# Bruhat-Tits geometry of the probability simplex and its connection to the maximum entropy method

Henryk Gzyl[†] and Frank Nielsen[⋆]

[†]Centro de Finanzas IESA, Caracas
[†]Venezuela
[†]henryk.gzyl@iesa.edu.ve
[⋆]Sony Computer Science Laboratories, Inc.
[⋆]Tokyo, Japan
[⋆]Frank.Nielsen@acm.org

**Abstract**

The use of geometrical methods in statistics has a long and rich history highlighting many different aspects. In this paper, we consider the finite dimensional case since the basic ideas can be extended similarly to the infinite-dimensional case. Our aim is to understand exponential families of probabilities on a finite set from a geometrical point of view.

For that purpose, we consider a Riemannian geometry defined on the set of positive vectors in a finite-dimensional space. In this space, the probabilities on a finite set comprise a submanifold in which exponential families correspond to geodesic surfaces. We shall also obtain a geometric/dynamic interpretation of Jaynes' method of maximum entropy.

**Keywords**: Geometry on positive vectors, geometry on the probability simplex, logarithmic distance on the class of positive vectors,The maximum entropy method.

# 1 Introduction and preliminaries

## 1.1 Geometry of the probability distributions

Geometry and statistics have been intertwined for some time already, mainly through the study of differential-geometric structures in the space of parameters that characterize families of distributions. Consider for example the seminal works of Hotelling [13] and Rao [28], and the works of Amari [2, 3], Amari et al. [4], Efron [9], Barndorff-Nielsen [5], Vajda [30], and more recently Pistole and Semi [27], Pistone and Rogatin [26]. In all of these works a special emphasis is laid upon *exponential families*. In information geometry, the geometry of exponential families is elucidated by a dually flat manifold [2] (that is, a pair of torsion-free flat affine connections that are metric-compatible). A categorical distribution is a discrete probability distribution that describes the possible results of a random variable that can take on one of $n$ possible choices $\mathcal{X} = \{1, \ldots, n\}$. The space of all categorical distributions form an exponential family commonly called the probability simplex: $\Delta_{n-1}$. In information geometry, the probability simplex can also be viewed as a mixture family [3]. Mixture families can also be modeled by dually flat manifolds [24].

Consider as well the more recent work by Imparato and Trivelato [14] and Pistone [25], which certainly belongs to the same class of models, and, even though the techniques are quite different from those developed here, the similarities are many. And to finish this short list of references, we cite the nice textbook by Calin and Udriste [8].

A non-parametric approach based on an intrinsic geometry on the space of probability densities, and to understand exponential families in that set up was put forward in Gzyl and Recht [10]-[11]. The approach considered in that work was too general, and less germane than the Riemannian approach considered below. The excess generality in those papers comes from the use of algebras over the complex field. Even so, quite a bit of interesting connections between families of exponential type, geodesics and polynomials of binomial type was established in the second of the two last references mentioned. Let us point out another recent approach which considered the Hilbert geometry for the probability simplex and its connection with the Birhoff cone projective geometry of the positive orthant [23].

Actually there has been much interest in that geometry, not in $\mathbb{R}^n$ but in the space of symmetric positive-definite matrices. The reader can check with Lang [17] in which a relation of this geometry to Bruhat-Tits spaces is explored, or in Lawson and Lim [18] or Moakher [21] were the geometric mean property for sets of symmetric positive-definite matrices is established. More recently Arsigny et al. [1] described the use of that geometry to deal with a variety of applications, and Schwartzman [29] used that geometric setup to study lognormal distributions with values in the class of symmetric matrices.

### 1.2 Paper outline

The paper is organized as follows. In Section 2 we present the essentials about the geometry on the set of strictly positive vectors in a finite dimensional space. Here we present the finite dimensional case only for two reasons: First because all geometric ideas are already present in this case, and second, for not to unencumber the manuscript with technical details pertinent to the infinite dimensional case needed to deal with probability densities. In Section 3 we regard probabilities on finite sets as a submanifold of the set of strictly positive numbers, and verify that exponential probability distributions correspond to geodesic hyper-surfaces in that manifold. We then provide a geometric interpretation for the method of maximum entropy [15]: The Lagrange multipliers (which are related to intensive magnitudes in statistical physics, correspond to travel time along a geodesic from an initial distribution to the distributions satisfying given constraints). In section 5 we recall, for the sake of completeness, the role of the logarithmic entropy function as a Lyapunov function for standard Markovian dynamics.

## 2 The geometry on the space of positive real-valued vectors

The results described next are taken almost verbatim from [10]. The basic idea for the geometry that we are to define on the positive vectors in $\mathbb{R}^n$, is that we can think about them as functions $\boldsymbol{\xi} : \mathcal{X} = \{1, ..., n\} \to \mathbb{R}$, and all standard arithmetical operations either as *componentwise operations* among vectors or *pointwise operations* among functions. We shall denote by $\mathcal{M} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}(i) > 0, i = 1, ...n\}$ the set of all positive vectors (the positive orthant cone). $\mathcal{M}$ is an open set in $\mathbb{R}^n$ which is trivially a manifold over $\mathbb{R}^n$, having $\mathbb{R}^n$ itself as tangent space $T_{\boldsymbol{x}}M$ at each point $x \in \mathcal{M}$.

The set $\mathcal{M}$ plays the role that the positive definite matrices play in the work by Lang, Lawson and Lim and Moakher mentioned above. The role of the *group* [19, **?**] of invertible matrices in those references is to be played by

$$G = \{\boldsymbol{g} \in \mathbb{R}^n \mid g(i) \neq 0,\ i = 1, ..., n\}.$$

Group $G$ clearly is an Abelian group with respect to the standard componentwise product. The identity $\boldsymbol{e} = (1, \ldots, 1) \in G$ is the vector with all its components equal to 1. In order to define a scalar product at each $T_{\boldsymbol{x}}\mathcal{M}$ we use a transitive action of $G : \mathcal{M} \to \mathcal{M}$ of $G$ on $\mathcal{M}$ as follows. Set

$$\tau_{\boldsymbol{g}}(\boldsymbol{x}) = \boldsymbol{g}^{-1}\boldsymbol{x}\boldsymbol{g}^{-1}.$$

This action is clearly transitive on $\mathcal{M}$, and can be defined in the obvious way as an action on $\mathbb{R}^n$. We transport the scalar product on $T_{\boldsymbol{e}}M$ to any $T_{\boldsymbol{x}}M$ as follows.

The scalar product between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ at $T_{\boldsymbol{e}}M$ is defined to be the standard Euclidean product $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle = \sum \xi_i \eta_i$, and we shall switch between $\xi(i)$ and $\xi_i$ whenever is convenient for typographical reasons. We now use the fact that $\boldsymbol{x} = \tau_{\boldsymbol{g}}(\boldsymbol{e})$ with $\boldsymbol{g} = \boldsymbol{x}^{-1/2}$ to define the scalar product transported to $T_{\boldsymbol{x}}M$ by

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{x}} \equiv \langle \boldsymbol{x}^{-1}\boldsymbol{\xi}, \boldsymbol{x}^{-1}\boldsymbol{\eta} \rangle = \langle \boldsymbol{x}^{-2}\boldsymbol{\xi}, \boldsymbol{\eta} \rangle.$$

That is, we transport the vectors back to $T_{\boldsymbol{e}}M$ and compute their scalar product there. That scalar product allows us to define the length of a differentiable curve as follows:

Let $\boldsymbol{x}(t)$ be a differentiable curve in $\mathcal{M}$, its length is given by

$$\int_0^1 \sqrt{\langle \dot{\boldsymbol{x}}, \dot{\boldsymbol{x}} \rangle_{\boldsymbol{x}}} dt.$$

With this, the distance between $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}$ is defined by the usual formula

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \inf \left\{ \int_0^1 \sqrt{\langle \dot{\boldsymbol{x}}, \dot{\boldsymbol{x}} \rangle_{\boldsymbol{x}}} \mathrm{d}t \,|\, \boldsymbol{x}(t) \text{ differentiable such that } \boldsymbol{x}_1 = \boldsymbol{x}(0) \;\; \boldsymbol{x}_2 = \boldsymbol{x}(1) \right\}. \tag{2.1}$$

Actually, it also happens that the geodesics minimize the *action functional*

$$\int_0^1 \mathcal{L}(\dot{\boldsymbol{x}}(t), \boldsymbol{x}(t)) dt, \quad \text{with} \quad \mathcal{L}(\dot{\boldsymbol{x}}(t), \boldsymbol{x}(t)) = \frac{1}{2} \langle \dot{\boldsymbol{x}}, \dot{\boldsymbol{x}} \rangle_{\boldsymbol{x}}, \tag{2.2}$$

It takes an application of the Euler-Lagrange formula to see that the equation of the geodesics in this metric is

$$\ddot{\boldsymbol{x}}(t) = \boldsymbol{x}^{-1} \dot{\boldsymbol{x}}^2, \;\; \boldsymbol{x}(0) = \boldsymbol{x}_1, \;\; \boldsymbol{x}(1) = \boldsymbol{x}_2, \tag{2.3}$$

the solution to which is

$$\boldsymbol{x}(t) = \boldsymbol{x}_1 e^{-t \ln(\boldsymbol{x}_1/\boldsymbol{x}_2)} = \boldsymbol{x}_2^t \boldsymbol{x}_1^{(1-t)}. \tag{2.4}$$

This is the $e$-geodesic in information geometry [3], also called a Bhattacharyya arc.

**Comments.** The choice of sign in the exponent is arbitrary. We choose the sign as negative now so that a negative sign does not occur when we deal with the maximum entropy method below. It should also be clear that the transport mentioned above coincide with the geodesic transport just defined.

The geometric construction carried out above was to render as natural the following distance between positive vectors in $\mathcal{M}$. The geodesic distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = \sum_{i=1}^{n} \left( \ln x_1(i) - \ln x_2(i) \right)^2. \tag{2.5}$$

Notice as well that instead of solving (2.3) with initial and final conditions, we might as well consider the solution to (2.3) subject to $\boldsymbol{x}(0) = \boldsymbol{x}$, and $\dot{\boldsymbol{x}}(0) = \boldsymbol{\xi}$, which is clearly given by the (exponential) mapping $\boldsymbol{x} e^{-t\boldsymbol{\xi}}$.

The following result is taken verbatim from Gzyl (2017). It summarizes the main results from Chapter 5 of Lang (1995).

**Theorem 2.1** *With the notations introduced above we have:*

**1)** *The exponential mapping is metric preserving through the origin.*

**2)** *The derivative of the exponential mapping is measure preserving, that is,* $\exp'(\boldsymbol{\xi})\boldsymbol{\nu} = \boldsymbol{\nu}e^{\boldsymbol{\xi}}$ *as a mapping* $T_{\boldsymbol{x}}M \to T_{\exp \boldsymbol{x}}M$ *satisfies*

$$\langle \boldsymbol{\nu}, \boldsymbol{\nu} \rangle = \langle \exp'(\boldsymbol{\xi})\boldsymbol{\nu}, \exp'(\boldsymbol{\xi})\boldsymbol{\nu} \rangle_{\exp(\boldsymbol{\xi})}$$

**3)** *With the metric given by (2.5),* $\mathcal{M}$ *is a Bruhat-Tits space, that is, it is a complete metric space in which the semi-parallelogram law holds. That is, given any* $\boldsymbol{x}_1$, $\boldsymbol{x}_2 \in \mathcal{M}$, *there exists a unique* $\boldsymbol{z} \in \mathcal{M}$ *such that for ant* $\boldsymbol{y} \in \mathcal{M}$ *the following holds*

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 + 4d(\boldsymbol{z}, \boldsymbol{y})^2 \leq 2d(\boldsymbol{y}, \boldsymbol{x}_1)^2 + 2d(\boldsymbol{y}, \boldsymbol{x}_2)^2. \tag{2.6}$$

**Comments:**

**1)** The proof of each assertion follows from calculus. In our framework, commutativity makes things much simpler. To obtain the completeness of $\mathcal{M}$ we transfer it from $\mathbb{R}^n$ via the exponential mapping.

**2)** The point $\boldsymbol{z}$ mentioned in item (3) is given by $\boldsymbol{z} = \sqrt{\boldsymbol{x}_1\boldsymbol{x}_2}$.

Along with the notion of geodesic curve, there is a notion of geodesic surface through (or containing) a point $\boldsymbol{x}(0)$. A parametric geodesic surface containing $\boldsymbol{x}(0) \in \mathcal{M}$ and having tangents $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_K$ there is a mapping $\boldsymbol{t} \in \mathbb{K} \to \mathcal{M}$ given by

$$\boldsymbol{t} = (t_1, ..., t_K) \in \mathbb{R}^K \to \boldsymbol{x}(\boldsymbol{t}) = e^{-\sum_i t_i \boldsymbol{\xi}_i}. \tag{2.7}$$

We leave for the reader to verify that we can reach any point of this surface traveling along the individual geodesics one at a time.

In the next section we shall see how this surface maps into a geodesic surface in the set of all probabilities on $[n] = \{1, ..., n\}$, and the probabilistic interpretation of the geodesic surface will be that of an exponential family.

# 3 The induced geometry on the set of discrete probabilities

If we think about the lines in $\mathcal{M}$ as the object of our interest, we may think about the probabilities on a discrete sample space of cardinality $n$ as the *representatives* of the rays in $\mathcal{M}$ (equivalence classes). Let us introduce the notation $\mathcal{P} = \{\boldsymbol{x} \in \mathcal{M} | \langle \boldsymbol{e}, \boldsymbol{x} \rangle = 1\}$. Clearly, if the point $\frac{\boldsymbol{x}}{\langle \boldsymbol{e}, \boldsymbol{x} \rangle}$ which lies in $\mathcal{P}$ is a representative of the line through $\boldsymbol{x}$ but the mapping

$$\mathcal{M} \to \mathcal{P} \quad \boldsymbol{x} \to \frac{\boldsymbol{x}}{\langle \boldsymbol{e}, \boldsymbol{x} \rangle},$$

is a *projection* but not an orthogonal projection. Similarly, a curve $t \in I \to \boldsymbol{x}(t) \in \mathcal{M}$ projects onto a curve in $\mathcal{P}$, and the question is: Do geodesics in $\mathcal{M}$ project onto geodesics in $\mathcal{P}$?

Before answering this question, note the following. If $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are two points in $\mathcal{P}$ the curve $\boldsymbol{x}(t) = \boldsymbol{p}_2^t \boldsymbol{p}_1^{(1-t)} = \boldsymbol{p}_1 \exp(-t\boldsymbol{\xi})$ that joins them is a geodesic in the ambient space $\mathcal{M}$, but it is not necessarily a curve in $\mathcal{P}$. To answer the question posed in the last paragraph consider $\boldsymbol{p}(t) = \boldsymbol{x}(t)/Z(t)$ with $Z(t) = \langle \boldsymbol{e}, \boldsymbol{x}(t) \rangle$ which certainly is a curve lying in $\mathcal{P}$. Note that $\langle \boldsymbol{e}, \boldsymbol{p}(t) \rangle = 1$, then

$$\dot{\boldsymbol{p}} = -\boldsymbol{p} \left( \boldsymbol{\xi} - \boldsymbol{p} \langle \boldsymbol{e}, \boldsymbol{p}\boldsymbol{\xi} \rangle \right), \tag{3.1}$$

clearly satisfies $\langle \boldsymbol{e}, \dot{\boldsymbol{p}}(t) \rangle = 0$. That is the velocity along $\boldsymbol{p}$ is tangent to $\mathcal{P}$ at every point. Differentiate with respect to $t$ once more and use the previous equation to obtain

$$\ddot{\boldsymbol{p}}(t) = \boldsymbol{p} \left( \frac{\dot{p}^2(t)}{\boldsymbol{p}^2(t)} - \left\langle \boldsymbol{e}, \frac{\dot{p}^2(t)}{\boldsymbol{p}^2(t)} \right\rangle \right).$$

Notice that $\boldsymbol{p}(t)$ satisfies the geodesic equation in the coordinates of the ambient space $\mathcal{M}$ corrected so that the acceleration is tangent to $\mathcal{P}$. That is, the projection of a geodesic *is* a geodesic. We can gather these comments under the following theorem:

**Theorem 3.1** *The geodesic between two points $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ in $\mathcal{P}$ can be obtained by projecting down to $\mathcal{P}$ the geodesic between the same points in the ambient space $\mathcal{M}$.*

The pending question is: How to choose coordinates in $\mathcal{P}$ is order to transport the whole geometric structure there instead of working with the coordinates of the ambient space ($\mathcal{M}$) in which it sits as a submanifold.
Note as well that instead of a geodesic joining two points, the same result applies to a geodesic issued from a point $\boldsymbol{p}_1$ in the direction of a tangent vector $\boldsymbol{\xi}$. And the same applies to a geodesic surface determined by a collection of vectors $\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_K\}$ parameterized by $\boldsymbol{t} \in \mathbb{R}^K$. The analogue of the previous result is now easy to establish.

**Theorem 3.2** *The geodesic surface $\boldsymbol{p}(\boldsymbol{t})$ containing the point $\boldsymbol{p}_1$ and tangent to the vectors $\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_K\}$ at $\mathcal{P}_{\boldsymbol{p}_1}$ is given by*

$$\boldsymbol{p}(\boldsymbol{t}) = \frac{\boldsymbol{p}_1 e^{-\sum_i t_i \boldsymbol{x}_i}}{Z(\boldsymbol{t})} \quad \textit{with} \quad Z(\boldsymbol{t}) = \left\langle \boldsymbol{e}, \boldsymbol{p}_1 e^{-\sum_i t_i \boldsymbol{x}_i} \right\rangle . \tag{3.2}$$

*is obtained by projecting the geodesic surface in $\mathcal{M}$ down to $\mathcal{P}$.*

**Comment:** Clearly (3.2) describes an *exponential family* of probabilities on $\{1, ..., n\}$, or in other words, exponential probabilities are geodesic surfaces in the metric described in Section 2.

Since the geodesics are defined for all values of the parameter $t$, a pending question is: What does $\boldsymbol{p}(t)$ tend to as $t \to \pm\infty$? To answer the first question, put $k_{\max} = \arg\max\{\ln(p_1(k)/p_2(k)) \mid k = 1, ..., n\}$ and similarly $k_{\min} = \arg\min\{\ln(p_1(k)/p_2(k)) \mid k = 1, ..., n\}$. Let us consider the sets of subscripts at which the maximum or the minimum are reached, that is $J_{max} = \{1 \leq k \leq n : k = m_{max}\}$ and $J_{min} = \{1 \leq k \leq n : k = m_{min}\}$, and let their cardinalities be, respectively, $M_{\max}$ and $M_{\min}$. Then as $t \to \infty$,

$$p_k(t) \to \begin{cases} 1/M_{\min} & k \in J_{\min} \\ 0 & \text{otherwise.} \end{cases}$$

When $t \to -\infty$ a similar result is obtained with $M_{\min}$ replaced by $M_{\max}$.

# 4 Geometric/dynamic interpretation of the maximum entropy method

Consider a random variable $\boldsymbol{\xi}$ as a tangent vector in $T_{\boldsymbol{p}_0}M$ and the class

$$\mathcal{P}_\mu = \left\{ \boldsymbol{p} \in \mathcal{P} \; : \; E_{\boldsymbol{p}}[\boldsymbol{\xi}] = \sum_i p(i)\xi_i = \langle \boldsymbol{e}, \boldsymbol{p}\boldsymbol{\xi} \rangle = \mu \right\},$$

for some given number $\mu$, that is, the class of all probabilities under which $\boldsymbol{\xi}$ has expected value $\mu$. Since $\mathcal{P}_\mu$ is a hyperplane in $\mathcal{P}$, we may wonder whether following the geodesic $\boldsymbol{p}(t) = \boldsymbol{p}_0 \exp(-t\boldsymbol{\xi})/Z(t)$ issued from $\boldsymbol{p}_0$ along $\boldsymbol{\xi}$, we might intercept $\mathcal{P}_\mu$. If the answer is yes, then so is the answer to our question.

If $t^*$ is such that $\boldsymbol{p}(t^*) \in \mathcal{P}_\mu$, then clearly

$$\boldsymbol{p}(t^*) \in \mathcal{P}_\mu \quad \Leftrightarrow \quad t^* = \operatorname{argmin} \{ t\mu + \ln Z(t) \; : \; t \in \mathbb{R} \}. \tag{4.1}$$

Clearly the function in curly brackets is strictly convex, and the first order condition for $t^*$ to be its minimizer is equivalent to the assertion on the left hand side. Consider now the relative entropy function

$$S(\boldsymbol{p}, \boldsymbol{p}_0) : \mathcal{P} \to \mathbb{R} \quad S(\boldsymbol{p}, \boldsymbol{p}_0) = -\langle \boldsymbol{e}, \boldsymbol{p} \ln(\boldsymbol{p}/\boldsymbol{p}_0) \rangle = \sum_i p(i) \ln \left( \frac{p(i)}{p_0(i)} \right).$$

Now, note that if we replace the generic $\boldsymbol{p}$ by a probability along the geodesic, we have $S(\boldsymbol{p}(t), \boldsymbol{p}_0) = \mu + \ln(Z(t)$. To complete the argument, note that the concavity of the logarithm implies that for any pair of probabilities $S(\boldsymbol{p}, \boldsymbol{q}) \leq 0$ (with equality whenever they are equal), implies that taking $\boldsymbol{q} = \boldsymbol{p}(t)$

$$S(\boldsymbol{p}, \boldsymbol{p}(t)) \leq 0 \quad \Rightarrow S(\boldsymbol{p}, \boldsymbol{p}_0) \leq t\mu + \ln(Z(t)) = S(\boldsymbol{p}(t), \boldsymbol{p}_0) \quad \text{for any} \quad t.$$

That is the entropy of any $p(t)$ along the geodesic bounds from above the entropy of any $\boldsymbol{p} \in \mathcal{P}_\mu$. What we do not know is whether there is a $t^*$ for which $\boldsymbol{p}(^*) \in \mathcal{P}_\mu$.

What (4.1) asserts is that if there is a $t^*$ minimizing $t\mu + \ln Z(t)$, then $p(t^*) \in \mathcal{P}_\mu$ and necessarily $\boldsymbol{p}(t^*)$ solves the following entropy maximization problem:

$$\text{Find} \quad \boldsymbol{p}^* \in \mathcal{P}_\mu \quad \text{such that} \quad \boldsymbol{p}^* \quad \text{maximizes} \quad S(\boldsymbol{p}, \boldsymbol{p}_0) \quad \text{over} \quad \mathcal{P}_\mu.$$

To sum up, whether or not the geodesic issued from $\boldsymbol{p}_1$ along $\boldsymbol{\xi}$ intersects the "plane" $\mathcal{P}_\mu$ is equivalent solvability of the entropy maximization problem. And since the dual entropy function $\Sigma(t) = t\mu + \ln Z(t)$ is interpreted as a *free energy* in statistical thermodynamics, the travel time $t^*$ has an interpretation as an "intensive" thermodynamical variable conjugate to $\boldsymbol{\xi}$.

# 5 Entropy: a Lyapunov function for Markovian dynamics

The results of the previous section, interesting as they may be, are not connected to a dynamics related to a physical process. For example, we may consider the case in which there is a Markovian process with state space $\{1, ...., n\}$ and transition rate matrix $Q$. When we suppose that the chain is irreducible (we can reach any state starting from any other state), it is well known that if the transition state is either symmetric or reversible, the entropy function is a Lyapunov function for the chain. The spell it out in symbols, note if $\boldsymbol{p}(0)$ is any initial distribution on the state space, then $\boldsymbol{p}(t) = e^{tQ}\boldsymbol{p}(0)$ describes the probability distribution at current time $t$.
The following was proved in Klein (1956) for the Ehrenfest urn model and extended in Moran (1960) as follows:

**Theorem 5.1** *With the notations introduced above, then:*
**1)** *If the Markov chain is symmetric, that is, $Q(i,j) = Q(j,i)$, or*
**2**) *If the chain is reversible, that is, if there is an equilibrium probability law $\boldsymbol{p}_e$ such that $\sum_j Q(i,j)p_e(j) = 0$,*
*then the entropy $S(\boldsymbol{p}(t)$ satisfies $dS(\boldsymbol{p}(t))/dt > 0$.*

These results are part of a large chain of results on the issue of time (ir)reversibility in statistical thermodynamics. From the mathematical point of view, the result is a particular case of a more general, and surprisingly simple to prove result for monotone continuous mappings, which applies to linear and non-linear dynamical systems. See

Brown (1985).

# 6 Concluding remarks

To sum up, there is a curious and nice relationship between a geometry on the set of positive real vectors and the exponential families of probability distributions on finite sets. In this setup, exponential families appear as *geodesic surfaces* in the set of probabilities. This Bruhat-Tits space is different than the Hilbert simplex/Birkhoff cone geometry proposed in [23] (Hilbert cross-ratio distance on the probability simplex fails the semi parallelogram law).

The logarithmic distance in the set of strictly positive vectors leads to notion of best predictor that complements the theory best prediction is square distance. For more on this see Gzyl (2017).

# References

[1] Arsigny, V., Fillard, P., Pennec, X. and Ayach, N. (2007). *Geometric Means in a Novel Vector Space Structure on Symmetric positive definite matrices*, SIAM J. Matrix Theory, **29**, 328-347.

[2] Amari, S.-i. *"Differential Geometric Methods in Statistics"*, Lecture Notes in Statistics, **28**, Berlin (1985).

[3] Amari, S.-i. *"Information Geometry and its Applications"*, Springer (2016).

[4] Amari, S.-i., Barndorff-Nielsen, O., Kass, R., Lauritzen, S. and Rao, C. *"Differential Geometry in Statistical Inference"* Institute of Mathematical Statistics Lecture Notes, Monograph Series, Vol. 10, Hayward, (1987).

[5] Barndorff-Nielsen, O. *"Information and Exponential Families in Statistical Theory"*, Chichester, Wiley (1978).

[6] Brown, C.C. (1985) *Entropy increase and measure theory*, Proc. Am. Math. Soc., **95**, 488-450.

[7] Casalis, M. (1991) *Familles exponentielles naturelles sur rd invariantes par un groupe.* International Statistical Review/Revue Internationale de Statistique, 241-262.

[8] Calin, O. and Udriste, C. *Geometric Modeling in Probability and Statistics*, Springer Internl. Pub., Switzerland, (2010).

[9] Efron, B. (1978) *The geometry of exponential families.* The Annals of Statistics, 6(2), 362-376.

[10] Gzyl, H. and Recht, L. (2006) *"A geometry in the space of probabilities II: Projective spaces and exponential families"* Rev. Iberoamericana de Matemáticas, **22**, 833-850.

[11] Gzyl, H. and Recht, L. (2007) *Intrinsic geometry on the class of probability densities and exponential families*, Public. Mathematiques, **51**, 309-322.

[12] Gzyl, H. (2017) *Prediction in logarithmic distance.* `http://arxiv.org/abs/1703.08696`.

[13] Hotelling, H (1930) *Spaces of statistical parameters.* Bulletin of the American Mathematical Society (AMS), 36:191

[14] Imparato, D. and Trivelato, B. *Geometry of extended exponential models*, in *Algebraic and Geometric Methods in Statistics*, Gibilisco, P., Riccomagno, E. Rogantin, M.P. and Wynn, H. eds., Cambridge Univ. Press, Cambridge, (2010).

[15] Jaynes, E. T. (1957). *Information theory and statistical mechanics.* Physical review, 106(4), 620.

[16] Klein, M. (1956) Entropy and the Ehrenfest urn model, Physica, **22**, 569-575,

[17] Lang, S. Math talks for undergraduates, Springer, New York, (1999).

[18] Lawson, J.D. and Lim, Y. (2001) *The Geometric mean, matrices, metrics and more*, Amer. Math.,Monthly, **108**, 797-812.

[19] Li, F., Zhang L. and Zhang Z. (2018) *Lie Group Machine Learning*, Walter de Gruyter GmbH & Co KG, ISBN9783110499506.

[20] Luenberger, D.G. *Investment Science*, Princeton Univ. Press, Princeton, (1980).

[21] Moakher, M. (2005) *A differential geometric approach to the geometric mean of symmetric positive definite matrices*, SIAM. J. Matrix Anal. & Appl., **26**, 735-747

[22] Moran, P.A.P. (1960) *Entropy, Markov processes and Boltzmann's H-theorem*, Proc. Camb. Phil. Soc., **57**, 833-842.

[23] Nielsen, F., and Sun, K. (2017). Clustering in Hilbert simplex geometry. preprint arXiv:1704.00454.

[24] Nielsen, F., and Nock R. (2018) *On the geometry of mixtures of prescribed distributions*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[25] Pistone, G. *Algebraic varieties vs. differentiable manifolds*, in *Algebraic and Geometric Methods in Statistics*, Gibilisco, P., Riccomagno, E. Rogantin, M.P. and Wynn, H. eds., Cambridge Univ. Press, Cambridge, (2010).

[26] Pistone, G. and Rogantin, M.P. *The exponential statistical manifold: mean parameters, orthogonality and space transformations"* Bernoulli, **5** (1999), 721-760.

[27] Pistone, G. and Sempi, C. *"An infinite dimensional geometric structure in the space of all probability measures equivalent to a given one"*. Ann. Statist., **23** (1995), 1543-1561.

[28] Rao, C. R. (1992). *Information and the accuracy attainable in the estimation of statistical parameters*. In Breakthroughs in statistics (pp. 235-247). Springer, New York, NY.

[29] Schwartzmazn, A. (2015) *Lognormal distribution and geometric averages of positive definite matrices*, Int. Stat. Rev., **84**, 456-486.

[30] Vajda, I. *"Theory of statistical inference and information"* Kluwer Acad., Dordrecht, (1989).