not for wide circulation

~/tech/tangent.tex, 1997.08.27

Geometries of Statistical Manifolds

Huaiyu Zhu
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
Email: zhuh@santafe.edu

August 30, 1997

Abstract

The geometries of information manifolds are developed in an invariant framework. The tangent spaces are constructed as Orlicz spaces and corresponds to the metric and affine connections. Convergences in these spaces correspond to convergences in information deviation.

Drafts 1996.06.17. 1997.05.10. 05.14. 07.05

Key words: Information geometry, infinite dimensional manifolds, tangent spaces, Orlicz spaces, nonparametric inference.

Contents

1 Introduction

1.1 Some fundamental statistical concepts

Statistical problems can be usefully viewed as problems concerning operations on probability measures which form various spaces. Such treatments are invariant to renaming of sample or parameter spaces [MS66]. Parametric problems can be viewed as problems concerning coordinates on these spaces. Models are subspaces, and robustness concerns the issue that the true distribution might lie outside this subspace we call a model. Asymptotic theories concern convergences and other local properties of these spaces. Regularity conditions are essentially smoothness conditions, one of the best known might be the Cramér-Wald regularity condition [LeC70, Condition A1].

Several concepts introduced by R. A. Fisher [Fis22, Fis25] are of paramount importance to statistics: consistency, efficiency, deficiency and sufficiency. Following Rao [Rao62], they may be called zeroth, first, second and the infinite orders of efficiency. We shall not consider it here as consistency is often trivial in practice.

Through the work of Fisher[Fis22], Rao[Rao45], Cencov [Čen82], Efron [Efr75], Amari [Ama82] and many others, it was eventually revealed that the first order efficiency concerns the geometric concept of metric ("distance") and the second order efficiency concerns in addition the concept of affine connection ("straightness"). There is a family of affine connections, ranging from e-connection (exponential) to m-connection (mixture). We label this family by $\delta \in [0, 1]$ where 0 corresponds to exponential and 1 corresponds to mixture. Under certain regularity conditions a model admit sufficient statistic if it is e-flat (the exponential families are), and a estimation method is most efficient if its "ancillary family" is m-flat (the maximum likelihood method is).

1.2 Local and global structures

These two concepts are local, hence only applicable to asymptotics. Asymptotic theories deal with the situation when the sample is such that the estimate lie close to the true distribution and certain expansions are available. The metric and affine connections corresponds to second and third order terms in these expansions. It was proved by Cencov that on finite sample spaces the Fisher metric and the δ-connections are, up to a multiplicative factor, all the possible such structures independent of structures in the sample space [Čen82]. There is evidence that this also applies to infinite sample spaces [Ama85] but the proof would presumably be more complicated than the already complicated proof on finite sample spaces.

A related approach to statistics rely on some global measures of "distance" between two probability measures. We call these information deviations following Cencov [Čen82]. They are generalizations of the concept of entropy, through the work of Jeffreys [Jef61], Kullback [Kul59], Renyi [Rén61], Csiszar [Csi67a], Cencov [Čen82], Amari [Ama85] and many others. One peculiarity of information deviation is that it is generally asymmetric, because the "distance"

one probability measure is from another depends on the point of view taken, as explained in [Čen82]. Some more intriguing counter-examples can be found in [Csi67b].

Amari [Ama82] showed that the metric and the family of affine connections naturally and uniquely define a family of information deviations as studied previously. Conversely Eguchi [Egu83] showed that the metric and affine connections are given by second and third order differentiations of the information deviation. Naturally all the information deviations are equivalent up to second order, as do the χ^2 distances. It was also shown that there is a beautiful dual-affine geometry underlying all these relations [Ama82, Ama85] which can also be generalized to abstract statistical manifolds [Lau87].

Remarkably, it turned out that statistical estimates based on minimization of information deviation in a Bayesian framework is not only second order but infinite order efficient, i.e. sufficient [ZR95a], if the estimates are allowed to be finite measures instead of restricted to probability measures. These estimates are called ideal estimates, and it turned out that the optimal estimates on any statistical model are given as projections of the ideal estimates onto the model, where the projections are defined through either the metric and connection or through the deviation. This may explain why there is virtually no study on asymptotic expansions beyond the order of three. The non-informative priors in Bayesian statistics turn out to correspond to δ-uniformity, and particular instances of the latter for $\delta \in \{0, 1/3, 1/2, 2/3, 1\}$ had been identified in non-Bayesian studies earlier [Kas84]. The 1-ideal estimate with 0-uniform prior happens to be the empirical distribution, and its 1-projection onto a model is nothing but the maximum likelihood estimate on that model. It is expected that most non-Bayesian theories in statistics corresponds to Bayesian theories with these uniform priors.

1.3 Tangent spaces and differentiation

To associate these abstract results with the very rich asymptotic theories, it is necessary to differentiate these objects and expand them in certain power series. In general differentiation requires a tangent space which is a linear space attached at the given point. For infinite sample spaces the space of all probability measures is infinite dimensional so are the tangent spaces. Although finite dimensional models would lead to finite dimensional tangent spaces, it is always useful to consider all such space as embedded in one and the same whole space.

Generally infinite dimensional linear spaces require additional topological structures to be of practical use. The most thoroughly studied and most useful such spaces are Banach spaces, in particular Hilbert spaces. Therefore it is interesting to see if some Banach spaces can serve as tangent spaces of statistical manifolds. It was noted in [ZR] that there are certain relations between the dual-affine geometry and the duality between Lebesgue spaces $L_{1/\delta}$. It turns out that the most natural tangent spaces are some Orlicz space which are equivalent to the Lebesgue spaces for $\delta \in (0, 1)$ but are different for $\delta \in \{0, 1\}$. Orlicz spaces are certain Banach spaces which generalize the Lebesgue spaces.

The use of Orlicz spaces is natural and unavoidable in view of the following facts. Function spaces have been well studied not only in isolation but more importantly as "scales" of spaces. All the function spaces of practical use are associated to each other by Sobolev-type embedding and interpolation relations [Tri83]. Such relations break down usually in two cases, either at integer order derivatives or for spaces associated with L_1 and L_∞. We shall not consider derivatives (with respect to sample space) here because it is only relevant when we also consider estimation of derivatives of density functions, which involves differentiable structures in the sample space. All our considerations here apply when the sample space is only measurable.

The other break down points can be removed by replacing L_1 and L_∞ with some Orlicz spaces. Remarkably, these spaces studied in pure functional analysis long ago happen to correspond to information deviations in statistics. It was shown by Csiszar that the topologies defined by δ-deviations are very different for $\delta \in \{0, 1\}$. This was generalized to the space of finite measures [Zhu96], and ratios between various deviations are given there. The bare minimum technical reason is associated with the following well-known anomaly

$$\int x^{n-1} \sim x^n, \qquad \text{but} \int x^{-1} \sim \log x \neq x^0. \tag{1.1}$$

It happens that those important concepts in statistics such as exponential families and log-likelihood are very closely associated with the logarithm appearing in the above formula. Considering the richness of asymptotic theory [LC86, IK81], the complication of Orlicz space appears to be a small price to pay.

1.4 Related approaches

Many attempts have been made in the past to construct a full non-parametric tangent space [Vaj89, Ama85]. Tangent spaces for finite sample spaces are studied in [Čen82]. Nonparametric studies of Fisher metric was undertaken in [KL76]. Various formulations of tangent spaces (mainly corresponding to information metric but sometimes equivalent to third order for $\delta \in \{0, 1\}$ can also be found in [Pfa82]. A Hilbert fiber bundle construction of tangent spaces for finite dimensional information manifolds was described in [Ama87]. Recently Orlicz spaces have been used to construct the e-geometry ($\delta = 0$) for spaces of probability measures equivalent to a given one [PS95]. Of course, most results in asymptotic theory depend implicitly on some unspecified tangent spaces. This is our main motivation for explicit construction of tangent spaces.

Our new formulation do not require any regularity conditions beyond that of a measurable space, such as dimensionality or smoothness. It is not restricted to the case where there exists a dominating finite measure, which is often not available in practice; e.g., the family of Gaussians is only dominated by the Lebesgue measure which is not finite. It also applies to finite measures as well as probability measures, which is useful because the ideal estimates are finite measures but not probability measures for any $\delta \in [0, 1)$. Our treatment includes all $\delta \in [0, 1]$, which is important considering the fact that the duality between δ and $1 - \delta$ plays an important

role in statistics. Most importantly, our Young function defining the Orlicz space completely corresponds to the information deviations. It is therefore expected that asymptotic expansion of the ideal estimation and error decomposition formula in these geometric structures will given all results of asymptotic theories which are invariant with respect to statistical isomorphisms.

2 Measures and some of their functions

Consider a measurable space $[Z, \mathcal{F}]$, where Z is a sample space and $\mathcal{F}$ is a σ-algebra of measurable sets. Denote by $\mathcal{M}(Z,\mathcal{F})$, $\mathcal{M}_+(Z,\mathcal{F})$, $\widetilde{\mathcal{P}}(Z,\mathcal{F})$, and $\mathcal{P}(Z,\mathcal{F})$ the space of charges (signed measures), measures, finite measures and probability measures on $(Z,\mathcal{F})$, respectively. Reference to $\mathcal{F}$ or $(Z,\mathcal{F})$ are omitted when there is no risk of confusion. Obviously,

$$\mathcal{M}_+ = \{p \in \mathcal{M} : \ p \geq 0\}, \tag{2.1}$$

$$\widetilde{\mathcal{P}} = \left\{p \in \mathcal{M}_+ : \ \int p < \infty\right\}, \tag{2.2}$$

$$\mathcal{P} = \left\{p \in \widetilde{\mathcal{P}} : \ \int p = 1\right\}. \tag{2.3}$$

A function $F : \ \mathbb{R}_+^n \to \mathbb{R}_+ \cup 0$ is called bounded and homogeneous (bh) if

$$\forall a_1, \ldots a_n, c \in \mathbb{R}_+ : \ F(ca_1, \ldots, ca_n) = cF(a_1, \ldots, a_n), \tag{2.4}$$

$$\exists C \in \mathbb{R}_+ : \ \forall a_1, \ldots, a_n \in \mathbb{R}_+ : \ F(a_1, \ldots, a_n) < C(a_1 + \cdots + a_n). \tag{2.5}$$

A bh function can be naturally extended to a bh function for finite measures, $F : \ \widetilde{\mathcal{P}}^n \to \widetilde{\mathcal{P}}$, by $\forall p_1, \ldots, p_n \in \widetilde{\mathcal{P}}$:

$$F(p_1, \ldots, p_n) := rF(p_1/r, \ldots, p_n/r), \qquad r \in \widetilde{\mathcal{P}} : r \ll p_1 + \ldots p_n \ll r, \tag{2.6}$$

eq:bhf

where p/r is the Radon-Nikodým derivative, which exists if and only if $p \ll r$, ie., p is dominated by r. The result $F(p_1, \ldots)$ is independent of the choice of r satisfying the condition in (2.6).

Unless otherwise indicated, it is assumed that $p, q, r \in \widetilde{\mathcal{P}}$, $\delta \in [0,1]$. In the following we shall often discuss $F(p,q)$ as if p and q are positive numbers. Such usage is justified by $F(p,q) = rF(f,g)$ where $r \equiv p+q$ and $f = p/r$, $g = q/r$ are density functions, when the discussion applies pointwise to $F(f,g)$.

We need fractional powers of measures as introduced in [Nev65, IV.1.4, p. 112–113]. Let $p \in \mathcal{M}_+$, $\delta \in (0,1]$. Define the Lebesgue spaces

$$L_{1/\delta}(p) := \left\{f : \ \int |f|^{1/\delta} p < \infty\right\}. \tag{2.7}$$

Define an equivalence relation among the couples $[f,p]$, where $p \in \mathcal{M}_+$, $f \in L_{1/\delta}(p)$, by $[f, gp] = [fg^\delta, p]$. Then the equivalence class of $[f,p]$ may be unambiguously denoted fp^δ. The space of δth power of finite measures

$$L_{1/\delta} := \left\{fp^\delta : \ p \in \widetilde{\mathcal{P}}, f \in L_{1/\delta}(p)\right\}, \tag{2.8}$$

is a Banach space with addition, multiplication and norm given by

$$fp^\delta + gq^\delta := \left(f\left(\frac{p}{r}\right)^\delta + g\left(\frac{q}{r}\right)^\delta\right) r^\delta, \tag{2.9}$$

$$afp^\delta = (af)p^\delta, \tag{2.10}$$

$$\|fp^\delta\|_{1/\delta} := \left(\int |f|^{1/\delta} p\right)^\delta. \tag{2.11}$$

The choice of r is irrelevant as long as it is equivalent to $p+q$. When $\delta = 1/2$ we have a Hilbert space L_2 with inner product

$$(fp^{1/2}, gq^{1/2}) := \int fg \left(\frac{p}{r}\right)^{1/2} \left(\frac{q}{r}\right)^{1/2} r. \tag{2.12}$$

Clearly for any $p \in \widetilde{\mathcal{P}}$, the space $L_{1/\delta}(p) \subset L_{1/\delta}$ isometrically. The mapping $p \to p^\delta$ maps $\widetilde{\mathcal{P}}$ onto $\widetilde{\mathcal{P}}^\delta \subset L_{1/\delta}$.

3 Information deviation

The concept of information deviation was generalized from the concept of entropy [Sha48, KL51, Rén61, Csi67a, Čen82, Ama82, Ama85, ZR95b]. It gives a "long range distance" between two distributions, but is generally non-symmetric.

Let $F: \mathbb{R}_+^2 \to \mathbb{R}_+$ be a homogeneous function $F(ac, bc) = cF(a, b)$. Then there is a canonical extension of F to measures $F: \mathcal{M}_+^2 \to \mathcal{M}_+$. It associates with a function $f: [-1, 1] \to \mathbb{R}_+$ by $f(u) := F(1 + u, 1 - u)$. The f-deviation for $p, q \in \mathcal{M}_+$ is defined as

$$D_f(p, q) := \int F(p, q) = \int r f(u), \tag{3.1}$$

$$r := (p + q)/2, \qquad u := (p - q)/(p + q). \tag{3.2}$$

A similar construct, $\int pf(q/p)$, was proposed in [Csi67a]. Our definition is slightly more general as it does not require $q \ll p$ or the special conventions such as $a \cdot f(a/b) = 0$ when a, b are 0 or ∞. Therefore all results concerning Csiszár's f-divergence can be directly applied to our f-deviation, with some trivial modification concerning the different appearances of f in the formulas. It was shown in [Čen82, p.???] to the effect that the measure rf^{-1} on $\mathbb{R}_+$

$$rf^{-1}(A) := \int_{f(u) \in A} r, \tag{3.3}$$

is minimum sufficient for specifying the deviation D_f. On the other hand, let $\widetilde{r} := ru^{-1}$, then

$$D_f(p, q) = \int_X r f(u) = \int_{[-1,1]} \widetilde{r} f. \tag{3.4}$$

Let $\delta \in [0, 1]$, $p, q \in \widetilde{\mathcal{P}}$. Then the δ-deviation is defined as

$$\ddot{(}p, q) := \begin{cases} \dfrac{\delta p + (1 - \delta)q - p^\delta q^{1-\delta}}{\delta(1 - \delta)} & \delta \in (0, 1), \\ \lim\limits_{\delta \to 0} \ddot{(}p, q) = p - q + q \log(q/p), & \delta = 0, \\ \lim\limits_{\delta \to 1} \ddot{(}p, q) = q - p + p \log(p/q), & \delta = 1. \end{cases} \tag{3.5}$$

$$D_\delta(p, q) := \int \ddot{(}p, q). \tag{3.6}$$

This definition of δ-deviation on $\widetilde{\mathcal{P}}$ was motivated by the need of δ-convexity [ZR95c] and inspired by [Ama85]. The definition of δ-deviation on $\mathcal{P}$ was given by [Čen82] from consideration of invariance, and by [Ama82, Ama85] from consideration of dual affine geometry. It is almost given in [LeC70]. [RV63] considers Hellinger distance between Gaussian processes. The explicit formulas for $\delta \in \{0, 1\}$ seems to first appeared in [ZR, ZR95c] but it is implicitly used in the proofs of positiveness of the entropies since [Sha48].

4 Orlicz spaces

4.1 General Orlicz spaces

Our treatment of Orlicz spaces follows [KJc77, Chap. 3]. A Young generating function is a non-decreasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, satisfying

$$\phi(0) = 0, \qquad \phi(\infty) = \infty. \tag{4.1}$$

The Young function

$$\Phi(s) := \int_s \phi(s) \tag{4.2}$$

is increasing and convex function and satisfies

$$\Phi(0) = 0, \qquad \Phi(\infty) = \infty. \tag{4.3}$$

Let $\psi = \phi^{-1}$ be a generalized inverse of ϕ,

$$\psi(t) = \sup\{s : \phi(s) \leq t\}, \tag{4.4}$$

then ψ is also a generating function and $\phi = \psi^{-1}$ in the same sense. The Young function

$$\Psi(t) := \int_t \psi(t) \tag{4.5}$$

is said to be complementary to Φ and they are said to form a Young complementary pair. For any $s, t \in \mathbb{R}_+$:

$$\Phi(s) + \Psi(t) \geq st. \tag{4.6}$$

For $p \in \widetilde{\mathcal{P}}$. The Orlicz class is defined as

$$\widetilde{L}_\Psi(p) := \left\{ u : \rho_{\Psi,p}(u) := \int p\Psi(|u|) < \infty \right\}. \tag{4.7}$$

The Orlicz space is a Banach space with Orlicz norm $\|u\|_\Phi$.

$$L_\Phi(p) := \left\{ u : \|u\|_{\Phi,p} := \sup_{\rho_{\Psi(u),p} \leq 1} \int p|uv| < \infty \right\}. \tag{4.8}$$

The subscript p is omitted when it is clear from context. The Orlicz norm is monotone

$$|u| \leq |v| \implies \|u\|_\Phi \leq \|v\|_\Phi. \tag{4.9}$$

From $\|u\|_\Phi \leq \rho_\Phi(u) + 1$, we see that $\widetilde{L}_\Phi(p) \subseteq L_\Phi(p)$.

The Young function Φ satisfies the Δ_2 condition, denoted $\Phi \in \Delta_2$, if either of the equivalent conditions holds

$$\exists k, T \in \mathbb{R}_+ : \forall s \geq T : \Phi(2s) \leq k\Phi(s), \tag{4.10}$$

$$\limsup_{t\to\infty} \frac{t\phi(t)}{\Phi(t)} < \infty. \tag{4.11}$$

The space $L_\Phi(p)$ is reflexive if and only if $\Phi, \Psi \in \Delta_2$, and $\widetilde{L}_\Phi(p) = L_\Phi(p)$.

4.2 Special Orlicz spaces

The following special family of Orlicz spaces will be of particular interest to information geometry. All the formulas apply to $\delta \in [0,1]$ and are continuous with respect to δ, although the formula for $\delta \in \{0,1\}$ is often formally different. We will not always mention limit of δ to the boundary of $[0,1]$. Define

$$\phi_\delta(s) := \frac{(1+\delta s)^{(1-\delta)/\delta} - 1}{1-\delta}, \tag{4.12}$$

$$\phi_0 = e^s - 1, \tag{4.13}$$

$$\phi_1 = \log(1+s). \tag{4.14}$$

The corresponding Young functions Φ_δ are

$$\Phi_\delta(s) := \frac{(1+\delta s)^{1/\delta} - s - 1}{1-\delta}, \tag{4.15}$$

$$\Phi_0 = e^s - s - 1, \tag{4.16}$$

$$\Phi_1 = (1+s)\log(1+s) - s. \tag{4.17}$$

For any $\delta \in [0,1]$, the following properties are easy to prove

- $\phi_\delta(s) \approx s$ and $\Phi_\delta(s) \approx s^2/2$ for $s \approx 0$.
- $\phi_{1-\delta}$ and ϕ_δ are inverse to each other; $\Phi_{1-\delta}$ and Φ_δ form a Young pair.
- For $\delta \in (0,1]$, $\Phi_\delta \in \Delta_2$, but $\Phi_0 \notin \Delta_2$. The spaces L_{Φ_δ} are reflexive if and only $\delta \in (0,1)$.
- For $\delta \in (0,1)$, the space L_{Φ_δ} is equivalent to the Lebesgue space $L_{1/\delta}$?

Therefore L_{Φ_0} and L_{Φ_1} can be regarded as substitutes of L_∞ and L_1 without the anomalies of the latter. For example, the Sobolev embedding theorem holds smoothly for the Orlicz space L_{Φ_0} but it does not hold for L_∞ [Ada75, p.242].

The e-geometry, i.e. 0-geometry was studied by [PS95] using $\Phi(s) = \cosh(s) - 1$. From

$$1 \le \frac{e^s - s - 1}{\cosh s - 1} < 2, \tag{4.18}$$

it is clear that their space $L^\Phi(p)$ is equivalent to a subspace of our $L_{\Phi_0}(p)$. It is a subspace because not all elements of the latter are probability measures. The convergence in these two spaces are equivalent up to second order, as can be seen by examining $\partial^k \Phi(0)$ for $k = 0,1,2,3$. It should be pointed out, however, that due to some technical complications the "e-convergence" defined there is in fact not equivalent to either of these convergences. We do not think such complications are relevant to statistics. Our choice of the Young function has the advantage of corresponding exactly to information deviations, besides being identical to that used in the study of function spaces.

5 Tangent spaces

sec:tang

5.1 δ-coordinates

Let $\delta \in [0,1]$. For given $r \in \widetilde{\mathcal{P}}$ define the δ, r-coordinate or δ, r-representation of any $p \in \widetilde{\mathcal{P}}$ as

$$l_{\delta,r}(p) := \frac{(p/r)^\delta - 1}{\delta}, \tag{5.1}$$

$$l_{0,r}(p) = \log(p/r), \tag{5.2}$$

$$l_{1,r}(p) = q/r - 1. \tag{5.3}$$

Then for all $\delta \in (0,1]$,

$$\int r(\delta l_{\delta,r}(p))^{1/\delta} = \int (p^\delta - r^\delta)^{1/\delta} \leq \left((\int p)^\delta + (\int r)^\delta\right)^{1/\delta} < \infty, \tag{5.4}$$

$$r\Phi_\delta(l_{\delta,r}(p)) = F_\delta(p,r), \tag{5.5}$$

$$\int r\Phi_\delta(l_{\delta,r}(p)) = D_\delta(p,r) < \infty, \tag{5.6}$$

$$D_\delta(p,r) + D_\delta(r,q) - \int r l_{\delta,r}(p) l_{1-\delta,r}(q) = D_\delta(p,q). \tag{5.7}$$

Therefore $l_{\delta,r}(p) \in L_{1/\delta}(r)$, $l_{\delta,r}(p) \subseteq \widetilde{L_{\Phi_\delta}}(r)$.

Remark 5.1 This means that it is more concenient to use q as an anchor in $D(p,q)$, as follows the convention of Csiszar.

To examine convergence in $\widetilde{L_{\Phi_\delta}}(r)$, it is convenient to define what can intuitively described as the "difference between p and q as reflected in r"

$$r_\delta(p,q) := \left(r^\delta + |p^\delta - q^\delta|\right)^{1/\delta}, \tag{5.8}$$

$$r_0(p,q) = r \exp|\log(p/q)|. \tag{5.9}$$

Then the "distance between p and q from point of view of r" is

$$\int r\Phi_\delta(|l_{\delta,r}(p) - l_{\delta,r}(q)|) = D_\delta(r_\delta(p,q), r). \tag{5.10}$$

- The spaces $L_{\Phi_\delta}(r)$ and $L_{1/\delta}(r)$ are linearly isomorphic and topologically equivalent.
- The convergence does not depend on r for $\delta \in (0,1)$.
- The topology thus defined is weaker than $D_\delta(p,q)$ for $\delta \in \{0,1\}$.

In analogy to the embedding of $L_{1/\delta}(r)$ to $L_{1/\delta}$, we can regard all the $L_{\Phi_\delta}(r)$ as embedded in a large Banach space L_{Φ_δ} which is then independent of any dominating measure,

$$L_{\Phi_\delta} := \left\{(1+\delta s)r^\delta : r \in \widetilde{\mathcal{P}}, s \in L_{\Phi_\delta}(r)\right\}, \tag{5.11}$$

$$L_{\Phi_0} := \left\{s + \log r : r \in \widetilde{\mathcal{P}}, s \in L_{\Phi_0}(r)\right\}. \tag{5.12}$$

Note that L_{Φ_0} is an affine space instead of a linear space. When only affine properties are considered, $T^\delta\mathcal{M} = L_{\Phi_\delta}$ is a tangent bundle of $\widetilde{\mathcal{P}}$ and $T^\delta_r\mathcal{M} = L_{\Phi_\delta}(r)$ is the tangent space at r.

For $\delta \in (0,1)$ all the topologies are the same. So the tangent bundle is in the ordinary sense, although being infinite dimensional. The tangent space is naturally embedded in the Banach space $L_{1/\delta}$,

For $\delta = 1$ the tangent space is spanned by likelihood functions and for $\delta = 0$ it is spanned by log-likelihood functions. However, the convergence in both $L_{\Phi_1}(r)$ and $L_{\Phi_0}(r)$ depends on r. The distance between p, q is given explicitly as

$$D_0(r_0(p,q),r) = \int r(e^s - s - 1), \qquad s = |\log(p/q)|, \tag{5.13}$$

$$D_1(r_1(p,q),r) = \int r((1+s)\log(1+s) - s), \qquad s = |p-q|/r. \tag{5.14}$$

For $s \approx 0$, we have $\Phi_\delta(s) \approx s^2/2$ regardless of δ, and

$$\int r\Phi_\delta(|l_{\delta,r}(p) - l_{\delta,r}(q)|) \approx \int \frac{r^{1-2\delta}}{2}\left|\frac{p^\delta - q^\delta}{\delta}\right|^2, \tag{5.15}$$

which corresponds to the Hilbert space given in [Ama87]

$$T^\delta_p\widetilde{\mathcal{P}} := \left\{ur^\delta : \int r(p/r)^{1-2\delta}u^2 < \infty\right\} = \left\{u \in L_{1/\delta} : \int p^{1-2\delta}u^{2\delta} < \infty\right\}. \tag{5.16}$$

For $u < C < 1$ only expansion at 0 matters. (?)

5.2 Alternative δ-coordinates

An equivalent way to construct tangent space was suggested in [Ama85, §3.1, p.66], using $L_{1/\delta}(r)$ as a tangent space with δ-coordinate $l_\delta(p) \in L_{1/\delta}$

$$l_\delta(p) := \frac{1}{\delta}p^\delta, \qquad \delta \neq 0, \tag{5.17}$$

$$l_0(p) := \log p. \tag{5.18}$$

If we define

$$\Phi_\delta(s) := \frac{(\delta s)^{1/\delta}}{1-\delta}, \tag{5.19}$$

then

$$\Phi_\delta(l_\delta(p)) = \frac{p}{1-\delta}, \tag{5.20}$$

$$\Phi_\delta(l_\delta(p)) + \Phi_{1-\delta}(l_{1-\delta}(q)) - l_\delta(p)l_{1-\delta}(q) = F_\delta(p,q). \tag{5.21}$$

Let $\mathcal{Q}$ denote a finite dimensional family dominated by $r \in \mathcal{P}$. Suppose p is parameterized by some coordinates. Using the tensor index notation generalized to Banach manifold [LL56], we may take the Fréchet derivatives

$$\partial_i l_\delta(p) = p^\delta \partial_i l_0(p). \tag{5.22}$$

$$\partial_i \partial_j l_\delta(p) = p^\delta \partial_i \partial_j l_0(p) + \delta p^\delta \partial_i l_0(p) \partial_j l_0(p). \tag{5.23}$$

The derivatives of log-likelihood are also known as Fisher's score functions. We can compute the metric tensor and the Christoffel symbol for the coordinate $l_\delta(p) \in L_{1/\delta}$. This computation requires Fréchet differentiation, but the index notation of tensor analysis is still usable [LL56], since $L_{1/\delta}$ is reflexive. For the Hilbert space case $\delta = 1/2$ a more intuitive interpretation of the index notation is possible [Iya80]. [Kal63] uses Volterra differentiation which is less restricted than Fréchet differentiation.

The infinite dimensional logarithmic differentiation of measures in [DM85] may be irrelevant as it appears to be differentiation with respect to the sample space. Other considerations include [Pfa82].

A non-parametric Fisher information metric was constructed in [KL76] which of course is equivalent to the metric defined by all the above geometries. In particular, it can be easily obtained from the Hellinger distance, which is the distance in L_2.

5.3 Affine connections

In general, for any $\delta \in [0, 1]$, the affine structure of L_{Φ_δ} naturally define the δ-affine connection. It can be seen that it coincides with Amari's α-connection with $\alpha = 1 - 2\delta$ when the latter is defined. It is interesting to give the explicit formula for the Riemannian metric and Christoffel symbols of δ-connection as expressed in ϵ-coordinates, $\epsilon \in [0, 1]$, for any finite dimensional model parameterized by θ^i, using

$$\partial_i l_\epsilon = \partial_i l, \qquad \partial_i \partial_j l_\epsilon = \partial_i \partial_j l + \epsilon \partial_i l \partial_j l, \tag{5.24}$$

which does not depend on r, we have

$$g_{ij} = \int p \partial_i l_\epsilon \partial_j l_\epsilon, \tag{5.25}$$

$$\overset{\delta}{\Gamma}_{ijk} = \int p \partial_i \partial_j l_\epsilon \partial_k l_\epsilon + (\delta - \epsilon) \int p \partial_i l_\epsilon \partial_j l_\epsilon \partial_j l_\epsilon. \tag{5.26}$$

From this it is obvious that δ-affine structure is exactly the natural affine structure of the δ-coordinates in $L_{1/\delta}$.

6 Convergence issues

Given a signature function f, for $\varepsilon \in \mathbb{R}_+$, the (f, ε)-ball centered at $r \in \widetilde{\mathcal{P}}$ is defined as

$$B_{f,\varepsilon}(r) := \left\{ p \in \widetilde{\mathcal{P}} : \ D_f(p, r) < \varepsilon \right\}. \tag{6.1}$$

The collection of these balls for all centers and radii,

$$\mathcal{B}_f := \left\{ B_{f,\varepsilon}(r) : \ r \in \widetilde{\mathcal{P}}, \varepsilon \in \mathbb{R}_+ \right\}, \tag{6.2}$$

forms a neighborhood subbase. The collection of finite intersections of these balls forms a neighborhood base which defines a topology called the f-topology. This topology is Hausdorff (T2) if $\mathcal{B}_f$ itself is a neighborhood base. It is however always a T1 space.

It was shown [Csi67b] that f-deviation defines a metrizable topology on $\mathcal{P}$ if f is convex, $f'(0) = 0$, $f''(0) > 0$ and $f(\pm 1) < \infty$. This is true for D_δ if $\delta \in (0, 1)$.

For all $\delta \in (0, 1)$, it was effectively proved in [Csi67b] and generalized in [Zhu96] that the topology on $\widetilde{\mathcal{P}}^\delta$ defined by D_δ is equivalent to that of $L_{1/\delta}$. Therefore $\widetilde{\mathcal{P}}$ is a Banach promanifold $L_{1/\delta}$. Using the techniques in [Zhu96] it is easy to show that the ratio of convergence diverges. Since these topologies are equivalent to that of L_2 they are also Hilbert promanifold. The concept of infinite dimensional manifolds are studied in [CBWMDB77]. See [AMR83] for an introduction.

Note that each δ defines a different linear structure. The spaces $\widetilde{\mathcal{P}}^\delta$ are not strictly manifolds because they are positive cones in a Banach space and it is known that in infinite dimensional spaces cones are generally not open subsets. We call this space a promanifold. All of its finite dimensional submanifolds are ordinary manifolds.

Since this topology corresponds to all the convergence concepts in asymptotic theories, we shall take it as default unless stated otherwise. Given the topology on $\widetilde{\mathcal{P}}$, the σ-algebra of Bair sets and Borel sets are well defined [Nev65, II.7]. The Borel measures on $\widetilde{\mathcal{P}}$ are well defined [Kol56, Hal50], which can be used to define the measurable space on which the prior and posterior are defined. For infinite sample space, the space $\widetilde{\mathcal{P}}$ is infinite dimensional. The prior and posterior as measures might be too coarse for applications and the conditional distribution space defined by [Rén56] may be useful.

It was discovered by Csiszár [Csi67b] that $\mathcal{B}_f$ may not be a neighborhood base. For $\delta \in \{0, 1\}$ the topologies are non-Hausdorff but still Fréchet (T1), [Csi67b, Zhu96]. This topology is stronger than the topology defined by D_δ, $\delta \in (0, 1)$. It was shown [Zhu96] that the same is true on $\widetilde{\mathcal{P}}$. Note that the notion of topology used by Csiszár followed [Sie56] which is somewhat different from current usage [Kel55].

The proper condition for a sequence p_n to converge to p under such D_f is that $\forall q \in \widetilde{\mathcal{P}}$:

$$D_f(p, q) < \infty \implies \exists m : \forall n > m : D_f(p_n, q) \leq D_f(p, q). \tag{6.3}$$

One consequence is that it is not feasible to define $p_n \to p$ simply by $D_f(p_n, p) \to 0$, which may not even be able to exclude $D_f(p_n, p_{n+1}) = \infty$. If the topology defined by D_f is stronger than that defined by D_g, we shall say that D_f is stronger than D_g, or simply f is stronger than g.

Such topologies are far too strong and not very convenient for applications. It effectively means convergence from the point of view of all r. Therefore the convergence in each tangent space at r may be more useful in practice, which however depends on r.

7 Asymptotics

It was proved in [Zhu96] that the δ-deviations are approximated to the second order by the "symmetric χ^2 deviation" and to the third order by the δ-χ^2 deviations

$$D_\delta(p,p) \approx \chi^2_{1/2}(p,q) := \int \frac{(p-q)^2}{(p+q)} = \int r 2x^2, \tag{7.1}$$

$$\chi^2_\delta(p,q) := \int (p-q)^2/2r_\delta, \qquad r_\delta = (p+q)/3 + ((1-\delta)p + \delta q)/3. \tag{7.2}$$

This implies that the asymmetric χ^2-deviations are more skewed than D_0 and D_1.

$$\chi^2_{-1}(p,q) = \int (p-q)^2/p, \qquad \chi^2_2(p,q) = \int (p-q)^2/q. \tag{7.3}$$

The following bounds were also proved in [Zhu96]

$$\min\left\{\frac{\delta_2}{\delta_1}, \frac{1-\delta_2}{1-\delta_1}\right\} \le \frac{D_{\delta_1}(p,q)}{D_{\delta_2}(p,q)} \le \max\left\{\frac{\delta_2}{\delta_1}, \frac{1-\delta_2}{1-\delta_1}\right\}. \tag{7.4}$$

$$1 \le \frac{D_\delta(p,q)}{\chi^2_\delta(p,q)} \le \frac{2}{3}\left(1 + \max\left\{\frac{1}{\delta}, \frac{1}{1-\delta}\right\}\right). \tag{7.5}$$

By differentiation we have

$$g_{ij} = -\partial_i \partial'_j D_f(p,q)|_{p=q} = \frac{f''(0)}{4}\int \frac{\partial_i p \partial_j p}{p} = \frac{f''(0)}{4}\int p \partial_i l \partial_j l, \tag{7.6}$$

$$\Gamma_{ijk} = -\partial_i \partial_j \partial'_k D_f(p,q)|_{p=q} \tag{7.7}$$

$$= \frac{f'''(0)}{8}\int \frac{\partial_i p \partial_j p \partial_k p}{p^2} + \frac{f''(0)}{4}\int \left(\partial_i \partial_j p - \frac{\partial_i p \partial_j p}{2p}\right)\frac{\partial_k p}{p} \tag{7.8}$$

$$= \frac{f'''(0)}{8}\int p \partial_i l \partial_j l \partial_k l + \frac{f''(0)}{4}\int p\left(\partial_i \partial_j l - \frac{1}{2}\partial_i l \partial_j l\right)\partial_k l. \tag{7.9}$$

Without loss of generality, we can set $f''(0) = 4$. Then

$$g_{ij} = \int p \partial_i l \partial_j l, \tag{7.10}$$

$$\Gamma_{ijk} = \frac{f'''(0)}{2f''(0)}\int p \partial_i l \partial_j l \partial_k l + \int p\left(\partial_i \partial_j l - \frac{1}{2}\partial_i l \partial_j l\right)\partial_k l. \tag{7.11}$$

For $f(x) = \ddot{(}1+x, 1-x)$, it is easy to show that

$$f''(0) = 4, f'''(0) = 8(\delta - 1/2), \tag{7.12}$$

so the geometry is exactly the δ-geometry.

$$g_{ij} = \int p \partial_i l \partial_j l, \tag{7.13}$$

$$\Gamma_{ijk} = (\delta - 1/2)T_{ijk} + \overset{1/2}{\Gamma}_{ijk}. \tag{7.14}$$

References

[ABNK+87] S. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R Rao, editors. *Differential Geometry in Statistical Inference*, volume 10 of *IMS Lecture Notes Monograph*. Inst. Math. Statist., Hayward, CA, 1987.

[Ada75] R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[Ama82] S. Amari. Differential geometry of curved exponential families—curvature and information loss. *Ann. Statist.*, 10(2):357–385, 1982.

[Ama85] S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Springer Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.

[Ama87] S. Amari. Differential geometrical theory of statistics. In Amari et al. [ABNK+87], chapter 2, pages 19–94.

[AMR83] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*. Addison-Wesley, London, 1983.

[CBWMDB77] Y. Choquet-Bruhat, C. De Witt-Morette, and M. Dillard-Bleick. *Analysis, Manifolds and Physics*. North Holland, Amsterdam, 1977.

[Čen82] N. N. Čencov. *Optimal Decision Rules and Optimal Inference*. Amer. Math. Soc., Rhode Island, 1982. Translation from Russian, 1972, Nauka, Moscow.

[Csi67a] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[Csi67b] I. Csiszár. On topological properties of f-divergences. *Studia Sci. Math. Hungar.*, 2:329–339, 1967.

[DM85] Yu. L. Daletskiĭ and B. D. Maryanin. Smooth measures on infinite-dimensional manifolds. *Dokl. Akad. Nauk SSSR*, 285(6):1297–1300, 1985.

[Efr75] B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *Ann. Statist.*, 3:1189–1242, 1975.

[Egu83] S. Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.*, 11:793–803, 1983.

[Fis22] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc., A*, 222:309–368, 1922.

[Fis25] R. A. Fisher. Theory of statistical estimation. *Proc. Camb. Phi. Soc.*, 22:700–725, 1925.

[Hal50] P. R. Halmos. *Measure Theory*. Van Nostrand, New York, 1950.

[IK81] I. A. Ibragimov and R. Z. Khasminskii. *Statistical Estimation : Asymptotic Theory.* Springer-Verlag, New York, 1981.

[Iya80] M. Iyanaga. A differential geometry on the Hilbert manifold. *Mathematical Reports, College of General Education, Kyushu Univ.*, 12(2):29–45, 1980.

[Jef61] H. Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, 1961. (First ed. 1939).

[Kal63] G. Kallianpur. Von Mises functionals and maximum likelihood estimation. In Rao [Rao63], pages 137–146.

[Kas84] R. E. Kass. Canonical parameterization and zero parameter effects curvature. *J. R. Statist. Soc., B*, 46:86–92, 1984.

[Kel55] J. L. Kelley. *General Topology.* University Series in Higher Mathematics. Van Nostrand, New York, 1955.

[KJc77] A. Kufner, O. John, and S. Fučik. *Function Spaces.* Neordhoff, Leyden, 1977.

[KL51] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[KL76] Yu. A. Koshevnik and B. Ya. Levit. On a non-parametric analogue of the information matrix. *Th. Prob. Appl.*, 21:738–753, 1976.

[Kol56] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea., New York, 1956. Translation of *Grundbegriffe der Wahrscheinlichkeitsrechnung, 1933.*

[Kul59] S. Kullback. *Information Theory and Statistics.* J. Wiley, New York, 1959.

[Lau87] S. L. Lauritzen. Statistical manifolds. In Amari et al. [ABNK$^+$87], chapter 4, pages 163–216.

[LC86] L. M. Le Cam. *Asymptotic Methods in Statistical Theory.* Springer-Verlag, New York, 1986.

[LeC70] L. LeCam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, 41(3):802–828, 1970.

[LL56] D. Laugwitz and E. R. Lorch. Riemann metrics associated with convex bodies and normed spaces. *Amer. J. Math.*, 78:889–894, 1956.

[MS66] N. Morse and R. Sacksteder. Statistical isomorphism. *Ann. Math. Statist.*, 37:203–214, 1966.

[Nev65] J. Neveu. *Mathematical Foundations of the Calculus of Probability.* Holden-Day, San Francisco, 1965. Translated from French, 1964, Masson.

[Pfa82] J. Pfanzagl. *Contributions to a General Asymptotic Statistical Theory*, volume 13 of *Lect. Note Statistics.* Springer-Verlag, New York, 1982.

[PS95] G. Pistone and C. Sempi. An infinite-dimensional geometric strucutre on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, 1995.

[Rao45] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

[Rao62] C. R. Rao. Efficient estimates and optimum inference procedures in large samples (with discussion). *J. R. Statist. Soc., B*, 24:46–72, 1962.

[Rao63] C. R. Rao, editor. *Contributions to Statistics.* Pergamon, 1963.

[Rén56] A. Rényi. On conditional probability spaces generated by a dimensionally ordered set of measures. *Th. Prob. Appl.*, 1(1):55–64, 1956.

[Rén61] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. on Math. Statist. Prob.*, volume 1, pages 547–561. Univ. California, 1961.

[RV63] C. R. Rao and V. S. Varadarajan. Discrimination of gaussian processes. In Rao [Rao63], pages 363–390.

[Sha48] C. E. Shannon. A mathematical theory of communication, I & II. *Bell Syst. Tech. J.*, 27:379–423, 623–656, 1948.

[Sie56] W. Sierpińsky. *General Topology.* Mathematical Expositions, 7. U. Toronto, Toronto, 2 edition, 1956. translated from Polish by C. Krieger, first ed 1952.

[Tri83] H. Triebel. *Theory of Function Spaces.* Monographs in Mathematics, 78. Birkhäuser Verlag, Basel, 1983.

[Vaj89] Igor Vajda. *Theory of Statistical Inference and Information.* Kluwer Academic, Dordrecht, 1989. Translation of: Teoria informacie a statistickeho rozhodovania.

[Zhu96] H. Zhu. On topologies and geometries of information deviations of finite measures. Submitted, 1996.

[ZR] H. Zhu and R. Rohwer. Measurements of generalisation based on information geometry. Presented at Math. of Neural Networks and Appl. Conf. (MANNA), Oxford, July 1995. To appear in *Ann. Math. Artif. Intell.*

[ZR95a] H. Zhu and R. Rohwer. Bayesian geometric theory of statistical inference. Submitted, 1995.

[ZR95b] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation. *Neural Proc. Lett.*, 2(6):28–31, 1995.

[ZR95c] H. Zhu and R. Rohwer. Information geometric measurements of generalisation. Technical Report NCRG/4350, Aston University, 1995. `ftp://cs.aston.ac.uk/neural/zhuh/generalisation.ps.Z`.