Information Geometry for Complex Systems. Background material

2nd PhD School on Mathematical Modeling of Complex Systems, Pescara, 16-28 July, 2012

Demetris P.K. Ghikas,*
Department of Physics University of Patras, Patras 26500, Greece

May 29, 2012

Abstract

Information Geometry offers the framework for a geometric formulation and quantification of entities related to information. The geometric structures are directly associated to information theoretic concepts which can be analyzed with concrete geometric tools. These concepts and tools give the means to visualize the information properties and relations and to derive quantitative results in practical applications. This Background Material is meant to give, to an unitiated reader, the absolute minimum background on some concepts and tools of Differential Geometry and Information Theory. It is hoped that this material will make the presentation of Information Geometry in the main talk, smoother and faster.
KEY WORDS : Differential Geometry, metrics, connections, information measures.

*ghikas@physics.upatras.gr, Tel: +302610-997460, FAX: +302610-997617

1 Motivation for reading this background

.

The term *Information Geometry* points to a conceptual construction which connects concepts of, a priori, totally different nature. But one can immediately imagine that some things related to information are points of a space on which some geometry can be assumed to provide distances between these things. This vague statement expressed in mathematical terms is indeed the framework of Information Geometry. The points related to information are probability distributions, the space is the manifold associated with their parameters and the geometry is coming from a unique metric with a statistical meaning. A useful mental picture is the set of normal distributions where the manifold coordinates are the mean value and the standard deviation. The question "how far are two normal distributions" requires an appropriate metric which is closely related to the estimation error. This is the *Fisher-Rao metric*. Thus we need to clarify the relation between probability distributions and various concepts of information. Then we need to understand how to work on manifolds whose points are these probability distributions. And then to be able to select an appropriate Riemannian metric. Living on a manifold we need to know the information theoretic meaning of travelling on this manifold. The most "conservative" or "controlled" way is to move our measuring stick *parallel to itself*. But how we may define and check this *parallelism*? Here comes the concept of *connection*. Connections help us to construct the *geodesics*. These are the *parallel lines* of the manifold. In general these are not the shortest lines. The connections whose geodesics are shortest lines are the *Riemannian Connections*. In applications of Information Geometry many non-Riemannian connections are used. But as it has turned out, for some special applications we need more general geometric tools, beyond those of a usual geometric manifold. That is, we need to go beyond metric properties, and ask how two distributions differ, not how far they are. This is offered by the so called *contrast functions*. These are generalizations of the *Entropy* and *Relative Entropy* functionals. All these

special information geometric concepts are going to be introduced in the main talk, where some of their application are also discribed. Thus the absolutely minimal background we need to present here concerns the concepts of

1) Differentiable Manifolds
2) Tangent Vectors
3) Riemannian Metrics
4) Connections, Parallel Transport and Curvature.
5) Information Theory and Entropic Functionals.

2 Differential Geometry

2.1 Differentiable manifolds

.

Suppose we are told that the outcomes of a test are values of a random variable normally distributed. With no other information we need to know the parameters (mean and standard deviation) of the gaussian density function. We may imagine for any pair of these parameters that there is a point on a two dimensional surface. To specify the point we need a coordinate system related to the parameters and a correspondence between the points of the surface and the two dimensional space of the parameter values. This gives us the idea of a *manifold*.

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

More generally let $\mathcal{M}$ be a set. Since we need the concepts of neighborhoods and continuity of maps we must assume some topology on $\mathcal{M}$. For the consistent construction of various geometric structures it is assumed that the topology has denumerable basis and it is Hausdorff. If you do not know these concepts, do not mind, since in the most common applications of Information Geometry everything works well. On the other hand, since information manifolds have been introduced for the statistical estimation problems of parameters, historically and in most current applications only local properties are needed. Thus there is no immediate need for global topological properties. Still, from a theoretical point of view the global properties of info

manifolds are interesting research problems.

Definition
A topological space $\mathcal{M}$ is called *m-dimensional Euclidean* if for each point $p \in M$ their exists a neighborhood of p that is homeomorphic to R^m.

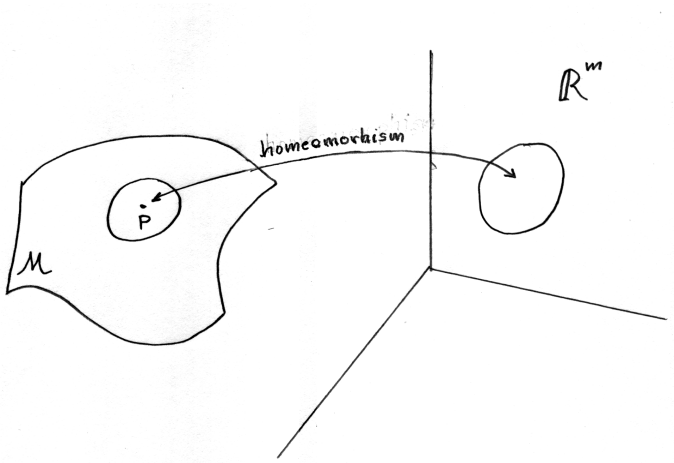

Figure 1: Definition of m-Euclidean manifold

Definition
A topological space $\mathcal{M}$ is an *m-dimensional topological manifold* if its topology is second countable, Hausdorff and $\mathcal{M}$ is *m-dimensional Euclidean.*

In order to be able to construct various differential geometric entities we need, first of all, to implement differentiation. This can be done through the above correspondence which associates to the points of $\mathcal{M}$, m coordinate functions. These functions are defined for a given coordinate system, but in order to have a definition of differentiability to be independent of the coordinate system, some additional compatibility properties must be assumed. Let ϕ and ψ be two coordinate systems. They are *compatible* if both $\phi \circ \psi^{-1}$ and $\psi \circ \phi^{-1}$ are smooth (infinitely differentiable) on their respective domains which are open subsets of R^m. This means that the two coordinate systems ϕ, ϕ are compatible if the transformation from the ϕ-coordinate system to

the ψ-coordinate system is a diffeomorphism.

Definition
A *local coordinate system* at a point p in an m-dimensional topological manifold is a pair (x,U) in which U is an open set containing p and x is a homeomorphism of U onto an open subset of R^m.

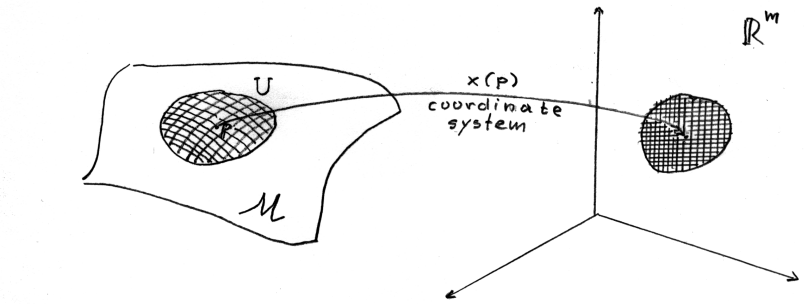

Figure 2: A local coordinate system

Definition
Two coordinate systems (x_α, U_α) and (x_β, U_β) on a topological manifold are *compatible* if both $x_\alpha \circ x_\beta^{-1}$ and $x_\beta \circ x_\alpha^{-1}$ are infinitely differentiable.

A *smooth structure* or *atlas* on a topological manifold $\mathcal{M}$ is a collection of compatible coordinate systems that cover M.

Definition
A *smooth manifold* or *differentiable manifold* is a topological manifold with a maximal smooth structure.

Information Geometry considers sets of parametric distributions. The number of parameters corresponds to the dimensionality of the manifolds. The usual distribution functions give smooth manifolds, as it is evident from the example of the family of normal distributions. But in the applications we

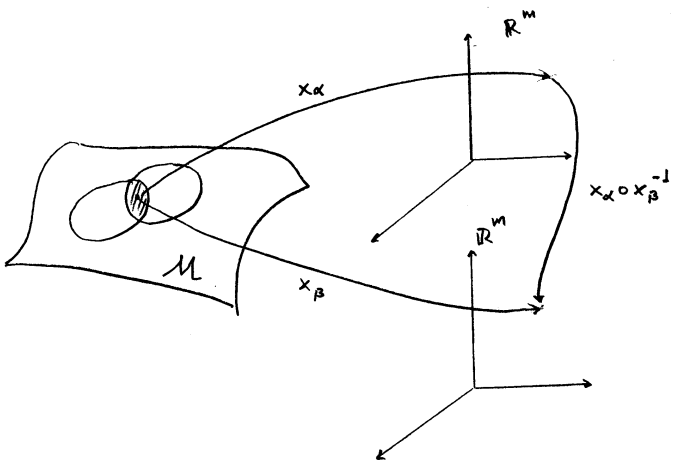

Figure 3: Compatible Coordinate Systems

need to consider subfamilies or to think of a family as a subfamily of a bigger one. This in turn corresponds to the consideration of subsets belonging to a given manifold.

To make this precise we must define first the concept of a function defined on a manifold. This will be a map between manifolds $f : \mathcal{M} \to S$ where a point $p \in \mathcal{M}$ is mapped to the point $f(p) \in \mathcal{S}$. But in order to be able to characterize this map we must use the corresponding coordinate systems. Thus we have

Definition
A function $f : \mathcal{M} \to S$ from an m-dimensional smooth manifold $\mathcal{M}$ into an k-dimensional smooth manifold $\mathcal{S}$ is s-fold continuously differentiable and of rank r at $p \in \mathcal{M}$ if for some coordinate system (x,U) at p and (y,V) at f(p) the function $y \circ f \circ x^{-1}$ is s-fold continuously differentiable and of rank r at x(p), where $r = 0, 1, ..., min\{k, m\}$. If f is infinitely differentiable at p it is called smooth at p. If it is smooth at every point of $\mathcal{M}$ it is called smooth. The set of all real smooth functions on $\mathcal{M}$ is denoted by $C^\infty(\mathcal{M})$

The meaning of this definition is that to characterize the map from p to f(p) we go from the point x(p) in R^m which is given by the m-tuple of the

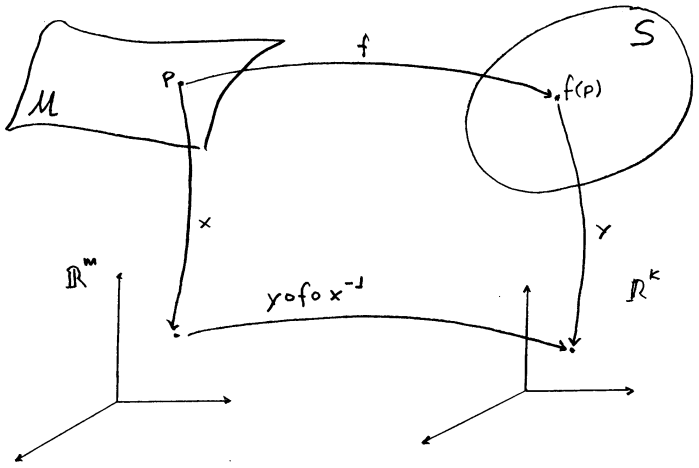

Figure 4: Maps between manifolds

coordinates of p , back to the manifold point p, then we map the point to f(p), and there we have the k-tuple y(f(p)) of the coordinate of f(p). Thus we use the known definition of functions between the Eucledian spaces $R^m \to R^k$. This definition gives us the main morphism used in Differential Geometry, namely

Definition
The function $f : \mathcal{M} \to S$ is a *diffeomorphism* if it is a homeomorphism of the smooth manifold $\mathcal{M}$ onto the smooth manifold S and both f and f^{-1} are smooth.

The importance of this definition is, that from the homeomorphisms which preserve the topological structure we go to diffeomorhisms which preserve the differential structure as well. Thus the generalization of the picture of a two dimenstional surface in R^3 is that of a submanifold.

Definition
A mapping $f : \mathcal{M} \to S$ of a m-dimensional manifold into a k-dimensional manifold is an *imbedding* if it is a smooth map of rank r that is homeomorphic to its image $f(M) \subset S$. If f is an imbedding then f(M) is an *imbedded*

submanifold of S.

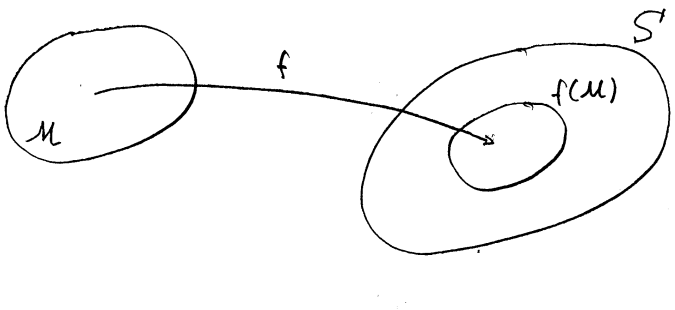

Figure 5: Imbedded Submanifold

2.2 Tangent Vectors

.

In Eucledian Geometry, "tangent vectors" are meant to be "tangents" to smooth curves. Suppose we are given a curve on $\mathcal{M}$. This is a one-to-one function $\gamma : I \to \mathcal{M}$ from some interval $I \subset R$ to M. For a selected coordinate system let $\gamma^i(t) = \xi^i(\gamma(t))$, Then $\bar{\gamma}(t) = [\gamma^1, ..., \gamma^m]$ is a curve in R^m. If this curve is C^∞ then it is said to be a C^∞ curve on $\mathcal{M}$. If $\mathcal{M}$ is a subset of R^n , then the tangent to this curve at a point a of $\mathcal{M}$ is defined through the limit

$$\dot{\gamma}(a) = \lim_{h \to 0} \frac{\gamma(a+h) - \gamma(a)}{h}. \tag{2}$$

But in this definition all the relevant vectors belong in the same vector space, and the limit is mathematically well defined. When $\mathcal{M}$ is a general manifold this definition of the tangent of a curve cannot be used. But, on the other hand, we may define the derivatives of functions defined on curves, because these functions are real-valued. Let f be a C^∞ on $\mathcal{M}$. $f(\gamma(t)) = \bar{f}(\bar{\gamma}(t)) = \bar{f}(\gamma^1(t), ..., \gamma^m(t))$ is a real-valued function and we have the usual derivative definition

$$\frac{d}{dt} f(\gamma(t)) = \left(\frac{\partial \bar{f}}{\partial \xi^i} \right)_{\gamma(\bar{t})} \frac{d\gamma^i(t)}{dt}. \tag{3}$$

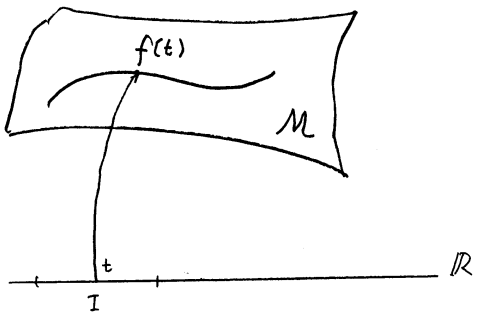

Figure 6: Functions on the Manifold

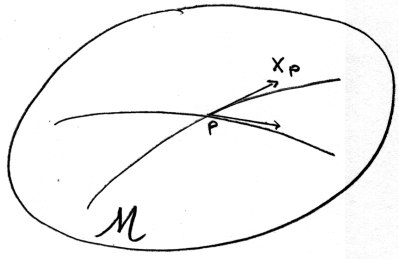

Figure 7: Tangent Vectors

This is the form of a directional derivative. Thus we may define the tangent vectors as *operators* acting on smooth functions defined on the manifold. Namely we may define the tangent on the curve $\gamma(t)$ as

$$\dot{\gamma}(a) = \left(\frac{d\gamma}{dt}\right)_p = \dot{\gamma}^i(a)\left(\frac{\partial}{\partial \xi^i}\right)_p \tag{4}$$

More formally we have

Definition
Let $\mathcal{M}$ be a smooth manifold. A *tangent vector* at a point $p \in M$ is a mapping $X_p : C^\infty(\mathcal{M}) \to R$ such that for all $f, g \in C^\infty$, and $a \in R$ the following properties hold
(i) $X_a(a \cdot f + g) = a \cdot X_p(f) + X_p(g)$
(ii) $X_p(f \cdot g) = g \cdot X_p(f) + f \cdot X_p(g)$

This means that tangent vectors satisfy the algebraic properties of vectors

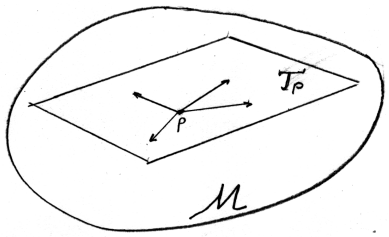

Figure 8: Tangent Space

and the Leibnitz rule for derivatives.

Definition
The set of tangent vectors at p is the *tangent space* of $\mathcal{M}$ at p.

This is denoted by $T_p\mathcal{M}$. For a given coordinate mapping x, defined on an open set $U \subseteq M$ with $x : U \to R^m$ we have $x = (x^1, ..., x^m)$. Then the a^{th} x-coordinate basis tangent vector at the point p is $\partial/\partial x^a|_p$. With this notation a general tangent vector is written as

$$X_p = \sum v^a(p) \cdot \frac{\partial}{\partial x^a}|_p \tag{5}$$

The functions v^a are the components of the tangent vector at p. A smooth family of tangent vectors is a vector field defined on $\mathcal{M}$. The space of all vector fields on $\mathcal{M}$ is denoted by $\mathcal{X}(\mathcal{M})$

2.3 Riemannian Metrics

.

For every point $p \in \mathcal{M}$ the tangent space $T_p\mathcal{M}$ is a vector space on which we may define an inner product. The collection of all these inner products $< \cdot, \cdot >= \{< \cdot, \cdot >_p \, | p \in \mathcal{M}\}$ gives a metric structure on the manifold. We have

Definition
A smooth *Riemannian metric* is a function $< \cdot, \cdot >$: $\mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \mapsto \mathcal{C}^\infty(\mathcal{M})$ satisfying for all $f \in \mathcal{C}^\infty(\mathcal{M})$ and all $X, Y, Z \in \mathcal{X}(\mathcal{M})$

$$< X, Y > \; = \; < Y, X > \tag{6}$$

$$< fX, Y > \; = \; f < X, Y > \tag{7}$$

$$< X + Z, Y > \; = \; < X, Y > \; + \; < Z, Y > \tag{8}$$

and to avoid degeneracy, for all $X \neq 0$ and all $p \in \mathcal{M}$

$$< X, X > \; > \; 0 \tag{9}$$

It is evident that these are properties that an inner product on a vector space must satisfy, namely to be bilinear, symmetric and positive definite. With this metric structure we have

Definition
A *smooth Riemannian manifold* is a pair $(\mathcal{M}, < \cdot, \cdot >)$ where $\mathcal{M}$ is a smooth manifold and $< \cdot, \cdot >$ is a Riemannian metric on $\mathcal{M}$.

For a given coordinate basis on $\mathcal{M}$ the metric evaluated on the basis elements is written as $g_{ab} =< \partial_a, \partial_b >$. Then the Riemannian manifold is denoted by $(\mathcal{M}, g)$. A diffeomorphism between two Riemannian manifolds which preserves the metric is called an *isometry*.

2.4 Connections, Parallel Transport and Curvature

.

Suppose we want to draw a straight line beween two points on a plane. There are two ways to do that. We start from one point with a small vector pointing to the direction of the other point and draw small vectors parallel to each other untill we reach the other point. An easier way is to stretch a string between the two points and draw the line parallel to the string which is the minimal line connecting the two points.. Of course we get the same result. Suppose now that we want to do the same thing on the surface of a sphere. The second method is straightforward, because we know that the great circles are the minimal lines connecting two points on the sphere. The equivalent to the fisrt method is that we must draw tangent vectors parallel to themselves. Again we obtain great circles. But now let us imagine the following experiment. We draw a triangle on the plane and move a vector parallel to itself, that is making constant angle with the sides of the triangle. Upon returning to the initial point we see that the final vector is identical to the initial. If we do the experiment on the sphere with a triangle formed by great circles, we discover that the final vector differs from the initial one. We will discover that , more generally, the difference between the initial and final vectors depend on the paths followed. This is due to the fact that the sphere is not "flat". This is not a metric property, and it is coming from the meaning of the concept "straight". Behind this is the fundamental geometric concept of connection and its role in the definition of parallel transport. It is a special case that the "straight lines", or *geodesics* of a connection are also minimal lines. This is the *Riemannian connection.*

Let us be more formal. In the definition of parallel transport of vectors defined on a plane we use the fact that these vectors belong, for any pair of points , to vector spaces which can be identified. That is, there is a natural linear mapping between the bases of these spaces. But for two points $p, p' \in \mathcal{M}$ there is no natural correspondence between the tangent spaces $T_p\mathcal{M}$ and $T_{p'}\mathcal{M}$. A *connection* is such a correspondence. Let $\Pi_{p,p'}$ be a linear mapping between the tangent spaces at p and p'. Let p and p' be infinitesimally close. This means that the corresponding coordinates are infinitesimaly close. If $[\xi^i]$ is a coordinate system for $\mathcal{M}$ let $d\xi^i = \xi^i(p') - \xi^i(p)$. Let $\{(\partial_1)_p, ..., (\partial_m)_p\}$ and $\{(\partial_1)_{p'}, ..., (\partial_m)_{p'}\}$ be the corresponding basis vectors of the tangent spaces. Then the linear mapping must give an expression of the basis vectors at p in

terms of the basis vectors at p'. We must have

$$\Pi_{p,p'}((\partial_j)_p) \;=\; (\partial_j)_{p'} \;+\; d\xi^i(\Gamma^k_{ij})_p(\partial_k)_{p'} \tag{10}$$

where $(\Gamma^k_{ij})_p : i,j,k = 1,...,m$ are m^3 real numbers, depending on the point p. If for each pair of neighboring points p and p' in $\mathcal{M}$ there is a linear mapping $\Pi_{p,p'} : T_p \to T_{p'}$ defined by the previous equation and if the $\Gamma' s$ as functions of p are all $\mathcal{C}^\infty$ then this linear mapping is an *affine connection.* The set of functions (Γ^k_{ij}) are called *connection coefficients.*

Now having a rule of correspondence between the vectors of the tangent spaces we may define the "parallel translation of tangent vectors". Let two points p,q on $\mathcal{M}$ and a curve which connects them $\gamma : [a,b] \to \mathcal{M}$ such that $\gamma(a) = p$ and $\gamma(b) = q$. A *tangent vector* along γ is a mapping from each point $\gamma(t)$ to the tangent vector $X(t) \in T_{\gamma(t)}$. If for infinitesimally close points on the curve we have

$$X(t+dt) \;=\; \Pi_{\gamma(t),\gamma(t+dt)}(X(t)) \tag{11}$$

then X(t) is defined to be *parallel* along γ. Using the connection coefficients,

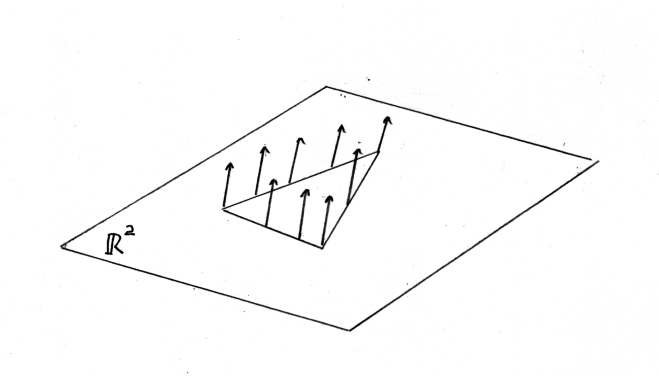

Figure 9: Moving parallel vectors defined on the plane

this condition on X(t) becomes

$$\dot{X}^k(t) \;+\; \dot{\gamma}^i(t)X^j(t)(\Gamma^k_{ij})_{\gamma(t)} \;=\; 0 \tag{12}$$

The solution of this ordinary differential equation is the *parallel translation along* γ of X(a) defined on p.

Such a parallel translation helps us to define the derivatives of tangent

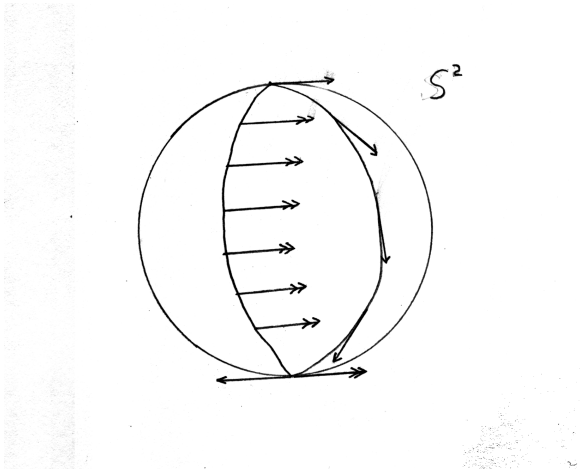

Figure 10: Moving parallel vectors on the sphere

vectors. In the definition of the derivative of vectors on the plane we construct the infinitesimal change of the vectors from the parallel translate of the vector from its initial value. We do exactly this here but using the definition of the parallel translation along the curve as defined above. This leads us to the definition of the *covariant derivative* $\frac{\delta X}{dt}$. We have

$$\delta X(t) \;=\; \Pi_{\gamma(t+dt),\gamma(t)}(X(t+dt)) \;-\; X(t) \tag{13}$$

and dividing by dt we obtain in a coordinate system

$$\frac{\delta X(t)}{dt} \;=\; \{\dot{X}^k(t) \;+\; \dot{\gamma}^i(t)X^j(t)(\Gamma^k_{ij})_{\gamma(t)}\}(\partial_k)_{\gamma(t)} \tag{14}$$

From this it is evident that parallel translation is equivalent to the vanishing of the covariant derivative. Now the concept of directional derivatives of functions on $\mathcal{M}$ implemented by the tangent vectors which are differential operators on $C^\infty(\mathcal{M})$ can be extended to "directional derivatives of tangent vectors". This variation of vector fields in various directions gives the essential information of the "shape" of the manifold. Let a vector field $X = X^i\partial_i$. For a vector field $Z = Z^i(\partial_i)_p \in T_p\mathcal{M}$ we want the derivative of X along Z. Considering a curve for which Z is the tangent vector we use the definition of covariant derivative along this curve. Denoting this covariant derivative with $\nabla_Z X$ we get from the previous equations

$$\nabla_Z X \;=\; Z^i\{(\partial_i X^k)_p \;+\; X^j_p(\Gamma^k_{ij})_p\}(\partial_k)_p \tag{15}$$

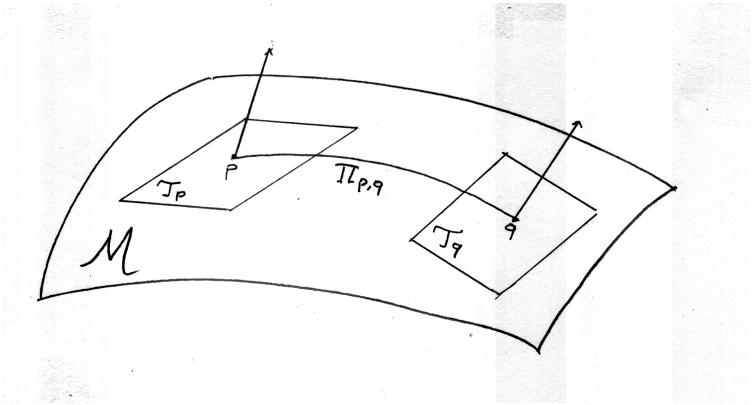

Figure 11: Associating vectors of different tangent spaces

We observe that $\nabla_Z X \in \mathcal{X}(\mathcal{M})$. Thus for any pair of vector fields $X, Y \in \mathcal{X}(\mathcal{M})$, given by $X = X^i\partial_i$ and $Y = Y^i\partial_i$ we may define the vector field $\nabla_X Y \in \mathcal{X}(\mathcal{M})$ with

$$\nabla_X Y \;=\; X^i\{\partial_i Y^k \;+\; Y^j\Gamma^k_{ij}\}\partial_k \tag{16}$$

We call this the *covariant derivative* of Y with respect of X. For the basis vectors this becomes

$$\nabla_{\partial_i}\partial_j \;=\; \Gamma^k_{ij}\partial_k \tag{17}$$

This equation expresses the action of the covariant derivative as a trasfor-

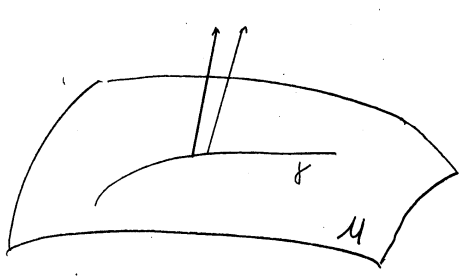

Figure 12: Using the connection for the parallel translation of vectors and the construction of derivatives.

mation of the basis of tangent vector as we move in the direction of each

basis vector.

It follows that a covariant derivative ∇, for $X, Y, Z \in \mathcal{X}(\mathcal{M})$ and $f \in C^{\infty}(\mathcal{M})$ satisfies the properties

$$i) \quad \nabla_{X+Y} Z = \nabla_X Z + \nabla_Y Z \tag{18}$$

$$ii) \quad \nabla_X (Y + Z) = \nabla_X Y + \nabla_X Z \tag{19}$$

$$iii) \quad \nabla_{fX} Y = f \nabla_X Y \tag{20}$$

$$iv) \quad \nabla_X (fY) = f \nabla_X Y + (Xf) Y \tag{21}$$

These properties can be used as the defining properties of a connection on a manifold. We note that the family of all connections on a smooth manifold forms an affine space. This is very important, and useful for Information Geometry, where, as we are going to see, there exist connections with statistical meaning. These may be combined linearly to form connections which play a particular role.

Now given a connection ∇ suppose that there exists a coordinate system with the property that all its basis vectors are parallel with respect to the connection. That is, all the covariant derivatives $\nabla_{\partial_i} \partial_j$ are zero. This means that, for this coordinate system, all connection coefficients vanish identically. Then this is an *affine coordinate system* for ∇, and the connection is said to be *flat*. The existence of such a coordinate system is a property of the connection, and though it is not trivial to find such coordinates, there exist geometric structures derived from the connection which are related to this existence. The idea is to have properties which characterize the connection but do not depend on the coordinate system. These are maps from pairs or triplets of vector fields with values in the space of vector fields. The important property is that they are tensors, and so their value does not depend on the coordinate system. Let $X, Y, Z \in \mathcal{X}(\mathcal{M})$. We have

$$R(X, Y) Z = \nabla_X (\nabla_Y Z) - \nabla_Y (\nabla_X Z) - \nabla_{[X,Y]} Z \tag{22}$$

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y] \tag{23}$$

where the commutator [X,Y], with $X = X^i \partial_i$ and $Y = Y^i \partial_i$ is the vector field

$$[X, Y] = (X^j \partial_j Y^i - Y^j \partial_j X^i) \partial_j \tag{24}$$

These tensor fields are called respectively *Riemann - Christoffel curvature tensor* and *torsion tensor*. In a specific coordinate system their tensor elements are

$$R^l_{ijk} = \partial_i \Gamma^l_{jk} - \partial_j \Gamma^l_{ik} + \Gamma^l_{ih} \Gamma^h_{jk} - \Gamma^l_{jh} \Gamma^h_{ik} \tag{25}$$

$$T^k_{ij} \;=\; \Gamma^k_{ij} \;-\; \Gamma^k_{ji} \tag{26}$$

As we said before, when the connection is flat, there exists and affine coordinate system in which the connection coefficients are zero. Then, in this system the tensors R and T are zero. But since they are tensors they are zero in all coordinate systems. In this case we say that the manifold is *flat* with respect with this connection. If the torsion is zero, namely $T^k_{ij} = 0$ then the connection coefficients are symmetric with respect to the indexes ij, and it is called *symmetric connection* or *torsion-free connection*. In the Classical Information Geometry mainly symmetric connections are introduced.

From the expressions above for the elements of R and T it follows that 1-dimensional manifolds are automatically flat. Now suppose that the manifold $\mathcal{N}$ is a submanifold of the manifold $\mathcal{M}$ and let a connection ∇ on $\mathcal{M}$. Then in general the covariant derivative $\nabla_X Y$ for vector fields of $\mathcal{N}$ would be outside of $\mathcal{X}(\mathcal{N})$. If for all $X, Y \in \mathcal{X}(\mathcal{N})$ we have

$$\nabla_X Y \in \mathcal{X}(\mathcal{N}) \tag{27}$$

the submanifold $\mathcal{N}$ is called *autoparallel*. 1-dimensional autoparallel submanifolds are the *geodesics* for this connection of $\mathcal{M}$. These curves then can be parametrized by an affine coordinate γ, and satisfy the ordinary differential equation

$$\ddot{\gamma}(t) \;+\; \dot{\gamma}^i(t)\dot{\gamma}^j(t)(\Gamma^k_{ij})_{\gamma(t)} \;=\; 0 \tag{28}$$

where $\gamma^i = \xi^i \circ \gamma$. Now having a definiton of geodesics as the *parallel lines for a given connection* we must face our "metric prejudices" that geodesics are lines between points with minimal lenght. Thus a natural question is what is the relation between the metric geodesics with those defined through a connection. To answer this we must think of a condition that combines the two independent geometric structures : metric and connection. A natural condition is that if two vector fields are parallel transported along a line on the manifold (action of a connection) their inner product should not change (metric condition). Thus if the covariant derivatives of the two vectors along the curve are zero, namely $\frac{\delta}{dt}X = 0$ and $\frac{\delta}{dt}Y = 0$ we get from the equation

$$\frac{d}{dt} < X(t), Y(t) > \;=\; \left\langle \frac{\delta X(t)}{dt}, Y(t) \right\rangle \;+\; \left\langle X(t), \frac{\delta Y(t)}{dt} \right\rangle \tag{29}$$

that the inner product remains constant. This special connection is unique and it is called *metric connection* or *Riemannian connection*. Its connection

coefficients are given by

$$\Gamma_{ij,k} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}) \tag{30}$$

where $\Gamma_{ij,k} = \Gamma^h_{ij} g_{hk}$.

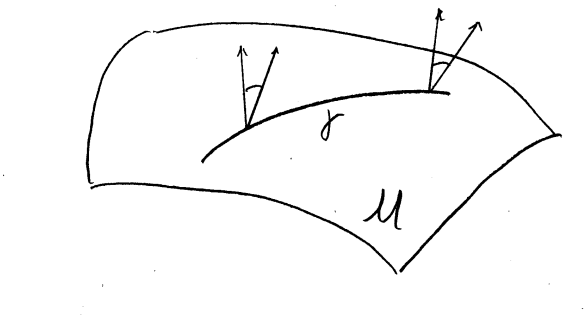

Figure 13: The metric connection leaves invariant the inner product

3 Classical Information Theory

3.1 Introduction

.

As it is going to be presented in the talk, historically, Information Geometry started from the observation that a certain metric on the manifold of parametric probability distribution has definite statistical meaning and usefulness. Then particular connections were introduced which are flat and play an important role in the developement of many applications. It also became obvious that more general geometries may be introduced. These geometries are based on the so called *contrast functions*. These are generalizations of the relative entropy concept of Information Theory. So , in this Background Material for the talk, we present some motivation and the most essential definitions of these functionals.

3.2 Uncertainty, Entropy and Information

.

In the search for a definition and quantification of the concept of "Information", it turned out to be natural to ponder about the opposite concept, that of "Uncertainty". In this way, we may accept the intuitive expectation that "the decrease of uncertainty means, or is equal to the increase of information". And the reason to follow this route is that it seems easier to formalize the concept of "uncertainty". Shannon, in 1948, introduced his definition of Entropy, as a concept essentially identified with Uncertainty. To be able to derive the mathematical form of Entropy, he started with certain axioms that a quantity of Uncertainty must satisfy. These axioms of Uncertainty are intuitively plausible. Asking the question "uncertainty of what?" , we may say "of the occurance or not of certain events". But this immediately brings us to randomness of events, and thus to probabilities that certain random variables take various values. But then the uncertainty must not depend on the particular values that the random variable takes, but only on the probabilities of the events. Thus suppose that we have a random variable X, which can take the values x_i with probabilities $p_i, 1 \leq i \leq n$. Let $H(p_1, ..., p_n)$ be the uncertainty of the outcome of the values of X. We expect the following to be the minimal properties that this function must satisfy.

A1 $H(p_1, ..., p_n)$ is maximum when $p_1 = p_2 = ... = p_n = 1/n$
A2 $H(p_1, ..., p_n)$ must be symmetric in its arguments.
A3 $H(p_1, ..., p_n) \geq 0$. Is zero only when one p_i is equal to 1.
A4 $H(p_1, ..., p_n, 0) = H(p_1, ..., p_n)$
A5 $H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}) \leq H(\frac{1}{n+1}, \frac{1}{n+1}, ..., \frac{1}{n+1})$
A6 $H(p_1, ..., p_n)$ should be a continuous function of its arguments.
A7 $H(\frac{1}{mn}, \frac{1}{mn}, ..., \frac{1}{mn}) = H(\frac{1}{m}, \frac{1}{m}, ..., \frac{1}{m}) + H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$
A8 $H(p_1, ..., p_m, q_1, ..., q_n) = H(p, q) + pH(p_1/p, ..., p_m/p) + qH(q_1/q, ..., q_n/q)$, where $p = p + ... + p_m$ and $q = q + ... + q_n$, with p+q = 1.

The intuitive origin of these axioms is the following:
A1 A fair die is more uncertain than a biased one.
A2 Uncertainty depends only on the probabilities, and not on the order of the appearance of the events (for independent events).
A3 Conventionally uncertainty is a positive quantity. There should be no

uncertainty if there is no randomness.
A4 An event with zero probability cannot affect uncertainty.
A5 More possible events should present more uncertainty.
A6 Small changes of the probabolities of the event should have small influence on the uncertainty.
A7 The uncertainty of two sequences of random events should be the sum of the uncertainties.
A8 Classifying the events into two categories, the total uncertainty should be equal to the uncertainty of the appearance of the two categories plus the weighted sum of the individual uncertainties of the events in the two categories.
Then the following theorem holds

Theorem
Let $H(p_1, ..., p_n)$ be a function defined for any integer n and for all values of $p_1, ..., p_n$, such that

$$p_i \geq 0 \quad \sum_{i=1}^{\infty} p_i = 1. \tag{31}$$

If H satisfies the axioms A1-A8, then

$$H(p_1, ..., p_n) \; = \; -\lambda \sum_k p_k log p_k \tag{32}$$

with λ any positive constant and the sum is over all values for which $p_k > 0$

Now given a random variable with a finite number of values with probabilities such that $\sum p_i = 1, \quad p_i > 0, \quad 1 \leq i \leq n$ the *entropy* of X is defined by the above function.

It seems to be an obvious fact that given two independent random variables, the total uncertainty about them should be the sum of their individual uncertainties, and if there is a dependence between them then the total uncertainty nust be less than the sum of uncertainties. Indeed this is the case as it can be proved that

$$H(X, Y) \; \leq \; H(X) \; + \; H(Y) \tag{33}$$

To make this dependence more quantitative we use conditional probabilities. Let an event A in the probability space of the random variable X. Then the

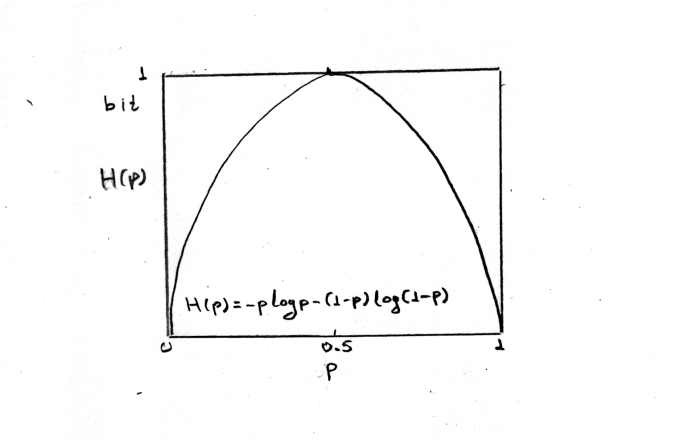

Figure 14: The Shannon Entropy is maximal for equal probabilities

conditional entropy is defined

$$H(X|A) \;=\; -\sum_{k=1}^{\infty} P(X = x_k|A) log P(X = x_k|A) \tag{34}$$

This gives the conditional entropy between two random variables

$$H(X|Y) \;=\; \sum_{j} H(X|Y = y_j) P(Y = y_j) \tag{35}$$

Obviously the following hold

$$H(X|X) \;=\; 0 \tag{36}$$

and for independent X and Y

$$H(X|Y) \;=\; H(X) \tag{37}$$

The Conditional entropy helps us to replace the inequality for the total uncertainty with an equality, as it can easily be proved that

$$H(X,Y) \;=\; H(Y) \;+\; H(X|Y) \tag{38}$$

There is an other route to the definiton of the concept of "Information".

This was proposed by Hartley in 1928. Let two events E_1 and E_2 on the same probability space with probabilities p_1 and p_2 respectively. A measure of information for two independent realizations of the experiment should satisfy the equation

$$I(p_1 p_2) \;=\; I(p_1) \;+\; I(p_2) \tag{39}$$

since for independent events the total probability is the product of the individual probabilities. A function that satisfies this equality is

$$I(E) \;=\; -log_2 P(E) \tag{40}$$

We define I(E) as the *information* of the event E. The logarithm with base 2 measures this information in units of *bits*. Now, for a random variable X, as above, we may define the *information* about X as the average of the information for its individual values, namely

$$H(X) \;=\; -\sum_k p_k log_2 p_k \;=\; -\sum_k p_k I(X = x_k) \tag{41}$$

which is equal to the entropy defined above. Thus we have

$$Information \quad = \quad Entropy \quad = \quad Uncertainty \tag{42}$$

3.3 Relative entropy and Mutual Information

.

We referred to the function H(X,Y) as the total uncertainty for the two random variables. Its "entropic name" is *joint entropy*. Explicitly it is defined as

$$H(X,Y) \;=\; -\sum_x \sum_y p(x,y) log p(x,y) \tag{43}$$

But in practice, in particular in statistical estimation tests, we want to estimate and compare probability distributions. For one random variable we want to quantify how inefficient is to assume that the probability distribution is given by p(x) while the true is q(x). We want to know "how far we are". This is quantified with the *relative entropy* or the *Kullback-Leibler distance*

$$D(p||q) \;=\; \sum_x p(x) log \frac{p(x)}{q(x)} \tag{44}$$

This is not a true distance, since it is not symmetric.

Now for two random variable X and Y a measure of dependence should be a functional of probability distributions which quantifies the difference between the joint distribution and the product of the individual ones. It turns out that the appropriate functional is the *mutual information*

$$I(X,Y) = \sum_x \sum_y p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{45}$$

The following properties of the mutual information reveal its relation to the joint entropy, the information content, and the concept of uncertainty

$$I(X,Y) = H(X) - H(X|Y) \tag{46}$$
$$I(X,Y) = H(Y) - H(Y|X) \tag{47}$$
$$I(X,Y) = H(X) + H(Y) - H(Y|X) \tag{48}$$
$$I(X,Y) = I(Y,X) \tag{49}$$
$$I(X,X) = H(X) \tag{50}$$

Many interesting identities and inequalities hold for these functions. They are the main objects that play a fundamental role in Information Theory. Their relevance to Information Geometry comes form the fact that they can be a starting point for the construction of "distance like" objects, contrast functions and more general geometric structures than the usual Riemannian ones.

4 Bibliographical Comments

.

This Background Material is a minimal introduction to the most elementary concepts of Differential Geometry and Information Theory. It gives the definitions of various concepts that are going to be used in the talk *Information Geometry for Complex Systems I, II*. The presentation has followed some standard books of Differential Geometry, Information Geometry and Information Theory. There are hundreds of introductory and advanced referencies which can be found in the internet. The search with the code words "information geometry" will make the reader dizzy, for sure, but looking at the google scholar for the name "amari" will get tens of papers and talks of Amari which containe both introductory and advanced material. The reader is strongly adviced to do this exercise.

4.1 Differentiale Geometry

.

Information Geometry started as an attempt to give geometrical meaning to the estimation theory of statistics. The differential geometric constructions started from classical parametric statistical models, considering them as points of differential manifolds. Then the interest turned to non-parametric statistics and to questions of quantum estimation theory. Here we presented the absolute minimum background for information geometry of the classical parametric models. The non-parametric and quantum extensions would require far more background which would not be necessary for the talk, which conforms to the philosophy of the School.

For the introduction of the concepts we used the books [1, 2, 3, 4], These are classical referencies for Information Geometry. Their introduction of Differential Geometric concepts are elementary, but good enough to present a foundation of the theory of Information Geometry. The book [5] is a very good introduction of Differential Geometry, with more advanced material. The interested reader will fined in this bool a deeper approach to the concept of connection.

4.2 Information Theory

.

The latest developments of Information Geometry which evolve arround the concepts of contrast functions, are based on ideas of entropy and relative entropy. The book [7] is a classic reference on Information Theory. Though there exist many books on Information Theory, this is introductory and very up to date and complete. The book [6] is concerned with cryptography, but its first chapter is an amazingly clear and intuitive introduction to the concept of entropy.

References

[1] Shun-ichi Amari, *Differential-Geometric Mathods in Statistics*, Lecture Notes in Statistics Vol 28, Springer-Verlag, Berlin, 1985.

[2] Shun-ichi Amari, *Methods of Information Geoemetry*, Translations of Mathematical Monographs Vol 191, AMS, Oxford University Press, Oxford, 2000.

[3] Michael K. Murray, John W. Rice, *Differential Geometry and Statistics*, Chapman & Hall , London, 1994.

[4] Robert E. Kass, Paul W. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York, 1997.

[5] M. Crampin, F.A.E Pirani, *Applicable Differential Geometry*, Cambridge University Press, Cambridge, 1986.

[6] Dominic Welsh, *Codes and Cryptography*, Oxford Science Publications, Clarenton Press, Oxford 1988.

[7] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*, Wiley Series in Communications, John Wiley & Sons, Inc., New York,1991