}{q(t)}\right)^2 - \mathbb{E}\left(\frac{\ddot{q}(t)}{q(t)} - \left(\frac{\dot{q}(t)}{q(t)}\right)^2\right).$$

In this context, $\frac{d}{dt}\frac{D}{dt}q(t)$ denotes the acceleration of the distribution $q(t)$, with $\frac{\ddot{q}(t)}{q(t)}$ representing the element-wise second derivative of $q(t)$ normalized by $q(t)$, $\left(\frac{\dot{q}(t)}{q(t)}\right)^2$ as the square of the element-wise first derivative term and $\mathbb{E}\left(\frac{\ddot{q}(t)}{q(t)} - \left(\frac{\dot{q}(t)}{q(t)}\right)^2\right)$ as the expectation of the acceleration term, ensuring that the result is centered.

Notice that the velocity is a deviation from a mean value in the uniform probability. The corresponding intensity is the variance of $\frac{d}{dt}\log q(t)$ in the uniform probability. This is a legitimate and consistent estimation of the rate of variation, which is different from the more standard estimation using Fisher's information, which is a variance computed in the current probability, not the uniform probability.

The curves of zero acceleration are the curves of constant velocity. Such curves are commonly called *geodesics*. A geodesic from $p = \mathrm{CLR}(u)$ and initial velocity v has the following form:

$$q(t) = \mathrm{CLR}^{-1}(tv + u) = \exp(tu - K_p(u)) \cdot p\,,$$

where CLR^{-1} denotes the inverse centered log-ratio invertible chart, u represents the initial point in CLR coordinates, v indicates the initial velocity in CLR coordinates, $K_p(u)$ serves as the normalizing constant, and $\exp(tu - K_p(u)) \cdot p$ defines the geodesic point at time t.

Most optimization algorithms look for a critical point of a scalar field, that is, a point where the gradient is null. Given a definition of the inner product in the vector space of an affine space, the gradient of a real function Φ on the base set (scalar field) is a mapping $\operatorname{grad}\Phi$ from the base set to the vector space or the tangent space, that is, a vector field, such that, for all smooth curves $q(t)$, the following holds:

$$\frac{d}{dt}\Phi(q(t)) = \left\langle \operatorname{grad}\Phi(q(t)), \frac{D}{dt}q(t)\right\rangle,$$

where $\operatorname{grad}\Phi$ denotes the gradient of Φ, mapping from the base set to the tangent space to create a vector field. For a smooth curve, $q(t)$, $\frac{d}{dt}\Phi(q(t))$ represents the time derivative of Φ along $q(t)$, while the inner product $\left\langle \operatorname{grad}\Phi(q(t)), \frac{D}{dt}q(t)\right\rangle$ quantifies the rate of change of Φ in the direction of the velocity $\frac{D}{dt}q(t)$.

In Aitchison geometry, if $\mathrm{CLR}(q(t)) = u(t)$ and $q(t) = \mathrm{CLR}^{-1}(u(t))$, then the following applies:

$$\frac{d}{dt}\Phi \circ \mathrm{CLR}^{-1}(u(t)) = \mathbb{E}\Big(\operatorname{grad}\Phi \circ \mathrm{CLR}^{-1}(u(t))\,\dot{u}(t)\Big) = \frac{1}{\#\Omega}\sum_{x\in\Omega} \operatorname{grad}\Phi \circ \mathrm{CLR}^{-1}(u(x;t))\,\dot{u}(x;t)\,.$$

The composition $\Phi \circ \mathrm{CLR}^{-1}$ represents the function Φ applied to the inverse CLR transformation, where $\mathbb{E}\Big(\operatorname{grad}\Phi \circ \mathrm{CLR}^{-1}(u(t))\,\dot{u}(t)\Big)$ denotes the expected value of the product of the gradient of Φ composed with CLR^{-1}, and the derivative of $u(t)$; $\frac{1}{\#\Omega}\sum_{x\in\Omega}$ represents the average over all elements x in Ω, with $\operatorname{grad}\Phi \circ \mathrm{CLR}^{-1}(u(x;t))$ being the gradient of Φ composed with CLR^{-1} at the point $u(x;t)$, and $\dot{u}(x;t)$ denoting the time derivative of $u(x;t)$.

It follows that

$$\frac{\partial}{\partial u(x)}\Phi \circ \mathrm{CLR}^{-1}(u) = \frac{1}{\#\Omega}\operatorname{grad}\Phi \circ \mathrm{CLR}^{-1}(u)(x).$$

or, equivalently,

$$\frac{1}{\#\Omega}\text{grad}\,\Phi(q) = \nabla(\Phi \circ \text{CLR}^{-1}) \circ \text{CLR}\,,$$

where $\frac{\partial}{\partial u(x)}\Phi \circ \text{CLR}^{-1}(u)$ represents the partial derivative of $\Phi \circ \text{CLR}^{-1}$ with respect to $u(x)$, $\frac{1}{\#\Omega}\text{grad}\,\Phi \circ \text{CLR}^{-1}(u)(x)$ is the averaged gradient of Φ composed with CLR^{-1} at $u(x)$, $\nabla(\Phi \circ \text{CLR}^{-1})$ denotes the gradient of the composed function $\Phi \circ \text{CLR}^{-1}$, and $\nabla(\Phi \circ \text{CLR}^{-1}) \circ \text{CLR}$ is the gradient applied to CLR-transformed variables.

4. Product-Sample Spaces

The traditional presentation of the probability simplex's convex geometry uses the language of linear programming (LP); see, for example, [7] (Ch. IV). The following section is an aside about LP from [7] (§ IV.10) to be skipped if the reader knows this topic.

4.1. Short Recap of Linear Programming

Linear programming deals with function optimization under equality and inequality constraints, where both the objective function and the constraints are linear functions. Every LP problem admits two forms: primal and dual forms.

The *primal problem in the canonical form* is

$$\begin{aligned} &\text{Find} \quad c = \inf \sum_x c(x)r(x) \\ &\text{Subject to} \quad \sum_{x\in X} A(y,x)r(x) = \beta(y)\,, \quad \text{for all } y \in Y \text{ and } r(x) \geq 0 \text{ for all } x \in X, \end{aligned}$$

where X and Y are finite sets, A, β, and c are given real functions that define the problem, and r is a vector is the *primal plan*, representing the decision variables in the optimization problem.

Y typically represents the index set for the constraints in the optimization problem. Specifically, Y is the set of indices that correspond to the linear constraints $\sum_{x\in X} A(y,x)r(x) = \beta(y)$ that must be satisfied for each $y \in Y$. Each y in Y corresponds to a different constraint, where $A(y,x)$ represents the coefficients in the constraint matrix, and $\beta(y)$ represents the right-hand side values for the constraints.

The objective is to find the value c, which is the lower bound of the total cost $\sum_x c(x)r(x)$, where $c(x)$ represents the cost associated with x, and $r(x)$ represents the amount or level of activity at x. A plan, r, is *feasible* if it satisfies both constraints. That is, the linear equations must hold for all $y \in Y$, and $r(x)$ must be non-negative for all $x \in X$.

The *dual problem in standard form* is as follows:

$$\begin{aligned} &\text{Find} \quad \beta = \sup \sum_y \beta(y)\lambda(y) \\ &\text{Subject to} \quad \sum_{y\in Y} A(y,x)\lambda(y) \leq c(x) \quad \text{for all } x \in X, \end{aligned}$$

where λ is the *dual plan*. The objective is to find the value β, which is the upper bound of the sum $\sum_y \beta(y)\lambda(y)$.

Here, $\beta(y)$ represents the benefit or value associated with y, and $\lambda(y)$ represents the level or intensity at y. The constraint $\sum_y A(y,x)\lambda(y) \leq c(x)$ ensures that the linear combination of $\lambda(y)$ with coefficients $A(y,x)$ does not exceed $c(x)$ for each $x \in X$.

A general theorem relates the primal and dual problems. The *strong duality theorem* states that the values are equal, $c = \beta$, provided that a feasible primal plan exists. If, moreover, $c > -\infty$, then primal and dual optimal plans exist.

4.2. Product Sample Space

In this section, we assume a product of two finite sample spaces, $\Omega = \Omega_1 \times \Omega_2$, and discuss the probability simplex on such a sample space, $\mathcal{P}(\Omega_1 \times \Omega_2)$.

The two *marginalization mappings* are,

$$\Pi_1 : \mathcal{P}(\Omega_1 \times \Omega_2) \ni q(\cdot,\cdot) \mapsto \sum_{\omega_2} q(\cdot,\omega_2) \in \mathcal{P}(\Omega_1)\,,$$
$$\Pi_2 : \mathcal{P}(\Omega_1 \times \Omega_2) \ni q(\cdot,\cdot) \mapsto \sum_{\omega_1} q(\omega_1,\cdot) \in \mathcal{P}(\Omega_2)\,,$$

and the product marginalization mapping is

$$\Pi\colon \mathcal{P}(\Omega) \ni q \mapsto (\Pi_1 q, \Pi_2 q) \in \mathcal{P}(\Omega_1) \times \mathcal{P}(\Omega_2)\,.$$

Given marginal distributions $q_1 \in \mathcal{P}(\Omega)_1$ and $q_2 \in \mathcal{P}(\Omega)_2$, their counter images for Π are called *transport plans*, and

$$\Pi(q_1, q_2) = \{q \in \mathcal{P}(\Omega_1 \times \Omega_2) \mid \Pi_1 q = q_1, \Pi_2 q = q_2\}. \tag{32}$$

is the *set of transport plans*. It contains the probability functions of all joint distributions on $\Omega_1 \times \Omega_2$ with the marginals specified as q_1 and q_2. For each $q \in \Pi(q_1, q_2)$ the conditional $\omega_1 \mapsto q(\omega_1,\cdot)/q_1(\omega)$ is seen as a rule to move a given mass from ω_1 to Ω_2, which justifies the name of the transport plan.

The set of transport plans, $\Pi(q_1, q_2)$, is non-empty (because it contains the product of the two given marginals), and it is a convex, closed, and bounded set in the real space (because it is the intersection of the probability simplex with a hyperplane).

We can rewrite the marginalization conditions using the indicator function $(\xi = \omega)$, which is 1 if $\xi = \omega$ and 0 otherwise,

$$\sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} (\xi_1 = \omega_1) q(\omega_1,\omega_2) = q_1(\xi_1)\,,$$
$$\sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} (\xi_2 = \omega_2) q(\omega_1,\omega_2) = q_2(\xi_2)\,.$$

Define a function, $A : (\Omega_1 \cup \Omega_2) \times (\Omega_1 \times \Omega_2) \to \mathbb{R}$, as

$$\sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} A(\xi;\omega_1,\omega_2) q(\omega_1,\omega_2) = (q_1 \cup q_2)(\xi)\,,$$

that is,

$$A(\xi;\omega_1,\omega_2) = \begin{cases} (\xi_1 = \omega_1), & \text{if } \xi \in \Omega_1\,, \\ (\xi_2 = \omega_2), & \text{if } \xi \in \Omega_2\,, \end{cases}$$

so that we can write the marginalization equations into a single equation,

$$\sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} A(\xi;\omega_1,\omega_2) q(\omega_1,\omega_2) = (q_1 \cup q_2)(\xi)\,,$$

which provides a better form of LP formalism.

Given the *transport cost* $c : \Omega_1 \times \Omega_2 \to \mathbb{R}$, the optimal transport problem or Kantorovich problem looks for a transport plan with a minimal expected cost for a transport plan with given marginals, that is, with value the $\inf \mathbb{E}_q[c]$, where q varies in $\Pi(q_1, q_2)$. It is a primal problem in canonical form,

$$\text{Find} \quad c = \inf \sum_{\omega_1,\omega_2} c(\omega_1,\omega_2) q(\omega_1,\omega_2)\,, \tag{33}$$

$$\text{Subject to} \sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} A(\xi;\omega_1,\omega_2) q(\omega_1,\omega_2) = (q_1 \cup q_2)(\xi)\,, \quad q(\omega_1,\omega_2) \geq 0\,. \tag{34}$$

The dual problem in standard form is

$$\text{Find} \quad \beta = \sup\left(\sum_{\xi_1\in\Omega_1} q_1(\xi_1)\lambda_1(\xi_1) + \sum_{\xi_2\in\Omega_2} q_2(\xi_2)\lambda_2(\xi_2)\right), \tag{35}$$

$$\text{Subject to} \quad \lambda_1(\omega_1) + \lambda_2(\omega_2) \le c(\omega_1,\omega_2)\,. \tag{36}$$

A feasible transport plan exists, and the set of transport plans is closed and bounded. Hence, according to the strong duality theorem, the optimal values of the primal and dual problems are equal, and both primal and dual optimal solutions exist.

We can derive a more explicit statement. The optimal value of the primal problem equals the optimal value of the dual problem, namely

$$\sum_{\omega_1,\omega_2} c(\omega_1,\omega_2)\bar{q}(\omega_1,\omega_2) = \sum_{\omega_1\in\Omega_1} q_1(\omega_1)\bar{\lambda}_1(\omega_1) + \sum_{\omega_2\in\Omega_2} q_2(\omega_2)\bar{\lambda}_2(\omega_2) = \sum_{\omega_1,\omega_2} (\bar{\lambda}_1(\omega_1) + \bar{\lambda}_2(\omega_2))\bar{q}(\omega_1,\omega_2)\,.$$

where $\bar{q}$ is the optimal transport plan, and $\bar{\lambda}_1(\omega_1)$ are the dual variables.

The inequality

$$c \ge \lambda_1 \oplus \lambda_2(\omega_1,\omega_2) = \lambda_1(\omega_1) + \lambda_2(\omega_2)$$

ensures that, for the optimal transport plan, the cost $c(\omega_1,\omega_2)$ is equal to the sum of the dual variables $\bar{\lambda}_1(\omega_1)$ and $\bar{\lambda}_2(\omega_2)$ whenever $\bar{q}(\omega_1,\omega_2) \neq 0$, namely

$$c(\omega_1,\omega_2) = \bar{\lambda}_1(\omega_1) + \bar{\lambda}_2(\omega_2) \quad \text{provided} \quad \bar{q}(\omega_1,\omega_2) \neq 0\,.$$

In the transport literature, the equation above is called Kantorovich's theorem, which specifies that the cost equals the sum of two single variable functions in support of an optimal plan. See, for example, [31].

4.3. ANOVA Splitting of the Statistical Bundle

In this section, we use the language of effects' decomposition to analyze categorical tabular data. A classical treatment of this topic is [32].

The space of real random variables on $\Omega_1 \times \Omega_2$ is endowed with the inner product with weights provided via a strictly positive density, $q \in \mathcal{P}_>(\Omega_1 \times \Omega_2)$, and the resulting Euclidean space is denoted as $L^2(q)$. The linear sub-spaces of $L^2(q)$, which respectively express the *q-grand mean*, the two *q-simple effects*, and the *q-interaction*, are

$$\begin{aligned} B_0(q) &\sim \mathbb{R}, \\ B_1(q) &= \{f_1 \circ \Pi_1 \mid f_1 \in L^2_0(q_1)\}, \\ B_2(q) &= \{f_2 \circ \Pi_2 \mid f_2 \in L^2_0(q_2)\}, \\ B_{12}(q) &= (B_0(q) + B_1(q) + B_2(q))^{\perp}, \end{aligned}$$

where $\circ$ stands for the operation of functions composition, $f \circ g(x) = f(g(x))$, and the orthogonality is computed in the q-probability function, that is, in the inner product of $L^2(q)$, for which $\langle u, v\rangle_q = \mathbb{E}_q[uv]$. Each element of the space $B_0(q) + B_1(q) + B_2(q)$ has the form $u = u_0 + f_1 \circ \Pi_1 + f_2 \circ \Pi_2$, where $u_0 = \mathbb{E}_q[u]$ and f_1, f_2 are uniquely defined.

For each $q \in \mathcal{P}_>(\Omega)$, there exists a unique orthogonal splitting:

$$L^2(q) = \mathbb{R} \oplus (B_1(q) + B_2(q)) \oplus B_{12}(q).$$

Namely, each $u \in L^2(q)$ can be written uniquely as follows:

$$u = u_0 + (u_1 + u_2) + u_{12}\,,$$

where $u_0 = E_q[u]$, and $(u_1 + u_2)$ is the q-orthogonal projection of $u - u_0$ onto $(B_1(q) + B_2(q))$. That is,

$$\mathbb{E}_q(u_{12}|\Pi_1) = 0 \quad \text{and} \quad \mathbb{E}_q(u_{12}|\Pi_2) = 0\,.$$

In conclusion, the ANOVA splitting of the statistical bundle is

$$S_q\mathcal{P}_>(\Omega) = (B_1(q) + B_2(q)) \oplus B_{12}(q), \tag{37}$$

and we can interpret each factor in the decomposition as a vector space of velocities.

Example 16. *Given $q_1 \in \mathcal{P}_>(\Omega_1)$ and $q_2 \in \mathcal{P}_>(\Omega_2)$, consider $\Pi^\circ(q_1, q_2)$ to be the open set consisting of all (strictly) positive probability densities in the set of transport plans $\Pi(q_1, q_2)$.*

Consider a smooth curve with values in the interior of the open set of transport plans $t \mapsto q(t) \in \Pi^\circ(q_1, q_2)$ with $q(0) = q$. The velocity at q is given by $\dot{q}(0)$. Such a velocity is an interaction, $\dot{q}(0) \in B_{12}(q(0))$. In fact, $\mathbb{E}_{q_0}\left[\dot{q}(0)\right]$, and for all $f_i = g_i \circ \Pi_i \in B_i$, $i = 1, 2$,

$$\left\langle \dot{q}(0), f_i \right\rangle_{q(0)} = \frac{d}{dt}\mathbb{E}_{q(t)}[f_i] = \frac{d}{dt}\mathbb{E}_{q_i}[g_i] = 0\,.$$

Conversely, given any interaction $v \in B_{12}(q)$, consider the curve $t \mapsto q(t) = (1 + tv)q$. This curve stays in $\Pi^\circ(q_1, q_2)$ for t in a neighborhood of 0. Moreover, the interaction v is equal to the velocity at $t = 0$; i.e., $v = \dot{q}(0)$.

We define the *statistical bundle of positive transport plans* as

$$S\Pi^\circ(q_1, q_2) = \{(q, v) \,|\, q \in \Pi^\circ(q_1, q_2), v \in B_{12}(q)\} = \\ \left\{(q, v) \,\middle|\, q \in \Pi^\circ(q_1, q_2), \mathbb{E}_q(v|\Pi_1) = \mathbb{E}_q(v|\Pi_2) = 0\right\}.$$

The mixture transport can be restricted to $S\Pi^\circ(q_1, q_2)$ to provide a parallel transport between the fibers. If $q, r \in \Pi^\circ(q_1, q_2)$ and $v \in B_{12}(q)$, let us compute $w = {}^{\mathrm{m}}\mathbb{U}_q^r v = \frac{q}{r}v$ and check that $w \in B_{12}(r)$. If f is a function of Ω_i only, $i = 1, 2$, then

$$\langle w, f \rangle_r = \mathbb{E}_r\left[\frac{q}{r}vf\right] = \mathbb{E}_q[vf] = \langle v, f \rangle_q = 0\,.$$

In particular, it follows that the mixture affine displacement $(q, r) \mapsto \eta_q(r) = \frac{r}{q} - 1$ is well defined in the open transport plan; that is, $\eta_q(r) \in B_{12}(q)$. In fact, for all $f = g \circ \Pi_i \in B_i(q)$, $i = 1, 2$,

$$\left\langle \frac{r}{q} - 1, f \right\rangle_q = \mathbb{E}_q\left[\left(\frac{r}{q} - 1\right)f\right] = \mathbb{E}_r[f] - \mathbb{E}_p[f] = \mathbb{E}_{q_i}[g] - \mathbb{E}_{q_i}[g] = 0\,.$$

The open transport statistical model is a sub-affine space of the mixture affine space. The literature frequently describes it as "flat" in the sense that each mixture's auto-parallel curve between two points with the same margins has constant margins.

Example 17 (Mixed parametrization)**.** *The ANOVA splitting of the statistical bundle* (37) *suggests a mixed parameterization of $\mathcal{P}_>(\Omega_1 \times \Omega_2)$. Mixed parametrization combines elements of different parameterization schemes to describe the space. This approach is classical in the statistics of contingency tables, where mixed parametrizations are often used to model complex dependencies and interactions between variables.*

For each reference $q \in \Pi^\circ(q_1, q_2)$, consider the probability densities of the form

$$r = \mathrm{e}^{v_1 + v_2 - K_p(v_1 + v_2)} \cdot q\,, \quad v_1 + v_2 \in B_1(q) + B_2(q)\,.$$

The set of such densities is a sub-manifold of the exponential affine manifold usually termed an additive model *(§ 17.5 in [18]). The set of velocities at q is q-orthogonal to the space of transport plans that contain q.*

Example 18 (Optimal transport)**.** *Kantorovich's problem can be studied as a gradient-flow problem. Let $c\colon \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a transport cost with ANOVA decomposition at q:*

$$c = c_{0,q} + c_{1,q} + c_{2,q} + c_{12,q}\,.$$

The function $q \mapsto C(q) = \mathbb{E}_q[c]$ restricted to the open transport model $\Pi^\circ(q_1, q_2)$ has a (statistical) gradient in $S\Pi^\circ(q_1, q_2)$. For each generic smooth curve in $\Pi^\circ(q_1, q_2)$ with $q(0) = q$, namely

$$\frac{d}{dt}C(q(t)) = \frac{d}{dt}\mathbb{E}_{q(t)}[c] = \langle \dot\star q(t), c\rangle_{q(t)} = \left\langle \dot\star q(t), c_{12,q(t)}\right\rangle_{q(t)}\,.$$

The natural gradient in the bundle is

$$\operatorname{grad} C\colon \Pi^\circ(q_1, q_2) \ni q \mapsto c_{12,q} = c - c_{0,q} - (c_{1,q} + c_{2,q})\,,$$

and the gradient flow equation is

$$\dot\star q(t) = -c_{12,q(t)}\,.$$

The natural gradient mapping $q \mapsto c_{12,q}$ is an orthogonal projection on the complete set of transport plans $\Pi(q_1, q_2)$. It is remarkable to note that, in the more extensive set, the condition of zero gradients is compatible with the form of the optimal transport cost, according to Kantorivich.

5. Second-Order Computations and Mechanical Models

The notion of acceleration in the statistical bundle was introduced in Example 12 via the simple idea of computing the velocity's velocity. This was feasible because we deal with affine spaces. This section analyzes the definition in detail and considers its possible application to models inspired by mechanics. The term "mechanics" refers to the applications discussed, for example, the notion of a curve with minimal Fisher information, where we interpret the Fisher information along a curve as a Lagrangian function. This section requires an introductory notion of the language of analytic mechanics. See, for example, the monograph [33] or the introductions of the papers [12,34].

Consider a statistical bundle, $S\mathcal{P}_>(\Omega)$, with a cocycle of parallel transports, $\mathbb{U}^q_p$, $p, q \in \mathcal{P}_>(\Omega)$, and a compatible affine atlas, s_p, $p \in \mathcal{P}_>(\Omega)$. The statistical bundle is an affine space for the charts

$$s_{p,u}\colon S\mathcal{P}_>(\Omega) \ni (q, v) \mapsto \left(s_p(q), \mathbb{U}^p_q u - v\right) \in S_p\mathcal{P}_>(\Omega) \times S_p\mathcal{P}_>(\Omega)\,.$$

If $\Gamma\colon t \mapsto (q(t), v(t))$ is a curve in the statistical bundle, its expression at (p, u) is

$$\Gamma_{p,u}\colon t \mapsto \left(s_p(q(t)), {}^{\mathrm{e}}\mathbb{U}^p_{q(t)} v(t) - u\right)$$

and the expression of the velocity at (p, u) is

$$\frac{d}{dt}\Gamma_{p,u}(t) = \left(\frac{d}{dt}s_p(q(t)), \frac{d}{dt}\mathbb{U}^p_{q(t)} v(t)\right)\,.$$

The velocity in the *moving frame* centered at $(p, u) = (q(t), v(t))$ is

$$\dot\star\Gamma(t) = \left(\dot\star q(t), \frac{d}{dt}\mathbb{U}^p_{q(t)} v(t)\Big|_{p=q(t)}\right)\,.$$

Especially if $\Gamma(t) = (q(t), \dot\star q(t))$, we have the definition of *affine acceleration*,

$$\overset{**}{q}(t) = \frac{d}{dt}\mathbb{U}^{p}_{q(t)}\overset{\star}{q}(t)\Big|_{p=q(t)} . \tag{38}$$

In conclusion, each system of parallel transports has its own acceleraation. Precisely, the following applies:

- The *exponential acceleration is*

$$\frac{d}{dt}\,{}^{e}\mathbb{U}^{p}_{q(t)}\overset{\star}{q}(t)\Big|_{p=q(t)} = \frac{d}{dt}\left(\frac{\dot{q}(t)}{q(t)} - \mathbb{E}_{p}\left[\frac{\dot{q}(t)}{q(t)}\right]\right)\Big|_{p=q(t)} = \frac{\ddot{q}(t)q(t) - (\dot{q}(t))^{2}}{q(t)^{2}} - \mathbb{E}_{p}\left[\frac{\ddot{q}(t)q(t) - (\dot{q}(t))^{2}}{q(t)^{2}}\right]\Big|_{p=q(t)} = \frac{\ddot{q}(t)}{q(t)} - \left((\overset{\star}{q}(t))^{2} - \mathbb{E}_{q(t)}\left[(\overset{\star}{q}(t))^{2}\right]\right) . \tag{39}$$

- The *mixture acceleration* is

$$\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^{p}_{q(t)}\overset{\star}{q}(t)\Big|_{p=q(t)} = \frac{d}{dt}\frac{q(t)}{p}\frac{\dot{q}(t)}{q(t)}\Big|_{p=q(t)} = \frac{d}{dt}\frac{\dot{q}(t)}{p}\Big|_{p=q(t)} = \frac{\ddot{q}(t)}{q(t)} . \tag{40}$$

Remark 1. *It should be noted that, because of the very definition, the acceleration is zero if and only if the curve has a constant velocity in the proper parallel transport; that is,* $\overset{\star}{q}(s) = \mathbb{U}^{q(s)}_{q(t)}\overset{\star}{q}(t)$. *In particular, the exponential acceleration is null for 1-dimensional exponential models, and the mixture acceleration is null for 1-dimensional mixture models.*

The acceleration is appropriately defined as a triple of curves: the model, the velocity, and the acceleration. It is convenient to define a vector bundle whose fibers are products of two fibers of the statistical bundle. Moreover, to identify each fiber and to prepare for the use of the covariance inner product as a duality pairing, the two fibers are denoted, respectively, as $S_{p}\mathcal{P}_{>}(\Omega)$ and ${}^{*}S_{p}\mathcal{P}_{>}(\Omega)$.

The *full bundle* is

$${}^{1}S^{1}\mathcal{P}_{>}(\Omega) = \left\{(q,\eta,v)\,\middle|\, q\in\mathcal{P}_{>}(\Omega), \eta\in{}^{*}S_{q}\mathcal{P}_{>}(\Omega), v\in S_{q}\mathcal{P}_{>}(\Omega)\right\} .$$

With such a notation, any twice-smooth curve, q, is lifted to a curve in the full bundle,

$$(q,\overset{**}{q},\overset{\star}{q})\in{}^{1}S^{1}\mathcal{P}_{>}(\Omega) .$$

Remark 2. *The derivation of curves in the statistical bundle is also called* covariant derivation. *The treatment of covariant derivatives in standard monographs of differential geometry, such as, for example, [14,35], is considerably more involved than ours because we are considering the particular case of affine manifold endowed with an inner product instead of a general differentiable manifold or a general Riemannian manifold. In some cases, we must distinguish between exponential and mixture covariant derivatives.*

Below, we provide an example of a problem of interest that requires a second-order formalism. After that, we discuss how general arguments originating in mechanics can help solve such problems.

5.1. Curves with Minimal Fisher's Information

Let us consider a smooth curve in the open probability simplex, $q\colon [0,1]\ni t\mapsto q(t)\in\mathcal{P}_{>}(\Omega)$, connecting $q_{0}=q(0)$ to $q_{1}=q(1)$. The velocity of the curve in the statistical bundle is $]0,1[\ni t\mapsto\overset{\star}{q}(t)=\frac{d}{dt}\log q(t)\in S_{q(t)}\mathcal{P}_{>}(\Omega)$, the Fisher's score. Each fiber $S_{q}\mathcal{P}_{>}(\Omega)$ of the statistical bundle $SP_{>}(\Omega)$ is endowed with the covariance inner product $\langle v,w\rangle_{q}=\mathrm{Cov}_{q}(v,w)$, which defines the Riemannian metric of the statistical bundle.

With a mechanical analogy, the (kinetic) *energy* of the curve $q(\cdot)$ is the squared norm of the velocity, while the integral of the energy along the curve is the *action*,

$$L(q(\cdot)) = \int_1^0 \langle \dot\!^\star_q(t), \dot\!^\star_q(t) \rangle_{q(t)}\, dt = \int_0^1 \sum_{\omega\in\Omega} \frac{\dot q(\omega;t)^2}{q(\omega;t)}\, dt\,. \tag{41}$$

We can find a necessary condition for a minimum of the action by performing the classical Euler computation in the current context. Consider a curve in the statistical bundle $t \mapsto (q(t), v(t)) \in S\mathcal{P}_>(\Omega)$, such that $v(0) = v(1) = 0$. For each real θ, let us consider

$$t \mapsto q_\theta(t) = e^{\theta v(t) - K_{q(t)}(\theta v(t))} \cdot q(t)\,.$$

Such a perturbed smooth curve has the same endpoints as $q(\cdot)$, and the velocity is

$$\dot\!^\star_{q_\theta}(t) = \frac{d}{dt}\log q_\theta(t) = \theta\frac{d}{dt}v(t) - \frac{d}{dt}K_{q(t)}(\theta v(t)) + \frac{d}{dt}\log q(t) = \\ \theta\frac{d}{dt}v(t) - \frac{d}{dt}K_{q(t)}(\theta v(t)) + \dot\!^\star_q(t)\,.$$

At any external point of the action, we must have

$$\frac{d}{d\theta}\int_0^1 \langle \dot\!^\star_{q_\theta}(t), \dot\!^\star_{q_\theta}(t)\rangle_{q_\theta(t)}\, dt\Big|_{\theta=0} = 0\,.$$

The derivative of the integrand is

$$\frac{d}{d\theta}\sum\Big(\theta\dot v(t) - \frac{d}{dt}K_{q(t)}(\theta v(t)) + \dot\!^\star_q(t)\Big)^2 q_\theta(t) \\ = \sum 2\Big(\theta\dot v(t) - \frac{d}{dt}K_{q(t)}(\theta v(t)) + \dot\!^\star_q(t)\Big)\Big(\dot v(t) - \frac{d}{dt}\frac{d}{d\theta}K_{q(t)}(\theta v(t))\Big)q_\theta(t) \\ + \sum\Big(\theta\dot v(t) - \frac{d}{dt}K_{q(t)}(\theta v(t)) + \dot\!^\star_q(t)\Big)^2\Big(v(t) - \frac{d}{d\theta}K_{q(t)}(\theta v(t))\Big)q_\theta(t)\,.$$

The derivatives $\frac{d}{dt}$, $\frac{d}{d\theta}$, $\frac{d^2}{dtd\theta}$ of the cumulant function are, respectively,

$$\frac{d}{dt}K_{q(t)}(\theta v(t)) = \frac{d}{dt}\log\sum e^{\theta v(t)}q(t) = \\ \Big(\sum e^{\theta v(t)}q(t)\Big)^{-1}\Big(\sum\theta\dot v(t)e^{\theta v(t)}q(t) + \sum e^{\theta v(t)}\dot\!^\star_q(t)q(t)\Big) = \\ \mathbb{E}_{q_\theta(t)}\big[\theta\dot v(t) + \dot\!^\star_q(t)\big]\,,$$

$$\frac{d}{d\theta}K_{q(t)}(\theta v(t)) = \log\sum e^{\theta v(t)}q(t) = \Big(\sum e^{\theta v(t)}q(t)\Big)^{-1}\sum v(t)e^{\theta v(t)}q(t) = \mathbb{E}_{q_\theta(t)}[v(t)]\,,$$

$$\frac{d^2}{dtd\theta}K_{q(t)}(\theta v(t)) = \frac{d}{dt}\mathbb{E}_{q_\theta(t)}[v(t)] = \frac{d}{dt}\sum v(t)q_\theta(t) = \\ \sum\dot v(t)q_\theta(t) + \sum v(t)\dot q_\theta(t) = \mathbb{E}_{q_\theta(t)}\big[\dot v(t) + v(t)\dot\!^\star_{q_\theta}(t)\big]\,.$$

The values at $\theta = 0$ are, respectively, 0, 0, and $\mathbb{E}_{q(t)}\big[\dot v(t) + v(t)\dot\!^\star_q(t)\big]$.
The θ-derivative of the energy at $\theta = 0$ is

$$\frac{d}{d\theta}\sum\Big(\theta\dot v(t) - \frac{d}{dt}K_{q(t)}(\theta v(t)) + \dot\!^\star_q(t)\Big)^2 q_\theta(t)\Big|_{\theta=0} \\ 2\sum\dot\!^\star_q(t)\Big(\dot v(t) - \mathbb{E}_{q(t)}\big[\dot v(t) + v(t)\dot\!^\star_q(t)\big]\Big) + \sum(\dot\!^\star_q(t))^2 v(t)q(t) =$$

$$2\Big\langle \overset{\star}{q}(t), \dot{v}(t) - \mathbb{E}_{q(t)}[\dot{v}(t)] \Big\rangle_{q(t)} - 2\Big\langle \overset{\star}{q}(t)^2 - \mathbb{E}_{q(t)}\Big[\overset{\star}{q}(t)^2\Big], \ \Big\rangle_{q(t)}.$$

The variation in the energy in the direction $v(\cdot)$ is

$$\int_0^1 2\Big\langle \overset{\star}{q}(t), \dot{v}(t) - \mathbb{E}_{q(t)}[\dot{v}(t)] \Big\rangle_{q(t)} dt + \int_0^1 \Big\langle \overset{\star}{q}(t)^2 - \mathbb{E}_{q(t)}[\overset{\star}{q}(t)^2], v(t) \Big\rangle_{q(t)} dt\,,$$

where the first integral vanishes,

$$\sum_{x\in\Omega} \int_0^1 2\dot{q}(x;t)\dot{v}(x;t)\,dt = \sum_{x\in\Omega} -\int_0^1 2\ddot{q}(x;t)v(x;t)\,dt = \int_0^1 \Big\langle -2\frac{\ddot{q}(t)}{q(t)}, v(t) \Big\rangle_{q(t)} = 0\,.$$

In conclusion,

$$\int_0^1 \Big\langle \frac{\ddot{q}(t)}{q(t)} - \frac{1}{2}\Big(\overset{\star}{q}(t)^2 - \mathbb{E}_{q(t)}[\overset{\star}{q}(t)^2]\Big), v(t) \Big\rangle_{q(t)} dt = 0$$

for all $v(\cdot)$. One can easily check that the Euler–Lagrange equation holds,

$$\frac{\ddot{q}(t)}{q(t)} - \frac{1}{2}\Big(\overset{\star}{q}(t)^2 - \mathbb{E}_{q(t)}\Big[\overset{\star}{q}(t)^2\Big]\Big) = 0\,.$$

Remark 3. *The LHS of the equation above is the semi-sum of the two accelerations of Equations (39) and (40). As S.-I Amari first discussed, this particular acceleration is the Levi–Civita acceleration of the open probability simplex. See the monograph [2]. The present paper does not discuss geometries other than the affine ones.*

5.2. Lagrangian and Hamiltonian Formalism on the Probability Simplex

Consider a smooth curve, $t \mapsto (q(t), \eta(t), w(t))$, in the full statistical bundle ${}^1S^1\mathcal{P}_>(\Omega)$. Here, $q(t)$ represents the probability distribution, $w(t)$ is an element of the fiber at $q(t)$, interpreted as a velocity, and $\eta(t)$ is an element of the fiber at $q(t)$, interpreted as an acceleration. This second instance of the fiber is interpreted as a cotangent space in duality with the tangent space via the covariance inner product.

The derivative of the inner product $\langle \eta(t), w(t)\rangle_{q(t)}$ is given by the sum of the mixture covariant derivative of $\eta(t)$ and the exponential covariant derivative of $w(t)$,

$$\begin{aligned}\frac{d}{dt}\langle \eta(t), w(t)\rangle_{q(t)} = \frac{d}{dt}\Big\langle {}^{\mathrm{m}}\mathbb{U}^p_{q(t)}\eta(t), {}^{\mathrm{e}}\mathbb{U}^p_{q(t)}w(t)\Big\rangle_p = \\ \Big\langle \frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}^p_{q(t)}\eta(t), {}^{\mathrm{e}}\mathbb{U}^p_{q(t)}w(t)\Big\rangle_p + \Big\langle {}^{\mathrm{m}}\mathbb{U}^p_{q(t)}\eta(t), \frac{d}{dt}\,{}^{\mathrm{e}}\mathbb{U}^p_{q(t)}w(t)\Big\rangle_p = \\ \Big\langle \frac{{}^m D}{dt}\eta(t), w(t)\Big\rangle_{q(t)} + \Big\langle \eta(t), \frac{{}^m D}{dt}w(t)\Big\rangle_{q(t)}\,. \quad (42)\end{aligned}$$

Recall that, for a given scalar field, $F\colon \mathcal{P}_>(\Omega) \to \mathbb{R}$, the natural gradient of F is a section of the statistical bundle, such that, for all smooth curves, $t \mapsto q(t)$,

$$\frac{d}{dt}F(q(t)) = \langle \operatorname{grad} F(q(t)), \overset{\star}{q}(t)\rangle_{q(t)}\,.$$

This formulation ensures that the gradient grad F captures the rate of change of the scalar field F along the curve $q(t)$. We will need a more general gradient form that includes both the natural gradient and Equation (42).

Let there be given a real function, $F\colon {}^1S^1\mathcal{P}_>(\Omega) \times D \to \mathbb{R}$, where D is a domain of $\mathbb{R}^k$. For a generic smooth curve,

$$t \mapsto (q(t), \eta(t), w(t), c(t)) \in {}^1S^1\mathcal{P}_>(\times)D,$$

we want to write

$$\begin{aligned}&\frac{d}{dt}F(q(t),\eta(t),w(t),c(t))\\&=\left\langle \operatorname{grad} F(q(t),\eta(t),w(t),c(t)),\overset{\star}{q}(t)\right\rangle_{q(t)}+\left\langle \frac{{}^{m}D}{d\eta}(t),\nabla_{m}F(q(t),\eta(t),w(t),c(t))\right\rangle_{q(t)}\\&\quad+\left\langle \nabla_{e}F(q(t),\eta(t),w(t),c(t)),\frac{{}^{m}D}{dt}w(t)\right\rangle_{q(t)}+\nabla F(q(t),\eta(t),w(t),c(t))\cdot\dot{c}(t)\end{aligned}$$

These terms collectively describe the total derivative: how the function F changes along the curve with values in the full bundle and a parametric set, D. The four gradient components can be computed in the proper chart by varying only one component at a time.

The dually affine geometry of the statistical bundle is naturally well suited to describing the dynamics of probability densities in Lagrangian and Hamiltonian formalism, as presented, for example, in the textbook [33]. The Lagrangian formulation of mechanics derives the fundamental laws of force equilibrium from variational principles. In this analogy, the open probability simplex $\mathcal{P}_{>}(\Omega)$ corresponds to the configuration space, while the statistical bundle $S\mathcal{P}_{>}(\Omega)$, which includes velocities, is associated with the mechanical phase space. In this section, we describe the general methodology in an analogy with the standard treatment in mechanics, and we apply it to a specific but quite general example. For the full proof, see [12].

Given a *Lagrangian function*,

$$L\colon S\mathcal{P}_{>}(\Omega)\times[0,1]\to\mathbb{R}\,,$$

for each smooth curve $q:[0,1]\ni t\mapsto q(t)\in\mathcal{P}_{>}(\Omega)$, we consider its lift $t\mapsto(q(t),\overset{\star}{q}(t))\in S\mathcal{P}_{>}(\Omega)$. The action is the integral of the Lagrangian along the lifted curve over the fixed time interval $[0,1]$,

$$q(\cdot)\mapsto A(q)=\int_{1}^{0}L(q(t),\overset{\star}{q}(t),t)\,dt\,.$$

In mechanical terms, the action encapsulates the system's dynamics in the Lagrangian function and the path taken by the curve in the configuration space. The discussion of Section 5.1 is an instance of this approach applied to statistics. Two further examples follow.

Example 19 (Mixture curve)**.** *The mixture curve $q(t)=q_0+t(q_1-q_0)$ has the velocity*

$$\overset{\star}{q}(t)=\frac{d}{dt}\log(q_0+t(q_1-q_0))=\frac{q_1-q_0}{q_0+t(q_1-q_0)}.$$

The action of this curve for the Lagrangian $L(q,\overset{\star}{q})=\langle\overset{\star}{q},\overset{\star}{q}\rangle_q$ is

$$\begin{aligned}L(q(\cdot))=\int_0^1\sum\frac{(q_1-q_0)^2}{q_0+t(q_1-q_0)}\,dt.=\\\sum(q_1-q_0)\log(q_0+t(q_1-q_0))\big|_{t=0}^{t=1}=\sum(q_1-q_0)\log\frac{q_1}{q_0}=\\\left\langle\eta_{q_0}(q_1),s_{q_0}(q_1)\right\rangle_{q_0}.\end{aligned}$$

Example 20 (Exponential curve)**.** *The exponential curve $q(t)=e^{tu-K_{q_0}(tu)}\cdot q_0$ connects q_0 to $q_1=q(1)$. The velocity is*

$$u-dK_{q_0}(tu)[u]=u-\mathbb{E}_{q(t)}[u]\,.$$

The action of this curve for the Lagrangian $L(q, \overset{\star}{q}) = \left\langle \overset{\star}{q}, \overset{\star}{q} \right\rangle_q$ *is*

$$L(q(\cdot)) = \int_0^1 \mathbb{E}_{q(t)} \left[(u - \mathbb{E}_{q(t)}[u])^2 \right] dt \, .$$

We use the equality

$$\mathbb{E}_{q(t)} \left[(u - \mathbb{E}_{q(t)}[u])^2 \right] = \frac{d}{dt} \mathbb{E}_{q(t)}[u]$$

and $\mathbb{E}_{q_0}[u] = 0$ *to get*

$$L(q(\cdot)) = \int_0^1 \frac{d}{dt} \mathbb{E}_{q(t)}[u] dt = \mathbb{E}_{q_1}[u] = \mathbb{E}_{q_1} \left[\log \frac{q_1}{q_0} - \mathbb{E}_{q_0} \left[\log \frac{q_1}{q_0} \right] \right] =$$
$$\mathbb{E}_{q_1} \left[\log \frac{q_1}{q_0} \right] - \mathbb{E}_{q_0} \left[\log \frac{q_1}{q_0} \right] = D(q_1 \parallel q_0) + D(q_0 \parallel q_1) \, .$$

Hamilton's principle states that the action $A(q)$ has a critical point at the solution q in the space of curves on $\mathcal{P}_>(\Omega)$ if the Euler–Lagrange equation holds,

$$\frac{{}^m D}{dt} \nabla_e L(q(t), \overset{\star}{q}(t), t) = \operatorname{grad} L(q(t), \overset{\star}{q}(t), t) \, .$$

This follows from a general principle in mechanics. See [36] for a detailed proof in the statistical case.

Let us define the Hamiltonian corresponding to a Lagrangian with the usual duality argument. The gradient mapping and Legendre transform are applied at each fixed density $q \in \mathcal{P}_>(\Omega)$ and each time t. The partial mapping $S_q\mathcal{P}_>(\Omega) \ni w \mapsto L_{q,t}(w) = L(q, w, t)$ has a gradient mapping in the covariance duality of ${}^*S_q\mathcal{P}_>(\Omega) \times S_q\mathcal{P}_>(\Omega)$, namely $w \mapsto \nabla_e L(q, w, t)$.

The Hamiltonian $H(q, \eta, t)$ is the Legendre transform of the Lagrangian,

$$H(q, \eta, t) = \left\langle \eta, (\nabla_e L_{q,t})^{-1}(\eta) \right\rangle_q - L(q, (\nabla_e L_{q,t})^{-1}(\eta)).$$

Here, $(\nabla_e L_{q,t})^{-1}(\eta)$ denotes the inverse of the exponential gradient mapping applied to η.

These components collectively describe the system's dynamics in terms of Hamilton's principle and Hamiltonian formalism.

If the curve $t \mapsto q(t)$ is a solution of the Euler–Lagrange equation, then the curve $t \mapsto \zeta(t) = (q(t), \eta(t))$ lies in the cotangent bundle ${}^*SE(\mu)$. The momentum $\eta(t)$ is defined as:

$$\eta(t) = \nabla_e L(q(t), \overset{\star}{q}(t), t).$$

In this context, the mixture bundle ${}^*SE(\mu)$ acts as the cotangent bundle in classical mechanics.

The Hamilton equations describe the evolution of the system:

$$\begin{cases} \frac{{}^m D}{dt} \eta(t) = -\nabla H(q(t), \eta(t), t) \\ \overset{\star}{q}(t) = \nabla_m H(q(t), \eta(t), t). \end{cases}$$

These equations describe the system's evolution in terms of the Hamiltonian function $H(q, \eta, t)$. The time derivative of the Hamiltonian H along the trajectory is given by the following:

$$\frac{d}{dt} H(q(t), \eta(t), t) = \frac{\partial}{\partial t} H(q(t), \eta(t), t).$$

This equation indicates that the system's total energy changes according to the Hamiltonian's explicit time dependence.

Example 21 (Compare with Section 5.1)**.** *Given the Lagrangian* $L(q, w) = \frac{1}{2}\langle w, w\rangle_q$*, the Hamiltonian obtained via the Legendre transform is* $H(q, \eta) = \frac{1}{2}\langle \eta, \eta\rangle_q$*. The Lagrangian represents the kinetic energy, while the Hamiltonian represents the system's total energy.*

The gradients of the Hamiltonian and Lagrangian are computed as follows:

$$\begin{aligned} \nabla H(q, \eta) &= -\frac{1}{2}(\eta^2 - E_q[\eta^2]) \,, \\ \nabla_m H(q, \eta) &= \eta \,, \\ \nabla L(q, w) &= \frac{1}{2}(w^2 - E_q[w^2]) \,, \\ \nabla_e L(q, w) &= w \,. \end{aligned}$$

The Euler–Lagrange equation for $\overset{\star}{q} = w \in {}^*SE(\mu)$ *is given by the following:*

$$\frac{{}^m D}{dt}\overset{\star}{q}(t) = \frac{1}{2}(\overset{\star}{q}(t)^2 - E_{q(t)}[\overset{\star}{q}(t)^2]) \,,$$

where

$$\frac{{}^m D}{dt}\overset{\star}{q}(t) = \frac{\ddot{q}(t)}{q(t)} \,.$$

The Hamilton equations are

$$\begin{cases} \frac{{}^m D}{dt}\eta(t) = \frac{1}{2}(\eta^2 - E_q[\eta^2]), \\ \overset{\star}{q}(t) = \eta(t), \end{cases}$$

The conserved energy is $H(q(t), \eta(t)) = \frac{1}{2}E_q\left[\frac{\dot{q}(t)^2}{q(t)}\right]$*; that is, the Fisher's information is constant in the solution.*

6. Conclusions

This research has provided a general and cohesive geometric framework for studying several geometries of the probability simplex, including Aitchison's and Amari's dual affine and transport geometries. By providing a cogent presentation of these various geometric approaches, the framework has the potential to significantly improve the comprehension and manipulation of probability distributions in statistical modeling and machine learning.

This unified approach is based on an axiomatic approach called the affine geometry of the statistical bundle. This approach is beneficial when the accurate handling of compositional data is crucial, such as when developing computational tools, statistical model optimization, and gradient flow optimization are involved. In particular, our formalism provides a principled treatment of second-order computations, such as accelerations.

As a notable contribution, we have shown that Aitchison CLR provides the charts of an affine space isomorphic to the information geometry space defined by the exponential presentation of probability functions. Moreover, both approaches allow for a statistically natural calculus concerning gradients of scalar fields and gradient flows.

The framework can increase performance in activities ranging from data processing to implementing sophisticated machine learning algorithms by facilitating more accurate and efficient modeling, providing a consistent geometric perspective. Moreover, we have shown how other geometries of great current interest, those associated with optimal transport, fit into the scheme of gradient flows in the statistical bundle.

The paper's final part contributed to discussing simple but relevant examples of statistical models controlled via second-order differential equations. This opens the discussion of dynamic statistical models with unusual time behavior, such as periodicity, without any reference to any dynamical system in the sample space.

Author Contributions: Conceptualization, G.P.; formal analysis, G.P.; investigation, G.P.; writing—original draft, G.P. and M.S.; writing—review and editing, G.P. and M.S.; supervision, G.P.; project administration, G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This project received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie GA No. 101034449.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The present paper is partly based on unpublished lecture notes by M. Shoaib after G. Pistone's lectures at the seminar that Eva Riccomagno organized in 2022–2023. The authors gratefully thank Eva Riccomagno for her invitation, support, supervision, and revision of an early version of the manuscript. Collegio Carlo Alberto, Italy, partially supports G. Pistone. He is a member of GNAFA-INDAM.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Amari, S. Dual connections on the Hilbert bundles of statistical models. In *Geometrization of Statistical Theory*; ULDM Publishing: Lancaster, PA, USA, 1987; pp. 123–151.
2. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Translated from the 1993 Japanese original by Daishi Harada; American Mathematical Society: Providence, RI, USA, 2000; Volume 191, 206p.
3. Pistone, G.; Sempi, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.* **1995**, *23*, 1543–1561. [CrossRef]
4. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. *Modelling and Analysis of Compositional Data*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2015. [CrossRef]
5. Aitchison, J. *The Statistical Analysis of Compositional Data*; Monographs on Statistics and Applied Probability; Chapman & Hall: London, UK, 1986; 416p.
6. Egozcue, J.J.; Pawlowsky-Glahn, V.; Tolosana-Delgado, R.; Ortego, M.I.; van den Boogaart, K.G. Bayes spaces: Use of improper distributions and exponential families. *Rev. Real Acad. Cienc. Exactas Fis. Nat. Ser. Mat.* **2012**, *107*, 475–486. [CrossRef]
7. Barvinok, A. *A Course in Convexity*; Graduate Studies in Mathematics; American Mathematical Society: Providence, RI, USA, 2002; Volume 54, 366p.
8. Lay, S.R. *Convex Sets and Their Applications*; Courier Corporation: North Chelmsford, MA, USA, 2007.
9. Amari, S.I. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016; 374p.
10. Nomizu, K.; Sasaki, T. *Affine Differential Geometry: Geometry of Affine Immersions*; Cambridge Tracts in Mathematics; Cambridge University Press: Cambridge, UK, 1994.
11. Pistone, G. Information Geometry of the Probability Simplex: A Short Course. *Nonlinear Phenom. Complex Syst.* **2020**, *23*, 221–242. [CrossRef]
12. Chirco, G.; Malagò, L.; Pistone, G. Lagrangian and Hamiltonian dynamics for probabilities on the statistical bundle. *Int. J. Geom. Methods Mod. Phys.* **2022**, *19*, 2250214. [CrossRef]
13. Bourbaki, N. *Variétés Differentielles et Analytiques. Fascicule de réSultats/Paragraphes 1 à 7*; Springer Science and Business Media: New York, NY, USA, 1971.
14. Lang, S. *Differential and Riemannian Manifolds*, 3rd ed.; Graduate Texts in Mathematics; Springer: New York, NY, USA, 1995; 364p.
15. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather. Rev.* **1950**, *78*, 1–3. [CrossRef]
16. Weyl, H. *Space—Time—Matter*; Translation of the 1921 raum zeit materie; Dover: New York, NY, USA, 1952.
17. Brown, L.D. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*; Institute of Mathematical Statistics: Beachwood, OH, USA, 1986; 283p.
18. Efron, B.; Hastie, T. *Computer Age Statistical Inference*; IMS Monographs; Cambridge University Press: Cambridge, UK, 2016; Volume 19, 475p.
19. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006; 748p.
20. Ohara, A.; Amari, S. Differential geometric structures of stable state feedback systems with dual connections. *Kybernetika* **1994**, *30*, 369–386.
21. Cafaro, C.; Alsing, P.M. Decrease of Fisher information and the information geometry of evolution equations for quantum mechanical probability amplitudes. *Phys. Rev. E* **2018**, *97*, 042110. [CrossRef] [PubMed]
22. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. *Information Geometry*; Ergebnisse der Mathematik und ihrer Grenzgebiete; Springer: Cham, Switzerland, 2017; 407p. [CrossRef]
23. Ohara, A. On Affine Immersions of the Probability Simplex and Their Conformal Flattening; In Proceedings of the Geometric Science of Information: Third International Conference, GSI 2017, Paris, France, 7–9 November 2017; Springer: Cham, Switzerland, 2017; pp. 247–254. [CrossRef]

24. Duy, T.T.; Nguyen, L.V.; Nguyen, V.; Trung, N.; Abed-Meraim, K. Fisher information neural estimation. In Proceedings of the 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 2111–2115. [CrossRef]
25. Ly, A.; Marsman, M.; Verhagen, J.; Grasman, R.; Wagenmakers, E. A Tutorial on Fisher Information. *J. Math. Psychol.* **2017**, *80*, 40–55. [CrossRef]
26. Wei, X.X.; Stocker, A. Mutual Information, Fisher Information, and Efficient Coding. *Neural Comput.* **2016**, *28*, 305–326. [CrossRef] [PubMed]
27. Gibilisco, P.; Pistone, G. Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* **1998**, *1*, 325–347. [CrossRef]
28. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
29. Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability* **1975**, *3*, 146–158. [CrossRef]
30. Landau, L.D.; Lifshits, E.M. *Statistical Physics*, 3rd ed.; Course of Theoretical Physics; Butterworth-Heinemann: Oxford, UK, 1980; Volume 5.
31. Peyré, G.; Cuturi, M. Computational Optimal Transport. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
32. Fienberg, S.E. *The Analysis of Cross-Classified Categorical Data*; MIT Press: Cambridge, MA, USA, 1980; 198p.
33. Arnold, V.I. *Mathematical Methods of Classical Mechanics*; Graduate Texts in Mathematics; Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein; Springer: New York, NY, USA, 1989; 516p.
34. Pistone, G. Lagrangian Function on the Finite State Space Statistical Bundle. *Entropy* **2018**, *20*, 139. [CrossRef] [PubMed]
35. Do Carmo, M.P. *Riemannian Geometry*; Mathematics: Theory & Applications; Translated from the second Portuguese edition by Francis Flaherty; Birkhäuser Boston Inc.: Basel, Switzerland, 1992; 300p.
36. Chirco, G.; Pistone, G. Dually affine Information Geometry modeled on a Banach space. *arXiv* **2022**, arXiv:2204.00917. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.