A TASTE OF INFORMATION GEOMETRY

KEVIN O'CONNOR

1. Introduction

Information geometry seeks to characterize the structure of statistical models from a differential-geometrical point of view. By considering families of probability distributions as manifolds with coordinate charts determined by the parameters of each individual model, the tools of differential geometry such as divergences and metric tensors naturally provide additional means of studying their characteristics. As will be described later, this perspective offers powerful insights into areas such as mean field theory and machine learning.

In the first part of this paper, we offer some motivation for the study of information geometry as well as a brief history of the field. Next we introduce statistical manifolds which is the primary object of study in information geometry. Moving on, we will define divergences and metric tensors and illustrate how they interact with statistical manifolds. Finally, we will highlight two particular areas of application for the tools of information geometry.

1.1. Motivating Example.

The motivation for information geometry primarily comes from a desire to analyze statistical models, or families of probability distributions, in a manner which is independent of the choice of coordinates. A particularly clear example of a situation when the representation affects the results of a statistical analysis is provided by Grosse [5]. He demonstrates how regularized polynomial regression depends on the chosen representation of the data. It is easy to see why the dependence of the results of a statistical procedure on the choices of the individual performing the analysis is unsatisfactory. Since information geometry provides the tools for characterizing the properties of a statistical space which are independent of the choice of coordinates, there is a natural motivation for viewing problems in statistics and probability through the lens of information geometry.

1.2. A Brief History.

Information geometry traces its roots back to the early work of C. R. Rao [1]. In his paper, *Information and the Accuracy Attainable in the Estimation of Statistical Parameters* (1945), Rao put forth a new way of measuring the statistical distance between two populations via a Riemannian metric which was shown to be equivalent to Fisher's information matrix [8]. This signaled the beginning of information geometry as it is known today. Further contributions were made by Jeffreys, Chentsov, Efron, Barndorff-Nielsen and Amari in the decades that followed [1]. Today, information geometry is an active area of research with international conferences held regularly. However, despite over a half century of development, it is still considered by many to be a relatively young field with real potential for future contributions to statistics and probability theory in general.

2. Statistical Manifolds

A central idea in Information Geometry is that of a *statistical manifold*. But first we should state formally what a manifold is.

Definition 2.1. [7] *An n-dimensional manifold is a Hausdorff, second-countable space such that every point has an open neighborhood which is homeomorphic to* $\mathbb{R}^n$.

Intuitively, this is a space which locally resembles $\mathbb{R}^n$. Identifying points in such a space requires a something called a *coordinate chart*. It is defined as follows,

Definition 2.2. [7] *A coordinate chart for an n-manifold,* M*, is a pair* (U, ϕ) *such that* $U \subset M$ *is open and* $\phi(U) \subset \mathbb{R}^n$.

Thus a coordinate chart gives us a means by which to label points in the space. Note that the coordinate chart is not unique. Often times it is convenient to transform the coordinate chart. In the case of a statistical manifold, this will amount to reparametrizing the distribution [2]. Additionally, a single coordinate chart usually does not provide coordinates for the entire manifold, unless the whole space is topologically isomorphic to Euclidean space.

The definition of a statistical manifold will also rely on the idea of a Riemannian manifold, so we define it next.

Definition 2.3. [7] *Let* M *be a manifold with an inner product,* g_p*, on the tangent space of every point* $p \in M$*. Then* M *is Riemannian if* $p \mapsto g_p(X(p), Y(p))$ *is smooth for any differentiable vector fields,* X *and* Y*. Furthermore, we say* g *is a Riemannian metric.*

Now we can define a statistical manifold.

Definition 2.4. [1] *A Riemannian manifold* M *is a statistical manifold, with coordinate chart* ϕ*, if* $M = \{P(\cdot;\theta)\}_{\theta\in\Theta}$*, for some family of distributions, and* $\phi(P(\cdot;\theta)) = \theta$.

In other words, we treat a family of distributions with parameters, $\theta \in \Theta$, as a manifold where each point on the manifold is a probability distribution. Furthermore, as the coordinate chart is defined, each point in the manifold is identified by its parameters.

Example 2.5. *The family of all normal distributions, that is distributions with density*

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

and parameter space,

$$\Theta = \{(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$$

is a 2-dimensional statistical manifold.

3. Divergences and Riemannian Metrics

Having established the central structure of a statistical manifold, we begin to consider a set of tools which allow us to act on this manifold. First we define a *divergence*.

Definition 3.1. [2] *Let* M *be a manifold. Then a function,* $D : M \times M \to \mathbb{R}$*, is a divergence on* M *if all three of the following conditions hold. Let* $P, Q \in M$.

1. $D(P,Q) \geqslant 0$.
2. $D(P,Q) = 0$ *iff* $P = Q$ *a.e.*
3. *For* P *close to* Q*,* $D(\xi_P, \xi_P + d\xi)$*, where* ξ_P *and* $\xi_P + d\xi$ *are coordinates of* P *and* Q *respectively, can be written as a Taylor expansion in terms of a positive definite matrix,* $G = (g_{ij})$*, as*

$$D(\xi_P, \xi_P + d\xi) = \frac{1}{2}\sum g_{ij}(\xi_P) d\xi_i d\xi_j + \mathcal{O}(|d\xi|^3)$$

Thus the divergence gives us a means of determining the degree of separation between two points on a manifold. Importantly, a divergence is not a metric since it is not necessarily symmetric and need not satisfy the triangle inequality. An important divergence in information geometry is the *Kullback-Leibler (KL) divergence*, or relative entropy.

Definition 3.2. [2] *The KL-divergence between two points $P, Q \in M$ with densities p and q respectively, is defined as*

$$D_{KL}(P,Q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

The KL-divergence gives us an idea of how much two distributions differ. More precisely, if X is a random variable and P and Q are two distributions for X, $D_{KL}(P,Q)$ gives the amount of information gained about X by using P instead of Q as its distribution [9]. As an example, consider the KL-divergence between two normal random variables.

Example 3.3. *Let $M = \{P(x;\mu,\sigma^2)\}$ be the statistical manifold of all univariate normal distributions and take $P, Q \in M$ with coordinates, $\phi(P) = (\mu, \sigma^2)$ and $\phi(Q) = (\mu', \sigma'^2)$. Then the KL-divergence is calculated as follows.*

$$\begin{aligned}
D_{KL}(P,Q) &= \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx \\
&= \int_{\mathbb{R}} p(x) \left(\frac{1}{2}\log(2\pi\sigma'^2) + \frac{(x-\mu')^2}{2\sigma'^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \right) dx \\
&= \log \tfrac{\sigma'}{\sigma} \int_{\mathbb{R}} p(x)dx + \frac{1}{2\sigma'^2} \int_{\mathbb{R}} (x-\mu')^2 p(x)dx - \frac{1}{2\sigma^2} \int_{\mathbb{R}} (x-\mu)^2 p(x)dx \\
&= \log \tfrac{\sigma'}{\sigma} + \tfrac{1}{2\sigma'^2} \int_{\mathbb{R}} (x^2 - 2x\mu' + \mu'^2) p(x)dx - \frac{1}{2\sigma^2}\sigma^2 \\
&= \log \tfrac{\sigma'}{\sigma} + \tfrac{1}{2\sigma'^2} \left(\sigma^2 + \mu^2 - 2\mu'\mu + \mu'^2\right) - \tfrac{1}{2} \\
&= \log \tfrac{\sigma'}{\sigma} + \tfrac{\sigma^2 + (\mu-\mu')^2}{2\sigma'^2} - \tfrac{1}{2}
\end{aligned}$$

heta) With another definition, we can show an important relationship between KL-divergence and the Fisher information.

Definition 3.4. [2] *Let $P \in M$ with density p, and suppose we have coordinates (parameters), $\theta = (\theta_1, ..., \theta_k)$. Then the Fisher information is defined by*

$$I_{jk}(\theta) = \int_X \frac{\partial \log p(x,\theta)}{\partial \theta_j} \frac{\partial \log p(x,\theta)}{\partial \theta_k} p(x,\theta) dx$$

In the case where θ is 1-dimensional, we can write this as

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right)^2 \right]$$

The Fisher information is a Riemannian metric on a statistical manifold and measures the curvature of the manifold at any given point [2]. It is referred to as *information* because high curvature reflects lower variability and therefore better knowledge about a parameter. As an example, we compute the Fisher information for the normal distribution below.

Example 3.5. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. The log-likelihood is given by*

$$\log f(x|\theta) = -\frac{(x-\mu)^2}{2\sigma^2} + C$$

where C is a constant. So

$$\frac{\partial^2}{\partial \mu^2} \log f(x|\theta) = -\frac{1}{\sigma^2}$$

and since this is constant,

$$I(\mu) = -\mathbb{E}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}$$

Note that the Fisher information of μ increases as σ^2 decreases. Intuitively, this makes sense because a distribution with smaller variance would give you a better estimate of μ. Whereas, as $\sigma^2 \to \infty$, the influence that μ has on the data effectively goes away and it becomes impossible to infer anything about μ from the data.

An important result shows a relationship between the KL-divergence and the Fisher information metric. Specifically, for sufficiently close points on a manifold, the KL-divergence between these points approximates the Fisher information between them. This is stated formally in the following theorem.

Theorem 3.6. [6] *Let $P_\theta, P_{\theta_0} \in M$ with coordinates θ and $\theta_0 = \theta + d\theta$ respectively. Then*

$$I_{jk}(\theta_0) = \frac{\partial^2}{\partial \theta_j \partial \theta_k} D_{KL}(P_\theta, P_{\theta_0})\Big|_{\theta_0 = \theta}$$

Proof. We will prove this in the 1-dimensional case following a proof given by Kullback [6]. Let $P_\theta, P_{\theta_0} \in M$ with densities, $p(x;\theta)$ and $p(x;\theta_0)$ respectively such that $\theta_0 = \theta + \delta\theta$. Writing the KL-divergence,

$$\begin{aligned} D_{KL}(P_\theta, P_{\theta_0}) &= \int p(x;\theta) \log \frac{p(x;\theta)}{p(x;\theta_0)} dx \\ &= \int p(x;\theta) \log \frac{p(x;\theta)}{p(x;\theta+\delta\theta)} dx \\ &= \int p(x;\theta) \left(\log p(x;\theta) - \log p(x;\theta+\delta\theta)\right) dx \end{aligned}$$

Then using a second-order Taylor expansion for $\log p(x;\theta+\delta\theta)$,

$$\begin{aligned}
D_{KL}(P_\theta, P_{\theta_0}) &= \int p(x;\theta)\left(\log p(x;\theta) - \log p(x;\theta) - \delta\theta\frac{d\log p(x;\theta)}{d\theta} - \frac{\delta\theta^2}{2}\frac{d^2\log p(x;\theta)}{d\theta^2}\right)dx \\
&= \int p(x;\theta)\left(-\delta\theta\frac{d\log p(x;\theta)}{d\theta} - \frac{\delta\theta^2}{2}\frac{d^2\log p(x;\theta)}{d\theta^2}\right)dx \\
&= \int p(x;\theta)\left(-\delta\theta\frac{1}{p(x;\theta)}\frac{dp(x;\theta)}{d\theta} - \frac{\delta\theta^2}{2}\left(\frac{1}{p(x;\theta)}\frac{d^2p(x;\theta)}{d\theta^2}\right.\right. \\
&\qquad \left.\left. - \tfrac{1}{p(x;\theta)^2}\left(\tfrac{dp(x;\theta)}{d\theta}\right)^2\right)\right)dx \\
&= -\delta\theta\int\frac{dp(x;\theta)}{d\theta}dx - \frac{\delta\theta^2}{2}\int\left(\frac{d^2p(x;\theta)}{d\theta^2} - \frac{1}{p(x;\theta)}\left(\frac{dp(x;\theta)}{d\theta}\right)^2\right)dx \\
&= -\delta\theta\tfrac{d}{d\theta}\int p(x;\theta)dx - \frac{\delta\theta^2}{2}\frac{d^2}{d\theta^2}\int p(x;\theta)dx + \frac{\delta\theta^2}{2}\int\frac{1}{p(x;\theta)}\left(\frac{dp(x;\theta)}{d\theta}\right)^2dx \\
&= 0 + 0 + \tfrac{\delta\theta^2}{2}\int p(x;\theta)\left(\frac{1}{p(x;\theta)}\frac{dp(x;\theta)}{d\theta}\right)^2dx \\
&= \tfrac{\delta\theta^2}{2}\int p(x;\theta)\left(\frac{d\log p(x;\theta)}{d\theta}\right)^2dx \\
&= \tfrac{1}{2}\delta\theta^2 I(\theta)
\end{aligned}$$

So

$$D_{KL}(P_\theta, P_{\theta_0}) = \frac{1}{2}\delta\theta^2 I(\theta)$$

and

$$I(\theta) = \frac{d^2}{d\theta^2}D_{KL}(P_\theta, P_{\theta_0})\Big|_{\theta_0=\theta}$$

■

This result should remind the reader of the third condition in the definition of a divergence as it shows that the KL-divergence can be Taylor expanded in terms of the Fisher information matrix, (I_{jk}).

4. Central Limit Theorem

In this section we present a proof of the Central Limit Theorem by means of the KL-divergence. We will show first that for a sequence of standardized partial sums, S_n, the KL-divergence between S_n and the normal distribution goes to 0 as $n \to \infty$. Then we will show that this implies that S_n converges weakly to the normal distribution. This proof will follow a paper by Andrew Barron [3].

In this section, we denote the standardized Fisher information as $J(Y)$. I.e. for a random variable Y with density $g(y)$,

$$J(Y) = \sigma^2\mathbb{E}\left[(\rho(Y) - \rho_\phi(Y))^2\right]$$

where $\rho = g'/g$ and $\rho_\phi = \phi'/\phi$. Furthermore, these proofs will make use of two lemmas which are proven in [3] and are provided here without proof.

Lemma 4.1. *Let X be a random variable with mean μ and variance $\sigma^2 < \infty$. Then*

$$D_{KL}(X, \phi) = \int_0^1 \frac{1}{2t} J\left(\sqrt{t}X + \sqrt{1-t}Z\right) dt$$

where $Z \sim \mathcal{N}(\mu, \sigma^2)$.

Lemma 4.2. *Let*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$$

and

$$S'_n = \sqrt{t} S_n + \sqrt{1-t} Z$$

for $0 \leqslant t < 1$. Then

(i) $(p+q)J(S'_{p+q}) \leqslant pJ(S'_p) + qJ(S'_q)$

(ii) $J(S'_{2n}) \leqslant J(S'_n)$

(iii) $\lim_{n\to\infty} J(S'_n) = 0$

Now we prove the first result.

Theorem 4.3. *Let $X_1, ..., X_n$ be a sequence of iid random variables with mean 0 and variance σ^2. Define their partial sum as*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$$

with density $f_n(x)$. Furthermore, let $\phi(x)$ be the density of the $\mathcal{N}(0, \sigma^2)$ distribution. Then

$$\lim_{n\to\infty} D_{KL}(f_n, \phi) = 0$$

Proof. First we prove that the limit exists. To begin we recognize that $D_{KL}(f_{mp}, \phi) \leqslant D_{KL}(f_p, \phi)$ for $m \in \mathbb{N}$. We will prove this by induction on m. In the case of $m = 2$, by lemma 4.2 (ii), $J(S'_{2p}) \leqslant J(S'_p)$ so by monotonicity, $D_{KL}(S_{2p}) \leqslant D_{KL}(S_p)$. Now suppose it holds for $m-1$. By lemma 4.2 (i),

$$mpJ(S'_{mp}) \leqslant pJ(S'_p) + (m-1)pJ(S'_{(m-1)p})$$

thus

$$J(S'_{mp}) \leqslant \frac{1}{m} J(S'_p) + \frac{m-1}{m} J(S'_{(m-1)p})$$

But by assumption, $J(S'_{(m-1)p}) \leqslant J(S'_p)$. So

$$\begin{aligned} J(S'_{mp}) &\leqslant \tfrac{1}{m} J(S'_p) + \tfrac{m-1}{m} J(S'_p) \\ &= J(S'_p) \end{aligned}$$

Thus by induction, $J(S'_{mp}) \leqslant J(S'_p)$ and by monotonicity, $D_{KL}(f_{mp}, \phi) \leqslant D_{KL}(f_p, \phi)$. Now choose p such that

$$D_{KL}(f_p, \phi) \leqslant \inf_{n \geqslant 1} D_{KL}(f_n, \phi) + \epsilon$$

for some $\epsilon > 0$. Define $r < p$ such that $n = mp + r$. Now, we make use of the inequality presented in [3],

$$D_{KL}(f_r, \phi) \leqslant -\frac{1}{2} \log\left(1 - \frac{r}{n}\right)$$

Then

$$\begin{aligned} D_{KL}(f_n, \phi) &\leqslant D_{KL}(f_{mp}, \phi) + D_{KL}(f_r, \phi) \\ &\leqslant D_{KL}(f_{mp}, \phi) - \tfrac{1}{2} \log\left(1 - \tfrac{r}{n}\right) \\ &\leqslant D_{KL}(f_p, \phi) - \tfrac{1}{2} \log\left(1 - \tfrac{p}{n}\right) \\ &\leqslant \inf_{n \geqslant 1} D_{KL}(f_n, \phi) + \epsilon - \tfrac{1}{2} \log\left(1 - \tfrac{p}{n}\right) \end{aligned}$$

So

$$\begin{aligned} \lim_{n \to \infty} D_{KL}(f_n, \phi) &\leqslant \liminf_{n \to \infty} D_{KL}(f_n, \phi) + \epsilon - \tfrac{1}{2} \lim_{n \to \infty} \log\left(1 - \tfrac{p}{n}\right) \\ &= \liminf_{n \to \infty} D_{KL}(f_n, \phi) + \epsilon \end{aligned}$$

But $\epsilon > 0$ is arbitrary, so sending $\epsilon \downarrow 0$ we get

$$\lim_{n \to \infty} D_{KL}(f_n, \phi) = \liminf_{n \to \infty} D_{KL}(f_n, \phi)$$

Thus the limit exists. Now we prove that the limit is 0. Consider the subsequence $\{n_k\}_{k \geqslant 1} = \{2^k\}_{k \geqslant 1}$. Then by lemma 4.2 (ii) and (iii), $J(S'_{n_k}) \downarrow 0$ and thus $D_{KL}(f_{n_k}, \phi) \downarrow 0$ by monotone convergence (assuming that $D_{KL}(f_{n_k}, \phi)$ is finite eventually for some $k \geqslant 1$). Since the limit exists, we conclude that

$$\lim_{n \to \infty} D_{KL}(f_n, \phi) = 0$$

■

Theorem 4.4. *For $X_1, ..., X_n, S_n, f_n$ and ϕ as defined in theorem 4.4, if*

$$\lim_{n \to \infty} D_{KL}(f_n, \phi) = 0$$

then

$$S_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Proof. This follows from the inequality due to Csiszar [4],

$$\left(\int |f_n - \phi|\right)^2 \leqslant 2 D_{KL}(f_n, \phi)$$

Writing,

$$-\sqrt{2 D_{KL}(f_n, \phi)} \leqslant \int |f_n - \phi| \leqslant \sqrt{2 D_{KL}(f_n, \phi)}$$

Taking limits of both sides gives us

$$\lim_{n\to\infty} \int |f_n - \phi| = 0$$

Thus $f_n \xrightarrow{\mathcal{L}^1} \phi$. Now we will show that this implies convergence in distribution. Let F_n and Φ be the cdf's of S_n and the $\mathcal{N}(0, \sigma^2)$ respectively. Note that Φ is continuous at all points $x \in \mathbb{R}$, thus it suffices to show convergence of the cdf's for an arbitrary $x \in \mathbb{R}$.

$$\begin{aligned} |F_n(x) - \Phi(x)| &= \left| \int_{-\infty}^{x} f_n - \int_{-\infty}^{x} \phi \right| \\ &= \left| \int_{-\infty}^{x} (f_n - \phi) \right| \\ &\leqslant \int_{-\infty}^{x} |f_n - \phi| \\ &\leqslant \int_{-\infty}^{\infty} |f_n - \phi| \\ &\to 0 \end{aligned}$$

Then by definition,

$$S_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

■

5. Applications

5.1. **Mean Field Theory.** [2] Motivated by interacting particle systems in physics, mean field theory is an area of probability theory which seeks to simplify the problem of characterizing stochastic systems with many interacting components. Writing the probability distribution of the system in the form

$$q(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\nu}) = \exp\left\{ \boldsymbol{\theta} \cdot \boldsymbol{x} + \sum_{r=1}^{L} \nu_r c_r(x_r) - \psi(\boldsymbol{\theta}, \boldsymbol{\nu}) \right\}$$

we can see that the interactions between components are represented by the second term in the exponential. So when this sum goes to 0, or $\boldsymbol{\nu} = \mathbf{0}$, the interdependence of the distribution disappears. If we define the submanifold of distributions with no interactions,

$$M_0 = \{q(\boldsymbol{x}, \boldsymbol{\theta}, \mathbf{0})\}$$

it can be proven that the m-projection of $q(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\nu})$ to M_0, i.e., minimizing $D_{KL}(q(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\nu}), p)$ over $p \in M_0$, retains the mean of $\boldsymbol{x}$. This allows one to consider a simplified system with no interactions without sacrificing too much of the information present in the original model. However, the m-projection is often computationally infeasible so often the e-projection, which minimizes $D_{KL}(p, q(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\nu}))$, is used as the mean field approximation.

5.2. **Machine Learning.** Machine learning is an exciting area of research at the intersection of statistics, computer science, and optimization in which researchers develop algorithms which can be trained on existing data to make accurate predictions about future data. One type of problem which is commonly addressed in a machine learning scenario is that of clustering. In this situation, algorithms are used to identify clusters of data based on their covariates and some measure of distance. Information geometry offers an alternative perspective from the standard view of clustering.

[2] Consider the k-means algorithm. In this clustering problem, we wish to assign each data point to one of k possible groups with unknown means. Typically, these group assignments and the resulting group means are chosen in order to minimize the total squared euclidean distance between each data point and its group's mean. However, information geometry allows us to extend this naturally to other distance measures by considering the general case of a dually flat divergence. This allows for a much more general and flexible formulation of the k-means algorithm. It can be shown that this formulation also terminates in finitely many steps but does not guarantee optimality.

References

[1] Amari, Shun-ichi. *Differential-geometrical methods in statistics*. Vol. 28. Springer Science & Business Media, 2012.

[2] Amari, Shun-ichi. *Information geometry and its applications*. Springer Japan, 2016.

[3] Barron, Andrew R. *Entropy and the central limit theorem*. The Annals of probability (1986): 336-342.

[4] Csiszar, I. *Information-type measures of difference of probability distributions and indirect observations*. Studia Sci. Math. Hungar. 2 299-318.

[5] Grosse, Roger. *Information geometry for machine learning*. <https://metacademy.org/roadmaps/rgrosse/dgml>.

[6] Kullback, Solomon. *Information theory and statistics*. Courier Corporation, 1997.

[7] Lee, John M. *Introduction to Smooth Manifolds*. Springer New York, 2003. 1-29.

[8] Nielsen, Frank. *Cramer-Rao lower bound and information geometry*. arXiv preprint arXiv:1301.3578 (2013).

[9] Cover, Thomas M., and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.