Rev. Mat. Iberoamericana **22** (2006), no. 2, 545–558

A geometry on the space of probabilities I. The finite dimensional case

Henryk Gzyl and Lázaro Recht

Abstract

In this note we provide a natural way of defining exponential coordinates on the class of probabilities on the set $\Omega = [1, n]$ or on $\mathbb{P} = \{p = (p_1, \ldots, p_n) \in \mathbb{R}^n | p_i > 0; \Sigma_{i=1}^n p_i = 1\}$. For that we have to regard $\mathbb{P}$ as a projective space and the exponential coordinates will be related to geodesic flows in $\mathbb{C}^n$.

1. Motivational preliminaries

Exponential families have been very much in use in Statistics and Information Theory, see for example Barndorff-Nielsen's classic [1]. They appear naturally when considering the problem of finding a probability q defined on some measurable space $(\Omega, \mathcal{B})$ when only the expected values

$$E_q[X_i] = m_i; \quad i = 1, \ldots, K. \tag{1.1}$$

are specified. From the work of Boltzmann (~ 1880) and Gibbs (~ 1900) the following technique evolved. One realizes that the class of probability measures

$$\mathcal{K} = \{q \,|\, (1.1) \text{ holds}\}$$

is a convex set. To systematically pick up points from a convex set, an efficient approach is to rephrase the problem as: Find the q that maximizes some concave function defined over $\mathcal{K}$.

In Statistical Physics, where the interest is to characterize equilibrium states of macroscopic systems, the obvious choice is to consider an "entropy"

2000 Mathematics Subject Classification: Primary: 46L05, 53C05, 53C56, 60B99, 60E05. Secondary: 53C30, 32M99, 62A25, 94A17.
Keywords: C^*-algebra, reductive homogeneous space, lifting of geodesics, exponential families, maximum entropy method.

function which describes a quantity that increases in time to a maximum value which is achieved when the system reaches equilibrium. By the way, the entropy function happens to be a Lyapunov function for the underlying dynamics of the system, a fact which closes a nice circle of ideas.

To be specific, let us consider a finite dimensional problem, that is $\Omega = \{1, 2, \ldots, n\}$ and $\mathcal{B}$ is the collection of all subsets of Ω. The class of all probability distributions on Ω is the following simplex in $\mathbb{R}^n$:

$$\mathbb{P} = \Big\{p = (p_1, \ldots, p_n) \,|\, \sum_{i=1}^{n} p_i = 1,\ p_i \geq 0\ \forall i\Big\}.$$

A fact that many people find curious about the solution to (1.1) is the appearance of exponential families of the type

$$q(i) = \frac{e^{-<\lambda, X>(i)}}{Z(\lambda)} p(i) \tag{1.2}$$

where $p = (p(i))$ is a probability or a positive vector,

$$< \lambda, X > \equiv \sum_{k=1}^{K} \lambda_k X_k,$$

and where

$$Z(\lambda) = \sum_{i=1}^{n} e^{-<\lambda, X>(i)} p(i)$$

is an obvious normalization factor. How this comes about, and how it is related to the entropy function is rather well known. For all the mathematical finesse involved see [2].

In other words, if one applies the methodology proposed by Boltzmann and Gibbs, plus basic mathematics, (1.2) is just a consequence. But there is a different way of looking at the problem, which makes it natural to think of exponential parameterization of the points of $\mathbb{P}$, that is, of mappings $\mathbb{R}^n \to \mathbb{P}$ sending $X \in \mathbb{R}^n$ onto $q(X)$ given by

$$q_i(X) = \frac{e^{-X_i}}{\sum_{k=1}^{n} e^{-X_k}} p_i\,. \tag{1.3}$$

Below we shall present a description of the maximum entropy method and see how is (1.3) related to (1.2).

The intention of this note is to make clear how the exponential coordinatization (1.3) is natural and to provide a geometric interpretation of the entropy function. To achieve this, a special geometry is introduced on the

set of positive vectors regarded as an special subset of the class of complex valued functions on a finite set. This is carried out in section 2, which is a particular case of a theory developed by Corach, Porta and Recht in [3], [7] and [8], from which we draw freely. Our case is simpler than the theory developed there, because all the Banach algebras we deal with here are commutative. The exponential coordinates are related to a natural projective structure on the class of positive vectors in $\mathbb{R}^n$, as we will see in section 3. Curiously enough, the results obtained here resemble some of the results obtained by Pistone and Sempi in [9]. More on this resemblance will be apparent in the second note of this series, in which we will develop the general case.

2. A geometry on the set of positive vectors

In this section, after briefly recalling some of the basic facts about $\mathbb{C}^n$ regarded as a C^*-algebra, we define a special connection and describe its geodesics and parallel transport along them.

2.1. The basic set up

For reasons that shall become clear below, we want to understand in which sense the curve

$$\gamma(t) = a(0)^{1-t}a(1)^t \tag{2.1}$$

where $a(0)$ and $a(1)$ are positive vectors, is a geodesic. For that we start considering the n-vectors with positive components as a special subset of $\mathbb{C}^n$.

To begin with, we regard $\mathcal{A} = \mathbb{C}^n$ as complex valued functions $X : [1, n] \to \mathbb{C}$ with the natural algebraic structure imposed on them in which $XY(k) \equiv X(k)Y(k)$. Actually this structure turns $\mathcal{A}$ into a commutative C^*-algebra. In this algebra the set of invertible vectors

$$G = \{X \in \mathcal{A} \,|\, X \not\equiv 0\}$$

is a (commutative) group and the class $G^+ \subset G$ denotes the class of positive elements. Note that G^+ is a homogeneous space for the group action defined by

$$L_g : G^+ \to G^+ \qquad L_g(a) = (g^*)^{-1}ag^{-1}, \ \ \forall\, a \in G^+$$

for any $g \in G$. Here x^* denote the complex conjugate of x. Since the product is commutative, $L_g(a) = |g|^{-2}a$. If we identify G^+ with the diagonal matrices with strictly positive elements, every $a \in G^+$ defines a scalar product on $\mathbb{C}^n$ by $< x, y >_a = \sum a(i)x^*(i)y(i)$. We may interpret the group action as an isometry $\mathbb{C}^n_a \to \mathbb{C}^n_{L_g(a)}$.

Interpretations aside, for any fixed $a_0 \in G^+$ we can define the projection operator

$$\pi_{a_0} : G \to G^+$$

by means of

$$\pi_{a_0}(g) = L_g(a_0)$$

and notice right away that the fiber (isotropy group) over a_0 defined by

$$I_{a_0} = \{g \in G \mid \pi_{a_0}(g) = a_0\} = \{g \in G \mid |g| = 1\} = \mathbb{T}^n$$

the n-dimensional torus.

Since G is clearly an open subset of $\mathbb{C}^n$, its tangent space at any point is $\mathbb{C}^n$, *i.e.*,

$$(TG)_1 = \mathbb{C}^n = \mathcal{A}\,,$$

and it is easy to see that

$$(TI_{a_0})_1 = V_{a_0} = 0 \oplus i\mathbb{R}^n\,,$$

which in the non-commutative case corresponds to the anti-hermitian elements in $\mathcal{A}$.

The derivative $(D\pi_{a_0})_1(X)$ of π_{a_0} at 1 in the direction of $X \in \mathcal{A}$ is easy to compute, and it is given by

$$(D\pi_{a_0})_1(X) = -a_0(X + X^*)\,.$$

Clearly

$$(D\pi_{a_0})_1 : \mathcal{A} \to (TG^+)_{a_0} \equiv \mathbb{R}^n \oplus 0\,.$$

Note as well that the horizontal space at a_0, defined by

$$H_{a_0} \equiv \{X \in \mathcal{A} \,|\, (a_0)^{-1}X^*a_0 = X\}$$

may be described as

$$H_{a_0} = \{X \in \mathcal{A} \,|\, X^* = X\} = \mathbb{R}^n \oplus 0.$$

Therefore, we have the obvious splitting

$$\mathcal{A} = H_{a_0} \oplus V_{a_0}\,.$$

Not only that, the map $(D\pi_{a_0})_1$ is invertible from the left. That is, there exists a mapping $\kappa_{a_0} : (TG^+)_{a_0} \to (TG)_1$, given by

$$\kappa_{a_0}(z) \equiv -\frac{a_0^{-1}}{2}\, z$$

such that $(TG^+)_{a_0} \overset{\kappa}{\to} (TG)_1 \overset{(D\pi_{a_0})}{\to} (TG^+)_{a_0}$ is the identity mapping.

The mapping κ is called the **structure 1-form** of the homogeneous space G^+, and it is an $\mathcal{A}$-valued linear mapping defined on $(TG^+)_{a_0}$. All the geometry on G^+ comes from κ. This whole construction makes (G, G^+, π_a) plus the datum κ an object that Kobayashi and Nomizu call a reductive homogeneous space. See chapter 11 in [6] for full details.

2.2. Lifting curves from G^+ to G and parallel transport

Let us begin with a basic lemma, in which one of the basic properties of the connection κ will become apparent.

Lemma 2.1 *Let $a(t) : [0,1] \to G^+$ be a continuous curve in G^+. There exists a curve $g(t) : [0,1] \to G$ (called **the lifting of** $a(t)$ **to** G) such that*

$$L_{g(t)}(a_0) = \pi_{a_0}(g(t)) = a(t)\,, \tag{2.2}$$

where the identification $a(0) = a_0$ will be used from now on.

Proof. Let us verify that the solution to the (transport) equation

$$\dot{g}(t) = \kappa_{a(t)}(\dot{a(t)})g((t) \tag{2.3}$$

satisfies (2.2). Here the commutativity makes things really simple. Equation (2.3), explicitly spelled out, is

$$\dot{g}(t) = -\frac{\dot{a}(t)}{2a(t)}g(t), \quad g(0) = 1\,;$$

which can be easily solved to yield

$$g(t) = \Big(\frac{a(0)}{a(t)}\Big)^{1/2}\,. \tag{2.4}$$

Note that

$$\pi_{a_0}(g(t)) = \Big(\frac{a(t)}{a(0)}\Big)^{1/2} a(0) \Big(\frac{a(t)}{a(0)}\Big)^{1/2} = a(t)\,.$$

■

The **parallel transport** along $a(t)$ (from $a(0)$ to $a(1)$) is defined in

Definition 2.1 *Let $a(t)$ be a curve in G^+ and let $g(t)$ be its lifting to G. The parallel transport along $a(.)$ is the mapping $\tau(a(.)) : (TG^+)_{a(0)} \to (TG^+)_{a(1)}$ defined by*

$$\tau(a(.))(X) = L_{g(1)}(X)\,. \tag{2.5}$$

We may now say that $a(t)$ is a **geodesic** if $\dot{a}(0)$ is transported onto $\dot{a}(t)$ by means of the (time) rescaled curve $b(s) := a(st),\ s \in [0,1]$. From (2.4) and (2.5) it is clear that this amounts to

$$\dot{a}(t) = \frac{a(t)}{a(0)}\dot{a}(0) \quad \Longleftrightarrow \quad \frac{\dot{a}(t)}{a(t)} = \frac{\dot{a}(0)}{a(0)} \equiv X.$$

Or equivalently,

Lemma 2.2 *The curve $a(t)$ is a geodesic if and only if there exists a (real) vector X such that*

$$a(t) = a(0)\, e^{tX}.$$

Comment 2.1 *This means that the lifted geodesic is given by*

$$g(t) = a(0)^{1/2}\, e^{-tX/2}.$$

To tie up with the construction in the previous and following sections, observe that if we specify the initial and final points of the geodesic, the vector X is automatically determined:

$$a(1) = a(0)e^{X} \Longrightarrow X = \ln\Big(\frac{a(1)}{a(0)}\Big),$$

and the equation of the geodesic can be rewritten as

$$a(t) = a(0)^{1-t}a(1)^{t} \tag{2.6}$$

which is nothing but (2.1) with some names changed.

2.3. A semi-norm and it associated distance

To begin with, recall that $\mathcal{A}$ is a Banach algebra, and thus comes endowed with some norm, which if needed will be denoted by a symbol different from the one we introduce next. Consider the following semi-norm: For any $Z \in \mathcal{A}$ define

$$\|Z\| = |<1, Z>| \equiv \Big|\sum_i Z(i)\Big|,$$

which is just the absolute value of the trace of X. This leads to a pseudo-metric on TG^{+}, that is, the distance between points may be zero without the points having to be equal. The reason behind this rather peculiar choice is that the length of the geodesics determined by this pseudo-metric leads to the classical Kullback distance between probabilities.

To define a metric on TG^+, we begin by defining it at $(TG^+)_1$ by $\|X\|_1 \equiv \|X\|$, and transporting it to any other $(TG^+)_{a_0}$ by means of the group action: that is, we set

$$\|X\|_{a_0} = | < a_0^{-1}, X > |$$

and it is easy to verify that this is a consistent definition. Note now that if $a(t)$ is a geodesic joining $a(0)$ to $a(1)$ in G^+, then the "length" of the velocity vector along the geodesic is constant, for

$$\|\dot{a}(t)\|_{a(t)} = | < a(t)^{-1}, a(t)X >= \|X\|_1 = \left\| \ln\left(\frac{a(1)}{a(0)}\right) \right\|$$

and therefore, the geodesic distance from $a(0)$ to $a(1)$ is given by

$$d(a(0), a(1)) = \int_0^1 \|\dot{a}(t)\|_{a(t)}\, dt = \|X\|_1 = \left\| \ln\left(\frac{a(1)}{a(0)}\right) \right\|.$$

Comment 2.2 *Note that the choice of 1 for the definition of the norm is not mandatory. For example, for a fixed $q \in G^+$ we may have defined*

$$\|X\|_{(q),a(0)} := | < qa(0)^{-1}, X > |$$

and we would have ended up with

$$\|\dot{a}(t)\|_{(q),a(t)} = | < qa(t)^{-1}, a(t)X > | = \|X\|_{(q)} = \left\| q \ln\left(\frac{a(1)}{a(0)}\right) \right\|,$$

which is still symmetric in $a(0), a(1)$. If we chose $q = a(1)$, we would end up with

$$\|X\|_{a(1)} = \sum_i a_i(1) \ln\left(\frac{a_i(1)}{a_i(0)}\right),$$

which is not symmetric anymore. It corresponds to the ***Kullback distance*** *between $a(0)$ and $a(1)$, or the entropy of $a(1)$ with respect to $a(0)$.*

2.4. Conditional expectations and additive and multiplicative decompositions

Consider a subalgebra $\mathcal{B}$ of the algebra $\mathcal{A}$. Think for instance on the functions measurable with respect to the σ-algebra generated by some partition of Ω, which in our finite dimensional setup consists of vectors with components constant on the blocks of the partition. Denote by p a fixed probability on Ω and by $E_{p,\mathcal{B}}$ the usual conditional expectation, that is an orthogonal projection of $\mathcal{A}$ onto $\mathcal{B}$ satisfying $E_{p,\mathcal{B}}[XY] = E_{p,\mathcal{B}}[X]Y$ for any Y in $\mathcal{B}$. This conditional expectation induces a decomposition $\mathcal{A} = \mathcal{B} \oplus K$, where to simplify notation we set $K \equiv \ker(E_{p,\mathcal{B}})$. It induces also a similar decomposition on $\mathcal{A}^s \simeq \mathcal{R}^n \simeq (TG^+)_1$ given by $\mathcal{A}^s = \mathcal{B}^s \oplus K^s$.

The interesting thing is that this decomposition can be lifted to the set of positive vectors, *i.e.*, to G^+ by means of the exponential map. That is, if $q \in G^+$ and $q = e^X$ for $X \in A^s$, then

$$q = e^X = e^{E_{p,\mathcal{B}}[X]} e^{X - E_{p,\mathcal{B}}[X]}$$

and, in the case that will be important for us below, when $\mathcal{B} = \mathbb{C}$, which corresponds to the case in which the σ-algebra defining $\mathcal{B}$ is trivial and $E_{p,\mathcal{B}} = E_p$, the usual expected value with respect to p, the last decomposition will look like

$$q = e^X = e^{E_p[X]} e^{X - E_p[X]} \,. \tag{2.7}$$

2.5. The $\mathcal{B}$-projective structure on the class G^+

We want to define equivalence classes (modulo $\mathcal{B}$) in such a way that they are preserved under the action of $G_{\mathcal{B}}$, the group of invertible elements in $\mathcal{B}$, that is, under the action of the mapping $a \to L_g(a) = (g^*)^{-1} a g^{-1}$, but for $g \in G_{\mathcal{B}}$. In particular we want the relation, denoted by $\sim_{\mathcal{B}}$ to be such that if $a(t)$ is a curve in G^+, then $\tilde{a}(t) = L_g(a(t)) \sim_{\mathcal{B}} a(t)$. In particular, since we shall be transporting tangent vector fields, we will want the tangent

$$\tilde{X} = \dot{\tilde{a}}(0) = \frac{1}{|g|^2} \left(X - a\Big(\frac{V}{g} + \frac{V^*}{g^*}\Big)\right)$$

to be somehow equivalent to X. Here, $V = \dot{g}(0)$. For that, note that the previous identity can be rewritten as

$$\frac{\tilde{X}}{\tilde{a}} = \frac{X}{a} + W \,,$$

where $W = -(\frac{V}{g} + \frac{V^*}{g^*})$ is a symmetric element in $\mathcal{B}$. Now we state

Definition 2.2 *With the notations introduced above, we say that a and $\tilde{a}$, both in G^+, are $\sim_{\mathcal{B}}$ if and only if*

$$\frac{\tilde{a}}{a} \in G^+_{\mathcal{B}} \,.$$

Comment 2.3 *Notice that if $\frac{\tilde{a}(t)}{a(t)} = h(t) \in G^+_{\mathcal{B}}$ and $g(t) \in G_{\mathcal{B}}$ is any square root of $h(t)^{-1/2}$, then, taking logarithms and differentiating at $t = 0$, we obtain*

$$\frac{\tilde{X}}{\tilde{a}} = \frac{X}{a} - \Big(\frac{V}{g} + \frac{V^*}{g^*}\Big).$$

That is, the equivalence relation may be lifted to $G^+ \times \mathcal{A}^+$, which should be regarded as a trivial tangent bundle.

We can form the quotient space $G^+/\sim_{\mathcal{B}}$ and veryfy that $G^+/\sim_{\mathcal{B}} \simeq \mathbb{P}_{\mathcal{B}} \equiv \{\alpha \in G^+ \mid E_{p,\mathcal{B}}[\alpha] = 1\}$, where the equivalence is brought about by the mapping $\Phi_p : G^+ \to \mathbb{P}_{\mathcal{B}}$ and the following

Lemma 2.3 *With the notations introduced above, $\tilde{a} \sim_{\mathcal{B}} a$ if and only if $\Phi_p(\tilde{a}) = \Phi_p(a)$, where*

$$\Phi_p(a) = \frac{a}{E_{p,\mathcal{B}}[a]}.$$

Proof. Let $\tilde{a} = ah$ where $h \in G^+_{\mathcal{B}}$. Therefore $E_{p,\mathcal{B}}[\tilde{a}] = hE_{p,\mathcal{B}}[a]$ and $\Phi_p(\tilde{a}) = \Phi_p(a)$. Conversely, if $\Phi_p(\tilde{a}) = \Phi_p(a)$, then

$$\tilde{a} = a\frac{E_{p,\mathcal{B}}[\tilde{a}]}{E_{p,\mathcal{B}}[a]},$$

thus $\frac{\tilde{a}}{a} \in G^+_{\mathcal{B}}$. ■

Comment 2.4 *When $\mathcal{B} = \mathbb{C}$ is the algebra of functions measurable with respect to the trivial σ-algebra $\{\emptyset, \Omega\}$, then $G^+_{\mathcal{B}} = [0, \infty)$, and we have*

$$\Phi_p(a) = \frac{a}{< p, a >} = \frac{a}{E_p[a]}.$$

Also in this case $\mathbb{P}_{\mathcal{B}} = \mathbb{P}$, which explains our choice of notation. In this case we shall put $\sim$ instead of $\sim_{\mathcal{B}}$ to simplify the notation.

To define the action of G on $G^+/\sim_{\mathcal{B}}$ we proceed as usual: if $[a]$ denotes the equivalence class of $a \in G^+$, then we put $L_g[a] = [L_g a]$, and we have the following simple result:

Lemma 2.4 *Let $g \in G$ and let $[a] = [b]$. Then $[L_g(a)] = [L_g(b)]$.*

Proof. Invoking Lemma 2.3, it suffices to see that $\Phi_p([L_g(a)]) = \Phi_p([L_g(b)])$. For that it is enough to note that $b = ha$, where $h \in G^+_{\mathcal{B}}$, from which the desired conclusion follows. ■

2.6. Geometry on $\mathbb{P}_{\mathcal{B}}$

We are interested in geodesics in $G^+/\sim_{\mathcal{B}}$, but for not to worry about independence of the constructions on the representative chosen, we will work with a given class of representatives, namely with curves in $\mathbb{P}_{\mathcal{B}}$. We already know how G acts on $\mathbb{P}_{\mathcal{B}}$. Let $\alpha \in \mathbb{P}_{\mathcal{B}}$ and let us consider the mapping $\hat{\pi}_\alpha : G \to G$ and define the isotropy group of this action by $\hat{I}_\alpha = \{g \in G \mid \hat{\pi}_\alpha(g) = \alpha\}$.

Clearly, the tangent space to $\mathbb{P}_{\mathcal{B}}$ is $\{X \in \mathcal{A}^s \,|\, E_{p,\mathcal{B}}[X] = 0\} = K^s$. Note that if $g(t)$ is any curve in G such that $g(0) = 1$ and $\dot{g}(0) = X$, then

$$(D\hat{\pi})_1(X) = -\alpha(X + X^*) + \alpha E_{p,\mathcal{B}}[(X + X^*)\alpha] \in K^s.$$

(This is easy to see differentiating

$$\frac{|g(t)|^{-2}\alpha}{E_{p,\mathcal{B}}[|g(t)|^{-2}\alpha]}$$

at $t = 0$.) Note also that if $X \in \mathcal{B}$, then $(D\hat{\pi})_1(X) = 0$. Note as well that the tangent space $(T\hat{I}_\alpha)_1 = K^a$, *i.e.*, it consists of those antisymmetric elements X of $\mathcal{A}$ that have zero trace ($E_{p,\mathcal{B}}[X] = 0$).

Therefore $\mathcal{A} = \mathcal{B} \oplus K^s \oplus K^a$. Starting from this (which is a decomposition of the tangent space to G at $g = 1$) we can define a distribution of horizontal spaces by $H_g = \mathcal{B} \oplus \{gX \,|\, X \in K^s\}$. Again, to lift curves in $\mathbb{P}_{\mathcal{B}}$, we need a connection, this time defined as follows: for $\alpha \in \mathbb{P}_{\mathcal{B}}$ and $Y \in T_\alpha \mathbb{P}_{\mathcal{B}}$, we put

$$\kappa_\alpha(Y) = -\frac{1}{2}\,\alpha^{-1}Y.$$

Clearly, for $Y \in \mathbb{P}_{\mathcal{B}}$ we have $(D\hat{\pi})_1(\kappa_\alpha(Y)) = Y$. Now, to lift curves and define geodesics we can proceed verbatim as above.

3. Coordinates on $\mathbb{P}$

The set of probabilities on a finite sample space $\Omega = \{1, \ldots, n\}$ can be described as the closure of the manifold

$$\mathbb{P} = \{p = (p_1, \ldots, p_n) \in \mathbb{R}^n |\, p_i > 0; \Sigma_{i=1}^n p_i = 1\}$$

in $\mathbb{R}^n$. We saw above that we can identify rays in G^+ with points in $\mathbb{P}$ via an equivalence relation. That is, we identify lines in $\mathbb{R}^n_+$ with the point they intersect at $\mathbb{P}$ (or $\mathbb{P}$ can be regarded as a projective space). In other words, as the quotient space $G^+/\sim$, where $\sim$ is the equivalence relation of Definition 2.2. In particular, we saw in Lemma 2.3 that

$$a(1) \sim a(2) \quad \text{whenever} \quad \frac{a(1)}{< 1, a(1) >} = \frac{a(1)}{< 1, a(2) >}\,.$$

We saw in (2.6) that for any two points a_0 and a_1, there is a vector field $X = \ln(a_1/a_0)$ such that $\gamma(t) = a_0 e^{tX}$ is the geodesic joining a_0 to a_1 in G^+ (and G).

Note that the trace on $\mathbb{P}$ of a geodesic given by (2.6), or if you prefer, the equivalence class of each point of the geodesic, is given by

$$\gamma(t) = \frac{p(t)}{< 1, p(t) >} = \frac{p(0)^{1-t}p(1)^t}{\sum_i (p_i(0))^{1-t}(p_i(1))^t}\,. \tag{3.1}$$

This provides a geometric interpretation for (3.1) as the representative in $\mathbb{P}$ of the rays through the geodesic given by (2.6).

3.1. Exponential Families

Let us now examine a bit further in what sense (1.3) is natural in our setup. Set $a(0) = 1$ and let $a(1)$ be any other point in G^+. We now know that there exists a real vector X, actually given by $X = \ln a(1)$, such that $a(t) = e^{tX}$ joins 1 geodesically to $a(1)$, and the trace on $\mathbb{P}$ of this geodesic is $p(t) = a(t)/ < 1, a(t) >$, also given by

$$\frac{e^{tX}}{\sum e^{tX(i)}}\,.$$

That is, we have a correspondence between vectors in $\mathbb{R}^n$ regarded as tangent vectors to TG^+ and probabilities in $\mathbb{P}$, which we shall now explore further.

We shall consider the mapping

$$\begin{aligned} \Phi : (TG^+)_1 \simeq \mathbb{R}^n &\rightarrow \mathbb{P} \simeq G^+/\sim \\ X &\rightarrow \Phi(X) = \frac{e^X}{E[e^X]} \end{aligned} \tag{3.2}$$

and now we shall examine some of the basic properties of this map. Observe first that $\Phi(X) = \Phi(X + \alpha 1)$. Thus Φ as defined cannot be a bijective map.

Comment 3.1 *It is at this point where the choice of 1 to define de norm coincides with its role in the definition of* Φ. *Collinear vectors differ in their* $\|.\|_1$*-norm by a factor of* $e^{\alpha 1}$ *for appropriate* α.

Recall that to understand this more algebraically we noted that $\mathbb{R}^n \simeq \mathcal{A}^s \simeq (TG^+)_1 = \mathcal{B} \oplus K^s$, where $K = \ker E[.]$ and K^s is the class of real, centered random variables, and for this we want to regard the expected value as a linear mapping from $\mathcal{A}^s$ onto a commutative algebra $\mathcal{B}^s$ (which in this case coincides with $\mathbb{R}$). This additive decomposition at the Lie algebras level induces a multiplicative decomposition at the group level. That is, we can write any positive element in $g \in G^+$ as $g = e^X = e^{<1,X>}e^{X-<1,X>}$.

This establishes a mapping from $\mathbb{R} \times C$ where $C = \{e^Y \,|\, E[Y] = 0\}$ onto G^+. We thus obtain a correspondence between the approach to exponential families by Pistone and Sempi in [9] and that of Porta and Recht in [8].

Note now that the projection

$$g = e^{<1,X>} e^{X-<1,X>} \longrightarrow \frac{e^{X-<1,X>}}{E[e^{X-<1,X>}]}$$

is independent of $e^{<1,X>}$. This motivates the following: To make the map Φ a bijection, we have to restrict its domain. Basically $\mathbb{R}^n$ is an n-dimensional manifold whereas $\mathbb{P}$ is only $(n-1)$-dimensional. Thus if we define

$$\begin{aligned} \Phi : K^s &\rightarrow \mathbb{P} \\ Y &\rightarrow \Phi(Y) = \frac{e^Y}{E[e^Y]} \end{aligned} \tag{3.3}$$

we have a bijection, the inverse mapping being given by

$$q(i) \rightarrow Y(i) = \ln q(i) - \frac{1}{n} \sum_{j=1}^{n} \ln q(j)\,. \tag{3.4}$$

To conclude, we note that the special role played by the vector 1 can be done away as follows: We could define expected values with respect to any given (and fixed) $p \in G^+$ by the standard

$$E_p[X] = \sum_i p(i) X(i) =< p, X > .$$

We note again that $e^X = e^{<p,X>} e^{X-<p,X>}$ and if we put $K_p \equiv (\ker E_p[.])^s$, or more explicitly:

$$K_p = \{X \in \mathbb{R}^n | \Sigma_{i=1}^n p(i) X(i) = 0\} = \{X : [1,n] \rightarrow \mathbb{R} | E_p[X] = 0\}$$

which is a linear subspace of $\mathbb{R}^n$, on which the following maps are defined

$$\begin{aligned} \Psi_p(Y) : K_p &\rightarrow \mathbb{R} \\ Y &\rightarrow \Psi_p(Y) = \ln E_p[e^Y], \\ \Phi_p(Y) : K_p &\rightarrow \mathbb{P} \\ Y &\rightarrow e^{Y - \Psi_p(Y)} p = \frac{e^Y}{E_p[e^Y]} p. \end{aligned}$$

Now it is not hard to see that making use of the collection $\{K_p, \Phi_p\}$ an atlas for $\mathbb{P}$ can be defined, and to compute the change of coordinates maps it helps to know that the inverse to Φ_p is given by

$$q(i) \rightarrow Y(i) = \ln(q(i)/p(i)) - \sum_j p(j) \ln\Big(q(j)/p(j)\Big)\,.$$

Comment 3.2 *Note that the mapping given in* (3.4) *also establishes a bijection between the manifold* $\mathbb{P}$ *and the non-euclidean* $n-1$ *dimensional (hyperbolic) manifold* $\mathcal{C} = \{\xi \in \mathbb{R}^n \mid \prod_j \xi_j = 1\}$. *Let us advance some of its basic properties. The original connection* κ *on* G^+ *can be restricted to* $\mathcal{C}$ *and it coincides with the pullback of the connection induced on* $\mathbb{P}$ *by the projection mapping. But more interesting from the probabilistic point of view is the fact that the Euclidean metric, defined at* $1 \in \mathcal{C}$, *coincides with the covariance. That is, if* $\xi(t) = e^{t(X-E[X])}$ *and* $\eta(t) = e^{t(Y-E[Y])}$ *are two curves passing through* 1, *then* $< \frac{d\xi(0)}{dt}, \frac{d\eta(0)}{dt} >_1 = Cov(X,Y)$. *This scalar product can be defined by geodesic translation at every point of* $\mathcal{C}$. *It so happens that the metric associated with it is positive semidefinite, and the connection associated with this metric coincides with* κ *(both are torsion free and invariant under parallel transport, thus they must coincide) Some general results in this direction appear in [8]. This is perhaps an interesting and unexplored connection between geometry and probability. This is a good point to stop and rapidly review the*

3.2. The maximum entropy method

Let us now consider the problem of finding a measure q on $\Omega = \{1, \ldots, n\}$ such that (1.1) is satisfied for a given random variable X. We set $K = 1$ to keep things really simple. A variational method proposed by Jaynes in [5], builded upon the intuition of Boltzmann and Gibbs, can be summarized as follows. Choose a (prior) measure (or probability) p on Ω, and consider the class $\mathcal{K}_p = \{q << p \mid (1.1) \text{ holds}\}$, which is easily seen to be convex when not empty. On it, denote (this is not Jaynes's notation but close enough) the function $K(q,p) = \sum_1^n q(i) \ln(\frac{q(i)}{p(i)})$, which is convex in q, positive and has a minimum at p. Now consider the exponential family (1.2), that is $q_i(t) = \frac{e^{-tX(i)}}{Z(t)} p(i)$, where $Z(t)$ is the obvious normalization factor (but for physicists this is where the thermodynamics comes from). The problem now is to find a value of the parameter t such that $q(t)$ lies in $\mathcal{K}_p$. This is the part of the story told in [2], say. We will only add that in [4], Escher considered such a method, but did not present it as a generic variational method.

References

[1] Barndorff-Nielsen, O.: *Information and exponential families in statistical theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester, 1978.

[2] Borwein, J. and Lewis, A.: *Convex Analysis and Nonlinear Optimization*. CMS Books in Mathematics **3**. Springer-Verlag, New York, 2000.

[3] Corach, G.; Porta, H. and Recht, H.: The geometry of the space of selfadjoint invertible elements in a C^*-algebra. *Integral Equations Operator Theory* **16** (1993), no. 3, 335–359.

[4] Escher, F.: On the probability function in the collective theory of risk. *Scandinavisk Aktuarietidskrift* **15** (1932), 175–195.

[5] Jaynes, E.: Information Theory and Statistical Physics. *Phys. Rev.* **106** (1957), 620–630.

[6] Kobayashi, S. and Nomizu, K.: *Foundations of Differential Geometry*, vol. II. Interscience Tracts in Pure and Applied Mathematics **15**, vol. II. Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1969.

[7] Porta, H. and Recht, L.: Conditional expectations and operator decompositions. *Ann. Global Anal. Geom.* **12** (1994), no. 4, 335–339.

[8] Porta, H. and Recht, L.: Exponential sets and their geometric motions. *J. Geom. Anal.* **6** (1996), no. 2, 277–285.

[9] Pistone, G. and Sempi, C.: An infinite dimensional geometric structure on the space of all probability measures equivalent to a given one. *Ann. Statist.* **23** (1995), no. 5, 1543–1561.

Recibido: 22 de marzo de 2004
Revisado: 2 de julio de 2005

Henryk Gzyl
Departamento de Estadística
Universidad Carlos III de Madrid
C/ Madrid, 126, 28903-Getafe, Madrid (Spain)
and
Departamento de Estadística y Cómputo Científico
Universidad Simón Bolívar
AP 89000,Caracas 1080-A,Venezuela
hgzyl@est-econ.uc3m.es, hgzyl@usb.ve

Lázaro Recht
Departamento de Matemática
Universidad Simón Bolívar
AP 89000, Caracas 1080-A (Venezuela)
recht@usb.ve