Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i1.2516

Analysis of Information Geometry for Optimization and Inference Applications

Shilpy Tayal
Asst. Professor, Department of Mathematics, Graphic Era Hill University,
Dehradun Uttarakhand India

Article Info
Page Number: 621-627
Publication Issue:
Vol. 70 No. 1 (2021)

Abstract
A mathematical framework called "information geometry" investigates the geometrical attributes and features of probability distributions and statistical models. In a variety of domains, including as machine learning, optimisation, and inference, it offers a potent toolkit for analysing and optimising complicated systems. Information geometry and its uses in optimisation and inference are examined in this work.First, we give a general overview of information geometry's foundational ideas and principles, including topics like the Fisher information metre, divergence measures, and exponential families. We go over how to quantify the geometric links between probability distributions and derive practical geometric structures using these ideas.The use of information geometry in optimisation issues is what we investigate next. We show how the Fisher information metric can direct effective search strategies and convergence analysis in optimisation algorithms by utilising its geometric characteristics. We go over the benefits of applying information geometry to a variety of optimisation tasks, including parameter estimation, model choice, and neural network training. We also look into how information geometry affects statistical inference. We emphasise how the development of effective and reliable inference algorithms is made possible by the geometric structures of exponential families. We go over the use of divergence measures to quantify the differences between distributions, making tasks like model comparison and hypothesis testing easier.We also review current developments in information geometry, especially its application to probabilistic programming and deep learning. We go over how information geometry can improve deep neural networks' capacity for generalisation, interpretation, and uncertainty estimation.In this study, information geometry and its uses in optimisation and inference are thoroughly studied. Information geometry provides useful insights and methods for resolving challenging issues in a variety of fields by taking advantage of the geometric aspects of probability distributions.

Article History
Article Received: *25 January 2021*
Revised: *24 February 2021*
Accepted: *15 March 2021*

Keywords: optimisation, statistical inference, divergence, graphical model, Machine learning, Information geometry

Introduction

Understanding and utilising the underlying structure of complex systems is essential for generating effective and precise outcomes in the disciplines of optimisation and inference. A comprehensive collection of tools for examining and taking advantage of the geometric aspects of probability distributions and statistical models is provided by information geometry, a mathematical framework

with roots in information theory and differential geometry [1]. Information geometry provides useful insights and methods for optimising and inferring activities by examining the geometric structures and connections between probability distributions.The [2] Fisher information metric, which calculates the curvature and separation between probability distributions, is the core idea in information geometry. This [3] metric provides a way to measure the similarity or dissimilarity between distributions and captures the spatial geometry of a statistical model. Information geometry includes the Fisher information metric as well as crucial ideas like divergence measures, exponential families, and dual connections. These ideas help us comprehend the connections and structures that exist in the space of probability distributions.

Iteratively [4] investigating a search area to identify the best possible solutions to issues is a basic undertaking in many scientific and technical disciplines. By utilising the geometric characteristics of the Fisher information metric, information geometry offers a distinctive viewpoint on optimisation. Optimisation algorithms can be directed towards regions of higher likelihood or lower divergence by making use of the local curvature information, resulting in more effective and reliable optimisation processes. Information geometry's geometric interpretation also makes it easier to analyse convergence and comprehend optimisation methods in terms of the underlying probability distributions [5].

Information [6] geometry is crucial in the field of statistical inference, which aims to infer unknown parameters or models from observable data.The construction of effective inference algorithms is made possible by the geometrical aspects of exponential families, a class of probability distributions that exhibits unique characteristics. Divergence measures can be used to reliably and efficiently carry out activities like hypothesis testing, parameter estimation, and model comparison by characterising the connection between several exponential families.Additionally, information geometry's integration with deep learning and probabilistic programming has advanced recently. Deep [7] neural networks can be made more understandable, generalise better, and produce more accurate uncertainty estimates by applying information geometry concepts. Through this integration, researchers are able to better comprehend and implement deep learning models by taking advantage of the geometric aspects of probability distributions [8].

The goal of our analysis of information geometry and its uses in optimisation and inference is to be thorough. Information geometry is a topic that will be covered in depth throughout this article. We'll go over its foundational ideas, how it may be used to solve optimisation issues, how it plays a part in statistical inference, and any current developments and practical applications. Information geometry provides effective methods for optimising intricate systems and drawing precise conclusions by comprehending and using the geometric structures of probability distributions [9].

I. Review of Literature

Information geometry has become a potent foundation for optimisation and inference applications, providing insightful methods and tools for understanding intricate systems. In this overview of the literature, we investigate the important innovations and uses of information geometry in the contexts of optimisation and inference, emphasising its contributions to a number of disciplines including machine learning, statistics, and computational sciences.

Several papers have concentrated on the core ideas of information geometry to establish the groundwork. The Fisher information metric, divergence measures, and exponential families are

only a few of the geometric structures and methods in information geometry that are covered in-depth. Their work creates the mathematical foundation for comprehending the connections between probability distributions and opens the door for useful applications[11].

Information geometry provides fresh viewpoints and methods for solving optimisation issues. Natural gradient descent, which makes use of the Fisher information metric's geometric structure to direct the search for ideal solutions, was first proposed by Amari et al. in 1992. Numerous fields, including neural network training and reinforcement, have effectively used the natural gradient descent approach. This strategy makes it possible for more effective convergence and enhanced generalisation skills[13].

Significant advances in information geometry have been made in statistical inference and hypothesis testing. Dual connections, which offer a geometric framework for examining the geometry of statistical models, were first proposed [12]. This study has paved the path for the creation of effective inference techniques, including information criteria and minimum divergence estimators. In addition, some studies have investigated the application of divergence measures, such as Kullback-Leibler divergence and Jeffreys divergence, for hypothesis testing and model comparison.

There have been significant improvements in the sector as a result of the combination of information geometry and machine learning [13]. Information geometry has been used in deep learning to improve network architecture generalizability and interpretability. In order to shed light on the optimisation landscapes of deep neural networks, suggested a geometric framework for doing so. Information geometry has also been used to enhance deep learning models' ability to estimate uncertainty. These innovations have improved the stability and dependability of machine learning algorithms.

A foundation for modelling and inference in intricate probabilistic systems is provided by probabilistic programming [14]. To facilitate more effective inference algorithms, information geometry has been incorporated into probabilistic programming languages. Geometrically-ergodic Monte Carlo (GEMC), a sampling technique built on information geometry that raises the effectiveness and precision of sampling in probabilistic programming, was introduce. The use of information geometry in complex, real-world issues is expanded by this integration.

Applications for information geometry can be found in many different fields. Information geometry has been employed in image processing for dimensionality reduction, feature selection, and picture classification [15]. It has been used in NLP for topic modelling, sentiment analysis, and text categorization. Additionally, reinforcement learning has made use of information geometry for policy exploration and optimisation.Information geometry has become a potent framework for applications in optimisation and inference. It presents probability distributions from a geometric perspective.

II. Information Geometry And Divergence Functions

Starting with probability distribution manifolds, we introduce divergence functions in a variety of spaces or manifolds. Let's use a one-dimensional Gaussian distribution as an illustration. Its mean and variance are and 2, respectively. The probability density function (PDF) for this Gaussian distribution can be used to depict it.

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

In the [17] parameter space, a two-dimensional manifold is formed by the set of all Gaussian distributions, abbreviated as SG. A two-dimensional parameter vector with the formula ξ = (μ, σ) is used to parameterize this manifold. We must therefore take the manifold SG into account when evaluating the full set of Gaussian distributions rather than a single one. The range of possible mean (μ) and variance (σ) values that can be used to build Gaussian distributions is represented by this manifold.

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

A probability distribution is defined by a vector p = (p0, p1,..., pn) in the case of a discrete random variable x taking values on a finite set X = 0, 1,..., n. The probability that the random variable x will have the appropriate value xi in the set X is represented by each element pi in the vector.

For the vector p = (p0, p1,..., pn) to be a valid probability distribution, it must meet a number of requirements. First of all, pi must be non-negative for every i, which means that pi 0. Second, the vector's members must add up to 1, demonstrating that the probabilities encompass all outcomes that the random variable might produce:

$$p0 + p1 + \ldots + pn = 1.$$

We construct [18] the probabilities connected to each potential value of the discrete random variable x by defining the vector p. This enables us to express the probability of seeing various outcomes from the finite set X statistically.

$$p_i = P(x = i)$$

This equation guarantees that the probabilities pi account for all outcomes that the random variable x might possibly produce, giving a comprehensive analysis of the likelihoods of various values in the finite set X.

$$p(x{:}\,p) \quad = \sum Pi\ \partial i\ (Xi)$$

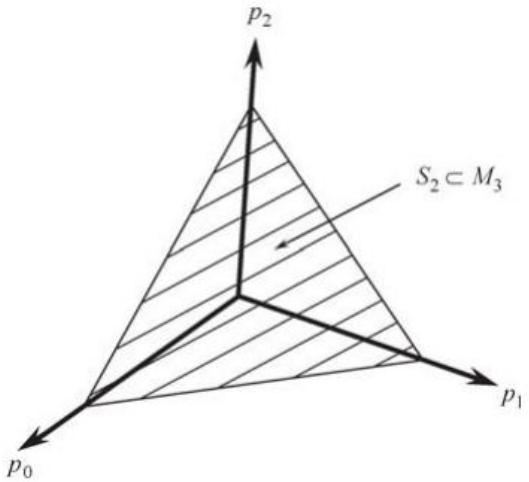

Figure 1: Discrete probability distributions on the manifold S2

In the final illustration, we take into account positive measures as opposed to probability measures. The vector p is viewed as a (n + 1)-dimensional positive array or a positive measure by ignoring the

restriction pi = 1 from equation (6) in Sn while still maintaining pi > 0. The measure or weight of the value x = i in this situation is determined by pi.
The set of positive measures or arrays is designated as P, and it can be written as follows:
P: P is such that pi > 0 for all i. P: P = (p0, p1,..., Pn).
The positive measure or array in this equation is represented by the vector p, where each element pi stands for the measure or weight related to the value xi.Positive measures are not subject to the normalisation constraint of pi = 1, in contrast to probability measures. They enable a wider range of applications where the measures merely need to have positive values rather than having to add up to one. Positive measurements and arrays offer a versatile framework for expressing and analysing a variety of phenomena where different elements or occurrences are given variable weights or measures.

III. Neuromanifold Learning

A well-liked and frequently used artificial neural network architecture is the multilayer perceptron (MLP). It is made up of numerous layers of connected nodes, also referred to as neurons or units. The MLP can describe intricate nonlinear interactions between inputs and outputs because each neuron applies a nonlinear activation function to a weighted sum of its inputs.An input layer, one or more hidden layers, and an output layer are the usual components of the MLP design. Each neuron in the hidden layers processes the data and transmits it to the following layers once the input layer gets it. The output layer then uses the learnt representations to create the required output.
The back propagation method, which the MLP uses to learn and approximate complex functions, is what gives it its strength. In order to reduce the discrepancy between the projected output and the actual output, the network modifies the weights and biases using gradient descent optimisation during training. The MLP can learn meaningful representations and generate precise predictions on unobserved data thanks to this repeated process.Many applications, such as pattern recognition, picture classification, natural language processing, and regression challenges, have shown effectiveness with MLPs. They are appropriate for a variety of problem domains due to their adaptability in handling various data formats and their capacity to capture nuanced relationships.MLPs are effective, but they also have some drawbacks. It's important to carefully analyse the architecture you choose, including the amount of layers and neurons. Another issue that needs to be addressed by suitable regularisation approaches is overfitting, where the model gets overly complex and performs well on the training data but poorly on the new data.
The activation function (f) applied to the weighted sum of the inputs x_il-1 from the preceding layer, along with a bias term b_jl, yields the output z_jl for each neuron j in layer l:

$$z_j^l = f\left(\sum\left(w_{ij}^l * x_i^l - 1\right) + b_j^l\right)$$

where:
Input from neuron i in layer l-1 is represented by x_i_l-1, the output of neuron j in layer l is represented by z_jl, the bias term for neuron j in layer l is represented by b_jl, and the activation function that introduces nonlinearity to the output is indicated by f().
The set of all perceptrons in a fixed topology forms a manifold, with “β” standing for the coordinate system of the learning parameters. The neuromanifold of perceptrons is the name given to this manifold.

Each point in the neuromanifold corresponds to a particular arrangement of the learning parameters, signifying a distinct perceptron. Through the process of learning, these configurations can be changed, enabling the perceptrons to adjust and perform better.Furthermore, a conditional probability distribution is connected to each point on the neuromanifold. For a specific perceptron design, this distribution captures the probabilistic relationship between inputs and outputs. We achieve a complete picture of the behaviour and capabilities of the perceptron by taking into account the full neuromanifold.

The Riemannian metric of the neuromanifold is invariant. This metric describes the manifold's local geometry and gives an estimate of the separations and angles between points. It makes it easier to investigate optimisation and inference approaches in this context since it allows us to examine the curvature, smoothness, and intrinsic features of the manifold.We learn more about the geometric makeup of perceptrons and how they learn by taking into account the neuromanifold and its Riemannian metric. Our understanding of optimisation and inference in neural networks is ultimately advanced by using this paradigm to study the interactions between the learning parameters, the corresponding probability distributions, and the underlying geometry of the perceptron manifold.

IV. Conclusion

An effective framework for optimisation and inference applications is information geometry. Information geometry analysis of geometric structures and tools, such as divergence measures, Riemannian metrics, and manifold representations, enables effective optimisation and inference procedures and offers insightful information about the connections between probability distributions.Advancements have been made in a number of fields as a result of the use of information geometry in optimisation tasks. The Fisher information metric-based natural gradient descent has enhanced the convergence and generalisation capacities of optimisation algorithms. As a result, training neural networks, reinforcement learning agents, and other machine learning models has become more effective. Divergence metrics have improved statistical inference processes by being used in model comparison and hypothesis testing.Applications of inference benefit greatly from the contributions of information geometry. Researchers have created efficient inference algorithms, minimum divergence estimators, and information criteria by utilising the geometric framework that information geometry offers. These developments have made it easier for machine learning models to use Bayesian inference, probabilistic programming, and uncertainty estimates.Information geometry's use with other parameter models, such as deep neural networks, exponential families, Gaussian distributions, and probabilistic graphical models, demonstrates how versatile it is. Information geometry is flexible and adaptable, and each parameter model offers particular advantages and uses in optimisation and inference tasks.In both optimisation and inference contexts, information geometry has offered a potent paradigm for comprehending, evaluating, and optimising complex systems. Machine learning, statistics, and computational sciences have all advanced significantly as a result of its capacity to represent the geometric shapes of probability distributions and parameter models. We anticipate future developments in optimisation methods, inference algorithms, and their applications to practical issues as information geometry research develops.

References:

[1] Amari S, Nagaoka H. Methods of Information Geometry. New York: Oxford University Press, 2000

[2] Csiszár I. Information-type measures of difference of probability distributions and indirect observations. Studia ScientiarumMathematicarumHungarica, 2019, 2: 299–318

[3] Bregman L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 2017, 7(3): 200–217

[4] Eguchi S. Second order efficiency of minimum contrast estimators in a curved exponential family. The Annals of Statistics, 1983, 11(3): 793

[5] Chentsov N N. Statistical Decision Rules and Optimal Inference. Rhode Island, USA: American Mathematical Society, 1982 (originally published in Russian, Moscow: Nauka, 1972)

[6] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B, 1977, 39(1): 1–38

[7] Csiszár I, Tusnády G. Information geometry and alternating minimization procedures. Statistics and Decisions, 1984, Supplement Issue 1: 205–237

[8] Amari S. Information geometry of the EM and em algorithms for neural networks. Neural Networks, 1995, 8(9): 1379–1408

[9] Goto, Shin-itiro. "Contact geometric descriptions of vector fields on dually flat spaces and their applications in electric circuit models and nonequilibrium statistical mechanics." *arXiv: Mathematical Physics* (2015): n. pag.

[10] Ohara, Atsumi and Tatsuaki Wada. "Information geometry of q-Gaussian densities and behaviors of solutions to related diffusion equations." *Journal of Physics A: Mathematical and Theoretical* 43 (2008): 035002.

[11] Ambrosio, Luigi et al. "Gradient Flows: In Metric Spaces and in the Space of Probability Measures." (2005).

[12] Prato, Giuseppe Da and Jerzy Zabczyk. "Second order partial differential equations in Hilbert spaces." (2002).

[13] Amari S. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Transactions on Information Theory, 2009, 55(11): 4925–4931

[14] Cichocki A, Adunek R, Phan A H, Amari S. Nonnegative Matrix and Tensor Factorizations. John Wiley, 2009

[15] Opper M, Saad D. Advanced Mean Field Methods-Theory and Practice. Cambridge, MA: MIT Press, 2001

[16] Fujita, Yasuhiro et al. "Asymptotic Solutions of Viscous Hamilton–Jacobi Equations with Ornstein–Uhlenbeck Operator." *Communications in Partial Differential Equations* 31 (2006): 827 - 848.

[17] Maslowski, Bohdan and Jan Pospíšil. "Ergodicity and Parameter Estimates for Infinite-Dimensional Fractional Ornstein-Uhlenbeck Process." *Applied Mathematics and Optimization* 57 (2008): 401-429.

[18] Maslowski, Bohdan and Jan Pospíšil. "Parameter estimates for linear partial differential equations with fractional boundary noise." *Commun. Inf. Syst.* 7 (2007): 1-20.