# A PRIMER ON INFORMATION GEOMETRY[*]

RAHUL BHADANI[†]

**Abstract.** Information Geometry is a non-traditional way of studying probability measures using statistical manifolds and geometry. It encompasses various established fields such as information theory, statistics, machine learning, and topology with applications in signal processing, pattern-matching, and quantum information science. While there are plenty of articles written on this topic, in this article, I touch on some basics needed for readers to kick-start their journey in the field of information geometry.

**Key words.** Information Geometry, Statistics, Manifold, Differential Geometry, Data Science

**MSC codes.** 62B10, 31C12, 46N30

## 1. Introduction.

Recent decades saw a boom in information technology followed by a craze for machine learning, artificial intelligence, and data science [1]. The cause has inspired many to dig deeper into some mystics of mathematics [2, 3] and information geometry is one of them. Information geometry can be used in statistical manifold learning [4] that has recently proven to be a useful tool for unsupervised learning on the high-dimensional dataset. It can also calculate the distance between two probability measures with applications in pattern matching, constructing alternative loss functions for the training of a neural network, belief propagation network, and optimization problems [5].

Information geometry is a mathematical tool for exploring the world of information using geometry. Information Geometry is also called Fisherian Geometry for the reason that will be obvious in the later part of the article. Information geometry is the study of decision-making using geometry that may also include pattern-matching, modeling-fitting, etc. But why the geometric approach? Geometry allows studying invariance in a coordinate-free framework, provides a tool to think intuitively, and allows us to study equivariance. As an example, a centroid of a triangle is equivariant under affine transformation.

Let's discuss some fundamentals to understand what is there in Information Geometry.

## 2. Fundamentals.

To understand differential geometry, and hence information geometry, we need to see what a manifold is.

### 2.1. Manifold.

An n-dimensional manifold is a most general mathematical space with limits, continuity, and correctness as well as allows the existence of the continuous inverse function with n-dimensional Euclidean spaces. Manifolds locally resemble

[*] Submitted to the editors DATE.

[†] Vanderbilt University, Nashville, TN, USA (rahul.bhadani@vanderbilt.edu).

Euclidean space but they may not be Euclidean space. Essentially, a manifold is a generalization of Euclidean space.

**2.2. Topological Spaces.** Consider a space $X$ is defined by a set of points $x \in X$ and a set of subsets of $X$ called neighborhoods $\mathcal{N}(x)$ for each point. We have the followings properties:

1. If $U$ is a neighborhood of $x$, and $x \in U$, and $V \subset X$ and $U \subset V$, then $V$ is also the neighborhood of $x$.
2. The intersection of two neighborhoods of $x$ is also a neighborhood of $x$.
3. Any neighborhood $U$ of $x$ includes a neighborhood $V$ of $x$, such that $U$ is a neighborhood of all points in $V$.

Any space $X$ satisfying the above properties can be called **topological space**.

**2.3. Homeomorphism.** Considering $f : X \to Y$ to be a function between two topological spaces, then $X$ and $Y$ are homeomorphic if $f$ is continuous, bijective, and the inverse of $f$ is also continuous. Consider a manifold $\mathcal{M}$, and for all points $x \in \mathcal{M}$, if $U$ is a neighborhood of $x$, and for an integer $n$ such that $U$ is homeomorphic to $\mathbb{R}^n$, then small $n$ is the dimension of the manifold.

**2.4. Charts.** Homeomorphism denoted by the function $\kappa : U \to \kappa(U)$ is called a chart where $U$ may be an open subset of $\mathcal{M}$. There can be many ways to construct a chart to define $\mathcal{M}$. A collection of such charts are called an **atlas**. The idea is presented in Figure 1. Mathematically, an atlas would be defined by Equation (2.1). A specific example of a chart is a coordinate system which can be a function that maps points on a manifold.

$$A = \{\kappa_i = U_i \to \kappa_i(U_i) \subset \mathbb{R}^n, U_i \in \mathcal{M}\}, i = 1, 2, 3, ... \tag{2.1}$$

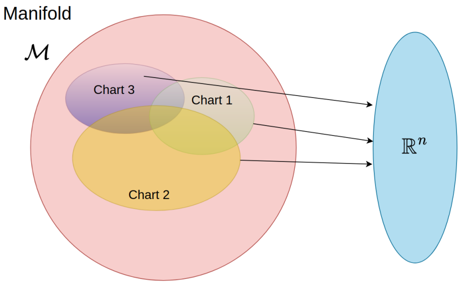

**Fig. 1:** An illustration of a manifold and charts

At this point, it is easy to define a **differentiable manifold**: a manifold whose transition maps (or functions) are infinitely differentiable.

That's all about manifolds in abstract mathematics. What about its understanding of statistics and data science? Remember, we deal with probabilities in statistics. This leads to the notion of statistical manifolds. In a statistical manifold, every point $p \in \mathcal{M}$ corresponds to a probability distribution over a domain $\mathcal{X}$. One can think of this with a specific example of a manifold formed by a family of normal distribution.

**2.5. Vectors and Tangents on a Manifold: how to define them on curved spaces.** In ordinary geometry, vectors are straight lines connecting two points but in curved spaces, this may not be true. Vectors on curved spaces are defined as tangents to a curve at a particular point on the manifold. If $u$ is a parameter that varies along the curve, then a curve may be defined as $x(u)$, often dropping the $u$ part and simply writing $x$. The vector in curved spaces becomes

$$X = \frac{\partial x}{\partial u}\bigg|_{u=0} \tag{2.2}$$

which is defined at point $p$ locally where $u = 0$. Note that the vector itself doesn't live on the manifold but it has Euclidean notion. Like charts, there can be many possible tangents at the point $p$. Yeah, this is mind-boggling if you are merely considering a 2D plane, but even in a 3D like on a sphere, there could be multiple tangent lines at a point — then we would talk about a tangent plane on a point of the sphere (Figure 2). Analogously, we can talk about a tangent space at a point $p$ for a manifold.

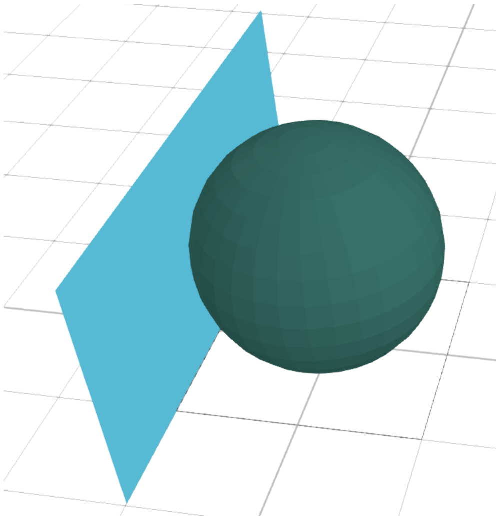

**Fig. 2:** A sphere with a tangent plane. A tangent plane consists of an infinite number of tangents at a point on a sphere.

If we go from one chart to another, it is the same as the coordinate transformation from, say, cartesian coordinates to polar coordinates. Suppose there is a transformation function $\phi$ to transform $x$ from one chart to another, then it can be written as $x' = \phi(x)$.

**2.6. Dual Space.** The dual space $V^*$ of a vector space $V$ is the space that contains all the linear functionals of $V$, i. e. all maps $T : V \mapsto F$, where $F$ is the field that $V$ is the vector space of. Therefore, the dual space contains all linear mappings from $V$ to $F$.

To understand a dual space, visualize a 2-dimensional real vector space. Let's consider a function $f_1$ that takes any vector and returns its $x$-coordinate. Take another function $f_2$ that takes any vector and returns its $y$-coordinate. Moving forward, imagine these two functions as vectors. Consider them as basis vectors in an arbitrary vector space. We can add them together, say as $f_1 + f_2$, a function that takes any vector and returns the sum of the $x$-coordinate and the $y$-coordinate. We can multiply them by numbers – $5 \cdot f_1$ as a function that takes any vector and returns the $x$ coordinate times 5. We can form a linear combination too – for example $4.5 \cdot f_1 - 10 \cdot f2$ as a function that takes any vector and returns the number 4.5 multiplied by the $x$ coordinate minus 10 multiplied by the y coordinate. This is what dual space does.

**2.7. Tensors.** Tensors in the case of the manifold are the most general. They can be considered mathematical multi-linear beasts that eat vectors from tangent spaces and their dual space and spit out real numbers. The total number of vectors from the tangent space and its dual space that are fed to the tensor is called the **rank of the tensor**. The number of these that come from the dual space gives what is called the **contravariant rank** and the number which comes from the tangent space is called the **covariant rank**. In essence, manifolds are geometric constructs and tensors are corresponding algebraic constructs.

**2.8. Metric.** A metric is a tensor field that induces an inner product on the tangent space at each point on the manifold. Any tensor field of the covariant rank two can be used to define a metric. Some sources call it the **Riemannian Metric**.

Now, we look at something useful in Information Geometry after a long and convoluted list of terminologies.

**3. Information Geometry.** Information geometry is a branch of mathematics that intersects statistics and differential geometry focusing on the study of probability distributions from a geometric perspective. Let's look at one of the most fundamental concepts in information geometry – information metric.

**3.1. Fisher Information Metric.** If we wish to find a suitable metric tensor at a point $\theta^*$ where $\theta^*$ corresponds to one of a family of distribution $p(x|\theta)$, then we need to look at a notion of distance between $p(x|\theta)$ and its infinitesimal perturbation $p(x|\theta + d\theta^*)$. The relative difference as a notion of distance is given by Equation (3.1).

$$\Delta = \frac{p(x|\theta + d\theta^*) - p(x|\theta)}{p(x|\theta)} = \frac{\frac{\partial p(x|\theta)}{\partial \theta^*} d\theta^*}{p(x|\theta)} = \frac{\partial \log p(x|\theta)}{\partial \theta^*} d\theta^* \tag{3.1}$$

Of course, relative distance depends on the random variable $x$. If you do the math correctly, then the expectation of $\Delta$, i.e., $\mathbb{E}(\Delta) = 0$. What about the variance? It turns out that the variance is non-zero. We could define $dl^2 = \mathbb{E}[\Delta^2]$. From the first principle, the length of an infinitesimal displacement between $\theta^*$ and $\theta^{**}$ for a metric $\mathcal{F}$ is given by $dl^2 = \mathcal{F} d\theta^* d\theta^{**}$. Solving for $dl^2 = \mathbb{E}[\Delta^2] = \mathcal{F} d\theta^* d\theta^{**}$ gives

$$\mathcal{F} = \int dx p(x|\theta) \frac{\partial \log p(x|\theta)}{\partial \theta^*} \frac{\partial \log p(x|\theta)}{\partial \theta^{**}} \tag{3.2}$$

which is what we call **Fisher Information Metric (FIM)**. FIM measures how much information an observation of the random variable $X$ carries about the parameter $\theta$ on average if $x \sim p(x|\theta)$. There is another way to arrive at Equation (3.2) up to a factor of $\frac{1}{2}$ using relative entropy. The burgeoning of Quantum Information Science has applications of the Fisher Information Metric [6, 7, 8].

I will conclude the article with a final thought that Fisher Information Matrix $\mathcal{I}$ which is a matrix version of $\mathcal{F}$ when we are dealing with multiple parameters, can be used in optimization similar to the gradient descent with update rule as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathcal{I}^{-1} \nabla J(\boldsymbol{\theta}) \tag{3.3}$$

where $\eta$ is the learning parameter, $\nabla J$ is the divergence of the scalar field $J$.

## 4. Conclusion.

In this article, I provided a brief overview of information geometry and related terms. A considerable amount of work in this discussion is omitted for clarity, especially focusing on beginners and I encourage readers to go through the literature for in-depth discussion.

## REFERENCES

[1] R. Agarwal and V. Dhar, "Big data, data science, and analytics: The opportunity and challenge for is research," pp. 443–448, 2014.

[2] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning.* Cambridge University Press, 2020.

[3] T. K. Dey and Y. Wang, *Computational topology for data analysis.* Cambridge University Press, 2022.

[4] K. Sun and S. Marchand-Maillet, "An information geometry of statistical manifold learning," in *International Conference on Machine Learning.* PMLR, 2014, pp. 1–9.

[5] S.-i. Amari, "Information geometry in optimization, machine learning and statistical inference," *Frontiers of Electrical and Electronic Engineering in China*, vol. 5, no. 3, pp. 241–260, 2010.

[6] W. A. Miller, "Quantum information geometry in the space of measurements," in *Quantum Information Science, Sensing, and Computation X*, vol. 10660. SPIE, 2018, pp. 102–117.

[7] L. Banchi, P. Giorda, and P. Zanardi, "Quantum information-geometry of dissipative quantum phase transitions," *Physical Review E*, vol. 89, no. 2, p. 022102, 2014.

[8] M. Hayashi, "Quantum information geometry and quantum estimation," in *Quantum Information Theory.* Springer, 2017, pp. 253–322.