Geometría de la información

Ángel Palacios Polo

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: Eduardo Martínez Fernández

Summary

Information Geometry (IG) is a branch of mathematics that uses differential geometry of manifolds in the field of probability and statistics. We consider a family of probability distributions specified by parameters such that it satisfies certain regularity conditions. These parameters are used as a coordinate system that allows to give the family a manifold structure that we call statistical manifold. In this way each point of the manifold represents a probability distribution. The objective of this work is to study the geometry of these manifolds, namely, metrics and connections taking into account their statistical nature, that is, each point of the manifold represents a probability distribution and show an application.

A metric is used to measure distances between points of a manifold, in our case, between probability distributions. Statistics has developed measures of distancing between probability distributions called divergences. Divergences can be defined in any manifold and are not necessarily metric, since among other things they are not asked for symmetry, but give an idea of the degree of separation between points of the manifold. The divergences are used in hypothesis contrast because their asymmetry allows to capture the fact of protecting more a hypothesis than other alternative. Derivatives of a divergence induce metrics and connections in the manifold. Taking into account the statistical nature of these manifolds and the estimation of parameters we can establish natural conditions to impose on divergences, metrics and connections in statistical manifolds. These conditions are known as monotonicity of information and invariance and are related to the ability of a statistician to take advantage of information that has a sample on an unknown parameter. These conditions restrict the study of divergences, metrics and connections to those that are statistically useful.

The first objective in the studies of the geometry of a statistical manifold is to find a metric that allows us to measure distances between probability distributions. The Fisher information matrix provides a metric in the statistical manifolds, called the Fisher information metric. The existence of a single invariant metric, except multiplicative constant, which is given precisely by Fisher's information metric, is a fundamental result in information geometry. This result shows the importance and singularity of the Fisher information metric, since it is not just an ordinary metric, it is the only reasonable metric in statistical manifolds taking into account their statistical nature.

A connection in a manifold allows to differentiate similarly to $\mathbb{R}^n$. Between the different connections that we can consider there is an important one, the connection of Levi-Civita. This connection always exists and is unique in any manifold equipped with a metric. The Levi-Civita connection is the most used because it has the distance minimization property. However for statistical manifolds, there is a family of invariant connections that depend on a parameter $\alpha \in \mathbb{R}$ called α-connections. If $\alpha = 0$ we obtain the Levi-Civita connection in the statistical manifolds.

An important class of divergences in statistical manifolds is f-divergences, which depend on the choice of a convex function f. The f-divergences are invariant and induce the Fisher information metric. The Kullback-Leibler divergence is a particular case.

One of the applications of information geometry is given in the area of optimization. Given a minimization problem in $\mathbb{R}^n$ we transform it into another equivalent minimization problem in a statistical manifold to choose. We solve the problem obtained by an iterative method based on the gradient descent method. The gradient depends on the chosen metric, which according to the theory will be the metric of

the Fisher information. The proposed method uses the geodesics of the chosen statistical manifold that are usually calculated numerically. In the case of normal distributions, the geodesics can be calculated exactly.

Below, we detail the contents of the chapters. In the first chapter, we present the results of Riemannian geometry needed to study statistical manifolds. We define the concept of tensor, in particular, define the Riemannian metric. A manifold equipped with a Riemannian metric is called Riemannian manifold. In manifolds we do not have a natural notion of derivative, so we need to define the connections that allow to derive tensor fields on manifolds. A connection is said to be isometric if the covariant derivative of the metric is zero. Moreover a connection defines two tensors: torsion and curvature. When the torsion of a connection is zero we say that the connection is simetric. When the curvature of a connection is zero we say that the connection is flat. We prove that there is a unique isometric and symetric connection known as the Levi-Civita connection. This connection allows to minimize distances as already mentioned. Finally we present an important example of Riemanniana manifold, hyperbolic space and give its geodesics.

In the second chapter, we introduce statistical models as families of probability distributions specified by parameters. These parameters are used as a coordinate system and then we have a manifold structure. We define divergences in manifolds and show that derivatives of a divergence induce metrics. Next, we prove that the Fisher information matrix defines Riemannian metric in statistical manifolds. Using statistical arguments we find the conditions of monotonicity of information and invariance. We define a family of divergences in statistical manifolds, the f-divergences, and prove that they are invariant. We prove that Fisher's metric is the only invariant metric, except multiplicative constant. We then calulated the Fisher's metric in some of the most important statistical models as binomial or normal among others. We emphasize that a normal model with multiple covariance matrix of the identity has the hyperbolic space geometry. In the penultimate, section we show that the divergences induce connections and we study with special attention to those induced by f-divergences. Finally we present an application of the theory in the area of optimization showing an iterative algorithm based on the descent of the gradient in statistical manifolds and using Fisher's metric and the geodesics of the chosen statistical manifold. If we choose the normal distributions we get exact solutions for the geodesics.

Índice general

Capítulo 1

Preliminares: geometría Riemanniana

Este capítulo contiene una introducción a la geometría Riemanniana. Consideramos variedades diferenciables de dimensión finita, sin borde y cuya topología sea Hausdorff y segundo numerable. Por «diferenciable» entenderemos siempre infinitamente diferenciable, es decir, de clase C^∞. Utilizaremos el convenio de suma de Einstein que funciona sumando solo en los índices que aparecen simultáneamente como subíndices y superíndices desde 1 hasta la dimensión n.

Sea M una tal variedad de dimensión finita n, $\mathscr{A}$ un atlas para M y $x \in \mathscr{A}$ una carta. Concretamente para cada punto $p \in M$ fijo existe una carta $x : U \subset M \to \mathbb{R}^n$ con U abierto de M, $p \in U$ con $x^i : U \subset M \to \mathbb{R}$, $i = 1, \ldots, n$ que llamamos funciones coordenadas o simplemente coordenadas (locales). Nos referimos a $(x^1, \ldots, x^n)$ como sistema de coordenadas locales. Denotamos por $\mathscr{F}(M) = \{f : M \to \mathbb{R} \mid f \text{ diferenciable}\}$ y por T_pM y T_p^*M al espacio tangente y cotangente de M en un punto $p \in M$ respectivamente. Recordamos que T_pM es un espacio vectorial de dimensión finita n y T_p^*M es su dual. Los elementos de T_pM se llaman vectores tangentes y son aplicaciones con valores reales que actúan sobre funciones de $\mathscr{F}(M)$ verificando linealidad y la regla de Leibniz, es decir, actúan sobre funciones como derivación. Una base de T_pM viene dada por $\{\partial_1 |_p, \ldots, \partial_n |_p\}$ donde $\partial_i |_p = \frac{\partial}{\partial x^i} |_p$. La familia de aplicaciones $\{dx^1 |_p, \ldots, dx^n |_p\}$ de T_p^*M verifican $dx^i |_p (\partial_j |_p) = \delta_{ij}$ donde δ_{ij} es la delta de Kronecker luego $\{dx^1 |_p, \ldots, dx^n |_p\}$ es la base dual de $\{\partial_1 |_p, \ldots, \partial_n |_p\}$ y viceversa. Sea $V \in T_pM$ entonces su expresión local es $V = \sum_{i=1}^n V^i \partial_i |_p = V^i \partial_i |_p$ donde $V^i = V(x^i) \in \mathbb{R}$.

Definición 1.1. Un campo vectorial X en M es una aplicación diferenciable que asigna a cada punto $p \in M$ un vector $X_p \in T_pM$ donde diferenciable significa que si $f \in \mathscr{F}(M)$ entonces $Xf \in \mathscr{F}(M)$. Denotamos por $\mathscr{X}(M)$ al conjunto de todos los campos vectoriales en M.

Si $X \in \mathscr{X}(M)$ entonces definimos $(Xf)_p = X_p f$ y la expresión local de X es $X = \sum_{i=1}^n X^i \partial_i$ donde $X^i = X(x^i) \in \mathscr{F}(M)$. Adoptamos la siguiente notación estándar:

$$(T_p^*M)^r = \underbrace{T_p^*M \times \cdots \times T_p^*M}_{r \text{ veces}}, \quad (T_pM)^s = \underbrace{T_pM \times \cdots \times T_pM}_{s \text{ veces}}, \quad r,s \in \mathbb{N}. \tag{1.1}$$

Definición 1.2. Un tensor de tipo (r,s) en un punto $p \in M$ es una aplicación $\mathbb{R}$-multilineal (lineal en todos sus argumentos) $T : (T_p^*M)^r \times (T_pM)^s \to \mathbb{R}$. Un campo tensorial $\mathscr{T}$ de tipo (r,s) es una aplicación diferenciable que asigna a cada punto $p \in M$ un tensor de tipo (r,s) en el punto p, equivalentemente, es una aplicación $\mathscr{F}(M)$-multilineal $\mathscr{T} : \mathscr{X}(M)^{*r} \times \mathscr{X}(M)^s \to \mathscr{F}(M)$. Un tensor o un campo tensorial de tipo (r,s) diremos que es s-covariante y r-contravariante.

Definición 1.3. Una métrica Riemanniana g en una variedad diferenciable M es un campo tensorial $g : \mathscr{X}(M) \times \mathscr{X}(M) \to \mathscr{F}(M)$ de tipo $(0,2)$ simétrico y definido positivo. Una variedad Riemanniana es una variedad diferenciable dotada de una métrica Riemanniana y la denotamos por (M,g).

Se puede demostrar que toda variedad diferenciable puede dotarse de una métrica Riemanniana. Una métrica Riemanniana g nos proporciona un producto escalar $g_p : T_pM \times T_pM \to \mathbb{R}$ en cada punto p

que depende diferenciablemente del punto $p \in M$. En un sistema de coordenadas locales tenemos

$$g = g_{ij} dx^i \otimes dx^j, \tag{1.2}$$

donde $g_{ij} = g(\partial_i, \partial_j)$ y $\otimes$ indica producto tensorial.

Definición 1.4. El corchete de Lie se define como la aplicación $[\,,\,] : \mathscr{X}(M) \times \mathscr{X}(M) \to \mathscr{X}(M)$ dada por

$$[X,Y]_p f = X_p(Yf) - Y_p(Xf), \ \forall f \in \mathscr{F}(M), p \in M. \tag{1.3}$$

El corchete de Lie satisface las siguientes propiedades:

1. $\mathbb{R}$-bilineal: $[aX+bY,Z] = a[X,Y] + b[Y,Z]$, $[Z,aX+bY] = a[Z,X] + b[Z,Y]$ $\forall a,b \in \mathbb{R}$.
2. Antisimetría: $[X,Y] = -[Y,X]$.
3. Identidad de Jacobi: $[X,[Y,Z]] + [Y,[Z,X]] + [Z,[X,Y]] = 0$.
4. $[fX,gY] = fg[X,Y] + f(Xg)Y - g(Yf)X$, $\forall f,g \in \mathscr{F}(M)$.

El corchete de Lie mide la no conmutatividad entre los flujos de los campos vectoriales. Si $[X,Y] = 0$ entonces los campos vectoriales X,Y conmutan. La expresión del corchete de Lie en coordenadas locales viene dada por

$$[X,Y] = \sum_{i,j=1}^{n} \left(\frac{\partial Y^i}{\partial x^j} X^j - \frac{\partial X^i}{\partial x^j} Y^j\right) \frac{\partial}{\partial x^i}. \tag{1.4}$$

En variedades no tenemos una noción natural de derivada como en $\mathbb{R}^n$. Con el propósito de tener una teoría de diferenciación en variedades similar a la de $\mathbb{R}^n$ se introduce el concepto de conexión, que puede verse como una extensión de la derivada direccional en $\mathbb{R}^n$. Una conexión permite diferenciar funciones, campos vectoriales y en general campos tensoriales respecto de un campo vectorial. Las conexiones dan lugar al transporte paralelo y permiten relacionar la geometría local en distintos puntos de la variedad.

Definición 1.5. Una conexión lineal o simplemente conexión ∇ en M es una aplicación $\nabla : \mathscr{X}(M) \times \mathscr{X}(M) \to \mathscr{X}(M)$ verificando:

1. $\nabla_X Y$ es $\mathscr{F}(M)$-lineal en X.
2. $\nabla_X Y$ es $\mathbb{R}$-lineal en Y.
3. Regla de Leibniz: $\nabla_X(fY) = (Xf)Y + f\nabla_X Y$, $\forall f \in \mathscr{F}(M)$.

La definición anterior es global. La expresión de una conexión en un sistema de coordenadas locales viene dada por

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k, \tag{1.5}$$

donde $\Gamma_{ij}^k \in \mathscr{F}(M)$ se denominan componentes de la conexión respecto de la base local $\{\partial_k\}$. Como $X = X^i \partial_i$, $Y = Y^j \partial_j$ entonces $\nabla_X Y = (\nabla_X Y)^k \partial_k$ donde $(\nabla_X Y)^k = X^i(\partial_i Y^k + Y^j \Gamma_{ij}^k)$. Una conexión permite diferenciar campos tensoriales. Sea T un tensor r-covariante entonces

$$\nabla_X T(Y_1,\ldots,Y_r) = XT(Y_1,\ldots,Y_r) - \sum_{i=1}^{n} T(Y_1,\ldots,\nabla_X Y_i,\ldots,Y_r). \tag{1.6}$$

Sean ∇,∇' dos conexiones entonces para todo $\alpha \in \mathbb{R}$ la combinación convexa $\alpha\nabla + (1-\alpha)\nabla'$ define otra conexión. La diferencia de dos conexiones es un campo tensorial de tipo $(1,2)$.

Definición 1.6. Sea (M,g) una variedad Riemanniana. Una conexión ∇ se dice isométrica si verifica

$$\nabla_Z g = 0, \quad \forall Z \in \mathscr{X}(M). \tag{1.7}$$

Equivalentemente, usando (1.6) y (1.7), una conexión isométrica queda caracterizada por la fórmula

$$Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y), \quad \forall X,Y,Z \in \mathscr{X}(M). \tag{1.8}$$

Eligiendo $X = \partial_i, Y = \partial_j, Z = \partial_k$ y tras unos cálculos la fórmula (1.8) se transforma en

$$\partial_k g_{ij} = \Gamma^p_{ki} g_{pj} + \Gamma^r_{kj} g_{ir}. \tag{1.9}$$

Una conexión define dos campos tensoriales: la torsión y la curvatura.

Definición 1.7. La torsión de una conexión ∇ se define como el campo tensorial $T : \mathscr{X}(M) \times \mathscr{X}(M) \to \mathscr{X}(M)$ de tipo $(1,2)$ dado por

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]. \tag{1.10}$$

La torsión cumple la propiedad antisimétrica $T(X,Y) = -T(Y,X)$. La expresión de la torsión en coordenadas locales viene dada por

$$T_{ij} = T(\partial_i, \partial_j) = \nabla_{\partial_i}\partial_j - \nabla_{\partial_j}\partial_i - [\partial_i, \partial_j] = (\Gamma^k_{ij} - \Gamma^k_{ji})\partial_k, \tag{1.11}$$

ya que $[\partial_i, \partial_j] = 0$. Por tanto las componentes de la torsión son $T^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji}$. Una conexión se dice **libre de torsión** o **simétrica** si $T = 0$ o equivalentemente si $\Gamma^k_{ij} = \Gamma^k_{ji}$.

Definición 1.8. La curvatura de una conexión ∇ se define como el campo tensorial $R : \mathscr{X}(M) \times \mathscr{X}(M) \times \mathscr{X}(M) \to \mathscr{X}(M)$ de tipo $(1,3)$ dado por

$$R(X,Y,Z) = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{1.12}$$

La curvatura también puede escribirse como

$$R(X,Y,Z) = ([\nabla_X, \nabla_Y] - \nabla_{[X,Y]})Z. \tag{1.13}$$

La expresión de la curvatura en coordenadas locales viene dada por

$$R(\partial_i, \partial_j, \partial_k) = R^p_{ijk}\partial_p \quad \text{donde} \quad R^p_{ijk} = \partial_i \Gamma^p_{jk} - \partial_j \Gamma^p_{ik} + \Gamma^p_{ih}\Gamma^h_{jk} - \Gamma^p_{jh}\Gamma^h_{ik}. \tag{1.14}$$

Una conexión simétrica se dice **plana** si $R = 0$. Es evidente que si existe un sistema de coordenadas locales en el que las componentes de la conexión son $\Gamma^k_{ij} = 0$ entonces por (1.14) las componentes de la curvatura son nulas y como la curvatura es un tensor entonces las componentes son nulas en cualquier sistema de coordenadas, es decir, $R = 0$. Recíprocamente, se puede demostrar que si $R = 0$ entonces existe un sistema de coordenadas locales en el que las componentes de la conexión son $\Gamma^k_{ij} = 0$.

El siguiente teorema es un resultado fundamental en geometría Riemanniana. Demuestra la existencia de una única conexión isométrica y simétrica y proporciona una fórmula explícita. Esta conexión se conoce como la **conexión de Levi-Civita**.

Teorema 1.1. *Sea (M,g) una variedad Riemanniana entonces existe una única conexión ∇ isométrica y simétrica. Explícitamente, ∇ queda determinada por la fórmula de Koszul:*

$$2g(\nabla_X Y, Z) = Xg(Y,Z) + Yg(X,Z) - Zg(X,Y) + g([X,Y],Z) - g([X,Z],Y) - g([Y,Z],X). \tag{1.15}$$

Demostración. Supongamos que existe una conexión ∇ isométrica y simétrica. Escribimos la condición de isometría para tres campos vectoriales X,Y,Z

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z), \tag{1.16}$$

con ecuaciones análogas para las permutaciones cíclicas de X,Y,Z. Sumando dos de las ecuaciones de la forma (1.16) y restando la tercera

$$\begin{aligned} &Xg(Y,Z) + Yg(Z,X) - Zg(X,Y) \\ &\qquad = g(\nabla_X Y, Z) + g(Y, \nabla_X Z) + g(\nabla_Y Z, X) + g(Z, \nabla_Y X) - g(\nabla_Z X, Y) - g(X, \nabla_Z Y). \end{aligned} \tag{1.17}$$

Por la linealidad y simetría de la métrica g tenemos

$$Xg(Y,Z)+Yg(Z,X)-Zg(X,Y)=g(\nabla_X Y,Z)+g(\nabla_Y X,Z)+g(\nabla_Y Z-\nabla_Z Y,X)+g(\nabla_X Z-\nabla_Z X,Y). \tag{1.18}$$

Como ∇ es simétrica (torsión $T=0$)

$$T(X,Y)=\nabla_X Y-\nabla_Y X-[X,Y]=0, \quad \text{o equivalentemente,} \quad \nabla_X Y-\nabla_Y X=[X,Y], \tag{1.19}$$

con ecuaciones similares para $T(Y,Z),T(Z,X)$, las permutaciones cíclicas de X,Y,Z. Utilizando las ecuaciones de la forma (1.19) para $T(Y,Z),T(Z,X)$ en (1.18)

$$Xg(Y,Z)+Yg(Z,X)-Zg(X,Y)=g(\nabla_X Y,Z)+g(\nabla_Y X,Z)+g([Y,Z],X)+g(-[Z,X],Y). \tag{1.20}$$

Por la propiedad antisimétrica del corchete de Lie

$$Xg(Y,Z)+Yg(Z,X)-Zg(X,Y)=g(\nabla_X Y,Z)+g(\nabla_Y X,Z)+g([Y,Z],X)+g([X,Z],Y). \tag{1.21}$$

En definitiva tenemos

$$Xg(Y,Z)+Yg(Z,X)-Zg(X,Y)-g([X,Z],Y)-g([Y,Z],X)=g(\nabla_X Y,Z)+g(\nabla_Y X,Z). \tag{1.22}$$

Despejando $\nabla_Y X=\nabla_X Y-[X,Y]$ en la ecuación (1.19) y sustituyendo en (1.22)

$$Xg(Y,Z)+Yg(Z,X)-Zg(X,Y)-g([X,Z],Y)-g([Y,Z],X)=g(\nabla_X Y,Z)+g(\nabla_X Y-[X,Y],Z). \tag{1.23}$$

Finalmente por la linealidad de g obtenemos la fórmula de Koszul

$$Xg(Y,Z)+Yg(Z,X)-Zg(X,Y)+g([X,Y],Z)-g([X,Z],Y)-g([Y,Z],X)=2g(\nabla_X Y,Z). \tag{1.24}$$

El procedimiento anterior garantiza la unicidad de tal conexión, supuesta su existencia. Para probar la existencia vemos que en efecto la fórmula de Koszul define una conexión, esto es comprobar las tres propiedades de la definición de conexión.

1. En primer lugar, ∇ es $\mathscr{F}(M)$-lineal en el primer argumento. Por la fórmula de Koszul

   $$2g(\nabla_{fX}Y,Z)=fXg(Y,Z)+Yg(fX,Z)-Zg(fX,Y)+g([fX,Y],Z)-g([fX,Z],Y)-g([Y,Z],fX). \tag{1.25}$$

   Por la $\mathscr{F}(M)$-linealidad de g y la propiedad número 4 del corchete de Lie queda

   $$\begin{aligned}2g(\nabla_{fX}Y,Z)=&fXg(Y,Z)+Y(fg(X,Z))-Z(fg(X,Y))\\&+g(f[X,Y]-(Yf)X,Z)-g(f[X,Z]-(Zf)X,Y)-fg([Y,Z],X).\end{aligned} \tag{1.26}$$

   De nuevo por la $\mathscr{F}(M)$-linealidad de g y la actuación de los campos vectoriales Y,Z como derivación

   $$\begin{aligned}2g(\nabla_{fX}Y,Z)=&fXg(Y,Z)+(Yf)g(X,Z)+fYg(X,Z)-(Zf)g(X,Y)-fZg(X,Y)\\&+fg([X,Y],Z)-(Yf)g(X,Z)-fg([X,Z],Y)+(Zf)g(X,Y)-fg([Y,Z],X).\end{aligned} \tag{1.27}$$

   Eliminando los sumandos que se cancelan, sacando f factor común y usando la fórmula de Koszul

   $$\begin{aligned}2g(\nabla_{fX}Y,Z)&=f[Xg(Y,Z)+Yg(X,Z)-Zg(X,Y)+g([X,Y],Z)-g([X,Z],Y)-g([Y,Z],X)]\\&=2fg(\nabla_X Y,Z).\end{aligned} \tag{1.28}$$

   En definitiva tenemos

   $$2g(\nabla_{fX}Y,Z)=2fg(\nabla_X Y,Z)=2g(f\nabla_X Y,Z), \quad \forall X,Y,Z\in\mathscr{X}(M), \tag{1.29}$$

   luego $\nabla_{fX}Y=f\nabla_X Y, \ \ \forall X,Y\in\mathscr{X}(M)$, es decir, ∇ es $\mathscr{F}(M)$-lineal en el primer argumento.

2. En segundo lugar, ∇ es $\mathbb{R}$-lineal en el segundo argumento. Esta propiedad se sigue del hecho de que la fórmula de Koszul viene dada en términos de la métrica, el corchete de Lie y los campos vectoriales que son todos $\mathbb{R}$-lineales en todos sus argumentos.

3. En tercer lugar, ∇ cumple la regla de Leibniz. Por la fórmula de Koszul

$$\begin{aligned} 2g(\nabla_X(fY),Z) &= Xg(fY,Z)+fYg(X,Z)-Zg(X,fY) \\ &\quad +g([X,fY],Z)-g([X,Z],fY)-g([fY,Z],X). \end{aligned} \tag{1.30}$$

Por la $\mathscr{F}(M)$-linealidad de g y la propiedad número 4 del corchete de Lie

$$\begin{aligned} 2g(\nabla_X(fY),Z) &= X(fg(Y,Z))+fYg(X,Z)-Z(fg(X,Y)) \\ &\quad +g(f[X,Y]+(Xf)Y,Z)-fg([X,Z],Y)-g(f[Y,Z]-(Zf)Y,X). \end{aligned} \tag{1.31}$$

Por la actuación de los campos vectoriales X,Z como derivación

$$\begin{aligned} 2g(\nabla_X(fY),Z) &= (Xf)g(Y,Z)+fXg(Y,Z)+fYg(X,Z)-(Zf)g(X,Y)-fZg(X,Y) \\ &\quad +fg([X,Y],Z)+(Xf)g(Y,Z)-fg([X,Z],Y)-fg([Y,Z],X)+(Zf)g(Y,X). \end{aligned} \tag{1.32}$$

Por la simetría de g dos sumandos se cancelan y utilizando la fórmula de Koszul

$$2g(\nabla_X(fY),Z)=2(Xf)g(Y,Z)+f2g(\nabla_XY,Z)=2g((Xf)Y+f\nabla_XY,Z), \tag{1.33}$$

donde la última igualdad se obtiene por la $\mathscr{F}(M)$-linealidad de g. En definitiva tenemos

$$2g(\nabla_X(fY),Z)=2g((Xf)Y+f\nabla_XY,Z),\ \ \forall X,Y,Z\in\mathscr{X}(M),\ \ \forall f\in\mathscr{F}(M), \tag{1.34}$$

luego $\nabla_XfY=(Xf)Y+f\nabla_XY,\ \forall X,Y\in\mathscr{X}(M),\forall f\in\mathscr{F}(M)$, es decir, ∇ cumple la regla de Leibniz.

□

Las componentes de la conexión de Levi-Civita se llaman **símbolos de Christoffel** que vienen dados por

$$\Gamma_{ij}^p=\frac{1}{2}g^{pk}(\partial_ig_{jk}+\partial_jg_{ik}-\partial_kg_{ij}), \tag{1.35}$$

donde (g^{pk}) denota la matriz inversa de la métrica (g_{ij}). Recíprocamente si las componentes de una conexión en una variedad Riemanniana (M,g) vienen dadas por la fórmula (1.35) entonces se trata de la conexión de Levi-Civita. Si la conexión de Levi-Civita es plana entonces por (1.9) tenemos $\partial_kg_{ij}=0$ y por tanto g_{ij} son constantes. Intuitivamente esto dice que localmente la variedad se comporta como el espacio euclídeo $\mathbb{R}^n$, que tiene $R=0$ y por tanto podemos interpretar la curvatura como una forma de medir cuanto se desvía la variedad de ser euclídea.

Sea $\gamma:I\to M$ una curva C^∞ y ∇ una conexión en una variedad M. Decimos que γ es una **geodésica** para la conexión ∇ si $\nabla_{\dot\gamma}\dot\gamma=0$. Sean Γ_{ij}^k las componentes de ∇ en un sistema de coordenadas locales entonces las geodésicas para dicha conexión son la solución del sistema de ecuaciones diferenciales ordinarias de segundo orden

$$\ddot{x}^k+\Gamma_{ij}^k\dot{x}^i\dot{x}^j=0, \tag{1.36}$$

donde $\gamma(t)=(x^1(t),\ldots,x^n(t))$. La conexión más utilizada es la de Levi-Civita, esto se debe a la siguiente propiedad: una curva que conecta dos puntos en una variedad por distancia mínima es una geodésica para la conexión de Levi-Civita. El recíproco es cierto para puntos suficientemente cercanos.

Presentamos a continuación un ejemplo importante de variedad Riemanniana que nos aparecerá en el siguiente capítulo, el espacio hiperbólico.

Ejemplo 1. El espacio hiperbólico de dimensión n es la variedad Riemanniana formada por el conjunto $H_n = \{(x_1,\ldots,x_{n-1},y) \in \mathbb{R}^n \mid y > 0\}$ dotado de la métrica dada por la matriz diagonal de dimensión n

$$\begin{pmatrix} \frac{1}{y^2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{y^2} \end{pmatrix}. \tag{1.37}$$

Si $n = 2$ se denomina plano hiperbólico. Un resultado conocido en geometría diferencial es que las geodésicas del plano hiperbólico están dadas por [10]

$$\gamma : t \to (\mathrm{Re}(z(t)), \mathrm{Im}(z(t))), \tag{1.38}$$

donde

$$z(t) = \frac{aie^{vt} + b}{cie^{vt} + d} \tag{1.39}$$

con $ad - bc = 1$, $v > 0$, $a,b,c,d \in \mathbb{R}$. Utilizando el teorema de Noether [10] se demuestra que cada geodésica del espacio hiperbólico permanece en un plano perpendicular al hiperplano $y = 0$, y que contiene a la velocidad inicial. La métrica inducida en dicho plano es la del plano hiperbólico, esto permite conocer las geodésicas del espacio hiperbólico, que están dadas por

$$\gamma : t \to (x_1(t),\ldots,x_{n-1}(t),y(t)) = (x(t),y(t)), \tag{1.40}$$

donde

$$x(t) = x_0 + \frac{\dot{x}_0}{\|\dot{x}_0\|}\tilde{x}(t), \quad y(t) = \mathrm{Im}(\gamma_{\mathbb{C}}(t)), \quad \text{con} \quad x_0 = x(0), \quad \tilde{x}(t) = \mathrm{Re}(\gamma_{\mathbb{C}}(t)), \tag{1.41}$$

y

$$\gamma_{\mathbb{C}}(t) = \frac{aie^{vt} + b}{cie^{vt} + d}, \tag{1.42}$$

con $a,b,c,d \in \mathbb{R}$ tal que $ad - bc = 1$, $v > 0$. Los valores de a,b,c,d,v, se determinan mediante la posición y velocidad iniciales de la geodésica.

Capítulo 2

Geometría de la información

2.1. Variedades estadísticas y divergencias

Sea X una variable aleatoria discreta o continua, escalar o vectorial que toma valores en un conjunto $\mathscr{X} \subset \mathbb{R}^m$ llamado espacio muestral y cuya distribución de probabilidad es $p : \mathscr{X} \to \mathbb{R}$ con

$$p(x) \geq 0, \quad \int_{\mathscr{X}} p(x)dx = 1. \tag{2.1}$$

Consideramos de forma unificada los casos discreto y continuo mediante notación integral, así $\int_{\mathscr{X}} p(x)dx$ en el caso discreto significa $\sum_{x\in\mathscr{X}} p(x)$. En el caso vectorial entendemos la notación anterior como sumas e integrales múltiples. Utilizamos la notación $\partial_i = \frac{\partial}{\partial \xi^i}$, salvo que haya que considerar distintos sistemas de coordenadas, en tal caso usaremos $\partial_{\xi^i} = \frac{\partial}{\partial \xi^i}$.

Un **modelo estadístico** M es una familia de distribuciones de probabilidad en $\mathscr{X}$ tal que cada distribución está parametrizada por n valores reales

$$M = \{p(x;\xi) \mid \xi = (\xi^1, \ldots, \xi^n) \in \Xi\}, \tag{2.2}$$

donde Ξ es un subconjunto de $\mathbb{R}^n$ que llamamos espacio paramétrico y la aplicación parametrización $\phi : \Xi \to M$ dada por $\xi \to p(x;\xi)$ es inyectiva. La aplicación $\varphi : M \to \mathbb{R}^n$ dada por $\varphi(p(x;\xi)) = \xi$ permite considerar $\xi = (\xi^1, \ldots, \xi^n)$ como un sistema de coordenadas para M, luego M es una variedad denominada **variedad estadística**. La aplicación entre variedades ϕ es diferenciable y de rango dim $\Xi = n$, es un encaje. En este trabajo estudiamos variedades estadísticas con varias suposiciones adicionales que facilitan su estudio y se cumplen en los modelos y aplicaciones sencillas:

1. El espacio paramétrico Ξ es un subconjunto abierto de $\mathbb{R}^n$. Como ϕ es diferenciable entonces diferenciamos libremente respecto de los parámetros.

2. El orden de integración y diferenciación se puede intercambiar. Por ejemplo, a menudo utilizaremos el intercambio
$$\int_{\mathscr{X}} \partial_i p(x;\xi)dx = \partial_i \int_{\mathscr{X}} p(x;\xi)dx = \partial_i 1 = 0. \tag{2.3}$$

3. El soporte de $p(x;\xi)$ dado por $\overline{\{x \mid p(x;\xi) > 0\}}$ no depende de ξ. Esto significa que M es un subconjunto de
$$\mathscr{P}(\mathscr{X}) = \left\{ p : \mathscr{X} \to \mathbb{R} \mid p(x) \geq 0, \int_{\mathscr{X}} p(x)dx = 1 \right\}, \tag{2.4}$$
que es un espacio de funciones de dimensión infinita. La condición de rango n se entiende como que $\{\partial_1 p(x;\xi), \ldots, \partial_n p(x;\xi)\}$ es un sistema de funciones linealmente independientes.

En una variedad estadística cada punto representa una distribución de probabilidad. Teniendo en cuenta la naturaleza estadística de estas variedades veremos que es natural estudiarlas bajo la llamada propiedad de invariancia. La geometría de la información estudia la geometría de las variedades estadísticas bajo la invariancia.

Ejemplo 2 (Distribuciones normales)**.** El conjunto de todas las distribuciones de probabilidad normales $X = N(\mu, \sigma)$ es una variedad estadística donde

$$\mathscr{X} = \mathbb{R}, \;\; n = 2, \;\; \xi = (\mu, \sigma), \;\; \Xi = \{(\mu, \sigma) \mid -\infty < \mu < \infty, 0 < \sigma < \infty\}, \;\; M \simeq \mathbb{R} \times \mathbb{R}_+, \tag{2.5}$$

$$p(x; \xi) = p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{2.6}$$

Ejemplo 3 (Distribuciones finitas)**.** Sea X una variable aleatoria que toma valores en el conjunto $\mathscr{X} = \{0, 1, \ldots, n\}$, $n \in \mathbb{N}$. Su distribución de probabilidad está determinada por $n+1$ probabilidades

$$p_i = P(X = i) \in (0, 1), \quad i = 0, 1, \ldots, n, \tag{2.7}$$

que representamos mediante el vector de probabilidades $p = (p_0, p_1, \ldots, p_n)$ tal que $\sum_{i=0}^{n} p_i = 1$. Llamamos simplex de dimensión n y lo denotamos por S_n al conjunto de todas las distribuciones de probabilidad con valores en $\{0, 1, \ldots, n\}$

$$S_n = \{p = (p_0, \ldots, p_n) \in \mathbb{R}^{n+1} \mid p_i > 0, \; \sum_{i=0}^{n} p_i = 1\}. \tag{2.8}$$

Entonces S_n es una variedad estadística de dimensión n donde un sistema de coordenadas viene dado por

$$\xi = (p_1, \ldots, p_n) \quad \text{con} \quad p_0 = 1 - \sum_{i=1}^{n} p_i. \tag{2.9}$$

Nota 1. Si permitimos $p_i \in [0, 1]$ entonces la variedad S_n tiene borde. En este trabajo nos limitamos al estudio de variedades sin borde.

El primer objetivo es encontrar una métrica Riemanniana en las variedades estadísticas que nos permita medir distancias entre distribuciones de probabilidad. La Estadística ha desarrollado medidas de distanciamiento entre distribuciones de probabilidad que usualmente no son métricas y que muestran como de distintas son dos distribuciones de probabilidad. Estas medidas reciben el nombre de **divergencias** y pueden definirse en variedades generales.

Definición 2.1. Una divergencia en una variedad M es una función diferenciable $D(\cdot \,||\, \cdot) : M \times M \to \mathbb{R}$ tal que para todo par de puntos $P, Q \in M$ verifica:

1. $D(P \,||\, Q) \geq 0$.
2. $D(P \,||\, Q) = 0$ si y solo si $P = Q$.
3. Para dos puntos suficientemente cerca, el desarrollo de Taylor de D es

$$D(\xi_P \,||\, \xi_P + \delta) = \frac{1}{2} \sum_{i,j=1}^{n} g_{ij}(\xi_P) \delta^i \delta^j + O(|\delta|^3), \tag{2.10}$$

donde ξ_P son las coordenadas de P en un sistema de coordenadas locales ξ, $\delta \in \mathbb{R}^n$ y $(g_{ij}(\xi_P))$ es una matriz definida positiva que depende de ξ_P.

Usamos indistintamente $D(P \mid\mid Q)$ o $D(\xi_P \mid\mid \xi_Q)$ dependiendo del uso que se va a hacer. Una divergencia es una medida del grado de separación entre dos puntos en una variedad, es decir, una medida de distinción o disparidad. Aunque se parece a una distancia ni ella ni su raíz cuadrada son una métrica en general, ya que no necesariamente satisface las propiedades simétrica y desigualdad triangular. Una métrica es un caso particular de divergencia. La asimetría de las divergencias es importante en estadística en problemas de contraste de hipótesis. Al enfrentar dos hipótesis alternativas la hipótesis nula se protege más que la alternativa dándole una mayor importancia y esto es captado por la asimetría de estas medidas. La tercera propiedad de la definición 2.1 nos dice que una divergencia induce una métrica en la variedad por derivación mediante la fórmula para sus componentes

$$g_{ij}^{(D)}(\xi_1) = \partial_{\xi_2^i}\partial_{\xi_2^j} D(\xi_1 \mid\mid \xi_2)_{|\xi_2=\xi_1}, \tag{2.11}$$

que llamamos **métrica inducida**. Una divergencia también induce conexiones en las variedades por derivación, tal como veremos en la sección 2.7, denominadas conexiones inducidas.

2.2. Métrica de la información de Fisher

El objetivo de esta sección es demostrar que la matriz de la información de Fisher dota a las variedades estadísticas de una métrica Riemanniana que llamamos **métrica de la información de Fisher** o simplemente métrica de Fisher. La importancia y singularidad de esta métrica se verá en secciones posteriores.

Definición 2.2. Sea $M = \{p(x;\xi)\}$ un modelo estadístico parametrizado por ξ. La matriz de la información de Fisher o simplemente matriz de Fisher de M en un punto ξ que denotamos por $G^F = (g_{ij}^F(\xi))_{i,j=1,\dots,n}$ está definida por

$$g_{ij}^F(\xi) = E_p[\partial_i \log p(x;\xi)\partial_j \log p(x;\xi)] = \int_{\mathscr{X}} \partial_i \log p(x;\xi)\partial_j \log p(x;\xi)\cdot p(x;\xi)dx, \tag{2.12}$$

donde E_p denota la esperanza respecto de $p(x;\xi)$, es decir, $E_p[f] = \int_{\mathscr{X}} f(x)p(x;\xi)dx$.

Consideramos modelos en los que la esperanza (2.12) es finita.

Lema 2.1. *La matriz de la información de Fisher admite las siguientes expresiones alternativas:*

$$g_{ij}^F(\xi) = 4\int_{\mathscr{X}} \partial_i \sqrt{p(x;\xi)}\partial_j \sqrt{p(x;\xi)}dx, \tag{2.13}$$

$$g_{ij}^F(\xi) = -E_p[\partial_i\partial_j \log p(x;\xi)]. \tag{2.14}$$

Demostración. Para la primera fórmula, derivando en la definición de matriz de Fisher

$$\begin{aligned} g_{ij}^F(\xi) &= \int_{\mathscr{X}} \partial_i \log p(x;\xi)\partial_j \log p(x;\xi)\cdot p(x;\xi)dx = \int_{\mathscr{X}} \frac{\partial_i p(x;\xi)}{p(x;\xi)}\frac{\partial_j p(x;\xi)}{p(x;\xi)}p(x;\xi)dx \\ &= 4\int_{\mathscr{X}} \frac{\partial_i p(x;\xi)}{2\sqrt{p(x;\xi)}}\frac{\partial_j p(x;\xi)}{2\sqrt{p(x;\xi)}}dx = 4\int_{\mathscr{X}} \partial_i\sqrt{p(x;\xi)}\partial_j\sqrt{p(x;\xi)}dx. \end{aligned} \tag{2.15}$$

Para la segunda fórmula, teniendo en cuenta (2.3) tenemos

$$E_p[\partial_i \log p(x;\xi)] = \int_{\mathscr{X}} \partial_i \log p(x;\xi)\cdot p(x;\xi)dx = \int_{\mathscr{X}} \partial_i p(x;\xi)dx = 0. \tag{2.16}$$

Derivando respecto de ξ^j en (2.16) y usando (2.3) tenemos

$$\partial_j \int_{\mathscr{X}} \partial_i \log p(x;\xi)\cdot p(x;\xi)dx = \int_{\mathscr{X}} \partial_j\partial_i \log p(x;\xi)\cdot p(x;\xi)dx + \int_{\mathscr{X}} \partial_i \log p(x;\xi)\partial_j p(x;\xi)dx = 0, \tag{2.17}$$

si y solo si

$$\int_{\mathscr{X}} \partial_j \partial_i \log p(x;\xi)\cdot p(x;\xi)dx + \int_{\mathscr{X}} \partial_i \log p(x;\xi)\partial_j \log p(x;\xi)\cdot p(x;\xi)dx = 0, \tag{2.18}$$

o equivalentemente

$$E_p[\partial_j \partial_i \log p(x;\xi)] + E_p[\partial_i \log p(x;\xi)\partial_j \log p(x;\xi)] = 0. \tag{2.19}$$

□

Teorema 2.1. *La matriz de la información de Fisher se transforma como un tensor 2-covariante bajo cambios de coordenadas.*

Demostración. Consideramos dos sistemas de coordenadas locales $\xi = (\xi^1,\ldots,\xi^n)$, $\theta = (\theta^1,\ldots,\theta^n)$ relacionados mediante el cambio de variable $\xi = \xi(\theta)$ con $\xi^j = \xi^j(\theta^1,\ldots,\theta^n)$. Sea $\bar{p}(x;\theta) = p(x;\xi(\theta))$, derivando con la regla de la cadena tenemos

$$\partial_{\theta^i}\bar{p}(x;\theta) = \frac{\partial \xi^k}{\partial \theta^i}\partial_{\xi^k} p(x;\xi), \quad \partial_{\theta^j}\bar{p}(x;\theta) = \frac{\partial \xi^r}{\partial \theta^j}\partial_{\xi^r} p(x;\xi). \tag{2.20}$$

Entonces

$$\begin{aligned} \bar{g}^F_{ij}(\theta) &= \int_{\mathscr{X}} \partial_{\theta^i}\log\bar{p}(x;\theta)\partial_{\theta^j}\log\bar{p}(x;\theta)\cdot\bar{p}(x;\theta)dx = \int_{\mathscr{X}} \frac{1}{\bar{p}(x;\theta)}\partial_{\theta^i}\bar{p}(x;\theta)\partial_{\theta^j}\bar{p}(x;\theta)dx \\ &= \left[\int_{\mathscr{X}} \frac{1}{p(x;\xi(\theta))}\partial_{\xi^k}p(x;\xi)\partial_{\xi^r}p(x;\xi)dx\right]\frac{\partial\xi^k}{\partial\theta^i}\frac{\partial\xi^r}{\partial\theta^j} = g^F_{kr}(\xi)\frac{\partial\xi^k}{\partial\theta^i}\frac{\partial\xi^r}{\partial\theta^j}. \end{aligned} \tag{2.21}$$

□

Teorema 2.2. *La matriz de la información de Fisher es una métrica Riemanniana en toda variedad estadística.*

Demostración. Comprobamos que la matriz de Fisher es simétrica y definida positiva. La simetría es evidente por la propia definición (2.12), veamos que es definida positiva. Sea M una variedad estadística, para todo punto ξ y vector tangente $v \in T_\xi M$ con $v \neq 0$ por (2.13) tenemos

$$\begin{aligned} g(v,v) &= \sum_{i,j=1}^n g^F_{ij}v^iv^j = 4\sum_{i,j=1}^n\left(\int_{\mathscr{X}} v^i\partial_i\sqrt{p(x;\xi)}v^j\partial_j\sqrt{p(x;\xi)}dx\right) \\ &= 4\int_{\mathscr{X}}\left(\sum_{i=1}^n v^i\partial_i\sqrt{p(x;\xi)}\right)\left(\sum_{j=1}^n v^j\partial_j\sqrt{p(x;\xi)}\right)dx = 4\int_{\mathscr{X}}\left(\sum_{i=1}^n v^i\partial_i\sqrt{p(x;\xi)}\right)^2 dx \geq 0, \end{aligned} \tag{2.22}$$

por tanto la matriz de Fisher es semidefinida positiva. Veamos que es definida positiva

$$\begin{aligned} g(v,v) = 0 &\Leftrightarrow \int_{\mathscr{X}}\left(\sum_{i=1}^n v^i\partial_i\sqrt{p(x;\xi)}\right)^2 dx = 0 \Leftrightarrow \left(\sum_{i=1}^n v^i\partial_i\sqrt{p(x;\xi)}\right)^2 = 0 \\ &\Leftrightarrow \sum_{i=1}^n v^i\partial_i\sqrt{p(x;\xi)} = 0 \Leftrightarrow \sum_{i=1}^n v^i\partial_i p(x;\xi) = 0 \Leftrightarrow v^i = 0,\ \forall i = 1,\ldots,n, \end{aligned} \tag{2.23}$$

ya que $\{\partial_1 p(x;\xi),\ldots,\partial_n p(x;\xi)\}$ es un sistema de funciones linealmente independientes. Así la matriz de Fisher es definida positiva y en definitiva una métrica Riemanniana. □

2.3. Monotonía de la información e invariancia

En esta sección consideramos condiciones que son naturales de imponer a las divergencias, métricas y en general a la geometría de las variedades estadísticas. Dichas condiciones se conocen como monotonía de la información e invariancia y se obtienen teniendo en cuenta la naturaleza estadística de estas variedades en las que cada punto representa una distribución de probabilidad.

Sea X una variable aleatoria con distribución de probabilidad paramétrica $p(x;\xi)$ que describe una población, $\tilde{X}$ una muestra aleatoria y T una función medible entonces $Y = T(X)$ es otra variable aleatoria y $T = T(\tilde{X})$ se denomina **estadístico**. Para hacer inferencias sobre el parámetro desconocido ξ partimos de la información que suministra la muestra aleatoria $\tilde{X}$ resumiendo esta información muestral en un estadístico $T(\tilde{X})$. El resumen que hace cualquier estadístico $T(\tilde{X})$ supone mantener o reducir la información que suministra la muestra acerca del parámetro desconocido ξ. Una propiedad deseable de un estadístico $T(\tilde{X})$ es que no pierda información, esta propiedad se conoce como **suficiencia**. Un estadístico es suficiente si aprovecha toda la información que suministra la muestra respecto al parámetro ξ. Formalmente un estadístico $T(\tilde{X})$ es suficiente respecto del parámetro ξ si dado el valor del estadístico $T(\tilde{X})$ la distribución condicional de la muestra aleatoria $\tilde{X}$ no depende de ξ, es decir,

$$P(\tilde{X} = \tilde{x} \mid T(\tilde{X}) = t) = P(\tilde{X} = \tilde{x} \mid T(\tilde{X}) = t; \xi). \tag{2.24}$$

Para determinar si un estadístico es suficiente contamos con un criterio más simple y eficaz que la definición anterior: **el teorema de factorización**. Sea $f(\tilde{x};\xi)$ la distribución de probabilidad conjunta de una muestra $\tilde{X}$. Un estadístico $T(\tilde{X})$ es suficiente para ξ si y solo si existen funciones $g(t;\xi)$ y $h(\tilde{x})$ tal que para toda muestra $\tilde{x} \in \mathscr{X}$ y todo valor del parámetro ξ se tiene

$$f(\tilde{x};\xi) = g(T(\tilde{x});\xi)h(\tilde{x}). \tag{2.25}$$

El término información está relacionado con la idea de disparidad. Cuanto mayor sea la variabilidad, es decir, las discrepancias en la población, mayor información debe contener la muestra aleatoria. Es evidente que si todos los resultados de un fenómeno aleatorio son equiprobables tenemos menos información para decidir sobre alguno de ellos que si sabemos que algunos tienen mayor probabilidad de suceder que otros, es decir, mayor variabilidad implica mayor información y menor información implica menor variabilidad. A continuación formulamos esta idea usando divergencias.

Sea $M = \{p(x;\xi)\}$ una variedad estadística donde $p(x;\xi)$ es la distribución de probabilidad de una variable aleatoria X. Un estadístico T define otra variedad estadística $M^T = \{q(y;\xi)\} = \{q(T(x);\xi\}$ donde $q(y;\xi)$ es la distribución de probabilidad de la variable aleatoria $Y = T(X)$ dada por

$$q(y;\xi) = \int_K p(x;\xi)dx, \tag{2.26}$$

con $K = \{x \mid T(x) = y\}$. Sean D, D^T dos divergencias en M, M^T respectivamente. Conviene recordar que una divergencia en una variedad mide la discrepancia entre sus puntos. Como a través de un estadístico la información se mantiene o se reduce y menor información implica menor variabilidad, es decir, menor discrepancia, entonces

$$\boxed{D^T(\xi_1 \mid\mid \xi_2) \leq D(\xi_1 \mid\mid \xi_2).} \tag{2.27}$$

La desigualdad (2.27) se conoce como **monotonía de la información**. Por definición, un estadístico es suficiente si y solo si no pierde información, entonces

$$\boxed{D^T(\xi_1 \mid\mid \xi_2) = D(\xi_1 \mid\mid \xi_2) \quad \Leftrightarrow \quad T \text{ es suficiente.}} \tag{2.28}$$

Una divergencia se dice **invariante** si cumple (2.27) y (2.28). Se puede demostrar [1] que dada una variedad Riemanniana con conexiones existe una divergencia canónica que induce la métrica y las conexiones dadas por derivación mediante las fórmulas de la métrica inducida y de las conexiones inducidas. Entonces una métrica o una conexión se dice invariante si está inducida por una divergencia invariante.

Por otro lado un estadístico T genera una partición del espacio muestral. Sea el espacio muestral $\mathscr{X} = \{x \mid x \text{ muestra observable de } X\}$, su imagen por T es $\mathscr{T} = \{t \mid t = T(x) \text{ para algún } x \in \mathscr{X}\}$. Entonces el estadístico T genera una partición de $\mathscr{X}$ en subconjuntos $A_t = \{x \mid T(x) = t\}$ con $t \in \mathscr{T}$ y así $T(x) = t$ es equivalente a $x \in A_t$. En términos de la particiones del espacio muestral, T es suficiente respecto del parámetro ξ si basta con conocer a que conjunto de la partición generada por T conduce la muestra obtenida, no añadiendo más información saber cual es la muestra concreta. Prestaremos especial atención a las variedades S_n, por lo que conviene detallar lo estudiado en esta sección para S_n.

Si X toma valores en el espacio muestral $\mathscr{X} = \{0,1,\ldots,n\}$ su distribución de probabilidad es $p = (p_0, p_1, \ldots, p_n)$ con $p_i = P(X = i)$, $i = 0,1,\ldots,n$. Un estadístico T genera una partición $\{A_j\}_{j=0,\ldots,m}$ de $\{0,1,\ldots,n\}$ con $m \leq n$ y una variable aleatoria $Y = T(X)$ con valores en $\{0,1,\ldots,m\}$ y distribución de probabilidad $\bar{p} = (\bar{p}_0, \bar{p}_1, \ldots, \bar{p}_m)$ con

$$\bar{p}_j = P(Y = j) = \sum_{i \in A_j} p_i, \tag{2.29}$$

ya que una variable aleatoria discreta (finita) solo puede transformarse en otra discreta (finita). El estadístico T nos conduce de S_n a S_m. Sean $p, q \in S_n$ y D una divergencia en S_n entonces la invariancia queda

$$D^T(\bar{p} \,||\, \bar{q}) \leq D(p \,||\, q), \tag{2.30}$$

satisfaciendo la igualdad si y solo si T es suficiente. Como Y es función de X y usando la descomposición de la distribución conjunta en producto de la marginal por la condicional se tiene

$$p(x;\xi) = p(x,y;\xi) = p(y;\xi)p(x \mid y;\xi), \quad \text{equivalentemente,} \quad p(x \mid y;\xi) = \frac{p(x;\xi)}{p(y;\xi)}, \tag{2.31}$$

con ξ un sistema de coordenadas en S_n. Por definición T es suficiente si y solo si la distribución condicional no depende de ξ, es decir,

$$p(x \mid y;\xi) = p(x \mid y;\xi'), \tag{2.32}$$

y por tanto

$$\frac{p(x;\xi)}{p(y;\xi)} = \frac{p(x;\xi')}{p(y;\xi')}. \tag{2.33}$$

2.4. Divergencias invariantes: f-divergencias

Una clase importante de divergencias en una variedad estadística son las f-divergencias. En esta sección estudiamos las métricas que inducen y su relación con la invariancia.

Definición 2.3. Sea M una variedad estadística y $f : (0,\infty) \to \mathbb{R}$ una función diferenciable y convexa tal que $f(1) = 0$ y $f''(1) = 1$. Llamamos f-divergencia a la función $D_f : M \times M \to \mathbb{R}$ definida por

$$D_f(p \,||\, q) = E_p\left[f\left(\frac{q(x)}{p(x)}\right)\right] = \int_{\mathscr{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) dx. \tag{2.34}$$

Se ha omitido la dependencia en ξ para simplificar la lectura. Para demostrar que toda f-divergencia es en efecto divergencia necesitamos recordar la desigualdad de Jensen de análisis convexo. Una función $f : \mathbb{R} \to \mathbb{R}$ es convexa si y solo si

$$f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n), \tag{2.35}$$

para cualesquiera $\lambda_i \geq 0$ tal que $\sum_{i=1}^n \lambda_i = 1$. La igualdad se satisface si y solo si f es lineal o $x_i = x_j$ para todo $i, j = 1,\ldots,n$. Además si f convexa y $E[|f(X)|] < \infty$ entonces

$$f(E[X]) \leq E[f(X)]. \tag{2.36}$$

Proposición 2.1. *Toda f-divergencia es divergencia.*

Demostración. Comprobamos las tres propiedades de la definición de divergencia. Recordar que para simplificar usamos $p(x) = p(x;\xi_1)$, $q(x) = q(x;\xi_2)$. Entonces:

1. Por la desigualdad de Jensen se tiene

$$D_f(p \,||\, q) = E_p\left[f\left(\frac{q(x)}{p(x)}\right)\right] \geq f\left(E_p\left[\frac{q(x)}{p(x)}\right]\right) = f\left(\int_{\mathscr{X}} p(x)\frac{q(x)}{p(x)}dx\right) = f(1) = 0. \quad (2.37)$$

2. Si $p = q$ entonces se tiene

$$D_f(p \,||\, p) = E_p\left[f\left(\frac{p(x)}{p(x)}\right)\right] = E_p[f(1)] = E_p[0] = 0. \quad (2.38)$$

   Recíprocamente si $p \neq q$ como $f''(1) = 1$, es decir, f es estrictamente convexa en 1 entonces

$$D_f(p \,||\, q) = E_p\left[f\left(\frac{q(x)}{p(x)}\right)\right] > f\left(E_p\left[\frac{q(x)}{p(x)}\right]\right) = f(1) = 0. \quad (2.39)$$

3. Derivando respecto de los parámetros $\xi_1, \xi_2 \in \Xi \subset \mathbb{R}^n$ se tiene

$$\partial_{\xi_1^j}\left(p(x)f\left(\frac{q(x)}{p(x)}\right)\right) = \partial_{\xi_1^j}p(x)f\left(\frac{q(x)}{p(x)}\right) - p(x)f'\left(\frac{q(x)}{p(x)}\right)\frac{q(x)\partial_{\xi_1^j}p(x)}{p(x)^2}, \quad (2.40)$$

$$\partial_{\xi_2^j}\left(p(x)f\left(\frac{q(x)}{p(x)}\right)\right) = p(x)f'\left(\frac{q(x)}{p(x)}\right)\frac{\partial_{\xi_2^j}q(x)}{p(x)} = f'\left(\frac{q(x)}{p(x)}\right)\partial_{\xi_2^j}q(x), \quad (2.41)$$

$$\begin{aligned}\partial_{\xi_2^i}\partial_{\xi_2^j}\left(p(x)f\left(\frac{q(x)}{p(x)}\right)\right) &= p(x)f''\left(\frac{q(x)}{p(x)}\right)\frac{\partial_{\xi_2^i}q(x)}{p(x)}\frac{\partial_{\xi_2^j}q(x)}{p(x)} + p(x)f'\left(\frac{q(x)}{p(x)}\right)\frac{\partial_{\xi_2^i}\partial_{\xi_2^j}q(x)}{p(x)} \\ &= f''\left(\frac{q(x)}{p(x)}\right)\frac{q(x)^2}{p(x)}\partial_{\xi_2^i}\log q(x)\partial_{\xi_2^j}\log q(x) + f'\left(\frac{q(x)}{p(x)}\right)\partial_{\xi_2^i}\partial_{\xi_2^j}q(x).\end{aligned} \quad (2.42)$$

   Usando (2.3), igualando $\xi_1 = \xi_2$, es decir, $p(x) = q(x)$ y como $f(1) = 0$ y $f''(1) = 1$ entonces

$$\partial_{\xi_1^j}D_f(\xi_1 \,||\, \xi_2)_{|\xi_2=\xi_1} = f(1)\int_{\mathscr{X}}\partial_{\xi_1^j}p(x)dx - f'(1)\partial_{\xi_1^j}\int_{\mathscr{X}}p(x)dx = 0, \quad (2.43)$$

$$\partial_{\xi_2^j}D_f(\xi_1 \,||\, \xi_2)_{|\xi_2=\xi_1} = f'(1)\partial_{\xi_2^j}\int_{\mathscr{X}}q(x)dx = 0, \quad (2.44)$$

$$\begin{aligned}\partial_{\xi_2^i}\partial_{\xi_2^j}D_f(\xi_1 \,||\, \xi_2)_{|\xi_2=\xi_1} &= f''(1)\int_{\mathscr{X}}\partial_{\xi_2^i}\log q(x)\partial_{\xi_2^j}\log q(x)\cdot q(x)dx + f'(1)\partial_{\xi_2^i}\partial_{\xi_2^j}\int_{\mathscr{X}}q(x)dx \\ &= E_q[\partial_{\xi_2^i}\log q(x)\partial_{\xi_2^j}\log q(x)] = g^F_{ij}(\xi_2),\end{aligned} \quad (2.45)$$

   y como la matriz de Fisher es definida positiva, el resultado queda demostrado.

□

Proposición 2.2. *Toda f-divergencia en S_n es invariante.*

Demostración. Sea D_f una f-divergencia en S_n, T un estadístico que genera la partición $\{0\}$, $\{1,2\}$, $\{3\},\ldots,\{n\}$ del espacio muestral $\{0,1,\ldots,n\}$ y D_f^T una f-divergencia inducida por T en S_{n-1}. Basta demostrar el resultado para la partición mencionada porque en S_n cualquier otra partición es composición de estas. Las divergencias D_f, D_f^T son de la forma

$$D_f(p \,||\, q) = p_0 f\left(\frac{q_0}{p_0}\right) + p_1 f\left(\frac{q_1}{p_1}\right) + p_2 f\left(\frac{q_2}{p_2}\right) + p_3 f\left(\frac{q_3}{p_3}\right) + \cdots + p_n f\left(\frac{q_n}{p_n}\right), \tag{2.46}$$

$$D_f^T(p \,||\, q) = p_0 f\left(\frac{q_0}{p_0}\right) + (p_1+p_2) f\left(\frac{q_1+q_2}{p_1+p_2}\right) + p_3 f\left(\frac{q_3}{p_3}\right) + \cdots + p_n f\left(\frac{q_n}{p_n}\right), \tag{2.47}$$

por tanto para demostrar la monotonía de la información, $D_f^T(p \,||\, q) \leq D_f(p \,||\, q)$ hay que demostrar

$$(p_1+p_2) f\left(\frac{q_1+q_2}{p_1+p_2}\right) \leq p_1 f\left(\frac{q_1}{p_1}\right) + p_2 f\left(\frac{q_2}{p_2}\right). \tag{2.48}$$

Introducimos la siguiente notación

$$u_1 = \frac{q_1}{p_1},\ u_2 = \frac{q_2}{p_2}, \tag{2.49}$$

y por la desigualdad de Jensen

$$\begin{aligned}(p_1+p_2) f\left(\frac{q_1+q_2}{p_1+p_2}\right) &= (p_1+p_2) f\left(\frac{p_1}{p_1+p_2}u_1 + \frac{p_2}{p_1+p_2}u_2\right) \\ &\leq (p_1+p_2)\left(\frac{p_1}{p_1+p_2} f(u_1) + \frac{p_2}{p_1+p_2} f(u_2)\right) \\ &= p_1 f(u_1) + p_2 f(u_2) = p_1 f\left(\frac{q_1}{p_1}\right) + p_2 f\left(\frac{q_2}{p_2}\right).\end{aligned} \tag{2.50}$$

Para demostrar la invariancia hay que ver que (2.48) se satisface con igualdad si y solo si T es suficiente. Como estamos en el caso S_n, por (2.33), T es suficiente si y solo si $u_1 = u_2$, o equivalentemente, la desigualdad de Jensen se verifica con igualdad. Por tanto T es suficiente si y solo si

$$(p_1+p_2) f\left(\frac{q_1+q_2}{p_1+p_2}\right) = p_1 f\left(\frac{q_1}{p_1}\right) + p_2 f\left(\frac{q_2}{p_2}\right), \tag{2.51}$$

y entonces T es suficiente si y solo si $D_f = D_f^T$. Así las f-divergencias son invariantes en S_n. □

Nota 2. Tomando $f(u) = -\log u$ obtenemos la divergencia de Kullback-Leibler. El tercer apartado de la proposición 2.1 demuestra que toda f-divergencia induce la métrica de Fisher. El resultado de la proposición 2.2 se mantiene en cualquier variedad estadística [3]. Por tanto la métrica de Fisher es invariante en cualquier variedad estadística.

2.5. Unicidad de la métrica invariante: la métrica de Fisher

La importancia y singularidad de la métrica de Fisher reside en el hecho de que es la única métrica invariante, salvo constante multiplicativa, en una variedad estadística, es decir, la única métrica razonable en estas variedades teniendo en cuenta su naturaleza estadística. Chentsov [7] demostró este resultado usando teoría de categorías, nosotros lo demostramos para el caso particular de las variedades S_n mediante una reformulación de la invariancia.

Consideramos S_n, es decir, una variable aleatoria X que toma valores en $\{0,1,\ldots,n\}$ y distribución de probabilidad $p = (p_0, p_1, \ldots, p_n) \in S_n \subset \mathbb{R}_+^{n+1}$ con $p_i = P(X=i)$, $i = 0,1,\ldots,n$. Un estadístico T genera una partición $\{A_j\}_{j=0,1,\ldots,m}$ del conjunto $\{0,1,\ldots,n\}$ con $m \leq n$ y una variable aleatoria $Y =$

$T(X)$ que toma valores en $\{0,1,\ldots,m\}$ y distribución de probabilidad $q=(q_0,q_1,\ldots,q_m)\in S_m\subset\mathbb{R}_+^{m+1}$ con

$$q_j=P(Y=j)=\sum_{i\in A_j}p_i,\ \ j=0,1,\ldots,m. \tag{2.52}$$

La partición $\{A_j\}$ permite definir la aplicación

$$f:S_n\to S_m;\ \ f:p\to f(p)=q;\ \ q_j=\sum_{i\in A_j}p_i, \tag{2.53}$$

que no es inyectiva, no tiene inversa. Una aplicación más interesante se puede definir en sentido contrario. Sea $(r_{ij})_{i=0,1,\ldots,n;j=0,1,\ldots,m}$ una distribución de probabilidad condicionada cualquiera

$$r_{ij}=P(X=i\mid Y=j)=\begin{cases}P(X=i\mid Y=j), & \text{si } i\in A_j,\\ 0, & \text{si } i\notin A_j,\end{cases} \tag{2.54}$$

que cumple $\sum_{i=0}^n r_{ij}=\sum_{i\in A_j}P(X=i\mid Y=j)=1$. Entonces podemos definir la aplicación

$$h:S_m\to S_n;\ \ h:q\to h(q)=p;\ \ p_i=\sum_{j=0}^m r_{ij}q_j=r_{ij}q_j, \tag{2.55}$$

donde la última igualdad se debe a que $r_{ij}=0$ a no ser que $i\in A_j$, es decir, en la suma solo hay un término no nulo. Por tanto la aplicación h es inyectiva, tiene inversa. Las aplicaciones h se denominan **aplicaciones de Markov** y son inmersiones de la variedad S_m en S_n. Las aplicaciones de Markov tienen la siguiente propiedad

$$\sum_{i=0}^n p_i=\sum_{i=0}^n\sum_{j=0}^m r_{ij}q_j=\sum_{j=0}^m q_j\sum_{i=0}^n r_{ij}=\sum_{j=0}^m q_j=1. \tag{2.56}$$

Consideramos ahora los espacios tangentes T_pS_n, T_qS_m y sus respectivas bases $\{e_i^n\}_{i=1,\ldots,n}$, $\{e_j^m\}_{j=1,\ldots,m}$ donde $e_i^n=\partial_i=\frac{\partial}{\partial p_i}$, $e_j^m=\partial_j=\frac{\partial}{\partial q_j}$. Si $U,V\in T_qS_m$, $u,v\in T_pS_n$ entonces son combinación lineal de la base

$$V=V^je_j^m=\sum_{j=1}^m V^je_j^m,\quad v=v^ie_i^n=\sum_{i=1}^n v^ie_i^n. \tag{2.57}$$

El producto escalar viene dado por

$$\langle U,V\rangle_q=g_{rk}^m(q)U^rV^k,\quad \langle u,v\rangle_p=g_{rk}^n(p)u^rv^k, \tag{2.58}$$

donde

$$g_{rk}^m(q)=\langle e_r^m,e_k^m\rangle_q,\quad g_{rk}^n(p)=\langle e_r^n,e_k^n\rangle_p. \tag{2.59}$$

Dada una aplicación de Markov $h:S_m\to S_n$, por teoría de variedades tenemos asociada una aplicación lineal entre los correspondientes espacios tangentes denominada aplicación diferencial definida por

$$dh_q:T_qS_m\to T_pS_n;\quad dh_q(e_j^m)=\sum_{i=1}^n\frac{\partial p_i}{\partial q_j}e_i^n, \tag{2.60}$$

Las variedades S_m, S_n son combinaciones convexas de puntos, luego combinaciones lineales y por tanto las variedades y sus correspondientes espacios tangentes coindicen. Entonces

$$dh_q(e_j^m)=\sum_{i=1}^n\frac{\partial p_i}{\partial q_j}e_i^n=\sum_{i=1}^n r_{ij}e_i^n, \tag{2.61}$$

y

$$dh_q(V)=v\quad\text{con}\quad v^i=r_{ij}V^j. \tag{2.62}$$

Reformulamos la invariancia de las métricas en las variedades S_n mediante la siguiente expresión que involucra aplicaciones de Markov

$$\langle U,V\rangle_q = \langle dh_q(U), dh_q(V)\rangle_p = \langle u,v\rangle_p. \tag{2.63}$$

Demostraremos que existe una única métrica invariante bajo aplicaciones de Markov en S_n, salvo constante multiplicativa, que es la métrica de Fisher. La reformulación se basa en que fijada una distribución condicional $\{r_{ij}\}$, fijada una aplicación de Markov h, las distribuciones $q \in S_m$ y su imagen $h(q) = p \in S_n$ representan la misma información. La imagen del simplex S_m en S_n por una aplicación de Markov h es idéntica estadísticamente a S_m, en el sentido que es tan fácil o difícil distinguir dos distribuciones en S_m como distinguir sus respectivas imágenes en S_n. Cuales sean las relaciones geométricas en S_m deben ser exactamente las mismas que en $h(S_m)$. Para demostrar el teorema de Chentsov, necesitamos calcular la métrica de Fisher en S_{m-1}. Sea $q = (q_0,\dots,q_{m-1}) \in S_{m-1}$, donde $q_k = P(X = k;\xi)$, con $k = 0,\dots,m$ y $\xi = (q_1,\dots,q_{m-1})$ un sistema de coordenadas. Derivamos respecto $q_1,\dots,q_m$ y tenemos

$$\partial_i q_0 = \partial_i(1 - \sum_{j=1}^{m-1} q_j) = -1, \quad i = 1,\dots,m-1, \quad \partial_i q_j = \delta_{ij} = \begin{cases} 0, & \text{si } i \neq j \\ 1, & \text{si } i = j \end{cases} \quad \text{para } i,j = 1,\dots,m-1. \tag{2.64}$$

Entonces para todo $i,j = 1,\dots,m-1$, la métrica de la información de Fisher es la matriz

$$\begin{aligned} g_{ij}(\xi) &= E[\partial_i \log P(X=k;\xi)\partial_j \log P(X=k;\xi)] = E[\partial_i \log q_k \partial_j \log q_k] = \sum_{k=0}^{m-1} q_k \partial_i \log q_k \partial_j \log q_k \\ &= \sum_{k=0}^{m-1} \frac{\partial_i q_k}{q_k}\frac{\partial_j q_k}{q_k} q_k = \frac{\partial_i q_0 \partial_j q_0}{q_0} + \sum_{k=1}^{m-1} \frac{\partial_i q_k}{q_k}\frac{\partial_j q_k}{q_k} q_k = \frac{1}{q_0} + \sum_{k=1}^{m-1} \frac{\delta_{ik}\delta_{jk}}{q_k} = \frac{1}{q_0} + \frac{\delta_{ij}}{q_j}. \end{aligned} \tag{2.65}$$

Teorema 2.3. *Existe una única métrica invariante en S_n salvo constante multiplicativa que viene dada por la métrica de la información de Fisher.*

Demostración. Consideramos el simplex S_{n-1} como subconjunto de $\mathbb{R}^n_+ = \{(x_1,\dots,x_n) \in \mathbb{R}^n \mid x_i > 0, i = 1,\dots,n\}$. Si $m = n$ entonces las aplicaciones f, h son permutaciones, que son producto de trasposiciones. Fijamos $a,b \in \{1,\dots,n\}$ y consideramos la trasposición que intercambia los lugares a,b. Entonces la matriz r_{ij} de h es una matriz de trasposición y utilizando (2.61) tenemos

$$dh_q(e^n_a) = e^n_b, \quad dh_q(e^n_b) = e^n_a \quad y \quad dh_q(e^n_j) = e^n_j, \quad \forall j \neq a,b. \tag{2.66}$$

La hipótesis de invarianza se escribe para $n = m$ como

$$\langle e^n_{j_1}, e^n_{j_2}\rangle = \langle dh_q(e^n_{j_1}), dh_q(e^n_{j_2})\rangle, \tag{2.67}$$

y por tanto llegamos a las expresiones

$$g^n_{aj}(q) = g^n_{bj}(p) \quad y \quad g^n_{bj}(q) = g^n_{aj}(p), \quad \forall j \neq a,b, \tag{2.68}$$

$$g^n_{aa}(q) = g^n_{bb}(p) \quad y \quad g^n_{bb}(q) = g^n_{aa}(p), \tag{2.69}$$

$$g^n_{ij}(q) = g^n_{ij}(p), \quad \forall i,j \neq a,b. \tag{2.70}$$

Las condiciones anteriores son útiles en el baricentro de S_{n-1} que es el punto

$$\bar{p} = \left(\frac{1}{n},\dots,\frac{1}{n}\right) \in \mathbb{R}^n_+, \tag{2.71}$$

ya que este punto $\bar{p}$ es invariante bajo permutaciones de sus componentes. Así para todas las elecciones posibles de pares (a,b) tenemos

$$\langle e_i, e_j\rangle^n_{\bar{p}} = g^n_{ij}(\bar{p}) = B(n), \ \forall i \neq j, \tag{2.72}$$

$$\langle e_i, e_i\rangle^n_{\bar{p}} = g^n_{ii}(\bar{p}) = C(n),\ \forall i. \tag{2.73}$$

En definitiva la expresión de la métrica en $\bar{p}$ viene dada por la matriz $G_n = (g^n_{ij}(\bar{p}))_{i,j=1,\dots,n}$ con

$$g^n_{ij}(\bar{p}) = A(n)\delta_{ij} + B(n),\ \forall i, j = 1,\dots,n, \tag{2.74}$$

donde $A(n) = C(n) - B(n)$, con $A(n), B(n), C(n) \in \mathbb{R}$ dependen de n y δ_{ij} es la Delta de Kronecker. En forma matricial

$$G_n = A(n)\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} + B(n)\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}. \tag{2.75}$$

Definimos la función $\phi : \mathbb{R}^n_+ \to \mathbb{R}$ dada por

$$\phi(p_1,\dots,p_n) = \sum_{i=1}^{n} p_i - 1. \tag{2.76}$$

Como ϕ es diferenciable y la matriz Jacobiana de ϕ en un punto $p = (p_1,\dots,p_n)$ es

$$\Phi = (\frac{\partial\phi}{\partial p_1},\dots,\frac{\partial\phi}{\partial p_n})_{|p} = (1,\dots,1), \tag{2.77}$$

que tiene rango 1 entonces por teoría de varieades $S_{n-1} = \phi^{-1}(0)$ es una variedad diferenciable de dimensión $n-1$. Tomando las cartas identidad, Φ es la matriz de la aplicación diferencial $d\phi_p$. Sea $z = (z_1,\dots,z_n) \in T_pS_{n-1}$ un vector tangente a S_{n-1} en un punto p entonces $d\phi_p(z) = 0$, es decir,

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = 0. \tag{2.78}$$

Por tanto al aplicar la métrica G_n a vectores tangentes de S_{n-1} el segundo sumando se anula

$$B(n)\begin{pmatrix} z_1 & \cdots & z_n \end{pmatrix}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = 0, \tag{2.79}$$

luego podemos escribir la métrica G_n así

$$g^n_{ij}(\bar{p}) = A(n)\delta_{ij}. \tag{2.80}$$

Consideramos puntos con coordenadas racionales

$$q = (q_1,\dots,q_m) = \left(\frac{k_1}{n}, \frac{k_2}{n}, \dots, \frac{k_m}{n}\right) \in S_{m-1}, \tag{2.81}$$

donde $k_j \in \mathbb{Z}_+$ tal que $\sum_{j=1}^m k_j = n$. Sea la aplicación de Markov h definida mediante la distribución condicional

$$r_{ij} = \begin{cases} \frac{1}{k_j}, & \text{si } i \in A_j, \\ 0, & \text{si } i \notin A_j, \end{cases} \tag{2.82}$$

donde $\{A_j\}_{j=1,\dots,m}$ es la partición del conjunto de valores $\{1,\dots,n\}$ dada por

$$A_1 = \{1,2,\dots,k_1\},\ A_2 = \{k_1+1,\dots,k_1+k_2\},\dots,A_m = \{k_1+\cdots+k_{m-1}+1,\dots,k_1+\cdots+k_m = n\}, \tag{2.83}$$

y cada A_j contiene exactamente k_j elementos. Entonces h aplica todo punto q en $\bar{p}$

$$h(q) = (p_1, \ldots, p_n) = (r_{1j}q_j, \ldots, r_{nj}q_j) = \left(\frac{1}{k_j}\frac{k_j}{n}, \ldots, \frac{1}{k_j}\frac{k_j}{n}\right) = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \bar{p}. \tag{2.84}$$

Sea $e_1^m \in \mathbb{R}_+^m$ un vector básico del espacio tangente de S_{m-1} entonces

$$dh_q(e_1^m) = \sum_{i=1}^{n} r_{i1}e_i^n = \frac{1}{k_1}e_1^n + \cdots + \frac{1}{k_1}e_{k_1}^n = \frac{1}{k_1}(e_1^n + \cdots + e_{k_1}^n). \tag{2.85}$$

De forma similar para cada $j = 2, \ldots, m$

$$dh_q(e_j^m) = \sum_{i=1}^{n} r_{ij}e_i^n = \frac{1}{k_j}(e_{k_1+\cdots+k_{j-1}+1}^n + \cdots + e_{k_1+\cdots+k_j}^n). \tag{2.86}$$

Por tanto aplicando la hipótesis de invariancia y utilizando (2.80) tenemos

$$\begin{aligned} g_{11}^m(q) &= \langle e_1^m, e_1^m \rangle = \langle h_*(e_1^m), h_*(e_1^m) \rangle = \left\langle \frac{1}{k_1}\sum_{i=1}^{k_1} e_i^n, \frac{1}{k_1}\sum_{i=1}^{k_1} e_i^n \right\rangle \\ &= \frac{1}{k_1^2}\left(\langle e_1^n, e_1^n \rangle + \cdots + \langle e_1^n, e_{k_1}^n \rangle + \langle e_2^n, e_1^n \rangle + \cdots + \langle e_2^n, e_{k_1}^n \rangle + \cdots + \langle e_{k_1}^n, e_1^n \rangle + \cdots + \langle e_{k_1}^n, e_{k_1}^n \rangle\right) \\ &= \frac{k_1 A(n)}{{k_1}^2} = \frac{A(n)}{k_1} = \frac{nc}{k_1} = \frac{c}{q_1}, \end{aligned} \tag{2.87}$$

donde la constante multiplicativa $c \in \mathbb{R}$ determina la escala de la métrica. Análogamente para cada $j = 2, \ldots, m$

$$g_{jj}^m(q) = \langle e_j^m, e_j^m \rangle = \langle h_*(e_j^m), h_*(e_j^m) \rangle = \left\langle \frac{1}{k_j}\sum_{i\in A_j} e_i^n, \frac{1}{k_j}\sum_{i\in A_j} e_i^n \right\rangle = \frac{k_j A(n)}{{k_j}^2} = \frac{A(n)}{k_j} = \frac{nc}{k_j} = \frac{c}{q_j}. \tag{2.88}$$

Hemos demostrado que toda métrica invariante en S_{m-1} es de la forma

$$g_{ij}^m(q) = \begin{cases} \frac{c}{q_j}, & \text{si } i = j, \\ 0, & \text{si } i \neq j. \end{cases} \tag{2.89}$$

que coincide con la métrica obtenida en (2.65) salvo la constante multiplicativa $c \in \mathbb{R}$. El teorema queda demostrado para puntos racionales de la forma (2.81). Como $\mathbb{Q}$ es denso en $\mathbb{R}$, por continuidad el resultado se mantiene para todo punto $q \in S_{m-1}$. □

2.6. Ejemplos de la métrica de Fisher

En esta sección calculamos la métrica de Fisher para algunos de los modelos estadísticos más importantes. Utilizamos en los cálculos la fórmula (2.14). La métrica de Fisher de un modelo finito general es la matriz de la fórmula (2.65). Presentamos la métrica de Fisher para los modelos binomial, geométrico, Poisson, normal y exponencial.

2.6.1. Modelo binomial

La distribución de probabilidad de una variable aleatoria $X = B(n,x)$ binomial de parámetros $n \in \mathbb{N}$, $x \in (0,1)$ viene dada por

$$P(X = k; x) = \begin{cases} \binom{n}{k}x^k(1-x)^{n-k}, & \text{si } k \in \{0, \ldots, n\}, \\ 0, & \text{si } k \notin \{0, \ldots, n\}. \end{cases} \tag{2.90}$$

Tomamos el logaritmo y derivamos dos veces respecto del parámetro x

$$\log P(X=k;x) = \log \binom{n}{k} + \log x^k + (n-k)\log(1-x), \tag{2.91}$$

$$\partial_x \log P(X=k;x) = \frac{k}{x} - \frac{n-k}{1-x}, \quad \partial_x^2 \log P(X=k;x) = -\frac{k}{x^2} - \frac{n-k}{(1-x)^2}. \tag{2.92}$$

Entonces la métrica de Fisher de un modelo binomial es el escalar

$$\begin{aligned} g_{11}^F(x) &= -E[\partial_x^2 \log p(X=k;x)] = \sum_{k=0}^{n} \left(\frac{k}{x^2} + \frac{n-k}{(1-x)^2} \right) P(X=k;x) \\ &= \frac{1}{x^2} \sum_{k=0}^{n} kP(X=k;x) + \frac{1}{(1-x)^2} \left(n \sum_{k=0}^{n} P(X=k;x) - \sum_{k=0}^{n} kP(X=k;x) \right) = \frac{n}{x(1-x)}, \end{aligned} \tag{2.93}$$

donde se ha usado que la media de X es $E[X] = \sum_{k=0}^{n} kP(X=k;x) = nx$. En particular para $n=1$ tenemos el modelo Bernoulli y su correspondiente métrica de Fisher.

2.6.2. Modelo geométrico

La distribución de probabilidad de una variable aleatoria $X = G(p)$ geométrica de parámetro $p \in (0,1)$ viene dada por

$$P(X=k;p) = \begin{cases} p(1-p)^{k-1}, & \text{si } k \in \{1,2,\dots\}, \\ 0, & \text{si } k \notin \{1,2,\dots\}. \end{cases} \tag{2.94}$$

Tomamos el logaritmo y derivamos dos veces respecto del parámetro p

$$\log P(X=k;p) = \log p + (k-1)\log(1-p), \tag{2.95}$$

$$\partial_p \log P(X=k;p) = \frac{1}{p} + \frac{k-1}{p-1}, \quad \partial_p^2 \log P(X=k;p) = -\frac{1}{p^2} - \frac{k-1}{(p-1)^2}. \tag{2.96}$$

Entonces la métrica de Fisher de un modelo geométrico es el escalar

$$\begin{aligned} g_{11}^F(p) &= -E[\partial_p^2 \log P(X=k;p)] = \sum_{k=1}^{\infty} \left(\frac{1}{p^2} + \frac{k-1}{(p-1)^2} \right) P(X=k;p) \\ &= \frac{1}{p^2} + \frac{1}{(1-p)^2} \left(\frac{1-p}{p} \right) = \frac{1}{p^2(1-p)}, \end{aligned} \tag{2.97}$$

donde se ha usado que la media $E[X] = \sum_{k=1}^{\infty} kP(X=k;p) = \frac{1}{p}$.

2.6.3. Modelo de Poisson

La distribución de probabilidad de una variable aleatoria $X = P(\lambda)$ de Poisson de parámetro $\lambda > 0$ viene dada por

$$P(X=k;\lambda) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!}, & \text{si } k \in \{0,1,\dots\}, \\ 0, & \text{si } k \notin \{0,1,\dots\}. \end{cases} \tag{2.98}$$

Tomamos el logaritmo y derivamos dos veces respecto del parámetro λ

$$\log P(X=k;\lambda) = -\lambda + k\log\lambda - \log k!, \;\; \partial_\lambda \log P(X=k;\lambda) = -1 + \frac{k}{\lambda}, \;\; \partial_\lambda^2 \log P(X=k;\lambda) = -\frac{k}{\lambda^2}. \tag{2.99}$$

Entonces la métrica de Fisher de un modelo de Poisson es el escalar

$$g_{11}^F(\lambda) = -E[\partial_\lambda^2 \log P(X=k;\lambda)] = \sum_{k=0}^{\infty} \frac{k}{\lambda^2} P(X=k;\lambda) = \frac{1}{\lambda^2} \sum_{k=0}^{\infty} kP(X=k;\lambda) = \frac{1}{\lambda}, \tag{2.100}$$

donde se ha usado que la media de X es $E[X] = \sum_{k=0}^{\infty} kP(X=k;\lambda) = \lambda$.

2.6.4. Modelo normal

La distribución de probabilidad de una variable aleatoria $X = N(\mu,\sigma)$ normal de parámetros $\mu \in \mathbb{R}$, $\sigma > 0$ viene dada por la función de densidad

$$p(x;\xi) = p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}. \tag{2.101}$$

Tomamos el logaritmo y derivamos parcialmente respecto de μ y σ:

$$\log p(x;\mu,\sigma) = -\log\sigma - \log(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \tag{2.102}$$

$$\partial_\mu \log p(x;\mu,\sigma) = \frac{x-\mu}{\sigma^2}, \quad \partial_\sigma \log p(x;\mu,\sigma) = \frac{-1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}, \tag{2.103}$$

$$\partial_\mu^2 \log p(x;\mu,\sigma) = -\frac{1}{\sigma^2}, \;\; \partial_\sigma\partial_\mu \log p(x;\mu,\sigma) = -\frac{2(x-\mu)}{\sigma^3}, \;\; \partial_\sigma^2 \log p(x;\mu,\sigma) = \frac{1}{\sigma^2} - \frac{3}{\sigma^2}\left(\frac{x-\mu}{\sigma}\right)^2. \tag{2.104}$$

Entonces la métrica de Fisher de un modelo normal es la matriz cuyas entradas son

$$g_{11}^F(\mu,\sigma) = -E[\partial_\mu^2 \log p(x;\mu,\sigma)] = \int_{-\infty}^{\infty} \frac{1}{\sigma^2}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sigma^2}. \tag{2.105}$$

$$\begin{aligned} g_{12}^F(\mu,\sigma) = g_{21}^F(\mu,\sigma) &= -E[\partial_\mu\partial_\sigma \log p(x;\mu,\sigma)] = \int_{-\infty}^{\infty} -\frac{2(x-\mu)}{\sigma^3}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\infty}^{\infty} -\frac{2y}{\sigma^3}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2} dy = 0, \end{aligned} \tag{2.106}$$

ya que es la integral de una función impar en un intervalo simétrico.

$$\begin{aligned} g_{22}^F(\mu,\sigma) &= -E[\partial_\sigma^2 \log p(x;\mu,\sigma)] = -\int_{-\infty}^{\infty} \left(\frac{1}{\sigma^2} - \frac{3}{\sigma^2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= -\frac{1}{\sigma^2} + \frac{3}{\sigma^2}\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{2}{\sigma^2}, \end{aligned} \tag{2.107}$$

donde hemos usado el cambio de variable $z = \frac{x-\mu}{\sigma}$ y $\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 1$. En efecto, el integrando es función par en un intervalo simétrico y aplicando los cambios de variable $y = z^2$, $t = \frac{1}{2}y$, tenemos

$$\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{2}{\sqrt{\pi}}\int_0^{\infty} t^{\frac{3}{2}-1} e^{-t} dt = \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{2}{\sqrt{\pi}}\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 1, \tag{2.108}$$

ya que $\Gamma(n+1) = n\Gamma(n)$ y $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. En definitiva la métrica de Fisher es la matriz

$$\begin{pmatrix} g_{11}^F(\mu,\sigma) & g_{12}^F(\mu,\sigma) \\ g_{21}^F(\mu,\sigma) & g_{22}^F(\mu,\sigma) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \tag{2.109}$$

Haciendo el cambio de parametrización $\mu \to \frac{\mu}{\sqrt{2}}$ la métrica (2.109) se transforma en $\frac{2}{\sigma^2}I$, que es un múltiplo positivo de la métrica del plano hiperbólico. Análogamente la métrica de Fisher de un modelo $N(\mu,\Sigma)$ normal multivariante de dimensión n con μ vector de medias y $\Sigma = \sigma^2 I$ matriz de covarianzas múltiplo de la identidad es

$$\begin{pmatrix} \frac{1}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \frac{1}{\sigma^2} & 0 \\ 0 & \cdots & 0 & \frac{2n}{\sigma^2} \end{pmatrix}. \tag{2.110}$$

Haciendo el cambio de parametrización $\mu \to \frac{\mu}{\sqrt{2n}}$ la métrica (2.110) se transforma en $\frac{2n}{\sigma^2}I$, que es un múltiplo positivo de la métrica del espacio hiperbólico. Por tanto con una parametrización adecuada, el modelo estadístico normal con matriz de covarianzas múltiplo de la identidad dotado de la métrica de Fisher tiene la geometría del espacio hiperbólico y por tanto las mismas geodésicas.

2.6.5. Modelo exponencial

La distribución de probabilidad de una variable aleatoria $X = E(\lambda)$ exponencial de parámetro $\lambda > 0$ viene dada por la función de densidad

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } x > 0, \\ 0, & \text{si } x \leq 0. \end{cases} \tag{2.111}$$

Tomamos el logaritmo y derivamos dos veces respecto del parámetro λ

$$\log p(x;\lambda) = \log\lambda - \lambda x, \quad \partial_\lambda \log p(x;\lambda) = \frac{1}{\lambda} - x, \quad \partial^2_\lambda \log p(x;\lambda) = -\frac{1}{\lambda^2}. \tag{2.112}$$

Entonces la métrica de Fisher de un modelo exponencial es el escalar

$$g^F_{11}(\lambda) = -E_p[\partial^2_\lambda p(x;\lambda)] = -\frac{1}{\lambda^2}\int_0^\infty -\lambda e^{-\lambda x}dx = -\frac{1}{\lambda^2}\left[e^{-\lambda x}\right]_{x=0}^{\infty} = \frac{1}{\lambda^2}. \tag{2.113}$$

2.7. Geometría invariante

Una vez estudiadas las divergencias y métricas invariantes en variedades estadísticas, es el momento de estudiar conexiones invariantes en estas variedades. Utilizaremos la siguiente notación

$$D_{i,}(\xi) = \partial_{\xi^i_1} D(\xi_1 \mid\mid \xi_2)_{|\xi_1=\xi_2\equiv\xi}, \quad D_{,i}(\xi) = \partial_{\xi^i_2} D(\xi_1 \mid\mid \xi_2)_{|\xi_1=\xi_2\equiv\xi}, \tag{2.114}$$

$$D_{ij,k}(\xi) = \partial_{\xi^i_1}\partial_{\xi^j_1}\partial_{\xi^k_2} D(\xi_1 \mid\mid \xi_2)_{|\xi_1=\xi_2\equiv\xi}. \tag{2.115}$$

Sabemos que una divergencia D en una variedad M induce una métrica $g^{(D)}$ cuyas componentes son

$$g^{(D)}_{ij}(\xi) = D_{,ij}(\xi). \tag{2.116}$$

Las divergencias verifican $D(P \mid\mid Q) = 0$ si y solo si $P = Q$ y por tanto $g^{(D)}$ admite las siguientes expresiones alternativas para sus componentes

$$g^{(D)}_{ij}(\xi) = D_{ij,}(\xi) = D_{,ij}(\xi) = -D_{i,j}(\xi) = -D_{j,i}(\xi). \tag{2.117}$$

Si Γ^p_{ij} son las componentes de una conexión ∇ en una variedad Riemanniana (M,g), denotamos por $\Gamma_{ijk} = \Gamma^p_{ij}g_{kp}$ que dependen del punto ξ. Se puede demostrar que D induce dos conexiones $\nabla^{(D)}$, $\nabla^{(D^*)}$ en M por derivación mediante la fórmula para sus componentes en un sistema de coordenadas locales

$$\Gamma^{(D)}_{ijk}(\xi) = -D_{ij,k}(\xi) = -\partial_{\xi^i_1}\partial_{\xi^j_1}\partial_{\xi^k_2} D(\xi_1 \mid\mid \xi_2)_{|\xi_2=\xi_1\equiv\xi}, \tag{2.118}$$

$$\Gamma^{(D^*)}_{ijk}(\xi) = -D_{k,ij}(\xi) = -\partial_{\xi^k_1}\partial_{\xi^i_2}\partial_{\xi^j_2} D(\xi_1 \mid\mid \xi_2)_{|\xi_2=\xi_1\equiv\xi}, \tag{2.119}$$

que llamamos **conexiones inducidas**. Las conexiones inducidas $\nabla^{(D)}$, $\nabla^{(D^*)}$ son simétricas

$$\Gamma^{(D)}_{ijk}(\xi) = -D_{ij,k}(\xi) = -D_{ji,k}(\xi) = \Gamma^{(D)}_{jik}(\xi) \quad \text{luego} \quad \Gamma^p_{ij}(\xi) = \Gamma^p_{ji}(\xi), \tag{2.120}$$

ya que las derivadas parciales son intercambiables. Análogamente para los $\Gamma^{(D^*)}_{ijk}(\xi)$.

Definición 2.4. Sea (M,g) una variedad Riemanniana. Dos conexiones ∇,∇^* en M se dicen duales respecto de la métrica g si

$$Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y), \quad \forall X,Y,Z \in \mathscr{X}(M). \tag{2.121}$$

Eligiendo $X=\partial_i$, $Y=\partial_j$, $Z=\partial_k$ y tras unos cálculos la fórmula (2.121) se transforma en

$$\partial_k g_{ij} = \Gamma^p_{ki} g_{pj} + \Gamma^{r*}_{kj} g_{ir} = \Gamma_{kij} + \Gamma^*_{kji}. \tag{2.122}$$

En la definición 2.4 dada una conexión ∇ el miembro de la izquierda y el primer sumando del miembro de la derecha son conocidos. El segundo sumando del miembro de la derecha determina totalmente la conexión ∇^*, luego dada ∇ existe una única conexión dual ∇^*.

Proposición 2.3. *Sean ∇, ∇^* dos conexiones simétricas y duales en una variedad Riemanniana (M,g). Entonces la conexión de Levi-Civita que denotamos por $\nabla^{(0)}$ viene dada por*

$$\nabla^{(0)} = \frac{1}{2}(\nabla + \nabla^*). \tag{2.123}$$

Demostración. Es evidente que $\nabla^{(0)}$ es conexión porque es combinación convexa de dos conexiones. Si ∇, ∇^* son duales respecto de g, utilizando la relación de dualidad

$$\begin{aligned} Xg(Y,Z) &= \frac{1}{2}Xg(Y,Z) + \frac{1}{2}Xg(Y,Z) = \frac{1}{2}[g(\nabla_X Y,Z) + g(Y,\nabla^*_X Z)] + \frac{1}{2}[g(\nabla^*_X Y,Z) + g(Y,\nabla_X Z)] \\ &= g\left(\frac{\nabla_X Y + \nabla^*_X Y}{2}, Z\right) + g\left(Y, \frac{\nabla_X Z + \nabla^*_X Z}{2}\right) = g(\nabla^{(0)}_X Y, Z) + g(Y, \nabla^{(0)}_X Z), \end{aligned} \tag{2.124}$$

luego $\nabla^{(0)}$ es isométrica. Sean $T, T^*, T^{(0)}$ las torsiones de $\nabla, \nabla^*, \nabla^{(0)}$ respectivamente. Entonces

$$\begin{aligned} T^{(0)}(X,Y) &= \nabla^{(0)}_X Y - \nabla^{(0)}_Y X - [X,Y] = \frac{\nabla_X Y + \nabla^*_X Y}{2} - \frac{\nabla_Y X + \nabla^*_Y X}{2} - [X,Y] \\ &= \frac{1}{2}(\nabla_X Y - \nabla_Y X - [X,Y]) + \frac{1}{2}(\nabla^*_X Y - \nabla^*_Y X - [X,Y]) = \frac{1}{2}T(X,Y) + \frac{1}{2}T^*(X,Y) = 0, \end{aligned} \tag{2.125}$$

ya que ∇, ∇^* son simétricas, sus torsiones son $T = T^* = 0$, luego $\nabla^{(0)}$ es simétrica. □

Teorema 2.4. *Las conexiones inducidas $\nabla^{(D)}$, $\nabla^{(D^*)}$ son duales respecto de la métrica inducida $g^{(D)}$.*

Demostración. Las componentes de la métrica inducida $g^{(D)}$ son

$$g^{(D)}_{ij}(\xi) = -D_{i,j}(\xi). \tag{2.126}$$

Derivando respecto de ξ^k obtenemos la relación de dualidad

$$\partial_k g^{(D)}_{ij}(\xi) = -D_{ki,j}(\xi) - D_{i,kj}(\xi) = \Gamma^{(D)}_{kij}(\xi) + \Gamma^{(D^*)}_{kji}(\xi). \tag{2.127}$$

□

Proposición 2.4. *Sea D_f una f-divergencia. Las componentes de las conexiones inducidas por D_f en un sistema de coordenadas locales son*

$$\Gamma^{(D_f)}_{ijk}(\xi) = E_p\left[\left(\partial_i\partial_j \log p(x;\xi) + \frac{1-\alpha}{2}\partial_i \log p(x;\xi)\partial_j \log p(x;\xi)\right)(\partial_k \log p(x;\xi))\right], \tag{2.128}$$

$$\Gamma^{(D^*_f)}_{ijk}(\xi) = E_p\left[\left(\partial_i\partial_j \log p(x;\xi) + \frac{1+\alpha}{2}\partial_i \log p(x;\xi)\partial_j \log p(x;\xi)\right)(\partial_k \log p(x;\xi))\right], \tag{2.129}$$

con $\alpha = 3 + 2f'''(1) \in \mathbb{R}$.

Demostración. La técnica de la demostración es análoga a la de proposición 2.1. Apoyándonos en los cálculos de la proposición 2.1, derivamos respecto de ξ_1^k

$$\begin{aligned}\partial_{\xi_1^k}\partial_{\xi_2^i}\partial_{\xi_2^j}\left(p(x)f\left(\frac{q(x)}{p(x)}\right)\right) &= \partial_{\xi_1^k}p(x)f''\left(\frac{q(x)}{p(x)}\right)\frac{\partial_{\xi_2^i}q(x)\partial_{\xi_2^j}q(x)}{p(x)^2}\\ &- f'''\left(\frac{q(x)}{p(x)}\right)\frac{q(x)\partial_{\xi_1^k}p(x)\partial_{\xi_2^i}q(x)\partial_{\xi_2^j}q(x)}{p(x)^3}\\ &- f''\left(\frac{q(x)}{p(x)}\right)\frac{2\partial_{\xi_2^i}q(x)\partial_{\xi_2^j}q(x)\partial_{\xi_1^k}p(x)}{p(x)^2}\\ &+ \partial_{\xi_1^k}p(x)f'\left(\frac{q(x)}{p(x)}\right)\frac{\partial_{\xi_2^i}\partial_{\xi_2^j}q(x)}{p(x)}\\ &- f''\left(\frac{q(x)}{p(x)}\right)\frac{q(x)\partial_{\xi_1^k}p(x)\partial_{\xi_2^i}\partial_{\xi_2^j}q(x)}{p(x)^2}\\ &- f'\left(\frac{q(x)}{p(x)}\right)\frac{\partial_{\xi_2^i}\partial_{\xi_2^j}q(x)\partial_{\xi_1^k}p(x)}{p(x)}.\end{aligned} \tag{2.130}$$

Usando el intercambio (2.3), igualando $\xi_1=\xi_2$, es decir, $p(x)=q(x)$ y como $f(1)=0$ y $f''(1)=1$ entonces

$$\begin{aligned}\Gamma_{ijk}^{(D_f^*)}(\xi) &= -\partial_{\xi_1^k}\partial_{\xi_2^i}\partial_{\xi_2^j}D_f(\xi_1\,||\,\xi_2)_{|\xi_1=\xi_2\equiv\xi}\\ &= \int_{\mathscr{X}}\left(\frac{(1+f'''(1))\partial_i p(x)\partial_j p(x)\partial_k p(x)}{p(x)^2}+\frac{\partial_k p(x)\partial_i\partial_j p(x)}{p(x)}\right)dx,\end{aligned} \tag{2.131}$$

que coincide con (2.129) para $\alpha=3+2f'''(1)$. Se procede análogamente para $\Gamma_{ijk}^{(D_f)}(\xi)$. □

Por el teorema 2.4, la proposición 2.3 y la proposición 2.4 las componentes de la conexión de Levi-Civita en una variedad estadística son

$$\frac{\Gamma_{ijk}^{(D_f)}(\xi)+\Gamma_{ijk}^{(D_f^*)}(\xi)}{2} = E_p\left[\left(\partial_i\partial_j\log p(x;\xi)+\frac{1}{2}\partial_i\log p(x;\xi)\partial_j\log p(x;\xi)\right)\partial_k\log p(x;\xi)\right], \tag{2.132}$$

que coincide con (2.128) y (2.129) para $\alpha=0$. Denotamos por $\nabla^{(\alpha)}=\nabla^{(D_f)}$, $\nabla^{(-\alpha)}=\nabla^{(D_f^*)}$ y se denominan **α-conexiones**. En resumen, las α-conexiones forman una familia de conexiones simétricas, invariantes y duales respecto de la métrica de Fisher en las variedades estadísticas, obteniéndose la conexión de Levi-Civita para $\alpha=0$.

2.8. Aplicación en optimización

La geometría de la información tiene aplicación en el área de la optimización. Sea una función objetivo diferenciable $f:\mathbb{R}^n\to\mathbb{R}$ y consideramos el problema de minimización

$$(P)\quad \min_{x\in\mathbb{R}^n} f(x). \tag{2.133}$$

Para resolver (P) elegimos una familia de distribuciones de probabilidad paramétrica $M=\{p(x;\xi)\}$ tal que formen una variedad estadística. Resultados de estadística [10, 11, 12] permiten reemplazar el problema original (P) por el siguiente problema equivalente llamado relajación estocástica,

$$(R)\quad \min_{\xi\in\Xi} F(\xi), \tag{2.134}$$

donde $F : \Xi \to \mathbb{R}$ es $F(\xi) = E_p[f(x)]$. Los problemas (P), (R), son equivalentes en determinadas condiciones, tienen el mismo valor mínimo y podemos recuperar la solución de (P) a partir de la solución de (R) y viceversa. Resolvemos el problema (R) utilizando el método de descenso del gradiente en la variedad M. Este método consiste en seguir la dirección dada por el vector gradiente que es la de máxima variación para encontrar óptimos locales. Como el objetivo es minimizar tomaremos el gradiente cambiado de signo para encontrar un mínimo local. Podemos plantearlo de forma continua mediante un problema de valor inicial como sistema de ecuaciones diferenciales ordinarias

$$\begin{cases} \dot{\xi} = -\nabla_{\xi} F(\xi) \\ \dot{\xi}(0) = \xi_0. \end{cases} \tag{2.135}$$

Por el teorema de exitencia y unicidad de solución de problemas de valor inicial existe una única solución local $\xi(t)$, que es la curva que nos conduce desde ξ_0 a un mínimo local. Por un lado en el planteamiento anterior la resolución exacta no siempre es posible, por lo que daremos un método iterativo. Por otro lado el gradiente de una función f se define como el vector $\nabla f(x)$ tal que

$$\langle \nabla f(x), v \rangle = df(x)v, \tag{2.136}$$

donde $df(x)v$ es la diferencial de f en el punto x aplicada al vector v y $\langle\,,\,\rangle$ es un producto escalar, luego el gradiente depende de la métrica. Por el teorema de Chentsov existe una única métrica razonable en las variedades estadísticas, la métrica de la información de Fisher, que por tanto es la que elegimos. Entonces tenemos el siguiente gradiente, llamado **gradiente natural**

$$\tilde{\nabla}_{\xi} = \left(G^F\right)^{-1}(\xi)\frac{\partial}{\partial \xi}. \tag{2.137}$$

Discretizamos t proponiendo el siguiente método iterativo conocido como IGO, basado en el descenso del gradiente natural

$$\boxed{\xi_{t+1} = \xi_t - \lambda_t \tilde{\nabla}_{\xi_t} F(\xi_t), \quad \lambda_t > 0, \quad t = 0,1,2\ldots} \tag{2.138}$$

Dado un punto inicial ξ_0 calculamos el gradiente natural $\tilde{\nabla}_{\xi_0} F(\xi_0)$ que nos da una dirección para una recta. Sobre dicha recta siguiendo la dirección de descenso del gradiente, el valor de la función objetivo mejora hasta un cierto punto en que puede empeorar y calculamos ese punto. Se trata de encontrar el mínimo de una función sobre una recta, es decir, el mínimo de una función de una variable que puede resolverse de forma exacta o numérica mediante bisección o Newton-Raphson. Esto determina el valor de λ_0 y volviendo a (2.138) tenemos lo necesario para obtener el siguiente punto ξ_1 y seguir iterando.

El método IGO busca el siguiente punto sobre la recta determinada por el vector gradiente natural, pero las rectas no son una noción intrínseca de la variedad, dependen de la parametrización, del sistema de coordenadas. Sin embargo la generalización de las rectas, es decir, las geodésicas sí son intrínsecas a la variedad, no dependen de la parametrización. Cambiamos en IGO las rectas por geodésicas y obtenemos el método conocido como GIGO. Necesitamos definir la aplicación exponencial de una variedad.

Definición 2.5. Sea M una variedad, $p \in M$ un punto y $v \in T_pM$ un vector tangente a M en p. Entonces existe una única geodésica $\gamma = (x^1(t), \ldots, x^n(t))$ que pasa por el punto p con vector tangente v dada por la solución del sistema

$$\begin{cases} \ddot{x}^k + \Gamma^k_{ij}\dot{x}^i\dot{x}^j = 0, \\ x(0) = p, \quad \dot{x}(0) = v. \end{cases} \tag{2.139}$$

Entonces llamamos aplicación exponencial en M a

$$\exp_p : T_pM \to M; \quad v \to \exp_p(v) = \gamma(1). \tag{2.140}$$

Una vez definida la aplicación exponencial, el método iterativo GIGO viene dado por

$$\boxed{\xi_{t+1} = \exp_{\xi_t}(\lambda_t Y) \quad \text{donde} \quad Y = -\tilde{\nabla}_{\xi_t} F(\xi_t), \quad \lambda_t > 0, \quad t = 0,1,2\dots} \tag{2.141}$$

El cálculo exacto de geodésicas en general es complicado y recurrimos a resoluciones numéricas, sin embargo, para la variedad de las distribuciones normales, podemos calcularlas de forma exacta. Denotamos por $\mathbb{G}_n$ la variedad estadística de distribuciones de probabilidad normales multivariantes de dimensión n dotada con la métrica de Fisher y por $\tilde{\mathbb{G}}_n$ la variedad estadística de distribuciones de probabilidad normales multivariantes de dimensión n con matriz de covarianzas múltiplo de la identidad dotada con la métrica de Fisher. Por la sección 2.6.4 las geodésicas de $\tilde{\mathbb{G}}_n$ coinciden con las del espacio hiperbólico que vimos en el primer capítulo. El siguiente resultado es ahora inmediato:

Teorema 2.5 (Geodésicas en $\tilde{\mathbb{G}}_n$)**.** *Sea $\gamma : t \to N(\mu(t), \sigma(t)^2 I)$ una geodésica en $\tilde{\mathbb{G}}_n$. Entonces existen $a,b,c,d \in \mathbb{R}$ con $ad - bc = 1$, $v > 0$ tal que*

$$\mu(t) = \mu(0) + \sqrt{2n}\frac{\dot{\mu}_0}{\|\dot{\mu}_0\|} r(t), \quad \sigma(t) = Im(\gamma_{\mathbb{C}}(t)), \quad r(t) = Re(\gamma_{\mathbb{C}}(t)), \quad \gamma_{\mathbb{C}}(t) = \frac{aie^{vt} + b}{cie^{vt} + d}. \tag{2.142}$$

Para las geodésicas de $\mathbb{G}_n$, usando el teorema de Noether se demuestra que si $\gamma : t \to N(\mu_t, \Sigma_t)$ es una geodésica en $\mathbb{G}_n$ entonces las cantidades

$$J_\mu = \Sigma_t^{-1}\dot{\mu}_t, \quad J_\Sigma = \Sigma^{-1}(\dot{\mu}_t \mu_t^T + \dot{\Sigma}_t), \tag{2.143}$$

no dependen de t, son constantes a lo largo de las geodésicas. Esto permite reducir el orden de las ecuaciones de las geodésicas de 2 a 1, llegando a que $\gamma : t \to N(\mu_t, \Sigma_t)$ es una geodésica en $\mathbb{G}_n$ si y solo si $\mu : t \to \mu_t$ y $\Sigma : t \to \Sigma_t$ satisfacen las ecuaciones con condiciones iniciales:

$$\begin{cases} \dot{\mu}_t = \Sigma_t J_\mu, \\ \dot{\Sigma}_t = \Sigma_t(J_\Sigma - J_\mu \mu_t^T) = \Sigma_t J_\Sigma - \dot{\mu}_t \mu_t^T, \\ J_\mu = \Sigma_0^{-1}\dot{\mu}_0, \quad J_\Sigma = \Sigma_0^{-1}(\dot{\mu}_0 \mu_0^T + \dot{\Sigma}_0). \end{cases} \tag{2.144}$$

Las ecuaciones (2.144) se pueden resolver analíticamente. Usando la factorización de Cholesky de la matriz de covarianzas $\Sigma_t = A_t A_t^T$, las ecuaciones (2.144) se reescribien en términos de A_t y se pueden resolver de forma exacta, obteniéndose el siguiente resultado:

Teorema 2.6 (Geodésicas en $\mathbb{G}_n$)**.** *La geodésica en $\mathbb{G}_n$ con punto inicial $N(\mu_0, \Sigma_0 = A_0 A_0^T)$ y velocidad inicial $(\dot{\mu}_0, \dot{\Sigma}_0) \in T_{N(\mu_0, A_0 A_0^T)}\mathbb{G}_n$ viene dada por*

$$\exp_{N(\mu_0, A_0 A_0^T)}(s\dot{\mu}_0, s\dot{\Sigma}_0) = N(\mu_1, A_1 A_1^T), \tag{2.145}$$

con

$$\mu_1 = 2A_0 R(s) \sinh\left(\frac{sG}{2}\right) G^- A_0^{-1}\dot{\mu}_0 + \mu_0, \quad A_1 = A_0 R(s), \tag{2.146}$$

donde exp *es la aplicación exponencial de $\mathbb{G}_n$ y G es una matriz que satisface*

$$G^2 = A_0^{-1}(\dot{\Sigma}_0 \Sigma_0^{-1}\dot{\Sigma}_0 + 2\dot{\mu}_0 \dot{\mu}_0^T)(A_0^{-1})^T, \tag{2.147}$$

$$R(s) = \left(\left(\cosh\left(\frac{sG}{2}\right) - A_0^{-1}\dot{\Sigma}_0 (A_0^{-1})^T G^- \sinh\left(\frac{sG}{2}\right)\right)^{-1}\right)^T, \tag{2.148}$$

y G^- es una pseudo-inversa de G.

Los detalles se pueden consultar en [10].

En conclusión, el método de optimización GIGO utiliza geodésicas, que en general son complicadas de calcular y se suele recurrir a resoluciones numéricas. Eligiendo las distribuciones normales encontramos expresiones exactas para las geodésicas. La geometría de $\tilde{\mathbb{G}}_n$ es la del espacio hiperbólico y por tanto tienen las mismas geodésicas, que son bien conocidas en geometría diferencial. Para las geodésicas en $\mathbb{G}_n$ obtenemos expresiones exactas pero de una complejidad mayor.

Bibliografía

[1] S. AMARI, *Information geometry and its applications*, Springer, 2016.

[2] O. CALIN, C. UDRISTE, *Geometric modeling in probability and statistics*, Springer, 2014.

[3] S. AMARI, H. NAGAOKA, *Methods of information geometry*, Translations of mathematical monographs, vol 191, 2000.

[4] F. BRICKELL, R. S. CLARK, *Differentiable manifolds: an introduction*, Van Nostrand Reinhold, 1970.

[5] G. CASELLA, R. L. BERGER, *Statistical inference*, Duxbury, 2002.

[6] R. T. ROCKAFELLAR, *Convex analysis*, Princeton University Press, Princeton, NJ, 1970.

[7] N. N. CHENTSOV, *Statistical decision rules and optimal inference*, AMS, 1982 (originally published in Russian, Nauka, 1972).

[8] L. L. CAMPBELL, *An extended Chentsov characterization of the information metric*, Proocedings of American Mathematical Society, 98, 135-141, 1986.

[9] A. CATICHA, *The basics of information geometry*, Physics Department, University at Albany-SUNY, Albany, NY-12222, USA, 17 December 2014.

[10] J. BENSADON, *Black box optimization using geodesics in statistical manifolds*, Entropy, 17, 304-345, 2015.

[11] L. MALAGO, M. MATTEUCCI, G. PISTONE, *Towards the geometry of estimation of distribution algorithms based on the exponential family*, In Proceedings of the 11th Workshop Proceedings on Foundations of Genetic Algorithms, Schwarzenberg, Austria, 5-9 January 2011.

[12] L. MALAGO, G. PISTONE, *Information geometry of the gaussian distribution in view of stochastic optimization*, In Proceedings of the 15th Workshop Proceedings on Foundations of Genetic Algorithms, Aberystwyth, UK, 17-20 January 2015.