An Overview of Information Geometry

Yuxi Liu

Abstract

This paper summarizes the basics concepts of information geometry, and gives as example applications, Jeffreys prior in Bayesian probability, and natural gradient descent in machine learning. Necessary concepts in probability and statistics are explained in detail.

The main purpose of the paper is to provide, for people familiar with differential geometry, an accessible overview and entry point into information geometry. No originality is claimed in content.

1 Introduction

Information geometry applies methods of differential geometry to problems in statistics, probability theory, data analysis, and many other related fields. Despite its wide reach, and , it is not as well-known or well-understood as it should. This is probably due to a lack of accessible introductions.

This paper aims to provide such an introduction. The intended audience has a good symbolic and intuitive understanding of modern differential geometry, but merely symbolic understanding of probability, statistics, and machine learning. As such, all such concepts used in the paper are explained in detail, from the bare basics, while the differential geometry is written in ruthless telegraphic style.

Standard references of information geometry are [Amari and Nagaoka, 2007] and [Amari, 2016].

We use Einstein summation throughout the paper.

2 Probability concepts

In this section, we define some probability concepts we will use in the paper. See [Cover and Thomas, 2006, chapter 2] for a detailed exposition.

We consider a probability space $(\Omega, \mathcal{B}, P)$, where Ω is the state space, and P is the probability measure on the measure space $(\Omega, \mathcal{B})$. The probability space is either discrete or continuous, depending on the context. If it is discrete, we use the notation $\Omega = \{0, 1, \cdots\}$, and $P_i = P(\{i\})$. If it is continuous, we assume that the probability measure has a density $p : \Omega \to [0, \infty)$, so that any measurable subset $A \subset \Omega, P(A) = \int_A p(x)dx$.

In information theory, the entropy of a random variable quantifies the amount of randomness in a random variable.

Definition 2.1. The **entropy** of a discrete random variable P is

$$H(P) = -\sum_i P_i \ln P_i. \tag{1}$$

For a continuous random variable, we can define an analogous quantity on its probability density:

Definition 2.2. The **differential entropy** or **continuous entropy** of a probability density p is

$$h(p) = -\int_\Omega p(x) \ln p(x)dx. \tag{2}$$

One should be cautious to note that the differential entropy is not defined for the random variable P itself, but for its probability density p. Consider for instance the uniform random variable on $[0, 1]$. Its probability density function is $p(x) = 1$. So we have $h(p) = 0$. However, we could trivially transform the random variable by stretching its state space to $[0, k]$, then we have transformed probability density $p(x) = \frac{1}{k}, h(p) = \ln k$ which could be negative. In contrast, the entropy of a discrete random variable is always nonnegative.

Philosophically, in probability theory, one is interested in the random variable itself, rather than its presentation. As such, if a quantity, such as the differential entropy, depends on how the random variable is presented, then it is not a concept in probability, but rather, a concept in mathematical analysis.

When there are two random variables P, Q, we can measure their difference by

Definition 2.3. The **relative entropy** or **Kullback–Leibler divergence** between two probability densities p, q is

$$D(p\|q) = \int_\Omega p(x) \ln \frac{p(x)}{q(x)} dx. \tag{3}$$

When both random variables are discrete, this reduces to

$$D(P\|Q) = \sum_i P_i \ln \frac{P_i}{Q_i}. \tag{4}$$

It can be proven that the Kullback–Leibler divergence is always nonnegative, and is zero iff $p = q$. This result is called the **Gibbs inequality**. It also does not depend on the presentation of P, Q, that is, under a change of state space $f : \Omega \to \Omega'$, p, q are changed to p', q', but as long as f does not "lose information",

$$D(p\|q) = D(p'\|q').$$

Note that this is not a distance, since in general, $D(p\|q) \neq D(q\|p)$, and it also does not satisfy the triangle inequality. This is why it's called a "divergence".

3 Defining the information manifold

For this section, we follow [Caticha, 2015], and [Leinster, 2018].

Now we consider a family of probability densities $\{p_\theta | \theta = (\theta^1, \cdots, \theta^n) \in M\}$, where θ is usually called the **parameter** of the family, and M is an open submanifold of $\mathbb{R}^n$. This M is the **information manifold**, or **statistical manifold**. In general, an information manifold can be constructed by multiple parametrizations, each parametrization by θ being one chart of the manifold. We will have no need for such generality however, and would develop the theory as if M is covered by one chart. We will also assume that $p(x|\theta)$ is smooth in θ.

In this view, we immediately see that the continuous entropy function h is a scalar function on the manifold. Differentiating under the integral sign shows that $h \in C^\infty(M)$.

We would endow it with a Riemannian metric that measures the distance between infinitesimally close probability densities $p(x|\theta)$ and $p(x|\theta + d\theta)$.

Example 3.1. Consider the family of normal distribution on the real line. Then each is characterized by its mean μ and variance σ^2, with $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, and

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

The Kullback–Leibler divergence between two points on an information manifold is an instance of a general concept, divergence, which can be thought of as an asymmetric generalization of distance.

Definition 3.1. The **divergence** on a manifold M is some function $D(\cdot\|\cdot) : M \times M \to [0, \infty)$, such that $D(q\|p) = 0$ iff $p = q$, and in any local chart θ, we have

$$D(\theta + d\theta\|\theta) = \frac{1}{2} g_{ij} d\theta^i d\theta^j + o(|d\theta|^2) \tag{6}$$

for some positive definite matrix $[g_{ij}]$.

A divergence has an associated Riemannian metric g_{ij} as given in the definition.

3.1 Metric via distinguishability

Consider the relative difference of $p(x|\theta)$ and $p(x|\theta + d\theta)$:

$$\Delta_\theta(d\theta) = \frac{p(x|\theta) - p(x|\theta + d\theta)}{p(x|\theta)} = \frac{\partial \ln p(x|\theta)}{\partial \theta^i} d\theta^i \tag{7}$$

We can interpret this as how distinguishable the two densities are, if all we have are their values at a particular $x \in \Omega$. If it is zero, then they appear the same. The bigger its absolute value is, the more they differ.

We define convenient notation $p_\theta(x) = p(x|\theta)$ and the log-density function $l_\theta(x) = l(x|\theta) = \ln p(x|\theta)$.

Fixing a parameter $\theta \in M$ and an infinitesimal variation $d\theta$, the expected value of $\Delta_\theta(d\theta)$ is

$$\mathbb{E}_\theta[\Delta_\theta(d\theta)] = \int_\Omega \Delta_\theta(d\theta) dx = \int_\Omega (p_{\theta+d\theta}(x) - p_\theta(x)) dx = 0 \tag{8}$$

which is unfortunate. The square, however, is nontrivial:

$$\mathbb{E}_\theta[\Delta_\theta(d\theta)^2] = \left(\int_\Omega p_\theta(x) \frac{\partial \ln p_\theta(x)}{\partial \theta^i} \frac{\partial \ln p_\theta(x)}{\partial \theta^j} dx \right) d\theta^i d\theta^j \tag{9}$$

This allows us to define a metric on M.

Definition 3.2. The **Fisher information metric** on M is defined by

$$g_{ij}(\theta) = \mathbb{E}_\theta \left[\partial_i l \partial_j l \right]. \tag{10}$$

Note that we always take partial derivatives over the parameters θ unless otherwise specified, so $\partial_i l$ is partial derivative by θ^i.

It's easy to verify that g_{ij} is a symmetric $(2, 0)$-tensor (check that it transforms covariantly under a coordinate change $\theta \mapsto \theta'$). It remains to show that it's positive-definite, that is, for all $d\theta$, $g_{ij} d\theta^i d\theta^j \geq 0$, with equality hold iff $d\theta = 0$.

From definition, $g_{ij} d\theta^i d\theta^j = \langle \Delta_\theta(d\theta)^2 \rangle$ is the expected value of a nonnegative number, so $g_{ij} d\theta^i d\theta^j \geq 0$. If it equals zero, then we have $\Delta_\theta(d\theta)^2 = 0$ almost surely, that is, $\frac{p_{\theta+d\theta}(x) - p_\theta(x)}{p_\theta(x)} = 0$ almost surely. So $p_{\theta+d\theta}(x) = p_\theta(x)$ almost surely.

We thus assume that the manifold coordinates are **regular**, that is, the Fisher metric is positive-definite in that coordinate.

3.2 Metric via Kullback–Leibler divergence

Consider the Kullback–Leibler divergence of two densities p, q:

$$D(p\|q) = \int_\Omega p(x) \ln \frac{p(x)}{q(x)} dx = -\int_\Omega p(x) \ln \left(1 + \frac{q(x) - p(x)}{p(x)}\right) dx$$

expand in Taylor series to second order,

$$\approx \int_\Omega (p(x) - q(x)) + \frac{1}{2} \frac{(p(x) - q(x))^2}{p(x)} dx. \tag{11}$$

The linear term integrates to zero, since $\int_\Omega p(x)dx = \int_\Omega q(x)dx = 1$, so

$$D(p\|q) \approx \frac{1}{2} \int_\Omega \left(\frac{p(x) - q(x)}{p(x)}\right)^2 p(x)dx. \tag{12}$$

Now let $p = p_{\theta+d\theta}, q = p_\theta$, we have

$$D(p_{\theta+d\theta}\|p_\theta) \approx \frac{1}{2} \left\langle \Delta_\theta (d\theta)^2 \right\rangle = \frac{1}{2} g_{ij} d\theta^i d\theta^j. \tag{13}$$

deriving the Fisher metric as the Kullback–Leibler divergence between infinitesimally close points on the information manifold, that is, it is the associated Riemann metric of Kullback–Leibler divergence.

We can recast this in another form. Start with

$$D(p_{\theta+d\theta}\|p_\theta) = \frac{1}{2} g_{ij}(\theta) d\theta^i d\theta^j \approx \frac{1}{2} g_{ij}(\theta + d\theta) d\theta^i d\theta^j = D(p_\theta\|p_{\theta+d\theta})$$

and perform Taylor expansion again, up to second order:

$$\ln p_{\theta+d\theta}(x) \approx \ln p_\theta(x) + s_\theta \cdot d\theta + \frac{1}{2} (\partial_i \partial_j l)\, d\theta^i d\theta^j$$

where $s_\theta = \nabla_\theta l_\theta$ is usually called the **score function**.

So

$$\begin{aligned} \frac{1}{2} g_{ij} d\theta^i d\theta^j &= D(p_\theta\|p_{\theta+d\theta}) \\ &= \int_\Omega p(x|\theta) \left(\ln p(x|\theta) - \ln p(x|\theta + d\theta)\right) dx \\ &\approx -\mathbb{E}_\theta \left[s_\theta \cdot d\theta + \frac{1}{2} (\partial_i \partial_j l)\, d\theta^i d\theta^j \right] \\ &= -\frac{1}{2} \mathbb{E}_\theta [\partial_i \partial_j l] d\theta^i d\theta^j \end{aligned} \tag{14}$$

where we used the fact that the expectation of score is zero:

$$\mathcal{E}_\theta[s_\theta] = \int_\Omega p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} dx = \nabla_\theta \int_\Omega p_\theta(x) dx = 0 \tag{15}$$

So we obtain

$$g_{ij} = -\mathbb{E}_\theta[\partial_i \partial_j l]. \tag{16}$$

3.3 Uniqueness of the Fisher metric

The Fisher metric has many derivations, and is in some sense the unique (up to a constant factor) Riemannian metric on the information manifold that is compatible with the probability distributions p_θ. More precisely, consider the idea of doing a transformation of the underlying state space Ω. Then this transforms the probability distributions p_θ too. If the transformation of state space is "lossy", that is, some different states are sent to the same states, then information is lost, otherwise, information is preserved. This can be formalized by the concept of **sufficient statistic** transform.

Now consider a divergence D and two points on the information manifold p_θ, p_δ. After a change of state space, the two points are transformed too, into some p'_θ, p'_δ. If the change of state space is lossy, then we expect that the difference between the two distributions would become blurred, that is, they look more similar. Otherwise, we expect their difference to look the same.

That is, we expect

$$D(p_\theta \| p_\delta) \geq D(p'_\theta \| p'_\delta) \tag{17}$$

where equality holds iff the transform is a sufficient statistic transform. This property of divergence D is called **information monotonicity**.

Then, we can show that any divergence that satisfies information monotonicity induces the Fisher metric. This was proved by Chentsov in 1972. For a detailed proof, see [Amari, 2016, chapter 3].

3.4 Examples

3.4.1 Normal distribution

Consider the one-dimensional normal distribution $p(x|\mu, \sigma^2)$ defined in Example 3.1, then the Fisher metric on the upper half plane $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ makes it into the Poincaré half-plane model (with a horizontal stretching). The geodesics are upper-half elliptic arcs or verticle half-lines, and given any two normal distributions, one can connect them with a geodesic line segment and interpolate between them using the arc length as a parameter. This interpolation by arc length is in some sense natural, even though its practical significance is unclear.

For a detailed calculation and some pictures, see [Costa et al., 2015].

3.4.2 Jeffreys prior

In Bayes statistics, one always starts with an assumption on the random events one wishes to study. For example, before one investigates a coin's flips, one must make an assumption on its behavior.

As usual, assume the coin has no memory, and has chance $\theta \in (0, 1)$ of coming up heads. One must make one further assumption on the probability distribution of θ on $(0, 1)$. This is the all-important **prior** distribution. After choosing the prior, one performs experiments, and after that, uses Bayes theorem to update the prior into a **posterior**.

In general, one is faced with the problem of determining a good prior before one starts the experiment. It must be in some sense be completely non-informative, so that it reflects the experimenter's maximal ignorance.

Given the Fisher metric on the information manifold, we can define the associated volume density $\sqrt{|\det(g)|}$ on the manifold. Since the Fisher metric is in some sense uniquely defined on the manifold, the volume density is also uniquely defined. This has led Jeffreys [Jeffreys, 1946] to propose it as a useful non-informative prior in Bayesian probability. This prior is called the **Jeffreys prior**.

For example, in our coin-flip experiment, we have

$$p(T|\theta) = 1 - \theta, \quad p(H|\theta) = \theta \tag{18}$$

giving the Fisher metric

$$g_{\theta\theta}(\theta) = \mathbb{E}_\theta[\partial_\theta l \partial_\theta l] = (1-\theta)\left(-\frac{1}{1-\theta}\right)^2 + \theta\left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta(1-\theta)} \tag{19}$$

and the Jeffreys prior

$$\rho(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}. \tag{20}$$

Intuitively, Jeffreys prior gives higher weight to regions on the information manifold with more distinguishability. In our coin-flip example, the prior is heavily weighted on the two extreme ends, which can be intuitively understood thus: whereas two coins with $\theta = 0.4, 0.5$ are quite similar, two coins with $\theta = 0.001, 0.101$ are very different, with one almost never coming up heads, the other once in a while. This shows that distinguishability is concentrated on the extreme ends.

4 Connections on the information manifold

Now we define families of connections on the information manifold. For this section we follow [Amari, 2016, chapter 6] and [Nielsen, 2018].

4.1 Levi-Civita connection

We set up the notation quickly. For details, refer to [Lee, 2018].

Consider a general smooth manifold M with a local coordinate chart $(x^1, \cdots x^n)$ around $p \in M$. Any connection on that region is defined by n^3 smooth functions Γ_{ij}^k, so that the covariant derivative induced by the connection gives $\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k \partial_k$. The connection is **symmetric** or **torsionfree** iff for any vector fields X, Y,

$$\nabla_X Y - \nabla_Y X - [X, Y] = 0. \tag{21}$$

In coordinates,

$$\Gamma_{ij}^k = \Gamma_{ji}^k. \tag{22}$$

Suppose M has Riemannian metric g, then we define

$$\Gamma_{ijk} = \Gamma_{ij}^l g_{lk} = \langle \nabla_{\partial_i}\partial_j, \partial_k \rangle \tag{23}$$

Take inner products of the two vector fields X, Y as a scalar field $\langle X, Y \rangle \in C^\infty(M)$. Then take another vector field Z, we can differentiate the scalar field along the vector field:

$$Z \langle X, Y \rangle = \nabla_Z \langle X, Y \rangle . \tag{24}$$

The connection ∇ **preserves** the metric g iff we have a "product rule":

$$\nabla_Z \langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle \tag{25}$$

In coordinates,

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma_{kji} \tag{26}$$

The **Levi-Civita connection** or **Riemannian connection** on M is the unique connection such that is symmetric and preserves the metric. In coordinates, we have

$$\Gamma_{ijk} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}) \tag{27}$$

4.2 Dual connection

We can generalize the concept of a metric-preserving connection by considering two connections $\nabla, \tilde{\nabla}$, such that a generalized form of Equation 25 holds:

$$Z \langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \left\langle X, \tilde{\nabla}_Z Y \right\rangle \tag{28}$$

In coordinates,

$$\partial_i g_{jk} = \Gamma_{ijk} + \tilde{\Gamma}_{ikj} \tag{29}$$

which shows that each connection has a unique dual connection.

Definition 4.1. Two connections are **dual connections** with respect to g iff Equation 28 holds for any smooth vector fields X, Y, Z.

Definition 4.2. A **dual manifold** is a Riemannian manifold equipped with a pair of dual connections $(M, g, \nabla, \tilde{\nabla})$.

Note that since the metric is symmetric: $\langle X, Y \rangle = \langle Y, Y \rangle$, duality is also symmetric, that is, if Equation 28 holds, then we also have

$$Z \langle X, Y \rangle = \left\langle \tilde{\nabla}_Z X, Y \right\rangle + \langle X, \nabla_Z Y \rangle$$

Just as Levi-Civita connection preserves the inner product of vectors parallelly transported along any path, we have an analogous result:

Theorem 4.1. *Given dual connections $\nabla, \tilde{\nabla}$, and two vectors $X(t), Y(t)$ that are respectively parallel-transported by $\nabla, \tilde{\nabla}$ along a path $\gamma(t) : I \to M$, then their inner product is constant along the path.*

Proof

$$\begin{aligned} \frac{d}{dt} \langle X(t), Y(t) \rangle &= \dot{\gamma}(t) \langle X(t), Y(t) \rangle \\ &= \left\langle \nabla_{\dot{\gamma}(t)} X(t), Y(t) \right\rangle + \left\langle X(t), \tilde{\nabla}_{\dot{\gamma}(t)} Y(t) \right\rangle \\ &= \langle 0, Y(t) \rangle + \langle X(t), 0 \rangle \end{aligned}$$

□

Now we consider curvatures on dual manifolds.

Definition 4.3. A connection ∇ is **flat** iff its Riemann curvature tensor is zero.

Recall that the Riemann curvature tensor is defined by

$$R_\nabla(X,Y,Z,W) = \left\langle (\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]})Z, W \right\rangle \tag{30}$$

Then we can expand out, using Equation 28,

$$0 = (XY - YX - [X,Y]) \langle Z, W \rangle = R_\nabla(X,Y,Z,W) + R_{\tilde{\nabla}}(X,Y,W,Z)$$

Recall the sectional curvature tensor is defined by

$$K_\nabla(X,Y) = R_\nabla(X,Y,X,Y) \tag{31}$$

Then, by antisymmetry of the Riemann curvature tensor in the first two inputs,

$$R_\nabla(X,Y,X,Y) = -R_{\tilde{\nabla}}(X,Y,Y,X) = R_{\tilde{\nabla}}(Y,X,Y,X)$$

So we have the "fundamental theorem of information geometry":

Theorem 4.2. *Given dual manifold $(M, g, \nabla, \tilde{\nabla})$, the Riemann curvature tensors of the dual connections satisfy*

$$R_\nabla(X,Y,Z,W) + R_{\tilde{\nabla}}(X,Y,W,Z) = 0 \tag{32}$$

and their sectional curvature tensors satisfy

$$K_\nabla(X,Y) = K_{\tilde{\nabla}}(Y,X) \tag{33}$$

In particular, if one of the dual connections have constant sectional curvature K, so does the other, and if one of them is flat, so does the other.

Note that if $\nabla = \tilde{\nabla}$, that is, ∇ preserves the metric g, then this means R_∇ is antisymmetric in its last two inputs, which is a standard result in Riemannian geometry, see for example [Lee, 2018, Proposition 7.12 (b)].

Definition 4.4. A dual manifold is flat iff one of its connections is flat, which means, in light of Theorem 4.2, iff both of its connections are flat.

Given a pair of dual connections, their average preserves the metric:

$$Z \langle X, Y \rangle = \left\langle \left(\frac{\nabla + \tilde{\nabla}}{2} \right)_Z X, Y \right\rangle + \left\langle X, \left(\frac{\nabla + \tilde{\nabla}}{2} \right)_Z Y \right\rangle$$

and if both connections are symmetric, their average is also symmetric, so it is the Levi-Civita connection.

We define a linear space of connections on a dual manifold by

Definition 4.5. Given dual manifold $(M, g, \nabla, \tilde{\nabla})$, for all $\alpha \in \mathbb{R}$, the associated **α-connection** is

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla + \frac{1-\alpha}{2}\tilde{\nabla} \tag{34}$$

So $\nabla^{(1)} = \nabla, \nabla^{(-1)} = \tilde{\nabla}$.

The connection coefficients of the α-connection are

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(0)}_{ijk} - \frac{\alpha}{2}T_{ijk} \tag{35}$$

We will restrict our attention to dual manifolds with symmetric connections, so that $\nabla^{(0)}$ is the Levi-Civita connection.

4.3 Cubic tensor

While the connection coefficients Γ_{ijk} are not tensors, the difference between any two is, so

Definition 4.6. The **cubic tensor** or **Amari–Chentsov tensor** of a dual manifold with symmetric connections $(M, g, \nabla, \tilde{\nabla})$ is

$$T_{ijk} = \tilde{\Gamma}_{ijk} - \Gamma_{ijk} \tag{36}$$

Or, in coordinate-free notation, for any vector fields X, Y, Z,

$$T(X, Y, Z) = \left\langle \nabla_X Y - \tilde{\nabla}_X Y, Z \right\rangle \tag{37}$$

Theorem 4.3. *The cubic tensor is symmetric, and in coordinates,*

$$\nabla_{\partial_i} g_{jk} = T_{ijk} \tag{38}$$

Proof By Equation 29, 36,

$$\partial_i g_{jk} = (\Gamma_{ijk} + \Gamma_{ikj}) + T_{ikj}$$

So

$$T_{ikj} = \partial_i g_{jk} - (\Gamma_{ijk} + \Gamma_{ikj}) = \nabla_{\partial_i} g_{jk}$$

So $T_{ikj} = T_{ijk}$. Also, by Equation 36, since $\nabla, \tilde{\nabla}$ are symmetric, T_{ijk} is symmetric in i, j, so T is totally symmetric. □

Theorem 4.4. *For all $\alpha \in \mathbb{R}$, $\nabla^{(\alpha)}, \nabla^{(-\alpha)}$ are dual.*

Proof By Equation 35 and symmetry of T,

$$\Gamma^{(\alpha)}_{ijk} + \Gamma^{(-\alpha)}_{ikj} = \Gamma^{(0)}_{ijk} + \Gamma^{(0)}_{ikj} = \partial_i g_{jk}$$

□

4.4 Divergence derives dual connections

Now we derive a metric and a pair of dual connections on any manifold M with a divergence D. This would then be applied to the information manifold with Kullback–Leibler divergence as the divergence, giving the Fisher metric and the dual connections on the information manifold.

First, we define some convenient notations:

$$\partial_{i,\cdot} = \frac{\partial}{\partial \theta^i}, \quad \partial_{\cdot,j} = \frac{\partial}{\partial \theta'^j} \tag{39}$$

Then recall that the divergence D, when written in local coordinates θ, has the form

$$D(\theta + d\theta \| \theta + d\theta') = \frac{1}{2} g_{ij}(\theta)(d\theta^i - d\theta'^i)(d\theta^j - d\theta'^j) + o(|d\theta - d\theta'|^2) \tag{40}$$

for infinitesimal $d\theta, d\theta'$, so we have

$$g_{ij}(\theta) = -\partial_{i,j} D(\theta \| \theta')|_{\theta=\theta'} \tag{41}$$

So far this is nothing new. But now we can define the dual connections

$$\begin{aligned} \Gamma_{ijk} &= -\partial_{k,ij} D(\theta \| \theta')|_{\theta=\theta'} \\ \tilde{\Gamma}_{ijk} &= -\partial_{ij,k} D(\theta \| \theta')|_{\theta=\theta'} \end{aligned} \tag{42}$$

Theorem 4.5. *The pair of connections defined by Equation 42 are dual with respect to the metric g.*

Proof Let $g_{ij}(\theta \| \theta') = -\partial_{i,j} D(\theta \| \theta')$, so that

$$g_{ij}(\theta) = g_{ij}(\theta \| \theta)$$

Then we have

$$\partial_k g_{ij}(\theta) = \partial_k g_{ij}(\theta \| \theta) = (\partial_{k,\cdot} + \partial_{\cdot,k})(g_{ij}(\theta \| \theta'))|_{\theta=\theta'} = \Gamma_{kij}(\theta) + \tilde{\Gamma}_{kji}(\theta)$$

□

4.5 Examples of dual connections

Now, for the particular case of M being the information manifold, and D being the Kullback–Leibler divergence, direct calculation gives

$$\begin{aligned} \Gamma_{ijk} &= \mathbb{E}_\theta[\partial_k l \partial_{ij} l] \\ \tilde{\Gamma}_{ijk} &= \mathbb{E}_\theta[\partial_k l(\partial_i l \partial_j l + \partial_{ij} l)] \\ T_{ijk} &= \mathbb{E}_\theta[\partial_i l \partial_j l \partial_k l] \end{aligned} \tag{43}$$

From this, we can obtain the family of $\nabla^{(\alpha)}$-connections, as defined in Equation 34. It turns out that they can all be derived by divergences, called α-**divergences**. To define it, we first define the f-**divergence**:

$$D_f(p\|q) = \int_\Omega q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{44}$$

where f is convex, and $f(1) = 0$.

Then, the α-divergences are defined by

$$f(t) = \begin{cases} \frac{4}{1-\alpha^2}(1 - t^{(1+\alpha)/2}), & \text{if } \alpha \neq \pm 1 \\ t \ln t, & \text{if } \alpha = 1 \\ -\ln t, & \text{if } \alpha = -1 \end{cases} \tag{45}$$

Direct computation shows that the α-connections are really induces by them.

5 Two important statistical families

Now we consider two important statistical families: the exponential and the mixture families. They are especially important in information geometry as the prototypical information manifolds, just as the Euclidean space is the prototypical manifold in differential geometry.

5.1 Mixture family

Consider two coins with different probabilities of coming up heads, so that their probability distributions can be modelled as p_1, p_2, with $p_1(H) = 0.55, p_1(T) = 0.45$, for example.

Then we can "mix" the two coins by using a third coin, which has a probability of θ of coming up heads. Then, we can mix the two coins by this procedure:

(1) Flip the third coin.

(2) If it comes up heads, choose the first coin, else, choose the second coin.

(3) Flip the chosen coin and use the result.

Then the probability distribution of the outcome is

$$\theta p_1 + (1-\theta)p_2 = \theta(p_1 - p_2) + p_2 \tag{46}$$

This kind of mixing can be generalized to the concept of

Definition 5.1. A **mixture family** is a statistical family parametrized by $\theta \in \Theta$ where Θ is some open subset of $\mathbb{R}^n$, such that its distributions are

$$p_\theta(x) = \theta^i F_i(x) + C(x) \tag{47}$$

where

$$\int_\Omega F_i(x)dx = 0 \qquad \int_\Omega C(x)dx = 1$$

and F_i are linearly independent with some common support.

By direct calculation,

$$\tilde{\Gamma}_{ijk} = 0$$

for mixture families. So, for a mixture family defined by Equation 47, the parameters θ are affine parameters for the connection $\tilde{\nabla}$ on the information manifold, and so the $\tilde{\nabla}$-geodesics are of the form

$$\gamma(t) = p_{\theta t + \theta'(1-t)}, \quad t \in (a, b) \subset \mathbb{R} \tag{48}$$

That is, they are the linear mixtures between two probability distributions.

5.2 Exponential family

Whereas the mixture family comes from linear sums, the exponential family comes from "linear products", that is, linear sums of logarithms.

Definition 5.2. An **exponential family** is a statistical family parametrized by $\theta \in \Theta$ where Θ is some open subset of $\mathbb{R}^n$, such that its distributions are

$$p_\theta(x) = \exp\big(C(x) + \theta^i F_i(x) - \phi(\theta)\big) \tag{49}$$

where F_i are linearly independent with some common support, and ϕ is a normalization function defined such that $\int_\Omega p_\theta(x)dx = 1$ for all θ.

One might motivate considering "linear products" by considering what happens when one attempts to combine the two previously mentioned coins "in disjoint product" instead of "in disjoint sum". That is, one throws both of them together and considers all four possible outcomes HH, HT, TH, TT. Then, the probability of this mixture of coins would be of the form

$$p(HH) = p_1(H)p_2(H), p(HT) = p_1(H)p_2(T), \cdots$$

That is, $p = p_1 p_2$. One can then generalize this to "fractional mixtures" like

$$p = C p_1^{\theta^1} p_2^{\theta^2} p_3^{\theta^3} \cdots$$

where C is a normalization constant. Then this leads directly to the definition of the exponential family.

Exponential families are very prevalent in statistics. Statistical models that can be written in the form of exponential families include Gaussian distributions, gamma distributions, beta distributions, etc.

Basically, any statistical model where we are considering a "product" mixture of several pure distributions, we would be considering exponential families. Contrast this with a "sum" mixture, where we don't know which distribution we would be drawing from, but we know that it's one of the pure ones. This gives a mixture family.

By taking the gradient of $\int_\Omega p_\theta(x)dx = 1$, we get

$$\partial_i \phi = \mathbb{E}_\theta[F_i] \tag{50}$$

If we combine the F_i into a vector $\mathbf{F} = (F_1, \cdots F_n)$, then we get the more suggestive form

$$\nabla \phi = \mathbb{E}_\theta[\mathbf{F}] \tag{51}$$

Taking gradient again, we get the hessian

$$\nabla^2 \phi = \mathrm{Var}_\theta[\mathbf{F}] \tag{52}$$

We can calculate the Christoffel symbols of ∇ in θ coordinates,

$$\Gamma_{ijk} = \mathbb{E}_\theta[(F_k - \partial_k \phi)\partial_{ij}\phi] = (\mathbb{E}_\theta[F_k] - \partial_k \phi)\partial_{ij}\phi = 0$$

Thus, just as the case of mixture families, the information manifold is flat, and θ provide an affine parametrization relative to the ∇-connection. The ∇-geodesics look like straight lines when drawn in θ parameters.

5.3 Notation: e- and m- prefixes

In the case of exponential families, $\Gamma = 0$, so the lines are ∇-geodesics. In the case of mixture families, $\tilde{\Gamma} = 0$, so the lines are $\tilde{\nabla}$-geodesics. Thus, in the context of information geometry, for any dual manifold $(M, g, \nabla, \tilde{\nabla})$, even if it's not the information manifold of an exponential or mixture family, the ∇ connection is still often called the **e-connection**, and $\tilde{\nabla}$, the **m-connection**; the ∇-geodesics are called **e-geodesics**, and $\tilde{\nabla}$-geodesics, **m-geodesics**.

If the dual manifold is flat, then one can find a coordinate θ on it such that $\Gamma = 0$ in these coordinates, and the coordinate system would be called a set of **e-affine parameters**. Similarly for **m-affine parameters**.

6 Application to machine learning

Information geometry has many applications. Here we will only discuss its basic application to machine learning. Many applications can be found in [Amari, 2016].

6.1 Learning as optimization on manifolds

In machine learning, such as deep learning, one starts with a statistical model with various parameters and some training data, and adjusts the parameters during a training process, so that the model fits the training data.

Formally, consider the problem of learning a function of form $f(\boldsymbol{x}) = y$. The problem is to learn a function f_θ that approximates f using some training data $\{(\boldsymbol{x}_i, y_i) | i = 1, \cdots N\}$.

First, we turn the problem into a statistical problem. $f : \mathbb{R}^n \to \mathbb{R}$ is specified by a subset of $\mathbb{R}^n \times \mathbb{R}$, and we turn it into a probability distribution on $\mathbb{R}^n \times \mathbb{R}$. This can be accomplished by, for example, imposing a probability distribution $p_{\boldsymbol{x}}$ on $\boldsymbol{x}$, and adding a noise ϵ to y, so that $y = f(\boldsymbol{x}) + \epsilon$. Then, the joint probability distribution is

$$p(\boldsymbol{x}, y) = p_{\boldsymbol{x}}(\boldsymbol{x}) p_\epsilon(y - f(\boldsymbol{x})) \tag{53}$$

where p_ϵ is the probability distribution that the noise satisfies.

Then, the training data is drawn from the distribution, and the problem is to learn the best θ from the training data, such that f_θ closely approximates f.

To measure the closeness, one introduces a **loss function** $l(\boldsymbol{x}, y, \theta)$ to represent how bad the mistake is, if we use f_θ and encounter a sample $(\boldsymbol{x}, y)$. The most commonly used loss function is the squared error:

$$l(\boldsymbol{x}, y, \theta) = (y - f_\theta(\boldsymbol{x}))^2 \tag{54}$$

The goal is to minimize **expected loss**:

$$L(\theta) = \mathbb{E}_{(\boldsymbol{x},y)\sim p}[l(\boldsymbol{x}, y, \theta)] \tag{55}$$

Now, fix the distribution p of $(\boldsymbol{x}, y)$, we can view the loss function as a random variable parametrized by θ. Thus, we obtain an information manifold M parametrized by θ. The expected loss L is a scalar function on M, and is usually smooth. Thus, the problem reduces to minimizing a scalar function on a Riemannian manifold.

6.2 The natural gradient method

The standard method for finding a minimum is the **gradient descent** method. Let

$$\nabla L = (\partial_i L)_i \tag{56}$$

be an n-tuple function defined on M. This is often mistakenly called a "vector", but it is not a vector on M. The ∇ symbol is not a connection, either. It is simply a notation inherited from its use in vector calculus in Euclidean space.

Anyway, ∇L points to the direction where L increases, so if we go against it, we decrease L. The gradient descent method is defined by

$$\theta_{t+1} = \theta_t - \alpha_t \nabla L(\theta_t) \tag{57}$$

where α_t is the **learning rate** parameter, and usually is set to decrease to zero as t increases.

In practice, one uses the **stochastic gradient descent**, which is explained in any machine learning textbook.

As warned, the problem with ∇L is that it is not a vector on M, and is thus unnatural. In order to do gradient descent naturally, one starts with $dL : TM \to \mathbb{R}$, then use the metric to map this 1-form to a vector field, via a musical isomorphism:

$$\nabla_N L = (dL)^\sharp, \quad dL = \langle \nabla_N L, \cdot \rangle \tag{58}$$

In matrix notation, practical for machine learning,

$$[\nabla_N L] = [g]^{-1}[\nabla L] \tag{59}$$

where $[g]$ is the Fisher information matrix.

One step of natural gradient descent with learning rate α is a step that minimizes the expected loss L, under the constraint that the step must have Kullback–Leibler divergence approximately equal to $\frac{1}{2}\alpha^2$. This forces the update to be of substance. Under unnatural gradient descent, there is no guarantee on the Kullback–Leibler divergence of the step, thus no guarantee that the update has substance.

Compared to the more popular method of **stochastic gradient descent**, in general, natural gradient descent converges in fewer steps, but each step takes more time. As of current writing, stochastic gradient descent is still the dominant method in practical machine learning.

6.3 Other natural methods

There are other similar "natural" methods, such as natural Newton method, natural conjugate gradient descent, natural policy descent for reinforcement learning, etc. These are methods of **Riemannian optimization**, which are developed theoretically and numerically in [Absil et al., 2009].

Compared to more "unnatural" (ungeometrical) methods, natural methods all share the general theme of being (approximately) unaffected by a change of parameters, more likely to converge, converging in fewer steps, but each step being harder to numerically approximate. As a result, they do not dominate over unnatural methods in practical machine learning.

6.4 Singularities on the information manifold

Throughout this paper, we have always treated the information manifold as a Riemannian manifold, but this can fail when the metric stops being positive-definite, but merely positive-semidefinite. This happens when the Fisher information matrix becomes singular.

Singularities are unavoidable in practical machine learning. For example, consider the **artificial neural network** (in particular, it is a **multilayer perceptron** network) shown in Figure 1.

Artificial neural networks are inspired by biological neural networks. A single neuron i is modelled as a cell with n inputs $x_1, \cdots x_n$, and 1 output y_i. The inputs are weighted by **neural connection strengths** $w_{i1}, \cdots w_{in}$, and the cell outputs according to an **activation function** $\phi : \mathbb{R} \to \mathbb{R}$, so that $y_i = \phi(\boldsymbol{w}_i \cdot \boldsymbol{x})$.

Then, all the cells' outputs are weighted and summed to give the final output $y = \sum_{i=1}^{m} v_i y_i$

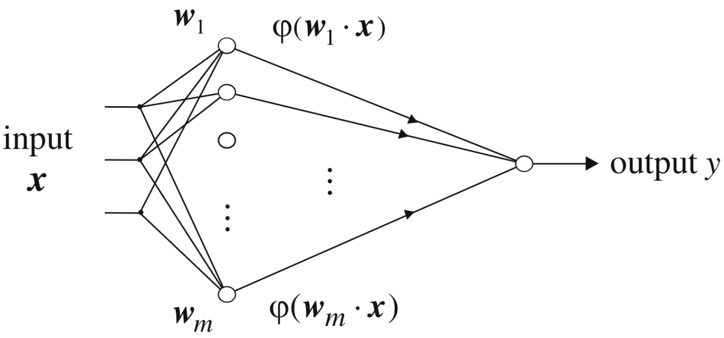
$\boldsymbol{w}_1$
$\varphi(\boldsymbol{w}_1 \cdot \boldsymbol{x})$
input
$\boldsymbol{x}$
output y
$\boldsymbol{w}_m$
$\varphi(\boldsymbol{w}_m \cdot \boldsymbol{x})$

Figure 1: A multilayer perceptron network with one hidden layer. Figure taken from [Amari, 2016, Fig. 12.4].

All together, the network is a graphical representation of a function f_θ, defined by

$$f_\theta(\boldsymbol{x}) = \sum_i^m v_i \phi(\boldsymbol{w}_i \cdot \boldsymbol{x}), \quad \theta = (\boldsymbol{w}_1, \cdots \boldsymbol{w}_m; v_1, \cdots v_m) \tag{60}$$

Here, θ is a tuple of the neural connection strengths. All boldfaced letters are n-tuples from $\mathbb{R}^n$.

Permutating the neurons in the hidden layer would not change the network's input-output behavior, so we obtain the same probability distribution for multiple parameters. We can imagine quotienting out these equivalent networks as $\overline{M} = M/\sim$. This produces a **neuromanifold**, which unfortunately is not a smooth manifold, as it has singularities.

The symmetry of neural networks with respect to permutating its hidden neurons, in particular, creates a singularity on the neuromanifold: the image of the neural networks that are completely unchanged by exchanging two hidden neurons. To see why, consider a toy example. $M = \mathbb{R}^2$, and $(x, y) \sim (x', y')$ iff $x = y', y = x'$, that is, the two points are symmetric across the diagonal line $x = y$. Then $M/\sim$ is a half-plane with a line of singularities: the image of the diagonal line.

Singularities are problematic in practice, where natural gradient descent fails, and stochastic gradient descent often plateaus for a long time. One way to deal with this is by **resolution of singularities**, which essentually means finding a Riemannian manifold **blowup** $\pi : N \to \tilde{M}$, such that π is bijective, smooth, and it is diffeomorphic everywhere except at the preimages of singularities. In other words, it's the bare minimum one must do to smoothen $\tilde{M}$.

Resolution of singularities has been extensively studied in algebraic geometry, and the methods from algebraic geometry, such as Hironaka's theorem, have been applied to machine learning. For more information, refer to the paper [Amari et al., 2018] or the standard reference book [Watanabe, 2009].

References

[Absil et al., 2009] Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds.* Princeton University Press.

[Amari, 2016] Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

[Amari and Nagaoka, 2007] Amari, S.-i. and Nagaoka, H. (2007). *Methods of information geometry*, volume 191. American Mathematical Soc.

[Amari et al., 2018] Amari, S.-i., Ozeki, T., Karakida, R., Yoshida, Y., and Okada, M. (2018). Dynamics of learning in mlp: Natural gradient and singularity revisited. *Neural computation*, 30(1):1–33.

[Caticha, 2015] Caticha, A. (2015). The basics of information geometry. AIP Publishing LLC.

[Costa et al., 2015] Costa, S. I., Santos, S. A., and Strapasson, J. E. (2015). Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, 197:59–69.

[Cover and Thomas, 2006] Cover, T. and Thomas, J. (2006). *Elements of information theory*. Wiley-Interscience.

[Jeffreys, 1946] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

[Lee, 2018] Lee, J. M. (2018). *Introduction to Riemannian Manifolds*. Springer Science & Business Media, 2nd edition.

[Leinster, 2018] Leinster, T. (2018). The Fisher metric will not be deformed. `https://golem.ph.utexas.edu/category/2018/05/the_fisher_metric_will_not_be.html`.

[Nielsen, 2018] Nielsen, F. (2018). An elementary introduction to information geometry. *CoRR*, abs/1808.08271.

[Watanabe, 2009] Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press.