UQFC: Geometrizations of Statistical Models

Jeffrey Epstein

March 3, 2014

Smooth Geometries

As a wise Austrian once said, "Let's start at the very beginning / A very good place to start / When you read you begin with A-B-C / When you study smooth geometry you begin with smooth manifolds $\mathcal{M}$ and put things on them."

Riemannian Metrics

A Riemannian metric g is a smooth section of the $(0,2)$-tensor bundle on $\mathcal{M}$ (i.e. $g : \mathcal{M} \ni p \mapsto g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$) in which for all $p \in \mathcal{M}$ the element g_p associated to p satisfies the following three conditions for all elements $u, v, w \in T_p\mathcal{M}$ of the tangent space at p and all $\lambda_1, \lambda_2 \in \mathbb{R}$:

i) $g_p(u, v) = g_p(v, u)$ (symmetry)
ii) $v \neq 0 \to g_p(v, v) > 0$ (positive definite)
iii) $g_p(\lambda_1 u + \lambda_2 v, w) = \lambda_1 g_p(u, w) + \lambda_2 g_p(v, w)$ (bilinearity)

A Riemannian manifold or Riemannian geometry is a pair $(\mathcal{M}, g)$ for $\mathcal{M}$ a smooth manifold and g a Riemannian metric. A Riemannian metric allows us to define lengths of curves on $\mathcal{M}$. A curve in $\mathcal{M}$ is defined as a smooth map $c : \mathbb{R} \to \mathcal{M}$ and a finite curve as a smooth map $c : [0, 1] \to \mathcal{M}$. Every curve c induces a vecotr field $\dot{c}(t) = dc(t)/dt \in T_{c(t)}\mathcal{M}$. The length of a finite curve connecting points $p = c(1)$ and $q = c(0)$ is defined as

$$d_c(p, q) = \int_0^1 dt \sqrt{g_c(t)\left(\dot{c}(t), \dot{c}(t)\right)}$$

The Riemannian distance between points p and q in $\mathcal{M}$ with respect to a Riemannian metric g is defined to be the length of the shortest finite curve connecting p and q:

$$d_g(p, q) = \inf_c \{d_c(p, q) | c(0) = q, c(1) = p\}$$

This is a metrical distance. Putting a Riemannian metric on a smooth manifold $\mathcal{M}$ turns it into a metric space and allows us to take inner products (and hence norms) of elements of the tangent space at a single point, but gives us no way to compare elements of the tangent spaces at different points.

Affine Connections

An affine connection ∇ is a map $\mathcal{M} \ni p \mapsto \nabla_p : T_p\mathcal{M} \times T_p\mathcal{M} \to T_p\mathcal{M}$ that obeys the following conditions for all $\lambda_1, \lambda_2 \in \mathbb{R}$, functions/scalar fields $f, h : \mathcal{M} \to \mathbb{R}$, and sections $u, v, w \in T\mathcal{M}$ of the tangent bundle on $\mathcal{M}$. We will use the notation $\nabla(u, v) = \nabla_u v$:

i)$\nabla_u(fv + hw) = u(f)v + f\nabla_u v + u(h)w + h\nabla_u w$ (Linear and Leibnitz in the second argument, recalling that sections of the tangent bundle are maps from scalar fields to scalar fields)
ii) $\nabla_{fu+hv} w = f\nabla_u w + h\nabla_v w$ (Linear in the first argument)

These conditions suggest that $\nabla_u v$ is a directional derivative of a vector field v along another vector field u. We call this the covariant derivative of v along u.

Parallel transport: Let $u(t) \in T_{c(t)}\mathcal{M}$ be a vector field defined along a curve $c(t)$. If $\nabla_{\dot{c}} u(t) = 0$, we say that u is parallelly transported along c. If this is the case, we define $t_c^{\nabla} u \equiv t^{\nabla}_{c(0),c(1)} := u(t_1) \in T_{c(t_1)}\mathcal{M}$ to be the parallel transport of $u = u(t_0) \in T_{c(t_0)}$ with respect to ∇. This is how we *define* what it means for two elements of different tangent spaces to be equal, since all we can a priori require is that the null elements of these spaces must be identified. A curve $c(t)$ is called a ∇-geodesic iff $\nabla_{\dot{c}(t)} \dot{c}(t) = 0$.

The Riemann-Christoffel curvature tensor for an affine connection ∇ is a map $R^{\nabla} : T\mathcal{M} \times T\mathcal{M} \times T\mathcal{M} \to T\mathcal{M}$. The torsion tensor of an affine connection ∇ is a map $T^{\nabla} : T\mathcal{M} \times T\mathcal{M} \to T\mathcal{M}$. An affine connection is torsion-free or symmetric if $T^{\nabla}(u, v) = 0$ for all u, v and flat if $R^{\nabla}(u, v, w) = 0$ for all u, v, w.

An affine manifold or affine geometry is a pair $(\mathcal{M}, \nabla)$ for ∇ flat and torsion-free. Equipping a smooth manifold $\mathcal{M}$ with such a connection allows us to compare elements of the tangent spaces at different points, but gives us no notion of distances between points on the manifold.

Metric-Affine Geometries

A triple $(\mathcal{M}, g, \nabla)$ where g is a Riemannian metric and ∇ is an affine connection is called a metric-affine geometry. Equipped with both a metric and a connection on the manifold, we are able both to define distances between points and to compare elements of the tangent spaces at different points. Note that the metric and the connection are a priori independent of each other - we have defined the two objects separateley.

A metric-compatible connection is one that satisfies the equivalent conditions

$$\begin{aligned} \nabla_u g(v, w) &= 0 \\ g(t_c^{\nabla} u, t_c^{\nabla} v) &= g(u, v) \forall u, v \in T\mathcal{M} \\ g(\nabla_u v, w) + g(v, \nabla_u w) &= u(g(v, w)) \end{aligned}$$

for all sections $u, v, w \in T\mathcal{M}$ of the tangent bundle on $\mathcal{M}$ and all curves c.

Every Riemannian manifold $(\mathcal{M}, g)$ defines a unique torsion-free metric-compatible affine connection called the Levi-Civita connection, giving a unique metric-affine geometry $(\mathcal{M}, g, \nabla_{\mathrm{LC}})$.

Norden-Sen Duals

A pair $(\nabla, \nabla^{\dagger})$ of affine connections over a smooth manifold $\mathcal{M}$ is called Norden-Sen dual with respect to a Riemannian metric g iff

$$g(t_c^{\nabla} u, t_c^{\nabla^{\dagger}} v) = g(u, v)$$

for all vector fields u, v and all curves c. A quadruple $(\mathcal{M}, g, \nabla, \nabla^{\dagger})$ is called a Norden-Sen manifold or Norden-Sen geometry. We have the equality

$$R^{\nabla}(u, v, w) = R^{\nabla^{\dagger}}(u, v, w)$$

Eguchi Geometries

A Norden-Sen geometry $(\mathcal{M}, g, \nabla, \nabla^{\dagger})$ with torsion-free connections is called an Eguchi Geometry. This quadruple defines a distance $D : \mathcal{M} \times \mathcal{M} \to [0, \infty]$, which is unique up to the third order term in its Taylor expansion.

We can also go the other way. The directional derivative at a point $p \in \mathcal{M}$ in the direction $v \in T_p\mathcal{M}$ is defined as $\mathcal{D}_v|_p F = \frac{d}{d\tau} F(u + \tau v)|_{\tau=0}$. Consider a triple-differentiable distance function D on $\mathcal{M}$ that satisfies $\mathcal{D}_v|_p\mathcal{D}_v|_pD(p,q)|_{q=p} \in (0,\infty)$ for all $p \in \mathcal{M}$ and $v \in T_p\mathcal{M}\backslash\{0\}$.

The directional derivative $\mathcal{D}_p^{(v)}$ is defined as $\mathcal{D}_p^{(v)}f(p) = \frac{d}{dt}f(p+vt)|_{t=0}$. We can apply this to a two-input function such as a distance $D(p,q)$ by fixing q as follows:

$$\mathcal{D}_p^{(v)}D(p,q)|_{p=q} = \frac{d}{dt}D(q+vt,q)|_{t=0}$$

We don't have to do this fixing $p=q$, of course. If we don't, we have

$$\mathcal{D}_p^{(v)}D(p,q) = \frac{d}{dt}D(p+vt,q)|_{t=0}$$

which is still a two-input function, and tells us how fast the distance between two points varies as we change the first one in a particular direction (remember that the distance need not be symmetric).

Consider a distance D on a smooth manifold $\mathcal{M}$ that is triple-differentiable and satisfies the condition

$$\mathcal{D}_p^{(v)}\mathcal{D}_p^{(v)}D(p,q)|_{p=q} \in (0,\infty) \quad \forall p \in \mathcal{M} \quad \forall v \in T_p\mathcal{M}\backslash\{0\}$$

This quantity is a function $\mathcal{M} \to \mathbb{R}$ which tells us the second derivative of the distance between two points as we move one away from the other.

Eguchi showed that given a smooth finite-dimensional manifold $\mathcal{M}$ and such a distance D (not necessarily symmetric!), a Riemannian metric g and a dual pair of torsion-free affine connections $(\nabla, \nabla^\dagger)$ on $\mathcal{M}$ are defined by the Eguchi equations:

$$\begin{aligned} g_p(u,v) &= \mathcal{D}_p^{(u)}\mathcal{D}_q^{(v)}D(p,q)|_{p=q} \\ g_p((\nabla_u)_pv,w) &= -\mathcal{D}_p^{(u)}\mathcal{D}_p^{(v)}\mathcal{D}_q^{(w)}D(p,q)|_{p=q} \\ g_p(v,(\nabla_u^\dagger)_pw) &= -\mathcal{D}_q^{(u)}\mathcal{D}_q^{(w)}\mathcal{D}_p^{(v)}D(p,q)|_{p=q} \end{aligned}$$

These define a torsion-free Norden-Sen geometry $(\mathcal{M},g,\nabla,\nabla^\dagger)$, which is an Eguchi geometry.

Proof of Eguchi Equations

Consider a smooth distance function (Eguchi calls this a contrast function) $\rho : \mathcal{M}\times\mathcal{M} \to \mathbb{R}$ such that $\rho(p,q) \geq 0$ for all $p,q \in \mathcal{M}$ with equality iff $p=q$. Define

$$\rho(X_1\ldots X_n|Y_1\ldots Y_m)(z) = (X_1)_p\ldots(X_n)_p(Y_1)_q\ldots(Y_m)_q\rho(p,q)|_{p,q=z}$$

Then $\rho(X_1\ldots X_n|Y_1\ldots Y_m)$ is a map $\mathcal{M} \to \mathbb{R}$. However, via this definition we can also consider ρ itself to be a map from pairs consisting of an n-tuple and an m-tuple of sections of the tangent bundle on $\mathcal{M}$ to scalar fields. The following equality holds because $\rho(x,y)$ has a minimum at $x=y$:

$$\rho(Y|\cdot)(z) = Y_p\rho(p,q)|_{p,q=z} = 0 \to \rho(Y|\cdot) = 0 \quad \forall Y \in T\mathcal{M}$$

Similarly, $\rho(\cdot|Y) = 0$ for all $Y \in T\mathcal{M}$. Using this result, we have

$$0 = X_p\left[\rho(Y|\cdot) + \rho(\cdot|Y)\right] = X_pY_p\rho(p,q) + X_pY_q\rho(p,q) = \rho(XY|\cdot) + \rho(X|Y)$$

yielding the relation

$$\rho(XY|\cdot) = -\rho(X|Y)$$

Let's consider a distance function such that $\rho(XX|\cdot) > 0$ for all $X \neq 0$ and define a function from pairs of sections of the tangent bundle on $\mathcal{M}$ to real-valued fields on $\mathcal{M}$:

$$g(X,Y) = -\rho(X|Y)$$

g is symmetric:

$$g(X,Y) - g(Y,X) = \rho(Y|X) - \rho(X|Y) = \rho(XY|\cdot) - \rho(YX|\cdot) = \rho([X,Y]\,|\cdot) = 0$$

g is bilinear:

$$g(\lambda X, Y) = \rho(\lambda X|Y) = -\lambda\rho(X|Y) = \lambda g(X,Y)$$

g is positive definite:

$$g(X,X) = -\rho(X|X) = \rho(XX|\cdot) > 0 \quad \forall X \neq 0$$

by assumption. We've shown that g is a symmetric, bilinear, positive definite map from pairs of sections of the tangent bundle on $\mathcal{M}$ to scalar fields on $\mathcal{M}$. This is a Riemannian metric! Or is it? We've defined a map from pairs of sections of the tangent bundle to scalar fields, but how do we know that the scalar fields agree at points at which X and Y agree? Suppose (X', Y') are vector fields such that $(X', Y')(z) = (X,Y)(z)$. Then

$$g(X', Y')(z) = -\rho(X'|Y')(z) = -X'_p Y'_q \rho(p,q)|_{p,q=z} = -X_p Y_q \rho(p,q)|_{p,q=z} = g(X,Y)(z)$$

So we're safe, and this is indeed a Riemannian metric.

The connections are defined via the equations

$$\begin{aligned} g(\nabla_X Y, Z) &= -\rho(XY|Z) \\ g(\nabla^*_X Y, Z &= -\rho(Z|XY) \end{aligned}$$

which are required to hold for all $Z \in T\mathcal{M}$. They are bilinear by definition. Moreover, we have

$$\begin{aligned} g(\nabla_{fX} Y, Z) &= -\rho((fX)Y|Z) = -f\rho(XY|Z) = fg(\nabla_X Y, Z) = g(f\nabla_X Y, Z) \\ g(\nabla_X fY, Z) &= -\rho(XfY|Z) = -\rho((Xf)Y + fXY|Z) = -(Xf)\rho(Y|Z) - f\rho(XY|Z) \\ &= (Xf)g(Y,Z) + fg(\nabla_X Y, Z) = g((Xf)Y + f\nabla_X Y, Z) \end{aligned}$$

which allows us to identify

$$\begin{aligned} \nabla_{fX} Y &= f\nabla_X Y \\ \nabla_X fY &= (Xf)Y + f\nabla_X Y \end{aligned}$$

The same relations hold for ∇^*, and these are precisely the defining properties of an affine connection, so ∇ and ∇^* are indeed affine connections on $\mathcal{M}$.

∇ and ∇^* are also metric-compatible:

$$Xg(Y,Z) = -X\rho(Y|Z) = -\rho(XY|Z) - \rho(Y|XZ) = g(\nabla_X Y, Z) + g(\nabla^*_X Y, Z)$$

Using the symmetry of g:

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

which is the condition for metric compatibility! Now note that

$$g(\nabla_X Y - \nabla_Y X, Z) = -\rho(XY - YX, Z) = g([X,Y], Z)$$

for all $Z \in T\mathcal{M}$. This implies

$$\nabla_X Y - \nabla_Y X - [X,Y] = T^{\nabla}(X,Y) = 0$$

The same holds for ∇^*, so both connections are torsion-free. Therefore the quadruple $(\mathcal{M}, g, \nabla, \nabla^*)$ is an Eguchi geometry. In addition, $\bar{\nabla}$ is torsion-free, and because there is a unique torsion-free metric compatible affine connection defined by a Riemannian metric, $\bar{\nabla}$ is the Levi-Civita connection for g.

Consider the case of a symmetric distance $\rho(p, q) = \rho(q, p)$. Then

$$\rho(A|B)(z) = A_p B_q \rho(p, q)|_{p,q=z} = A_p B_q \rho(q, p)|_{q,p=z} = A_q B_p \rho(p, q)|_{p,q=z} = \rho(B|A)(z)$$

Then

$$g(\nabla_X Y, Z) = -\rho(XY|Z) = -\rho(Z|XY) = g(\nabla^*_X Y, Z) \quad \forall Z \in T\mathcal{M}$$

This implies $\nabla = \nabla^* = \bar{\nabla}$, so the symmetric distance case reduces to a the Riemannian geometry equipped with the Levi-Civita connection. Notice what we (where we are Shinto Eguchi) have done here. We've started with a smooth manifold and a possibly asymmetric distance function, and we've built a Riemannian geometry on the manifold, which induces a symmetric distance function!

Statistical Models

A premeasurable space $(\chi, \Omega(\chi))$ consists of a set χ and a σ-algebra $\Omega(\chi)$ of subsets. A (σ-)algebra of subsets is a collection of sets closed under complementation, finite (countable) unions, and finite intersections.

A countably additive measure on a premeasurable space is a function $\mu : \Omega(\chi) \to [0, \infty]$ such that $\mu(\varnothing) = 0$ $\mu(\cup_{i\in\mathbb{N}} E_i) = \sum_{i\in\mathbb{N}} \mu(E_i)$. Define $\mathrm{Meas}^+(\chi, \Omega(\chi))$ to be the set of all countably additive positive measures on the premeasurable space $(\chi, \Omega(\chi))$. A statistical model is a subset $\mathcal{M}(\chi, \Omega(\chi)) \subseteq \mathrm{Meas}^+(\chi, \Omega(\chi))$.

Radon-Nikodym Derivatives

Given two measures μ and ν on a premeasurable set, we say that ν is absolutely continuous with respect to μ if ν assigns measure zero to all sets assigned measure zero by μ. This means essentially that when we use the measure ν to integrate over a set, there are no "jumps" in the integral compared to the integral using μ. The Radon-Nikodym Theorem states that any complex measure λ absolutely continuous with respect to a positive measure μ may be expressed as $\lambda(E) = \int_E f\mu$ for some $L_1(\mu)$ function f, called the Radon-Nikodym derivative $d\lambda/d\mu$. Recall (or learn) that $L_p(\mu)$ is the space of functions f such that

$$\left(\int_\chi |f|^p \mu\right)^{1/p} < \infty$$

These derivatives can be thought of as "measure densities" on the space χ. In the spirit of full disclosure, I don't understand how these integrals work. How what if the algebra is just $\{\varnothing, \chi\}$ or something similarly coarse?

We will assume that there is a countably additive measure $\tilde{\mu}$ on our space $(\chi, \Omega(\chi))$ so that $\mathcal{M}(\chi, \Omega(\chi))$ may be represented as a set of Radon-Nikodym derivatives, which are just $L_1(\tilde{\mu})$ functions on the set. We write

$$\mathcal{M}(\chi, \Omega(\chi)) \cong \mathcal{M}(\chi, \Omega(\chi), \tilde{\mu}) \subseteq L_1(\chi, \Omega(\chi), \tilde{\mu})^+$$

Probabilistic models are simply statistical models all of whose elements are normalized to 1

$$\mathcal{M}(\chi, \Omega(\chi), \tilde{\mu}) \subseteq S(\chi, \Omega(\chi)\tilde{\mu}) = \{p \in L_1(\chi, \Omega(\chi), \tilde{\mu})^+ | \int_\chi \tilde{\mu} p = 1\}$$

The full set S is known as a probability simplex.

Finite Models

For finite sets χ with n elements and $\Omega(\chi) = 2^{\chi}$, $\mathrm{Meas}^{+}(\chi, \Omega(\chi))$ is isomorphic to $\mathbb{R}^n$. The finite probability simplex is

$$S(\chi, 2^{\chi}) = \{(p_1, \ldots, p_n) \in [0,1]^n \,|\, \sum_{i=1}^{n} p_i = 1\}$$

Exponential Models

An important class of statistical models is the exponential family, defined as

$$\mathcal{M}(\chi, \Omega(\chi), \tilde{\mu}) = \{\exp\left(-\log Z(\theta) - \theta^i f_i(x)\right) | \theta = (\theta^1, \ldots, \theta^n) \in \Omega \subseteq \mathbb{R}^n\}$$

where the $f_i : \chi \to \mathbb{R}$ are a linearly independent set of functions all also linearly independent from $f = 1$. Z is a normalization factor, sometimes called a partition function. Exponential coordinates are the maps

$$\theta^i : \mathcal{M}_{\exp}(\chi, \Omega(\chi), \tilde{\mu}) \ni p \mapsto \theta(p) \in \Theta \subseteq \mathbb{R}^n$$

Many important classes of probability distributions are exponential families, such as normal and multivariate normal distributions, Poisson distributions, and the complete set of probability distributions on a finite set.

Coarse Grainings

A coarse graining is a positive linear function

$$T_\star : L_1(\mathcal{A}_2, \mu_2) \to L_1(\mathcal{A}_1, \mu_1)$$

such that $\|f\| = \|T_\star(f)\|$ for all $f \in L_1(\mathcal{A}_2, \mu_2)^+$. This condition may be equivalently stated

$$\int \mu_1 T_\star(f) = \int \mu_2 f$$

We only impose this condition on the positive cone $L_1(\mathcal{A}_2, \mu_2)^+$ because that's what we care about, and we want to be as general as possible. Finite coarse grainings between finite probability simplices S_1 and S_2 with sample spaces of the same size n correspond to $n \times n$ right stochastic matrices, which are matrices of non-negative real entries with rows summing to one. In this representation, probability distributions are row vectors and $T_\star$ acts by right multiplication.

f-Distances

For any set X, a distance is a map $D : X \times X \to [0, \infty]$ such that $D(x, y) = 0 \leftrightarrow x = y$. It is bounded if it takes only finite values. It may be symmetric, and it is metrical if it is bounded, symmetric, and satisfies the triangle inequality. A statistical distance is a distance on a statistical model $\mathcal{M}(\mathcal{A}) \subseteq L_1(\mathcal{A})^+$.

For a non-empty convex subset C of a vector space X, a function $f : C \to \mathbb{R}$ is called convex if

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

and strictly convex if the inequality is strict. Consider a function $f : \mathbb{R}^+ \to \mathbb{R}$ convex on $(0, \infty)$ with $f(1) = 0$ and strictly convex at 1. For such a function f, the f-divergence or f-distance is a map $D_f : \mathcal{M}(\mathcal{A}) \times \mathcal{M}(\mathcal{A}) \to [0, \infty]$ defined as (now using Wikipedia's language)

$$D_f(P \| Q) = \int_\Omega f\left(\frac{dP}{dQ}\right) dQ = \int_\Omega f\left(\frac{p}{q}\right) q\mu$$

where P and Q are probability distributions with P absolutely continuous with respect to Q. The second equality follows if P and Q are both absolutely continuous with respect to some dominating measure μ.

Here, p and q are the Radon-Nikodym derivatives of P and Q with respect to μ. The condition $f(1) = 0$ implies $D(P\|P) = 0$.

The f-distance is non-negative, since by Jensen's inequality

$$D_f(P\|Q) = \int_\Omega f\left(\frac{p}{q}\right) q\mu \geq f\left(\int_\Omega \frac{p}{q} q\mu\right) = f(1) = 0$$

The f-distance is jointly convex in both variables, i.e

$$D_f(\lambda\omega_1 + (1-\lambda)\omega_2, \lambda\phi_1 + (1-\lambda)\phi_2) \leq \lambda D_f(\omega_1, \phi_1) + (1-\lambda) D_f(\omega_2, \phi_2)$$

for all $\lambda \in [0, 1]$.

The f-distance is also monotonic with respect to coarse-graining:

$$D_f(\omega, \phi) \geq D(T_\star(\omega), T_\star(\phi))$$

An example of an f-distance is the total variation distance, given by $f(x) = |x - 1|$:

$$D_{\text{TV}}(P\|Q) = \int_\Omega f\left(\frac{p}{q}\right) q\mu = \int_\Omega \left|\frac{p-q}{q}\right| q\mu = \int_\Omega |p - q|\mu$$

The Kullback-Leibler distance can also be obtained this way using $f(x) = x\ln(x)$:

$$D_{\text{KL}}(P\|Q) = \int_\Omega f\left(\frac{p}{q}\right) q\mu = \int_\Omega \frac{p}{q} \ln\left(\frac{p}{q}\right) q\mu = \int_\Omega p \ln\left(\frac{p}{q}\right) \mu$$

Stepping away from the nirvanic peace of pure form, let's see what the monotonicity with respect to coarse graining means in the case of probability distributions. Let p and q be probability distributions in a statistical model, and let $\kappa = \{\kappa(y|x) \geq 0; x \in X, y \in Y\}$ be an arbitrary transition probability distribution. Denote by p_κ and q_κ the probability distributions on Y induced from p and q by the mapping κ. Then we have

$$\begin{aligned}
D_f(p\|q) &= \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx = \int q(x) \left(\int \kappa(y|x) dy\right) f\left(\frac{p(x)}{q(x)}\right) dx = \int\int q(x)\kappa(y|x) f\left(\frac{p(x)}{q(x)}\right) dxdy \\
&= \int\int q_\kappa(y) q_\kappa(x|y) f\left(\frac{p(x)}{q(x)}\right) dxdy = \int q_\kappa(y) \left(\int q_\kappa(x|y) f\left(\frac{p(x)}{q(x)}\right) dx\right) dy \\
&\geq \int q_\kappa(y) f\left(\int q_\kappa(x|y) \frac{p(x)}{q(x)} dx\right) dy = \int q_\kappa(y) f\left(\int \frac{\kappa(y|x)q(x)}{q_\kappa(y)} \frac{p(x)}{q(x)} dx\right) dy \\
&= \int q_\kappa(y) f\left(\frac{1}{q_\kappa(y)} \int \kappa(y|x) p(x) dx\right) dy = \int q_\kappa(y) f\left(\frac{p_\kappa(y)}{q_\kappa(y)}\right) dy = D_f(p_\kappa\|q_\kappa)
\end{aligned}$$

where I've used $\kappa(y|x)q(x) = q_\kappa(x|y)q_\kappa(y)$ and Jensen's inequality. What this means is that the f-distance reflects our intuition that you ought not be able to distinguish more easily between probability distributions by performing operations on the measurements.

Interlude: Information Theory

Mutual Information

The mutual information of two random variables X and Y with joint pdf $p(x, y)$ and marginal pdfs $p(x)$ and $p(y)$ is the Kullback-Leibler distance between the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$:

$$I(X;Y) = D_{\text{KL}}(p(x,y)\|p(x)p(y)) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Note that although $D_{\mathrm{KL}}(p,q)$ is not symmetric in p and q, the mutual information $I(X;Y)$ is symmetric in X and Y.

To illustrate this concept, let's consider the case of finite sample spaces $\mathcal{X}$ and $\mathcal{Y}$ with joint probababilities $p(x,y)$. First, consider the case in which there is a one-to-one pairing between $\mathcal{X}$ and $\mathcal{Y}$, so that whenever x_i is selected, so is y_i. Then

$$I(X;Y) = \sum_i p_i \log \frac{p_i}{p_i^2} = -\sum_i p_i \log p_i = H(X) = H(Y)$$

which is the Shannon entropy of the individual distributions. Now consider the case in which the selection of x is totally independent from the selection of y. In other words, $p(x,y) = p(x)p(y)$. Then

$$I(X;Y) = \sum_{ij} p_i p_j \log \frac{p_i p_j}{p_i p_j} = 0$$

So in some sense the mutual information tells us how much we can learn about the outcome of experiments governed by one probability distribution by looking at the results of another experiment governed by another probability distribution.

Data Processing

Markov Chain: Random variables X, Y, Z form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is conditionally independent of X, i.e., if we can write $p(x,y,z) = p(x)p(y|x)p(z|y)$.

Data Processing Inequality: If $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$. In particular, if $Z = g(Y)$, then $X \to Y \to g(Y)$ and $I(X;Y) \geq I(X;g(Y))$.

Sufficient Statistics: Consider two random variables X and Y with sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and probability distributions $p(x;\xi)$ and $q(y;\xi)$ drawn from some statistical model parametrized by ξ. Now consider a map $F : \mathcal{X} \to \mathcal{Y}$. F and $p(x;\xi)$ together determine $q(x;\xi)$. Defining $r(x;\xi) = p(x;\xi)/q(F(x);\xi)$:

$$p(x;\xi) = q(y;\xi)p(x|y;\xi) \to p(x|y;\xi) = \frac{p(x;\xi)}{q(y;\xi)} = r(x;\xi)\delta(y - F(x))$$

If for all events $A \subseteq \mathcal{X}$ and all $y \in \mathcal{Y}$ $Pr(A|y;\xi)$ does not depend on ξ, or equivalently $r(x;\xi)$ does not depend on ξ, F is called a sufficient statistic for the model S parametrized by ξ. Roughly, this means that the sample gives no more information about ξ than does the statistic. Let some statistic F be sufficient. Then

$$p(x;\xi) = q(y;\xi)p(x|y;\xi) = q(y;\xi)r(x;\xi)\delta(y - F(x)) = q(F(x);\xi)r(x;\xi) = q(F(x);\xi)r(x)$$

so the ξ-dependence of p is contained entirely in the distribution q of the RV Y. Therefore, if our goal is to estimate ξ, we get no more information from knowing X than from knowing Y. In fact, the above condition is both necessary and sufficient for sufficiency. This is known as the factorization theorem. Returning to the data-processing inequality, we have $I(\Xi;X) = I(\Xi;T(X))$ for a sufficient statistic T.

Heading Back to Geometry

γ-Distances

A subset of the f-distances are the γ-distances, which are defined by

$$f_\gamma(t) = \begin{cases} \frac{1}{\gamma} + \frac{1}{1-\gamma}t - \frac{1}{\gamma(1-\gamma)}t^\gamma & \gamma \in (0,1) \\ t\log t - (t-1) & \gamma = 1 \\ -\log t + (t-1) & \gamma = 0 \end{cases}$$

For $\gamma \in (0,1)$, the distance D_γ satisfies the generalized cosine equation

$$D_\gamma(\omega,\psi) = D_\gamma(\omega,\phi) + D_\gamma(\phi,\psi) - \frac{1}{\gamma(1-\gamma)} \int (\omega^\gamma - \phi^\gamma)(\psi^{1-\gamma} - \phi^{1-\gamma})$$

Generalized Pythagorean Theorem

Given a statistical model $\mathcal{M}$ and a submodel Q, define

$$P_Q^D(p) = \operatorname{arginf}_{q\in Q}\{D(p,q)\}$$

The generalized Pythagorean theorem states

$$D(q, P_Q^D(p)) + D(P_Q^D(p), p) = D(q,p) \quad \forall (p,q) \in Q \times \mathcal{M}$$

This does not hold for all submodels Q and distances D, but does hold for, for example, $D = D_{\mathrm{KL}}$ and Q an exponential model such as Gaussians.

A Simpler Derivation of Eguchi

Consider a distance function D and Taylor expand:

$$\begin{aligned} D(P(\theta)\|P(\theta_0)) &= D(P(\theta_0)\|P(\theta_0) + \Delta\theta^i \partial_i D(P(\theta)\|P(\theta_0)|_{\theta=\theta_0} + \Delta\theta^i\Delta\theta^j \partial_i\partial_j D(P(\theta)\|P(\theta_0)|_{\theta=\theta_0} + \mathcal{O}(\Delta\theta^3) \\ &= \Delta\theta^i\Delta\theta^j \partial_i\partial_j D(P(\theta)\|P(\theta_0)|_{\theta=\theta_0} + \mathcal{O}(\Delta\theta^3) \\ &:= \Delta\theta^i\Delta\theta^j g_{ij}(\theta_0) \end{aligned}$$

Fisher Metric

The Fisher metric on a statistical manifold is defined using this relation from the Kullback-Leibler distance.

$$g_{ij}(\theta) = \int_{\mathcal{X}} \partial_i \left(\ln p(x;\theta)\right) \partial_j \ln\left(p(x;\theta)\right) p(x;\theta)$$

This metric is monotonic with respect to mappings $F : \mathcal{X} \to \mathcal{Y}$ is the sense that $g_F(\theta) \leq g(\theta)$, i.e. $g - g_F$ is positive semidefinite. A necessary and sufficient condition for equality is that F be a sufficient statistic. The Fisher metric also satisfies an additivity condition. Given $p_{12}(x_1,x_2;\theta) = p_1(x_1;\theta)p_2(x_2;\theta)$ it is the case that $g_{12}(\theta) = g_1(\theta) + g_2(\theta)$.

α-Connections

An affine connection $\nabla^{(\alpha)}$ may be defined on a statistical manifold by

$$g_p\left((\nabla^{(\alpha)}_{\partial_i})_p \partial_j, \partial_k \right) = \int \left(\partial_i\partial_j \ln p(x;\theta) + \frac{1-\alpha}{2} \partial_i \ln p(x;\theta) \partial_j \ln p(x;\theta) \right) \partial_k \ln p(x;\theta) p(x;\theta)$$

A nice relation is

$$\nabla^{(\alpha)} = (1-\alpha)\nabla^{(0)} + \alpha\nabla^{(1)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}$$

The 0-connection is Riemannian with respect to the Fisher metric.

Cencov's Theorem

We might wonder if the Fisher metric and the α-connection are special. In fact, they are. In particular, if F is a sufficient statistic for S, then these objects are invariant under F. That is to say:

$$\begin{aligned} \langle X, Y\rangle_p &= \langle \lambda_*(X), \lambda_*(Y)\rangle'_{\lambda(p)} \\ \lambda_*\left(\nabla^{(\alpha)}_X Y\right) &= \nabla'^{(\alpha)}_{\lambda_*(X)} \lambda_*(Y) \end{aligned}$$

for all $X, Y \in \mathcal{T}(S)$. Here, the Fisher connection is denoted by $\langle \cdot, \cdot \rangle$. λ is a diffeomorphism from S to S_F and $\lambda_* : \mathcal{T}(S) \to \mathcal{T}(S_F)$ is defined by

$$(\lambda_*(X))_{\lambda(p)} = (d\lambda)_p(X_p)$$

Cencov's theorem states that given a sequence $\{(g_n, \nabla_n)\}_{n=1}^{\infty}$ defined on the statistical manifolds $\mathcal{P}(\mathcal{X}_n)$, if for all $n, m, S \subset \mathcal{P}_n$ and $F : \mathcal{X}_n \to \mathcal{X}_m$ for $n \geq m$ and F a sufficient statistic for S the induced metrics and connections on S and S_F are invariant, then there is some $c > 0$ and some $\alpha \in \mathbb{R}$ such that $g_n = cg_n^{\text{Fisher}}$ and $\nabla_n = \nabla_n^{(\alpha)}$

Norden-Sen Again

Recall that we have defined a notion of Norden-Sen dual pairs (∇, ∇^*) of affine connections with respect to a metric g as pairs that obey the equality

$$Z \langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z^* Y \rangle$$

for all $X, Y, Z \in \mathcal{T}(S)$. Theorem: fora ny statistical model S, the triple $(G_{\text{Fisher}}, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$ is a Norden-Sen geometry.

Back to Pythagoras

Let p, q, r be points in a statistical manifold S such that p and q are connected by a ∇-geodesic γ_1 and q and r are connected by a ∇^*-geodesic γ_2. If γ_1 and γ_2 are orthogonal with respect to the metric g, then we have the Pythagorean relation

$$D(p \| r) = D(p \| q) + D(q \| r)$$

Note that this holds even if D is not a symmetric distance!

A corollary relies on the notion of auto-parallel submanifolds. Let S be a manifold and M a submanifold. M is called auto-parallel with respect to a connection ∇ if $\nabla_X Y \in \mathcal{T}(M)$ for all $X, Y \in \mathcal{T}(M)$. Now let p be a point in S and M a ∇^*-auto-parallel submanifold of S. Then a necessary and sufficient condition for a point $q \in M$ to satisfy $D(p \| q) = \min_{r \in M} D(p \| r)$ is for the ∇-geodesic connection p and q to be orthogonal to M at q. The point q is called the ∇-projection of p onto M.