Information Geometry

A Crash Course

Erick Galinkin

January 14, 2022

What is Information Geometry?

Definition

Information Geometry is a method of exploring the world of information by means of modern geometry.

In many cases, using a geometric foundation instead of a measure-theoretic foundation allows for generalization and extension of existing results. In other cases, geometric tools allow us to break open problems that are difficult to solve analytically or algebraically.

Basic Idea

There is a space of models that describe a system we seek to understand. Information geometry allows us to use a notion of "distance" to determine "how far away" our best guess is from reality.

e.g. In statistical inference, using a space of probability distributions to infer what distribution the data we have is sampled from.

Manifolds

A manifold $\mathscr{M}$ is a locally "flat" or Euclidean space. Think of curves in 2-dimensions or surfaces in 3-dimensions... but they may be in really high dimensions.

- A 1 dimensional manifold is a line or a curve - $\mathbb{R}$ – the real line, or S^1, a circle.
- Manifolds of dimension 2 are surfaces - planes like $\mathbb{R}^2$ – the Cartesian plane or a sphere, S^2.
- Higher dimensional hypersurfaces, hyperplanes, etc. also exist – these are exciting objects!

Metrics

Definition

A **metric** d is a function that gives a distance between each pair of elements in a set. Specifically:

$$d : X \times X \to [0, \infty)$$

and $\forall x, y, z \in X$:

1. $d(x, y) = 0 \iff x = y$ Identity of indescernibles
2. $d(x, y) = d(y, x)$ Symmetry
3. $d(x, y) \leq d(x, z) + d(z, y)$ Triangle inequality

A metric induces a topology on a set! (But not all topologies are metrizable.)

Making our Definitions (Topologically) Formal

Definition

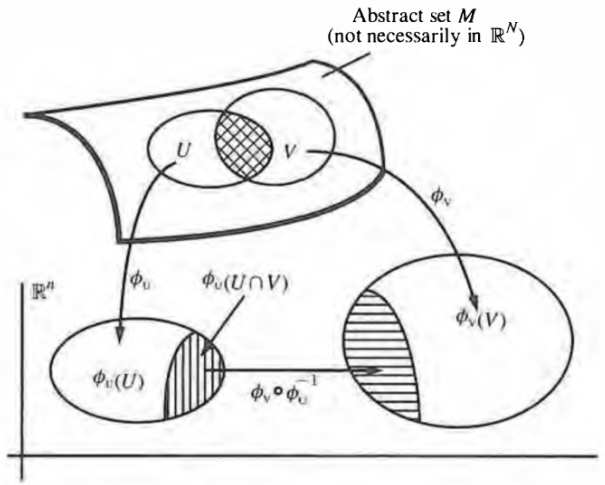
A **manifold** $\mathscr{M}$ is a topological surface such that for all points $x \in \mathscr{M}$, there is a neighborhood U of x and some integer n such that U is homeomorphic (topologically equivalent) to a subset of $\mathbb{R}^n$.

Note that there are many definitions of manifold that are a bit more technical, but are useful in their respective fields. In information geometry, the fact that we can imagine "close" points as if they are on a flat plane is all we really need.

Charts and Atlases

Definition

A **chart** on a topological space $\mathscr{M}$ is an open subset U of $\mathscr{M}$ together with an open embedding $\phi : U \to \mathbb{R}^n$.
An **atlas**, A, is a collection of charts that covers the manifold. That is:
$A = \{\kappa_i : U_i \to \kappa_i(U_i) \subset \mathbb{R}^n\}$ and $\bigcup_i U_i = \mathscr{M}$



We say a manifold is **differentiable** if it has an atlas whose charts are all differentiable.

Parallel Transport

We know that the shortest distance between any two points on a plane is a straight line. What is a straight line on a curved surface?

In geometry, we call the shortest path between two points on a surface a **geodesic**.

To find the geodesic, we can take derivatives on the surface and use the notion of an **affine connection** – an object that connects tangent spaces – by way of **parallel transport**. Essentially, parallel transport is when we move the tangent vectors along a curve and the geodesic is the "straight line" induced by that movement.

Divergence

A **divergence** is a statistical "distance": how "far away" one probability distribution is from another.

Divergence and Metrics

A divergence is not necessarily a metric! Metrics are symmetric, most divergences are asymmetric.
That is, for a metric μ, $\mu(x, y) = \mu(y, x)$ but given a divergence D, $D(X, Y) \neq D(Y, X)$ in general!

Depending on your manifold, you may need to consider different divergences:

- Euclidean divergence (If your coordinate system is orthonormal!)
- KL divergence $D_{KL}[p(x)||q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx$
- Bregman divergence for a convex function ψ:
  $D_\psi[\xi||\xi_0] = \psi(\xi) = \phi(\xi) - \nabla\psi(\xi_0) \cdot (\xi - \xi_0)$

Statistical manifolds

Definition

The **probability density function** of a Gaussian random variable X is

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

Where μ is the mean and σ^2 is the variance.

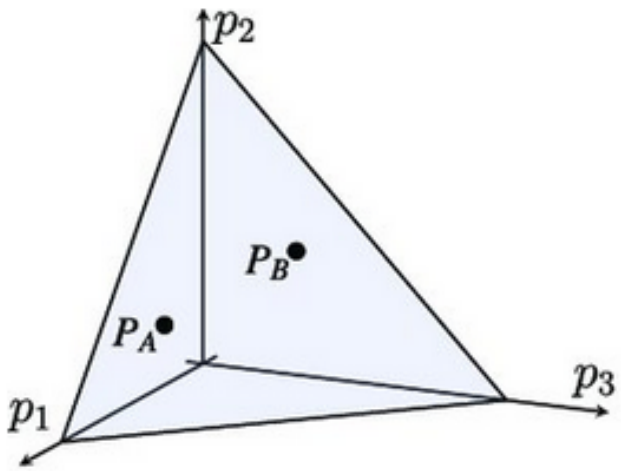
One good example of a **statistical manifold** is the set of all Gaussian distributions. Each point is a probability density function and $\xi = (\mu, \sigma)$, $\sigma > 0$ is the coordinate system on that manifold.

More Statistical Manifolds

Let X be a discrete random variable taking values in $\{0, 1, ..., n\}$. Then

$$p_i = Pr\{x = i\}, \quad i = 0, 1, ..., n$$

so $p(x)$ can be represented by the vector $\vec{p} = (p_0, p_1, ..., p_n)$, just as if we were in $\mathbb{R}^n$!



Applications

Lots of applications for Information Geometry exist:

- Game Theory
- Statistical Learning
- Biology
- Finance
- Physics

I don't know about most of these things.

Motivation: Statistical Learning

Loads of work by Shun-Ichi Amari [Ama12] and Sumio Watanabe [Wat09] demonstrate the efficacy of information geometry in statistical (machine) learning. Much of statistical learning theory depends on **maximum likelihood estimation**, which requires finding parameters for a statistical model which are most probable for the observed distribution. (we'll talk about this later)

Note!

In most cases, we do not know the real distribution, so we need to use a parametrized model to approximate it.

We'll call this parameter θ.

Determining Distribution Parameters

Let $f(x;\theta)$ denote an indexed family of probability densities. An estimator for θ for sample size n is a function $T : \chi^n \to \Theta$. The estimator should approximate of the parameter and so we call $T - \theta$ the error of the estimator and define the bias of the estimator as the expected value of the error:

$$E[T(x_1, x_2, ..., x_n) - \theta]$$

We define the score:

Definition

The **score** of a random variable $X \sim f(x;\theta)$ is defined as:

$$V = \frac{\partial}{\partial\theta} \ln f(X;\theta) = \frac{\frac{\partial}{\partial\theta} f(X;\theta)}{f(X;\theta)}$$

Fisher Information

Definition

Fisher Information is a way of measuring the information that an observable random variable X carries about an unknown parameter θ of a distribution that models X. We define the Fisher information $I(\theta)$ as the variance of the score:

$$J(\theta) = E_\theta \left[\frac{\partial}{\partial \theta} \ln f(X; \theta) \right]^2$$

Definition

The **Cramer-Rao bound** of an unbiased estimator T is a lower bound on the variance of that estimator which is equal to the inverse of the Fisher information.

$$var(T) \geq \frac{1}{J(\theta)}$$

More Statistical Learning

Definition

Let $q(x)$ be the true distribution, $p(x|\theta)$ be a statistical model parametrized by θ, and $\phi(x)$ be the a priori probability density function. Then the **negative log likelihood** is given by:

$$R_n(w) = -\sum_{i=1}^{n} \log p(X_i|\theta) - a_n \log \phi(\theta)$$

where $\{a_n\}$ is a sequence of nonnegative real values.

Then we can seek to define a statistical estimation method to optimize θ in a way that minimizes $R_n(\theta)$. But we have a problem. Sometimes, our model is singular.

Singular Models

In cases where our model is singular, our log likelihood can get nasty.

Definition

A model is **regular** if the Fisher information matrix $J(\theta)$ is positive definite for all $\theta \in \Theta$. A model is **singular** if it is not regular.

So what do we do? Watanabe shows us that we need only find a change of variables that makes the KL-divergence locally monomial on the manifold. This follows from (and always exists due to) the resolution of singularities theorems proved by Heisuke Hironaka .

Theorem

The Fundamental Theorem of Singular Learning (Watanabe) Given mild conditions on a model $\mathcal{M}$, there exists a change of variables $\rho : \mathscr{M} \to \Omega$ such that

$$D_{KL}(\rho(\mu)) = \mu^{2\kappa} - \frac{1}{\sqrt{N}}\mu^{\kappa}\xi_N(\mu)$$

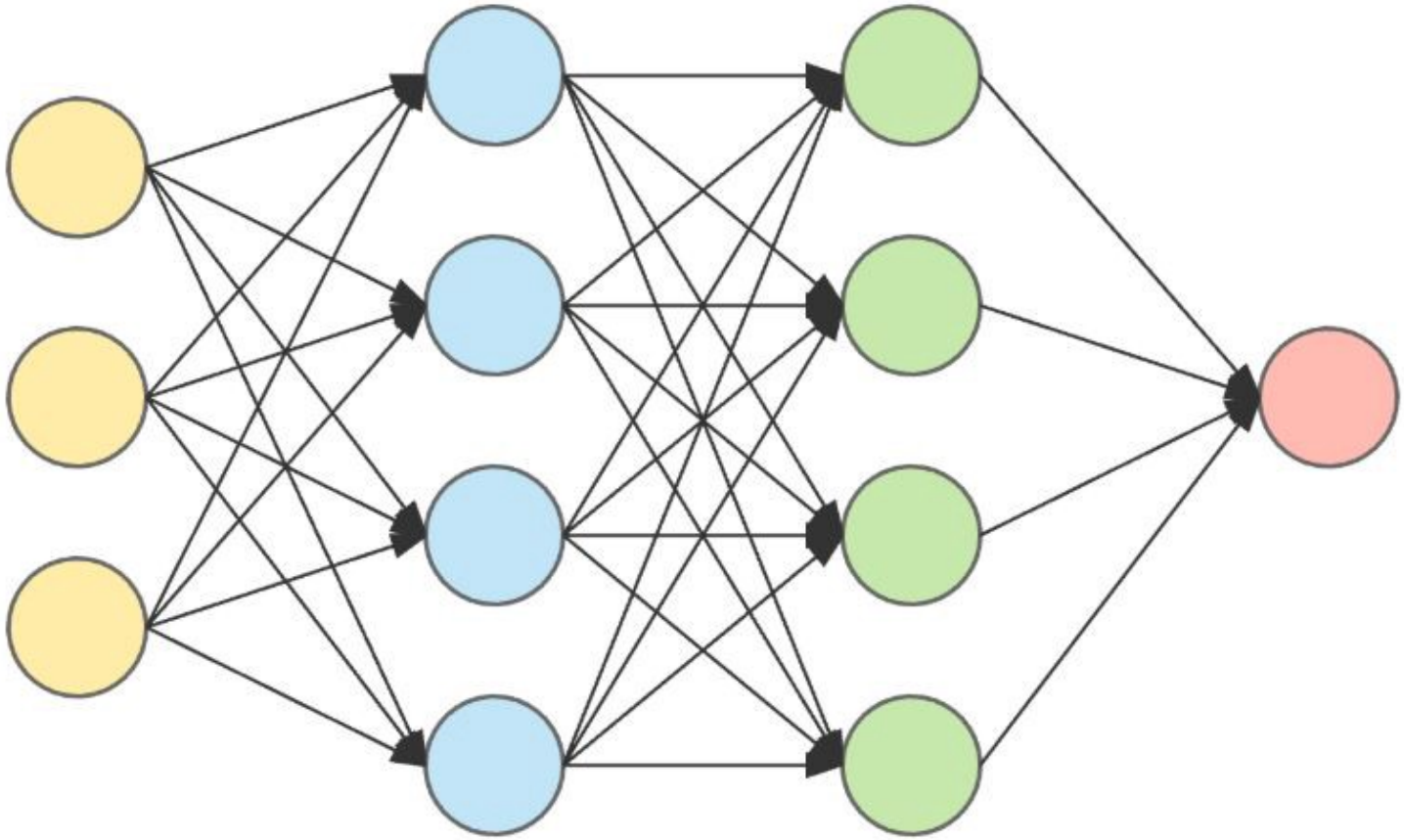
where $\xi_N(\mu)$ converges to a Gaussian process on $\mathscr{M}$

Neural Networks - a quick background

- Neural networks are a really exciting machine learning model because it's very good at lots of tasks.
- Each individual neuron[1] consists of an affine transformation according to a weight w and bias β and a nonlinearity σ, so the output of each neuron is $\sigma(w^\intercal x + \beta)$.
- Two common activation functions are $ReLU = \max\{0, w^\intercal x + \beta\}$ and sigmoid $S(x) = \frac{1}{1+e^{-(w^\intercal x+\beta)}}$

[1] Kind of - there are lots of complicated inner workings in neurons

Neural Networks



Geometry of Neural Networks

Definition

Tropical geometry is a branch of algebraic geometry where we study geometric structures over the **max-plus** (sometimes min-plus) or "Tropical" semiring $\mathbb{T} = \{\mathbb{R} \cup \{-\infty\}, \bigoplus, \bigotimes\}$:
Addition: $x \bigoplus y = \max\{x, y\}, \quad x \bigoplus -\infty = x$
Multiplication: $x \bigotimes y = x + y, \quad x \bigotimes 0 = x$

ReLU networks have deep connections to Tropical geometry! This is because the $\bigoplus$ operator is the same as the ReLU function with one operand set to 0. One interesting result from Zhang *et al.* [ZNL18]: All feedforward *ReLU* networks are equivalent to tropical rational maps.

Neuromanifolds

Definition

A **neuromanifold** is a (Riemannian) manifold associated with a neural network such that $y = f_\theta(x)$ is the input-output mapping of a neural network with input distribution $p_X(x)$ and output distribution $p_Y(y;\theta)$ where θ is a vector parameter consisting of all the network's weights and biases. The statistical manifold associated with the network is $\mathcal{S} = \{p(x;y;\theta)\}$ and the metric $g = J(\theta)$ defined on the manifold is the Fisher information.

Because the Fisher Information is the negative Hessian of log likelihood, we can use this to enhance gradient descent.

Natural Gradient Descent

In a sort of hand-wavy way, natural gradient descent is the ability to optimize our loss $\mathcal{L}(\theta)$ with respect to the inputs, and outputs of a network by using the Fisher Information F to optimize our parameter θ.
This matrix can be hard to compute for large networks, but approximation approaches do exist for computing the natural gradient [PB13], which is particularly nice for online learning! This can offer a lot of the benefits of second-order optimization without having to compute the Hessian (which is hard).

Game Theory

We paraphrase Jurgen Jost's [JBOW16] geometric restatement of Nash's theorem:

Theorem

Given the $(m_i, 1)$-dimensional simplex of mixed strategies, $\Sigma_i := \{p_i \in \mathbb{R}_+^{m_i} : \sum_{\alpha=1}^{m_i} p_i^\alpha = 1\}$, let $\rho_i : \Sigma_{-i} \rightrightarrows \Sigma_i$. If ρ_i were single-valued, then it could be depicted as a graph in Σ_{-i} with values in Σ_i. Correspondingly, ρ_{-i} would be a graph over Σ_i with values in Σ_{-i}. It is geometrically clear that these two graphs must intersect somewhere and that point is the Nash equilibrium.

This simple statement of Nash's theorem makes clear the power of geometric language!

Works cited I

- Shun-ichi Amari, *Differential-geometrical methods in statistics*, vol. 28, Springer Science & Business Media, 2012.
- Jürgen Jost, Nils Bertschinger, Eckehard Olbrich, and David Wolpert, *Information geometry and game theory*, Information Geometry and its Applications IV, Springer, 2016, pp. 19–46.
- Razvan Pascanu and Yoshua Bengio, *Revisiting natural gradient for deep networks*, arXiv preprint arXiv:1301.3584 (2013).
- Sumio Watanabe, *Algebraic geometry and statistical learning theory*, no. 25, Cambridge university press, 2009.
- Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim, *Tropical geometry of deep neural networks*, International Conference on Machine Learning, PMLR, 2018, pp. 5824–5832.