INFORMATION GEOMETRY

ENTROPIC SMASH

Contents

1. Introduction

The field of *Information Geometry* has been defined by some of its pioneers as "a method of exploring the world of information by means of modern geometry" **?**, "the field that studies the geometry of decision making" **?**, and "the differential geometric treatment of statistical models" **?**. In general, it is a discipline that lies at the intersection of mathematics and statistics and uses the language of differential geometry.

The language of information geometry is the same as that of differential geometry, but the purpose is to study the instersection between mathematics and statistics.

The key geometric object is the so-called *statistical manifold*, and it is formalized as a triple

$$(\mathcal{M}, g, T),$$

where $\mathcal{M}$ is just a smooth manifold, usually finite-dimendional, and g is the so-called *Fisher-Rao metric* which, in coordinates, we can just think of as a matrix. The pair $(\mathcal{M}, g)$ is the standard framework in Riemannian geometry. Here, we have an additional object, T, which is the *Amari Chenstov tensor* and is

slightly more complicated than a metric: it takes as input three vectors and outputs one number, while the metric takes as input of two vectors.

There is an equivalent phrasing of this framework. We can use the tensor T and the metric g to build an object which is called a *connection*. In Riemannian geometry, given a metric, we can always find a connection ∇, the Levi-Civita connection, which induces the parallel transport, the geodesics, and so on ... and the study of the Levi-Civita connection connection is essentially Riemannian geometry. Here, we don't study the Levi-Civita connection, but we rather build another object, another connection, which depends on both g and T. And then, we find its dual ∇^*, which turns out to be another connection, and the equality

$$\nabla = \nabla^*$$

is trivial, or we get something self-adjoint, only when the connection we started with was the Levi-Civita connection associated with g. But this is just a very particular case and, in general, ∇ is not Levi-Civita. We can check that this is a real duality, in the sense that taking the dual is an involution, i.e. $\nabla^{**} = \nabla$ always holds. And the case where ∇ is flat, this duality is exactly the well-known Legendre duality from convex analysis.

The natural question is: why do we care about this very abstract structure? The motivation comes from statistics. The natural metric to put on our manifold and the natural three-tensorial Amari Chenstov tensor are essentially universal objects coming out of the *Fisher information*. Let's consider the standard framework in parametric statistics: pick a sample space Ω, with some sigma algebra, $\mathscr{F}$, then the pair $(\Omega, \mathscr{F})$ is our *sample*, or *state*, *space*. Then we consider a subset $\mathcal{M} \subset \mathcal{P}(\Omega)$ of the space of probability measures over Ω, with some substructure. We assume these subsets are very regular, they are smooth finite-dimensional submanifolds, which we can think of a sphere or embedded in the Euclidean space of dimension n. Given a manifold, it's rather natural to use coordinates or parameterize, at least locally, our probability measures over it. We consider good subsets $\Omega \subseteq \mathbb{R}^n$, such as an open set, and the mapping:

$$\theta \in \Theta \longmapsto P_\theta \in \mathcal{M}$$

into probability measures in our $\mathcal{M}$. We want a few good properties for this parameterization: we will assume that this is very smooth with respect to the θ variable, but, for now, let's only assume that this is injective, so that we can look at the inverse. In many applications, we are able to parameterize the whole manifold with only one *chart*.

[DRAWING]

The main conceptual point to be kept in mind is that we can distinguish two different levels in this framework:

- the manifold of probability distributions M is an intrinsic object and it is given, it constitutes the *abstract level*;
- the parameter space Θ is a chart we choose and use to parametrize the abstract object, it lies at the *concrete level*.

Now we need the metric g and the tensor T. Actually, g comes from the Fisher information and T a derived measure. We will see that this is essentially the only reasonable choice from the point of view of statistics.

Let's now see some example of objects in this framework that are already kind of interesting from the geometric standpoint.

Example (Normal distributions). Let $\Omega = \mathbb{R}$ and the family $\mathcal{M} \subset \mathcal{P}(\mathbb{R})$ be the space of normal distributions, namely we parametrize $\mathcal{M}$ with $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+$ and the mapping

$$P_\theta = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\omega-\mu)^2/2\sigma^2} \, \mathrm{d}\omega.$$

So $\mathcal{M}$ is literally the space of measures of the Gaussians, and it is diffomorphic to the half space, so it is flat. Morover, this is one of the examples where you can parametrize everything with only one chart.

Example (Exponential family). Let $\Omega = \mathbb{R}$, $\theta = \mathbb{R}^n$ and again we parametrize everything with just one chart via

$$P_\theta = \exp(k(\omega) + \langle V(\omega), \theta\rangle - F(\theta))\,\mathrm{d}\omega$$

where $k : \mathbb{R} \longrightarrow \mathbb{R}$ is just a nice function: smooth with compact support, while $V : \mathbb{R} \longrightarrow \mathbb{R}^n$ is a similarly nice vector-valued object and F is essentially given. We want this to be a probability measure, so if we integrate in ω, we want to get 1, and this forces the definition

$$F(\theta) = \log \int_\Omega \exp(k(\omega) + \langle V(\omega), \theta\rangle)\,\mathrm{d}\omega.$$

So are we have two degrees of freedom, k and V, and we can then parametrize the probability measures over $\mathbb{R}$

Let's now write down, informally, the tensor objects g and T. We assume that, for every θ, our probability P_θ is absolutely continuous with respect to a certain measure μ, i.e.

$$P_\theta \ll \mu \in \mathcal{P}(\Omega)$$

so that there is a measure dominating all the probabilities p_θ, or at least locally, and this is the same as writing $P_\theta = p_\theta \mu$.

Now, we want to define g, which is essentially a metric, so we have two indices:

$$\begin{aligned} g_{ij}(\theta) &= \int_\Omega \frac{\partial}{\partial\theta_i} \log p_\theta(\omega) \frac{\partial}{\partial\theta_j} \log p_\theta(\omega)\, p_\theta(\omega)\,\mathrm{d}\mu(\omega) \\ &= \int_\Omega \frac{\partial}{\partial\theta_i} \log p_\theta(\omega) \frac{\partial}{\partial\theta_j} \log p_\theta(\omega)\, P_\theta(\omega) \\ &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i} \log p_\theta(\omega) \frac{\partial}{\partial\theta_j} \log p_\theta(\omega)\right] \end{aligned}$$

so we are taking the log-likelihood and differentiating with respect to the variable θ, as $\theta \subset \mathbb{R}^n$, which has coordinates, so we can differentiate with respect to this variable. We then do the same with respect to another variable, and integrate with respect to the same probability distribution.

When $i = j$, this is the standard definition of Fisher information. Now, the point is that, given coordinates, again, where Θ is our choice and is not intrinsic, we can check that this quadratic form is symmetric and positive definite so it induces a metric or a scalar product, which depends on the point θ we are studying. This is essentially the definition of Riemannian metric. Pick two vectors $v, w \in \mathbb{R}^n$ and think of them as tangent vectors at a given precise $\theta \in \Theta$, then the inner product between the two is going to be

$$g_\theta(v, w) = \sum_{i,j} v_i w_j g_{ij}(\theta)$$

where we wrote the two vectors in the standard Euclidean coordinates. In the current frame, we should think of abstract manifolds and abstract tangent spaces attached to these manifolds, but when we write things in coordinates, we are just looking at open sets of an $\mathbb{R}^n$, and these concepts kind of collapse into basic concepts of linear algebra. In the abstract picture, $\mathcal{M}$ is a manifold of measures, and g will always take that Fisher-like form.

The 3-tensor T depends on three indices:

$$T_{ijk}(\theta) = \int_\Omega \frac{\partial}{\partial\theta_i} \log p_\theta(\omega) \frac{\partial}{\partial\theta_j} \log p_\theta(\omega) \frac{\partial}{\partial\theta_k} \log p_\theta(\omega)\, p_\theta(\omega)\,\mathrm{d}\mu(\omega)$$

and can no longer be represented as a matrix.

It is now clear that we can go on with this kind of definition and define some k-tensor depending on k indices as the integral of $p_\theta \prod_{\ell=1}^{k} \partial \log p_\theta / \partial \theta_\ell$. We would still get universal invariant objects as g and T. The first order one is, instead, simply 0:

$$\int_\Omega \frac{\partial \log p_\theta}{\partial \theta_i} p_\theta = \int_\Omega \frac{1}{p_\theta} \frac{\partial p_\theta}{\partial \theta_i} p_\theta = \frac{\partial}{\partial \theta_i} \int_\Omega p_\theta = \frac{\partial}{\partial \theta_i} 1 = 0.$$

So, the first meaningful tensor is g.

For the statistical invariance we were hinting to, the Fisher-Rao metric g and the Amari Chenstov tensor T are, up to scaling factors, the only reasonable choices and, a posteriori, we will see that for vectors X, Y, Z:

$$T(X, Y, Z) = g(\nabla_X Y - \nabla^*_X Y, Z)$$

for a connection ∇_X.

Let's go back to our previous examples to see the form these objects take.

Example. In the Gaussians example, g is a metric in this space of normal distributions, which, in coordinates, is a metric in the half space. It turns out that this is the hyperbolic metric

$$g = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

with constant curvature equal to -1. So, Gaussians with this natual metric are like a model for the hyperbolic geometry.

Moreover, this metric blows up as $\sigma \uparrow +\infty$ which, geometrically means that the boundary of this half space is being sent to infinity. Namely, we should think of it as a two-dimensional plane with a metric with constant negative curvature, and not as a half space. However, the constant -1 curvature is the geometry induced in terms of the Levi-Civita connection associated to g, but this is not the framework of information geometry, where we should ask about the geometry of the already cited connection ∇, which we shall define. the right connection is something else and also depends on the Amari tensor.

The fact that the mean has no appearance in the metric, is justified by the hyperbolicity in the σ-direction of the half-plane, while it is cylindrical in μ so that, shifting μ is sort of an isometry in terms of the form of the bell-shaped Gaussian.

Example. For the exponential family case,

$$g_{ij} = \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} F(\theta)$$

so the metric is given by the variation of the function F. This is one of the models where the connection ∇ is flat, and its duality can be interpreted as the Legendre duality. The dual model to the exponential family is the so-called *mixture family*, which are those probabilities that are linear: their density is just a linear function.

1.1. **Statistical invariance.** Suppose, now, having a few transformations that we can apply to our model, one of them being to change the sample space or, some mappign

$$\kappa : \Omega \longmapsto \Omega$$

we use to modify all the measure in the model by pushing-forward all of them. Otherwise, can do it stochastically using Markov kernels.

If κ is somehow a bijection, there's essentially no loss of information, as we are just re-parameterizing Ω: thus, a good metric and Amari tensor, sensible of the statistics of the problem, should have the property that if we compute it in the original model or in the re-parameterized model, we get the same object. From the point of view of differential geometry, say Ω is a manifold itself, we pick κ to be a diffeomorphism the we use to change the model by pushing-forward all the measures, then it is easy to check that g and T don't

change. Moreover, they are the only objects invariant under this type of transformation, so they are very, very universal.

There's some precise results: if we consider a tarnsformation where we're losing information, be something that is highly not injective or coarse, then we have an inequality: in any model, we have something more than that.

More precisely, in coordinates, pick a diffeomorphism κ acting on Ω and then use this to parameterize probability measures by means of a push-forward measure

$$\theta \in \Theta \longmapsto \kappa_{\#} p_\theta \in \mathcal{P}(\Omega)$$

where, for every set $A \subset \Omega$, the push-forward measure applied to A is the measure

$$\kappa_{\#} p_\theta(A) = p_\theta(\kappa^{-1}(A))$$

of the pre-image through κ^{-1} of the set A. So, given another transformation of Ω, we can use it to transform any model by pushing-forward.

1.2. f**-divergences.** As S. Amari states, information geometry has emerged from studies of invariant geometrical structure involved in statistical inference. In this section we link together some concepts mentioned above, namely that the choices of g and T are motivated by their property of invariance. Indeed, these metrics can be derived from an f-divergence, which is invariant.

Consider the map we previously defined

$$\Theta \ni \theta \longrightarrow p_\theta \in M$$

and a function f convex, positive, and such that $f(1) = 0$. Then, the *f-divergence* between two maps p_θ and $p_{\theta'}$ is:

$$D_f(p_\theta, p_{\theta'}) = \int_\Omega f\Big(\frac{p_\theta}{p_{\theta'}}\Big) p_{\theta'}\, d\mu$$

We want to look at a curve γ parametrized by t on the parameter space Θ and consider $\theta = \theta(t)$ and an increment along the curve $\theta' = \theta(t+h)$. We can compute the divergence between the distributions at these two values, namely $D_f(p_\theta, p_{\theta'})$.
Our question is: what happens if the increment h is very very little? That is, what happens if we Taylor expand the divergence D_f?
Of course, the 0-th order term is 0. It turns out that the first-order term is 0 as well, and the first meaningful terms are the second-order term, that depends on the Fisher information g and on f only by a number, and the third-order term that depends on the 3-tensor T.
Let us compute this surprising result. Looking at the first-order term we have:

$$\frac{d}{dh} D_f(p_{\theta(t)}, p_{\theta(t+h)}) = \int \frac{d}{dh} p_{\theta(t+h)} f\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big)\, d\mu - \int p_{\theta(t+h)} f'\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p_{\theta(t)}}{{p^2}_{\theta(t+h)}} \frac{d}{dh} p_{\theta(t+h)}\, d\mu$$

Now, evaluating this at $h = 0$, in the second term we get $f'(1)$ evaluated at h, which is 0, bringing the whole term to 0; in the first term we exchange the integral and the derivative and get $\frac{d}{dh}\int p_{\theta(t)}\, d\mu = 0$ as the integral of a density function is equal to 1. Hence, the whole term goes to 0.
Let us look at the second-order term:

$$\begin{aligned}
\frac{d^2}{d^2h} D_f(p_{\theta(t)}, p_{\theta(t+h)}) &= \frac{d}{dh} \int \frac{d}{dh} p_{\theta(t+h)} f\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big)\, d\mu - \frac{d}{dh} \int f'\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p_{\theta(t)}}{p_{\theta(t+h)}} \frac{d}{dh} p_{\theta(t+h)}\, d\mu \\
&= -\int \Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 f'\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p_{\theta(t)}}{{p^2}_{\theta(t+h)}}\, d\mu \\
&+ \int \Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 \left[f''\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p_{\theta(t)}}{p_{\theta(t+h)}} \frac{p_{\theta(t)}}{{p^2}_{\theta(t+h)}} \frac{d}{dh} p_{\theta(t+h)} + f'\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p_{\theta(t)}}{{p^2}_{\theta(t+h)}} \frac{d}{dh} p_{\theta(t+h)} \right] d \\
&= \int \Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 f''\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{{p^2}_{\theta(t)}}{{p^3}_{\theta(t+h)}}\, d\mu
\end{aligned}$$

Where between the first and second line we evaluate the two derivatives of a product only when the derivative doesn't fall at $\frac{d}{dh} p_{\theta(t)}$ because we already now it goes to 0 for $h = 0$; then, between the second and forth line we have a nice cancellation and we are left only with one term of the sum. Evaluating this at

$h = 0$, we are left with:

$$\begin{aligned}\frac{d^2}{d^2h} D_f(p_{\theta(t)}, p_{\theta(t+h)})\Big|_{h=0} &= f''(1) \int \Big(\frac{d}{dh} p_{\theta(t)}\Big)^2 \frac{1}{p_{\theta(t)}}\, d\mu \\ &= f''(1) \int \frac{\frac{d}{dt} p_{\theta(t)}}{p_{\theta(t)}} \frac{\frac{d}{dt} p_{\theta(t)}}{p_{\theta(t)}} p_{\theta(t)} d\mu \\ &= f''(1) \int \frac{d}{dt} \log[p_{\theta(t)}] \frac{d}{dt} \log[p_{\theta(t)}]\, p_{\theta(t)} d\mu \\ &= f''(1)\, g(\gamma^{\cdot}, \gamma^{\cdot})\end{aligned}$$

Where at this point we can equivalently evaluate the derivative at t, we multiply above and below by $p_{\theta(t)}$ and use the derivative of the log $\frac{d}{dt} \log p_{\theta(t)} = \frac{1}{p_{\theta(t)}} \frac{d}{dt} p_{\theta(t)}$. We get a term depending only on a constant $f(1)$ and the Fisher information g evaluated at a very specific derivative, which is the derivative along the curve, which is nothing but the velocity of the curve.

Now, turn to the third-order term, we have to take the derivative of what we found above:

$$\begin{aligned}\frac{d^3}{d^3h} D_f(p_{\theta(t)}, p_{\theta(t+h)}) &= \frac{d}{dh} \int \Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 f''\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p^2{}_{\theta(t)}}{p^3{}_{\theta(t+h)}}\, d\mu \\ &= \int \frac{d}{dh} \left[\Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 f''\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p^2{}_{\theta(t)}}{p^3{}_{\theta(t+h)}} \right] d\mu \\ &= \int \frac{d}{dh} \Big[f''\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p^2{}_{\theta(t)}}{p^3{}_{\theta(t+h)}} \Big] \Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 + \frac{d}{dh} \Big[\Big(\frac{d}{dh} p_{\theta(t+h)}\Big)^2 \Big] f''\Big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\Big) \frac{p^2{}_{\theta(t)}}{p^3{}_{\theta(t+h)}}\, d\mu\end{aligned}$$

Let's compute the first term, when the derivative falls on $f''\big(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\big) \frac{p^2{}_{\theta(t)}}{p^3{}_{\theta(t+h)}}$.

$$\int \left(\frac{d}{dh} p_{\theta(t+h)}\right)^2 \left[-f'''\left(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\right) \frac{p^3_{\theta(t)}}{p^5_{\theta(t+h)}} \frac{d}{dh} p_{\theta(t+h)} - 3f''\left(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\right) \frac{p^2_{\theta(t)}}{p^4_{\theta(t+h)}} \frac{d}{dh} p_{\theta(t+h)} \right] d\mu$$

Evaluating it at $h = 0$:

$$\begin{aligned}&\int \Big(\frac{d}{dt} p_{\theta(t)}\Big)^3 \Big(-f'''(1) \frac{1}{p^2{}_{\theta(t)}} - 3f''(1) \frac{1}{p^2{}_{\theta(t)}} \Big)\, d\mu \\ &= -\big(f'''(1) + 3f''(1)\big) \int \Big(\frac{d}{dt} p_{\theta(t)}\Big)^3 \frac{1}{p^2{}_{\theta(t)}}\, d\mu\end{aligned}$$

And using the same trick as above we get:

$$\begin{aligned}&= -\big(f'''(1) + 3f''(1)\big) \int \frac{d}{dt} \log p_{\theta(t)} \frac{d}{dt} \log p_{\theta(t)} \frac{d}{dt} \log p_{\theta(t)}\ p_{\theta(t)}\, d\mu \\ &= -\big(f'''(1) + 3f''(1)\big)\, T(\gamma^{\cdot}, \gamma^{\cdot}, \gamma^{\cdot})\end{aligned}$$

i.e., we get exactly a term depending on the Amari-Chentsov tensor evaluated at $\gamma^{\cdot}$

Now, let's go back and analyze the case when the derivative falls on $\big(\frac{d}{dh} p_{\theta(t+h)}\big)^2$:

$$\int \frac{d}{dh} \left(\frac{d}{dh} p_{\theta(t+h)}\right)^2 f''\left(\frac{p_{\theta(t)}}{p_{\theta(t+h)}}\right) \frac{p^2{}_{\theta(t)}}{p^3{}_{\theta(t+h)}}\, d\mu$$

Evaluating this term at $h = 0$ we get:

$$= 2 \int \frac{d}{dt} p_{\theta(t)} \frac{d^2}{d^2 t} p_{\theta(t)} f''(1) \frac{1}{p_{\theta(t)}} \, d\mu$$
$$= 2 f''(1) \int \frac{d}{dt} \log p_{\theta(t)} \frac{d^2}{d^2 t} p_{\theta(t)} \, d\mu$$

Aaaaaand bo :) [...]

Overall, the cool result we get is that Taylor expanding the f-divergence along a curve γ on the parameter space Θ, the first significant terms we get are exactly the Fisher-Rao metric g and the Amari-Chentsov T evaluated at the derivative along the curve.

2. Fundamental concepts in geometry

· This is the 2nd reading seminar of the series, held by Prof. Pigati on April 1st ·

What follows is a set of informal definitions of objects like manifolds, charts and tangent space that are at the basis of the geometric framework introduced above.

Given $n \in \mathbb{N}$, a n-dimensional *manifold* $\mathcal{M}^n$ is, roughly, something that looks like the Euclidean space $\mathbb{R}^n$ at small scale. When defining manifolds, we don't actually use a notion of distance (metric), instead we use topological notions and define a "small scale" in terms of open sets.

Definition 2.1 (Manifold)**.** Given a topological space M^n (Hausdorff*, second countable), a structure of an n-dimensional manifold is given by:

i. a collection $\{U_i\}_{i\in I}$ of open sets $U_i \subseteq M$ such that $M = \bigcup_{i\in I} U_i$;
ii. for each i, a bijective map $\varphi_i : U_i \longrightarrow V_i \subseteq \mathbb{R}^n$;
iii. $\varphi_i \circ \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \longrightarrow \varphi_i(U_i \cap U_j)$ is C^∞ *(compatibility property)*

Where you can think of the property of being "Hausdorff" as the points in the space being distinguishable; and the compatibility property ensures we can define what is a smooth function on our manifold.

Example. Take $U \subset \mathbb{R}^n$ open. $\mathcal{M}^n := U$ is a n-dimensional manifold with open cover $\{U\}$ and chart $\{id : U \longrightarrow U\}$.

Example. Take the unitary n-dimensional sphere $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$. Given $p \in S^n$, we define the emisphere related to p by cutting the sphere in half with the hyperplane generated by the vector p (the hyperplane orthogonal to it) where the origin 0 lies, i.e. $U_p := \{x \in S^n : (x, p) > 0\}$. Then, $\varphi_p : U_p \longrightarrow B_1(0)$ is a projection that "flattens" the emisphere onto the hyperplane.

Example (Guess the manifold). Take $M^2 = U_1 \cup U_2$.
$\varphi_1 : U_1 \longrightarrow \mathbb{R} \times (0, 2\pi)$
$\varphi_2 : U_2 \longrightarrow \mathbb{R} \times (0, 2\pi)$
$\varphi_1(U_1 \cap U_2) = \mathbb{R} \times (\pi, 2\pi)$
$\varphi_2 \circ \varphi_1^{-1}(x, y) := (x, y - \pi)$

Solution: We can deduce that our manifold M is a cylinder. The charts φ_1 and φ_2 will be maps from the manifold to the cylinder minus one line (to ensure bijection).

Example (Guess the manifold). Take $M^2 = U_1 \cup U_2$.
$\varphi_1 : U_1 \longrightarrow \mathbb{R}^2$
$\varphi_2 : U_2 \longrightarrow \mathbb{R}^2$
$\varphi_1(U_1 \cap U_2) = \varphi_2(U_1 \cap U_2) = \mathbb{R}^2 \backslash \{0\}$
$\varphi_2 \circ \varphi_1^{-1}(x) = \frac{x}{|x|^2}$

Solution: We can deduce that our manifold is the 2-dimensional sphere S^2, where $\varphi_1$is given by the sphere without the point at the north pole, and φ_2 is given by removing the south pole. The two maps project the manifold onto the plane cutting it in half (the plane where the equator lies).

Definition 2.2 (Charts)**.** The maps $\varphi_i : U_i \longrightarrow V_i$ used in definition 2.1 are called charts. The inverse maps φ_i^{-1} are called parametrizations.

The main conceptual idea is that for any point $x \in M^n$ in the manifold, you can find an open set U_i that is omeomorphic to an open subset of the Euclidean space $\mathbb{R}^n$. Figure 1 helps visualizing this idea. Then, we can think of charts as maps allowing us to find (multiple sets of) "coordinates" for points on the manifold.

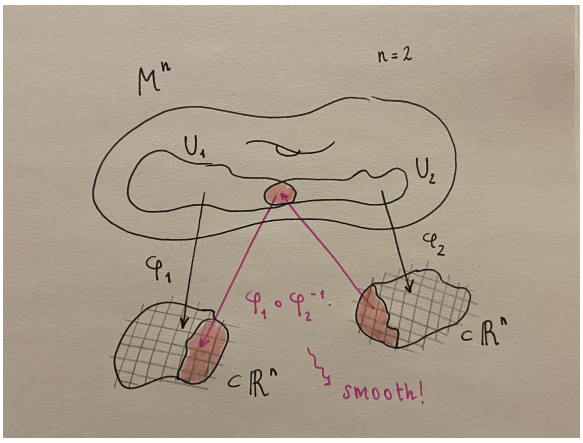

FIGURE 1. Visual representation of a manifold. The two sets in the lower part are subsets of the Euclidean space, i.e. charts. What you would like is that the pink map ($\varphi_2 \circ \varphi_1^{-1}$) is smooth, so that the two coordinate system defined on the pink region of the manifold ($U_1 \cap U_2$) are compatible.

Indeed, the inverse of a chart is a parametrization in the sense that it maps some parameters - vectors in $\mathbb{R}^n$ - to points in our manifold. What we hope is that the combination of charts defined on the same region on the manifold is smooth (compatibility property), in order to define a *smooth* manifold (we could hope for any other property like continuity or C^7, and to enforce it we would impose it on charts in the same way).

Definition 2.3 (Atlas and maximal atlas)**.** A collection of charts is called an *atlas*. Any atlas can be uniquely extended to a maximal atlas, which is one that is not contained in any other atlas.

Remark. Recall the examples above: the two structures we have defined on the manifold S^n are the same, i.e. they give the same maximal atlas.

Example. Here's instead an example of incompatible charts. Take $M = \mathbb{R}$ and the charts $id : M \longrightarrow \mathbb{R}$ and $\psi : M \longrightarrow \mathbb{R} : \psi(x) = x^3$. Then we get $id \circ \psi^{-1}(x) = \sqrt[3]{x} \notin C^\infty$.

Definition 2.4 (Smoothness)**.** Let M, N be two manifolds such that:

M has atlas $\{\varphi_i\}_{i\in I} : U_i \longrightarrow \mathbb{R}^n$
N has atlas $\{\psi_j\}_{j\in J} : V_j \longrightarrow \mathbb{R}^n$

Given $p \in M$, $f : M \longrightarrow N$ is smooth if $\exists U_i \ni p$ and $\exists W_j \ni f(p), W_j \subseteq \mathbb{R}^n$, such that the map

$$\psi_j \circ f \circ \varphi_i^{-1} : \varphi_i(f^{-1}(W_j)) \longrightarrow W_j$$

is smooth.
Moreover, using compatibility of charts we can replace "$\exists$" with "$\forall$".

Now, we turn to defining what a tangent space is. approximated to a tangent space.

2.1. **Tangent spaces.** We can get an intuition of *tangent spaces* by means of a useful analogy. Just like we can approximate a smooth function to a linear one, by taking its Taylor expansion, any manifold can be approximated to a tangent space.

Let $p \in M^n$. If $n = 0$ we simply define $T_pM \coloneqq \{0\}$. Now, assume $n > 0$. We give two possible definitions of the tangent space at p and we will later comment on their equivalence. The first one is more visual, you think of vectors in the tangent space as velocities of curves starting at p. The second approach is more algebraic, you think of the tangent space as all the possible ways to perform a directional derivative of a function.

First approach (visual): Consider the space of all smooth curves $\gamma : [0, \varepsilon] \to M$, $\gamma(0) = p$ starting from p. We declare two such curves equivalent if their velocity is equal when read in every chart. Formally, for any $\gamma_1 : [0, \varepsilon_1] \to M$ and $\gamma_2 : [0, \varepsilon_2] \to M$, $\gamma_1 \sim \gamma_2$ if for any chart $\varphi : U \to V \subseteq \mathbb{R}^n$ around p, it holds $(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0)$. Then T_pM will be defined as the set of all equivalence classes.

A few remarks are in order.

Remark. It is enough to check the condition in the definition only for one chart. Indeed, when changing chart, the velocity is transformed by applying the differential of the change of chart map (which we recall to be smooth by the definition of smooth manifold).

Remark. T_pM is indeed a vector space and its dimension is the same as the one of the manifold. To see this, fix any chart centered at p (we can always choose such a chart up to translation and moreover the choice is irrelevant by the previous remark) $\varphi : U \to V$ with $p \in U$ and consider for any $w \in \mathbb{R}^n$ the curve $\gamma_w(t) \coloneqq \varphi^{-1}(tw)$, that is, the preimage through the chart of the line spanned by w in $\mathbb{R}^n$. One can show that the map $w \mapsto [\gamma_w]$ is a bijection between $\mathbb{R}^n$ and T_pM. The sum operation on the equivalence classes is then defined by $[\gamma_w] + [\gamma_v] \coloneqq [\gamma_{w+v}]$. The vector space structure hereby defined does not depend on the choice of chart.

Remark. When the manifold is immersed in some Euclidean space $\mathbb{R}^n$, we can view T_pM as a vector subspace. The affine subspace $p + T_pM$ corresponds to our intuition of a "Taylor approximation" of M. For a surface (2-dimensional manifold) in $\mathbb{R}^3$ that would be the tangent plane to the surface.

Second approach (algebraic) : Consider the space of real valued smooth functions on M, denoted $C^\infty(M)$. We call a map $D : C^\infty(M) \to \mathbb{R}$ a *derivation* at p if it is linear and it satisfies the following product rule: $D(fg) = D(f)g(p) + f(p)D(g)$. Derivations are essentially directional derivatives of functions, evaluated at p. We define the tangent space as the set of all derivations at p, that is, all possible ways of performing a directional derivative of a function.

Remark. The vector space structure in this second definition is clearer (sums and scalar multiples of derivations are still derivations) but proving that it has dimension n still requires some work. A possible strategy would be proving that, given a derivation D at p and a chart $\varphi : U \to V$ centered at p, there exists a vector $w \in \mathbb{R}^n$ such that $D(f)$ is the directional derivative of $f \circ \varphi^{-1}$ in direction w evaluated at 0. The proof of this fact uses the following factorization lemma:

Lemma 2.5. For any $f \in C^\infty(M)$, $f(x) = f(0) + \sum_{i=1}^n x_i g_i(x)$ for some smooth functions $\{g_i\}$.

The connection between the two definitions is the following. Given an equaivalence class $[\gamma]$ as in the first definition, we get a derivation D by defining $D(f) \coloneqq (f \circ \gamma)'(0) \in \mathbb{R}$. Using local charts one can check that this is indeed a bijection.

Example. Consider the sphere $S^n \subseteq \mathbb{R}^{n+1}$. It is conveniently defined as the zero set of a smooth function, namely $h(x) = |x|^2 - 1$. In such a case, one can show that the tangent space is the zero set of its differential:

$$T_pS^n = \ker Dh(p) = \ker(v \mapsto 2\langle p, v\rangle)$$

Hence, the tangent space to the sphere at p is none other than the orthogonal complement $p^\perp$.

2.2. **Vector bundles.** We now turn to a different construction, ubiquitous in differential geometry, which is the one of vector bundles. Informally, they provide a smooth way to glue together different vector spaces at each point of a manifold.

Definition 2.6 (Vector bundle)**.** Let M be an n-dimensional manifold and k a positive integer. A *vector bundle* of rank k over M is a manifold E^{n+k} together with a smooth map $\pi : E \to M$, called projection, such that:

(1) each fiber $E_p := \pi^{-1}(p)$ has the structure of a k-dimensional vector space
(2) for each $p \in M$ there exists a neighborhood U of p and a diffeomorphism $\Phi : \pi^{-1}(U) \to \mathbb{R}^k \times U$, called *local trivialization*, of the form $\Phi = (F, \pi)$, where F is a vector space isomorphism with $\mathbb{R}^k$ when restricted to each fiber.

Notice that in the definition above, we assumed a smooth structure on the total space E as a datum. However, the way vector bundles arise in practice is from disjoint unions of vector spaces assigned to each point of the manifold. With this in mind, let us see a more operational, yet equivalent, characterization of vector bundles which gives a smooth structure on the total space almost for free, provided that the local trivializations overlap in a specific way.

Proposition 2.7. Suppose that for each $p \in M$ we have a real vector space E_p of some fixed dimension k and let $E := \bigsqcup_{p\in M} E_p$. If moreover we have an open cover $\{U_\alpha\}_{\alpha\in A}$ of M and a bijective map $\Phi_\alpha : \pi^{-1}(U_\alpha) \to \mathbb{R}^k \times U_\alpha$ of the same form as in Definition 2.6 (ii) for each $\alpha \in A$ and if it holds that for each $\alpha, \beta \in A$ with $U_\alpha \cap U_\beta \neq \varnothing$, the map $\Phi_\alpha \circ \Phi_\beta^{-1}$ from $\mathbb{R}^k \times (U_\alpha \cap U_\beta)$ to itself has the form

$$\Phi_\alpha \circ \Phi_\beta^{-1}(v, p) = (L_p v, p)$$

for some $L_p \in GL(k, \mathbb{R})$ depending smoothly on p, then E has a unique topology and smooth structure making it into a smooth manifold and a smooth rank-k vector bundle over M, with the obvious projection π and $\{\Phi_\alpha\}$ as smooth local trivializations.

Example (Tangent Bundle). Denote $TM = \bigsqcup_{p\in M} T_pM$ and by π the obvious projection. We want to put a structure of vector bundle of rank $2n$ on TM which will then take the name of *tangent bundle.* This is the prototypical example of a situation where we want to glue together different vector spaces, in this case the tangent spaces, at each point of M. In principle, we should construct a topology and a differentiable structure on TM, but using the equivalent characterization we just gave, we can bypass this step and directly provide the local trivializations and transition functions, which incidentally are fairly simple in this case.

Take a smooth atlas of M (to fix ideas you can take the maximal one) and in the domain of each chart define the map $\Phi : \pi^{-1}(U) \to \mathbb{R}^n \times U$ by

$$\Phi\left(\sum_i v^i \left.\frac{\partial}{\partial x^i}\right|_p\right) := ((v^1, \dots, v^n), p)$$

where $\{ \left.\frac{\partial}{\partial x^i}\right|_p \}$ is the basis of T_pM induced by the coordinates of the chart. This is clearly bijective and restricts to a linear isomorphism with $\mathbb{R}^n$ on fibers. Finally, one can easily check that for any two such maps the transition function is given by the differential of the change of coordinates, which is indeed an invertible linear map:

$$\Phi_\alpha \circ \Phi_\beta^{-1}(v, p) = (D(\varphi_\alpha \circ \varphi_\beta^{-1})(x)[v], p).$$

We can define a suitable notion of *isomorphism of vector bundles*

Definition 2.8. We call two vector bundles E, E' isomorphic if there exists a diffeomorphism $F : E \to E'$ that respects the fibers, that is E_p goes to $E'_{F(p)}$, or more compactly $\pi_{E'} \circ F = \pi_E$, and gives a linear isomorphism on each fiber. Moreover we say that E is *trivial* if is isomorphic to $\mathbb{R}^k \times M$.

Example (TS^1 is trivial). Recalling that the tangent space to a point on the sphere is its orthogonal complement we define the vector bundle isomorphism $F : \mathbb{R} \times S^1 \to TS^1$ by $F(t, p) = tRp \in T_pS^1$ where R is a counterclockwise 90 degrees rotation.

Remark. In light of this, we may hypothesize that the tangent bundle of any sphere is trivial. It turns out that it is not true and this fails already in dimension 2. However, proving that TS^2 is not trivial requires a very deep theorem in differential topology, called the *hairy ball theorem* which states that there is no nonvanishing smooth (actually continuous) tangent *vector field* on even-dimensional n-spheres. The connection with triviality of vector bundles comes from the fact that being trivial is equivalent to having a smoothly varying choice of basis of each fiber. If this were the case on the $2-$sphere we would have a smooth nonvanishing vector field on it, thus contradicting the hairy ball theorem.

In full generality, TS^n is trivial if and only if $n \in \{1, 3, 7\}$

Example (TS^3 is trivial). To show this, it is convenient to view S^3 as a subset of the quaternions: $S^3 := \{a + ib + jc + kd | a^2 + b^2 + c^2 + d^2 = 1\}$. The map $F : \mathbb{R}^3 \times S^3 \to TS^3$ defined by

$$F((t_1, t_2, t_3), p) = t_1pi + t_2pj + t_3pk \in T_pS^3$$

for example, is a vector bundle isomorphism.

3. Riemannian geometry

Seminar 15 apr, Prof. Pigati. [Next meeting is May 13, then May 27]
Connections: a way to perform derivatives on a manifold

Definition 3.1 (Section of a map)**.** A smooth section $s : M \to E$ is a smooth map such that $s(p) \in \pi^{-1}(p)\ \forall p \in M$

When $E = TM$, a section is called a vector field. (Recall E is our vector bundle with associated map π, and $X = \sum_{i=1}^{n} X^i \frac{\partial}{\partial x_i}$.

Definition 3.2 (Connection)**.** A connection on E, denoted ∇, takes a section s of E and a vector field X, and gives a new section $\nabla_X s$ that satisfies the following properties:

- $C^\infty(M)$-linear in X: $\nabla_{fX+gY} s = f\nabla_X s + g\nabla_Y s \quad \forall f, g : M \to \mathbb{R}$
- $\mathbb{R}$-linear in s
- Leibniz rule: $\nabla_X(fs) = f\nabla_x s + X(f)s$

Where $X(f)$ indicates differentiating f along direction X.

Remark. $\nabla_X s(p)$ depends only on $s|_U$ for an arbitrary small neighborhood U and on $X(p)$.

Remark. In a trivialization $\pi^{-1} \cong \mathbb{R}^k \times U$ we always have the following structure:

$$s(x) = (v(x), x)$$

$$\nabla_X s(x) = Ds(x)[X(x)] + X^i A_i(x) s(x)$$

Where A is a $k \times k$ matrix, not necessarily invertible.

Remark. If ∇ is a connection, any other connection has the form $\tilde{\nabla} = \nabla + \alpha$ where $\alpha(x) : \pi^{-1}(x) \otimes T^*_x M \to \pi^{-1}(x)$ is a linear map. Namely, $\tilde{\nabla}_X s = \nabla_X s + \alpha(X)s$, where $\alpha(X)s$ is the "0-th order term".

Hence, the space of all connections is an **affine space**.
In computations, assuming U trivializes E and is included in a chart, we can write

$$\nabla_{\frac{\partial}{\partial x_i}} e_j = \sum_{l=1}^{\text{rank}} \Gamma^l_{ij} e_l$$

where $\{e_1, \dots, e_k\}$ is a canonical basis of $\mathbb{R}^k \cong \pi^{-1}(x)$. $\Gamma^l_{ij}(x)$ are called *Christoffel symbols.*
$A_i(x)$ has coefficient Γ^l_{ij} in position (l, j).

$$\nabla_{\frac{\partial}{\partial x^i}} \begin{pmatrix} c_1(x) \\ \vdots \\ c_k(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial c_1}{\partial x^i} \\ \vdots \\ \frac{\partial c_k}{\partial x^i} \end{pmatrix} + A_i \begin{pmatrix} c_1(x) \\ \vdots \\ c_k(x) \end{pmatrix}.$$

Remark. For $E = TM$ we just need to take U=domain of chart. Then $TM|_U$ is trivialized by $c_1(x)\frac{\partial}{\partial x^1} + \dots + c_n(x)\frac{\partial}{\partial x^n}$ and $\nabla_{\frac{\partial}{\partial x^i}}(\frac{\partial}{\partial x^j}) = \sum_{l=1}^{n} \Gamma^l_{ij} \frac{\partial}{\partial x^l}$

3.1. Parallel Transport.

Given a smooth curve $\gamma : [0, T] \to M$, a section along γ is $s : [0, T] \to E$ (smooth) s.t. $s(t) \in \pi^{-1}(\gamma(t))$

Example. Given any $v_0 \in \pi^{-1}(\gamma(0)$ there is a unique section $s(t)$ along γ s.t. $\nabla_{\gamma'(t)} s(t) = 0$. What does $\nabla_{\gamma'(t)}$ mean? In Rimeannian geometry, it is called the *covariant derivative* along a curve γ.

Locally in t we can always write

$$s(t) = \sum_m c_m(t) s_m(\gamma(t))$$

with s_m section in the usual sense, define near $\gamma(t)$, and we let

$$\nabla_{\gamma'(t)} := \sum c_m(t) \nabla_{\gamma'(t)} s_m + \sum c'_m(t) s_m(\gamma(t))$$

What we get $v_0(s(0) \mapsto s(T))$ is called *parallel transport* along γ and gives an isomorphism $\pi^{-1}(\gamma(0)) \xrightarrow{\sim} \pi^{-1}(\gamma(T))$. However, two curves will produce different isomorphisms $\pi^{-1}(p) \to \pi^{-1}(q)$.

What is the *curvature* of a connection? It measure the failure of parallel transport along loops to be an identity.

$$F_\nabla\Big(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\Big) s := \nabla_{\frac{\partial}{\partial x^i}}(\nabla_{\frac{\partial}{\partial x^j}} s) - \nabla_{\frac{\partial}{\partial x^j}}(\nabla_{\frac{\partial}{\partial x^i}} s) = B_{ij} s$$

is a 0-order thing. Hence, parallel transport = Id $+ \varepsilon^2 \cdot F_\nabla\big(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\big)$

3.2. **Riemannian manifold.** A Riemannian manifold (M, g) is a manifold endowed with a positive definite scalar product $g(p)$ on each T_pM, depending smoothly on p.

Example. IF $M \subseteq \mathbb{R}^N$, we can take $g(p) :=$ restriction of the inner product in $\mathbb{R}^N$, e.g. $S^N \subset \mathbb{R}^{N+1}$

Theorem 3.3. On TM, connection, called *Levi-Civita connection*, s.t.

- $X(g(Y, Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$

The Riemannian tensor on (M, g) is defined as

$$R(X, Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z$$

and is a section of $TM \otimes T^*M \otimes T^*M \otimes T^*M$.

[...]

[Missing lecture May 13]

Recall that our framework is that of (M, g, ∇)

[two theorems explained last time] [Fenkel duality explained last time]

Definition 3.4 (Bregman divergence)**.** Consider $p, q \in M$, $\theta p = \theta(p)$, $\theta q = \theta(q)$, and a strongly-convex function f. Then we define Bregman divergence as

$$D_f(p, q) := f(\theta p) - f(\theta q) - \nabla f(\theta p)(\theta p - \theta q)$$

Remark. (i) $D_f(p, q) \geqslant 0$ $\forall p, q \in M$ and $D_g = 0$ if and only if $p = q$;
(ii) D_f is not symmetric.

Analogously, we can define the *dual divergence* as

$$D_{f*}(p, q) := f^*(\theta^* p) - f^*(\theta^* q) - \nabla^* f^*(\theta^* p)(\theta^* p - \theta^* q)$$

Proposition 3.5. $D_f(p, q) = D_{f*}(q, p) \quad \forall p, q \in M$

Proof. Notice that $\nabla f(\theta q) = \theta^* q$ by definition of θ^* (Definition **??**).

$$\begin{aligned} D_f(p, q) &= f(\theta p) - f(\theta q) - \nabla f(\theta p)(\theta p - \theta q) \\ &= f(\theta p) - f(\theta q) + \theta q \theta^* q - \theta^* q \theta p \\ &= f(\theta p) + f^*(\theta^* q) - \theta^* q \theta p \end{aligned}$$

Where the last equality comes from $-f(\theta q) + \theta q \theta^* q = f^*(\theta^* q)$ by Definition **??**. By playing the same game with $D_{f*}(p, q)$ we get

$$D_{f*}(p, q) = f^*(\theta^* p) + f(\theta q) - \theta q \theta^* p$$

by using the third point in Definition **??** and that the * operation is an involution. The result follows. □

3.2.1. *Generalized Pythagorean Theorem.* Consider3 point $p, q, r \in M$ and 3 curves connecting them. We look at the remainder of the divergence between the 3 in a cycle, in the same fashion of what the Pythagorean theorem does with Euclidean distances. Indeed, we can think of the divergence as a generalized *squared* distance.

$$\begin{aligned} D_f(p, q) + D_f(q, r) - D_f(p, r) =& \big[f(\theta_p) - f(\theta_q) - \theta_q^*(\theta_p - \theta_q)\big] + \big[f(\theta_q) - f(\theta_r) - \theta_r^*(\theta_q - \theta_r)\big] \\ &- \big[f(\theta_p) - f(\theta_r) - \theta_r^*(\theta_p - \theta_r)\big] \\ =& \theta_r^*(\theta_p - \theta_r) - \theta_q^*(\theta_p - \theta_q) \\ =& (\theta_r^* - \theta_q^*)(\theta_p - \theta_q) \end{aligned}$$

Hence, the remainder of this commutator is the inner product between $(\theta_r^* - \theta_q^*)$ and $(\theta_p - \theta_q)$. Let us make sense of this number.

Consider the ∇-geodesic connecting p and q:

$$\gamma_{p,q}(t) = \theta p(t) + t\theta q$$

and the ∇^*-geodesic connecting q and r:

$$\gamma_{q,r}(t) = \theta^* q(1 - t) + t\theta^* r$$

Now, velocities of the two curves are explicit:

$$\dot{\gamma}_{p,q}(t) = \sum_i (\theta q - \theta p)_i \frac{\partial}{\partial \theta_i}$$

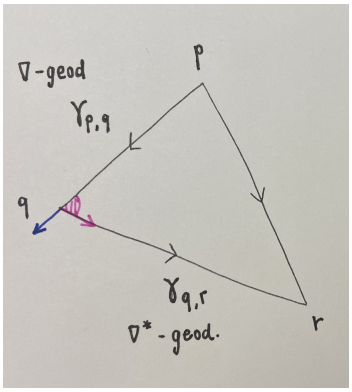

FIGURE 2.

$$\dot{\gamma}_{q,r} = \sum_i (\theta^* r - \theta^* q)_i \frac{\partial}{\partial \theta_i^*}$$

and similarly for $\dot{\gamma}_{q,r}$.

By inspecting Figure 2, we see that the pink angle between the two geodesics is the inner product between (minus) the velocity of $\gamma_{p,q}$ at $t = 1$ (pink arrow) and the velocity of $\gamma_{q,r}$ at $t = 0$ (blue arrow), i.e.

$$g\Big(-\sum_i (\theta q - \theta p)_i \frac{\partial}{\partial \theta_i}, \sum_j (\theta^* r - \theta^* q)_j \frac{\partial}{\partial \theta_j^*}\Big) = (\theta_p - \theta_q)(\theta_r^* - \theta_q^*)$$

Where the result is just the Euclidean inner product between the two quantities. Hence, if $\dot{\gamma}_{p,q} \perp \dot{\gamma}_{q,r}$, their inner product equals 0.

The result is that if the ∇-geodesic $\gamma_{p,q}$ is orthogonal to the ∇^*-geodesic $\gamma_{q,r}$, then:

$$D_f(p,q) + D_f(q,r) - D_f(p,r) = (\theta_p - \theta_q)(\theta_r^* - \theta_q^*) = 0$$

Implying:

$$D_f(p,q) + D_f(q,r) = D_f(p,r) \tag{1}$$

Remark. We can state a more general statement. Indeed, to get result 1 we just need that the velocities of the curves connecting p and r to q are orthogonal at q (at $t = 1$ and $t = 0$ respectively). This means that any two curves connecting p to q and r to q, that have the same first-order behavior of the ∇ and ∇^* geodesics at $t = 1$ and $t = 0$ respectively, are suitable.

This result is widely used in optimization, to characterize minimizers in terms of velocity of geodesics, for instance minimizing the entropy as a specific kind of divergence.

[parte di ottimizzazione]

[parte degli esponenziali]