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" **?**. The language of information geometry is the same as that of differential geometry, but the purpose is to study the intersection 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 a smooth manifold, usually finite-dimensional, 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. The additional object, T, 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*, indicated as ∇. We build ∇ starting from g and T, and then we find its dual connection ∇^*, which is a well-defined connection. We can check duality by verifying that taking the dual is an involution, i.e.

$$\nabla^{**} = \nabla$$

always holds. When ∇ is flat, this duality is exactly the well-known Legendre duality from convex analysis.

In Riemannian geometry, given a metric, we can always find a special type of connection, the Levi-Civita connection ∇^{LC}, which induces the parallel transport, the geodesics, and so on ... and the study of the Levi-Civita connection is fundamental to Riemannian geometry. In our framework, the equality

$$\nabla = \nabla^*$$

holds if and only if the connection we started with is exactly the Levi-Civita connection associated with g, $\nabla = \nabla^{LC}$. In this case, ∇ is self-adjoint. In general, ∇ is not Levi-Civita.

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 universal objects induced by *Fisher information*. Recall the definition of Fisher information.

Let's consider the standard framework in parametric statistics: pick a sample space Ω, with a sigma algebra $\mathscr{F}$, then the pair $(\Omega, \mathscr{F})$ is our *sample* or *state space*. 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, in particular they are smooth finite-dimensional submanifolds, which we can think of a sphere or embedded in the Euclidean space of dimension n. We use coordinates to parametrize locally the probability measures over $\mathcal{M}$ with parameters $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^n$is a nice subset, such as an open set. Finally, we have the mapping

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

that obtains probability measures that live on our $\mathcal{M}$. We ask for a few good properties of this parameterization: we will assume that this map is 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*.

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

- the manifold of probability measures M is the 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*.

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 $\mathbb{R}$

Obtaining tensors. Now, we need to find the metric g and the tensor T. Actually, g comes directly from the Fisher information and T is a derived measure. We will see that this is essentially the only reasonable choice from the point of view of statistics. Let's start by writing down the tensor objects g and T informally. We assume that, for every θ, our probability P_θ is absolutely continuous with respect to a certain measure μ, i.e.

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

meaning that there is a measure μ dominating all the probability measures P_θ, or at least locally. Therefore, each probability measure has a density $p_\theta \coloneqq \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$, hence

$$P_\theta = p_\theta \mu$$

and we can work with densities from now on.

Now, we define g, which is a metric hence it has two indices:

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

where $\partial_i \log p_\theta$ is the *score*, the derivative of the log-likelihood with respect to θ_i. Therefore, g arises as the *covariance matrix* of the score (up to imposing that the score has 0 expectation). When $i = j$, g is exactly the Fisher information (the variance of the score).[1]

Now, given coordinates $\theta \in \Theta$, it can be checked that this quadratic form g is symmetric and positive definite, i.e. it is a Riemannian metric, so it induces a scalar product, which depends on the point θ we are

[1] Let us try to clarify a potential confusion: the Fisher information is generally defined as the variance of the score, hence g_{ii}, but we usually consider also the Fisher information *matrix*, which is the whole covariance matrix and in that case coincides with g. In one dimension the two objects coincide.

studying. Pick two vectors $v, w \in \mathbb{R}^n$ and think of them as tangent vectors at a given $\theta \in \Theta$, 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 the 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 natural metric are like a model for the hyperbolic geometry.

Moreover, this metric blows up as $\sigma \uparrow +\infty$ that 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 correct connection for our framework is something else and depends also on the Amari tensor.

The fact that the mean has no appearance in the metric is justified by the fact that the half-plane is hyperbolic in the σ-direction and 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.

0.1. **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.