An Introduction to Information Geometry

Giulia Bertagnolli

2023-03-29T00:00:00+02:00

Table of contents

Preface

This are the lecture notes for the short course (8 hours) *The Geometry of Statistical Models* (Trento - March, 2023). Feedback, as well as reports on typos and errors, are welcome.

1 Introduction

The name C. R. Rao, Professor Emeritus of Statistics at Penn State University, is ubiquitous in statistics and it was him in 1945, who firstly understood the geometrical meaning of Fisher's information (Rao, 1945). Some results by Efron (Efron, 1975), in 1975, inspired Shun-ichi Amari, who discovered the family of affine $\alpha-$connections. Chentsov independently obtained the same results in 1972 (his work became known to the community only in 1982, with the English version of his work (Chentsov, 1982)). Other recurring names in the field, just to name a few, are the ones of A. P. David, Lauritzen—who formalised the concept of a *statistical manifold* for finite sample spaces—Nagaoka, co-author, with Amari, of the very first book on information geometry (Amari & Nagaoka, 2000), Giovanni Pistone for his works on non-parametric IG, see e.g. (Pistone, 2013), and Ay-Jost-Lê-Schwachhöfer, for their recent book (Ay, Jost, Vân Lê, & Schwachhöfer, 2017). Let us start with a very brief and informal introduction of the contents of this short course.

Intuitively, we will start from a sample space Ω and define a differentiable structure on the set $\mathscr{P}(\Omega)$ of probability measures on the sample space. Curves on this (probability/statistical) manifold are 1-dimensional parametric, statistical models. If I is an open interval of $\mathbb{R}$ and the mapping

$$I \ni \theta \mapsto p(\cdot;\theta)\nu$$

is smooth, then we can compute the velocity, acceleration, etc. of the curve and, consequently, we can describe the geometry of the statistical model. (I, Ω, p, ν) is called a (regular) 1-dimensional statistical model.

Observations

i. We have only introduced the sample space Ω, but we will need also a $\sigma-$algebra, i.e. $(\Omega, \mathcal{E})$ ans a $\sigma-$finite measure ν on this space. Then, as we will see, $p(\cdot;\theta)$ is a *density* w.r.t. the reference/dominating measure ν (i.e. we are in an absolute-continuous framework).
ii. When writing $p(x;\theta)$, $x \in \Omega$ is a sample, but we may also consider a random variable $X : \Omega \to \mathbb{R}$, and x represents an *observable* on Ω.
iii. When Ω is infinite, $\mathscr{P}(\Omega)$ is infinite-dimensional.
iv. Given a statistic $\kappa : \Omega \to \Omega'$, what happens to the geometric structure on $\mathscr{P}(\Omega)$? It turns out that the Fisher metric is invariant under sufficient statistics.

Let us look at some examples of manifolds of interest in IG: the set of positive-definite matrices of dimension $n \times n$ is a sub-manifold of dimension $\frac{n(n+1)}{2}$ of the n^2-dimensional manifold of all real matrices of that dimension; the set of neural networks, identified by the connection weights $\mathbf{W}$...

In the remaining of this section, we provide a brief recap of the main definitions of differential geometry, which are *useful* for understanding IG. *Useful*, but not mandatory, as we will see in Section 3.

1.1 Differential Geometry Recap

The core objects of IG are manifolds, more specifically *differentiable* manifolds. So, we need a brief recap of some concepts and tools of differential geometry. For more details see (Lang, 2012; Petersen, 2006; Sernesi, 1994) or the lecture notes of your favourite "Geometric analysis" course (also Moretti 2020).

A manifold is a set M endowed with a manifold structure, which is defined as a collection of *local charts*, an *atlas*.
A *local chart* is a pair (U, φ) where $U \subset M$ and $\varphi : U \to \varphi(U) \subset \mathbb{R}^n$[1] is a bijection and $\varphi(U)$ is open in $\mathbb{R}^n$.
Two charts (U, φ), (V, ψ) are said to be $\mathcal{C}^k-$*compatible* if either $U \cap V = \emptyset$, or the map $\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V) \subset \mathbb{R}^n$ is a bijection, and both this and its inverse $\varphi \circ \psi^{-1} : \psi(U \cap V) \to \varphi(U \cap V) \subset \mathbb{R}^n$ are of class $\mathcal{C}^k$, i.e. $\psi \circ \varphi^{-1}$ is a *diffeomorphism* of class $\mathcal{C}^k$ between open sets of $\mathbb{R}^n$.

An *atlas* of class $\mathcal{C}^k$ is a collection of charts $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$, where $\cup_{\alpha \in I} U_\alpha = M$ and the transition maps are pair-wise $\mathcal{C}^k-$compatible. Finally, we say that the atlas $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ defines a structure of $\mathcal{C}^k-$manifold on M and $\dim M = n$. If the charts are $\mathcal{C}^\infty-$compatible we talk about smooth charts, atlas, and manifold. On the other hand, if $k = 0$ we call the manifold a *topological* manifold.

Remarks

i. One can give M a topology in a unique way such that each U_α is open and the φ_α are topological isomorphisms (or *homeomorphism*, i.e. bijectinve and bi-continuous). In other words, a differentiable structure on M induced a topology on it.

[1] Here, φ could go, in general, to a topological linear space, i.e. a linear space with a topology making the operations of sum and scalar multiplication, continuous(Lang, 2012) (e.g. a Banach space). In this case, the transition map $\psi \circ \varphi^{-1}$ would be an $\mathcal{C}^k-$isomorphism of topological spaces. Here you might ask what is the differentiability for a map between topological spaces, for which a good reference(Lang, 2012).

ii. Given two atlases of class $\mathcal{C}^k$, they are *equivalent* if their union is still an atlas of class $\mathcal{C}^k$ and it is the equivalent class of atlases of class $\mathcal{C}^k$ that defines a $\mathcal{C}^k-$manifold on M (or the maximal family, see books on differential geometry for these technicalities).

iii. (Carmo, 1992) gives an equivalent definition of differentiable manifold, which is more intuitive when we look at parametric models. In this reference a *parametrisation* of M, a set, at a point $p \in M$ is a pair (U_α, x_α), where $p \in U_\alpha \subset \mathbb{R}^n$, U_α is an open set in $\mathbb{R}^n$ and $x_\alpha :\to M$ is an injective map. Then, as before we form an atlas and this provides the differentiable structure to the set M.

iv. We assume here that everyone has some familiarity with the fundamentals of differential geometry, so we do not make examples. For a more thorough introduction on differential geometry, see(Lang, 2012; Sernesi, 1994).

Given a chart at $p \in M$, i.e. $U \ni p$ and a $\varphi : U \in \mathbb{R}^n$, this is determined by its n component functions $\{\xi^i : U \to \mathbb{R}\}_{i=1}^n$, such that $\varphi(p) = (\xi^1(p), \dots, \xi^n(p))$. These are called the n local coordinates on U defined by the chart φ. Given two local charts at $p \in U \cap V \subset M$, (U, φ), (V, ψ), with coordinate systems $[\xi^i]$, $[\rho^i]$ respectively, the compositions $\psi \circ \varphi^{-1}$ and $\varphi \circ \psi^{-1}$ are the change of coordinates maps.

Let us look at an example, which will play an important role in understanding *affine connections.*

Example 1.1 (Affine manifold)**.** A real affine space of dimension n $\mathbb{A}^n$ is a triplet $(\mathbb{A}^n, V, \vec{\ })$, where $\mathbb{A}^n$ is the set of points, V is an $n-$dimensional vector space over $\mathbb{R}$–called the space of translations– and $\vec{\ }$is a map from $\mathbb{A}^n \times \mathbb{A}^n$ to V satisfying the following properties:

(i) for each fixed $p \in \mathbb{A}^n$ and vector $v \in V$ there exists a unique $q \in \mathbb{A}^n$ such that $\overrightarrow{pq} = v$
(ii) $\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$.

Each affine space is a connected and path-connected topological manifold with a natural $\mathcal{C}^\infty$ differential structure. For each point $O \in \mathbb{A}^n$ (the origin) and vector basis $\{e_i\}_{i=1}^n \subset V$ we can consider the map $f : \mathbb{A}^n \to \mathbb{R}^n$ which takes a point $p \in \mathbb{A}^n$ into the n coordinates of $\overrightarrow{Op}$ w.r.t. the basis $\{e_i\}_{i=1}^n \subset V$, which is a bijection. Furthermore the Euclidean topology on $\mathbb{R}^n$ induces a topology on $\mathbb{A}^n$, which does not depend on the choice of the origin and basis. f defines a global chart on $\mathbb{A}^n$–called the Cartesian coordinate system with origin $O \in \mathbb{A}^n$ and axes $\{e_i\}_{i=1}^n \subset V$–and each mapping f defines a smooth atlas on the affine space. Given two of these maps f, g which are determined by different origins and bases in V, $g \circ f^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ and $f \circ g^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ are linear and non-homogeneous coordinate transformations and are hence smooth.

Example 1.2 (Example: Projective manifold (Carmo, 1992))**.** The real projective space $P^n(\mathbb{R})$ is the set of straight lines of $\mathbb{R}^{n+1}$ passing through the origin $0 \in \mathbb{R}^{n+1}$—the set of linear subspaces of dimension 1 of $\mathbb{R}^{n+1}$ (Sernesi, 1994).

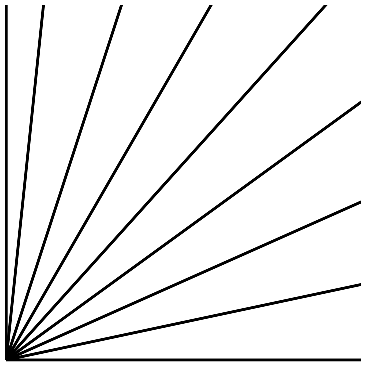

Figure 1.1: Some points in the projective plane.

For each $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$, the equivalence class $[x_1, \dots, x_{n+1}] = [\lambda(x_1, \dots, x_{n+1}) : \lambda \in \mathbb{R}]$ defines a point of the projective space. Consider the subsets (coordinate neighbourhoods) $V_i = \{[x_1, \dots, x_{n+1}] : x_i \neq 0\}$, $i = 1, \dots, n+1$, of directions not belonging to the $i-$th coordinate hyperplane. In V_i we have that $[x_1, \dots, x_{n+1}] = [\frac{x_1}{x_i}, \dots, 1, \dots, \frac{x_{n+1}}{x_i}]$, so we can define the mapping $\mathbf{x}_i : \mathbb{R}^n \to V_i$ by $\mathbf{x}_i(y_1, \dots, y_n) = [y_1, \dots, y_{i-1}, 1, y_i, \dots, y_n]$. The family $\{(R^n, \mathbf{x}_i)\}_{i=1,\dots,n+1}$ defines a differentiable structure on $P^n(\mathbb{R})$. The coordinates in each V_i are

$$\left(\frac{x_1}{x_i}, \dots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \dots, \frac{x_{n+1}}{x_i}\right)$$

and are called *inhomogeneous coordinates* corresponding to the homogeneous ones $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1}$. See (Carmo, 1992) for more details.

Now, $P^n(\mathbb{R})$ can also be seen as the quotient space of the unit sphere $S^n = \{p \in \mathbb{R}^{n+1} : \|p\| = 1\}$ by the equivalence relation $p \sim -p$ (identification of antipodal points). We can hence introduce another differentiable structure on the n-dimensional projective space, taking advantage from the parametrisation of S^n and then projecting it to $P^n(\mathbb{R})$ through the canonical projection $\pi : S^n \to P^n(\mathbb{R})$, see (Carmo, 1992) for details.

Let us now introduce the concept of differentiability of functions on a manifold.

A continuous map $f : M \to N$ between two differentiable manifolds of dimension n and m resp. is *smooth* (or also *differentiable*, or a *morphism*) at $p \in M$ if $\psi \circ f \circ \varphi^{-1} : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable for all charts (U, φ), (V, ψ) such that $p \in U$ and $f(p) \in V$. We indicate by $D(M|N)$ the class of smooth functions between M and N, or just by $D(M)$, when $N = \mathbb{R}$.

1.1.1 Tangent spaces and differentials

Let us begin with derivations and differentiations in $\mathbb{R}^n$.

The directional derivative of a function $f : \mathbb{R}^n \supseteq A \to \mathbb{R}^n$, with A open, at $x_0 \in A$, in direction $v \in \mathbb{R}^n$ such that $\|v\| = 1$, is defined, if the limit exists and is finite, as

$$D_v f(x_0) := \lim_{t\to 0} \frac{f(x_0 + tv)}{t}$$

In the same way we can define the partial derivatives w.r.t. the $i-$th coordinate, as $\frac{\partial}{\partial x_i} f(x_0) := D_{e_i} f(x_0)$ and the gradient in of f is then

$$\nabla f(x) = \left(\frac{\partial}{\partial x_1} f(x), \dots, \frac{\partial}{\partial x_n} f(x) \right)$$

Recall the difference between *derivabilità [IT]* at a point and *differenziabilità [IT]* (differentiability).

With this identification of vectors with (directional) derivatives in mind, let us define the tangent spaces of a manifold M. Here, the charts allow us to carry over ideas of the "usual" differential calculus in the Euclidean space to our manifold.

Definition 1.1 (Derivations)**.** Given a smooth manifold M, a *derivation* in $p \in M$ is a $\mathbb{R}-$linear map $D_p : D(M) \to \mathbb{R}$ such that for all $f, g \in D(M)$

$$D_p fg = f(p) D_p g + g(p) D_p f.$$

With the following linear structure

$$\left(aD_p + bD'_p \right) f := aD_p f + bD'_p f \quad \forall a, b \in \mathbb{R},\ \forall f \in D(M)$$

the set $\mathcal{D}_p(M)$ of of all derivations at p becomes an $\mathbb{R}-$vector space.

We first observe that the space of derivations is not empty. Given a chart (U, φ) at p with coordinates $[\xi^i]$ the operators

$$\left.\frac{\partial}{\partial \xi^i}\right|_p : f \mapsto \left.\frac{\partial f \circ \varphi^{-1}}{\partial \xi^i}\right|_{\varphi(p)}$$

are derivations. The subspace of $\mathcal{D}_p(M)$ spanned by $\frac{\partial}{\partial \xi^i}$ has the same dimension as M and does not depend on the choice of the chart at p. Let $[\rho^i]$ be another local coordinate system at p defined by the chart (ψ, V), then we have[2]

$$\left.\frac{\partial}{\partial \rho^k}\right|_p = \left.\frac{\partial \xi^r}{\partial \rho^k}\right|_{\psi(p)} \left.\frac{\partial}{\partial \xi^r}\right|_p \tag{1.1}$$

where the terms $\left.\frac{\partial \xi^r}{\partial \rho^k}\right|_{\psi(p)}$ are the coefficients of the Jacobian J of the change of coordinates transformation, which is non singular. By definition, indeed, we have that $\left.\frac{\partial \xi^r}{\partial \xi^s}\right|_p = \delta^r_s$ and we can compose the maps as follows $\varphi \circ \psi^{-1} \circ \psi \circ \varphi^{-1}$ which is the identity on $\varphi(U \cap V)$, so that $\delta^r_s = \left.\frac{\partial \xi^r}{\partial \xi^s}\right|_p = \left.\frac{\partial \xi^r}{\partial \rho^k}\right|_{\psi(p)} \left.\frac{\partial \rho^k}{\partial \xi^s}\right|_{\varphi(p)}$, i.e. the matrix J is invertible, hence non singular. Therefore the spaces spanned by $\left.\frac{\partial}{\partial \xi^i}\right|_p$ and $\left.\frac{\partial}{\partial \rho^k}\right|_p$ coincide. It remains to prove that the dimension of the span of the n derivations is n, i.e. that the n derivations are linearly independent. But we refer to any book on differential geometry for this.

Definition 1.2 (Tangent space)**.** The tangent space of M at $p \in M$ is indicated by T_pM and is the subspace of $\mathcal{D}_p(M)$ spanned by the n derivations $\frac{\partial}{\partial \xi^i}$. It has dimension n and does not depend on the choice of the chart at p.

It can be proved that the space of all derivations on M at p coincides with T_pM, i.e. $\left\{\frac{\partial}{\partial \xi^i}\right\}_i$ is a basis for $\mathcal{D}_p(M)$. T_pM has the same dimension of the manifold.

The motivation for calling these spaces, *tangent spaces* comes from the following equivalent definition (Carmo, 1992)

Definition 1.3 (Tangent vector to a curve)**.** Let $\gamma : (-\epsilon, \epsilon) \to M$ be a differentiable curve on M such that $\gamma(0) = p \in M$. The tangent vector to γ at $t = 0$ is the linear map $\gamma'(0) : D(M) \to \mathbb{R}$ defined as follows

$$\gamma'(0)f := \left.\frac{d}{dt}(f \circ \gamma)\right|_{t=0}$$

for each $f \in D(M)$.

A tangent vector at $p \in M$ is then the tangent vector at $t = 0$ to some curve γ satisfying $\gamma(0) = p$.

[2] Einstein summation is use throughout these notes.

The equivalence between the two definitions of tangent spaces can be seen choosing a parametrisation $(U, \mathbf{x})$ at $\mathbf{x}(0) = p \in M$ and expressing both the function f and the curve γ in the parametrisation:

$$\begin{aligned} f \circ \mathbf{x} : &\quad (x^1, \dots, x^n) \mapsto f(x^1, \dots, x^n) \\ \mathbf{x}^{-1} \circ \gamma : &\quad t \mapsto (x^1(t), \dots, x^n(t)) \,. \end{aligned}$$

Then, $f \circ \gamma = f \circ \mathbf{x} \circ \mathbf{x}^{-1} \circ \gamma = f\left(x^1(t), \dots, x^n(t)\right)$ and

$$\gamma'(0) f = \frac{d}{dt} f\left(x^1(t), \dots, x^n(t)\right)\Big|_{t=0} = \dot{x}^i(0) \left.\frac{\partial}{\partial x^i}\right|_0 f$$

and again we find the vectors $\left.\frac{\partial}{\partial x^i}\right|_o$, which are here the tangent vectors at p of the coordinate curves $x^i \to \mathbf{x}(0, \dots, 0, x^i, 0, \dots, 0)$.

Let us go back to the affine manifold and consider its tangent space at $p \in \mathbb{A}^n$, $T_p\mathbb{A}^n$. It turns out that there is a natural *isomorphism* between $T_p\mathbb{A}^n$ and V.

Definition 1.4 (Cotangent space)**.** The dual space T_p^*M is called cotangent space of M at p and its elements are called 1-forms in p, or covectors, or covariant vectors. If $[\xi^i]$ are local coordinates at p, the dual basis to $\left\{ \left.\frac{\partial}{\partial \xi^i}\right|_p \right\}$ is denoted by $\{d\xi^i|_p\}$ and it holds $d\xi^i\left(\frac{\partial}{\partial \xi^j}\right) = \delta^i_j$.

Remark. Recall the duality between measures and functions [ADD CITATION]. Forms can be seen as (signed) measures.

Definition 1.5 (Tangent and cotangent bundles)**.**

$$\begin{aligned} TU &:= \left\{ (p, v) \mid p \in U, \quad v \in T_pM \right\}, \\ T^*U &:= \left\{ (p, \omega) \mid p \in U, \omega \in T_p^*M \right\} \end{aligned}$$

Definition 1.6 (Differential of a mapping or push forward)**.** Let M, N be two smooth manifolds and $f : M \to N$ a smooth function. The differential of f at $p \in M$ or push forward of f at p is the linear mapping

$$\begin{aligned} df_p : T_pM &\to T_{f(p)}N \\ X_p &\mapsto df X_p \end{aligned} \tag{1.2}$$

defined by $dfX_p(g) := X_p(g \circ f)$ for all vectors $X_p \in T_pM$ and all smooth functions $g \in D(N)$.

Observe that $g \circ f \in D(M|\mathbb{R})$.

1.1.2 Vector and tensor fields

A vector field is a mapping $X : p \mapsto X_p \in T_pM$, which associates to each point p in the manifold M a tangent vector. We indicate by $\mathfrak{X}(M)$ the set of all vector fields on M. Observe that this set is not empty, for instance, the $n-$mappings defined by $\frac{\partial}{\partial \xi^i} : p \to \left.\frac{\partial}{\partial \xi^i}\right|_p$ are vector fields, formed by the natural basis given by the coordinate system $[\xi^i]$. Each vector field X may be written as $X_p = X^i_p \partial_i|_p$, where $\partial_i := \frac{\partial}{\partial \xi^i}$ and X^i_p, for $i = 1, \dots, n$, are the scalar components of X w.r.t. the coordinate system $[\xi^i]$. The change of coordinates, here, is the same as in (Equation 1.1).

If the components of the vector field are C^∞ w.r.t. some coordinate system, then they are smooth w.r.t. any coordinate system, and X is then called a *smooth vector field.* With the following structure

$$X + Y : p \mapsto X_p + Y_p \qquad cX : p \mapsto cX_p$$

the set $\mathfrak{X}(M)$ becomes a linear space. More generally, a mapping $t : M \to \mathcal{A}_\mathbb{R}(T_pM)$ which associates to a point $M \ni p$ a tensor t_p in the tensor algebra generated by T_pM, T^*_pM, and $\mathbb{R}$, is said to by a tensor field.

[TBD] Recap on: - multi-linear maps - tensor products of vector spaces - tensor algebra generated by a vector space, its dual space, and its field.

In the framework of tensors, each $\left.\frac{\partial}{\partial \xi^i}\right|_p$ is a tensor of T_pM, i.e. a contravariant vector. Tensors of the dual tangent space T^*_pM are covariant vectors. The canonical bases of $\mathcal{A}_\mathbb{R}(T_pM)$ are given by tensor products of $\left\{ \left.\frac{\partial}{\partial \xi^i}\right|_p \right\}$ and $\{d\xi^i|_p\}$.

Assigning a smooth tensor field T on M is equivalent to assign a set of smooth functions which map

$$(\xi^1, \dots, \xi^n) \mapsto T^{i_1 \cdots i^m}{}_{j_1 \cdots j_k}(\xi^1, \dots, \xi^n)$$

in every local coordinate system of M such that they satisfy the rules of transformation of the components of a tensor, i.e.

$$T^{i_1 \cdots i_m}{}_{j_1 \dots j_k}, (\xi^1, \dots, \xi^n) = \left.\frac{\partial \xi^{i_1}}{\partial \rho^{k_1}}\right|_p \cdots \left.\frac{\partial \xi^{i_m}}{\partial \rho^{k_m}}\right|_p \left.\frac{\partial \rho^{l_1}}{\partial \xi^{j_1}}\right|_p \cdots \left.\frac{\partial \rho^{l_m}}{\partial \xi^{j_m}}\right|_p T'^{k_1 \cdots k_m}{}_{l_1 \dots l_k}(\rho^1, \cdots, \rho^n)$$

Remark. Each vector field $X \in \mathfrak{X}(M)$ defines a derivation at each point $p \in M$: take any differentiable $f \in D(M)$ then $X_p(f) := X^i(p) \left.\frac{\partial f}{\partial \xi^i}\right|_p$. In general, every smooth vector field X defines a linear mapping from $D(M)$ to $D(M)$ by $f \mapsto X(f)$, where $X(f)(p) =: X_p(f)$ for every $p \in M$.

The differential of $f \in D(M)$ at p is the 1-form defined, in local coordinates, by

$$df_p = \left.\frac{\partial f}{\partial \xi^i}\right|_p d\xi^i|_p.$$

Varying $p \mapsto df_p$ we have defined a smooth covariant vector field df, called the differential of f (note the absence of "at p").

Definition 1.7 (Vector field along a curve)**.** A vector field X along a differentiable curve $\gamma : I \to M$ from $I \subset \mathbb{R}$ open to the manifold is a differentiable mapping $t \mapsto X(t) \in T_{\gamma(t)}M$. *Differentiable* here means that for any $f \in D(M)$, the function $t \mapsto X(t)f$ is a differentiable function on I.

In the special case $X(t) = Y_{\gamma(t)}$ for some vector field Y, we say that $X(t)$ is induced by Y.

A particularly important tensor of covariant degree 2, i.e. a tensor in $[T_pM]_2^0$ is the Riemannian metric tensor, which we are introduce in the following section.

1.1.3 Riemannian manifolds

Assume that, for each $p \in M$, an inner product $\langle \ , \ \rangle_p$ is defined on T_pM.
The mapping $g : p \mapsto \langle \ , \ \rangle_p \in [T_pM]_2^0$ is a covariant tensor field of order 2. Equivalently, assume we have a smooth covariant tensor field g on M of degree 2, determining a symmetric, positive definite quadratic form $g_p : T_pM \times T_pM \to \mathbb{R}$.
g is called a **Riemannian metric** on M and (M, g) is then called **Riemannian manifold**. Observe that this metric is, in general, not unique and it is not naturally determined by the structure of M as a manifold.

Example 1.3. (1) The canonical metric in the Euclidean space $\mathbb{R}^n$ is $g_{ij} = \langle e_i, e_j \rangle = \delta_{ij}$ so that the matrix representation of the metric is the identity and, in Cartesian coordinates, $g = \delta_{ij} dx^i dx^j = \sum_{i=1}^n \left(dx^i\right)^2$.

(2) The product metric. Let M_1, M_2 be Riemannian manifolds and consider their product $M_1 \times M_2$ with the natural projections $\pi_i : M_1 \times M_2 \to M_i$ for $i = 1, 2$. Then a Rimannian metric on $M_1 \times M_2$ can be introduced by

$$\langle u, v \rangle_{(p,q)} = \langle d\pi_1(u), d\pi_1(v) \rangle_p + \langle d\pi_2(u), d\pi_2(v) \rangle_q \,,$$

where we use the differentials of the projections to push forward the derivations/vectors in the tangent space $T_{(p,q)}M_1 \times M_2$ to M_1 and M_2 accordingly.

We assume the existence of a Riemannian metric on M, but the following can be proved.

Theorem 1.1. *If M is a connected, smooth manifold, it is possible to define a Riemannian metric g on M.*

Proof. To proof this result one needs to introduce a smooth partition of the unity. We refer the interested reader to (Lang, 2012; Sernesi, 1994).

□

Given a coordinate system $[\xi^i]$ at p, using our usual notation $\partial_i := \frac{\partial}{\partial \xi^i}$ (more precisely, we should write $\partial_i|_p$ but it should be obvious from the context), we can see that the components (also called *local representation of the Riemannian metric* in the chart) g_{ij}, for $i, j = 1, \dots, n$, of g at p, are determined by

$$g_{ij}(p) = \langle \partial_i, \partial_j \rangle_p, \tag{1.3}$$

so that :

- the tensor at p is written as $g(p) = g_{ij}(p) d^i|_p \otimes d^j|_p$, where $\{d^i|_p\} = \{d\xi^i|_p\}$, for $i = 1, \dots, n$ is the dual basis of $\{\partial_i\}$ in the cotangent space $T_p^* M$;
- the scalar product between two tangent vecctor at p is $\langle v, w \rangle_p = g_{ij}(p) v^i w^j$, for any two vectors $v = v^i \partial_i|_p, w = w^i \partial_i|_p \in T_p M$;
- and the norm of any $v^i \partial_i|_p = v \in T_p M$ is given by $\|v\|_p^2 = g_{ij}(p) v^i v^j$.

Furthermore, we can define the **length of a (piecewise) smooth curve** $\gamma : I \ni t \mapsto \gamma(t) \in M$, where $I \subset \mathbb{R}$ is a bounded interval, as

$$L_g(\gamma) = \int_I \sqrt{|g(\gamma'(t), \gamma'(t))|} dt$$

and its **energy**

$$E_g(\gamma) = \int_I |g(\gamma'(t), \gamma'(t))| \, dt.$$

Remark. $L_g(\gamma)$ is re-parametrisation invariant.

Given the length of a curve, we can define a distance function in (M, g) so that (M, d_g) is a metric space, in the following way:

$$d_g(p, q) := \inf \left\{ L_{\mathbf{g}}(\gamma) \mid \gamma : [a, b] \to M, \gamma \text{ piecewise smooth}, \gamma(a) = p, \gamma(b) = q \right\}. \tag{1.4}$$

A curve γ achieving the minimum in (Equation 1.4) is called *geodesic*, although the proper definition of a geodesic, i.e. of the generalisation of a straight line, is that of a curve which does change direction, i.e. with constant velocity or, also zero acceleration.

Now, we can ask: how does a change of basis modify the metric tensor? Suppose we are given another coordinate system $[\rho^i]$ at p and let us define $\tilde{\partial}_k = \frac{\partial}{\partial \rho^k}$, then, simply recalling (Equation 1.1), we have:

$$\left\langle \tilde{\partial}_k, \tilde{\partial}_\ell \right\rangle = \tilde{g}_{k\ell} = g_{ij} \left(\frac{\partial \xi^i}{\partial \rho^k} \right) \left(\frac{\partial \xi^j}{\partial \rho^\ell} \right)$$

and

$$g_{ij} = \tilde{g}_{k\ell} \left(\frac{\partial \rho^k}{\partial \xi^i} \right) \left(\frac{\partial \rho^\ell}{\partial \xi^j} \right),$$

(observe that there is a dependence on p everywhere in the previous formulas, but we will often “forget” to write it explicitly).

The coefficients $g_{ij}(p)$ form a square matrix $G(p)$, which is symmetric and positive definite, so, its inverse $G(p)^{-1}$ exists. Let $g^{ij}(p)$ be its ij—th element, then

$$g_{ij} g^{jk} = \delta_i^k = \begin{cases} 1 & (k = i) \\ 0 & (k \neq i) \end{cases}$$

from which we can also obtain the change-of-coordinates relations (as exercise).

On a Riemannian manifold we can also define the *gradient* of a differentiable f, denoted here by $\operatorname{grad} f$, as the vector field satisfying

$$g(v, \operatorname{grad} f) = df(v) \tag{1.5}$$

for all $v \in TM$.

Those who are familiar with exterior calculus will notice that the metric one to transform a 1-form, the differential, into a 1-vector, the gradient.

1.1.4 Affine connections and covariant derivatives

In this section our goal is to compare tangent spaces $T_p(M)$ and $T_q(M)$, and the respective vectors, when $p \neq q \in M$ or, in general, to compare vector field $X, Y \in \mathfrak{X}(M)$ by giving a meaning to the derivative $\nabla_X Y$ of a vector field X w.r.t. the vector field Y.

Let us start with our an affine manifold $\mathbb{A}^n$.

Example 1.4. Let $(\mathbb{A}^n, V, \vec{\cdot})$ be an affine manifold, as in Exm. Example 1.1. Once we fix an origin $0 \in \mathbb{A}^n$ and a basis $\{e_i\}$ for the vector space of translations V, we have a global smooth chart from $\mathbb{A}^n$ to $\mathbb{R}^n$. A change of chart corresponds to a change of origin and, given two charts at p their Cartesian coordinates transform as

$$\tilde{x}^i = A^i_j x^i + B^i,$$

with A, a non-singular matrix, that does **not** depend on p.
The same applies to the components of a vector field w.r.t. two Cartesian coordinate systems at p, that is

$$X = X'^i \frac{\partial}{\partial \rho^i} = A^i_j X^j \frac{\partial}{\partial \xi^i}.$$

At each $p \in \mathbb{A}^n$ we have the tangent space $T_p\mathbb{A}^n$. It is easy to show that these tangent spaces are isomorphic to V through the natural isomorphism

$$\chi_p : T_p\mathbb{A}^n \to V; \quad v = v^i \frac{\partial}{\partial \xi^i}\bigg|_p \mapsto v^i e_i.$$

Then, $\chi_q^{-1} \circ \chi_p : T_p\mathbb{A}^n \to V \to T_q\mathbb{A}^n$ is a well-defined vector space isomorphism. Let us denote it by $\mathcal{A}^q_p$. Take, now, a point $q \in \mathbb{A}^n$ such that $q = p + hX_p$ for some $X_p \in T_p\mathbb{A}^n$, and consider a vector field $Y \in \mathfrak{X}(M)$, then the following is well-defined

$$\lim_{h \to 0} \frac{\mathcal{A}^p_q Y_{p+hX_p} - Y_p}{h} =: (\nabla_X Y)_p$$

Thanks to the affine structure we can consider the variation of the vector field Y in the direction given by the vector field X in a neighbourhood of p. Issues arise when we lack this affine-space structure.

Intuitively, we want to consider $\frac{dX}{dt}(t)$, the rate of change of a vector field $X(t) \in T_{\gamma(t)}M$ along a curve γ on M, see def. Definition 1.7. Unfortunately, $\frac{dX}{dt}(t) \notin T_{\gamma(t)}M$ in general. Hence, instead of considering $\frac{dX}{dt}(t)$, we will take the orthogonal projection of $\frac{dX}{dt}(t)$ to $T_{\gamma(t)}M$, which we indicate by $\frac{DX}{dt}(t)$ following the notation of (Carmo, 1992). $\frac{DX}{dt}(t)$ is called the covariant derivative of X. Taking the covariant derivative of the velocity of a curve γ, allows us to evaluate the acceleration of the curve and show that the curves with zero acceleration are the geodesics of M. A vector field X along γ is said to be **parallel** if $\frac{DX}{dt}(t) = 0$ for all $t \in I$. And, vice versa, we can start from the notion of parallelism and recover the one of covariant derivative.

Definition 1.8 (Affine connection and covariant derivative)**.** Let M be a differentiable manifold. An affine connection or covariant derivative operator ∇, is a map

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \ni (X, Y) \mapsto \nabla_X Y \in \mathfrak{X}(M)$$

which satisfies the following properties for every $p \in M$

i. $\left(\nabla_{fY+gZ} X\right)_p = f(p)\left(\nabla_Y X\right)_p + g(p)\left(\nabla_Z X\right)_p$ for all $f, g \in D(M)$ and vector fields $X, Y, Z \in \mathfrak{X}(M)$;
ii. $\left(\nabla_Y fX\right)_p = Y_p(f)X_p + f(p)\left(\nabla_Y X\right)_p$ for all $X, Y \in \mathfrak{X}(M)$ and $f \in D(M)$;

iii. $\left(\nabla_Y(aX+bZ)\right)_p = a(\nabla_Y X)_p + b(\nabla_Y Z)_p$ for all scalars $a, b \in \mathbb{R}$ and $X, Y, Z \in \mathfrak{X}(M)$.

The contravariant vector field $\nabla_Y X$ is called the covariant derivative vector of X with respect to Y and the affine connection ∇.

Firstly, observe that $Y_p(f)$ indicates the directional derivative of a differentiable (real-valued) function f in the direction of the vector field Y, in $p \in M$. $Y_p(f) = df_p(Y) = df(Y_p)$, where $df_p : T_pM \to \mathbb{R}$ is the differential of f at p, see Definition 1.6.

Remark. The properties listed in the definition are **pointwise** properties. If two vector fields X and X' have the same value at p, i.e. $X_p = X'_p$ then $(\nabla_X Z)_p = (\nabla_{X'} Z)_p$. Similarly if $Y = Y'$ in a neighbourhood of p, then $(\nabla_X Y)_p = (\nabla_X Y')_p$.
In other words, the notion of affine connection is a local notion.

With the notation introduced by a connection ∇ we can write $(\nabla_Y f)_p = Y_p(f) = df_p(Y)$ for the usual *directional derivative* of $f \in D(M)$ at p in the direction given by the vector field $Y \in \mathfrak{X}(M)$.

Let us see a proposition (Carmo, 1992, prop. 2.2), without proof.

Proposition 1.1. *Let M be a manifold with a connection ∇. There exists a unique correspondence which associates to a vector field X along a differentiable curve $\gamma : I \to M$ another vector field $\frac{DX}{dt}(t)$ along γ, called the covariant derivative of X along γ, such that:*

i. $\frac{D(X+Y)}{dt} = \frac{DX}{dt} + \frac{DY}{dt}$
ii. $\frac{D}{dt}(fX) = \frac{df}{dt}X + f\frac{DX}{dt}$ *for a differentiable function f on I*
iii. if X is induced by a vector field $Y \in \mathfrak{X}(M)$, i.e. $X(t) = Y(\gamma(t))$, then $\frac{DX}{dt} = \nabla_{\gamma'(t)}Y$.

We do not look at the whole proof, but for the uniqueness part, which is useful for the understanding of the covariant derivative.

Let $X(t) = X^i(t)\partial_i|_{\gamma(t)}$ be a local expression of X along the curve γ, where we composed both the coordinate functions and the components functions with the local parametrisation $\mathbf{x}$. By Proposition 1.1-i. and ii. we get

$$\frac{DX}{dt} = \frac{dX^i(t)}{dt}\partial_i|_{\gamma(t)} + X^i(t)\frac{d\partial_i|_{\gamma(t)}}{dt}.$$

By iii.

$$\frac{d\partial_i|_{\gamma(t)}}{dt} = \nabla_{\gamma'(t)}\partial_i = \nabla_{\frac{dx^j(t)}{dt}\partial_j}\partial_i = \frac{dx^j(t)}{dt}\nabla_{\partial_j}\partial_i.$$

where, in the second equality, we used the definition of the tangent vector to a curve Definition 1.3, and, in the third, i. of Definition 1.8. Putting together the two equations:

$$\frac{DX}{dt} = \frac{dX^i(t)}{dt}\partial_i + X^i(t)\frac{dx^j(t)}{dt}\nabla_{\partial_j}\partial_i.$$

This proves the uniqueness of the correspondence, if it exists.

Connection coefficients

Let us consider, as usual, a local chart (U, ϕ) at $p \in U \subset M$ with coordinate $[\xi^i]$ for $i = 1, \dots, n$ and two vector fields $X, Y \in \mathfrak{X}(M)$, which we decompose w.r.t. $\partial_i|_p$. Then

$$\begin{aligned} \left(\nabla_X Y\right)_p &= X^i(p)Y^j(p)\nabla_{\partial_i|_p}\partial_j + X^i(p)\partial_i|_p Y^j \partial_j|_p \\ \text{setting: } \nabla_{\partial_i|_p}\partial_j &= \left\langle \nabla_{\partial_i}\partial_j, d^k\right\rangle \partial_k|_p := \Gamma^k_{ij}(p)\partial_k|_p \\ \left(\nabla_X Y\right)_p &= X^i_p\left(\partial_i|_p Y^k + \Gamma^k_{ij}(p)Y^j_p\right)\partial_k|_p, \end{aligned} \tag{1.6}$$

where $\partial_i Y^j|_p = \frac{\partial Y^j}{\partial \xi^i}\Big|_p$.

For fixed $X \in \mathfrak{X}(M)$ and $p \in M$, the linear map $Y_p \mapsto (\nabla_{Y_p} X)_p$ (and a known result which guarantees that, for $p \in M$, if $t \in \mathscr{A}_{\mathbb{R}}(T_p M)$ then there exists a differentiable tensor field Ξ in M such that $\Xi_p = t$ defines a tensor $(\nabla X)_p \in T^*_p M \otimes T_p M$ such that the only possible contraction of Y_p and $(\nabla X)_p$ is $(\nabla_{Y_p} X)_p$. Varying $M \ni p \mapsto (\nabla X)_p$ defines a smooth (1, 1) tensor field ∇X, which in local coordinates reads

$$\partial_i X^k + \Gamma^k_{ij} X^j$$

and is called **covariant derivative tensor** of X w.r.t. the affine connection ∇.

It can be proved that assigning an affine connection on a manifold M of dimension n is completely equivalent to giving the n^3 coefficients $\Gamma^k_{ij}(p)$ in each local coordinate system, as smooth functions w.r.t. p and transform according to the appropriate transformation rules.

Example 1.5. Let us consider again the affine manifold $\mathbb{A}^n$ and take an affine connection ∇ on it. It is easy to see that $\Gamma^k_{ij}(p) = 0$ for each i, j, k and p in every Cartesian coordinate system, that $\left(\nabla_X Y\right)_p$ is a contravariant vector field, and that the components $\left(\nabla_X Y\right)_p = X^i(p)\partial_i|_p Y^j$ are the usual directional derivatives in $\mathbb{R}^n$ and change according to the transformation rule $\left(\nabla_X Y\right)^{'j}_p = A^j_k \left(\nabla_X Y\right)^k_p$.

Hence, the affine structure $\mathbb{A}^n$ provides automatically an affine connection, simply by the standard derivative in every fixed Cartesian coordinate system.

Now that we have defined the concepts of *affine connection* and *derivation of vectors fields along a curve*, let us introduce the *parallel transport.*

Definition 1.9 (Parallelism)**.** Let M be a differentiable manifold with an affine connection ∇. A vector field X along $\gamma : I \to M$ is called *parallel* if $\frac{DX}{dt} \equiv 0$ on I.

Proposition 1.2 (Parallel transport)**.** *Let M be a differentiable manifold with an affine connection ∇. Let γ be a differentiable curve in M and X_0 a vector tangent to M at $\gamma(t_0)$. Then, there exists a unique parallel vector field X along γ such that $X(t_0) = X_0$. $X(t)$ is called the parallel transport of X_0 along γ.*

For the detailed proof, see (Carmo, 1992, prop. 2.6). Here, let us just suppose that a such a vector field $X(t)$ exists, then

$$\begin{aligned} 0 = \frac{DX}{dt} &= \frac{dX^i(t)}{dt}\partial_i + X^i(t)\frac{dx^j(t)}{dt}\nabla_{\partial_j}\partial_i \\ &= \left(\frac{dX^k(t)}{dt} + X^i(t)\frac{dx^j(t)}{dt}\Gamma^k_{ij}\right)\partial_k \end{aligned}$$

so, we have a system of n differential equations of the first order in $X^k(t)$. Then, for any initial condition, i.e. vector $X(t_0)$, we have the existence of a unique parallel vector field $X(t)$ with the desired characteristics, which is defined on the whole interval I and is smooth. Moreover the map which assigns to each initial vector $X(t_0)$ the parallel vector field X is linear.

The linear map $\Pi(\gamma)^t_{t_0} : T_{\gamma(t_0)}M \to T_{\gamma(t)}M$ defined by $\Pi(\gamma)^t_{t_0}(X(t_0)) = X(t)$ is called the *parallel transportation* along γ. $\Pi(\gamma)^t_{t_0}$ is an isomorphism.

The notion of covariant derivative can be extended to tensor fields, defining $\nabla_X u$ for vector field X and a smooth tensor field u.

The coefficient $T^i_{jk} := \Gamma^i_{jk} - \Gamma^i_{kj}$ define the components of a tensor field, called the **torsion tensor field of the connection**

$$T = \left(\Gamma^i_{jk} - \Gamma^i_{kj}\right)\partial_i \otimes d^j \otimes d^k.$$

The torsion tensor at p is then, a bilinear map from $\mathfrak{X}(M) \times \mathfrak{X}(M)$ to a smooth vector field (same as for ∇), defined as

$$T_p(\nabla)\left(X_p, Y_p\right) = \nabla_{X_p}Y - \nabla_{Y_p}X - [X,Y]_p$$

If the tensor field T vanishes on M for every $X, Y \in \mathfrak{X}(M)$, i.e. $[X, Y] = \nabla_X Y - \nabla_Y X$ or, in terms of the connection coefficients $\Gamma^k_{ij} = \Gamma^k_{ji}$, then ∇ is said to be **torsion free**, or **symmetric**.

Given $X, Y \in \mathfrak{X}(M)$, the term $[X, Y]_p$ is called *bracket* (or Lie bracket) and it is defined as the unique contravariant smooth vector field Z such that $Zf = (XY - YX)f = X(Y(f)) - Y(X(f))$ for each $f \in D(M)$. The bracket exists and is unique, see e.g. [Lemma 5.2; Carmo (1992)]. Or [Gallot, Hulin, Lafontaine].

Given a Riemannian manifold (M, g), there is a preferred—exactly one (Petersen, 2006, p. Thm.1 (The fundamental theorem of Riemannian geometry))—affine connection ∇, which is torsion free and is completely determined by the metric, i.e. $\nabla g = 0$. Let us see what this means in other terms.

Definition 1.10 (Riemannian connection)**.** An affine connection ∇ on a manifold Riemannian manifold (M, g) is compatible with the metric when, for any smooth curve γ and any pair of parallel vector fields X, X' along the curve, we have $g(X, X') =$ constant.

In terms of the covariant derivative along a curve, ∇ is compatible with $\langle , \rangle$ if and only if

$$\frac{d}{dt}\langle V, W\rangle = \left\langle \frac{DV}{dt}, W \right\rangle + \left\langle V, \frac{DW}{dt} \right\rangle, \quad t \in I$$

i.e. the usual product rule holds. Here, we used the notation $\langle , \rangle$ for the metric to enphatise the "usual product rule".
If V, W are parallel along a curve γ then $\frac{d}{dt}\langle V, W\rangle = 0$ and we recover the definition. For the proof in the other direction, see (Carmo, 1992).

Corollary 1.1. *A connection ∇ on a Riemannian manifold M is compatible with the metric if and only if*

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$$

for all $X, Y, Z \in \mathfrak{X}(M)$.

This unique metric, torsion-free connection is the Riemannian, or Levi-Civita, connection. Its coefficients, called Christoffel's coefficients, are:

$$\Gamma^i_{jk} = \left\{ {i \atop jk} \right\} := \frac{1}{2} g^{is} \left(\frac{\partial g_{ks}}{\partial \xi^j} + \frac{\partial g_{sj}}{\partial \xi^k} - \frac{\partial g_{jk}}{\partial \xi^s} \right).$$

$\left\{ {i \atop jk} \right\}$ is called Christoffel's symbol.

1.1.5 Flatness

Let $X \in \mathfrak{X}(M)$ be a vector field on a differentiable manifold M with an affine connection ∇. If for any curve γ on M, X is parallel along γ w.r.t. ∇, then X is parallel on M. Equivalently X is parallel on M if $\nabla_Y X = 0$ for all $Y \in \mathfrak{X}(M)$.

Definition 1.11 (Affine coordinate system and flat manifold)**.** Let M be a differentiable manifold with a connection ∇ and let $[\xi^i]$ be a coordinate system of M. If the coordinate tangent vectors $\frac{\partial}{\partial \xi^i}$ are parallel on M, then $[\xi^i]$ is an *affine coordinate system* for ∇.

If an affine coordinate system for ∇ exists, then we say that ∇ is *flat*, or that M is flat with respect to the connection ∇.

It can be shown, using the transformation rules for the connection coefficients, that any affine transformation of the coordinates $[\xi^i]$ is, again, flat w.r.t. the connection ∇.

If a connection is flat, then parallel translation does not depend on the curve selected to connect the two points.

1.1.6 Geodesics

Definition 1.12 (Geodesic)**.**

2 Information geometry of statistical models

Let Ω be a set, we will now consider probability functions on Ω, i.e.

$$p : \Omega \to \mathbb{R}$$

such that $p(x) \geq 0$ for all $x \in \Omega$ and

i. $\sum_{x\in\Omega} p(x) = 1$ if Ω is a discrete set, or
ii. $\int_\Omega p(x)dx = 1$ (in this case p is a density function).

In general $(\Omega, \mathcal{B}, \nu)$ is a measurable space with $\sigma-$algebra (or Borel field) $\mathcal{B}$, ν a $\sigma-$finite measure. Given a probability measure P on Ω which is absolutely continuous w.r.t. ν, $p = \frac{dP}{d\nu} : \Omega \to \mathbb{R}$ is the Radon-Nikodym derivative of P.

Now, we have to distinguish different case:

i. If Ω is finite, then we can consider the real algebra of functions $f : \Omega \to \mathbb{R}$, which is a finite dimensional vector space over $\mathbb{R}$, denoted also by $\mathbb{R}^\Omega$, with the additional product operation (and additional axioms). Linear forms over $\mathbb{R}^\Omega$ are interpreted as signed measures (usual duality between functions and measures) and, since in $\mathbb{R}^\Omega$ there exists a preferred basis of function—$e_x(y) \in \{0, 1\}$ if x is different or equal to y, respectively—there is a natural isomorphism between the dual space of signed measures and $\mathbb{R}^\Omega$. Here we will introduce the probability simplex in section Section 3.
ii. If Ω is an infinite sample space, we will need also the $\sigma-$algebra $\mathcal{B}$ of subsets of Ω. In this case, $\mathbb{R}^\Omega$ is an infinite-dimensional functional space and we can consider either finite dimensional manifolds, through charts to/parametrisations from $\mathbb{R}^n$, or consider the infinite-dimensional non-parametric case. We will focus on the first.

Consider now a family M of probability distributions parametrised over a set of parameters $\Xi \subset \mathbb{R}^k$

$$M = \{p_\xi = p(\cdot; \xi) : \xi = [\xi^1, \dots, \xi^k] \in \Xi\}$$

where the mapping $p : \xi \mapsto p_\xi = p(\cdot; \xi)$ is injective (and we will also assume to have some regularity property). M is called an $k-$dimensional (parametric) **statistical model** on Ω. p plays the role of a parametrisation of the differentiable structure of (Carmo, 1992) and the $[\xi^i]$ define a (global) coordinate system for M. Each re-parametrisation of the model $\psi : \xi \to \psi(\xi) \subset \mathbb{R}^k$, where ψ is a smooth diffeomorphism, provides another equivalent (global) coordinate system for M, i.e. $\rho = \psi(\xi)$ and $M = \{p_{\psi^{-1}(\rho)} : \rho \in \psi(\xi)\}$. We can then consider M as a differentiable manifold, called a **statistical manifold**.

In the following we will mainly write $\xi^i(p) \in \mathbb{R}^k$ for $p \in M$, so that the ξ^is are the coordinates of the chart, instead of the parametrisation, to use a consistent notation with (Amari, 2016; Amari & Nagaoka, 2000). We will also, as it is commonly done in the field, denote by $T_\xi M$ the tangent space of M at p_ξ, and simply say "at ξ".

Observations

i. As we have defined it $p(\cdot, \xi)$ is a density function, the measure is $\mathbf{p}(\xi) = p(\cdot, \xi)\nu$ and ν does not depend on the parameter.
ii. $p : \Omega \times \Xi \to \mathbb{R}$ and we need some regularity assumptions w.r.t. x to compute, e.g., expected values. One possibility is $p_\xi \in L^1(\Omega, \nu)$ for all ξ. See (Ay et al., 2017) for more details.

Before moving on, some motivations. A typical problem in statistics is:

- given observations $x_1, \dots, x_n$ estimate the distribution generating the data p^* (true underlying distribution).
- p^* is unknown, but we often assume that it comes from a family of distributions, a model M, and the problem becomes a parameter estimation problem (assuming, possibly approximately, *faithfulness*).

Hence, it seems reasonable to focus on geometric properties of such **models**. By means of tangent spaces and their geometry (e.g. the Riemannian metric), we can study local properties of a statistical model. Affine connections, on the other hand, allow us to say something about the global geometry (e.g. curvature) of the model, since the establish a 1-to-1 affine correspondence between tangent spaces at different points. Originally, the works by Chentsov (1972), Efron (1975, 1978) and also Amari (1980, 1982) were motivated by the interest in higher-order asymptotic properties of inference. Nowadays, the affine/differential geometric approach of IG shows its usefulness in many applications, e.g. the natural gradient is well-known in optimisation and machine learning.

Assumptions on M

We assume that for each $x \in \Omega$ the parametrisation $\xi \to p(x; \xi)$ from the open (parameter) set $\Xi \subseteq \mathbb{R}^n$ to $\mathbb{R}$ is smooth (C^∞). This allows us to differentiate w.r.t. the parameter. We also assume that the order of integration and differentiation may be swapped, so that

$$\int_\Omega \frac{\partial}{\partial \xi_i} p(x; \xi)\nu(dx) = \frac{\partial}{\partial \xi_i} \int_\Omega p(x; \xi)\nu(dx) = 0 \tag{2.1}$$

Finally, we also assume that support of p, $\operatorname{supp}(p) = \{x \in \Omega : p(x; \xi) > 0\}$ does not depend on ξ. This means that, if we choose $\Omega = \operatorname{supp}(p)$, then $M \subset \mathcal{P}_>(\Omega)$, where

$$\mathcal{P}(\Omega) := \left\{ p : \Omega \to \mathbb{R} : p(x) > 0 \ \forall x \in \Omega, \int_\Omega p(x)\nu(dx) = 1 \right\}.$$

2.1 Fisher information

Recall the definition of the Fisher information matrix of a statistical model M.

Definition 2.1 (Fisher information)**.** Let $M = \{p(x;\xi) : \xi \in \Xi\}$ be an $k-$dimensional statistical model. The Fisher information matrix (FIM) of M at $\xi \in \Xi$ is the $k \times k$ matrix $I(\xi)$, whose $ij-$element is given by

$$g_{ij}(\xi) = \mathbb{E}_\xi \left[\frac{\partial}{\partial \xi^i} \ell(\xi; x) \frac{\partial}{\partial \xi^j} \ell(\xi; x) \right] = \int_\Omega \partial_i \ell(\xi; x) \partial_j \ell(\xi; x) p(x; \xi) \nu(dx) \tag{2.2}$$

where we use the notation $\partial_i = \frac{\partial}{\partial \xi^i}$ and $\ell(\xi; x)$ denotes the log-likelihood, or $\ell(\xi; x) = \log p(x; \xi)$, depending on whether we have a random variable x or a random sample $(x_1, \dots, x_n)$.

Remarks

i. It is possible to write $g_{ij}(\xi)$ in other forms:
   - $g_{ij}(\xi) = -E_\xi \left[\partial_i \partial_j \ell(\xi; x) \right]$[1]—use eq. Equation 2.1 and assume that we can exchange twice integration and derivation order to prove;
   - $g_{ij}(\xi) = 4 \int \partial_i \sqrt{p(x;\xi)} \partial_j \sqrt{p(x;\xi)} \nu(dx)$.

ii. The FIM $I(\xi)$ is symmetric and positive semi-definite, i.e. for $\mathbb{R}^k \ni v = v^i e_i$ with components:
$$v^t I(\xi) v = \int_\Omega \left\{ v^i \partial_i \ell(\xi; x) \right\}^2 p(x; \xi) \nu(dx) \geq 0,$$

iii. Assuming $I(\xi)$ be positive definite, means $v^i \partial_i \ell(\xi; x) \neq 0$, i.e. the vectors $\partial_i \ell(\xi; x) = \frac{\partial_i p(x;\xi)}{p(x;\xi)}$ be linearly independent, i.e. $\partial_i p(x; \xi)$ are linearly independent.

iv. For $g_{ij}(\xi)$ to be finite, we need at least $\partial_i \ell(\xi; \cdot) \in L^2(\Omega, \nu)$ for all ξ. We also assume also that $g_{ij} : \Xi \to \mathbb{R}$ is smooth (i.e. that we have a smooth covariant tensor of order 2).

Now, each $\xi \mapsto p(\cdot; 0, \dots, 0, \xi^i, 0, \dots, 0)$ is both, a one-dimensional model, and a curve on M, which can be used as coordinate curves on the statistical manifold. Let us re-write it in a more compact form: consider a differentiable curve from an open set $\mathbb{R} \supset I \ni \theta \mapsto p(\theta)$ to M and the tangent vector to the curve at $\theta = \theta_0$,

$$\dot{p}(\theta_0) = \lim_{h \to 0} \frac{p(\theta_0 + h) - p(\theta_0)}{h}.$$

[1] Second Bartlett's identity (Bartlett, 1953)

This tangent vector measures the variation w.r.t. the base measure ν. If, otherwise we consider $\ell(\theta; \cdot) = \log p(\theta; \cdot)$ as a coordinate curve, then

$$\frac{d}{d\theta} \log p(\theta; \cdot) = \frac{\dot{p}(\theta; \cdot)}{p(\theta; \cdot)}$$

the tangent vector is Fisher's score, which measures the relative variation of p w.r.t. itself. That is, by assuming the existence of Fisher's score, we are assuming that $\dot{p}$ is absolutely continuous w.r.t. p ($\dot{p}(\theta) \ll p(\theta)$). We will formalise it in the next chapter.

By Remark-iii. we have that both systems of tangent vectors are good candidates for a basis of the tangent space of M at ξ. Choosing as basis the one provided by Fisher's score, we obtain that the FIM is the matrix representation of **the Fisher metric**, which is a Riemannian metric for Remark-ii. It can be seen that the Fisher metric is invariant over the choice of coordinate system. Observe that $\mathbb{E}_\xi[\partial_i \ell(\xi; x)] = 0$[2], this gives us an intuition about the probabilistic/statistical meaning of tangent vectors: tangent vectors are random variables, which are centred w.r.t. the distribution at ξ.

Recall the definition of **sufficient statistic**: let $\mathbf{x} = (x_1, \dots, x_n)$ be a random sample and let κ be a statistic. Then κ is sufficient for ξ if the density of the vector $\mathbf{x}$ conditioned on the value of the statistic, $p(\mathbf{x}|\kappa(x); \xi)$ does not depend on ξ. Then, from Neyman's factorisation theorem we have the characterisation

$$\begin{aligned} \kappa : \Omega \to \Omega' \text{ sufficient for } \xi \;\Leftrightarrow\; & \exists\, s : \Omega' \times \Xi \to \mathbb{R}, \exists\, t : \Omega \to \mathbb{R} \text{ s. t.} \\ & p(x; \xi) = s(\kappa(x); \xi) t(x),\ \forall x, \xi. \end{aligned}$$

There is also a characterisation of a sufficient statistic in terms of the FIM, that is $I(\xi)$ is invariant under sufficient-statistic transformation, see e.g., (Amari & Nagaoka, 2000, thm. 2.1).

Other well-known results are

- Monotonicity, or chain rule: $I_\kappa(\xi) \leq I(\xi)$ i.e. the difference matrix $I(\xi) - I_\kappa(\xi)$ is positive semi-definite.
- Rao-Cramér inequality: $\mathbb{V}(\hat{\xi}(X)) \geq I(\xi)^{-1}$, where $\hat{\xi}$ is an estimator of ξ. When the equality holds, $\hat{\xi}$ is an efficient estimator of ξ.
- An efficient estimator does not always exist, unless we impose additional assumptions on the model M and on the parametrisation $\xi \mapsto p(\cdot; \xi)$.
- There always exists a sequence of asymptotically efficient estimators. The matrix $I(\xi)^{-1}$ quantifies the fluctuations of the asymptotic efficient estimator around the true value ξ.

[2] This is known as first Bartlett's identity (Bartlett, 1953)

2.2 Affine connections

Suppose that M is a $k-$dimensional statistical manifold and consider a function $\Gamma^{(\alpha)}_{ij,k}$ which maps each point ξ to the following

$$\left(\Gamma^{(\alpha)}_{ij,k}\right)_\xi \overset{\text{def}}{=} E_\xi\left[\left(\partial_i\partial_j\ell_\xi + \frac{1-\alpha}{2}\partial_i\ell_\xi\partial_j\ell_\xi\right)\left(\partial_k\ell_\xi\right)\right],$$

for some arbitrary $\alpha \in \mathbb{R}$. These are k^3 functions which define an affine connection $\nabla^{(\alpha)}$ on M, called the $\alpha-$connection,

$$\left\langle \nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k \right\rangle = \Gamma^{(\alpha)}_{ij,k}$$

where $g = \langle , \rangle$ is the Fisher metric.

- The $\nabla^{(\alpha)}-$connection is symmetric (i.e. torsion free).

Given another value $\beta \in \mathbb{R}$

$$\Gamma^{(\beta)}_{ij,k} = \Gamma^{(\alpha)}_{ij,k} + \frac{\alpha-\beta}{2}T_{ijk}$$

where T_{ijk} is a symmetric covariant tensor of order three defined by

$$\left(T_{ijk}\right)_\xi \overset{\text{def}}{=} E_\xi\left[\partial_i\ell_\xi\partial_j\ell_\xi\partial_k\ell_\xi\right],$$

a triple-covariance.

It also holds that

$$\begin{aligned}\nabla^{(\alpha)} &= (1-\alpha)\nabla^{(0)} + \alpha\nabla^{(1)} \\ &= \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}.\end{aligned}$$

Now, let us take the partial derivative of the elements of the Fisher information matrix:

$$\begin{aligned}\partial_k g_{ij} &= E_\xi\left[\left(\partial_k\partial_i\ell_\xi\right)\left(\partial_j\ell_\xi\right)\right] + E_\xi\left[\left(\partial_i\ell_\xi\right)\left(\partial_k\partial_j\ell_\xi\right)\right] + E_\xi\left[\left(\partial_i\ell_\xi\right)\left(\partial_j\ell_\xi\right)\left(\partial_k\ell_\xi\right)\right] \\ &= \Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i}.\end{aligned}$$

This means—recall Corollary 1.1—that the following results holds

Theorem 2.1. *The 0−connection is the Riemannian connection w.r.t. the Fisher metric.*

Let us now introduce particular parametric families.

2.3 The exponential and mixture families

Recall that a probability density p on Ω belongs to an the exponential family with k parameters if there exist functions $C, \{B_i\}_{i=1}^k$ on Ω and ψ on Θ such that

$$p(x;\theta) = \exp\{C(x) + \theta^i B_i(x) - \psi(\theta)\}.$$

(Remember that we use the Einstein summation over repeated indices convention).

$[\theta^i]$ are called natural, or canonical, parameters.

$$\psi(\theta) = \log \int_\Omega \exp\left[C(x) + \theta^i B_i(x)\right] \mathrm{d}x$$

From the definition of an exponential family, we have

$$\begin{aligned} \partial_i \ell(x;\theta) &= B_i(x) - \partial_i \psi(\theta) \quad \text{and} \\ \partial_i \partial_j \ell(x;\theta) &= -\partial_i \partial_j \psi(\theta). \end{aligned}$$

and, hence, $\Gamma^{(1)}_{ij,k} = -\partial_i \partial_j \psi(\theta) E_\theta \left[\partial_k \ell_\theta\right]$, which is zero for the first Bartlett's identity. This means that the system of coordinates given by the canonical parameters $[\theta^i]$ is an affine coordinate system w.r.t. $\nabla^{(1)}$, or a 1-affine coordinate system, and the exponential family is flat w.r.t. $\nabla^{(1)}$, or 1-flat.

Definition 2.2 (Exponential connection)**.** $\nabla^{(1)} =: \nabla^e$ is called exponential connection.

On the other hand, if p can be expressed as

$$p(x;\theta) = C(x) + \theta^i B_i(x)$$

than M is a mixture family with mixture parameters $[\theta^i]$. These probability functions form an affine subspace of $\mathcal{P}_>(\Omega)$.

When Ω is finite, $\mathcal{P}_>(\Omega)$ is a mixture family (Amari & Nagaoka, 2000, pag. 35).

We have:

- $\partial_i \ell(\theta;x) = \frac{B_i(x)}{p(x;\theta)}$
- $\partial_i \partial_j \ell(\theta;x) = -\frac{B_i(x)B_j(x)}{[p(x;\theta)]^2}$

Hence: $\Gamma^{(-1)}_{ij,k} = 0$ and $[\theta^i]$ is a $(-1)-$affine coordinate system and the mixture family is $(-1)-$flat. That is why, $\nabla^{(-1)} =: \nabla^m$ is called mixture connection.

We have then a key result in information geometry.

Theorem 2.2 (Affine coordinate systems)**.** *An exponential family (resp. a mixture family) is $e-$flat ($m-$flat) and its natural parameter form an $e-$affine ($m-$affine) coordinate system.*

It turns out that an exponential family is also $m-$flat and a mixture family is also $e-$flat. To show this result we have to introduce the concept of duality of connections.

2.4 Duality of connections

Through the metric there emerges a duality between connections. Let us consider $\left(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)}\right)$.

Definition 2.3 (Dual connections)**.** Let (M, g) be a Riemannian manifold and let ∇, ∇^* be two affine connections. If for every $X, Y, Z \in \mathfrak{X}(M)$ it holds

$$Zg(X,Y) = g\left(\nabla_Z X, Y\right) + g\left(X, \nabla^*_Z Y\right). \tag{2.3}$$

In terms of their local expression, given a coordinate system $[\xi^i]$

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}$$

In general, we have

- If ∇ is compatible with the metric, then ∇ is self-dual $\nabla = \nabla^*$;
- given a metric g and a connection ∇, there is a unique dual connection ∇^* of ∇ w.r.t. g;
- $(\nabla^*)^* = \nabla$;
- $\frac{\nabla+\nabla^*}{2}$ is metric.

Theorem 2.3 (Duality of $\alpha-$connections)**.** *For any statistical model the $\alpha-$ and $(-\alpha)-$connection are dual with respect to the Fisher metric.*

Let γ be a differentiable curve from $I \subset \mathbb{R}$ to M and consider the covariate derivatives of a vector field X along γ w.r.t. the two connections, i.e. $\frac{DX}{dt}, \frac{D^*X}{dt}$. Then, by Equation 2.3,

$$\frac{\mathrm{d}}{dt} g(X(t), Y(t)) = g\left(\frac{DX(t)}{dt}, Y(t)\right) + g\left(X(t), \frac{D^*Y(t)}{dt}\right) \tag{2.4}$$

where we recall that the derivation $\frac{d}{dt}$ corresponds here to the derivation along the curve.

If X and Y are parallel along γ, respectively w.r.t. ∇ and ∇^*, then $0 = \frac{DX(t)}{dt} = \frac{D^*Y(t)}{dt}$, which means that the RHS of Equation 2.4 is 0 and $g(X(t), Y(t))$ is constant along the curve and

$$g_q\left(\Pi_\gamma(X), \Pi^*_\gamma(Y)\right) = g_p(X, Y)$$

2.5 Dually flat spaces

Let (g, ∇, ∇^*) be a dualistic structure on M.

If the connections are both symmetric, then

$$M \text{ is } \nabla - \text{flat} \iff M \text{ is } \nabla^* - \text{flat}.$$

In terms of our two particular connections, the exponential family if 1-flat and is, hence, also (-1)-flat, while the mixture family is (-1)-flat and is also 1-flat.

If the dual connections ∇, ∇^* are both flat, then (M, g, ∇, ∇^*) is a **dually flat space**.

If (M, g, ∇, ∇^*) is a **dually flat space**, then there exist:

i. a $\nabla-$affine coordinate system $[\theta^i]$ and
ii. a ∇^*-affine coordinate system $[\eta_j]$, for which $\partial_i = \frac{\partial}{\partial \theta^i}$ and $\partial^j = \frac{\partial}{\partial \eta_j}$.

So, $g(\partial_i, \partial^j)$ is constant over M; moreover, we can choose one of the two connection-affine coordinate system (any regular affine transformation of an affine coordinate system works), say we choose the $[\eta_j]$, so that $g(\partial_i, \partial^j) = \delta_i^j$ and $\left[\theta^i\right], \left[\eta_j\right]$ are said **mutually dual** or **dual coordinate systems**.

Vice versa, if two coordinate systems $\left[\theta^i\right], \left[\eta_j\right]$ for a Riemannian manifold (M, g) exist are mutually dual, then the connections ∇, ∇^* for which they are affine are determined and the space (M, g, ∇, ∇^*) is flat.

Let us recall our local expressions:

i. $g_{ij} := \langle \partial_i, \partial_j \rangle$ and $g^{ij} := \langle \partial^i, \partial^j \rangle$;
ii. $\partial^j = \left(\partial^j \theta^i\right) \partial_i$ and $\partial_i = \left(\partial_i \eta_j\right) \partial^j$;
iii. then, we derive: $g_{ij} = \frac{\partial \eta_j}{\partial \theta^i}$ and $g^{ij} = \frac{\partial \theta^i}{\partial \eta_j}$.

Suppose, now, that we are given mutually dual coordinate systems $\left[\theta^i\right], \left[\eta_j\right]$ and, for some $\psi : M \to \mathbb{R}$, we consider the PDE

$$\partial_i \psi = \eta_i \tag{2.5}$$

or, in terms of the differential $d\psi = \eta_j d\theta^i$. A solution to Equation 2.5 exists is and only if $\partial_i \eta_j = \partial_j \eta_i$. But here, $\partial_i \eta_j = g_{ji} = g_{ij} = \partial_j \eta_i$ and so Equation 2.5 has a solution.

Using iii. we have $\partial_i \partial_j \psi = g_{ij}$, implying that the second derivative of ψ is a positive definite matrix, i.e. ψ is strictly convex in the coordinates $\left[\theta^i\right]$. In the same way, we find the solution, for a function φ, to the PDE $\partial^i \varphi = \theta^i$.
Using a solution ψ to Equation 2.5, let

$$\varphi = \theta^i \eta_i - \psi$$

whose differential is

$$d\varphi = \theta^i d\eta_i + \eta_i d\theta^i - d\psi$$

but we already derived the differential of ψ, $d\psi = \eta_j d\theta^i$, so that the equation becomes

$$d\varphi = \theta^i d\eta_i$$

and the two differential have the same form. Finally, also for φ holds $\partial^i \partial^j \varphi = g^{ij}$ and φ is strictly convex in the coordinates $[\eta_j]$.

Combining these results (and the usual "confusion" between probability points and their local coordinates/parameters), we write

$$\varphi(\eta) = \max_{\theta \in \Theta} \left\{\theta^i \eta_i - \psi(\theta)\right\}$$
$$\psi(\theta) = \max_{\eta \in \mathrm{H}} \left\{\theta^i \eta_i - \varphi(\eta)\right\}$$

where φ, ψ are convex functions on convex subsets of $\mathbb{R}^k$. This coordinate transformation between $\left[\theta^i\right]$ and $\left[\eta_j\right]$ (and vice versa) is an instance of **Legendre transformation**.

3 Non-parametric information geometry

Geometrising a problem or a field should, in principle, provide tools which do not depend from parametrisations. Hence, it makes sense that non-parametric models should be the main object of interest in IG. Of course, dealing with infinite-dimensional spaces is not always easy (or at our reach), but we can still introduce the methods and results of a non-parametric IG in the finite-dimensional case. In this way, the finite-dimensional (parametric) theory is derived from the infinite-dimensional (non-parametric) one. Here, we focus only on finite sample spaces (Pistone, 2020), but the theory has been developed in the infinite-dimensional case and you can find all the details and further references in (Chirco & Pistone, 2022; Pistone, 2013).

Let Ω be our finite sample space, so that the space of real functions on Ω, i.e. $\mathbb{R}^\Omega$ is a finite dimensional vector space. We are interested in the geometry of sets of probability functions

$$\mathcal{P}(\Omega) = \{q \in \mathbb{R}^\Omega : p(x) \geq 0, \sum_{x\in\Omega} p(x) = 1\}$$
$$\mathcal{P}_>(\Omega) = \{q \in \mathbb{R}^\Omega : p(x) > 0, \sum_{x\in\Omega} p(x) = 1\}.$$

The first is called *the probability simplex*[1] and is generated by $\delta-$functions, centred at each point $x \in \Omega$, i.e. $\delta_x(y) = 1$ if and only if $x = y$ and is zero otherwise. $\mathcal{P}(\Omega)$ is a convex subset of $\mathbb{R}^\Omega$, or, also, a convex subset of the affine space $\mathcal{A}(\Omega) = \{p \in \mathbb{R}^X : \sum_{x\in X} p(x) = 1\}$. Depending on which "outer" space we consider $\mathcal{P}(\Omega)$ to be embedded into, we can induce different geometries on the simplex.

The second set, $\mathcal{P}_>(\Omega)$ is called *the open probability simplex* and, as in the parametric case, it will be the main object of our study.

We will find that the geometric structure defined in this section is similar to the one of the previous section, but there will also be substantial differences. We will try to highlight both similarities and differences.

Let us start with from the intuitive idea. Recall the definition of *affine space* (see Example 1.1): we need a vector space/spaces of translations and the *dispalcement* mapping from the points set to the translations, such that the parallelogram law holds.

A natural affine structure on $\mathcal{P}(\Omega)$ could be given considering

[1] A simplex is a particular case of polytope, where a polytope is convex hull of a set of points. An $n-$simplex is the convex hull of exactly $n + 1$ points, meaning that these points are affinely independent.

- the vector space: $B_1 = \{v \in \mathbb{R}^\Omega : \sum_{x\in\Omega} v(x) = 0\}$, which is parallel to $\mathcal{A}(\Omega)$,
- the displacement: $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to B_1$ which maps a pair of probability functions to their difference $(p, q) \mapsto p - q$, indeed $\sum (p(x) - q(x)) = 0$.

This is the *flat* affine geometry inherited from the bigger affine space $\mathcal{A}(\Omega)$ (Chirco & Pistone, 2022).

Observe that, if we take a smooth curve on the simplex $I \ni \theta \mapsto p(\theta)$, i.e. a 1-dimensional model, and assume I to be an open subset of $\mathbb{R}$, the derivative $\frac{d}{d\theta}p(\theta) = \lim_{h\to 0} \frac{p(\theta+h)-p(\theta)}{h} \in B_1$. For this reason, we call B_1 the tangent space of the (open) probability simplex and denote it by $\mathcal{T}\mathcal{P}(\Omega) = \mathcal{T}\mathcal{P}_>(\Omega)$.

Again, Fisher suggests another way to take derivatives in $\mathcal{P}_>(\Omega)$, by Fisher's score (i.e. log-derivatives)

$$\frac{d}{d\theta}\log p(\theta) = \frac{\dot{p}(\theta)}{p(\theta)} =: \overset{\star}{p}(\theta). \tag{3.1}$$

Notation: we use the notation $\dot{p}$ to denote the "usual" derivative $\frac{dp}{d\theta}$ and $\overset{\star}{p}$ for the log-derivative (Fisher's score, in statistics).

Let us define the space B_θ containing all random variables u over Ω such that

$$\sum_{x\in\Omega} u(x)p(x;\theta) = 0,$$

that is, u is a centred random variable (zero expectation) w.r.t. the probability distribution $p(\theta)$, or a *contrast* for $p(\theta)$. Then $\frac{d}{d\theta}\log p(\theta) \in B_\theta$. Using this a definition for a tangent vector (velocity) to a curve (1-dimensional model) we have that each probability function has its own tangent space.

These examples motivate the definition of an affine structure on the probability simplex.

3.1 Affine structure of the probability simplex

Let us go back to the running example of the affine manifold, in particular, to Example 1.4 and Example 1.5. From these examples we learned that, given an affine structure on a set of points, (i) we automatically have a differentiable structure; (ii) we can confuse the space of translations with each tangent space at a point; (iii) if we give the mapping $\mathcal{A}_p^q$, which we now know is the parallel transport, we can get rid of the space of translations and re-define the affine structure in terms of, and work only with, the tangent spaces.

Definition 3.1 (Affine manifold - II)**.** Let M be a set and let $(B_q)_{q \in M}$ be a family of topological linear (abbrev. top-linear) spaces[2]. Let $(\mathbb{U}_p^q)_{p,q \in M}$ be a family of top-linear isomorphisms $\mathbb{U}_p^q : B_p \to B_q$ satisfying the cocycle condition

- (AF1) $\mathbb{U}_p^p = I$ and $\mathbb{U}_q^r \, \mathbb{U}_p^q = \mathbb{U}_p^r$.

$\mathbb{U}_p^q$ is the *parallel transport* from B_p to B_q.

Let $\mathbb{S} : (p, q) \mapsto s_p(q) \in B_p$ be the displacement mapping, such that

- (AF2) for each $p \in M$, s_p is injective from a neighbourhood of p in M to B_p;
- (AF3) $\mathbb{S}(p, q) + \mathbb{U}_q^p \mathbb{S}(q, r) = \mathbb{S}(p, r)$, or equivalently, for any $p, q, r \in M$, $s_p(q) + \mathbb{U}_q^p s_q(r) = s_p(r)$.

The structure $\left(M, (B_q)_{q \in M}, (\mathbb{U}_p^q)_{p,q \in M}, \mathbb{S}\right)$ is an affine space.

Finally, we assume that for each $p \in M$, the image $\mathrm{Im}(s_p)$ is a neighbourhood of $0 = s_p(p)$ (i.e. it contains an open subset and we can properly define the coordinate domains).

- Each s_p, together with the (open) subset of M on which it is defined, is a chart.
- Form (AF3) we have $s_p(q) + \mathbb{U}_q^p s_q(p) = s_p(p) = 0$
- The change-of-chart mapping at $r = s_q^{-1}(v)$, with $v \in B_q$ is

$$s_p \circ s_q^{-1}(v) = s_p(r) = s_p(q) + \mathbb{U}_q^p v$$

Observe that if $B_q = V$ for all $q \in M$, then the parallel transports are the identity map and we have the usual translation in the change-of-charts transformation (resulting in a change of origin).

Consider the displacement $s_p(q) = p - q$ on the open probability simplex $\mathcal{P}_>(\Omega)$: it satisfies the axioms of Definition 3.1, $\mathrm{Im}(s_p) = \mathcal{P}_>(\Omega) - p$ is open in the tangent space

$$\mathcal{T}\mathcal{P}(\Omega) = \mathcal{T}\mathcal{P}_>(\Omega) = \left\{ v \in \mathbb{R}^\Omega : \sum_{x \in \Omega} v(x) = 0 \right\}$$

and so, it is a chart on $\mathcal{P}_>(\Omega)$ (but not on $\mathcal{P}(\Omega)$ for $\mathcal{P}(\Omega) - p$ is closed in $\mathcal{T}\mathcal{P}(\Omega)$). The inverse $s_q^{-1}(v) = v + q$ and the change-of-chart is $s_p \circ s_q^{-1}(v) = s_p(v + q) = v + q - p = v + s_p(q)$.

This, in our statistical setting, is the *non interesting* displacement. Let us look at Fisher's score.

[2] A top-linear space is a linear space endowed a topology, which makes the operations of a vector space continuous.

3.1.1 Fisher's score on the open probability simplex

Remember that, in IG, we are interested in a particular flavour of differential geometry, thus, we always use a special set of charts and always give a presentation compatible with the specific statistical intuition of IG.

Take, again, a smooth curve on the probability simplex $I \ni \theta \mapsto q(\theta) \in \mathcal{P}(\Omega)$ and assume I to be an open subset of $\mathbb{R}$. To be very precise: for each θ we have

$$\theta \mapsto [q(\theta) = p(\cdot;\theta) : \Omega \to \mathbb{R}].$$

Assume that there exists a pair (x, θ_0) such that $q(x, \theta_0) = 0$, that is the curve $\theta \mapsto q(x, \theta)$, which is a curve in $\mathbb{R}$, has a minimum in θ_0, implying that also $\dot{q}(\theta_0) = 0$. Regarding (finite) $q_\theta, \dot{q}_\theta$ as measures on Ω, we can say that $\dot{q}_\theta$ is absolutely continuous w.r.t q_θ, written is short $\dot{q}_\theta \ll q_\theta$. Then, by Radon-Nikodym theorem, there exists the Radon-Nikodym derivative of $\dot{q}_\theta$ w.r.t. q_θ:

$$\overset{\star}{q}_\theta := \frac{\dot{q}_\theta}{q_\theta} \in L^1(\Omega, q_\theta).$$

$\dot{q}_\theta(x) = \dot{q}(x;\theta)$ is exactly the Fisher's score, which is defined for smooth statistical models[3]. If, in particular, the curve is on the open probability simplex, then $\overset{\star}{q}_\theta := \frac{d}{d\theta} \log q(\theta)$ is the log-derivative.

Notice that $\mathbb{E}_{q(\theta)}[\overset{\star}{q}\ (\theta)] = 0$, that is: Fisher's score is a contrast for the "true" probability. Thus, we consider for each $q \in \mathcal{P}_>(\Omega)$ the vector spaces of $q-$contrasts, i.e. random variables centred at zero w.r.t. q

$$B_q = \left\{ v \in \mathbb{R}^\Omega : \sum_{x \in \Omega} v(x) q(x) = 0 \right\}.$$

These play, both, the role of spaces of translations in the Definition 3.1 and of tangent spaces. We define also the **statistical bundle** as

$$\mathcal{S}\mathcal{P}_>(\Omega)\{(q, v) \in \mathcal{P}_>(\Omega) \times \mathbb{R}^\Omega : \mathbb{E}_q[v] = 0\}.$$

B_q is called a **fibre**[4] of the bundle at q.

Example 3.1 (A statistical bundle)**.** Let us draw a picture of $\mathcal{S}\mathcal{P}(\Omega)$ for $\Omega = \{1, 2\}$:

Now that we have tangent spaces, or spaces of translations, or fibres of velocities, at each $q \in \mathcal{P}_>(\Omega)$, we need to introduce the parallel transports and displacement the define our statistical affine manifolds.

[3] This holds also for non-finite sample spaces Ω, see (Chirco & Pistone, 2022) and for general signed measures (Ay et al., 2017).

[4] Or *fiber* which is the US version of the UK "fibre". The same happens with "centre" (UK) and ceter (US).

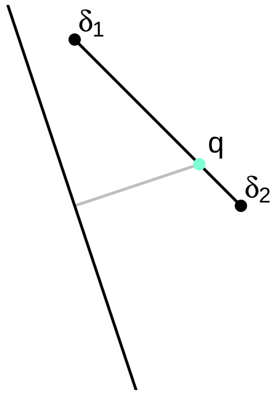

Figure 3.1: Visualising the statistical bundle. The segment is the probability simplex, while the line through, which is perpendicular to the vector connecting the origin to the point q, is the (moving) vector space of q–contrasts.

3.2 Exponential and mixture affine spaces

Let us start with the transports.

3.2.1 Parallel transport

For each $(p,u),(q,v) \in \mathcal{SP}_{>}(\Omega)$ we introduce two families of parallel transports:

- ${}^{\mathrm{e}}\mathbb{U}_p^q : B_p \ni u \mapsto u - \mathbb{E}_q[u] \in B_q$ — the *exponential transport*
- ${}^{\mathrm{m}}\mathbb{U}_q^p : B_q \ni v \mapsto \frac{q}{p} v \in B_p$ — the *mixture transport*.

It holds, for all $(p,u),(q,v) \in \mathcal{SP}_{>}(\Omega)$

$$\langle u, {}^{\mathrm{m}}\mathbb{U}_q^p v\rangle_p = \langle {}^{\mathrm{e}}\mathbb{U}_p^q u, v\rangle_q, \tag{3.2}$$

i.e., the two transports are dual to each other. It is easy to verify.

3.2.2 Fisher's information

Observe that, in stating the last result, we assumed that there is an inner product on each fibre. Given a 1-dimensional model $\theta \mapsto q(\theta)$ we defined its velocity as Fisher's score $\overset{\star}{q}(\theta)$.

Then, Fisher's information defines an inner product:

$$\mathbb{E}_{q(\theta)}\left[|\overset{\star}{q}(\theta)|^2\right] = \langle \overset{\star}{q}(\theta), \overset{\star}{q}(\theta)v\rangle_{q(\theta)}. \tag{3.3}$$

3.2.3 Displacements

We also define two displacement mappings.

Definition 3.2 (Exponential displacement)**.** The exponential displacement of $\mathcal{P}_>(\Omega)$ is the map

$$\mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto s_p(q) := \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \in B_p.$$

It is easy to show (exercise) that, together with the exponential transports, it defines the structure of an affine manifold, called the exponential affine space.

Let us compute the inverse of the chart s_p, which is defined on the whole fibre B_p.

$$\begin{aligned} v &= \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \\ e^v &= \frac{q}{p} e^{-\mathbb{E}_p\left[\log \frac{q}{p}\right]} \longrightarrow \mathbb{E}[e^v] = e^{-\mathbb{E}_p\left[\log \frac{q}{p}\right]} \\ \Rightarrow \quad e^v &= \frac{q}{p} e^{-\log \mathbb{E}[e^v]} \\ q &= e^{v - \log \mathbb{E}[e^v]} \cdot p = e^{v - K_p(v)} \cdot p \end{aligned}$$

Hence $s_p^{-1}(v) = e^{v - K_p(v)} \cdot p$, where $K_p : v \mapsto \log \mathbb{E}[e^v]$ is the cumulant functional.

Some important properties of the cumulant:

- The Kullback-Leibler divergence $\mathrm{D}\left(p \| q\right) = \mathbb{E}_p\left[\log \frac{p}{q}\right] = \mathbb{E}_p\left[\log \frac{p}{\exp(v - K_p(v)) \cdot p}\right] = K_p(v)$, i.e.
$$K_p(s_p(q)) = \mathrm{D}\left(p \| q\right) \tag{3.4}$$
- $dK_p(v)[h] = \mathbb{E}_{e_p(v)}[h] = \left\langle \frac{e_\rho(v)}{p} - 1, h \right\rangle_{e_p(v)}$, where $d \cdot [h]$ denotes the directional derivative
- $d^2 K_p(v)[h, k] = \mathrm{Cov}_{e_p(v)}(h, k) = \left\langle {}^{\mathrm{e}}\mathbb{U}_p^{e_p(v)} h, {}^{\mathrm{e}}\mathbb{U}_p^{e_p(v)} k \right\rangle_{e_p(v)} = \left\langle h, {}^{\mathrm{m}}\mathbb{U}_{e_p(v)}^p \, {}^{\mathrm{e}}\mathbb{U}_{e_p(v)}^p k \right\rangle_p.$

Definition 3.3 (Mixture displacement)**.** The mixture displacement of $\mathcal{P}_>(\Omega)$ is the map

$$\mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto \eta_p(q) := \frac{q}{p} - 1 \in B_p.$$

The inverse chart is $\eta_p^{-1}(v) = (1 + v) \cdot p$.

4 Applications of IG

4.1 The natural gradient

Here I re-write an interesting exercises shown by Giovanni Pistone at Genua, March, 6 2023. You can find his slides on his website.

The context is the usual finite non-parametric setting of the previous section, Section 3. Here, we denote B_q the tangent space at q, i.e. the fibre at q of the statistical bundle, by $\mathcal{S}_q\mathcal{P}_>(\Omega)$.

A **section** in $\mathcal{SP}_>(\Omega)$ is a vector field over $\mathcal{P}_>(\Omega)$, i.e. it is a function

$$A : \mathcal{SP}_>(\Omega) \ni q \mapsto A(q) \in \mathcal{S}_q\mathcal{P}_>(\Omega).$$

The **gradient** in $\mathcal{SP}_>(\Omega)$ w.r.t. the Fisher metric of a real function $\Phi \in D(\mathcal{P}_>(\Omega))$ is a section $A = \operatorname{grad}\Phi$ such that

$$\frac{d}{dt}\Phi(q(t)) = \langle A(q(t)), \overset{\star}{q}(t)\rangle_{q(t)}, \tag{4.1}$$

where on the right-hand-side of Equation 4.1 we have the usual differential of a differentiable function Φ, see Definition 1.6.

In the parametric case, i.e. when M is a $k-$dimensional statistical manifold in $\mathcal{P}_>(\Omega)$, we have see that $\frac{\partial}{\partial\theta_j}\log p(\theta)$ for $j = 1, \dots, k$ is a basis for the tangent spaces (sections) at $p(\theta)$, or, simply at θ. Then, the **gradient** on M of $\Phi \in D(M)$ is

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

where the $\operatorname{grad}\Phi$ is the global gradient on $\mathcal{P}_>(\Omega)$.

Hence, the gradient on M is defined as

$$\operatorname{grad}_{\mathcal{M}}\Phi(q) = \Pi(q)\operatorname{grad}\Phi(q),$$

where $\Pi(q)$ is the projection on the fibre of M.

Let us do some computations.

$\mathcal{S}_p\mathcal{P}_>(\Omega) \ni u = \sum_{j=1}^{k} u^j \frac{\partial}{\partial\theta^j}\log p(\theta)$.

4.2 Information geometry of graphical models

5 Summary

In summary, this book has no content whatsoever.

References

Abraham, R., Marsden, J. E., & Ratiu, T. (2012). *Manifolds, tensor analysis, and applications* (Vol. 75). Springer Science & Business Media.

Amari, S. (2016). *Information geometry and its applications* (Vol. 194). Springer. http://doi.org/10.1007/978-4-431-55978-8

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry* (Vol. 191). American Mathematical Soc. http://doi.org/10.1090/mmono/191

Ay, N., Jost, J., Vân Lê, H., & Schwachhöfer, L. (2017). *Information geometry* (Vol. 64). Springer. http://doi.org/10.1007/978-3-319-56478-4

Bartlett, M. S. (1953). Approximate confidence intervals. *Biometrika*, *40*(1/2), 12. http://doi.org/10.2307/2333091

Carmo, M. P. do. (1992). *Riemannian geometry.* Boston, Mass. [etc: Birkhäuser. http://doi.org/10.1007/978-1-4757-2201-7

Chentsov, N. N. (1982). Statiscal decision rules and optimal inference. *Monog*, *53*.

Chirco, G., & Pistone, G. (2022). Dually affine Information Geometry modeled on a Banach space. Retrieved from http://arxiv.org/abs/2204.00917

Efron, B. (1975). Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency). *The Annals of Statistics*, *3*(6), 1189–1242. http://doi.org/10.1214/aos/1176343282

Grady, L. J., & Polimeni, J. R. (2010). *Discrete calculus: Applied analysis on graphs for computational science* (Vol. 3). Springer.

Grinspun, E., Desbrun, M., Polthier, K., Schröder, P., & Stern, A. (2006). Discrete differential geometry: An applied introduction. *ACM Siggraph Course*, *7*(1).

Lang, S. (2012). *Differential and riemannian manifolds* (Vol. 160). Springer Science & Business Media.

Petersen, P. (2006). *Riemannian geometry* (Vol. 171). Springer.

Pistone, G. (2013, July). Nonparametric Information Geometry. arXiv. Retrieved from https://arxiv.org/abs/1306.0480

Pistone, G. (2020). Information geometry of the probability simplex: A short course. *Nonlinear Phenomena in Complex Systems*, *23*(2), 221–242. http://doi.org/10.33581/1561-4085-2020-23-2-221-242

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Reson. J. Sci. Educ*, *20*, 78–90.

Sernesi, E. (1994). Geometria 2 bollati boringhieri. Torino.

.1 Exterior calculus

One possibility for generalising the differential calculus to discrete domains, is through the *exterior calculus*, see (Abraham, Marsden, & Ratiu, 2012; Grady & Polimeni, 2010; Grinspun, Desbrun, Polthier, Schröder, & Stern, 2006).