Differential Geometric Approaches to Machine Learning

Pouplin, Alison Marie Sandrine

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Differential Geometric Approaches to Machine Learning

Alison Pouplin

Technical University of Denmark

Alison Pouplin: Differential Geometric Approaches to Machine Learning, supervised by Prof. Søren Hauberg and Asst. Prof. Georgios Arvanitidis, DTU Compute, Technical University of Denmark, 2021.
July 2023.

Summary in English

Differential geometry is a branch of mathematics that focuses on the study of manifolds and their properties, through the use of calculus and algebra. In machine learning, various high-dimensional spaces, including data space and parameter space, can be effectively treated as manifolds. Gaining insights into the structure of these manifolds is crucial for addressing numerous machine learning challenges.

This thesis demonstrates how differential geometry can be applied to tackle machine learning problems. It is divided into two parts, with the first part providing a general introduction to three subfields of differential geometry, and the second part highlighting the research conducted during this thesis. The first chapter delves into Riemannian geometry, offering an intuitive understanding of Riemannian objects and curvature concepts. This chapter lays the foundation for further investigation into the curvature of loss landscapes in the context of neural network generalization. The second chapter sheds light on Finsler geometry, drawing comparisons with Riemannian geometry. Finsler geometry has proven valuable in comparing expected lengths derived from a stochastic manifold. The third chapter introduces the Fisher-Rao metric in information geometry. This metric is employed to explore the latent space of Variational Auto-Encoders decoding to various distributions.

Resumé

Differentialgeometri er en gren af matematik, der fokuserer på studiet af mangfoldigheder og deres egenskaber gennem brug af calculus og algebra. I maskinlæring kan forskellige højdimensionelle rum, inklusive dataspace og parameterspace, effektivt behandles som mangfoldigheder. At opnå indsigt i strukturen af disse mangfoldigheder er afgørende for at kunne håndtere adskillige udfordringer inden for maskinlæring.

Denne afhandling demonstrerer, hvordan differentialgeometri kan anvendes til at tackle problemer inden for maskinlæring. Den er opdelt i to dele, hvor den første del giver en generel introduktion til tre underfelter af differentialgeometri, og den anden del fremhæver den forskning, der er udført i løbet af denne afhandling. Det første kapitel dykker ned i Riemannsk geometri, der tilbyder en intuitiv forståelse af Riemannske objekter og krumningsbegreber. Dette kapitel lægger grundlaget for yderligere undersøgelse af tab landskabets krumning i konteksten af neural netværks generalisering. Det andet kapitel kaster lys over Finsler-geometri og trækker sammenligninger med Riemannsk geometri. Finsler-geometri har vist sig værdifuld i sammenligningen af forventede længder afledt fra en stokastisk mangfoldighed. Det tredje kapitel introducerer Fisher-Rao metrikken i informationsgeometri. Denne metrik bruges til at udforske det latente rum af Variational Auto-Encoders, der dekoder til forskellige distributioner.

Preface

The present thesis was written at the Section for Cognitive Systems, DTU Compute, Technical University of Denmark in fulfillment of the requirements for acquiring a PhD degree at the Technical University of Denmark. Professor Søren Hauberg and Assistant Professor Georgios Arvanatidis supervised the project. The project was funded by the European Research Council (ERC) under the European Union Horizon 2020 research, innovation programme (757360).

The project was carried out from April 2020 to June 2023 at the Technical University of Denmark, with an exception to a four months external stay at the University of Cambridge under the supervisor of Associate Professor Carl Henrik Ek, and a three months internship at Relation Therapeutics, in London.

This thesis introduces of three different fields of differential geometry, leading to three papers, included in this thesis.

Alison Pouplin
Nantes, 14 of July 2023

Contents

Part I

Introduction

Prologue, Motivation and Approach 1

1.1 Differential geometry in machine learning

Self-supervised learning and data generation often involve mapping learned low-dimensional representations back to the original data space, via generative models. If we want to accurately navigate this latent space while preserving the underlying geometrical structure of the data, we should not assume it to be Euclidean. Instead, we need to draw upon tools from differential geometry. This statement can be better understood through a simple example.

Travelling the world: The surface of the Earth, a sphere embedded in $\mathbb{R}^3$, can be represented on a large map, a plane of 2 dimensions. When travelling from one country to another, we want to follow the shortest path that will lead us to our destination. It is tempting to draw a straight line on this map that represents our journey. Yet, this straight line on the map is not the shortest path on Earth. This is because the sphere cannot be represented, without any deformation, on a plane.

Figure 1.1: The shortest path between Cayenne and Paris is not a straight line on the map.

Whenever we are dealing with a non-Euclidean space, we need to borrow tools from differential geometry, and non-Euclidean manifolds are present everywhere in machine learning. They can be a low-dimensional latent space learned through generative models, the space of probability distributions, or the space of parameters of a neural network.

In this thesis, we study different aspects of differential geometry for machine learning. Specifically, we use tools from Riemannian geometry, Finsler geometry, and information geometry to solve both practical and theoretical problems. Each topic form its own chapter, and serves as an introduction to the research carried out in this thesis.

This research is based on two hypotheses, one that is essential to perform geometry, and another one that is practical to construct a metric.

1. **Our data points should lie on a manifold.**
   We assume that the data we are working with lie on a smooth surface, which can be highly complex and non-trivial but shouldn't have any irregularities. This condition is called the *manifold hypothesis* (Fefferman et al. 2016). All the differential geometry is built upon it and, depending on the data, it can be difficult to prove or disprove it in practice. As long as our observations have some inherent structure, the hypothesis is very likely to be true.

2. **Our generative models form an immersion.**
   To navigate the latent space effectively, we must equip it with a metric. This metric is obtained through the map that learns the latent space, typically a generative model. The metric functions as a norm, and for it to be non-degenerate, the map must be an *immersion*: a smooth function whose derivatives are injective.

1.2 Outline of this thesis

An accessible introduction of diverse subfields in differential geometry

This manuscript aims to provide an **accessible introduction** to three promising fields that could benefit the machine learning community. It is crafted for any curious reader, regardless of their background. The chapters are written in a way that tries to convey both the intuition and the motivations behind the numerous mathematical definitions.

A subfield of geometric deep learning

This research falls within **geometric deep learning**, is a field focusing on applying deep learning methods to data with a non-Euclidean structure, such as graphs or manifolds. As such, it covers many subfields including graph theory, gauge theory, group theory, and manifold learning. Due to the increasing popularity of graph neural networks, the term geometric deep learning has often been used interchangeably with graph theory in deep learning. However, this thesis does not cover graph theory, nor does it cover gauge theory or group theory.

Yet, these fields are deeply interconnected. In particular, differential geometry and graph theory share numerous concepts, such as the notion of curvature or the Laplacian operator. When a graph can be viewed as a mesh - a discrete approximation of a manifold - those connections deepen.

Such conceptual bridges may enable us to tackle more complex problems that may not be easily approached through a single mathematical lens. My humble hope is that this manuscript may provide new perspectives within the geometric deep learning community.

Thesis layout

To fully comprehend the content of chapters 2 and 3, the readers would benefit from a basic understanding of Riemannian geometry. The first half of chapter 1 conveniently provides an overview of essential concepts. While the papers are summarized at the end of each chapter, they are not extensively discussed. Instead, the reader is encouraged to refer to the original papers for a more comprehensive understanding.

In Chapter 2, we introduce the basic concepts of Riemannian geometry, where we focus on the first part on the notion of Riemannian metric and on the second part, the notion of curvature. In Chapter 3, we introduce the basic concepts of Finsler geometry, where we motivate the definition of the Finsler metric and compare the Finsler geometry with Riemannian geometry. In Chapter 4, we introduce the Fisher-Rao metric and look at its properties.

The papers are summarized at the end of each chapter, and references in Chapter 5, Chapter 6, and Chapter 7.

Contributions

This thesis is build upon the following papers, which are summarized at the end of each chapter, and references as chapters.

1. **On the curvature of the loss landscape**
   Alison Pouplin, Hrittik Roy, Sidak Pal Singh, Georgios Arvanitidis.
   Preprint, arXiv:2307.04719, 2023.
2. **Identifying latent distances with Finslerian geometry**
   Alison Pouplin, David Eklund, Carl Henrik Ek, Søren Hauberg.
   NeurIPS 2022 Workshop "Symmetry and Geometry in Neural Representations".
   Currently, under review at TMLR. Available at arXiv:2212.10010.
3. **Pulling back information geometry**
   Georgios Arvanatidis*, Miguel González-Duque*, Alison Pouplin*, Dimitris Kalatzis*, Søren Hauberg*. * Equal contribution
   Published in AISTATS 2022, and available at arXiv:2106.05367.

Besides these, I also participated in the following research projects:

1. **PyRelationAL: a python library for active learning research and development**
   Paul Scherer, Thomas Gaudelet, Alison Pouplin, Alice Del Vecchio, Suraj M S, Oliver Bolton, Jyothish Soman, Jake Taylor-King, Lindsay Edwards.
   Preprint, arXiv:2205.11117.
2. **Density estimation on smooth manifolds with normalizing flows**
   Dimitris Kalatzis, Johan Ziruo Ye, Alison Pouplin, Jesper Wohlert, Søren Hauberg.
   Preprint, arXiv:2106.03500.

1.3 Before embarking on a journey on curved spaces

The aim of the following section is to give the reader enough intuition to understand the bases of differential geometry.

Before exploring a manifold of interest, it is essential for it to meet certain requirements to be considered a **smooth manifold**. Once these criteria are satisfied, the next step involves equipping the manifold with geometric tools that enable differentiation through a **connection**, and measurements with a **metric**.

Where can we navigate? A manifold theory

What is a manifold? At its essence, a manifold is like a curve or a surface defined in any dimension. It needs to have a coherent structure such that, if we divide the manifold into different pieces, each piece should be well-defined, without peculiarities. Such manifolds are **topological manifolds**, and they are extensively studied in topology. In geometry, additional calculus operations, including differentiation, need to be performed on those manifolds. We need to be able to smoothly transition from one patch to another, leading us to define **smooth manifolds**, the central objects of our study.

Topological manifolds

At its core, a topological manifold is a **topological space**. It means that it can be fully covered using patches, allowing for different possible arrangements. We have three further requirements: first, we want to be able to count the number of patches used (**second countable**). Second, we want to be sure that the space does not fold or overlap on itself (**Hausdorff**). Finally, when we zoom in close enough to see a point, then it should look like the Euclidean space (**locally Euclidean**).

Being **second countable** ensures that the manifold is of manageable size. Although the number of patches can be infinite, it must be countably infinite, establishing a one-to-one correspondence with the natural numbers. The **Hausdorff** condition guarantees the ability to separate distinct points on the manifold, preventing scenarios where, for example, two points from different lines become indistinguishable at their intersection. Lastly, when observed at a sufficiently close scale, the manifold should resemble the **Euclidean** space, allowing for operations to be conducted locally.

Smooth manifolds

A smooth manifold, also called differentiable manifold, is a topological manifold on which we can perform calculus. To define a smooth manifold, we introduce the concept of charts and atlases. A **chart** is a pair (U, φ) consisting of an open subset U (a patch) of the manifold and a homeomorphism φ between that subset and a corresponding subset of Euclidean space. An **atlas** is a collection of charts that covers the entire manifold.

To qualify as a **smooth**, the transition functions between overlapping charts in the atlas must be smooth. In other words, if we switch from one chart to another, the transition should be seamless and allow for smooth calculations and differentiability.

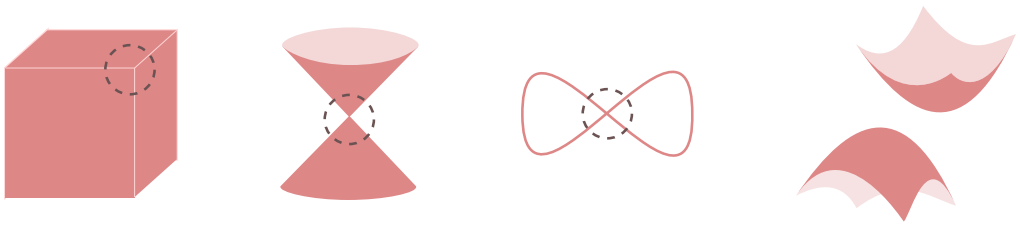


Figure 1.2: Understanding what is considered a non-smooth manifold can be easier than understanding a smooth manifold. **The first three figures** on the left are not classified as smooth manifolds because they have critical points. Specifically, the double cone and the line with a node are not Hausdorff, and so they are not considered as topological manifold. The cube can be considered as a topological manifold, but we cannot navigate smoothly from one side to another, so it is not a smooth manifold. **The last figure** represents the union of two smooth manifolds, making it a smooth manifold even though the two manifolds are disjoint.

Variations of smooth manifolds

In this manuscript, and very often in differential geometry, unless explicitly stated, the term **smooth manifolds** denotes manifolds that are homeomorphic to a designated modelling space – the Euclidean plane in $\mathbb{R}^d$, with d a finite dimension. Variations to this modeling space can yield diverse interesting manifolds. Notably, if the manifold is homeomorphic to a Hilbert, Banach or Frechet space, then we respectively define a **Hilbert, Banach or Frechet manifolds**, which have the advantage of being infinite dimensional, and are used for example in Information geometry. If the manifold is homeomorphic to a Euclidean plane in $\mathbf{C}^d$, then we obtain a **complex manifold**, and

equipped with a compatible hermitian metric, they are called **Kähler manifolds**. They are used in mathematics and mathematical physics.

Navigating by sight: The connection approach

Now that we have established what a manifold is, we can equip it with additional structure to perform differentiation. Yet, extending the notion of directional derivatives from the Euclidean space to manifolds is challenging: not only do the vectors change from one point to another, but their bases also vary along the manifold. To ensure consistent vector differentiation, we must have a way to compare the vectors and connect the bases as we move along the manifold. In other words, we need a connection.

What is a connection? A connection ∇ is a mathematical operator that generalizes the usual directional derivatives from the Euclidean space to manifolds. The operator aims to connect frames on the manifold so that we are able to compare vectors. When the connection is linear and torsion-free (no line on the manifold can be twisted), the connection is called an **affine connection**. The result of the affine connection applied to vector fields is called a **covariant derivative**.

Parallel transport

With a connection, we can ensure that the vectors are defined the same way along a path, by transporting them parallelly along a curve: this is **parallel transport**.

For a smooth curve γ on the manifold, a vector $\mathbf{x}$ is said to be parallel transported along the curve if $\nabla_{\dot{\gamma}}\mathbf{x} = 0$. For any manifold, parallel transport can be uniquely defined. It is either characterized directly by a connection, and conversely, we can derive a connection through parallel transport.

In summary, parallel transport enables a coherent comparison of vectors at distinct locations on a manifold by moving them along curves. However, this process is not path-independent: two vectors parallel-transported from and to the same points, but along different paths will be different. This effect is called **holonomy**, and it arises directly from the curvature of the manifold. Studying the holonomy of a manifold enables to quantify its curvature, and in particular, in Riemannian geometry, its Riemann curvature.

Navigating efficiently: A metric perspective

In the Euclidean space, we can effortlessly calculate the length, volume, and angle between vectors. The foundation of these computations is the dot product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$ that quantifies the interaction between two vectors, and remain unchanged across the space. We need a similar metric that accommodates for the curvature of the manifold. This is the idea behind the **Riemannian metric**.

Riemannian metric

The Riemannian metric is a positive-definite inner product, historically noted g by physicist studying gravity, defined for each point of the manifold. It is a fundamental quantity in differential geometry, so that it was previously called the first fundamental form. For two vectors $\mathbf{x}_1, \mathbf{x}_2$, we have: $g(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{G}\ \mathbf{x}_2$, with $\mathbf{G}$ the metric tensor that varies across the manifold. There are two things to note about the metric tensor: (1) it always needs to be a symmetric positive-definite matrix, for the scalar product to be

well-defined and (2) because it reflects the local properties of the manifold, the metric will vary smoothly from point to point.

The Riemannian metric plays the role of the scalar product on the manifold, and it allows for the definition of length $\mathcal{L}[\gamma] = \int \sqrt{g(\dot{\gamma}, \dot{\gamma})}dt$, and volume $\mathcal{V} = \sqrt{\det \mathbf{G}}$.

Pullback metric

In theory, any metric that is symmetric positive-definite is a Riemannian metric. If our goal is to navigate a low dimensional version $\mathcal{Z}$ of a high dimensional manifold $\mathcal{X}$ learned via a mapping $f : \mathcal{Z} \to \mathcal{X}$, a relevant metric can be constructed by *pulling back* the vectors through this said mapping. The metric tensor is defined as $\mathbf{G} = J^\top J$, with J the Jacobian of f. This metric is called the **pullback metric**.

To define a pullback metric, we need the function f to be always differentiable, and its Jacobian J to be injective. Combined, these conditions classify f as an **immersion**.

Other metrics

By relaxing some conditions, we can define other non-Riemannian metrics. The **pseudo-Riemannian metric** is a metric that is not necessarily positive-definite. This means that the distance between two distinct points can be null or even negative. In general relativity, such a metric allows for the treatment of relativistic spacetime phenomena. The **Finsler metric** is a metric that is defined as a norm that can be asymmetric and satisfies an additional convexity criterion. We will introduce Finsler geometry in Chapter 3. Finsler geometry can be seen as a generalization of Riemannian geometry.

Part II

Differential geometry

Riemannian geometry 2

Riemannian geometry is the geometrical exploration of smooth manifolds that are equipped with a Riemannian metric. It is certainly the oldest and most popular branch of differential geometry, overly present in the theory of general relativity, and increasingly prevalent in machine learning. When studying a manifold, two questions arise:

1. How do I perform operations on the manifold?
2. How curved is the manifold?

Those two questions split up the chapter into two parts. In the first part, we will reintroduce, more formally, the metric and the connection, to compute the geodesic, a recurring theme of this manuscript. In the second part, we will introduce the notion of curvatures and its derivatives, via the Riemannian curvature tensor.

2.1 Riemannian metric, connections and volume forms

Riemannian metrics

A Riemannian metric is a smoothly varying, positive-definite, symmetric bilinear form on the manifold. Essentially, it is an assignment of an inner product to each tangent space of the manifold, thereby providing a way to measure lengths and angles locally.

Definition 2.1.1 Riemannian metric
Let $\mathcal{M}$ be a differentiable manifold. A **Riemannian metric** is a map assigning at each point $x \in \mathcal{M}$ a scalar product $g(\cdot,\cdot) : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$, with $\mathbf{G}(x)$ a positive definite bilinear map, called the **metric tensor**, which varies smoothly with respect to x.
A smooth manifold equipped with a Riemannian metric is called a **Riemannian manifold**. We usually express the metric as $g(\mathbf{u},\mathbf{v}) = \langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{G}} = \mathbf{u}^\top \mathbf{G}\mathbf{v}$.
The **induced norm** is defined as: $\mathbf{v} \in \mathcal{T}_x\mathcal{M}, \|\mathbf{v}\|_{\mathbf{G}} = \sqrt{g(\mathbf{v},\mathbf{v})}$.

If the Riemannian metric tensor is the identity, we have the Euclidean metric, and the inner product becomes the usual dot product. We can construct the Riemannian metric such that we have distances that suit our goal.

Example 2.1.1 Mahalanobis distance
In statistics, the Mahalanobis distance determines how far a point $\mathbf{x}$ is from a distribution of mean $\boldsymbol{\mu}$ and covariance Σ. The Mahalanobis distance is defined as:

$$d(\mathbf{x},\boldsymbol{\mu}) = \|\boldsymbol{\mu}-\mathbf{x}\|_{\Sigma^{-1}} = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

We can see that the Mahalanobis distance is a Riemannian metric, with the metric tensor $\mathbf{G} = \Sigma^{-1}$. Note that here, the metric is constant, as the metric tensor does not vary on the manifold.

Pullback metric

We can also infer a Riemannian metric on a *source manifold* $\mathcal{M}$ by pulling back the metric of a *target manifold* $\mathcal{N}$ via a smooth map $f : \mathcal{M} \to \mathcal{N}$. The pullback operation provides a mechanism for transferring the geometric structure of one manifold onto another. The metric induced in the source manifold is called the *pullback metric*.

Definition 2.1.2 Pullback metric
Let $\mathcal{M}$ and $\mathcal{N}$ be two smooth manifolds, and $f : \mathcal{M} \to \mathcal{N}$ an immersion. If $\mathcal{N}$ is equipped with a Riemannian metric g, then the **pullback metric** f^*g on $\mathcal{M}$ is defined as:

$$f^*g(\mathbf{u}, \mathbf{v}) = g(f_*\mathbf{u}, f_*\mathbf{v}), \quad \forall(\mathbf{u}, \mathbf{v}) \in \mathcal{T}_x\mathcal{M},$$

where f_* is the pushforward operation.

In practice, the pushforward operation is the differential of the map f, i.e. $f_* = df = J$, with J, the Jacobian of f. For example, the pushforward of a vector $\mathbf{v} \in \mathcal{T}_x\mathcal{M}$ is $J\mathbf{v} \in \mathcal{T}_{f(x)}\mathcal{N}$. When the target manifold is equipped with the Euclidean metric, the pullback metric tensor is defined as $\mathbf{G} = J^\top J$.

We can understand where this result comes from with the Taylor expansion of the norm locally defined on the target manifold. We have $f : \mathcal{M} \to \mathcal{N} : \mathbf{x} \to \mathbf{y}$. On the target manifold, we can approximate the surface by its tangent plane: $f(\mathbf{x}+\Delta\mathbf{x}) \approx f(\mathbf{x}) + J\Delta\mathbf{x}$ and compute the Euclidean norm of $\Delta\mathbf{y}$:

$$\|\Delta\mathbf{y}\|_2^2 = \|f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x})\|_2^2 \approx \Delta\mathbf{x}^\top J^\top J\ \Delta\mathbf{x} = \|\Delta\mathbf{x}\|_{\mathbf{G}}^2\,,$$

with the metric tensor $\mathbf{G} = J^\top J$.

Example 2.1.2 Riemannian metric on the sphere
We consider a sphere parameterized with

$$f(\theta, \varphi) = r(\cos(\theta)\cos(\varphi), \cos(\theta)\sin(\varphi), \sin(\theta)),$$

with θ the polar angle and φ the azimuth angle. The Jacobian and the metric tensor are:

$$J = r\begin{pmatrix} -\sin(\theta)\cos(\varphi) & -\cos(\theta)\sin(\varphi) \\ -\sin(\theta)\sin(\varphi) & \cos(\theta)\cos(\varphi) \\ \cos(\theta) & 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{G} = r^2\begin{pmatrix} 1 & 0 \\ 0 & \cos^2(\theta) \end{pmatrix}.$$

Indicatrices

It is sometimes difficult to get some intuition about the Riemannian metric. The **indicatrices**, which are the set of vectors in the tangent plane such that their Riemannian norm is 1, provide a way to visualize the metric, and showcase the deformation of the plane.

Definition 2.1.3 Indicatrices
For a given point x on the Riemannian manifold $(\mathcal{M}, g)$, the **indicatrix** represents the set of vectors in the tangent space $\mathcal{T}_x\mathcal{M}$ such that the induced norm $\|\mathbf{u}\|_{\mathbf{G}} = \sqrt{g(\mathbf{u}, \mathbf{u})}$ that equipped the manifold is equal to 1:

$$\text{indicatrix at } x = \{\mathbf{v} \in \mathcal{T}_x\mathcal{M} |\ \|\mathbf{v}\|_{\mathbf{G}} = 1\}).$$

In the Euclidean plane in 2-dimensions, the indicatrices represent a circle of radius 1. In a Riemannian manifold, the indicatrices are ellipses, whose semi-axis's lengths

and directions are related to the eigenvalues and eigenvectors of $\mathbf{G}$. It represents how stretched or shrunk is the manifold compared to the Euclidean plane.

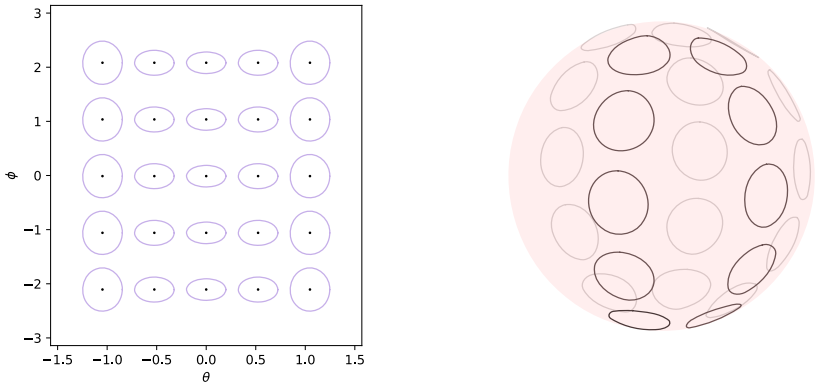


Figure 2.1: The plane (ϕ, θ) parameterizes the sphere, and the deformation of the plane are illustrated with the ellipses, representing the Riemannian metric. When pushforward to the sphere, those ellipses become circles, representing the Euclidean norm.

Connections

In curved manifolds, we can't use ordinary derivatives to differentiate vector fields due to the coordinates changing along the manifolds. We solve this by introducing a **connection** - a mathematical object that compares vectors at different points on the manifold, by connecting their coordinate's basis. Let's imagine that we would like to differentiate a vector $\mathbf{v}$ along another vector $\mathbf{u}$, such that, in a $\{e_i, e_j, e_k\}$ basis, we have: $\mathbf{u} = u^i e_i$ and $\mathbf{v} = v^j e_j$. Then:

$$\begin{aligned}\nabla_{\mathbf{u}}\mathbf{v} &= u^i \partial_i (v^j e_j) \\ &= u^i \partial_i (v^j) e_j + u^i v^j \partial_i e_j \\ &= u^i \partial_i (v^j) e_j + u^i v^j \Gamma_{ij}^k e_k\end{aligned}$$

The coefficients $\Gamma_{ij}^k e_k$ are called the **Christoffel symbols**. They represent a corrective term, added on the usual Euclidean derivative $u^i \partial_i (v^j) e_j$ to differentiate $\mathbf{v}$ along $\mathbf{u}$ on the manifold. In general, the derivative of e_j along e_i depends on another direction e_k: $\partial_i e_j = \Gamma_{ij}^k e_k$. On a flat Euclidean plane, the basis vectors e_i do not change along any direction, so we always have $\Gamma_{ij}^k = 0$.

Levi-Civita connection

A connection is **compatible** with a metric if the metric is preserved under parallel transport along any curve on the manifold. Parallel transport, in differential geometry, refers to the process of transporting a vector or tensor along a curve while preserving its direction and orientation. This means that if we take two vectors at a point, and we transport them along a curve such that their direction stays constant, the metric properties will remain consistent along the curve. In Riemannian geometry, the **Levi-Civita connection** is the unique connection that is compatible with the metric.

Theorem 2.1.1 The fundamental theorem of Riemannian geometry
On a Riemannian manifold $(\mathcal{M}, g)$, there exists a unique symmetric connection which is compatible with the metric g. This connection is called the **Levi-Civita connection.**

Proof. Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be vector fields on $\mathcal{M}$. We denote $\mathbf{x}g(\mathbf{y}, \mathbf{z})$ the derivative of the scalar field $g(\mathbf{y}, \mathbf{z})$ in the direction of the vector field $\mathbf{x}$. The Levi-Civita has two desirable properties:

1. Symmetry: $\nabla_{\mathbf{x}}\mathbf{y} - \nabla_{\mathbf{y}}\mathbf{x} = [\mathbf{x}, \mathbf{y}]$.
2. Metric-compatibility: $\mathbf{x}g(\mathbf{y}, \mathbf{z}) = g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) + g(\mathbf{y}, \nabla_{\mathbf{x}}\mathbf{z})$.

We can write the previous properties as:

1. $\mathbf{x}g(\mathbf{y}, \mathbf{z}) = g(\nabla_{\mathbf{x}}\mathbf{y}, \mathbf{z}) + g(\mathbf{y}, \nabla_{\mathbf{x}}\mathbf{z}) = g([\mathbf{x}, \mathbf{y}], \mathbf{z}) + g(\nabla_{\mathbf{y}}\mathbf{x}, \mathbf{z}) + g(\mathbf{y}, \nabla_{\mathbf{x}}\mathbf{z})$.
2. $\mathbf{y}g(\mathbf{z}, \mathbf{x}) = g(\nabla_{\mathbf{y}}\mathbf{z}, \mathbf{x}) + g(\mathbf{z}, \nabla_{\mathbf{y}}\mathbf{x})$
3. $\mathbf{z}g(\mathbf{x}, \mathbf{y}) = g(\nabla_{\mathbf{z}}\mathbf{x}, \mathbf{y}) + g(\mathbf{x}, \nabla_{\mathbf{z}}\mathbf{y})$

Adding the three equations (1)+(2)-(3), we get the Koszul formula:

$$2g(\nabla_{\mathbf{y}}\mathbf{x}, \mathbf{z}) = \mathbf{x}g(\mathbf{y}, \mathbf{z}) + \mathbf{y}g(\mathbf{z}, \mathbf{x}) - \mathbf{z}g(\mathbf{x}, \mathbf{y}) - (g([\mathbf{x}, \mathbf{y}], \mathbf{z}) + g([\mathbf{y}, \mathbf{z}], \mathbf{x}) + g([\mathbf{x}, \mathbf{z}], \mathbf{y})) .$$

The right-hand side of the equation does not depend on the connection ∇. If we had two similar connections ∇_1 and ∇_2, then $g((\nabla_1 - \nabla_2)_{\mathbf{y}}\mathbf{x}, \mathbf{z}) = 0 \implies \nabla_1 = \nabla_2$. So the Levi-Civita connection exists and is unique. □

The Christoffel symbols derived from the Levi-Civita connection can be expressed in function of the metric $g_{ij} = g(e_i, e_j)$:

$$\begin{aligned}
\nabla_k g_{ij} &= \nabla_k(g)(e_i, e_j) + g(\nabla_k e_i, e_j) + g(e_i, \nabla_k e_j) \\
0 &= \nabla_k g_{ij} - g(\nabla_k e_i, e_j) - g(e_i, \nabla_k e_j) \\
0 &= \partial_k g_{ij} - g(\Gamma^m_{ik} e_m, e_j) - g(e_i, \Gamma^m_{kj} e_m) \\
0 &= \partial_k g_{ij} - \Gamma^m_{ik} g_{mj} - \Gamma^m_{kj} g_{im}
\end{aligned}$$

Because the Levi-Civita connection is also torsion-free, we have: $\nabla_i e_j = \nabla_j e_i$, which translates into: $\Gamma^k_{ij} = \Gamma^k_{ji}$. When we permute the index and sum all the equations, we obtain: $\Gamma^m_{ik}\ g_{jm} = \frac{1}{2}(\partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ik})$, and with g^{jm} the inverse of g_{jm}:

$$\Gamma^m_{ik} = \frac{1}{2} g^{jm}(\partial_k g_{ij} + \partial_i g_{jk} - \partial_j g_{ik})$$

The Christoffel symbols are useful to determine the Ricci curvature of a manifold, and also to find a geodesic, using the geodesic equation. Let's say that we want to find the curve γ parameterized by t, such that along the curve, its velocity is constant. We note $\dot{\gamma} = \frac{\partial \gamma^i}{\partial t} e_i$:

$$\begin{aligned}
\nabla_{\dot{\gamma}}\dot{\gamma} &= \nabla_{\dot{\gamma}} \left(\frac{\partial \gamma^i}{\partial t} e_i \right) = \frac{\partial}{\partial t} \left(\frac{\partial \gamma^i}{\partial t} \right) e_i + \frac{\partial \gamma^j}{\partial t} \frac{\partial \gamma^i}{\partial t} \nabla_j(e_i) \\
&= \left(\frac{\partial^2 \gamma^k}{\partial t^2} + \Gamma^k_{ij} \frac{\partial \gamma^i}{dt} \frac{\partial \gamma^j}{\partial t} \right) e_k
\end{aligned}$$

Which leads to the **geodesic equation**:

$$\frac{\partial^2 \gamma^k}{\partial t^2} + \Gamma^k_{ij} \frac{\partial \gamma^i}{dt} \frac{\partial \gamma^j}{\partial t} = 0,$$

with γ a geodesic defined as the **straightest** possible curve.

Functional on curves

Now that we have defined the Riemannian metric, we can define the length and the energy of a curve $\gamma(t)$ parameterized by $t \in [0, 1]$ on a manifold.

Definition 2.1.4 Length and energy functional
We consider a curve γ and its derivative $\dot{\gamma}$ on a Riemannian manifold $\mathcal{M}$ equipped with the metric g. Then, we define the **length and the energy of the curve**:

$$\mathcal{L}(\gamma) = \int \|\dot{\gamma}(t))\|_{\mathbf{G}} \, dt \qquad \text{and} \qquad \mathcal{E}(\gamma) = \frac{1}{2} \int \|\dot{\gamma}(t))\|_{\mathbf{G}}^2 \, dt,$$

with $g_t = g_{\gamma(t)}$ and $\|\cdot\|_{\mathbf{G}}$ the norm induced by the metric g. Locally length-minimizing curves between two connecting points are called **geodesics**.

Two important characteristics stand out when it comes to the length of a curve and its energy.

First, the length is **reparameterization invariant:** let φ be a diffeomorphism mapping one basis to another such that $\gamma' = \varphi(\gamma)$. Then $\mathcal{L}(\gamma') = \mathcal{L}(\gamma)$. This property is quite important in machine learning, when we will want to compute length-minimizing curves in the latent space of a generative model. In other words, for two different runs of the same model, the geodesics γ and γ' computed in the latent spaces $\mathcal{Z}$ and $\mathcal{Z}'$ will lead to the same curve in the data space $\mathcal{X}$. We say that it solves the **identifiability problem** (Hauberg 2019).

Secondly, using the Cauchy-Schwarz inequality, we have: $\mathcal{L}(\gamma)^2 \leq 2\mathcal{E}(\gamma)$, with equality when γ minimizes the energy. Because of the reparameterization invariance of the curve length, a solver can find an infinite number of solutions to minimize it. On the other hand, the curve energy is convex and will converge to a unique solution. In practice, when the goal is to compute a geodesic, instead of solving the relevant system of ordinary differential equations, it can be easier to **minimize the curve energy**.

The shortest paths are also the straightest ones

When minimizing the energy functional, we are seeking to find the geodesic, defined as the **shortest** possible curve. It is interesting to note that the shortest curve is also the straightest curve when considering the Levi-Civita connection.

To achieve this minimization, we start with the general integral $\mathcal{E}(\gamma) = \int_c F(\gamma_t, \dot{\gamma}_t)dt$ where the integrand is given by: $F(\gamma_t, \dot{\gamma}_t) = \frac{1}{2} g_{ij} \dot{\gamma}_t^i \dot{\gamma}_t^j$. By applying the Euler-Lagrange equation, we obtain:

$$\frac{\partial}{\partial t} \left(\frac{\partial F}{\partial \dot{\mathbf{x}}^k} \right) - \frac{\partial F}{\partial \mathbf{x}^k} = 0,$$

with:

$$\begin{aligned}
\frac{\partial F}{\partial \dot{\mathbf{x}}^k} &= \frac{1}{2} g_{ij} \left(\frac{\partial \dot{\mathbf{x}}^i}{\dot{\mathbf{x}}^k} \dot{\mathbf{x}}^j + \frac{\partial \dot{\mathbf{x}}^j}{\dot{\mathbf{x}}^k} \dot{\mathbf{x}}^i \right) = \frac{1}{2} g_{ij} \left(\delta_{ik} \dot{\mathbf{x}}^j + \delta_{jk} \dot{\mathbf{x}}^i \right) \\
&= \frac{1}{2} (g_{ij} \dot{\mathbf{x}}^j + g_{ij} \dot{\mathbf{x}}^i) = g_{ij} \dot{\mathbf{x}}^i \\
\frac{\partial}{\partial t} \left(\frac{\partial F}{\partial \dot{\mathbf{x}}^k} \right) &= g_{ij} \frac{\partial \dot{\mathbf{x}}^i}{\partial t} + \frac{\partial g_{ij}}{\partial t} \dot{\mathbf{x}}^i = g_{ij} \frac{\partial^2 \mathbf{x}^i}{\partial t^2} + \frac{\partial g_{ik}}{\partial \mathbf{x}^j} \frac{\partial \mathbf{x}^j}{\partial t} \frac{\partial \mathbf{x}^i}{\partial t} \\
&= g_{ij} \frac{\partial^2 \mathbf{x}^i}{\partial t^2} + \frac{1}{2} \left(\frac{\partial g_{ik}}{\partial \mathbf{x}^j} + \frac{\partial g_{jk}}{\partial \mathbf{x}^i} \right) \frac{\partial \mathbf{x}^j}{\partial t} \frac{\partial \mathbf{x}^i}{\partial t} \\
\frac{\partial F}{\partial \mathbf{x}^k} &= \frac{1}{2} \frac{\partial g_{ij}}{\partial \mathbf{x}^k} \frac{\partial \mathbf{x}^j}{\partial t} \frac{\partial \mathbf{x}^i}{\partial t}
\end{aligned}$$

Notice that, using the Einstein summation, we have: $g_{ij}\dot{\mathbf{x}}^j = g_{ij}\dot{\mathbf{x}}^i$, since $g_{ij} = g_{ji}$, and indices are silent so: $g_{ij}\frac{\partial^2 \mathbf{x}^i}{\partial t^2} = g_{mk}\frac{\partial^2 \mathbf{x}^k}{\partial t^2}$. Then we recover the geodesic equation:

$$\frac{\partial}{\partial t} \left(\frac{\partial F}{\partial \dot{\mathbf{x}}^k} \right) - \frac{\partial F}{\partial \mathbf{x}^k} = g_{mk} \left(\frac{\partial^2 \mathbf{x}^k}{\partial t^2} + \Gamma^k_{ij} \frac{\partial \mathbf{x}^j}{\partial t} \frac{\partial \mathbf{x}^i}{\partial t} \right) = 0$$

Example 2.1.3 Geodesics on the sphere
We consider a sphere parameterized with $f(\theta, \varphi) = r(\cos(\theta)\cos(\varphi), \cos(\theta)\sin(\varphi), \sin(\theta))$, with θ the polar angle and φ the azimuth angle. To compute geodesics on the sphere, we can either we solve the ODE system, or we minimize the curve energy, and both methods lead to the same result. The curve $\gamma(t) = (\gamma_\theta(t), \gamma_\varphi(t))$ is parameterized in the plane. The non-null Christoffel symbols for the sphere are:

$$\Gamma^\theta_{\phi\phi} = \sin\theta\cos\theta \quad \text{and} \quad \Gamma^\varphi_{\theta\varphi} = -\tan\theta.$$

γ is a geodesic if its energy is minimized or if it satisfies a system of equations:

$$\mathcal{E}(\gamma) = \frac{1}{2} \int_c \left(\frac{\partial \gamma_\theta}{\partial t} \right)^2 + \cos^2\theta \left(\frac{\partial \gamma_\varphi}{\partial t} \right)^2 dt \qquad \text{or} \qquad \begin{cases} \frac{\partial^2 \gamma_\theta}{\partial t^2} + \sin\theta\cos\theta \left(\frac{\partial \gamma_\varphi}{\partial t} \right)^2 = 0 \\ \frac{\partial^2 \gamma_\varphi}{\partial t^2} - \tan\theta \frac{\partial \gamma_\varphi}{\partial t} \frac{\partial \gamma_\theta}{\partial t} = 0 \end{cases}$$

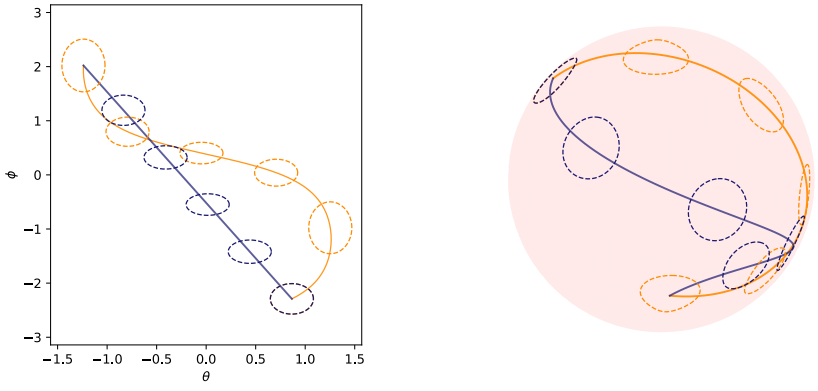


Figure 2.2: A geodesic is shown in orange and a straight line in blue. In this figure, the geodesic is obtained by minimizing the energy of a curve. The geodesics on the sphere are the great circles. Notice that they follow the direction of the indicatrices in the plane.

Volume forms

The determinant of the metric tensor $\det \mathbf{G}$ quantifies how the metric tensor distorts or stretches volumes. This determinant defines the volume form on a Riemannian manifold.

Akin to the change of variables, it is used to integrate a function from one coordinate system to another, when $f : \mathcal{M} \to \mathcal{N}$:

$$\int_{\mathcal{M}} h \circ f(\mathbf{x}) \sqrt{\det \mathbf{G}} d\mathbf{x} = \int_{\mathcal{N}} h(\mathbf{y}) d\mathbf{y}$$

Definition 2.1.5 Volume form
In local coordinates $(e^1, \cdots, e^d)$, the **volume form** of the Riemannian manifold $\mathcal{M}$, equipped with the metric tensor G, is defined as: $dV_{\mathbf{G}} = V_{\mathbf{G}}(x) e^1 \wedge \cdots \wedge e^d$, with:

$$V_{\mathbf{G}}(x) = \sqrt{\det \mathbf{G}}.$$

2.2 Computing geodesics on stochastic manifolds

Often, in machine learning, our data manifold $\mathcal{X}$ is composed of random variables obtained via a stochastic mapping $f : \mathcal{Z} \to \mathcal{X}$. When we pullback the metric from $\mathcal{X}$ to $\mathcal{Z}$, we obtain a random metric tensor, and so a random Riemannian metric.

Unfortunately, we are not yet equipped to derive geometric objects. A solution is to seek for a deterministic approximation of the metric. When the mapping f follows either Gaussian process or a Gaussian distribution, the expected metric tensor can be conveniently computed, despite some limitations discussed in Section 3.4 and Section 4.3.

Before tackling those issues, let us introduce the notion of random Riemannian metric, when f outputs a random variable and when f describes a stochastic process. Here, and in the rest of this manuscript, the terms *stochastic* and *random* are used interchangeably.

Random manifolds

A **stochastic process** is a collection of random variables $\{X(t, \omega), t \in T\}$ indexed by an index set T defined on a sample space Ω, which represents the set of all possible outcomes. An outcome in Ω is denoted by ω, and a realization of the stochastic process is the sequence of $X(\cdot, \omega)$ that depends on the outcome ω.

When we learn a low-dimensional representation of our data via a stochastic generative model $f : \mathcal{Z} \to \mathcal{X}$, the latent manifold is the index set defined above $\mathcal{Z} = T$ and the sample space Ω can be seen as the set of the model evaluations. In other words, for every point $\mathbf{z} \in \mathcal{Z}$, every time we run our model, the output $\mathbf{x} = f(\mathbf{z})$ is a random variable following a specific distribution.

Definition 2.2.1 Random Riemannian metric via a stochastic process
A **random Riemannian metric tensor** is a matrix-valued random field (i.e.: a collection of matrix-valued random variables $\{G(z, \omega), z \in \mathcal{Z}\}$), whose realization for a specific evaluation $\omega \in \Omega$ is a Riemannian metric tensor. We define the **random Riemannian metric** as the metric induced by a random Riemannian metric tensor: $g : (\mathbf{u}, \mathbf{v}) \to \mathbf{u}^\top \mathbf{G}\ \mathbf{v}$.

Similarly, if f follows a single distribution (e.g., $f(\mathbf{z}) \sim \mathcal{N}(\mu(\mathbf{z}), \sigma^2(\mathbf{z}))$), instead of a representation of a distribution over functions (e.g., $f \sim \mathcal{GP}(\mu, k)$), it will still induce a random Riemannian metric tensor, also defined as a random matrix whose sample paths are Riemannian metrics.

Because the metric is stochastic, all the functionals derived from the metric are stochastic as well, including the length of a curve, its energy or even the Christoffel symbols.

Minimizing a stochastic quantity, or solving a system of stochastic differential equations is far from trivial. Another solution is to seek for a deterministic approximation of the metric tensor.

A deterministic approximation of the metric tensor

Definition 2.2.2 Expected Riemannian metric
The **expected Riemannian metric** is the metric induced by the expected metric tensor: $g : (\mathbf{u}, \mathbf{v}) \to \mathbf{u}^\top \mathbb{E}[\mathbf{G}]\ \mathbf{v}$.

Let's consider two cases: when the mapping f is a Gaussian process, and when the mapping f follows a Gaussian distribution.

When the mapping $f \sim \mathcal{GP}(\mu, k)$ is a Gaussian process with the kernel k being differentiable, then its Jacobian exists, and it follows a Gaussian distribution, $J \sim \mathcal{N}(\mathbb{E}[J], \Sigma)$. As a consequence, $\mathbf{G}$ is a random matrix following a non-central Wishart distribution: $\mathbf{G} \sim \mathcal{W}_q(d, \Sigma, \Lambda)$ (Kent and Muirhead 1984, See definition 10.3.1), with a non-centrality term $\Lambda = \mathbb{E}[J]^\top \Sigma^{-1} \mathbb{E}[J]$. Then, the expected metric tensor can be computed analytically: $\mathbb{E}[\mathbf{G}] = \mathbb{E}[J]^\top \mathbb{E}[J] + d\Sigma$. These results have been first introduced by Tosi et al. (2014) to compute geodesics on a manifold modelled by a GP-LVM (Lawrence 2003).

Arvanitidis et al. (2018) also used the expectation of the metric pulled back from the decoder of a Variational Auto-Encoder (Kingma, Welling, et al. 2019). While the decoder cannot be modelled as a stochastic process, the data still follows a Gaussian distribution such that: $f(\mathbf{z}) = \mu(\mathbf{z}) + \sigma(\mathbf{z}) \odot \epsilon$, with $\epsilon \sim \mathcal{N}(0, \mathbb{I}_d)$. In that case, $\mathbb{E}[\mathbf{G}] = J_\mu(\mathbf{z})^\top J_\mu(\mathbf{z}) + J_\sigma(\mathbf{z})^\top J_\sigma(\mathbf{z})$, with $J_\mu(\mathbf{z}) = \nabla_{\mathbf{z}} \mu(\mathbf{z})$ and $J_\sigma(\mathbf{z}) = \nabla_{\mathbf{z}} \sigma(\mathbf{z})$.

In both cases, we obtained a closed-form expression of the expected metric tensor that conveniently helps the computations for different generative models. However, we also face some limitations: (1) the expected length does not correspond to the length of the expected metric, so the geodesics obtained with the expected metric do not correspond to the expectation of the stochastic geodesics; and (2) we don't have access to closed-form expressions when decoding to distributions that are not location-scale. The first problem is addressed in Section 3.4, using Finsler geometry, while the second one is tackled in Section 4.3, borrowing tools from information geometry.

2.3 Riemannian curvature: from tensors to scalars

Historically, Gauss (1827) and Riemann (1867) introduced two tools to study a curved surface:

- the first fundamental form, now referred as the Riemannian metric, allows us to conduct operations on the manifold;
- the second fundamental form, an extrinsic quantity that captures the curvature of the manifold in an ambient space.

As time passed and particularly with the development of tensor calculus by Ricci and Levi-Civita (1900) and the theory of connections by Cartan (1926), the understanding of curvature changed from being *extrinsic* to *intrinsic*. This means that the curvature can be measured independently of the surrounding space.

A primary intrinsic tool that evaluates how much the manifold differs from a flat Euclidean plane is the **Riemannian curvature tensor**. However, the 4-dimensional mathematical object can be challenging to understand intuitively. Therefore, additional

quantities, matrices and scalar, have been derived to better illustrate specific aspects of the manifold's curvature.

The Riemann curvature tensor

The **Riemann curvature** measures how the manifold deviates from being flat by using the concept of **holonomy**. In simple terms, it measures how much a vector changes its alignment when it is parallel transported around an infinitesimal loop on the manifold. This measurement tells us how the direction of vectors is affected, and so, how locally the surface is curved. Equivalently, we can think of the Riemannian curvature tensor as a way to measure the failure of the second covariant derivative to commute.

Definition 2.3.1 Riemann curvature tensor
Let $(\mathcal{M}, g, \nabla)$ be a Riemannian manifold equipped with ∇ the Levi-Civita connection. The **Riemann curvature tensor** is a (1,3)-tensor field on $\mathcal{M}$ defined by:

$$\mathrm{R}(\mathbf{x}, \mathbf{y}; \mathbf{z}) = \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} \mathbf{z} - \nabla_{\mathbf{z}} \nabla_{\mathbf{y}} \mathbf{z} - \nabla_{[\mathbf{x},\mathbf{y}]} \mathbf{z},$$

for any vector fields $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathfrak{X}(\mathcal{M})$, with $[\cdot, \cdot]$ the Lie bracket.

The Lie bracket is a way to measure the failure of the commutativity of two vector fields. It is defined as: $[\mathbf{x}, \mathbf{y}] = \nabla_{\mathbf{x}} \mathbf{y} - \nabla_{\mathbf{y}} \mathbf{x}$.

Proof. The Riemann curvature $\mathrm{R}(\mathbf{x}, \mathbf{y}; \mathbf{z})$ measures the deviation induced on $\mathbf{z}$ when parallel transported along a loop on a small section spanned by $\{\mathbf{x}, \mathbf{y}\}$. We consider the loop, denoted π, to be around a small parallelogram of surface area(π). The final vector transported along the parallelogram is denoted $\pi(\mathbf{z})$. We define:

$$\mathrm{R}(\mathbf{x}, \mathbf{y}; \mathbf{z}) = \lim_{\pi \to 0} \frac{\mathbf{z} - \pi(\mathbf{z})}{\text{area}(\pi)}.$$

The parallelogram is spanned by $\mathbf{x}$ and $\mathbf{y}$, and the loop can be decomposed in small loops: $\pi(\mathbf{z}) = (\pi_{-\mathbf{y}} \pi_{-\mathbf{x}} \pi_{\mathbf{y}} \pi_{\mathbf{x}})(\mathbf{z})$, with $\pi_{\mathbf{v}}(\mathbf{u})$ the parallel transport of $\mathbf{u}$ along $\mathbf{v}$. We denote area$(\pi) = |\pi_{\mathbf{x}}||\pi_{\mathbf{y}}|$, and for any vectors $\mathbf{u}$ and $\mathbf{v}$: $\pi_{\mathbf{v}}^{-1}(\mathbf{u}) = \pi_{-\mathbf{v}}(\mathbf{u})$:

$$\begin{aligned}
\mathbf{z} - \pi(\mathbf{z}) &= \mathbf{z} - (\pi_{-\mathbf{y}} \pi_{-\mathbf{x}} \pi_{\mathbf{y}} \pi_{\mathbf{x}})(\mathbf{z}) \\
&= \pi_{-\mathbf{y}} \pi_{-\mathbf{x}} \left((\pi_{-\mathbf{y}} \pi_{-\mathbf{x}})^{-1}(\mathbf{z}) - (\pi_{\mathbf{y}} \pi_{\mathbf{x}})(\mathbf{z}) \right) \\
&= \pi_{\mathbf{y}}^{-1} \pi_{\mathbf{x}}^{-1} \left(\pi_{\mathbf{x}} \pi_{\mathbf{y}}(\mathbf{z}) - \pi_{\mathbf{y}} \pi_{\mathbf{x}}(\mathbf{z}) \right)
\end{aligned}$$

Then, when the loop becomes infinitesimally small, we have:

$$\begin{aligned}
\lim_{\pi \to 0} \frac{\mathbf{z} - \pi(\mathbf{z})}{\text{area}(\pi)} &= \lim_{\pi_{\mathbf{x}} \to 0} \lim_{\pi_{\mathbf{y}} \to 0} \frac{\pi_{\mathbf{y}}^{-1} \pi_{\mathbf{x}}^{-1}}{|\pi_{\mathbf{y}}||\pi_{\mathbf{x}}|} \left((\pi_{\mathbf{x}} \pi_{\mathbf{y}}(\mathbf{z}) - \mathbf{z}) - (\pi_{\mathbf{y}} \pi_{\mathbf{x}}(\mathbf{z}) - \mathbf{z}) \right) \\
&= \lim_{\pi_{\mathbf{x}} \to 0} \lim_{\pi_{\mathbf{y}} \to 0} \left((\pi_{\mathbf{x}} \pi_{\mathbf{y}}(\mathbf{z}) - \mathbf{z}) - (\pi_{\mathbf{y}} \pi_{\mathbf{x}}(\mathbf{z}) - \mathbf{z}) \right) \\
&= \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} \mathbf{z} - \nabla_{\mathbf{y}} \nabla_{\mathbf{x}} \mathbf{z}.
\end{aligned}$$

This is a simple example, where we considered the vector fields $\mathbf{x}$ and $\mathbf{y}$ to commute. We have the Lie brackets $[\mathbf{x}, \mathbf{y}] = 0$. In the general case, they don't commute, and so we need to take into account this gap, by adding an extra term:

$$\lim_{\pi \to 0} \frac{\mathbf{z} - \pi(\mathbf{z})}{\text{area}(\pi)} = \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} \mathbf{z} - \nabla_{\mathbf{y}} \nabla_{\mathbf{x}} \mathbf{z} - \nabla_{[\mathbf{x},\mathbf{y}]} \mathbf{z}.$$

□

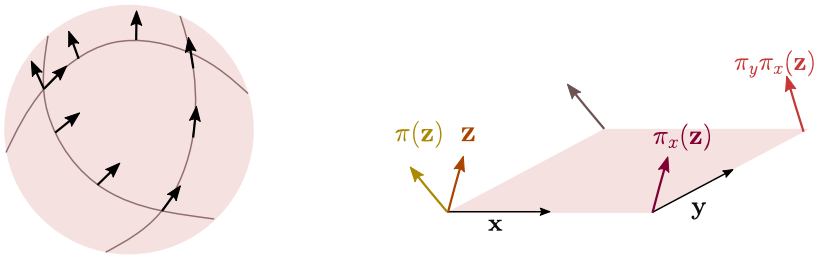
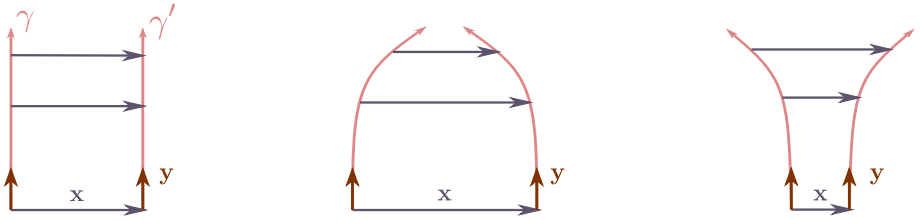


Figure 2.3: On the left, we illustrate the concept of holonomy of a connection: a vector parallel transported around a small loop on a curved manifold will fail to preserve its direction, unless the manifold is flat. **On the right**, the figure is a visual aid for the proof of definition 2.3.1, where we look at an infinitesimally small loop on the tangent space of a manifold, and we measure the deviation of the vector.

The sectional curvature

Another way to measure how a manifold differs from a flat space is to look at the deviation of two parallel geodesics on the space. Let's consider γ and γ' such that at a specific point t_0, they are next to each other and $\dot{\gamma}(0) = \dot{\gamma}'(0) = \mathbf{y}$. Now, we introduce another vector $\mathbf{x}$ with $|\mathbf{x}|_t = \gamma'(t) - \gamma(t)$ measuring the distance between those two geodesics at a time t. We want to quantify how fast the geodesics diverge from each other, which is the same as measuring the acceleration of $\mathbf{x}$ along $\mathbf{y}$. This is exactly: $\nabla_{\mathbf{y}}\nabla_{\mathbf{y}}\mathbf{x} = -\mathrm{R}(\mathbf{x},\mathbf{y};\mathbf{y})$. When projecting this acceleration over $\mathbf{x}$, we get the scalar quantity $\kappa = -\langle\nabla_{\mathbf{y}}\nabla_{\mathbf{y}}\mathbf{x},\mathbf{x}\rangle = \langle\mathrm{R}(\mathbf{x},\mathbf{y};\mathbf{y}),\mathbf{x}\rangle$.

- If $\kappa = 0$, then the geodesics always stay parallel on the manifold, which is the case in the flat plane.
- If $\kappa \geq 0$, then the norm $|\mathbf{x}|_t$ becomes smaller as we move along the geodesics. This happens on the sphere, until the geodesics intersect.
- If $\kappa \leq 0$, the gap $|\mathbf{x}|_t$ increases between the geodesics. They diverge from one another, which happens in the hyperbolic manifold.



$\langle\nabla_{\mathbf{y}}\nabla_{\mathbf{y}}\mathbf{x},\mathbf{x}\rangle = 0$ $\mathrm{K}(\mathbf{x},\mathbf{y}) = 0$ | $\langle\nabla_{\mathbf{y}}\nabla_{\mathbf{y}}\mathbf{x},\mathbf{x}\rangle < 0$ $\mathrm{K}(\mathbf{x},\mathbf{y}) > 0$ | $\langle\nabla_{\mathbf{y}}\nabla_{\mathbf{y}}\mathbf{x},\mathbf{x}\rangle > 0$ $\mathrm{K}(\mathbf{x},\mathbf{y}) < 0$

Figure 2.4: The sectional curvature is represented for the plane, the sphere and the hyperbolic space.

This deviation measure defines the sectional curvature upon normalization. The **sectional curvature** offers a way to measure locally the normalized deviation between two geodesics:

Definition 2.3.2 Sectional curvature
Let $(\mathcal{M}, g)$ be a Riemannian manifold with R its Riemann curvature tensor. The

sectional curvature is defined as:

$$\mathrm{K}(\mathbf{x},\mathbf{y}) = \frac{\langle \mathrm{R}(\mathbf{x},\mathbf{y};\mathbf{y}),\mathbf{x}\rangle}{|\mathbf{x}\wedge\mathbf{y}|^2},$$

where $\wedge$ is the exterior product between two vectors, and $|\mathbf{x}\wedge\mathbf{y}|^2 = \langle\mathbf{x},\mathbf{x}\rangle\langle\mathbf{y},\mathbf{y}\rangle - \langle\mathbf{x},\mathbf{y}\rangle^2$ is the area of the parallelogram spanned by $\mathbf{x}$ and $\mathbf{y}$.

As a side note, if we have access to different planes $\{\mathbf{x},\mathbf{y}\}$ on the manifold, one can actually recover the Riemann curvature tensor from the sectional curvature (Kühnel 2015).

Ricci curvature tensor

Another way to express the curvature is by taking the trace of the Riemann curvature tensor. This defines the Ricci curvature:

Definition 2.3.3 Ricci curvature
Let $(\mathcal{M}, g)$ be a Riemannian manifold with R its Riemann curvature tensor. The **Ricci curvature** is defined as:

$$\mathrm{Rc}(\mathbf{x},\mathbf{y}) = \sum_i \langle \mathrm{R}(\mathbf{e}_i,\mathbf{x};\mathbf{y}),\mathbf{e}_i\rangle$$

Along a unit vector $\mathbf{v}$, the Ricci curvature is linked to the sectional curvature: $\mathrm{Rc}(\mathbf{v},\mathbf{v}) = \sum_i \langle \mathrm{R}(\mathbf{e}_i,\mathbf{v};\mathbf{v}),\mathbf{e}_i\rangle = \sum_i \mathrm{K}(\mathbf{v},\mathbf{e}_i)$.

The Ricci curvature is also used to compute the Jacobi field, a vector field along a geodesic γ. We can think of the Jacobi fields as the infinite difference of the volume δV of a unit ball along γ, enclosed by geodesics abusively noted $\gamma + \delta\gamma$. Then, the Ricci curvature appears in the ordinary differential equation to give us the evolution of the volume of a ball along the geodesic γ:

$$\delta\ddot{V}(t) + \mathrm{Rc}(\dot\gamma_t,\dot\gamma_t)\delta V(t) = 0.$$

Scalar curvature

The scalar curvature is the contraction of the Ricci curvature tensor.

Definition 2.3.4 Scalar curvature
Let $(\mathcal{M}, g)$ be a Riemannian manifold with Rc its Ricci curvature tensor. The **scalar curvature** is defined as:

$$\mathrm{S} = \sum_i \mathrm{Rc}(\mathbf{e}_i,\mathbf{e}_i)$$

The scalar curvature can be written with respect to the sectional curvature: $\mathrm{S} = \sum_{i\neq j} \mathrm{K}(\mathbf{e}_i,\mathbf{e}_j)$. The Ricci curvature along a vector $\mathbf{v}$ is equal to the average of all the sectional curvatures spanned by planes $\{\mathbf{v},\mathbf{e}_i\}$, with $\{\mathbf{e}_1,\cdots,\mathbf{e}_n\}$ an orthonormal basis for the tangent plane of the manifold. The scalar curvature represents the average of all the sectional curvatures spanned by the pairs $\{\mathbf{e}_i,\mathbf{e}_j\}$. For a surface in $\mathbb{R}^3$, we actually have: $\mathrm{S} = 2\mathrm{K}(\mathbf{e}_1,\mathbf{e}_2)$

More intuitively, for each point of the manifold, the scalar curvature approximates the difference between the Riemannian volume and the Euclidean volume of a ball:

Proposition 2.3.1 Scalar curvature (asymptotic expansion of the volume of

geodesic balls)
Let $(\mathcal{M}, g)$ be a d-dimensional Riemannian manifold. The **scalar curvature** S is related to the asymptotic expansion of the Euclidean volume, denoted vol, of a geodesic ball on the manifold $\mathcal{B}_g(r)$ compared to the volume of the ball in the Euclidean space $\mathcal{B}_e(r)$, when the radius r tends to 0.

$$\text{vol}(\mathcal{B}_g(r)) = \text{vol}(\mathcal{B}_e(r)) \left(1 - \frac{\text{S}}{6(q+2)} r^2 + o(r^2)\right)$$

Proof. The proof, non-trivial, can be found in Loveridge (2004). □

Example 2.3.1 Scalar curvature for surfaces in $\mathbb{R}^3$
Let us have a look at different surfaces in $\mathbb{R}^3$ and their scalar curvature, represented in Figure 2.5.

1. The **sphere** parameterized with: $f(\mathbf{u}, \mathbf{v}) = r(\sin \mathbf{u} \cos \mathbf{v}, \sin \mathbf{u} \sin \mathbf{v}, \cos \mathbf{u})$, with $\mathbf{u}$ the polar angle and $\mathbf{v}$ the azimuth angle. The scalar curvature is: $\text{S} = 2r^{-2}$.
2. The **torus** parameterized with: $f(\mathbf{u}, \mathbf{v}) = ((R + r\cos \mathbf{v}) \cos \mathbf{u}, (R + r \cos \mathbf{v}) \sin \mathbf{u}, r \sin \mathbf{v})$, with R is the distance from the center of the tube to the center of the torus, r is the radius of the tube, $\mathbf{u}$ is the angle around the center of the torus, and $\mathbf{v}$ is the angle around the tube. The scalar curvature is $\text{S} = \frac{2 \cos \mathbf{v}}{r(R + r \cos \mathbf{v})}$
3. The **hyperboloid** parameterized with: $f(\mathbf{u}, \mathbf{v}) = (\cosh \mathbf{u} \cos \mathbf{v}, \cosh \mathbf{u} \sin \mathbf{v}, \sinh \mathbf{u})$, with $\mathbf{u}$ the hyperbolic angle and v the azimuth angle. The scalar curvature is $\text{S} = -2(\cosh(2\mathbf{u}))^{-2}$.
4. A surface is parameterized by $f(\mathbf{u}, \mathbf{v}) = (\mathbf{u}, \mathbf{v}, \mathbf{z})$, with $\mathbf{z} = e^{-c\mathbf{u}} \sin(\mathbf{u}) \sin(\mathbf{v})$. The scalar curvature can be computed, and is: $\text{S} = \frac{(c^2-1)\cos(2\mathbf{u}) - \cos(2\mathbf{v}) - c(c - 2\sin(2\mathbf{u}))}{e^{2c\mathbf{u}} + \cos(\mathbf{v}) \sin(\mathbf{u})^2 + (\cos(\mathbf{u}) - c\sin(\mathbf{u}))\sin(\mathbf{v})^2}$.

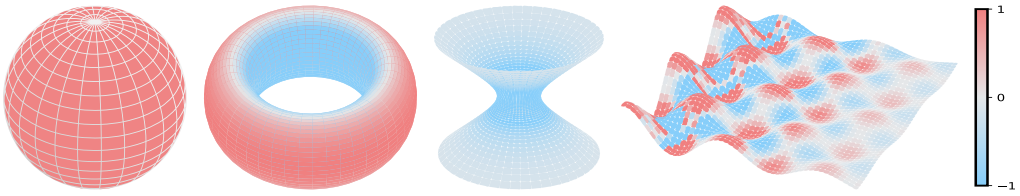


Figure 2.5: The scalar curvature is plotted for different surfaces from the Example 2.3.1 in $\mathbb{R}^3$. The sphere has always a positive scalar curvature, the hyperboloid one is always negative. For every surface, we can see that the scalar curvature is positive when the surface seems convex, and it is negative when we are at a saddle point. In those figures, the values of the scalar curvature have been normalized between -1 and 1.

2.4 Summary: On the curvature of the loss landscape

Generalization properties and flatness

The model's ability to generalize well to unseen data may be linked to the flatness of its loss landscape. The *flatness hypothesis* is based on the observation that flat minima in the loss function, characterized by wide and shallow dips, promote better generalization by allowing the use of less precise weights, thereby enhancing the model's robustness. This hypothesis is supported by extensive empirical and theoretical evidence from various studies that have explored different aspects of machine learning. However, within the community, the various definition of flatness are not always satisfactory.

In this paper, we treat the loss landscape as a Riemannian manifold and utilize tools from Riemannian geometry to investigate its curvature. We found that at minima, the scalar curvature has a straightforward expression related to the norm of the Hessian. While the norm itself does not always accurately defines flatness, it remains a valuable measure for understanding optimization. Our results suggest that the scalar curvature aligns with all the results involving the norm of the Hessian and suggesting that flatter minima improve generalization, while defining rigorously a measure of flatness.

Metric of the parameters space

The loss function f of a model is a smooth scalar-valued function defined on the parameter space $\mathcal{M} \subset \mathbb{R}^q$, where q is the number of parameters. In order to study the loss landscape of a model, we can look at the geometry of the graph of the loss function, a hypersurface embedded in $\mathbb{R}^{q+1}$.

Definition 2.4.1 Metric of a graph
Let $f : \mathcal{M} \subset \mathbb{R}^q \to \mathbb{R}$ be a smooth function. We call **graph of a function** the set:

$$\Gamma_f = \{(\mathbf{x}, y) \in \mathcal{M} \times \mathbb{R} \mid y = f(\mathbf{x})\}.$$

The graph Γ_f is a topological smooth manifold embedded in $\mathbb{R}^{q+1}$, and it is isometric to the Riemannian manifold $(\mathcal{M}, g)$ with $\mathcal{M} \subset \mathbb{R}^q$ and the induced Riemannian metric tensor:

$$\mathbf{G} = \mathbf{I}_q + \nabla f \nabla f^\top, \tag{2.1}$$

Instead of working in the ambient space $\mathbb{R}^{q+1}$, it is more convenient to study the intrinsic geometry of the loss function in the parameter space $(\mathcal{M}, g)$. We denote ∇ the Euclidean gradient operator of the loss function f, and $\mathbf{H}$ the Euclidean Hessian of f.

From this metric, we can define the scalar curvature:

Proposition 2.4.1 Scalar curvature
The scalar curvature is given by:

$$\mathrm{S} = \beta \left(\mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H}^2)\right) + 2\beta^2 \left(\nabla f^\top (\mathbf{H}^2 - \mathrm{tr}(\mathbf{H})\mathbf{H}) \nabla f\right), \tag{2.2}$$

with $\beta = (1 + \|\nabla f\|^2)^{-1}$.

Proof. See Appendix A.2. □

This expression simplifies when the gradient is zero, which corresponds to a critical point of the loss function. In this case, the scalar curvature is given by:

Corollary 2.4.2
When an extremum is reached ($\nabla f = 0$), the scalar curvature becomes:

$$\begin{aligned} \mathrm{S}(\mathbf{x}_{\min}) &= \mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H}^2) \\ &= \|\mathbf{H}\|_*^2 - \|\mathbf{H}\|_F^2, \end{aligned}$$

with $\|\cdot\|_*$ the nuclear norm and $\|\cdot\|_F$ the Frobenius norm.

Proof. This is a direct result of Proposition 2.4.1, when $\nabla f = 0$. □

Comparison with the norm of the Hessian

In this work, we have focused on comparing the scalar curvature with the norm of the Hessian, often used as a flatness measure. Our findings reveal that the trace of the Hessian matrix does not accurately assess flatness in some scenarios. Yet, it remains a valuable tool for demonstrating convergence results in optimization. While the scalar curvature serves as a better measure of flatness, it can reduce to the norm of the Hessian and aligns with previous findings in optimization.

The scalar curvature, norm of the Hessian and flatness

The accuracy of the trace of the Hessian in representing flatness can be compromised for two main reasons. Firstly, a zero trace of the Hessian does not necessarily indicate a flat region, as it could also signify a saddle point where there is a combination of positive and negative eigenvalues. In contrast, at a saddle point, the scalar curvature will be negative, but not zero. Secondly, when the dataset is divided into smaller batches, averaging the flatness over these batches can result in incorrect assessments of the overall flatness. Improper partitioning of the dataset may lead to each batch having a zero trace of the Hessian, falsely indicating that the entire dataset is flat. This is not the case with the scalar curvature. Therefore, relying solely on the trace of the Hessian may lead to misinterpretations of the flatness of a region.

The scalar curvature, norm of the Hessian and optimization

The norm of the Hessian is still associated with the optimization process. Notably, at a convex minimum, if $0 \leq \mathrm{S}(\mathbf{x}_1) \leq \mathrm{S}(\mathbf{x}_2)$, then $\mathrm{tr}(\mathbf{H}(\mathbf{x}_1)^2) \leq \mathrm{tr}(\mathbf{H}(\mathbf{x}_2)^2)$ and $\mathrm{tr}(\mathbf{H}(\mathbf{x}_1))^2 \leq \mathrm{tr}(\mathbf{H}(\mathbf{x}_2))^2$.

Proposition 2.4.3
Let $\mathbf{x}_{\min}$ an extremum, ε, a small scalar ($\varepsilon \ll 1$) and $\mathbf{x}$ a normalized vector ($\|\mathbf{x}\| = 1$). The trace of the square of the Hessian is an upper bound to the difference of the loss functions when perturbed by the weights:

$$\|f(\mathbf{x}_{\min} + \varepsilon\mathbf{x}) - f(\mathbf{x}_{\min})\|_2^2 \leq \frac{1}{4}\varepsilon^4 \,\mathrm{tr}(\mathbf{H}_{\min}^2) \tag{2.3}$$

Proof. This is obtained by applying the Taylor expansion, for a very small perturbation $\varepsilon \ll 1$. See Appendix A.2 for the full proof. □

Proposition 2.4.4
Zhu et al. (2018) used this definition and the expression of the gradient descent process (Equation 5.10) to approximate the escaping efficiency: $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}_{\min})] \approx \frac{t}{2}\,\mathrm{tr}(\mathbf{H}^2)$.

Proof. See Zhu et al. (2018) for the full proof. □

Proposition 2.4.5
We note $\mathbf{H}$ the Hessian of the loss of a model with q parameters, and S the scalar curvature, obtained in Proposition 2.4.1 and Corollary 5.2.4. When we reach a flat minimum, supposing the eigenvalues of $\mathbf{H}$ are similar, for a high number of parameters q, we have:

$$\mathrm{S}(\mathbf{x}_{\min}) \underset{q\to\infty}{\sim} \mathrm{tr}(\mathbf{H})^2$$

Proof. Let us suppose that, at a flat minimum, all the eigenvalues are similar: $\lambda_1 = \cdots = \lambda_q = \lambda \geq 0$. Then we, have $\|\mathbf{H}\|_*^2 = q^2\lambda^2$ and $\|\mathbf{H}\|_F^2 = q\lambda^2$. When the number of parameters increases, $\|\mathbf{H}\|_F^2 = o(\|\mathbf{H}\|_*^2)$, and as a consequence $\|\mathbf{H}\|_*^2 - \|\mathbf{H}\|_F^2 \sim \|\mathbf{H}\|_*^2$. □

In this proposition, we assume that all the eigenvalues are similar, hypothesis that is supported by empirical results (Ghorbani et al. 2019) when the number of parameters increases.

Conclusion

Our research analyzed the loss landscape as a Riemannian manifold and its connection to optimization generalization. We examined the scalar curvatures of the loss landscape, specifically at minima, which are defined as the difference between the nuclear and Frobenius norm of the Hessian of the loss function.

Motivated by the flatness hypothesis, our study explored the relationship between flat minima and better generalization. While the commonly used Hessian norm may not effectively gauge flatness in certain scenarios, the scalar curvature accurately captures flatness while still satisfying the characteristics associated with the Hessian norm, such as stability against perturbations and convergence of the optimization algorithm.

Future research could investigate the curvature within stochastic optimization and explore the scalar curvature as a random variable influenced by the underlying data and batch distribution. Understanding the relationship between the scalar curvature and the stochastic process, as well as its connection to implicit regularization in the model, would also be valuable.

Finsler geometry 3

> "Riemannian geometry studies spaces with only black and white colors, while Finsler geometry studies a colorful world. "
>
> - Shen and Shen (2016)

Finsler geometry is an extension of Riemannian geometry, where the metric is defined as a more or less restrictive norm instead of an inner product. Chern, who made numerous contributions to the field, also nicely summarized that "Finsler geometry is Riemannian geometry without the quadratic assumption" (Chern 1996). In short, using a Finsler metric allows us to extend our geometrical tools to explore spaces that could not be studied with the classic inner product while keeping our knowledge close to Riemannian geometry.

In Section 3.1, we will define the Finsler metric. We will then study this metric as a solution of an historical example in Section 3.2. In Section 3.3, we will compare and highlight the differences between Riemannian and Finsler geometry. Finally, in Section 3.4, we will talk about our research on stochastic manifolds, where we defined the expected lengths as a Finsler metric.

3.1 Definition

Initially considered by Riemann (1867) himself, Finsler geometry was set aside for half a century due to the unclear geometrical interpretations of this generalized metric, coupled with the challenges involved in performing calculations. Surprisingly, it was not in geometry but in the field of calculus of variations that Hilbert revived this problem as one of the most significant unsolved mathematical questions of the time (Hilbert et al. 1900). And this research interested Cathedory, who advised his doctoral student, Finsler, to work on curves and surfaces in general spaces.

The simplest problem in the calculus of variations is to maximize or minimize a functional L modelled by a function F that depends on the curve $\gamma(t)$ and its derivative $\dot{\gamma}(t)$:

$$L = \int F(t, \gamma(t), \dot{\gamma}(t))dt.$$

In geometry, and often in calculus of variations, the functional describes a length to be minimized. To ensure the existence of extrema, the integrand F must satisfy some conditions that are later used to define a Finsler metric. As we are interested in the length of a curve, the integrand F describes a norm defined at each point $\gamma(t)$ of the manifold. We will drop the argument γ, and denote F as $F_t(\dot{\gamma}(t))$ or equivalently $F_x(\mathbf{v})$ on a point x and a vector $\mathbf{v}$.

By definition, the distance between two points is a positive quantity that is null only when those points are equal. This translates with the integrand F being **positive definite**: $\forall\gamma,\ F_t(\dot{\gamma}(t)) \geq 0$ with equality if and only if $\dot{\gamma}(t) = 0$. As we want to find the minimum of the length, the functional should be at least twice differentiable, and the integrand should be at least C^1. We will usually assume that F is C^∞, and so, that it is **smooth**.

We also require the length of any given curve to remain unchanged irrespective of its reparameterization. This is a vital property in differential geometry, echoing the idea that the choice of basis should have no influence. Let's have, for example, a curve γ,

parameterized in Cartesian coordinates, and γ', the same curve but parameterized in polar coordinates, with $\gamma' = \varphi(\gamma)$, and φ is a diffeomorphism mapping one basis to another. Then, we require that $L(\gamma) = L(\gamma')$. L is invariant under reparameterization if and only if F is **homogeneous** of degree 1. In other words, to ensure that the length of a curve is independent of its parameterization, we require the integrand F to satisfy: $\forall \lambda > 0,\ F(\lambda \mathbf{v}) = \lambda F(\mathbf{v})$.

So far, the previous criteria helped us to produce a set of candidate functions that satisfy the necessary conditions for an extremum. But this extremum is not guaranteed to actually be the minimum of the length: it can also be a maximum or a saddle point. In the calculus of variations, the **Legendre criterion** is a second-order condition used to determine whether a solution is a minimum. The condition states that the determinant of the Hessian matrix of the integrand must be positive at the extremum for it to be a minimum. Here is a twist, we will not minimize the length functional, which has an infinity of local minimum since it is invariant under reparameterization. Instead, we will minimize the energy functional $E = \frac{1}{2} \int F_t^2(\dot{\gamma}(t))dt$, that is the lower bound of the squared length, thanks to Cauchy-Schwarz: $2E \leq L^2$, with equality at the minimum. The Legendre condition is satisfied if:

$$\frac{1}{2}\partial_i \partial_j F^2(\dot{\gamma}(t)) \quad \text{is positive definite.}$$

In Finsler geometry, the Legendre condition is renamed the *strong convexity criterion*. The Hessian $g_{ij}(\mathbf{v}) = \frac{1}{2}\partial_i \partial_j F^2(\mathbf{v})$, with $\partial_i = \frac{\partial}{\partial v^i}$, is also called the fundamental tensor, and is noted g like the Riemannian metric since it also needs to be positive definite. We can now state the definition of a Finsler metric:

Definition 3.1.1 Finsler metric
Let $F : \mathcal{TM} \to \mathbb{R}_+$ be a function defined on the tangent bundle $\mathcal{TM}$ of a differentiable manifold $\mathcal{M}$. We say that F is a **Finsler metric** if, for each point x of $\mathcal{M}$ and for any vector $\mathbf{v}$ on $\mathcal{T}_x\mathcal{M}$, the function F satisfies the following conditions:

1. Positive: $F(\mathbf{v}) \geq 0$, with equality if and only if $\mathbf{v} = 0$.
2. Smooth: F is a C^∞ function on the tangent bundle $\mathcal{TM} \setminus \{0\}$.
3. Positive homogeneous: $\forall \lambda \in \mathbb{R}_+,\ F(\lambda \mathbf{v}) = \lambda F(\mathbf{v})$.
4. Strong convexity criterion, or Legendre condition: the fundamental tensor field g defined as the Hessian of $\frac{1}{2}F(\cdot)^2$ is positive definite for $\mathbf{v}$ away from 0.

A differentiable manifold $\mathcal{M}$ equipped with a Finsler metric is called a **Finsler manifold**.

With this definition in mind, let's take a moment to discuss a few observations about the metric, particularly noting how it almost defines a norm.

The Finsler metric is less restrictive than a norm, since we require positive homogeneity instead of absolute homogeneity. This means that F does not need to be symmetric: for any point x in $\mathcal{M}$ and for a vector $\mathbf{v}$ in $\mathcal{T}_x\mathcal{M}$, we can have $F_x(\mathbf{v}) \neq F_x(-\mathbf{v})$. If F is symmetric, the metric is called *reversed*. This asymmetric property becomes incredibly relevant when working with anisotropic media, which happens in biology and physics (Antonelli and Miron 2013) and even when studying wildfire (Markvorsen 2016). We will also see that the Finsler metric is a solution to a navigation problem, described in Section 3.2.

The Finsler metric is also more restrictive than a norm, since it satisfies the Legendre condition, a stronger criterion than the triangle inequality. A norm that is smooth, positive homogeneous, and satisfies the strong convexity criterion is called Minkowski, and often, we would say that the Finsler metric is a *Minkowski norm*.

Adopting the nomenclature of Javaloyes and Sánchez (2011), a Finsler metric that relaxes the Legendre condition for the triangle inequality is called *pseudo-Finsler*, and a Finsler metric that is well-defined on a certain conic domain (open, non-empty, and if $\mathbf{v} \in A$, then $\lambda\mathbf{v} \in A$) instead of the tangent bundle is called a *conic Finsler* metric.

Finally, to prove that a metric satisfies the Legendre condition, we only need its fundamental tensor to be definite, instead of being positive definite, since it is already positive, as shown by Lovas (2007).

3.2 Motivation: Zermelo's navigation problem

> "In an unbounded plane where the wind distribution is given by a vector field as a function of time and position, a ship moves at a constant speed relative to the surrounding air mass. How must the ship be steered in order to come from a starting point to a given destination in the shortest time?"
>
> - Zermelo (1931)

In Prague, Zermelo (1931) gave a lecture on the solution of what is now called the Zermelo navigation problem. This problem, standard in theory of control, has been solved in the plane by Zermelo himself. Bao et al. (2004) solved it when the plane becomes a manifold.

Imagine a ship sailing in a calm sea without any current or wind. The trajectory of the journey can be obtained using a Riemannian metric, where, in the most simple case, it would be a geodesic on a 2-sphere (As studied in Example 2.1.3). When a wind or a current comes into play, the time required to travel in one direction is different from the time needed to travel in the opposite direction. Without wind, the time taken by our ship to traverse the direction $\mathbf{u}$ is $|\mathbf{u}|^2 = h(\mathbf{u}, \mathbf{u})$, with h a Riemannian metric. This metric will be perturbed by the wind, whose velocity can be described by a vector field, $\mathbf{w}$. For a unit of time, we assume that the ship used to traverse $\mathbf{u}$ without any hindrance. Now, the wind will modify its course, and, again for a unit of time, it will now traverse $\mathbf{u} + \mathbf{w} = \mathbf{v}/F(\mathbf{v})$, with $F(\cdot)$ a norm to define.

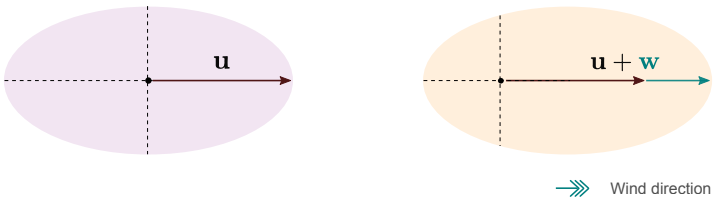


Figure 3.1: On the left, we have a Riemannian indicatrix, representing the length crossed by a ship in a unit of time, without any wind. **On the right**, we have a Finsler indicatrix, representing the length crossed by the ship in a unit of time, with a wind component.

We want to find F such that $|\mathbf{v}/F(\mathbf{v}) - \mathbf{w}|^2 = \mathbf{u}^2$. Using the definition of the inner product, the equation becomes: $|\mathbf{v}|^2/F^2(\mathbf{v}) - 2h(\mathbf{v}, \mathbf{w})/F(\mathbf{v}) + |\mathbf{w}|^2 - |\mathbf{u}|^2 = 0$. Multiplying by $-F^2(\mathbf{v})$, we obtain:

$$(|\mathbf{u}|^2 - |\mathbf{w}|^2)F^2(\mathbf{v}) + 2h(\mathbf{v}, \mathbf{w})F(\mathbf{v}) - |\mathbf{v}|^2 = 0.$$

This second order equation has a solution, if the discriminant is positive, which always happens when $|\mathbf{w}| < |\mathbf{u}|$. Bao et al. (2004) assumed that $|\mathbf{u}|^2 = 1$, and noted $\lambda = |\mathbf{u}|^2 - |\mathbf{w}|^2$. Then, using those notations, the positive root of the equation is obtained with:

$$F(\mathbf{v}) = \frac{\sqrt{h(\mathbf{v}, \mathbf{w})^2 + \lambda|\mathbf{v}|^2}}{\lambda} - \frac{h(\mathbf{v}, \mathbf{w})}{\lambda}$$

This norm satisfies all the conditions to be a Finsler metric. It is also asymmetric: $F(\mathbf{v}) \neq F(-\mathbf{v})$. This Finsler metric actually belongs to a specific type of metric called the *Randers* metric. It was first introduced by Randers (1941) to describe space-time in general relativity. The function F is written as $F(\mathbf{v}) = \alpha(\mathbf{v}) + \beta(\mathbf{v})$, with $\alpha(\mathbf{v}) = \sqrt{a_{ij}v^i v^j}$ a Riemannian metric, $\beta(\mathbf{v}) = b_i v^i$ a function such that b is a one-form, with $|b| < 1$.

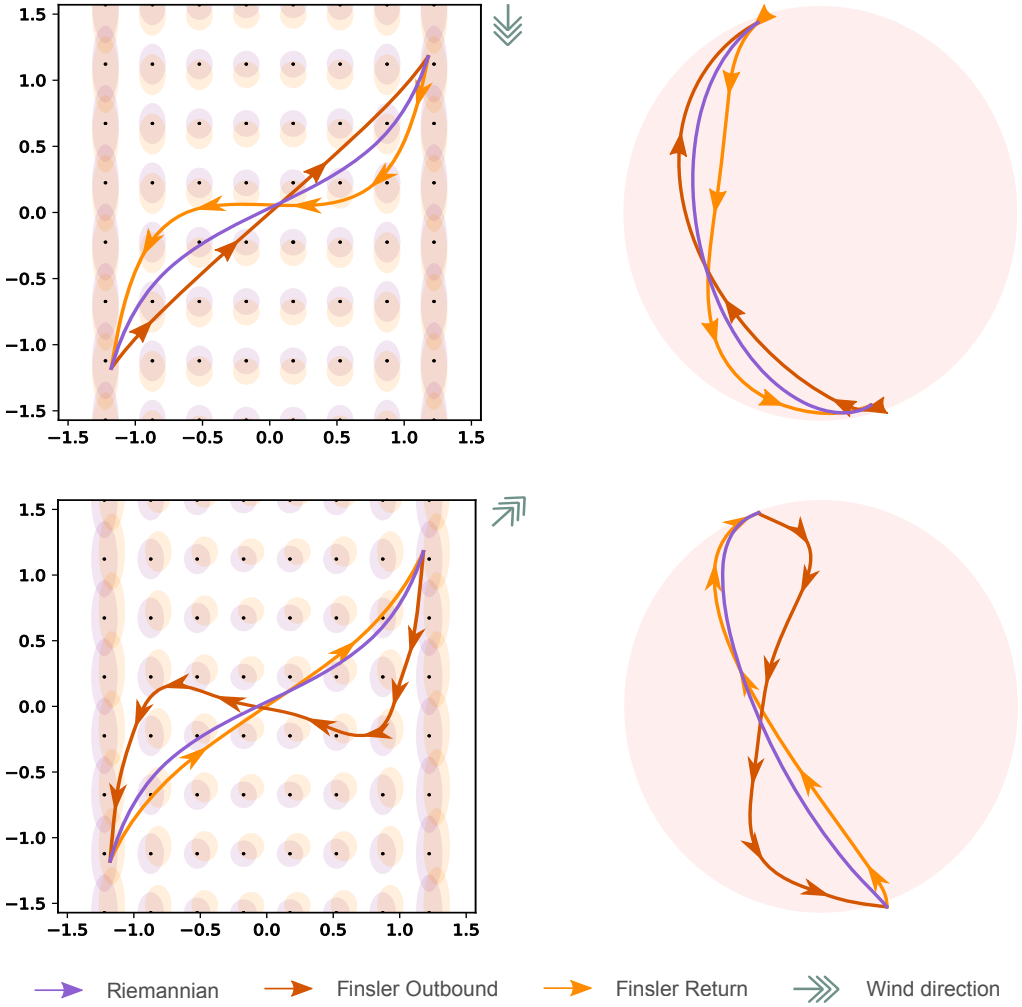


Figure 3.2: The figure represents three geodesics: the Riemannian, the Finsler *inbound* and the Finsler *outbound* geodesics, for two different wind directions on the sphere. The Finsler indicatrices (orange) are slightly shifted compared to the Riemannian indicatrices (purple), due to the wind. From the center of the indicatrices, the geodesics are looking to cover the longest distance, and as a consequence, the paths will be different for the inbound and outbound.

3.3 Finsler versus Riemann

To better compare Riemannian and Finsler geometry, we will change the usual notations and define the fundamental tensor $\mathbf{G}^{\mathbf{v}}_{ij}$ the Cartan tensor $\mathbf{C}^{\mathbf{v}}_{ijk}$, and their respective bilinear and trilinear form $g^{\mathbf{v}}$ and $C^{\mathbf{v}}$:

$$\mathbf{G}^{\mathbf{v}}_{ij}(x,\mathbf{v}) = \frac{1}{2}\partial_{\mathbf{v}^i}\partial_{\mathbf{v}^j}F^2_x(\mathbf{v}) \quad \text{and} \quad g^{\mathbf{v}}(\mathbf{x},\mathbf{y}) = \mathbf{x}^i\mathbf{y}^j g_{ij}(x,\mathbf{v})$$
$$\mathbf{C}^{\mathbf{v}}_{ijk}(x,\mathbf{v}) = \frac{1}{4}\partial_{\mathbf{v}^i}\partial_{\mathbf{v}^j}\partial_{\mathbf{v}^k}F^2_x(\mathbf{v}) \quad \text{and} \quad C^{\mathbf{v}}(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathbf{x}^i\mathbf{y}^j\mathbf{z}^k C_{ijk}(x,\mathbf{v})$$

We note $g^{\mathbf{v}}$ the Riemannian metric obtained with the fundamental tensor $\mathbf{G}^{\mathbf{v}}$, to not be confused with g the Riemannian metric associated with its tensor $\mathbf{G}$. While the Riemannian metric tensor $\mathbf{G}(x)$ only depends on the position $x \in \mathcal{M}$, the fundamental tensor $\mathbf{G}^{\mathbf{v}}(x, \mathbf{v})$ depends on both the position and the direction $\mathbf{v} \in T_x\mathcal{M}$. If the tensor $\mathbf{G}^{\mathbf{v}}$ is independent of the direction $\mathbf{v}$ then, it means that the Finsler metric is quadratic in $\mathbf{v}$ and so, Finsler geometry reduces to Riemannian geometry. Any Finslerian object depends on $\mathbf{v}$, and automatically reduces to their Riemannian counterpart otherwise.

Functionals

Similarly to Riemannian geometry, both the length and energy of the curve are defined with respect to the norm of the metric.

Definition 3.3.1 Curve length and curve energy
Let $\gamma : [a, b] \to \mathcal{M}$ be a smooth curve in a Finsler manifold $(\mathcal{M}, F)$. The curve length and the curve energy are defined as:

$$\mathcal{L}(\gamma) = \int_a^b F(\gamma'(t))dt \quad \text{and} \quad \mathcal{E}(\gamma) = \frac{1}{2}\int_a^b F^2(\gamma'(t))dt$$

By definition, they also share the same properties: the curve length is reparameterization invariant, and the curve energy is related to the curve length by the Cauchy-Schwarz inequality: $\mathcal{L}^2(\gamma) \leq 2\mathcal{E}(\gamma)$.

Volume measures

In Riemannian geometry, the volume form is uniquely defined by the metric. In Finsler geometry, however, there exists a multitude of definitions for the volume form, some being more justified than others, depending on the metric and the topological manifold (Wu 2011). The most common choices of volume measures are the **Busemann-Hausdorff measure** and the **Holmes-Thompson measure**.

The Busemann-Hausdorff measure acts as a scaling factor correcting the deformation of the Euclidean unit ball. In Riemannian geometry for example, the Riemannian volume measure can be defined as:

$$V_{\mathbf{G}} = \frac{\text{vol}(\mathcal{B})}{\text{vol}(\mathcal{E})},$$

where $\mathcal{B}$ is the Euclidean unit ball, $\mathcal{E}$ is an ellipse, and vol the euclidean volume. If λ_i are the eigenvalues of $\mathbf{G}$, the semi-axis of the ellipse are $(\lambda_i)^{-\frac{1}{2}}$. In Finsler geometry, we don't necessarily have an ellipse, but a convex indicatrix. The Busemann-Hausdorff measure is defined as the ratio between the volume of the Euclidean unit ball and the volume of the convex set, denoted $\mathcal{C} = \{\mathbf{v} \in \mathcal{T}_x\mathcal{M} \mid F_x(\mathbf{v}) \leq 1\}$.

Definition 3.3.2 Busemann-Hausdorff measure
Let $(\mathcal{M}, F)$ be a Finsler manifold. The Busemann-Hausdorff measure is defined as:

$$d\mu_{BH} = \frac{\text{vol}(\mathcal{B})}{\text{vol}(\mathcal{C})}$$

where $\mathcal{B}$ is the Euclidean unit ball, with $\mathcal{C} = \{\mathbf{v} \in \mathcal{T}_x\mathcal{M} \mid F_x(\mathbf{v}) \leq 1\}$ the Finsler indicatrix.

In Riemannian geometry, we also have $V_{\mathbf{G}} = \sqrt{\det \mathbf{G}}$. Similarly, the Holmes-Thompson measure involves taking the determinant of the fundamental tensor:

Definition 3.3.3 Holmes-Thompson measure
Let $(\mathcal{M}, F)$ be a Finsler manifold. The Holmes-Thompson measure is defined as:

$$d\mu_{HT} = \sqrt{\det(\mathbf{G}^{\mathbf{v}})}$$

where $\mathbf{G}^{\mathbf{v}}$ is the fundamental tensor of the Finsler metric.

While those two volume measures don't agree with each other, when the Finsler metric is a Riemannian metric, they both agree with the Riemannian volume measure.

Connections

In order to study geodesics and curvature, we need a way to compare vectors at different points on the manifold. And so, just like in Riemannian geometry, we need to define a connection. However, unlike Riemannian geometry, there is no canonical connection associated with a Finsler metric. Instead, there are several types of connections that can be constructed on a specific vector bundle of a Finsler manifold. Some examples of these connections are the Chern connection, the Cartan connection, the Berwald connection or the Hashiguchi connection, to name a few. The **Chern connection** is the most common one, as it is a natural generalization of the Levi-Civita connection: it is a linear connection, defined on the pullback bundle, almost compatible with the fundamental metric tensor, and it has zero torsion:

Theorem 3.3.1 Chern connection
Let $(\mathcal{M}, F)$ be a Finsler manifold, and $\mathbf{v}$ a non-vanishing vector field. There exists a unique linear connection $\nabla^{\mathbf{v}}$ on the pullback bundle π^*TM such that:

1. $\nabla^{\mathbf{v}}$ has zero torsion: $\nabla^{\mathbf{v}}_{\mathbf{x}}\mathbf{y} - \nabla^{\mathbf{v}}_{\mathbf{y}}\mathbf{x} = [\mathbf{x}, \mathbf{y}]$,
2. $\nabla^{\mathbf{v}}$ is almost compatible with the metric $g^{\mathbf{v}}$:

$$\mathbf{x} g^{\mathbf{v}}(\mathbf{y}, \mathbf{z}) = g^{\mathbf{v}}(\nabla^{\mathbf{v}}_{\mathbf{x}}\mathbf{y}, \mathbf{z}) + g^{\mathbf{v}}(\mathbf{y}, \nabla^{\mathbf{v}}_{\mathbf{x}}\mathbf{z}) + 2C^{\mathbf{v}}(\nabla^{\mathbf{v}}_{\mathbf{x}}\mathbf{v}, \mathbf{y}, \mathbf{z}),$$

with $\mathbf{x}, \mathbf{y}, \mathbf{z}$ vector fields on $\mathcal{M}$ and $[\mathbf{x}, \mathbf{y}]$ the Lie bracket of $\mathbf{x}$ and $\mathbf{y}$.

Proof. See Chern and Shen (2005, Theorem 2.1.1). □

Using the Chern connection, concepts like geodesics and curvature, which are well-understood in Riemannian geometry, can be generalized within the Finsler framework. We will focus on the similarities without delving into the technical details.

In Riemannian geometry, the Christoffel symbols, coefficients of the Levi-Civita connection, are related to the Riemannian metric. Similarly, in Finsler geometry, the Chern connection coefficients ($\Gamma^{\mathbf{v}}$) are related to the fundamental tensor ($\mathbf{G}^{\mathbf{v}} = \mathbf{g}^{\mathbf{v}}$) of the Finsler metric:

$$\Gamma^{\mathbf{v}i}_{jk} = \frac{1}{2}\mathbf{g}^{\mathbf{v}il}\left(\frac{\delta \mathbf{g}^{\mathbf{v}}_{lj}}{\delta x^k} + \frac{\delta \mathbf{g}^{\mathbf{v}}_{lk}}{\delta x^j} - \frac{\delta \mathbf{g}^{\mathbf{v}}_{jk}}{\delta x^l}\right),$$

where $\delta_{x^j} = \partial_{x^j} - N^i_j \partial_{\mathbf{v}^i}$ is a change of basis with N the non-linear connection, which comes from the fact that the Finslerian objects depend on both the position $x \in \mathcal{M}$ and the direction of the tangent vector $\mathbf{v} \in \mathcal{T}_x\mathcal{M}$, while Riemannian objects only depend on the position $x \in \mathcal{M}$. The Chern connection reduces to the Levi-Civita connection if it is independent of $\mathbf{v}$ as we have $\delta_{x^j} = \partial_{x^j}$.

The geodesic equation of a curve γ is defined as:

$$\frac{\partial^2 \gamma^k}{\partial t^2} + \Gamma^{\mathbf{v}k}_{ij} \frac{\partial \gamma^i}{dt} \frac{\partial \gamma^j}{\partial t} = 0,$$

which is similar to the Riemannian case.

Curvature

The Chern curvature is defined similarly to the Riemann curvature:

Definition 3.3.4 Chern curvature
Let $(\mathcal{M}, F)$ be a Finsler manifold, and $\mathbf{v}$ a non-vanishing vector field. The **Chern curvature** is defined as:

$$\mathrm{R}^{\mathbf{v}}(\mathbf{x}, \mathbf{y}; \mathbf{z}) = \nabla^{\mathbf{v}}_{\mathbf{x}} \nabla^{\mathbf{v}}_{\mathbf{y}} \mathbf{z} - \nabla^{\mathbf{v}}_{\mathbf{y}} \nabla^{\mathbf{v}}_{\mathbf{x}} \mathbf{z} - \nabla^{\mathbf{v}}_{[\mathbf{x},\mathbf{y}]} \mathbf{z}.$$

For a flag $(\mathbf{v}, \mathbf{w})$ consisting of the pole $\mathbf{v} \in \mathcal{T}_x\mathcal{M}$ and a vector $\mathbf{w}$ describing the flag (a 2-plane $P \subset \mathcal{T}_x\mathcal{M}$), the **flag curvature** is defined as:

Definition 3.3.5 Flag curvature
Let $(\mathcal{M}, F)$ be a Finsler manifold, and $\mathbf{v}$ a non-vanishing vector field. The flag curvature is defined as:

$$\mathrm{K}^{\mathbf{v}}(\mathbf{v}, P) = \mathrm{K}^{\mathbf{v}}(\mathbf{v}, \mathbf{w}) = \frac{g^{\mathbf{v}}(R^{\mathbf{v}}(\mathbf{v}, \mathbf{w}; \mathbf{w}), \mathbf{v})}{g^{\mathbf{v}}(\mathbf{v}, \mathbf{v}) g^{\mathbf{v}}(\mathbf{w}, \mathbf{w}) - g^{\mathbf{v}}(\mathbf{v}, \mathbf{w})^2}.$$

When the flag curvature is independent of $\mathbf{v}$, it coincides with the sectional curvature.

Finally, the Ricci curvature is defined as:

Definition 3.3.6 Ricci curvature
Let $(\mathcal{M}, F)$ be a Finsler manifold, and $\mathbf{v}$ a non-vanishing vector field. The **Ricci curvature** is defined as:

$$\mathrm{Rc}^{\mathbf{v}} = \sum_{i=1}^{n} K^{\mathbf{v}}(\mathbf{v}, \mathbf{e}_i),$$

with $\mathbf{e}_i$ an orthonormal basis of $\mathcal{T}_x\mathcal{M}$.

Table 3.1: Comparison of Riemannian and Finslerian geometry

Riemannian	Finslerian	Objects
$\mathcal{M}$	$(\mathcal{M}, \mathcal{T}_x\mathcal{M})$	Domain
inner product $\|\cdot\|_G : \mathbf{v} \to (\mathbf{v}^\top \mathbf{G}\ \mathbf{v})^{\frac{1}{2}}$	Minkowski norm $\|\cdot\|_F : \mathbf{v} \to F_x(\mathbf{v})$	Metric
$\mathbf{G}$ Metric tensor	$\mathbf{G}^{\mathbf{v}} = \frac{1}{2}\,\mathrm{Hess}(F^2)$ Fundamental tensor	Tensors
Ellipsoid $(\mathcal{E})$	Convex set $(\mathcal{C})$	Indicatrix [1]
$\mathcal{L}(\gamma) = \int_0^1 \|\dot\gamma(t)\|\,dt$ $\mathcal{E}(\gamma) = \frac{1}{2}\int_0^1 \|\dot\gamma(t)\|^2\,dt$		Length functional Energy functional
$\sigma_G(x) = \sqrt{\det \mathbf{G}}$ $= \frac{\mathrm{vol}(\mathcal{B})}{\mathrm{vol}(\mathcal{E})}$	$\sigma_{HT}(x) = \frac{1}{\mathrm{vol}(\mathcal{B})}\int_{\mathcal{C}} \det(\mathbf{G}^{\mathbf{v}})d\mathbf{v}$ $\sigma_{BH}(x) = \frac{\mathrm{vol}(\mathcal{B})}{\mathrm{vol}(\mathcal{C})}$	Volume form [2]
Levi-Civita connection $\nabla_{\mathbf{x}}\mathbf{y} = \partial_{\mathbf{x}}\mathbf{y} + x^i y^j \Gamma^k_{ij}\mathbf{e}_k$ $\Gamma^i_{jk} = \frac{1}{2}g^{il}\left(\frac{\partial g_{lj}}{\partial x^k} + \frac{\partial g_{lk}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^l}\right)$	Chern connection $\nabla^{\mathbf{v}}_{\mathbf{x}}\mathbf{y} = \partial_{\mathbf{x}}\mathbf{y} + x^i y^j \Gamma^{\mathbf{v}k}_{ij}\mathbf{e}_k$ $\Gamma^{\mathbf{v}i}_{jk} = \frac{1}{2}\mathbf{g}^{\mathbf{v}il}\left(\frac{\delta \mathbf{g}^{\mathbf{v}}_{lj}}{\delta x^k} + \frac{\delta \mathbf{g}^{\mathbf{v}}_{lk}}{\delta x^j} - \frac{\delta \mathbf{g}^{\mathbf{v}}_{jk}}{\delta x^l}\right)$	Connections and their coefficients [3]
$\mathrm{R}(\mathbf{x},\mathbf{y};\mathbf{z}) =$ $(\nabla_{\mathbf{x}}\nabla_{\mathbf{y}} - \nabla_{\mathbf{y}}\nabla_{\mathbf{x}} - \nabla_{[\mathbf{x},\mathbf{y}]})\mathbf{z}$	$\mathrm{R}^{\mathbf{v}}(\mathbf{x},\mathbf{y};\mathbf{z}) =$ $(\nabla^{\mathbf{v}}_{\mathbf{x}}\nabla^{\mathbf{v}}_{\mathbf{y}} - \nabla^{\mathbf{v}}_{\mathbf{y}}\nabla^{\mathbf{v}}_{\mathbf{x}} - \nabla^{\mathbf{v}}_{[\mathbf{x},\mathbf{y}]})\mathbf{z}$	Riemannian and Chern curvatures
$\mathrm{K}(\mathbf{v},\mathbf{u}) = \frac{g(\mathrm{R}(\mathbf{v},\mathbf{u};\mathbf{u}),\mathbf{v})}{g(\mathbf{v},\mathbf{v})g(\mathbf{u},\mathbf{u}) - g(\mathbf{v},\mathbf{u})^2}$	$\mathrm{K}^{\mathbf{v}}(\mathbf{v},\mathbf{u}) = \frac{g^{\mathbf{v}}(\mathrm{R}^{\mathbf{v}}(\mathbf{v},\mathbf{u};\mathbf{u}),\mathbf{v})}{g^{\mathbf{v}}(\mathbf{v},\mathbf{v})g^{\mathbf{v}}(\mathbf{u},\mathbf{u}) - g^{\mathbf{v}}(\mathbf{v},\mathbf{u})^2}$	Sectional and flag curvatures
$\mathrm{Rc}(\mathbf{v}) = \sum_{i=1}^n \mathrm{K}(\mathbf{v},\mathbf{e}_i)$	$\mathrm{Rc}^{\mathbf{v}}(\mathbf{v}) = \sum_{i=1}^n \mathrm{K}(\mathbf{v},\mathbf{e}_i)$	Ricci curvature

[1] We respectively denote $\mathcal{E} = \{\mathbf{v} \in \mathcal{T}_x\mathcal{M} \mid \|\mathbf{v}\|_G \leq 1\}$ represents an ellipsoid, and $\mathcal{C} = \{\mathbf{v} \in \mathcal{T}_x\mathcal{M} \mid \|\mathbf{v}\|_F \leq 1\}$ represents a convex set. The unit ball, denoted $\mathcal{B}$ represents the set of all unit vectors in the Euclidean space.

[2] We have different definitions of the Finsler volume form including the Holmes-Thompson volume form σ_{HT} and the Busemann-Hausdorff volume form σ_{BH}. In Riemannian geometry, we have a unique volume form σ_G.

[3] We have a unique linear connection that is torsion-free and metric-compatible in Riemannian geometry: the Levi-Civita connection. In Finsler geometry, we have multiple connections. The Chern connection $(\nabla^{\mathbf{v}})$ is the unique linear connection that is torsion-free and almost compatible with the fundamental tensor $\mathbf{G}^{\mathbf{v}}$.

3.4 Summary: Identifying latent distances with Finslerian geometry

In machine learning, generative models are useful to learn low-dimensional latent representations $\mathbf{z}$ of the data $\mathbf{x}$, through a smooth mapping $f : \mathcal{Z} \subset \mathbb{R}^q \to \mathcal{X} \subset \mathbb{R}^d$, with $q \leq d$. To explore this latent space, we equip the manifold with the stochastic pullback metric $\mathbf{G}$ (See Section 2.2).

Manipulating random variables is often unfeasible, and a solution to circumvent this problem has been to take the expectation of the metric tensor: $\mathbb{E}[\mathbf{G}]$. But the geodesics derived from this expected Riemannian metric do not correspond to the expected length-minimizing curves. In other terms, we have:

$$\mathcal{L}(\gamma)|_{\mathbb{E}[\mathbf{G}]} = \int \|\dot{\gamma}_t\|_{\mathbb{E}[\mathbf{G}]}\, dt \qquad \neq \qquad \mathbb{E}\left[\mathcal{L}(\gamma)|_{\mathbf{G}}\right] = \int \mathbb{E}\left[\|\dot{\gamma}_t\|_{\mathbf{G}}\right] dt.$$

Instead of taking the expectation of the metric tensor, which serves as a surrogate to define a norm, we wanted to take the expectation of the norm directly, and compare this new norm with the expected Riemannian norm. This research has been described in details in Pouplin et al. (2022) (Appendix **??**).

In our paper, we found that:

- The expected norm defines a Finsler metric.
- When f is a Gaussian process, the stochastic norm obtained through the pullback metric follows a noncentral Nakagami distribution, so the Finsler metric has a closed-form expression.
- In data space of high dimensions, for Gaussian processes, the Finsler metric and the expected Riemannian metric converge to each other at a rate of $\mathcal{O}(\frac{1}{d})$.

We conclude that, in practice, when working with high-dimensional data, the Riemannian metric serves as a good approximation of the Finsler metric, which is theoretically more grounded. It further justifies the use of the Riemannian metric in practice when exploring stochastic manifolds.

Notations

The paper tries to use consistent symbols for two fields of geometry that have different conventions. In Riemannian geometry, the metric is an inner product which can induce a norm, while in Finsler geometry, we are directly working with norms. We also assume that all our metric are always defined for a specific point p on our manifold $\mathcal{Z}$, and so we will drop this index. The following notations will be used:

Stochastic pullback metric tensor	$\mathbf{G} = \mathbf{J}^\top \mathbf{J}$
Expected Riemannian metric	$g : (\mathbf{u}, \mathbf{v}) \to \mathbf{u}^\top \mathbb{E}[\mathbf{G}]\ \mathbf{v}$
Stochastic pullback induced norm	$\|\cdot\|_{\mathbf{G}} : \mathbf{u} \to \sqrt{\mathbf{u}^\top \mathbf{G}\ \mathbf{u}}$
Expected Riemannian induced norm	$\|\cdot\|_R : \mathbf{u} \to \sqrt{\mathbf{u}^\top \mathbb{E}[\mathbf{G}]\ \mathbf{u}} \coloneqq \sqrt{g(\mathbf{u}, \mathbf{u})} = \|\cdot\|_{\mathbb{E}[\mathbf{G}]}$
Finsler metric	$\|\cdot\|_F : \mathbf{u} \to \mathbb{E}[\sqrt{\mathbf{u}^\top \mathbf{G}\ \mathbf{u}}] \coloneqq F(\mathbf{u}) = \mathbb{E}\left[\|\cdot\|_{\mathbf{G}}\right]$

Results and contributions

Proposition 3.4.1 The expected norm is a Finsler metric
Let $\mathbf{G}$ be a stochastic metric tensor, with $\mathbb{E}[\mathbf{G}]$ its expectation. The norm induced by the

Riemannian metric tensor defined on the manifold $\mathcal{Z}$ is denoted $\|\cdot\|_{\mathbf{G}} : \mathbf{z} \to (\mathbf{z}^\top \mathbf{G}\ \mathbf{z})^{\frac{1}{2}}$.

Then, the expected norm $\|\cdot\|_{\mathbb{E}[\mathbf{G}]} : \mathbf{z} \to \mathbb{E}\left[\mathbf{z}^\top \mathbf{G}\ \mathbf{z})^{\frac{1}{2}}\right]$ is not induced by a Riemannian metric and defines a Finsler norm.

Proof. If F was induced by a Riemannian metric, then this metric would be defined as: $g : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}_+ : (\mathbf{u}, \mathbf{v}) \to \mathbb{E}[\sqrt{\mathbf{u}^\top \mathbf{G} \mathbf{v}}]^2$. A Riemannian metric is an inner product, so it should be bilinear. Here, we can see that g is not linear: $g(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}) \neq g(\mathbf{u}_1, \mathbf{v}) + g(\mathbf{u}_2, \mathbf{v})$, so g is not a Riemannian metric. However, we can prove that $F : \mathbb{R}^q \to \mathbb{R} : \mathbf{v} \to \mathbb{E}[\sqrt{\mathbf{v}^\top \mathbf{G}\ \mathbf{v}}]$ is positive, homogeneous, and smooth. The difficult part is to prove that F satisfies the Legendre condition: $\frac{1}{2}\,\mathrm{Hess}(F^2)$ is positive definite and non-degenerate. We prove it, in the full proof of the paper, See Appendix **??**. □

This proposition states that the expectation of a norm induced by a Riemannian metric is a non-Riemannian Finsler metric. We were then interested in studying the properties of this Finsler metric. When the mapping f is a Gaussian process, the induced norm $\|\cdot\|_{\mathbf{G}}$ follows a non-central Nakagami distribution and its expectation has a closed-form expression.

Proposition 3.4.2 Closed-form expression for Gaussian processes
Let f be a Gaussian process and J its Jacobian, with $J \sim \mathcal{N}(\mathbb{E}[J], \Sigma)$. The Finsler norm can be written as:

$$F_z : \mathcal{T}_z\mathcal{Z} \to \mathbb{R}_+ : \|v\|_F \coloneqq v \to \sqrt{2}\sqrt{v^\top \Sigma v}\frac{\Gamma(\frac{d}{2} + \frac{1}{2})}{\Gamma(\frac{d}{2})}{}_1F_1\left(-\frac{1}{2}, \frac{d}{2}, -\frac{\omega}{2}\right),$$

with ${}_1F_1$ the confluent hypergeometric function of the first kind and ω, a noncentrality term with: $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J] v)$.

Proof. We suppose that f is a Gaussian process, and so is its Jacobian. $\mathbf{G}$ follows a non-central Wishart distribution: $\mathbf{G} = J^\top J \sim \mathcal{W}_q(d, \Sigma, \Sigma^{-1}\mathbb{E}[J]^\top \mathbb{E}[J])$. $\mathbf{v}^\top \mathbf{G}\ \mathbf{v}$ is a scalar and also follows a non-central Wishart distribution: $\mathbf{v}^\top \mathbf{G}\ \mathbf{v} \sim \mathcal{W}_1(d, \sigma, \omega)$, with $\sigma = \mathbf{v}^\top \Sigma\ \mathbf{v}$ and $\omega = (\mathbf{v}^\top \Sigma\ \mathbf{v})^{-1}(\mathbf{v}^\top \mathbb{E}[J]^\top \mathbb{E}[J]\ \mathbf{v})$ (Kent and Muirhead (1984, Definition 10.3.1)). The square-root of a non-central Wishart distribution follows a non-central Nakagami distribution (Hauberg 2018b). Then, by construction, the stochastic norm $\|\cdot\|_{\mathbf{G}}$ follows a non-central Nakagami distribution. The expectation of this distribution is known, and it has a closed-form expression. □

Proposition 3.4.3 Convergence of the Riemannian and Finslerian norms
Let f be a Gaussian Process. In high dimensions, the relative ratio between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$\frac{\|v\|_R - \|v\|_F}{\|v\|_R} = \mathcal{O}\left(\frac{1}{d}\right)$$

And, when d tends to infinity: $\forall v \in \mathcal{T}_z\mathcal{Z},\ \|v\|_R \underset{+\infty}{\sim} \|v\|_F$.

Proof. We arrive at this result by bounding the relative difference between the two norms, using a sharper version of the Jensen's inequality (Liao and Berg 2019). The upper bound is then expressed with respect to the non-central Nakagami distribution parameters. We also assume the latent manifold to be bounded, so the quantities do not diverge towards infinity. The full proof can be found in Appendix **??**. □

Experiments

We want to illustrate cases where these metrics differ in practice. In one case, we use synthetic data of low dimension and force the geodesics to go through region of high variance, and we use a toy dataset, FashioMNIST, of higher dimensions. The latent representations are learned with a GPLVM.

Experiments with synthetic data showing high variance

In those examples, we are plotting curves that does not follow the data points and so, go through regions of high variance. Notably, the background of the latent manifold represents the variance of the posterior distribution in logarithmic scale.

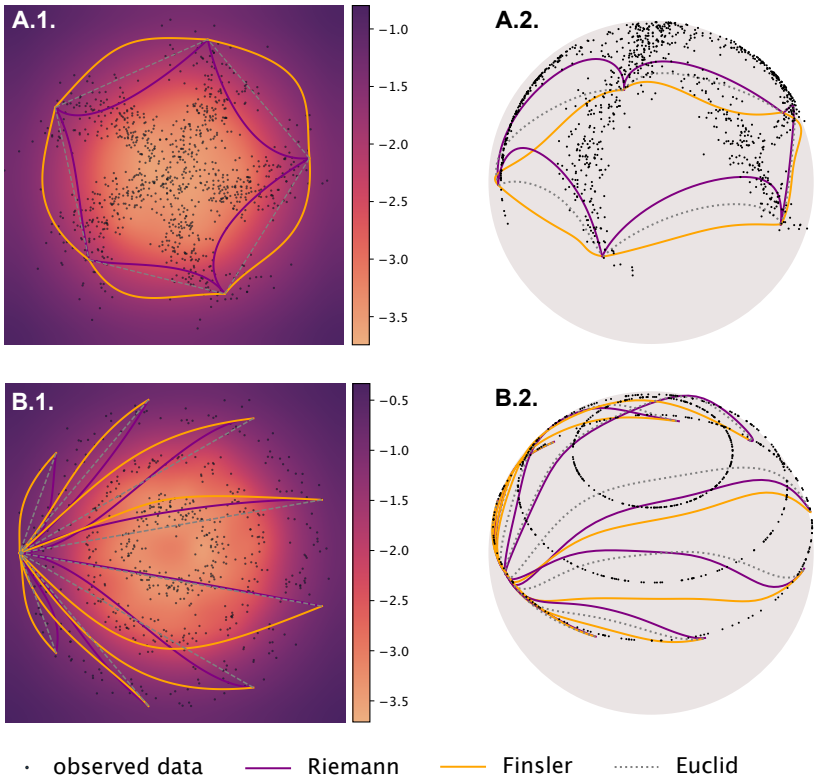


Figure 3.3: The Riemannian (purple), Finslerian (orange) and Euclidean (dotted gray) geodesics obtained by pulling back the metric through the Gaussian processes of a trained GPLVM. The models, trained on the 3-dimensional synthetic data (Figures A.2, B.2) learned their latent representations (Figures A.1, B.1)

The Riemannian geodesics avoid area of high variance, at the cost of following longer paths. The Finsler geodesics are not perturbed by the variance term, and will explore regions without any data, following shorter paths. This can be explained as the variance term acts as a regularizer between the Riemannian and Finslerian energy:

$$\mathcal{E}_R(\gamma) = \mathcal{E}_F(\gamma) + \int_0^1 \mathrm{var}(\|\dot{\gamma}_t\|_{\mathbf{G}})dt$$

Experiments with FashionMNIST

In these experiments, because the difference of the variance of the posterior learned by the GPLVM is low and because the data is high-dimensional, we cannot see any difference

between the Finslerian and the Riemannian geodesics. In practice, the Riemannian metric is a good approximation of the Finslerian metric.

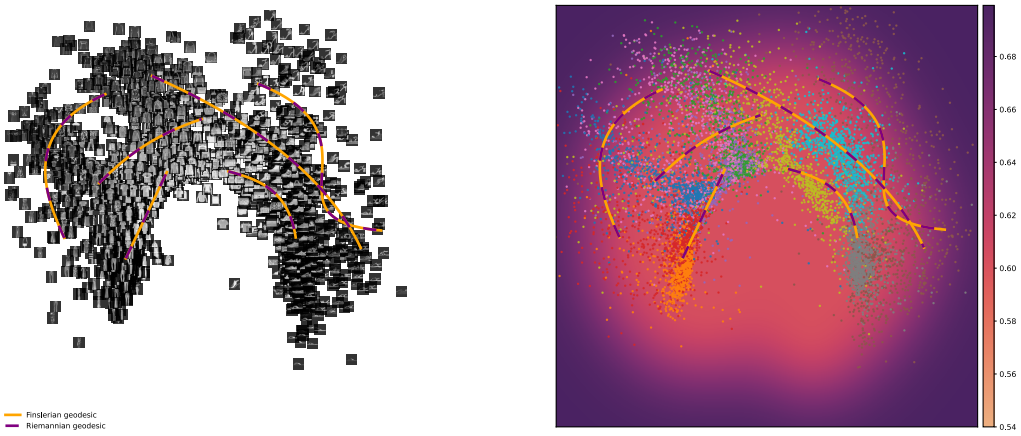


Figure 3.4: The Riemannian (purple) and the Finslerian (orange) geodesics are plotted in the learned latent space with the images (left) and data points (right) of FashionMNIST.

A note on the asymptotic behavior of Wishart matrices and a conjecture

It is suspected that the observed result can be considered as a specific case of a more general result, suggested by the following conjecture:

Conjecture 3.4.4
Let $f : \mathbb{R}^q \to \mathbb{R}^d$ be a noncentral Gaussian process, and J its Jacobian. The entries of the Jacobian are correlated and non-centered. For the dimension q fixed, the noncentral Wishart matrix $\mathbf{G} = J^\top J$ converges to its expectation $\mathbb{E}[\mathbf{G}]$, when d goes to infinity.

To prove this conjecture, one should look at the recent advancements in matrix theory. It is well known that for a $d \times q$-random matrix X whose entries are independent, centered (i.e. of mean zero) of unit variance, and stationary, the central Wishart matrix $d^{-1}X^\top X$ converges almost surely to the identity matrix, when d goes to infinity. The problem arises when the entries are simultaneously not independent, not centered, and could be non-stationary. In our case, we could assume the entries to be stationary, which happens when the kernel of the Gaussian process is, for example, distance based (Noack and Sethian 2022). Yet, we can neither assume the entries to be independent nor central, since we are looking at the Jacobian of a Gaussian process.

Some progress has been made to relax the independence assumption. Nourdin and Zheng (2018) studied the asymptotic behavior of Wishart matrices, assuming correlation between rows only and overall correlation. Bourguin and Dang (2022) extended this work further by studying different asymptotic regimes, and proved some convergence results with Gaussian entries that are non-stationnary. Yet, in all those cases, the Gaussian entries are always assumed to be centered. The non-central case is still an open problem.

Information geometry 4

Information geometry is a branch of Riemannian geometry that deals with the study of the space of probability distributions. It considers the manifold to be the space of probability distributions and defines the Riemannian metric as the Fisher-Rao metric. The Fisher-Rao metric was introduced Rao (1945) and further studied by Amari and Nagaoka (2000). This metric enables us to gain a probabilistic understanding from a geometric perspective.

The main emphasis of this chapter will be directed towards the Fisher-Rao metric. In Section 4.1, we will focus on its construction and gain insight into its conceptual underpinnings from the perspective of information theory. Then, in Section 4.2, we will analyze its properties and explore how it facilitates the navigation within the space of probability distributions.

4.1 Fisher-Rao metric

Space of probability measures

So far, we have assumed our Riemannian manifolds to be homeomorphic to $\mathbb{R}^n$, with a metric defined as an inner product between vectors of the tangent space. In information geometry, we are working on the space of probability measures. We consider the manifold to be homeomorphic to an infinite-dimensional **Banach space**. The inner product between two real-value functions f and g in the Lebesgue space $L^2(X, \mu)$, with X a set and μ a measure, is defined as: $\langle f, g \rangle = \int_X fg d\mu$. Before defining the Fisher-Rao metric, we need to introduce some concepts of measure theory and define the space of probability measures and its tangent space. This part was inspired by Itoh and Satoh (2022).

Let Ω be a manifold, which serves as the support of a distribution, and $\mathcal{B}(\Omega)$ its Borel set. We have a volume form λ on Ω, such that the manifold is normalized: $\int_\Omega d\lambda = 1$. λ is also said to be a *reference probability measure* on Ω. We can now define all the probability measures as the set of the measures absolutely continuous with respect to λ. A measure μ is absolutely continuous with respect to another finite measure λ ($\mu \ll \lambda$) if and only if there exists a measurable function $f \in L^2(\Omega, \lambda)$ such that, for any Borel set $\mathcal{X} \in \mathcal{B}(\Omega)$, we have: $\mu(\mathcal{X}) = \int_{x \in \mathcal{X}} f(x) d\lambda(x)$, also written as $\mu = f\lambda$.

The function f is a density function that measures the rate at which μ changes with respect to λ. In measure theory, f is also called the *Radon-Nikodym derivative* of μ with respect to λ and denoted $d\mu/d\lambda$. We will be working on the **space of probability measures** defined as:

$$\mathcal{P}(\Omega) = \left\{ \mu = f\lambda \mid f \in L^2(\Omega, \lambda), f \in C^0(\Omega), \int_\Omega \mu = 1 \right\}$$

We also define the **tangent space of the probability measures** at $\mu \in \mathcal{P}(\Omega)$:

$$\mathcal{T}_\mu \mathcal{P}(\Omega) = \left\{ \tau = h\lambda \mid h \in L^2(\Omega, \lambda), h \in C^0(\Omega), \int_\Omega \tau = 0 \right\}$$

Fisher-Rao metric from a measure theory perspective

Definition 4.1.1 Fisher-Rao metric as an inner product in $\mathcal{P}(\Omega)$
Let Ω be a compact bounded manifold, $\mathcal{P}(\Omega)$ the space of probability measures on Ω and $\mu = f(x)\lambda \in \mathcal{P}(\Omega)$ a probability measure. $\tau_i = h_i(x)\lambda$ and $\tau_j = h_j(x)\lambda$ are two elements of the tangent space $\mathcal{T}_\mu\mathcal{P}(\Omega)$.

The Fisher-Rao metric is defined as:

$$g_\mu(\tau_i, \tau_j) = \int_\Omega \frac{d\tau_i}{d\mu}\frac{d\tau_j}{d\mu}d\mu = \int_{x\in\Omega} \frac{h_i(x)}{f(x)}\frac{h_j(x)}{f(x)}f(x)d\lambda(x)$$

The Radon-Nikodym derivatives $d\tau_i/d\lambda$ are also defined as $h_i \in L^2(\Omega, \lambda)$ which is the density function of τ_i with respect to μ, and similarly, $d\mu/d\lambda$ is defined as $f \in L^2(\Omega, \lambda)$.

Now, the probability density functions depend on two parameters: a variable $x \in \Omega$, where the manifold Ω is the support of the distribution, and a parameter $\eta \in \mathcal{H}$ that characterizes the probability density function f. For example, if f is a Gaussian distribution, then $\eta = \{\mu, \sigma^2\}$ would be the mean and the variance of the distribution. We would write $f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp -\frac{1}{2}(\frac{x-\mu}{\sigma})^2$. $\mathcal{H}$ is called the **parameter space**, and it is chosen such that the map $\eta \to f(\cdot; \eta)$ is smooth and injective.

The density function h is the Radon-Nikodym derivative of $\tau \in \mathcal{T}_\mu\mathcal{P}(\Omega)$. If we consider $h_i = \partial_i f = \frac{\partial f}{\partial_{\eta_i}}$ as the derivative of f with respect to the parameter η_i, then, the inner product can be defined by the probability density function f:

$$\begin{aligned} g_\mu(\eta_i, \eta_j) &= \int_{x\in\Omega} \frac{\partial_i f(x,\eta)}{f(x,\eta)}\frac{\partial_j f(x,\eta)}{f(x,\eta)}f(x,\eta)d\lambda(x) \\ &= \int_{x\in\Omega} \partial_i \ln f(x,\eta)\partial_j \ln f(x,\eta)\ f(x,\eta)d\lambda(x). \end{aligned}$$

A space that is equipped with the Fisher-Rao metric is called a **statistical manifold**.

Table 4.1: Review of information geometry objects within Riemannian geometry

Objects	Riemannian geometry	Information geometry
Manifold	$\mathcal{M}$	$\mathcal{P}(\Omega) = \{f_\eta = f(\cdot, \eta) \mid \eta \in \mathcal{H}\}$ Space of probability distributions
Homeomorphic to	Euclidean space	Banach space
Tangent manifold	$\mathcal{T}_x\mathcal{M}$	$\mathcal{T}_\mu\mathcal{P}(\Omega)$
Metric tensor	$\mathbf{G}$ Riemannian metric	$\mathbb{E}[\nabla_\eta \ln f\ \nabla_\eta \ln f^\top]$ Fisher Rao metric
Length [1]	$\mathcal{L}(\gamma) = \int_0^1 \|\dot\gamma(t)\|\, dt$	$\mathcal{L}(\gamma) \approx \sqrt{2}\sum_{\eta=1}^{T} \sqrt{\mathrm{KL}(f_\eta, f_{\eta+\delta\eta})}$
Energy	$\mathcal{E}(\gamma) = \frac{1}{2}\int_0^1 \|\dot\gamma(t)\|^2\, dt$	$\mathcal{E}(\gamma) \approx \frac{2}{\delta\eta}\sum_{\eta=1}^{T} \mathrm{KL}(f_\eta, f_{\eta+\delta\eta})$

[1] See Section 4.3 for the derivation of length and the energy functional.

Fisher-Rao metric from an information theory perspective

Before the work of Amari and Nagaoka (2000) in information geometry, the Fisher-Rao metric was first defined by Rao (1945) as a multidimensional generalization of the Fisher information, a way to measure how sensitive is a model f when its parameters η vary. It was used by Fisher (1922) to define the quality of a measurement.

We have a random variable x on which we perform measurements. It could be a set of observed data points. η, a parameter that we are trying to estimate, such as the mean or the variance, related to x through a probability density function $f(x, \eta)$, which specifies the likelihood of observing x given η. We further call the *efficient estimator*, the best unbiased estimate $\hat{\eta}$ of the true parameter η after many independent measurements on a random variable x. In particular, we have $\mathbb{E}[\hat{\eta}(x)] = \int \hat{\eta}(x) f(x, \eta) dx = \eta$. We define the error of the estimate $\hat{\eta}$ as $\text{var}(\eta) = \int f(x, \eta)(\hat{\eta}(x) - \eta)^2 dx$. The Fisher information, denoted $\mathbf{I}_F$, is related to the error of the estimate $\text{var}(\eta)$ as:

$$\mathbf{I}_F \ \text{var}(\eta) \geq 1.$$

This equation is the Cramer-Rao inequality (Rao 1945; Cramér 1946) and expresses the reciprocity between the mean-square error $\text{var}(\eta)$ and the Fisher information $\mathbf{I}_F$ in the intrinsic data. The lowest the error $\text{var}(\eta)$ the more informative is our estimate, and the more sensitive is the Fisher information.

The Fisher information $\mathbf{I}_F$ and the Cramer-Rao inequality are obtained by differentiating the expectation of a class of unbiased estimators:

$$\begin{aligned}
\mathbb{E}[\hat{\eta}(x) - \eta] = 0 &\iff \int (\hat{\eta}(x) - \eta) f(x, \eta) dx = 0 \\
\partial_\eta \mathbb{E}[\hat{\eta}(x) - \eta] = 0 &\iff \int (\hat{\eta}(x) - \eta) \partial_\eta f(x, \eta) dx - \int f(x, \eta) dx = 0 \\
&\iff \int (\hat{\eta}(x) - \eta) \partial_\eta (\ln f(x, \eta)) f(x, \eta) dx = 1 \\
&\iff \int \left((\hat{\eta}(x) - \eta) \sqrt{f(x, \eta)} \right) \left(\partial_\eta \ln f(x, \eta) \sqrt{f(x, \eta)} \right) dx = 1
\end{aligned}$$

And using the Cauchy-Schwarz inequality:

$$\int (\hat{\eta}(x) - \eta)^2 f(x, \eta) dx \int (\partial_\eta \ln f(x, \eta))^2 f(x, \eta) dx \geq 1,$$

with $\text{var}(\eta) = \int f(x, \eta)(\hat{\eta}(x) - \eta)^2 dx$ and $\mathbf{I}_F = \int (\partial_\eta \ln f(x, \eta))^2 f(x, \eta) dx$.

The Fisher information measures how sensitive the expected statistics are to changes in the parameters. If a small change in parameters causes a large change in the expected statistics, it means that the estimator is informative and the Fisher information will be large. Another way to see that the Fisher information, and the Fisher-Rao metric measures the sensitivity of the model f to the parameters η is to consider the small perturbations of the probability density function $f(x, \eta)$ around η.

Definition 4.1.2 Fisher-Rao metric as the covariance matrix.
We define the small perturbations Δ on the probability density function f:

$$\Delta_i = \frac{f(x, \eta + d\eta^i) - f(x, \eta)}{f(x, \eta)} = \frac{\partial_i f(x, \eta)}{f(x, \eta)} d\eta^i = \partial_i \ln f(x, \eta) \ d\eta^i$$

The covariance matrix is: $\text{Cov}(\Delta_i, \Delta_j) = g_{ij} \ d\eta^i d\eta^j$, with the Fisher-Rao metric: $g_x(\partial_i, \partial_j) = \int \partial_i \ln f(x, \eta) \ \partial_j \ln f(x, \eta) f(x, \eta) dx$.

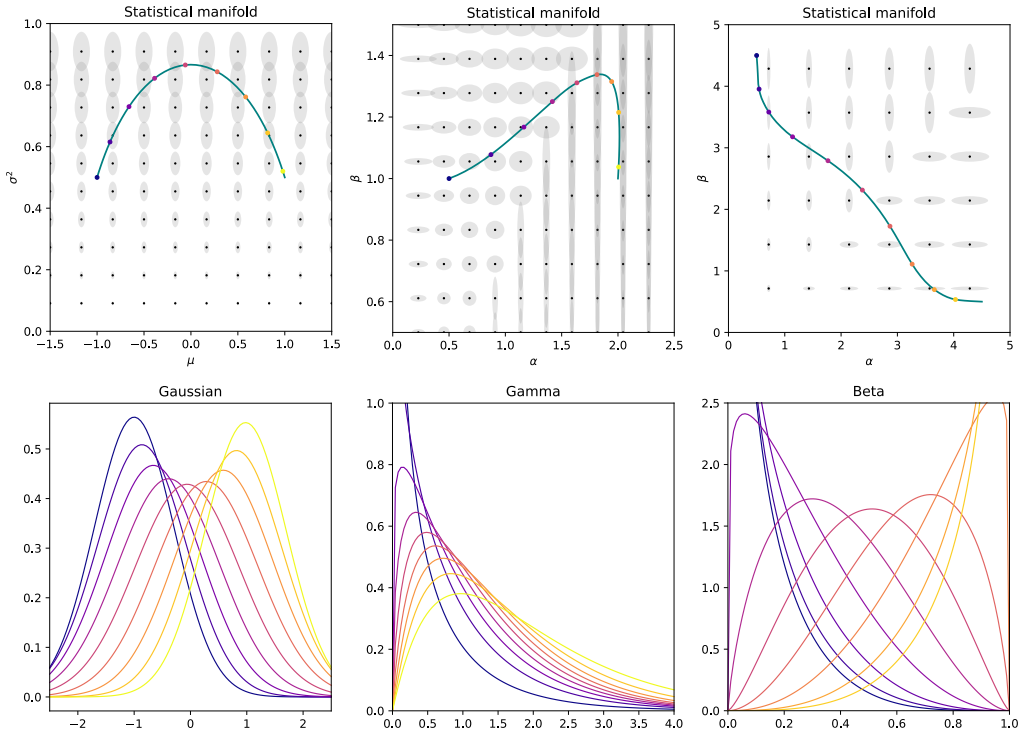


Figure 4.1: Three families of probability distribution: a Gaussian distribution, a Gamma distribution and a Beta distribution. In those figure, the Fisher-Rao metric is represented by its indicatrix. The geodesics have been plotted such that it minimizes the curve length.

Example 4.1.1 Fisher-Rao metric for a Gaussian, Gamma and Beta distributions

Distributions	Probability density function	Fisher-Rao matrix
Gaussian(μ, σ^2)	$(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$
Gamma(α, β)	$\Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1}e^{-\beta x}$	$\begin{pmatrix} \frac{\alpha}{\beta^2} & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \Psi_1(\alpha) \end{pmatrix}$
Beta(α, β)	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\begin{pmatrix} \Psi_1(\alpha) - \Psi_1(\alpha+\beta) & -\Psi_1(\alpha+\beta) \\ -\Psi_1(\alpha+\beta) & \Psi_1(\beta) - \Psi_1(\alpha+\beta) \end{pmatrix}$

The Fisher-Rao metric is plotted in Figure 4.1 for the three distributions. Notice that the Fisher-Rao for the Gaussian distribution is similar to the Riemannian metric of a half-Poincaré plane. To go from one normal distribution to another, we need to increase the variance first, and so increase the entropy, before reaching the target distribution. In each case, the geodesics follow the path that locally minimizes the KL-divergence. For the full derivations of the Fisher-Rao metric, see Paper 7 and Appendix C.

4.2 Statistics and information geometry

In statistics, one typically wishes to estimate some parameters $\eta \in \mathcal{H}$ based on random samples drawn from unknown probability distributions $f(x, \eta)$. Sometimes, we might not access the original data $x \in \Omega$, but an observable $x' \in \Omega'$, such that $x' = \kappa(x)$, with $\kappa : \Omega \to \Omega'$ a mapping. κ is called a **statistic**, and we say that κ is *sufficient* if there is

no loss of information to compute any estimate of the parameters η, and so we have:

$$f(x, \eta) = h(x) \cdot f(\kappa(x), \eta),$$

with $h(x)$ a constant that only depends on x.

Example 4.2.1 Sufficient statistics
Let us assume that we have independent identically distributed variables $x = \{x_1, \cdots, x_n\}$ following the exponential law with $f(x_i, \lambda) = \lambda e^{-\lambda x_i}$. Then we have: $f(x, \lambda) = \prod_i^n f(x_i, \lambda) = \lambda^n e^{-\lambda \sum_i x_i}$.

Instead of observing x, we observe the sum: $\kappa(x) = \sum_i^n x_i$. Then, our probability density function becomes: $f(\kappa(x), \lambda) = \lambda^n e^{-\lambda \kappa(x)}$. In this case we can see that:

$$f(x, \lambda) = f(\kappa(x), \lambda),$$

the sum is a sufficient statistic for the exponential distribution.

Let g be the Fisher-Rao metric. Then if κ is a sufficient statistic, and $h(x)$ a scaling factor, we have:

$$g_x(\partial_i, \partial_j) = h(x) \cdot g_{\kappa(x)}(\partial_i, \partial_j)$$

Not only the Fisher-Rao metric is invariant under sufficient statistics, it is also the unique Riemannian metric with this property in the space of probability densities. This result has been shown by Chentsov (1982) on finite sample spaces who gave his name to the theorem. It has been extended on infinite sample space by Ay et al. (2015) and Bauer et al. (2016).

Theorem 4.2.1 Chentsov's theorem
The Fisher-Rao metric is the **unique** Riemannian metric that is **invariant under sufficient statistics**, up to a scaling factor. We also say that the metric is invariant under the action of the diffeomorphism group, or Markov morphisms.

Proof. See Chentsov (1982), Ay et al. (2015) and Bauer et al. (2016). □

The Fisher-Rao metric is then twice invariant: it is co-invariant under reparameterization since it is a Riemannian metric, and invariant under sufficient statistics.

The Fisher-Rao metric is also related to the KL-divergence. The KL divergence (Kullback and Leibler 1951) is a measure of dissimilarity between two probability distributions, and quantifies the amount of information lost when one distribution is used to approximate another. It is defines between two probability measures p and q as:

$$KL(p, q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$$

The Fisher-Rao metric actually appears when we compute the Taylor expansion of the KL-divergence between two distributions locally.

We introduce the score function $s_i(x, \eta) = \partial_i \ln p(x, \eta)$. We have:

$$\begin{aligned} \text{The mean of score function:} \quad & \mathbb{E}[s_i(x, \eta)] = 0 \\ \text{The Fisher-Rao metric:} \quad & \mathbb{E}[s_i(x, \eta) s_j(x, \eta)] = g_{ij} \\ \text{The Amari-Chentsov tensor:} \quad & \mathbb{E}[s_i(x, \eta) s_j(x, \eta) s_k(x, \eta)] = T_{ijk} \end{aligned}$$

Theorem 4.2.2 Taylor expansion of the KL-divergence
We define KL as being the Kullback Leber divergence, $\mathbf{g}$ is our Fisher-Rao metric and $\mathbf{T}$ is the Amari Chentsov tensor. When $\|z\| = \|\eta - \eta'\|$ is small, we have the Taylor expansion of the canonical divergence:

$$KL[\eta : \eta'] = \frac{1}{2}\mathbf{g}_{ij}(\eta)z^i z^j + \frac{1}{6}\mathbf{T}_{ijk}(\eta)z^i z^j z^k + O(\|z\|^4).$$

Proof. See Felice and Ay (2021, Proposition 6) and Ay et al. (2015, Proposition 2.125) □

4.3 Summary: Pulling back information geometry

Generative models such as Variational Auto-Encoders (VAEs) are powerful tools to model high dimensional data. The data $\mathbf{x}$ is assumed to lie on a high-dimensional manifold $\mathcal{X}$, and it is parameterized by a low-dimensional latent representation $\mathbf{z} \in \mathcal{Z}$.

In the specific case of a Gaussian distribution, the decoder $f : \mathcal{Z} \subset \mathbb{R}^q \to \mathcal{X} \subset \mathbb{R}^D$, can be expressed as:

$$f(\mathbf{z}) = \mu(\mathbf{z}) + \sigma(\mathbf{z}) \odot \epsilon,$$

with $\epsilon \sim \mathcal{N}(0, \mathbb{I}_D)$. We can explore the latent space with the pullback metric: $\mathbf{G} = J^\top J$, with J the Jacobian of f. Since the decoder is stochastic, the metric pulled back from the ambient space is stochastic as well. A good deterministic approximation is to take its expectation, which is conveniently computed as:

$$\mathbb{E}[\mathbf{G}] = J_\mu(\mathbf{z})^\top J_\mu(\mathbf{z}) + J_\sigma(\mathbf{z})^\top J_\sigma(\mathbf{z}),$$

with $J_\mu(\mathbf{z}) = \nabla_{\mathbf{z}}\mu(\mathbf{z})$ and $J_\sigma(\mathbf{z}) = \nabla_{\mathbf{z}}\sigma(\mathbf{z})$.

We can see that, when a VAE is equipped with a Gaussian decoder, we can easily access the expected metric tensor, and compute geodesics. Yet, a major advantage of a VAE is its ability to handle data from various types of distributions, such as gamma, Poisson, or categorical, depending on the problem at hand. The difficulty comes when we try to calculate the expected metric tensor for these distributions - it is not a straightforward task. We cannot easily treat the random manifold as deterministic in a way that is convenient for computation.

Instead of focusing on the data manifold itself $\mathcal{X}$, we chose to direct our attention to the underlying structure that describes the probability distributions $\mathcal{D} \in \mathcal{P}(\mathcal{X})$ from which the data is drawn: $\mathbf{x} \sim \mathcal{D}(\mathbf{z})$. We can pullback the metric from the space of probability distributions $\mathcal{P}(\mathcal{X})$ to the parameter space $\mathcal{Z}$. This metric is precisely the Fisher-Rao metric.

In this work, we borrow tools from information geometry, and we equip the latent space of a VAE with the Fisher-Rao metric. The contributions are the following:

- We show that the geodesics derived with the Fisher-Rao metric are the KL-divergence minimizing curves, which greatly simplifies the computation of the geodesics in practice.
- We illustrate our findings with two real-life examples: one that uses a von Mises Fisher distribution to model human rigid motion, and one that uses a categorical distribution to model a recommender system.

Information geometry and generative models

In information geometry, we often look at the simple case when the data $\mathbf{x}$ is drawn from a parametric distribution $p(\mathbf{x}|\eta)$, with $\eta \in \mathcal{H}$ the parameter space. This parameter space, equipped with the Fisher-Rao metric $\mathbf{I}_{\mathcal{H}}$, is called a **statistical manifold**. For example, if we assume that the data is drawn from a Gaussian distribution, then $\eta = (\mu, \sigma)$, with μ the mean and σ the standard deviation.

In practice, for a VAE, this parameter space $\mathcal{H}$ can be parameterized from the latent space $\mathcal{Z}$, with a non-linear function h such that: $h(\mathbf{z}) = \eta$. Then, the decoder of a VAE can be represented in two different ways by a succession of mappings:

$$\underset{\mathbf{z}\in\mathcal{Z}}{\mathbf{z}} \xrightarrow{\quad p(\mathbf{x}|\mathbf{z}) \quad} (\mathbf{x} \sim \mathcal{D}(x)) \quad \underset{\mathbf{x}\in\mathcal{X},\ \mathcal{D}\in\mathcal{P}(\mathcal{X})}{}$$
$$\xrightarrow{\quad h \quad} \underset{\eta\in\mathcal{H}}{\eta} \xrightarrow{\quad p(\mathbf{x}|\eta) \quad}$$

The Fisher-Rao metric $\mathbf{I}_{\mathcal{Z}}$ directly pulled back from the probability space $\mathcal{P}(\mathcal{X})$ to $\mathcal{Z}$ is equal to the pulled back metric $J_h^\top \mathbf{I}_{\mathcal{H}} J_h$ from $\mathcal{M}$ to $\mathcal{Z}$, with $\mathbf{I}_{\mathcal{H}}$ equipping the parameter space $\mathcal{H}$. In other terms, we can equivalently compute geodesics on $\mathcal{Z}$ or $\mathcal{H}$.

Efficient geodesic computation

A well known result in information geometry is that the Fisher-Rao metric is the first order approximation of the KL-divergence between two infinitesimally close distribution:

$$\mathrm{KL}(p(\mathbf{x}|\eta), p(\mathbf{x}|\eta + \delta\eta)) = \frac{1}{2}\delta^\top \mathbf{I}(\eta)\delta\eta + o(\delta\eta^2),$$

with $\mathbf{I}(\eta) = \int p(\mathbf{x}|\eta)\nabla_\eta \log p(\mathbf{x}|\eta)\nabla_\eta \log p(\mathbf{x}|\eta)^\top d\mathbf{x}$, the Fisher-Rao metric tensor.

We consider a curve $\gamma(t)$ and its derivative $\dot{\gamma}(t)$ on the statistical manifold such that, $\forall t \in [0,1], \gamma(t) = \eta_t \in \mathcal{H}$. The manifold is equipped with the Fisher-Rao metric. The length and the energy functionals are defined with respect to the metric $\mathbf{I}(\eta)$:

$$\mathrm{Length}(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^\top \mathbf{I}(\eta)\dot{\gamma}(t)}dt \quad \text{and} \quad \mathrm{Energy}(\gamma) = \int_0^1 \dot{\gamma}(t)^\top \mathbf{I}(\eta)\dot{\gamma}(t)dt.$$

Length-minimizing curves between two connecting points can be found by minimizing the energy functional. In our case, the geodesics on the statistical manifold locally minimize the KL-divergence between two distributions.

Proposition 4.3.1 Geodesics on the statistical manifold
The KL-divergence between two close elements of the curve γ is defined as: $\mathrm{KL}(p_t,\ p_{t+\delta t}) = \mathrm{KL}(p(\mathbf{x}|\gamma(t)),\ p(\mathbf{x}|\gamma(t+\delta t)))$. The length $\mathcal{L}$ and the energy $\mathcal{E}$ functionals can be approximated with respect to this KL-divergence:

$$\mathcal{L}(\gamma) \approx \sqrt{2}\sum_{t=1}^{T} \sqrt{\mathrm{KL}(p_t, p_{t+\delta t})} \quad \text{and} \quad \mathcal{E}(\gamma) \approx \tfrac{2}{\delta t}\sum_{t=1}^{T} \mathrm{KL}(p_t, p_{t+\delta t})$$

Proof. We replace $\gamma(t+\delta t) = \gamma(t) + \delta t\ \dot{\gamma}(t)$ in the previous equations. See Appendix C, for the detailed proof. □

As we can directly access the distributions $p(\mathbf{x}|\eta)$ of a VAE, we can conveniently compute the KL-divergence between two close distributions, and so, compute geodesics on our latent space.

Staying within the support of the data

So far, we have described a convenient way to explore the probability space, by computing geodesics on the statistical manifold. However, we don't exactly have access to $\mathcal{P}(\mathcal{X})$, but rather $\mathcal{P}(\Omega)$, with Ω the set of all the possible outcomes, such that the probability density function of $\mathbf{x}$ satisfies:

$$\int_{\Omega} p(\mathbf{x}) d\mathbf{x} = 1.$$

The manifold Ω that we are exploring is much larger than the data manifold $\mathcal{X} \subset \Omega$. This is also reflected in the latent space $\mathcal{Z}$, and we want to ensure that the geodesics computed in our latent space stay within the support of the data.

To achieve this, we propose to reparameterize the non-linear map h such that, within the support of the data, the data is drawn from a chosen distribution $\mathcal{D}$, $p(\mathbf{x}|h(\mathbf{z})) = \mathcal{D}_{\mathcal{X}}$, and outside the support of the data, the uncertainty is maximized: $p(\mathbf{x}|h(\mathbf{z})) = \mathcal{U}_{\Omega \setminus \mathcal{X}}$, with $\mathcal{U}$ the uniform distribution:

$$p(\mathbf{x}|h(\mathbf{z})) = \begin{cases} \mathcal{D}_{\mathcal{X}} & \text{if} \quad \mathbf{x} \in \mathcal{X} \\ \mathcal{U}_{\Omega \setminus \mathcal{X}} & \text{if} \quad \mathbf{x} \notin \mathcal{X} \end{cases} \quad \Rightarrow \quad h(\mathbf{z}) = \begin{cases} h(\mathbf{z}) & \text{if} \quad \mathbf{z} \in \mathcal{Z} \\ u(\mathbf{z}) & \text{if} \quad \mathbf{z} \notin \mathcal{Z} \end{cases},$$

with u a function that maximizes the entropy of the distribution:

$$u(\mathbf{z}) = \begin{cases} \sigma(\mathbf{z}) \to \infty & \text{for a Normal distribution} \\ p(\mathbf{z}) = \frac{1}{2} & \text{for a Bernoulli distribution} \\ (\alpha(\mathbf{z}), \beta(\mathbf{z})) = (1, 1) & \text{for a Beta distribution} \\ \alpha(\mathbf{z}) = 1 & \text{for a Gamma distribution} \\ \lambda(\mathbf{z}) \to 1 & \text{for a Poisson distribution} \end{cases}$$

If d is the distance between a latent code $\mathbf{z}$ and the support of the data, then h is reparameterized as: $\tilde{h}(\mathbf{z}) = (1 - d(\mathbf{z})) \cdot h(\mathbf{z}) + d(\mathbf{z}) \cdot u(\mathbf{z})$.

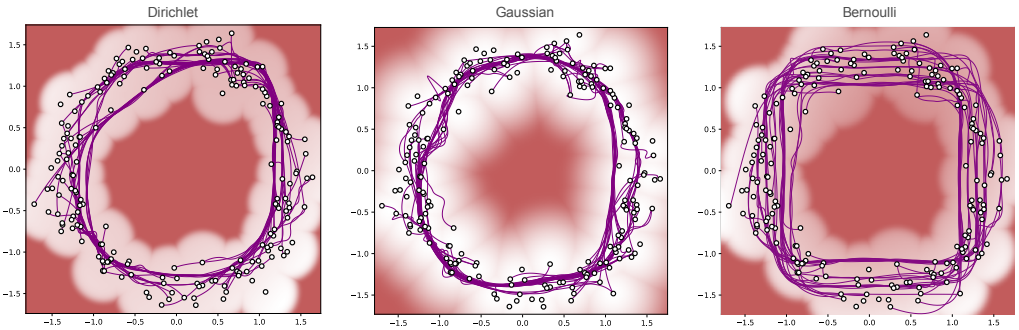


Figure 4.2: From left to right: Dirichlet, Gaussian and Bernoulli. The geodesics are plotted for those distributions when the support of the data is a circle. The white areas represent low entropy of the decoded distribution, while purple areas represent higher entropy.

Experiments: Motion capture data with a von Mises-Fisher decoder

We consider a model of human motion capture data. Here we observe a time series, where each time point corresponds to a human pose composed of multiple receptors. Over time, the individual limbs on the body only change their position and orientation. Each limb can be seen as a point on a sphere in $\mathbb{R}^3$ with radius given by the limb length. We can

view the entire skeleton representation as a product of spheres. From this, we build a VAE where the decoder distribution is a product of von Mises-Fisher distributions. In this case, we do not have easily accessible Fisher-Rao metric tensor, so we plot the geodesics by minimizing the curve-length obtained with the KL-divergence from Sec. 4.3. While the KL does not have a closed-form expression for the von Mises-Fisher distribution, we use a Monte Carlo estimate.

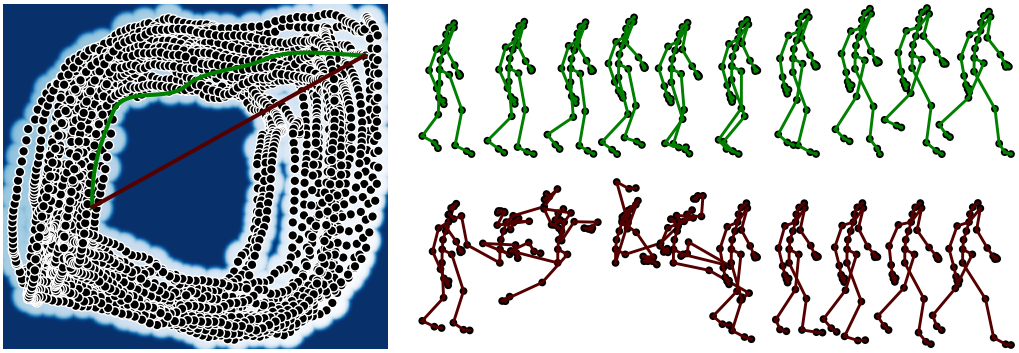


Figure 4.3: The figure shows the latent representation of a motion capture sequence of a person walking with the shortest paths superimposed. We see that the geodesics reflect the underlying periodic nature of the observed walking motion. As we traverse the latent space with a straight line and the geodesic, we sample the data from the decoder, producing new motion sequences. As can be seen, the straight line ends up creating implausible motions, while the geodesics from the Fisher-Rao metric generates meaningful poses.

Part III

Papers

On the curvature of the loss landscape

5

The paper *On the curvature of the loss landscape*, has been written with Hrittik Roy, Sidak Pal Singh, and Georgios Arvanitidis. The work is currently in progress and available as a preprint at https://arxiv.org/abs/2307.04719.

Abstract

One of the main challenges in modern deep learning is to understand why such over-parametrized models perform so well when trained on finite data. A way to analyze this generalization concept is through the properties of the associated loss landscape. In this work, we consider the loss landscape as an embedded Riemannian manifold and show that the differential geometric properties of the manifold can be used when analyzing the generalization abilities of a deep net. In particular, we focus on the scalar curvature, which can be computed analytically for our manifold, and show connections to several settings that potentially imply generalization.

5.1 Flatness and generalization in machine learning

The relationship between the generalization ability of a model and the flatness of its loss landscape has been a subject of interest in machine learning. **Flatness** refers to the shape of the hypersurface representing the loss function, parametrized by the parameters of the model. Flat minima are characterized by a wide and shallow basin. **Generalization** refers to the ability of a model to perform well on unseen data. A widely accepted hypothesis, proposed by various research groups (Hochreiter and Schmidhuber 1997; Hinton and Van Camp 1993; Buntine 1991) several decades ago, suggests that flat minima are associated with better generalization compared to sharp minima. The basis of this hypothesis stems from the observation that when the minima of the optimization landscape are flatter, it enables the utilization of weights with lower precision. This, in turn, has the potential to improve the robustness of the model.

The notion of flatness has been challenged by Dinh et al. (2017), who argued that the different flatness measures proposed are not invariant under reparameterization of the parameter space and questioned the assumption that flatness directly causes generalization.

Yet, numerous empirical and theoretical studies have presented compelling evidence that supports the relationship between flatness and enhanced generalization. This relationship has been observed in various contexts, by averaging weights (Izmailov et al. 2018), studying inductive biases (Neyshabur et al. 2017; Imaizumi and Schmidt-Hieber 2022), introducing different noise in gradient descent (Chaudhari et al. 2019; Pittorino et al. 2021), adopting smaller batch sizes (Keskar et al. 2016), and investigating ReLU Neural networks (Yi et al. 2019).

The exact relationship between flatness and generalization is still an open problem in machine learning. In this preliminary work, we build upon the *flatness hypothesis* as a primary motivation to investigate the curvature of the loss landscape, approaching it from a differential geometric perspective.

In this preliminary work, we analyze the loss landscape as a Riemannian manifold and derive its *scalar curvature*, an intrinsic Riemannian object that characterizes the

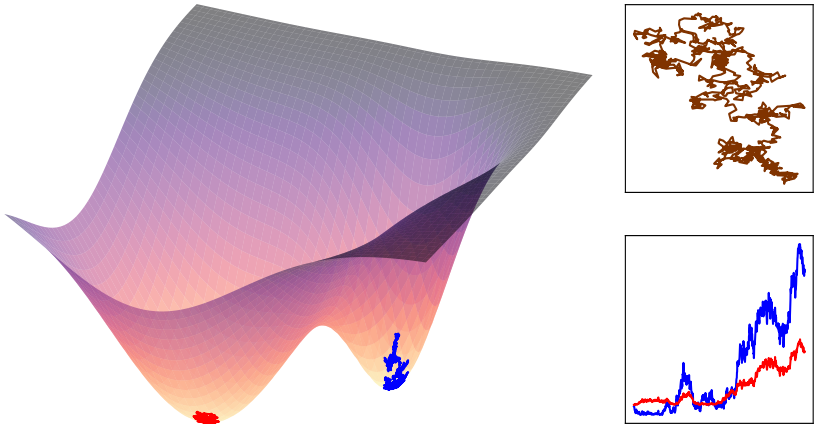


Figure 5.1: On the left, a surface represents a loss function $f(\mathbf{u}, \mathbf{v})$ on its parameter space $\{\mathbf{u}, \mathbf{v}\}$. We can see two minima, a sharp minimum and a flatter minimum. A Brownian motion navigates the parameter space around those two minima, in blue for the sharp one, and red for the shallow one. **On the right**, the upper figure represents the Brownian motion navigating in the parameter space. The same is used for both minima. The lower figure represents the perturbations of the loss f in both the sharp (blue) and flat (red) minima. The loss is more robust to perturbation in the flatter minima.

local curvature of the manifold. We found that the scalar curvature, at minima, has a straightforward expression and can be related to the norm of the Hessian. While the norm of the Hessian may not always accurately measure flatness, it remains a valuable indicator for understanding optimization. Our findings demonstrate that the scalar curvature possesses all the benefits of the Hessian norm without its limitations.

5.2 Geometry of the loss landscape and curvature

We are interested in finding the parameters $\mathbf{x}$ of a model that minimizes the loss function denoted f. The loss function is a smooth function defined on the parameter space $\mathcal{M} \subset \mathbb{R}^q$, where q is the number of parameters. In order to study the loss landscape of a model, we can look at the geometry of the graph of the loss function, which is a hypersurface embedded in $\mathbb{R}^{q+1}$.

Definition 5.2.1 Metric of a graph
Let $f : \Omega \subset \mathbb{R}^q \to \mathbb{R}$ be a smooth function. We call **graph of a function** the set:

$$\Gamma_f = \{(\mathbf{x}, y) \in \Omega \times \mathbb{R} \mid y = f(\mathbf{x})\}.$$

The graph Γ_f is a topological smooth manifold embedded in R^{q+1}, and it is isometric to the Riemannian manifold $(\mathcal{M}, g)$ with $\mathcal{M} \subset \mathbb{R}^q$ and the induced metric

$$g_{ij} = \delta_{ij} + \partial_i f \partial_j f. \tag{5.1}$$

The metric is obtained by pulling back, in one case, the loss function to the parameter space $(\partial_i f \partial_j f)$, and in another case, the parameter space to itself (δ_{ij}), (Lee 2018).

Instead of working in the ambient space $\mathbb{R}^{q+1}$, it is more convenient to study the intrinsic geometry of the loss function in the parameter space $(\mathcal{M}, g)$. In particular, knowing the Riemannian metric, we can compute the associated geometric quantities of the loss landscape as the Christoffel symbols, the Riemannian curvature tensor, and the scalar curvature (See Appendix A.1 for an introduction of those quantities). In the following,

we will denote ∇ the Euclidean gradient operator of the loss function f, and $\mathbf{H}$ the Euclidean Hessian of f.

Gradient(f): $(\nabla f)_i = \mathbf{J}_i = \partial_i f = f_{,i}$
Hessian(f): $(\mathbf{H})_{ij} = \partial_i \partial_j f = f_{,ij}$

Curvature in Riemannian geometry

The Christoffel symbols define a corrective term used to compute covariant derivatives in a curved space. They can be derived from the Riemannian metric.

Proposition 5.2.1 Christoffel symbols
The Christoffel symbols are given by:

$$\Gamma^i_{kl} = \beta f_{,i} f_{,kl}, \tag{5.2}$$

with $\beta = (1 + \|\nabla f\|^2)^{-1}$.

Proof. See Appendix A.2. □

Using those Christoffel symbols, we can directly compute the Riemannian curvature tensor. Using the Einstein summation convention, the Riemannian curvature tensor is an intrinsic mathematical object that characterizes the deviation of the curved manifold from the flat Euclidean manifold.

Proposition 5.2.2 Riemannian curvature tensor
The Riemannian curvature tensor is given by:

$$\mathrm{R}^i_{jkm} = \beta(f_{,ik} f_{,jm} - f_{,jm} f_{,jk}) - \beta^2 f_{,i} f_{,r} (f_{,rk} f_{,im} - f_{,rm} f_{,jk}), \tag{5.3}$$

with $\beta = (1 + \|\nabla f\|^2)^{-1}$.

Proof. See Appendix A.2. □

While those four-dimensional tensor gives us a complete picture of the curvature of a manifold, it can be difficult to interpret in practice. Instead, a scalar object, the scalar curvature, can be derived from the Riemannian curvature tensor. The scalar curvature quantifies locally how curved is the manifold.

Proposition 5.2.3 Scalar curvature
The scalar curvature is given by:

$$\mathrm{S} = \beta \left(\mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H}^2)\right) + 2\beta^2 \left(\nabla f^\top (\mathbf{H}^2 - \mathrm{tr}(\mathbf{H})\mathbf{H}) \nabla f\right), \tag{5.4}$$

with $\beta = (1 + \|\nabla f\|^2)^{-1}$.

Proof. See Appendix A.2. □

This expression simplifies when the gradient is zero, which corresponds to a critical point of the loss function. In this case, the scalar curvature is given by:

Corollary 5.2.4

When an extremum is reached ($\nabla f = 0$), the scalar curvature becomes:

$$\mathrm{S}(\mathbf{x}_{\min}) = \mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H}^2) \tag{5.5}$$

Proof. This is a direct result of Proposition 5.2.3, when $\nabla f = 0$. □

Note that we can also write, at the minimum, $\mathrm{S}(\mathbf{x}_{\min}) = \|\mathbf{H}\|_*^2 - \|\mathbf{H}\|_F^2$, with $\|\cdot\|_*$ the nuclear norm and $\|\cdot\|_F$ the Frobenius norm.

The scalar curvature as the deviation of the volume of geodesic balls

This scalar curvature has a simple interpretation, as it corresponds to the difference in volume between a geodesic ball embedded in the Riemannian manifold and a ball of reference, the Euclidean ball. In hyperbolic spaces, the Riemannian ball will be bigger than the Euclidean one, and in spherical spaces, it will be smaller. If the curved space is flat, they are both equal in volume, and the scalar curvature is null.

Proposition 5.2.5
(Gallot et al. 1990, Theorem 3.98)
The scalar curvature $\mathrm{S}(\mathbf{x})$ at a point $\mathbf{x} \in \mathcal{M}$ of the Riemannian manifold of dimension q is related to the asymptotic expansion of the volume of a ball on the manifold $\mathcal{B}_g(r)$ compared to the volume of the ball in the Euclidean space $\mathcal{B}_e(r)$, when the radius r tends to 0.

$$\mathrm{vol}(\mathcal{B}_g(r)) = \mathrm{vol}(\mathcal{B}_e(r)) \left(1 - \frac{\mathrm{S}(\mathbf{x})}{6(q+2)} r^2 + o(r^2)\right)$$

5.3 Scalar curvature and optimization

Corollary 5.2.4 establishes a connection between the scalar curvature at each peak or valley in the loss landscape and the magnitude of the Hessian: $\mathrm{S}(\mathbf{x}) = \|\mathbf{H}\| *^2 - \|\mathbf{H}\|_F^2$. Although the Hessian norm plays a key role in optimization tasks, we contend that it is not the most reliable gauge of flatness in *all* situations. On one hand, will delve into some issues that arise from only using the Hessian norm in Section 5.3.1. On the other hand, we will see how the scalar curvature reduces to the Hessian norm in some cases and supports theoretical findings in optimization in Section 5.3.2.

5.3.1 Limitations of the trace of the Hessian as a measure of flatness

The Hessian of the loss function, specifically its trace, has been shown to influence the convergence of optimization algorithms. For instance, Wei and Schwab (2019) revealed that stochastic gradient descent (SGD) reduces the trace of the loss function's Hessian in the context of over-parametrized networks. In a similar vein, Orvieto et al. (2022) discovered that SGD with anti-correlated perturbations enhances generalization due to the induced noise reducing the Hessian's trace. They also identified that the trace serves as an upper limit on the mean loss over a posterior distribution. Furthermore, within Graphical Neural Networks, Ju et al. (2023) demonstrated that the trace of the Hessian can evaluate the model's resilience to noise.

The saddle point problem

Yet, relying solely on the trace of the Hessian may not provide an accurate measure of flatness. For instance, if half of the eigenvalues are positive and the other half are negative, with their sum equaling zero, the trace of the Hessian will also be zero. This is misleading as it suggests a flat region, when in reality it is a saddle point.

Example 5.3.1 Curvature of a parametrized function
Let us imagine that the loss is represented by a function taking in inputs two weights u and v such that:

$$f(u, v) = e^{-cu} \sin(u) \sin(v),$$

with c a positive constant. We notably have $\lim_{u \to \infty} f(u, v) = 0$, and so the surface tends to be flatter with u increasing.

The trace of the Hessian of f and its scalar curvature can be computed analytically, and we have at a point $\mathbf{x} = (u, v)$:

$$\begin{aligned} \mathrm{tr}(\mathbf{H})(\mathbf{x}) =& e^{-cu}(-2u\cos(u) + (c^2 - 2)\sin(u))\sin(v) \\ \mathrm{S}(\mathbf{x}) =& \frac{(c^2 - 1)\cos(2u) - \cos(2v) - c(c - 2\sin(2u))}{e^{2cu} + \cos(v)\sin(u)^2 + (\cos(u) - c\sin(u))\sin(v)^2} \end{aligned}$$

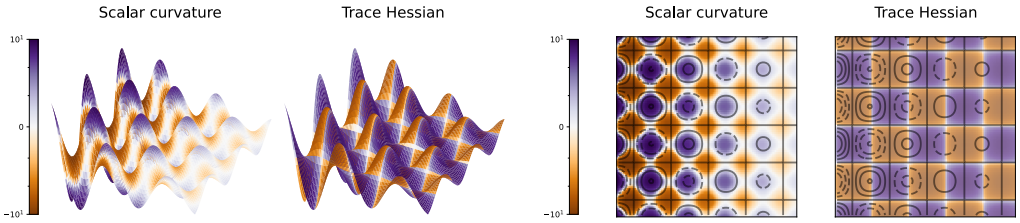


Figure 5.2: In this figure, the loss function is defined as $f(u, v) = e^{-cu} \sin(u) \sin(v)$, with $c = 0.1$. The first two figures represent the surface in 3d, while the two last figures represent the surface seen from above, in the $\{u, v\}$-space. Both the scalar curvature and the trace of the Hessian are shown through the gradient of color.

The expected flatness over mini-batches

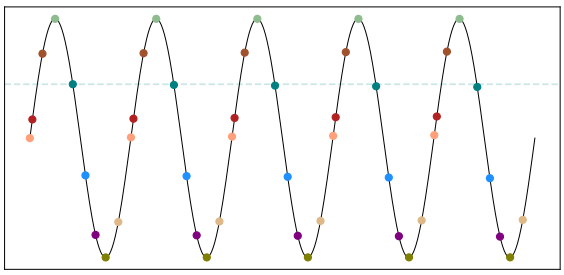


Figure 5.3: The data points fit a sinus. The dataset is split into 7 batches of different colors. If the flatness is defined as $\mathrm{tr}(\mathbf{H})$, the flatness over the entire dataset is equal to the expectation of the flatness of a batch. Thus, the curve is considered flat.

Another challenge emerges when the dataset is divided into small batches. If we choose the Hessian's trace as the measure of flatness, the overall flatness of the entire dataset equals the average flatness over these batches (Equation 5.6). This could potentially induce the wrong conclusion depending on the method used to partition the dataset: In Figure 5.3, the dataset is split in such a way that the trace of the Hessian is null for each batch, which means that the curve is considered as flat over the entire dataset.

The dataset, denoted $\mathcal{D}$, is split into k mini-batches: $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$. By linearity, the Hessian of the loss function over the entire dataset can be written as the mean of the

Hessian of mini-batches i.e.:

$$\mathbf{H}_{\mathcal{D}} = \frac{1}{k}\sum_i \mathbf{H}_{\mathcal{B}_i}$$

As a consequence, since the trace commutes with a summation, we have: $\mathrm{tr}(\mathbf{H}_{\mathcal{D}}) = \mathrm{tr}(\frac{1}{k}\sum_i \mathbf{H}_{\mathcal{B}_i}) = \frac{1}{k}\sum_i \mathrm{tr}(\mathbf{H}_{\mathcal{B}_i}) = \mathbb{E}[\mathrm{tr}(\mathbf{H}_{\mathcal{B}_i})]$. The trace of the Hessian of the loss function over the entire dataset is the expectation of the Hessian over mini-batches:

$$\mathrm{tr}(\mathbf{H}_{\mathcal{D}}) = \mathbb{E}[\mathrm{tr}(\mathbf{H}_{\mathcal{B}_i})] \tag{5.6}$$

The corresponding result does not hold for the scalar curvature in general.

Proposition 5.3.1
The scalar curvature of the hessian of the full dataset is not equal to the expectation of the Scalar curvature over mini-batches. That is there exists a dataset, $\mathcal{D}$, and mini-batches, $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ such that:

$$\mathrm{S}(\mathbf{H}_{\mathcal{D}}) \neq \mathbb{E}[\mathrm{S}(\mathbf{H}_{\mathcal{B}_i})]$$

Proof. See Appendix A.2. □

5.3.2 The scalar curvature supports previous theoretical findings through the Hessian norm

Although the two previous given examples suggest that in some cases, the trace of the Hessian is not a good definition of flatness, it is associated with the optimization process and the model's capacity to generalize in various ways. We will observe that under certain circumstances, the scalar curvature simplifies to the Hessian norm.

Perturbations on the weights

Seong et al. (2018) showed that the robustness of the loss function to inputs perturbations is related to the Hessian. We similarly show that the resilience of the loss function to weights perturbations is upper bounded by the norm of the Hessian. Additionally, a smaller scalar curvature implies stronger robustness.

Proposition 5.3.2
Let $\mathbf{x}_{\min}$ an extremum, ε, a small scalar ($\varepsilon \ll 1$) and $\mathbf{x}$ a normalized vector ($\|\mathbf{x}\| = 1$). The trace of the square of the Hessian is an upper bound to the difference of the loss functions when perturbed by the weights:

$$\|f(\mathbf{x}_{\min} + \varepsilon\mathbf{x}) - f(\mathbf{x}_{\min})\|_2^2 \leq \frac{1}{4}\varepsilon^4 \,\mathrm{tr}(\mathbf{H}_{\min}^2) \tag{5.7}$$

Proof. This is obtained by applying the Taylor expansion, for a very small perturbation $\varepsilon \ll 1$. See Appendix A.2 for the full proof. □

Let us assume two minima $\mathbf{x}_1$ and $\mathbf{x}_2$, and we suppose that the loss function at $\mathbf{x}_1$ is flatter than the one at $\mathbf{x}_2$ in terms of scalar curvature so $0 \leq \mathrm{S}(\mathbf{x}_1) \leq \mathrm{S}(\mathbf{x}_2)$. Being

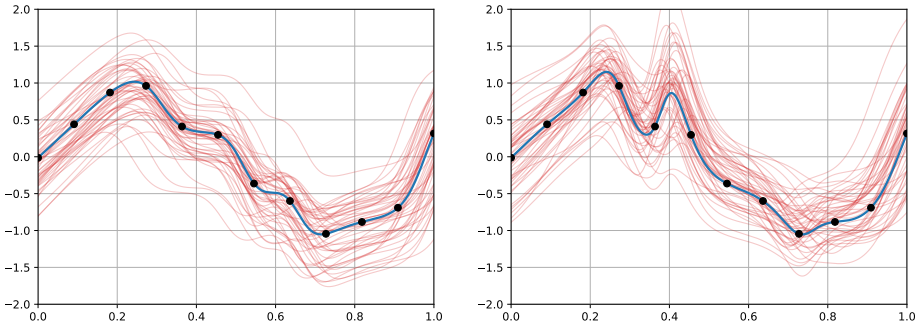


Figure 5.4: Empirical demonstration of Proposition 5.3.2. We train two identical and differently initialized deep nets using the same optimizer (Adam). We then perturb point wise the learned weights using Gaussian noise $\mathcal{N}(0, 0.1^2)$. As expected the model on the left with scalar curvature ≈ 430 is more robust to perturbations compared to the right model with scalar curvature ≈ 610.

at the minimum implies that $\mathrm{S}(\mathbf{x}_1) = \mathrm{tr}(\mathbf{H}_1)^2 - \mathrm{tr}(\mathbf{H}_1^2)$ and $\mathrm{S}(\mathbf{x}_2) = \mathrm{tr}(\mathbf{H}_2)^2 - \mathrm{tr}(\mathbf{H}_2^2)$ respectively. Then:

$$0 \leq \mathrm{S}(x_1) \leq \mathrm{S}(x_2) \iff 0 \leq \mathrm{tr}(\mathbf{H}_1)^2 - \mathrm{tr}(\mathbf{H}_1^2) \leq \mathrm{tr}(\mathbf{H}_2)^2 - \mathrm{tr}(\mathbf{H}_2^2) \tag{5.8}$$

$$\Rightarrow \mathrm{tr}(\mathbf{H}_1^2) \leq \mathrm{tr}(\mathbf{H}_2^2). \tag{5.9}$$

A flatter minimum $\mathrm{S}(\mathbf{x}_1) \leq \mathrm{S}(\mathbf{x}_2)$ leads to more robustness of the loss function to weights perturbations: $\|f(\mathbf{x}_1 + \varepsilon\mathbf{x}) - f(\mathbf{x}_1)\|_2^2 \leq \|f(\mathbf{x}_2 + \varepsilon\mathbf{x}) - f(\mathbf{x}_2)\|_2^2$.
In Figure 5.4, we consider $\varepsilon \sim \mathcal{N}(0, 0.01)$ to be a small perturbation, and we plotted the original loss function with the perturbed losses. We computed the $\mathrm{tr}\,\mathbf{H}^2$ at the minimum. When the scalar curvature is smaller, the variance across the perturbations at the minimum is smaller and the perturbations are more centered around the original loss function.

Efficiency of escaping minima

Stochastic gradient descent can be conceptualized as an Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein 1930), which is a continuous-time stochastic process that characterizes the behavior of a particle influenced by random fluctuations (Mandt et al. 2017). By considering the non-linear relationship between the weights and the covariance, the update rules in gradient descent resemble the optimization approach employed in the multivariate Ornstein-Uhlenbeck process. When approximating the covariance by the Hessian (Jastrzebski et al. 2017, Appendix A), the gradient descent can be seen as an Ornstein-Uhlenbeck process with:

$$\mathrm{d}\mathbf{x}_t = -\mathbf{H}\mathbf{x}_t \mathrm{d}t + \mathbf{H}^{\frac{1}{2}} \mathrm{d}W_t \tag{5.10}$$

The escaping efficiency measure is a metric used to evaluate the performance of optimization algorithms, including gradient descent, in escaping from local minima and finding the global minimum of the loss function, and is defined as $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}_{\min})]$. Zhu et al. (2018) used this definition and the expression of the gradient descent process (Equation 5.10) to approximate the escaping efficiency:

$$\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}_{\min})] \approx \frac{t}{2} \mathrm{tr}(\mathbf{H}^2). \tag{5.11}$$

Similar to the example above, gradient descent will have more difficulties to escape from a minimum with a small scalar curvature, and so it will converge more quickly to the

flat minima.

The scalar curvature is the squared norm of the Hessian in over-parametrized neural networks

Proposition 5.3.3
We note $\mathbf{H}$ the Hessian of the loss of a model with q parameters, and S the scalar curvature, obtained in Proposition 5.2.3 and Corollary 5.2.4. When we reach a flat minimum, supposing the eigenvalues of $\mathbf{H}$ are similar, for a high number of parameters q, we have:

$$\mathrm{S}(\mathbf{x}_{\min}) \underset{q\to\infty}{\sim} \mathrm{tr}(\mathbf{H})^2$$

Proof. Let us suppose that, at a flat minimum, all the eigenvalues are similar: $\lambda_1 = \cdots = \lambda_q = \lambda \geq 0$. Then we, have $\|\mathbf{H}\|_*^2 = q^2\lambda^2$ and $\|\mathbf{H}\|_F^2 = q\lambda^2$. When the number of parameters increases, $\|\mathbf{H}\|_F^2 = o(\|\mathbf{H}\|_*^2)$, and as a consequence $\|\mathbf{H}\|_*^2 - \|\mathbf{H}\|_F^2 \sim \|\mathbf{H}\|_*^2$. □

In this proposition, we assume that all the eigenvalues are similar. This strong assumption is supported by empirical results (Ghorbani et al. 2019). The empirical results show that during the optimization process, the spectrum of the eigenvalues becomes entirely flat, especially when the neural network includes batch normalization.

5.3.3 Reparameterization of the parameter space

The main argument challenging the link between flatness and generalization is that the flatness definitions, so far, are not *invariant under reparameterization*. Reparameterization refers to a change in the parametrization of the model, which can be achieved by transforming the original parameters (θ) into a new set of parameters (η). Even if we assume that the models have the same performance: $\{f_\theta, \theta \in \Theta \subset \mathbb{R}^q\} = \{f_{\varphi(\eta)}, \eta \in \varphi^{-1}(\Theta)\}$, this reparameterization alters the shape of the loss function landscape in $\mathbb{R}^q$. This is the core of the problem: Dinh et al. (2017) compared the flatness of f_θ and $f_{\varphi(\eta)}$ with respect to the same ambient space $\mathbb{R}^q$, while each measure should be defined, and compared, relative to their respective parameter space, and not to an arbitrary space of reference.

The scalar curvature is not invariant under reparameterization of the parameter space, and it should not be. It is, however, an **intrinsic** quantity, which means that it does not depend on an ambient space. As a consequence, it is also equivariant under diffeomorphism, and notably, if $\mathcal{M}$ and $\mathcal{M}'$ are two Riemannian manifolds related by an isometry $\Psi : \mathcal{M} \to \mathcal{M}'$, then $\mathrm{S}(\mathbf{x}) = \mathrm{S}(\Psi(\mathbf{x}))$, for all $\mathbf{x} \in \mathcal{M}$.

In the case of the scalar curvature, if we apply a diffeomorphism to the parameters space with $\varphi : \mathcal{M} \to \mathcal{M}'$, and $f : \mathcal{M}' \subset \mathbb{R}^q \to \mathbb{R}$ the loss function, then:

$$\mathbf{H}(f \circ \varphi) = \mathbf{J}(\varphi)^\top \mathbf{H}(f)\mathbf{J}(\varphi) + \mathbf{J}_k(f)\mathbf{H}^k(\varphi),$$

with $\mathbf{J}(\varphi)$, $\mathbf{J}(f)$ the Jacobian of φ and f, and $\mathbf{H}(f \circ \varphi)$, $\mathbf{H}(f)$ and $\mathbf{H}^k(\varphi)$ the Hessian of $f \circ \varphi$ and f. We note $\mathbf{H}^k(\varphi)_{ij} = \partial_i \partial_j \varphi^k$ the Hessian of the k-th component of φ.

At the minimum of the loss function, $\mathbf{J}(f) = 0$, with $\varphi : \mathcal{M} \to \mathcal{M}'$ a diffeomorphism, and $\mathbf{x}' = \varphi(\mathbf{x})$, the scalar curvatures on $\mathcal{M}$ and $\mathcal{M}'$ is derived as:

$$\begin{aligned}
\mathrm{S}(\mathbf{x}) &= \|\mathbf{H}_f\|_*^2 - \|\mathbf{H}_f\|_F^2 \,, \\
\mathrm{S}(\mathbf{x}') &= \left\|\mathbf{J}_\varphi \mathbf{J}_\varphi^\top \mathbf{H}_f\right\|_*^2 - \left\|\mathbf{J}_\varphi \mathbf{J}_\varphi^\top \mathbf{H}_f\right\|_F^2 .
\end{aligned}$$

5.4 Discussion

Our research focused on analyzing the loss landscape as a Riemannian manifold and its connection to optimization generalization. We introduced a Riemannian metric on the parameter space and examined the scalar curvatures of the loss landscape. We found that the scalar curvature at minima is defined as the difference between the nuclear and Frobenius norm of the Hessian of the loss function.

The *flatness hypothesis* forms the basis of our study, suggesting that flat minima lead to better generalization compared to sharp ones. The Hessian of the loss function is known to be crucial in understanding optimization. However, analyzing the spectrum of the Hessian, particularly in over-parametrized models, can be challenging. As a result, the research community has started relying on the norm of the Hessian. We show that, in certain scenarios, the Hessian norm doesn't effectively gauge flatness, whereas scalar curvature does. Despite this, the Hessian norm is still relevant to theoretical results in optimization, including the model's stability against perturbations and the algorithm's ability to converge. Similarly, these characteristics are also satisfied by the scalar curvature. In essence, the scalar curvature combines all the advantages of the Hessian norm while accurately describing the curvature of the parameter space.

Future research could explore the curvature within stochastic optimization and investigate the scalar curvature as a random variable affected by the underlying data and batch distribution. It would also be interesting to understand how the scalar curvature relates to the stochastic process and whether it is connected to any implicit regularization in the model.

Overall, our study contributes to the understanding of the loss function's parameter space as a Riemannian manifold and provides insights into the curvature properties that impact optimization and generalization.

Identifying latent distances with Finslerian geometry

6

The paper *Identifying latent distances with Finslerian geometry* has been written with David Eklund, Carl Henrik Ek, and Søren Hauberg. It was presented at the NeurIPS 2022 Workshop "Symmetry and Geometry in Neural Representations". The paper is currently under review at the Transactions on Machine Learning Research, and it is available publicly at https://arxiv.org/abs/2212.10010.

Abstract

Riemannian geometry provides us with powerful tools to explore the latent space of generative models while preserving the underlying structure of the data. The latent space can be equipped it with a Riemannian metric, pulled back from the data manifold. With this metric, we can systematically navigate the space relying on geodesics defined as the shortest curves between two points.

Generative models are often stochastic causing the data space, the Riemannian metric, and the geodesics to be stochastic as well. Stochastic objects are at best impractical, and at worst impossible, to manipulate. A common solution is to approximate the stochastic pullback metric by its expectation. But the geodesics derived from this expected Riemannian metric do not correspond to the expected length-minimizing curves.

In this work, we propose another metric whose geodesics explicitly minimize the expected length of the pullback metric. We show this metric defines a Finsler metric, and we compare it with the expected Riemannian metric. In high dimensions, we prove that both metrics converge to each other at a rate of $\mathcal{O}\left(\frac{1}{D}\right)$. This convergence implies that the established expected Riemannian metric is an accurate approximation of the theoretically more grounded Finsler metric. This provides justification for using the expected Riemannian metric for practical implementations.

6.1 Introduction

Generative models provide a convenient way to learn low-dimensional latent variables z corresponding to data observations x through a smooth function $f : \mathcal{Z} \subset \mathbb{R}^q \to \mathcal{X} \subset \mathbb{R}^D$, such that $x = f(z)$. Through this learned manifold, one can generate new data or compare observations by interpolating or computing distances. However, doing so by using the Euclidean distance in the latent space is misleading (Hauberg 2018a), because the latent variables are not statistically identifiable. If our observations are lying near a manifold (Fefferman et al. 2016), we want to equip our latent space with a metric that preserves distance measures on it. Figure 6.1 (left panel) illustrates the need for defining geometric-aware distances on manifolds.

Distances on a manifold can be precisely defined using a norm, which is a mathematical function that exhibits several desirable properties such as non-negativity, homogeneity, and the triangle inequality. In particular, a norm can be induced by an inner product (i.e., a quadratic function) that associates each pair of points on the manifold with a scalar value.

To derive the standard Riemannian interpretation of the latent space, we first compute the infinitesimal Euclidean norm according to the data space. Using the Taylor expansion, we have: $\|f(z+\Delta z) - f(z)\|_2^2 \approx \|f(z) + J(z)\Delta z - f(z)\|_2^2 = \Delta z^\top J(z)^\top J(z)\Delta z$. As a first approximation, the norm defined in the latent space locally preserves the Euclidean norm defined in the data space. The curvature of our data manifold is condensed in the Riemannian metric tensor $G_z = J^\top(z)J(z)$, which serves as a proxy to define the Riemannian metric: $g_z : (u, v) \to u^\top G_z v$. In mathematical jargon, we say that the Riemannian manifold $(\mathcal{Z}, g)$ is obtained by pulling back the Euclidean metric through the map f.

Riemannian geometry enables the exploration of the latent space in precise geometric terms, and quantities of interest such as the length, the energy or the volume can be directly derived from the pullback metric. These geometric quantities are, by construction, known to be invariant to reparameterizations of the latent space $\mathcal{Z}$, and are thus statistically identifiable (Hauberg 2018a). While this geometric framework exclusively handles deterministic objects, generative models are often stochastic. The learned map f, that mathematically describes those models, is stochastic too. For example, in the celebrated Gaussian Process Latent Variable Model (GPLVM) (Lawrence 2003), this function is a Gaussian process. The pullback metric is then stochastic, and standard differential geometric constructions no longer applies. Navigating the latent manifolds through geodesics (length-minimizing curves) becomes practically infeasible.

Previous research tried to circumvent this problem by approximating the stochastic pullback metric with the expected value of the Riemannian metric tensor, but the derived length are unintuitive quantities that do not correspond to the expected length:

$$\mathcal{L}(\gamma)\,|_{\mathbb{E}[G]} := \int \|\gamma(t)\|_{\mathbb{E}[G]}\; dt \quad \neq \quad \mathbb{E}\left[\mathcal{L}(\gamma)\,|_G\right] := \int \mathbb{E}\,\|\gamma(t)\|_G\; dt.$$

Instead of taking the expectation of the metric tensor, which serves as a surrogate to define a norm, we propose to take the expectation of the norm directly.

In this paper, we compare our expected norm with the norm induced by the commonly used expected metric tensor. The main findings are:

1. The expected norm defines a Finsler metric. Finsler geometry is a generalization of Riemannian geometry.
2. For Gaussian processes, the stochastic norm obtained through the pullback metric follows a non-central Nakagami distribution, so our Finsler metric has a closed-form expression.
3. In high dimensions, for Gaussian processes, our Finsler metric and the previously studied Riemannian metric converge to each other at a rate of $\mathcal{O}\left(\frac{1}{D}\right)$, with D the dimension of the data space.

We conclude that the Riemannian metric serves as a good approximation of the Finsler metric, which is theoretically more grounded. It further justifies the use of the Riemannian metric in practice when exploring stochastic manifolds.

6.1.1 Outline of the paper

The paper explores the geometry of latent spaces learned by generative models, which encode a latent low-dimensional manifold that represents observed high-dimensional data. The latent manifold is denoted $\mathcal{Z} \subset \mathbb{R}^q$ and the data manifold is denoted $\mathcal{X} \subset \mathbb{R}^D$.

Assuming the manifold hypothesis holds, we need to define an infinitesimal norm in the latent manifold to compute distances that respect the underlying geometry of the data. Such a norm can be constructed by pulling back the Euclidean distance through

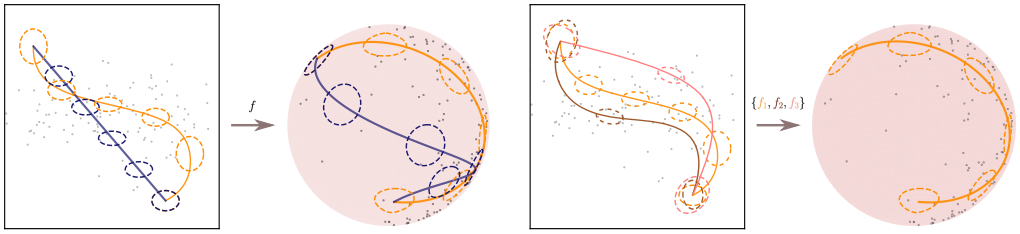


Figure 6.1: In this illustration, we show the 2-dimensional representation (θ, φ) of the sphere parametrized in $\mathbb{R}^3$ with $f : (\theta, \varphi) \to \cos(\theta)\sin(\varphi), \sin(\theta)\cos(\varphi), \sin(\varphi)$. In practice, we don't have access to well-parametrized manifolds, but instead those manifolds are being shaped by data points in $\mathbb{R}^D$. The data points are depicted as tiny dots for illustrative purposes. On the left figure, the map f is deterministic, and on the right figure, we imagine that f is stochastic. When we pullback, through f, a circle from the sphere to the plane, the circle is deformed to an ellipse. Those ellipses, called indicatrices, are the fingerprints of the metric, and they showcase the deformation of the plane. **Left figure**: In blue, the Euclidean distance, is represented by the straight line in the plane. It is not identifiable, and it does not represent a geodesic on the sphere. In orange, the length-minimizing curve obtained through the pullback metric of the mapping f leads to a great circle on the sphere. Following the great circle is the fastest way to go from one point to another. Effectively, the pullback metric leads to a geodesic, contrary to the Euclidean metric. Notice that, in the plane, the orange geodesic also follows the direction of the indicatrices, since it is length-minimizing. **Right figure**: We imagine that a generative model maps latent space to data space to the data space using a stochastic process $f = \{f_1, f_2, \dots\}$. This leads to a stochastic pullback metric, stochastic indicatrices, and stochastic paths in the latent space.

the smooth function that maps the latent manifold to the data manifold. This map, $f : \mathcal{Z} \to \mathcal{X}$, mathematically describes the decoder of a trained generative model. Those models being often stochastic, we consider f being a stochastic process. It means that the pullback metric tensor, $G = J^\top J$, and its induced norm, $\|\cdot\|_G : u \to \sqrt{u^\top G u}$, are also stochastic. In section 6.2, we mathematically define the notion of stochastic pullback metric and stochastic manifolds.

To circumvent all the challenges posed by this stochastic component, a deterministic approximation of the norm is needed. It can be defined by taking the expectation of the metric tensor. This norm, that we will note $\|\cdot\|_R : u \to \|u\|_{\mathbb{E}[G]}$, has been studied before by Tosi et al. (2014), and is explained in section 6.2.2. In this paper, we propose instead to directly take the expectation of the stochastic norm. This expected norm, noted $\|\cdot\|_F : u \to \mathbb{E}\left[\|u\|_G\right]$, is introduced in section 6.2.4. The norm $\|\cdot\|_R$ is defined by a Riemannian metric, and we show that the norm $\|\cdot\|_F$ defines a Finsler metric. We explain the general difference between Finsler and Riemannian geometry in section 6.3.

The aim of this paper is to compare those two norms. We first draw absolute bounds in section 6.3.2, and then relative bounds in 6.3.3. We also investigate the relative difference of the norms when the dimension of the data space increase, in section 6.3.4. Finally, we perform some experiments in Section 6.4, that illustrates the Riemannian and Finsler norm in the same latent space.

6.1.2 Related works

Riemannian geometry for machine learning

When navigating a learned latent space, the geodesics obtained through the pullback metric not only follow geometrically coherent paths, they also remain invariant under different representations. While two runs of the same model will produce distinct latent manifolds, the geodesics connecting the same chosen points should have the same length. We say that the pullback metric effectively solved the identifiability problem (Hauberg

2018a). This has led to the growing adoption of Riemannian geometry in machine learning applications. In robotics, Beik-Mohammadi et al. (2021) used the pullback metric from a variational autoencoder to safely navigate the space of motion patterns, and Scannell et al. (2021a) use geodesics under the expected metric in a GPLVM to control quadrotor robot. In proteins modelling, Detlefsen et al. (2022) showed that the derived geodesics on the space of beta-lactamase follow their phylogenetic tree structure. In game content generation, González-Duque et al. (2022) generated game levels more coherent and reliable by interpolating data along the geodesics. Jørgensen and Hauberg (2021) used the expected Riemannian metric from observations of pairwise distances through a GPLVM to build a probabilistic dimensionality reduction model.

Finsler geometry in machine learning

Our work crucially relies on Finslerian geometry, which has been well-studied mathematically, but has only seen very limited use in machine learning and statistics. We point to two notable exceptions, which are quite distinct from our work. Lopez et al. (2021) use symmetric spaces to represent graphs and endow these with a Finsler metric to capture dissimilarity structure in the observational data. Ratliff et al. (2021) discuss the role of differential geometry in motion planning for robotics. Along the way, they touch upon Finslerian geometry, but mostly as a neat tool to allow for generalizations. To the best of our knowledge, no prior work has investigated the links between stochastic and Finslerian geometry.

Strategies to deal with stochastic Riemannian geometry

Tosi et al. (2014) and Arvanitidis et al. (2018) introduced approximation of the pullback metric by taking the expectation of the metric tensor. In those two cases, the map f is respectively a trained Gaussian process, or the decoder of a VAE. In this paper, the derivations only hold if f is a smooth stochastic process (Definition 6.2.2), which is not the case* of the VAEs, and hence, our results are not directly applicable to those models.

In addition to the work of Tosi et al. (2014), a solution to circumvent the randomness of the metric tensor is to consider that the data follows a specific probability distribution. Instead of looking at the shortest path on the data manifold, Arvanitidis et al. (2022) borrow tools from information geometry and consider the straightest paths on the manifold whose elements are probability distributions.

6.2 Expectation on random manifolds

The metric pulled back by a stochastic mapping is, de facto, stochastic and endows a random manifold. Unfortunately, we are not yet equipped to derive geometric objects on a random manifold. Instead, we dodge this problem by seeking a deterministic approximation of this stochastic metric. As mentioned above, a common solution is to approximate such a metric by its expectation. In section 6.2.2, we study the expected Riemannian metric.

* The decoder of a VAE, while it decodes to a Gaussian, cannot be considered as a differentiable stochastic process. One reason is because the independence of the probability of the data: $p(x|z) = \prod_{i=1}^{n} p(x_i|z_i)$. Let us assume the opposite: the decoder is a Gaussian process. The covariance of the Gaussian process would be a diagonal matrix because of the independence of the probability of the data. The covariance would correspond to a Dirac distribution: $\text{cov}(x_i, x_j) = \delta_{ij}$. However, a stochastic process is differentiable only if the covariance is differentiable, which is not the case of the Dirac distribution.

The solution suggested by this paper is to approximate the expectation of the lengths instead of the random metric itself. In section 6.2.4, we show that this new metric is not Riemannian but Finslerian (Proposition 6.2.2), and it has a closed-form expression when the map f is a Gaussian process (Proposition 6.2.3).

6.2.1 Random Riemannian geometry

The pullback metric is defined as a Riemannian metric if and only if the mapping f is an immersion, which is a differentiable function whose derivatives are injective everywhere on the manifold (Lee 2013, Proposition 13.9). A manifold equipped with a Riemannian metric is called a Riemannian manifold.

Definition 6.2.1
The pullback of the Euclidean metric through the immersion $f : \mathcal{Z} \to \mathcal{X}$ is a **Riemannian metric**. It is defined as the inner product $g_z : (\mathcal{T}_z\mathcal{Z}, \mathcal{T}_z\mathcal{Z}) \to \mathbb{R}_+ : (u, v) \to u^\top G v$, at a specific point z in the manifold $\mathcal{Z}$. u and v are vectors lying in the tangent plane $\mathcal{T}_z\mathcal{Z}$ (i.e.: the set of all tangent vectors) of the manifold. $G = J^\top J$, with J the Jacobian of f.

Since a Riemannian metric is an inner product, it induces a norm: $\|\cdot\|_G$. We can then define the **curve length** and **curve energy** on a manifold: $L_G(\gamma) = \int_0^1 \|\dot\gamma(t)\|_G \, dt$ and $E_G(\gamma) = \frac{1}{2}\int_0^1 \|\dot\gamma(t)\|_G^2 \, dt$, with γ a curve defined on $\mathcal{Z}$, and $\dot\gamma$ its derivative. A locally length-minimizing curve between two connecting points is a **geodesic**. To obtain a geodesic, we can minimize the curve length, but in practice minimizing the curve energy is more efficient. On the manifold, we also define a **volume measure** in order to integrate probability functions: for $\mathcal{U} \subset \mathcal{Z}$, $\int_{f(\mathcal{U})} h(x)\, dx = \int_{\mathcal{U}} h(f(z)) V_R \, dz$, with $V_R(z) = \sqrt{G_z}$ the volume measure.

In addition, we are considering the case where the immersion f is a stochastic process. The outputs of our trained model, $x \in \mathcal{X}$, which represent our data, are random variables.

Definition 6.2.2
A **stochastic process** is a collection of random variables $\{X(t, \omega), t \in T\}$ indexed by an index set T defined on a sample space Ω, which represents the set of all possible outcomes. An outcome in Ω is denoted by ω, and a realization of the stochastic process is the sequence of $X(\cdot, \omega)$ that depends on the outcome ω.

In this framework, our index set is our latent manifold $T = \mathcal{Z}$, and our sample space Ω is defined as the set of the model evaluations. For every point $z \in \mathcal{Z}$, every time we execute our model, the output $x = f(z)$ is a random variable following a specific distribution. When the data x follow a Gaussian distribution, the stochastic process is called a **Gaussian process**. A GP-LVM (Lawrence 2003) is a model that learns how to map the data from a latent space to a data space through a Gaussian process.

When f is a stochastic immersion, the metric tensor becomes a random matrix. In this paper, we call a manifold equipped with the stochastic pullback metric a **random manifold**, noted $(\mathcal{Z}, g)$. As a consequence of the stochastic aspect of the metric, all the functionals are stochastic themselves, and they are no longer trivial to manipulate.

Definition 6.2.3
A **random Riemannian metric tensor** is a matrix-valued random field (i.e.: a collection of matrix-valued random variables $\{G(z, \omega), z \in \mathcal{Z}\}$), whose realization for

a specific evaluation $\omega \in \Omega$ is a Riemannian metric tensor. A **random Riemannian metric** is a metric induced by a random Riemannian metric tensor: $g_z : (\mathcal{T}_z\mathcal{Z}, \mathcal{T}_z\mathcal{Z}) \to \mathbb{R}_+ : (u, v) \to u^\top G v$. For the rest of the paper, the associated **stochastic norm** is noted:

$$\|\cdot\|_G : \mathcal{T}_z\mathcal{Z} \to \mathbb{R}_+ : u \to \sqrt{g_z(u,u)} \coloneqq \sqrt{u^\top G u}$$

If this stochastic norm is induced by f defined as a Gaussian process, then $\|\cdot\|_G$ follows a non-central Nakagami distribution. This is explained in the proof of Proposition 6.2.3.

6.2.2 Norm induced by the expected metric tensor

One way to approximate a random metric tensor is to take its expectation with respect to the collection of random metrics induced by the stochastic process. This has been introduced before by Tosi et al. (2014) GP-LVMs.

Definition 6.2.4
Let G be a stochastic Riemannian metric tensor on the manifold $\mathcal{Z}$. We refer to $\mathbb{E}[G]$ as the **expected metric tensor**. It induces a Riemannian metric and a norm on $\mathcal{Z}$. We will note the **norm induced by the expected metric tensor** as:

$$\|\cdot\|_R : \mathcal{T}_z\mathcal{Z} \to \mathbb{R}_+ : u \to \|u\|_{\mathbb{E}[G]} \coloneqq \sqrt{u^\top \mathbb{E}[G] u}$$

Like any Riemannian metric, we can define the following functionals: $L_R(\gamma) = \int_0^1 \sqrt{\dot\gamma(t)^\top \mathbb{E}[G]\dot\gamma(t)}\, dt$, $E_R(\gamma) = \int_0^1 \dot\gamma(t)^\top \mathbb{E}[G]\dot\gamma(t)\, dt = \mathbb{E}[E_R(\gamma)]$, and $V_R(z) = \sqrt{\det \mathbb{E}[G]}$.

6.2.3 Expected paths on random manifolds

Approximating the stochastic metric by its expectation seems a natural but also ad-hoc solution. If we want to explore a manifold, we might prefer to use a representative quantity, such as the lengths between data points. The expectation of the lengths can give us an idea about how, on average, two points are connected on a random manifold. The **expected curve length**, and its corresponding **curve energy** on the random manifold (Z, g) are defined as: $L_F(\gamma) = \int_0^1 \mathbb{E}\left[\sqrt{\dot\gamma(t)^\top G \dot\gamma(t)}\right] dt = \mathbb{E}[L_R(\gamma)]$, and $E_F(\gamma) = \int_0^1 \mathbb{E}\left[\sqrt{\dot\gamma(t)^\top G \dot\gamma(t)}\right]^2 dt$.

One observation made by Eklund and Hauberg (2019) is that the length (L_R) derived from the expected Riemannian metric is not equal to the expected curve length (L_F), and their respective energy curves differ by a variance term:

$$\begin{aligned} E_R(\gamma) - E_F(\gamma) &= \int_0^1 \dot\gamma(t)^\top \mathbb{E}[G]\dot\gamma(t) - \mathbb{E}\left[\sqrt{\dot\gamma(t)^\top G \dot\gamma(t)}\right]^2 dt \\ &= \int_0^1 \mathbb{E}\left[\|\dot\gamma(t)\|_G^2\right] - \mathbb{E}\left[\|\dot\gamma(t)\|_G\right]^2 dt = \int_0^1 \mathrm{Var}\left[\|\dot\gamma(t)\|_G\right] dt \end{aligned}$$

This term can be regarded as a **regularization term** for the Riemannian energy curve: the curve energy E_R might be penalized when the curve goes through regions with high-variance. In practice, for a Gaussian process with a stationary kernel, this variance term is upper bounded by the posterior variance that is relatively low next to the training points and is high outside the support of the data. Later, we will also see that the functionals agree in high dimensions, leading to the same geodesics (Section 6.3).

Eklund and Hauberg (2019) also noted that these quantities are bounded by the number of dimensions:

Proposition 6.2.1
(Eklund and Hauberg 2019) Let $f : \mathbb{R}^q \to \mathbb{R}^D$ be a stochastic process such that the sequence: $\{f'_1, f'_2, \ldots, f'_D\}$ has uniformly bounded moments. There is then a constant C such that:

$$0 \leq \frac{L_R - L_F}{L_R} \leq \frac{C}{8D}$$

6.2.4 Expected norm and Finsler geometry

Our work builds on Eklund and Hauberg (2019)'s research. We are interested in approximating the stochastic norm instead of the metric tensor, and by doing so, the derived curve length and curve energy are the same ones studied by Eklund and Hauberg (2019). We go further as we not only compare curve lengths, but the deterministic norms obtained with the stochastic metric.

Definition 6.2.5
Let G be a stochastic Riemannian metric tensor on the manifold $\mathcal{Z}$. It induces a stochastic norm, $\|\cdot\|_G$ on $\mathcal{Z}$. We will note the **expected norm** as:

$$\|\cdot\|_F : \mathcal{T}_z\mathcal{Z} \to \mathbb{R}_+ : u \to \mathbb{E}[\|u\|_G] \coloneqq \mathbb{E}\left[\sqrt{u^\top G u}\right]$$

While it cannot be induced by an inner-product, it is sufficiently convex to be defined as a **Finsler metric**.

Definition 6.2.6
Let $F : \mathcal{T}\mathcal{Z} \to \mathbb{R}_+$ be a continuous non-negative function defined on the tangent bundle $\mathcal{T}\mathcal{Z}$ of a differentiable manifold $\mathcal{Z}$. We say that F is a **Finsler metric** if, for each point z of $\mathcal{Z}$ and v on $\mathcal{T}_z\mathcal{Z}$, we have:

1. **Positive homogeneity**: $\forall \lambda \in \mathbb{R}_+$, $F(\lambda v) = \lambda F(v)$.
2. **Smoothness**: F is a C^∞ function on the slit tangent bundle $\mathcal{T}\mathcal{Z} \setminus \{0\}$.
3. **Strong convexity criterion**: the Hessian matrix $g_{ij}(v) = \frac{1}{2}\frac{\partial^2 F^2}{\partial v^i v^j}(v)$ is positive definite for non-zero v.

A differentiable manifold equipped with a Finsler metric is called a Finsler manifold. Finsler geometry can be seen as an extension of Riemannian geometry, since the requirements for defining a metric are less restrictive.

Proposition 6.2.2
Let G be a stochastic Riemannian metric tensor. Then, the function $F_z : \mathcal{T}_z\mathcal{Z} \to \mathbb{R} : u \to \|u\|_F$ defines a Finsler metric, but it is not induced by a Riemannian metric.

Proof. If F was induced by a Riemannian metric, then this metric would be defined as: $f_z : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}_+ : (v_1, v_2) \to \mathbb{E}[\sqrt{v_1^\top G v_2}]^2$. Since a Riemannian metric is an inner product, it should be symmetric, positive, definite and bilinear. Here, we can see that f_x is not bilinear, so f_z is not a Riemannian metric. However, we can prove that $F_z : \mathbb{R}^q \to \mathbb{R} : v \to \mathbb{E}[\sqrt{v^\top G v}]$ is positive, homogeneous, smooth and strongly convex, and so F_z is a Finsler metric (Shen and Shen 2016, Definition 2.1). For the full proof, see Section B.2.1. □

So far, we have assumed that f is an immersion and a stochastic process. If we consider f to be a **Gaussian Process** in particular, the Finsler norm can be rewritten in a closed form expression.

Proposition 6.2.3
Let f be a Gaussian process and J its Jacobian, with $J \sim \mathcal{N}(\mathbb{E}[J], \Sigma)$. The Finsler norm can be written as:

$$F_z : \mathcal{T}_z\mathcal{Z} \to \mathbb{R}_+ : \|v\|_F \coloneqq v \to \sqrt{2}\sqrt{v^\top \Sigma v}\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})}{}_1F_1\left(-\frac{1}{2}, \frac{D}{2}, -\frac{\omega}{2}\right),$$

with $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J]v)$ a noncentral term, and ${}_1F_1$ the confluent hypergeometric function of the first kind.

Proof. We suppose that f is a Gaussian process, and so is its Jacobian. G follows a non-central Wishart distribution: $G = J^\top J \sim \mathcal{W}_q(D, \Sigma, \Sigma^{-1}\mathbb{E}[J]^\top \mathbb{E}[J])$. $v^\top Gv$ is a scalar and also follows a non-central Wishart distribution: $v^\top Gv \sim \mathcal{W}_1(D, \sigma, \omega)$, with $\sigma = v^\top \Sigma v$ and $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J]v)$ (Kent and Muirhead 1984, Definition 10.3.1). The square-root of a non-central Wishart distribution follows a non-central Nakagami distribution (Hauberg 2018b). Then, by construction, the stochastic norm $\|\cdot\|_G$ follows a non-central Nakagami distribution. The expectation of this distribution is known, and it has a closed-form expression. □

The confluent hypergeometric function of the first kind, also known as the Kummer function, is a special function that is defined as the solution of a specific second-order linear differential equation. The term ω appears from the non-central Wishart distribution. When ω is non-zero, the distribution of the Jacobian shifts away from the origin, and ω represents the magnitude and the direction of this shift, balanced by the correlation between the variables. In Section 6.3.4, to prove our results in high-dimensions, we will assume that our manifold $\mathcal{Z}$ is bounded, and so is ω.

6.3 Comparison of Riemannian and Finsler metrics

6.3.1 Theoretical comparison

In geometry, we need to define a metric (a norm) to compute functionals, and the Riemannian metric is conveniently obtained by constructing an inner product. Because of its bilinearity, it greatly simplifies subsequent computations, but it is also restrictive. Relaxing this assumption and defining a metric as a more general[†] norm has been studied by Finsler (1918), who gave his name to this discipline.

Finsler geometry is similar to Riemannian geometry without the bilinear assumption, most of the functionals (curve length and curve energy) are defined similarly to those obtained in Riemannian geometry. However, the volume measure is different, and there are at least two definitions of volume measure used in Finsler geometry: the Busemann-Hausdorff volume and the Holmes-Thomson volume measure (Wu 2011). In this paper, we decided to focus on the Busemann-Hausdorff definition (Definition 6.3.1), which is more intuitive and easier to derive. If the Finsler metric is a Riemannian metric, the

[†] The norm actually needs to be strongly convex, but it is not necessary symmetric. This means that, for a vector v, we can have a non-reversible Finsler metric: $F_x(v) \neq F_x(-v)$. Intuitively, this means that the path used to connect two points would be different depending on the starting point. This asymmetric property becomes valuable when studying the geometry of anisotropic media (Markvorsen 2016), for example. In our case, our Finsler metric is reversible.

definition of volume naturally coincides with the Riemannian volume measure.

Definition 6.3.1
For a given point z on the manifold, we define the **Finsler indicatrix** as the set of vectors in the tangent space such that the Finsler metric is equal to: $\{v \in \mathcal{T}_z\mathcal{Z} | F_z(v) = 1\}$). We call $\mathbb{B}^n(1)$ the Euclidean unit ball, and vol$(\cdot)$ the standard Euclidean volume. In local coordinates $(e^1, \cdots, e^d)$ on a Finsler manifold $\mathcal{M}$, the **Busemann-Hausdorff volume** form is defined as $dV_F = V_F(z)e^1 \wedge \cdots \wedge e^d$, with:

$$V_F(z) = \frac{\text{vol}(\mathbb{B}^n(1))}{\text{vol}(\{v \in \mathcal{T}_z\mathcal{Z} | F_z(v) < 1\})}.$$

In the definition above, we introduce the notion of *indicatrix*. An indicatrix is a way to represent the distortion induced by the metric on a unit circle. If our metric is Euclidean, we will only have a linear transformation between the latent and the observational spaces, and the indicatrix would still be a circle. Because the Riemannian metric is quadratic, it will always generate an ellipse in the latent space. The Finsler indicatrix, however, would have a convex, even asymmetrical, shape. This difference can be observed in the indicatrix-field represented in Figure 6.2: The Finsler indicatrices in purple can have almost rectangular shape, while the Riemannian indicatrices, in orange, are ellipses.

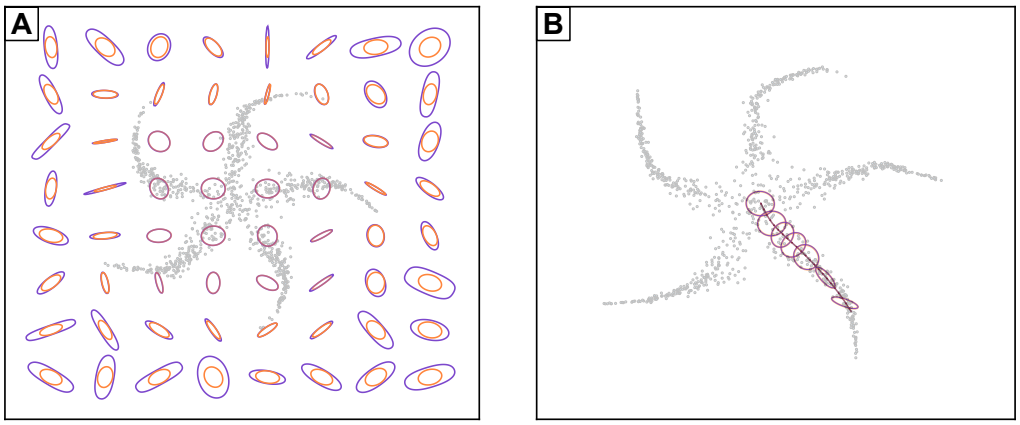


Figure 6.2: Indicatrix field over the latent space of the pinwheel data (in gray) representing the Riemannian (in orange) and Finslerian (in purple) metrics (See Section B.4). (A) The indicatrices are computed over a grid in the latent space. (B) The indicatrices are computed along a geodesic: the Riemannian and Finslerian metrics coincide.

There are also a few observations to note in Figure 6.2. First, in the area of low predictive variance (where data points lie in the latent space), the Finsler and Riemannian indicatrices are alike. This follows from the preceding comment that the metrics diverge by a variance term. If our mapping f was deterministic, both metrics would agree. Second, for every point, the Riemannian indicatrices are always contained by the Finslerian ones, illustrating Proposition 6.3.1 on our absolute bounds in the following section.

6.3.2 Absolute bounds on the Finsler metric

The Finsler norm is upper bounded with the Riemannian norm obtained from the expected metric tensor. It is also lower bounded:

Proposition 6.3.1

We define $\alpha = 2\left(\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})}\right)^2$. The Finsler norm: $\|\cdot\|_F$ is bounded by two norms, $\|\cdot\|_{\alpha\Sigma}$ and $\|\cdot\|_R$, induced by the two respective Riemannian metric tensors: the covariance tensor $\alpha\Sigma_z$ and the expected metric tensor $\mathbb{E}[G_z]$.

$$\forall (z, v) \in \mathcal{Z} \times \mathcal{T}_z Z : \ \|v\|_{\alpha\Sigma} \leq \|v\|_F \leq \|v\|_R$$

Proof. The full proof is detailed in Section B.2.2, and it can be summarized the following way. The upper bound $\|v\|_F \leq \|v\|_R$, also rewritten as: $\mathbb{E}[\sqrt{v^\top G v}] \leq \sqrt{v^\top \mathbb{E}[G] v}$, is obtained by applying Jensen's inequality, knowing that the square root $x \to \sqrt{x}$ is a concave function. The lower bound $\|v\|_{\alpha\Sigma} \leq \|v\|_F$, rewritten as $\sqrt{v^\top \alpha\Sigma v} \leq \mathbb{E}[\sqrt{v^\top G v}]$, is obtained using the closed form expression of the Finsler function. □

The result is illustrated in Figure 6.4 (lower right). Four metric tensors (G_1, G_2, G_3, G_4), each following a non-central Wishart distribution with a specific mean and covariance matrix, have been computed. For each of them, we have drawn the indicatrices $(\{v \in \mathcal{T}_z\mathcal{Z} \mid \|v\| = 1\})$ induced by the norms: $\|\cdot\|_F$, $\|\cdot\|_R$ and $\|\cdot\|_{\alpha\Sigma}$. As expected, we can notice that the $\alpha\Sigma$-indicatrix contains the Finsler indicatrix, itself containing R-indicatrix.

By bounding the Finsler metric, we are able to bound their respective functionals:

Corollary 6.3.2
The length, the energy and the Busemann-Hausdorff volume of the Finsler metric are bounded respectively by the Riemannian length, energy and volume of the covariance tensor $\alpha\Sigma$ (noted $L_{\alpha\Sigma}, E_{\alpha\Sigma}, V_{\alpha\Sigma}$) and the expected metric $\mathbb{E}[G]$ (noted L_R, E_R, V_R):

$$\begin{aligned} \forall z \in \mathcal{Z}, \ L_{\alpha\Sigma}(z) &\leq L_F(z) \leq L_R(z) \\ E_{\alpha\Sigma}(z) &\leq E_F(z) \leq E_R(z) \\ V_{\alpha\Sigma}(z) &\leq V_F(z) \leq V_R(z) \end{aligned}$$

Proof. The full proof is detailed in Section B.2.2. From Proposition 6.3.1, we need to integrate each term of the inequality to obtain the length and the energy. The volume is less trivial, since we use the Busemann-Hausdorff definition for measuring V_F. We have to place ourselves in hyperspherical coordinates, and show that the Finsler indicatrix is still bounded. □

6.3.3 Relative bounds on the Finsler metric

Proposition 6.3.3
Let f be a stochastic immersion. f induces the stochastic norm $\|\cdot\|_G$, defined in Section 6.2. The relative difference between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$0 \leq \frac{\|v\|_R - \|v\|_F}{\|v\|_R} \leq \frac{\mathrm{Var}\left[\|v\|_G^2\right]}{2\mathbb{E}\left[\|v\|_G^2\right]^2}.$$

Proof. This proposition is a direct application of the Sharpened Jensen's inequality (Liao and Berg 2019). □

The previous proposition is valid for any stochastic immersion. We can see that the metrics become equal when the ratio of the variance over the expectation shrinks to zero. This happens in two cases: when the variance converges to zero, which is similar to

having a deterministic immersion, and when the number of dimensions increases. The latter case is investigated below for a Gaussian process [‡].

Proposition 6.3.4
Let f be a Gaussian process. We note $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J]v)$, with J the Jacobian of f, and Σ the covariance matrix of J.
The relative ratio between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$0 \leq \frac{\|v\|_R - \|v\|_F}{\|v\|_R} \leq \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2}.$$

Proof. $v^\top G v$ follows a one-dimension non-central Wishart distribution: $v^\top G_z v \sim \mathcal{W}_1(D, \sigma, \omega)$, with $\sigma = v^\top \Sigma v$ and $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J]v)$. We use the theorem of the moments to obtain both the expectation and the variance, which leads us to the result. □

As we have seen that the metrics are bounded, it is easy to show that the functionals derived from those metrics are also bounded:

Corollary 6.3.5
When f is a Gaussian Process, the relative ratio between the length, the energy and the volume of the Finsler norm (noted L_F, E_F, V_F) and the Riemannian norm (noted L_R, E_R, V_R) is:

$$0 \leq \frac{L_R(z) - L_F(z)}{L_R(z)} \leq \max_{v \in \mathcal{T}_z\mathcal{Z}} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\}$$
$$0 \leq \frac{E_R(z) - E_F(z)}{E_R(z)} \leq \max_{v \in \mathcal{T}_z\mathcal{Z}} \left\{ \frac{2}{D+\omega} + \frac{1+2\omega}{(D+\omega)^2} + \frac{2\omega}{(D+\omega)^3} + \frac{\omega^2}{(D+\omega)^4} \right\}$$
$$0 \leq \frac{V_R(z) - V_F(z)}{V_R(z)} \leq 1 - \left(1 - \max_{v \in \mathcal{T}_z\mathcal{Z}} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\}\right)^q$$

Proof. We directly use Proposition 6.3.4. To obtain the inequalities with the lengths and the energies, we first multiply all the terms by the Riemannian metric, and we integrate every term. To obtain the inequality with the volume, similarly to Corollary 6.3.2, we place ourselves in hyperspherical coordinates and bound the radius of the Finsler indicatrix. The full proof is in Section B.2.2. □

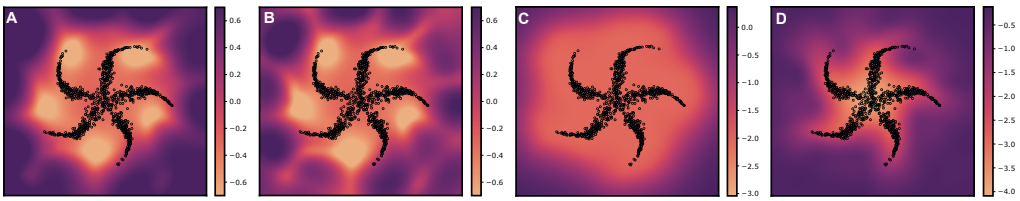


Figure 6.3: Difference of volume for data embedded in the latent space. (A) Riemannian volume measure, (B) Finslerian (Busemann-Hausdorff) volume measure, (C) Variance of the Gaussian process, (D) Ratio between the Riemannian and Finslerian volume: $(V_R(z) - V_F(z))/V_R(z)$. All heatmaps are computed in logarithm scale.

[‡] Interestingly, the term $\mathbb{E}[\nu]^2/\mathrm{Var}[\nu]$, when ν follows a central Nakagami distribution, is called a shape parameter. It has been introduced by Nakagami himself to study the intensity of fading in radio wave propagation (Nakagami 1960). When $\nu := \sqrt{\xi}$ with $\xi \sim \mathcal{W}_1(\Sigma, D)$, then $m = D/2$. This is a particular result obtained from Proposition 6.3.4, when $\mathbb{E}[J] = 0$.

In Figure 6.3, we can compare the volume measures obtained from the Riemannian and Finsler metrics, and in particular, their ratio in the top right image. When the metrics are computed next to the data points in area where the variance is very low, we can see that the ratio of the volume measure is at the order of magnitude 10^{-4}. Further away from the data points, the variance increases and so does the difference between the Riemannian and Finsler volume measures.

6.3.4 Results in high dimensions

Proposition 6.3.4 and Corollary 6.3.5 indicate that the metrics become similar when the dimension (D) of the observational space increases. If we assume that the latent space is a bounded manifold, the metrics converge to each other at a rate of $\mathcal{O}\left(\frac{1}{D}\right)$, as do their functionals.

We assume that the latent manifold is bounded. Then, we can deduce that (1) the term ω, which represents the non-centrality of the data, does not grow faster than the number of dimensions (See lemma B.2.7, in Section B.2.2) and (2) we that the metrics are finite.

Corollary 6.3.6
Let f be a Gaussian Process. In high dimensions, we have:

$$\frac{L_R(z) - L_F(z)}{L_R(z)} = \mathcal{O}\left(\frac{1}{D}\right), \quad \frac{E_R(z) - E_F(z)}{E_R(z)} = \mathcal{O}\left(\frac{1}{D}\right),$$
$$\text{and} \quad \frac{V_R(z) - V_F(z)}{V_R(z)} = \mathcal{O}\left(\frac{q}{D}\right).$$

When D converges toward infinity: $L_R \underset{+\infty}{\sim} L_F$, $E_R \underset{+\infty}{\sim} E_F$ and $V_R \underset{+\infty}{\sim} V_F$.

Proof. This result follows from Corollary 6.3.5, assuming the latent manifold is bounded. The full proof can be found in Section B.2.2. □

Corollary 6.3.7
Let f be a Gaussian Process. In high dimensions, the relative ratio between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$\frac{\|v\|_R - \|v\|_F}{\|v\|_R} = \mathcal{O}\left(\frac{1}{D}\right)$$

And, when D converges toward infinity: $\forall v \in \mathcal{T}_z\mathcal{Z}$, $\|v\|_R \underset{+\infty}{\sim} \|v\|_F$.

Proof. Similarly, from Proposition 6.3.4, in a bounded manifold, both metrics converge to each other in high dimensions. □

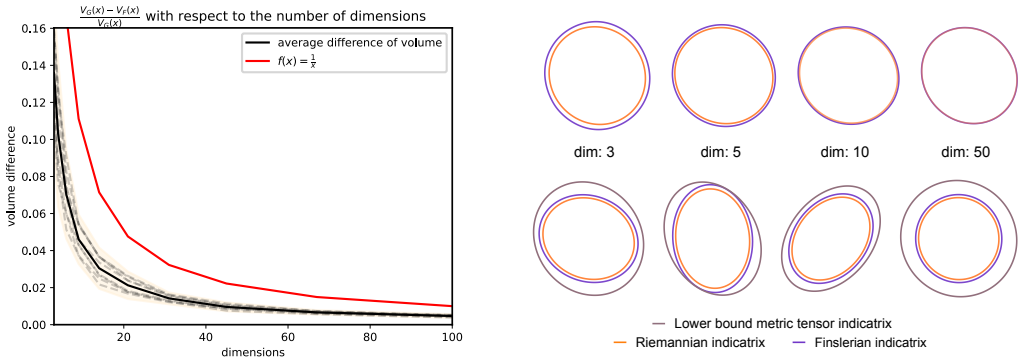


Figure 6.4: Left: Ratio of volumes $(V_R - V_F)/V_R$ decreasing with respect to the number of dimensions. The results were obtained from using a collection of matrices $\{G_i\}$ following a non-central Wishart distribution. Upper right: The Finsler and Riemannian indicatrices converge towards each other when increasing the number of dimensions. Lower right: Illustration of the absolute bounds in Proposition 6.3.1 with the $\alpha\Sigma$-indicatrices, Riemannian indicatrices and Finsler indicatrices.

6.4 Experiments

We want to illustrate cases where these metrics differ in practice. For this, we use two synthetic datasets, consisting of pinwheel and concentric circles mapped to a sphere, and four real-world datasets: a font dataset (Campbell and Kautz 2014), a dataset representing single-cells (Guo et al. 2010), MNIST (LeCun 1998) and fashionMNIST (Xiao et al. 2017). We trained a GPLVM with and without using stochastic active sets (Moreno-Muñoz et al. 2022) to learn a latent manifold. From the learned model, we can access the Riemannian and Finsler metrics, and minimize their respective curve energies to obtain the corresponding geodesics. All the code has been built using Stochman (Detlefsen et al. 2021), a python library to efficiently compute geodesics on manifolds.

6.4.1 Experiments with synthetic data showing high variance

The synthetic data correspond to simple patterns – a pinwheel and concentric circles – that have been projected onto a sphere. In those examples, we are plotting curves that does not follow the data points and so, go through regions of high variance. Notably, the background of the latent manifold represents the variance of the posterior distribution in logarithmic scale. For both cases, we have the 3-dimensional data space and the corresponding learned latent space. The curves are mapped from the latent space to the data space using the forward pass of the GPLVM.

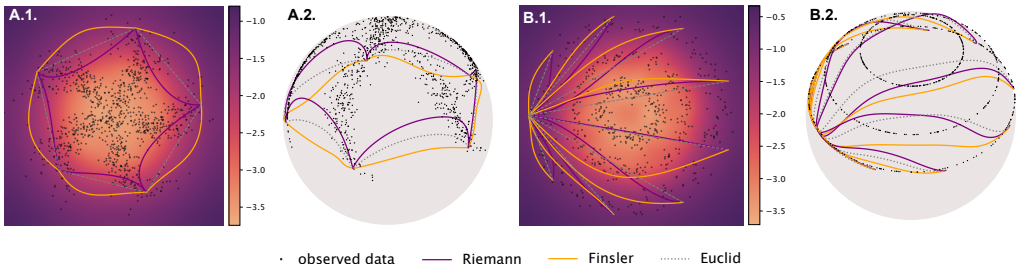


Figure 6.5: The Riemannian (purple), Finslerian (orange) and Euclidean (dotted gray) geodesics obtained by pulling back the metric through the Gaussian processes of a trained GPLVM. The models, trained on the 3-dimensional synthetic data (Figures A.2, B.2) learned their latent representations (Figures A.1, B.1)

We can see that the Riemannian geodesics tend to avoid area of high variance in both cases: they are attracted to the data, at the detriment of following longer paths in $\mathbb{R}^3$.

The Finslerian geodesics, on the opposite, are not perturbed by the variance term, and will explore regions without any data, following shorter paths in $\mathbb{R}^3$. The geodesics have been plotted by computing a discretized manifold, obtained from a 10x10 grid with Detlefsen et al. (2021). This approach proved to be more effective than minimizing the energy along a spline, which was prone to getting trapped in local minima. The GPLVM has been coded in Pyro (Bingham et al. 2019). All the implementation details can be found in the Appendix B.4.

6.4.2 Experiments with a font dataset and qPCR dataset

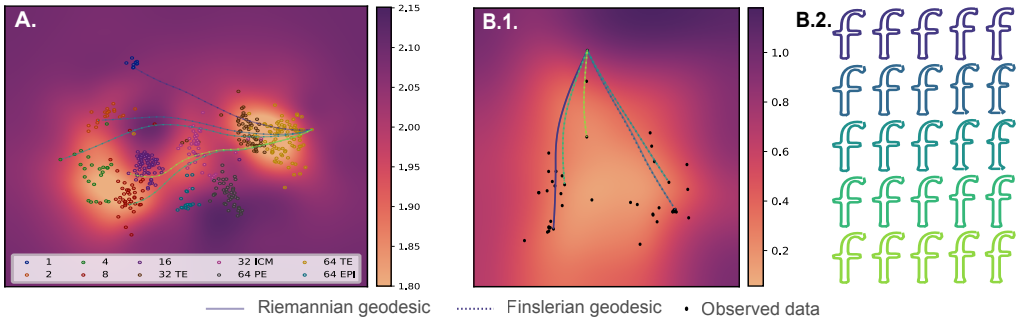


Figure 6.6: The Riemannian (plain line) and Finslerian (dotted line) geodesics obtained by pulling back the metric through the GPLVM. The models trained on single-cells data (which cannot be represented) and font data (Figure B.2) learned their respective 2-dimensional latent representation (Figures A, B.1).

The font dataset (Campbell and Kautz 2014) represents the contour of letters in various font, and the qPCR dataset (Guo et al. 2010) represents single-cells stages. We trained a GPLVM model, also using Pyro (Bingham et al. 2019), to learn the latent space. From the optimized Gaussian process, we can access the Riemannian and Finsler metric, and minimize their respective curve energies to obtain geodesics.
Similar to the previous experiment, the background color represents the variance of the posterior distribution in logarithmic scale. We can notice that the Riemannian and the Finsler geodesics agrees with each other. This experiment also agrees with our finding that in high dimensions (the dimensions of the single-cells data is 48, the font data is 256), both metrics converge to each other. This concludes that, in practice, the Riemannian metric is a good approximation of the Finsler metric, and Finsler geometry doesn't need to be used in high dimensions.

6.4.3 Experiments with MNIST and FashionMNIST

GPLVMs are often hard to train, but they can be scaled effectively using variational inference and inducing points. Using stochastic active sets has been shown to give reliable uncertainty estimates, and the model also learn more meaning latent representations, as shown in the original paper by Moreno-Muñoz et al. (2022). For those experiments, we used those models on two well-known benchmarks: MNIST and FashionMNIST.

In both experiments, because the difference of the variance of the posterior learned by the GPLVM across the latent manifold is low, as we can notice in the background of the plots on the right, and because the data is high dimensional, we cannot see any difference between the Finslerian and the Riemannian geodesics. Again, in practice, the Riemannian metric is a good approximation of the Finslerian metric.

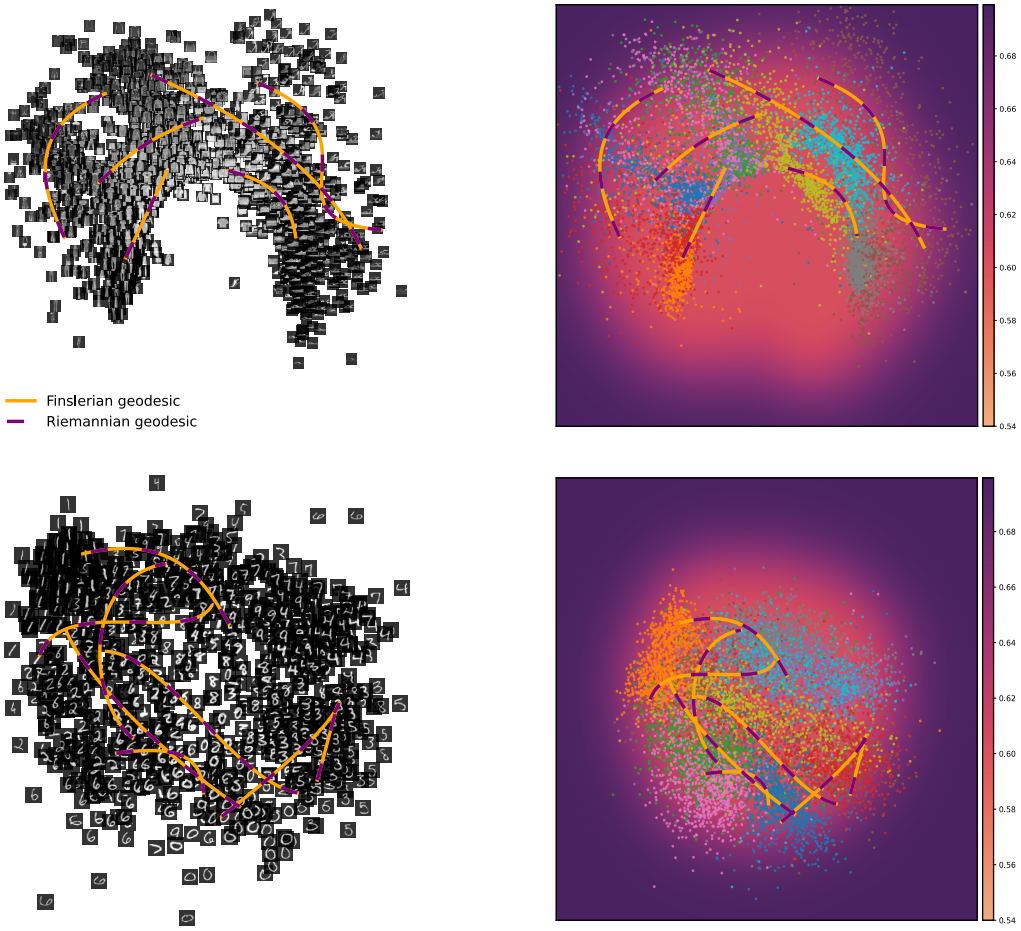


Figure 6.7: The Riemannian (purple) and the Finslerian (orange) geodesics are plotted. **Upper figure:** learned latent space with the images (left) and data points (right) of Fashion MNIST. **Lower figure:** learned latent space with the images (left) and data points (right) of MNIST.

6.5 Discussion

Generative models are often used to reduce data dimension in order to better understand the mechanisms behind the data generating process. We consider the general setting where the mapping from latent variables to observations is driven by a smooth stochastic process, and the sample mappings span Riemannian manifolds. The Riemannian geometry machinery has already been used in the past to explore the latent space.

In this paper, we have shown how curves and volumes can be identified by defining the length of a latent curve as its expected length measured in the observation space. This is a natural extension of classical differential geometric constructions to the stochastic realm. Surprisingly, we have shown that this does not give rise to a Riemannian metric over the latent space, even if sample mappings do. Rather, the latent representation naturally becomes equipped with a Finsler metric, implying that stochastic manifolds, such as those spanned by Latent Variable Models (LVMs), are inherently more complex than their deterministic counterparts.

The Finslerian view of the latent representation gives us a suitable general solution to explore a random manifold, but it does not immediately translate into a practical computational tool. As Riemannian manifolds are better understood computationally than Finsler manifolds, we have raised the question: How good an approximation of the Finsler metric can be achieved by a Riemannian metric? The answer turns out to be: quite good. We have shown that as data dimension increases, the Finsler metric becomes

increasingly Riemannian. Since LVMs are most commonly applied to high-dimensional data (as this is where dimensionality reduction carries value), we have justification for approximating the Finsler metric with a Riemannian metric such that computational tools become more easily available. In practice, we find that geodesics under the Finsler the Riemannian metric are near identical, except in regions of high uncertainty.

Acknowledgments.

This work was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). It also received funding from the European Research Council (ERC) under the European Union Horizon 2020 research, innovation program (757360). SH was supported in part by research grants (15334, 42062) from VILLUM FONDEN. The authors would like to thank Prof. Stein Markvorsen for valuable early discussions.

Notations

$\mathcal{Z}, \mathcal{X}$	Smooth differentiable latent ($\mathcal{Z}$) and data ($\mathcal{X}$) manifold,
f	A stochastic immersion $f : \mathcal{Z} \subset \mathbb{R}^q \to \mathcal{X} \subset \mathbb{R}^D$,
J	Jacobian of the stochastic function f,
G	Stochastic metric tensor defined as the pullback metric through f: $G = J^\top J$,
$\mathcal{T}_z\mathcal{Z}$	Tangent space of the manifold $\mathcal{Z}$ at a point z,
Σ	IF f is a Gaussian process, then $J \sim \prod_{i=1}^{D} \mathcal{N}(\mu_i, \Sigma)$,
$\lVert\cdot\rVert_G$	Stochastic induced norm: $\lVert v\rVert_G \coloneqq \sqrt{v^\top G v}$,
$\lVert\cdot\rVert_R$	Riemannian induced norm: $\lVert v\rVert_R \coloneqq \sqrt{v^\top \mathbb{E}[G] v} \coloneqq \sqrt{g(v,v)}$,
$\lVert\cdot\rVert_F$	Finsler norm: $\lVert v\rVert_F \coloneqq \mathbb{E}[\sqrt{v^\top G v}] = F(v)$,
L_R, E_R, V_R	Length, energy and volume obtained from the Riemannian induced norm $\lVert\cdot\rVert_R$,
L_F, E_F, V_F	Length, energy and Busemann Hausdorff volume obtained from the Finsler norm $\lVert\cdot\rVert_F$.

Pulling back information geometry 7

The paper *Pulling back information geometry* has been written with the co-first authors Georgios Arvanatidis, Miguel González-Duque, Dimitris Kalatzis, and Søren Hauberg. The paper has been published for the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022). The paper is available at https://arxiv.org/abs/2106.05367.

Abstract

Latent space geometry has shown itself to provide a rich and rigorous framework for interacting with the latent variables of deep generative models. The existing theory, however, relies on the decoder being a Gaussian distribution as its simple reparameterization allows us to interpret the generating process as a random projection of a deterministic manifold. Consequently, this approach breaks down when applied to decoders that are not as easily reparametrized. We here propose to use the Fisher-Rao metric associated with the space of decoder distributions as a reference metric, which we pull back to the latent space. We show that we can achieve meaningful latent geometries for a wide range of decoder distributions for which the previous theory was not applicable, opening the door to 'black box' latent geometries.

7.1 Introduction

Generative models such as *variational autoencoders (VAEs)* (Kingma and Welling 2013; Rezende et al. 2014) and *generative adversarial networks (GANs)* (Goodfellow et al. 2014) provide state-of-the-art density estimators for high dimensional data. The underlying assumption is that data $\mathbf{x} \in \mathcal{X}$ lie near a low-dimensional manifold $\mathcal{M} \subset \mathcal{X}$, which is parametrized through a low-dimensional *latent representation* $\mathbf{z} \in \mathcal{Z}$. As data is finite and noisy, we only recover a probabilistic estimate of the true manifold, which, in VAEs, is represented through a decoder distribution $p(\mathbf{x}|\mathbf{z})$. Our target is the geometry of this *random manifold*.

The geometry of the manifold has been shown to carry great value when systematically interacting with the latent representations, as it provides a stringent solution to the *identifiability problem* that plagues latent variable models (Tosi et al. 2014; Arvanitidis et al. 2018; Hauberg 2018a). For example, this geometry has allowed VAEs to discover latent evolutionary signals in proteins (Detlefsen et al. 2020), provide efficient robot controls (Scannell et al. 2021b; Chen et al. 2018a; Beik-Mohammadi et al. 2021), improve latent clustering abilities (Yang et al. 2018; Arvanitidis et al. 2018) and more. The fundamental issue with these geometric approaches is that the studied manifold is inherently a stochastic object, but classic differential geometry only supports the study of *deterministic* manifolds. To bridge the gap, Eklund and Hauberg (2019) have shown how VAEs with a Gaussian decoder family can be viewed as a random projection of a deterministic manifold, thereby making the classic theories applicable to the random manifold.

A key strength of VAEs is that they can model data from diverse modalities through the choice of decoder distribution $p(\mathbf{x}|\mathbf{z})$. For discrete data, we use categorical decoders, while for continuous data we may opt for a Gaussian, a Gamma or whichever distribution best suits the data. However, for non-Gaussian decoders, there exists no useful approach

for treating the associated random manifold as deterministic, which prevents us from systematically interacting with the latent representations without being subjected to identifiability issues. This limitation motivates the current work.

In this paper, we provide a general framework that allows us to interact with the geometry of almost any random manifold. The key and simple idea is to reinterpret the decoder as spanning a deterministic manifold in the space of probability distributions $\mathcal{H}$, rather than a random manifold in the observation space (see Fig. 7.1). Calling on classical *information geometry* (Amari and Nagaoka 2000; Nielsen 2020), we show that the learned manifold is a Riemannian manifold of $\mathcal{H}$, and provide the corresponding computational tools. The approach is applicable to any family of decoders for which the KL-divergence can be differentiated, allowing us to work with a wide range of models from a single codebase.

7.2 The geometry of generative models

As a starting point, consider the deterministic generative model given by a prior $p(\mathbf{z})$ and a decoder $f : \mathcal{Z} = \mathrm{R}^d \to \mathcal{X} = \mathrm{R}^D$, which is assumed to be a smooth immersion. The latent representation $\mathbf{z}$ of an observation $\mathbf{x}$ is generally not *identifiable*, meaning that one can recover different latent representations that give rise to equally good density estimates. For example, let $g : \mathcal{Z} \to \mathcal{Z}$ be a smooth invertible function such that $\mathbf{z} \sim p(\mathbf{z}) \Leftrightarrow g(\mathbf{z}) \sim p(\mathbf{z})$, then the latent representation $g(\mathbf{z})$ coupled with the decoder $f \circ g^{-1}$ gives the same density estimate as $\mathbf{z}$ coupled with f (Hauberg 2018a). Practically speaking, the identifiability issue implies that it is improper to view the latent space $\mathcal{Z}$ as being Euclidean, as any reasonable view of $\mathcal{Z}$ should be invariant to reparameterizations g.

The classic geometric solution to the identifiability problem is to define any quantity of interest in the observation space $\mathcal{X}$ rather than the latent space $\mathcal{Z}$. For example, the length of a curve $\gamma : [0, 1] \to \mathcal{Z}$ in the latent space can be defined as its length measured in $\mathcal{X}$ on the manifold $\mathcal{M} = f(\mathcal{Z})$ with $N \to +\infty$ as:

$$\begin{aligned} \mathrm{L}(\gamma) &= \sum_{n=1}^{N-1} \|f(\gamma(t_{n+1})) - f(\gamma(t_n))\| = \int_0^1 \left\|\dot{f}(\gamma(t))\right\| dt \\ &= \int_0^1 \sqrt{\dot{\gamma}(t)^\top \mathbf{J}_f(\gamma(t))^\top \mathbf{J}_f(\gamma(t))\dot{\gamma}(t)} dt, \end{aligned} \tag{7.1}$$

where $t_n = {}^n\!/_N$ and $t_{n+1} = {}^{n+1}\!/_N$, and we used the chain rule $\partial_t f(\gamma(t)) = \mathbf{J}_f(\gamma(t))\dot{\gamma}(t)$ with $\dot{\gamma}(t) = \partial_t \gamma(t)$ being the curve derivative, and $\mathbf{J}_f(\gamma(t)) \in \mathrm{R}^{D \times d}$ the Jacobian of f at $\gamma(t)$. This construction shows how we may calculate lengths in the latent space with respect to the metric of the observation space, which is typically assumed to be the Euclidean, but other options exist (Arvanitidis et al. 2021). In this way, the symmetric positive definite matrix $\mathbf{J}_f(\gamma(t))^\top \mathbf{J}_f(\gamma(t))$ is denoted by $\mathbf{M}(\gamma(t)) \in \mathrm{R}^{d \times d}_{\succ 0}$ and captures the geometry of $\mathcal{M}$ in $\mathcal{Z}$. This is known as the *pullback metric* as it pulls the Euclidean metric from $\mathcal{X}$ into $\mathcal{Z}$. As the Jacobian spans the d-dimensional tangent space at the point $\mathbf{x} = f(\mathbf{z})$, we may interpret $\mathbf{M}(\mathbf{z})$ as an inner product $\langle \vec{u}, \vec{v} \rangle_{\mathbf{M}} = \vec{u}^\top \mathbf{M}(\mathbf{z}) \vec{v}$ over this tangent space, given us all the ingredients to define *Riemannian manifolds*:

Definition 7.2.1
A Riemannian manifold is a smooth manifold $\mathcal{M}$ together with a Riemannian metric $\mathbf{M}(\mathbf{z})$, which is a positive definite matrix that changes smoothly throughout space and defines an inner product on the tangent space $\mathcal{T}_{\mathbf{z}}\mathcal{M}$.

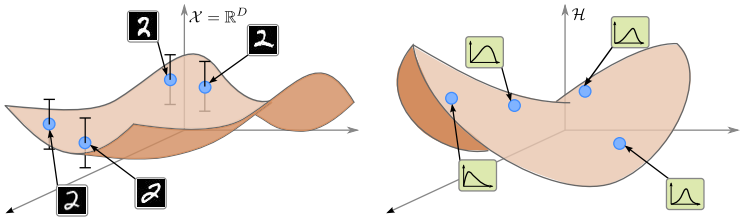


Figure 7.1: Traditionally (left), we view the learned manifold as a stochastic manifold in the observation space. We propose (right) to view the learned manifold as a deterministic manifold embedded in the space of decoder distributions, which is equipped with a Fisher-Rao metric based on information geometry.

We see that the decoder naturally spans a Riemannian manifold and the latent space $\mathcal{Z}$ can be considered as the *intrinsic coordinates*. Technically, we can consider any Euclidean space as the intrinsic coordinates of an abstract $\mathcal{M}$ using a suitable metric $\mathbf{M}(\mathbf{z})$, which is implicitly induced by an abstract f. Since the Riemannian length of a latent curve (7.1), by construction, is invariant to reparameterizations, it is natural to extend this view with a notion of *distance*. We say that the distance between two points $\mathbf{z}_0, \mathbf{z}_1 \in \mathcal{Z}$ is simply the length of the shortest connecting path, $\text{dist}(\mathbf{z}_0, \mathbf{z}_1) = \min_\gamma \text{L}(\gamma)$. Calculating distances implies finding the shortest path. One can show (Gallot et al. 2004) that length minimizing curves also have minimal *energy*:

$$\text{E}(\gamma) = \int_0^1 \|\dot{f}(\gamma(t))\|^2 dt = \int_0^1 \dot{\gamma}(t)^\top \mathbf{M}(\gamma(t))\dot{\gamma}(t) dt, \tag{7.2}$$

which is a locally convex functional. Shortest paths can then be found by direct energy minimization (Yang et al. 2018) or by solving the associated system of ordinary differential equations (ODEs) (Hennig and Hauberg 2014; Arvanitidis et al. 2019) (see supplementary materials for additional details).

7.2.1 Stochastic decoders

As previously discussed, deterministic decoders directly induce a Riemannian geometry in the latent space. However, most models of interest are stochastic and there is significant evidence that this stochasticity is important to faithfully capture the intrinsic structure of data (Hauberg 2018a). When the decoder is a smooth stochastic process, e.g. as in the Gaussian Process Latent Variable Model (GP-LVM) (Lawrence 2003), Tosi et al. (2014) laid the foundations for modeling a stochastic geometry. Most contemporary models, such as VAEs, assume independent noise, making this theory inapplicable. Arvanitidis et al. (2018) proposed an extension of this stochastic geometry to VAEs with Gaussian decoders, which take the form

$$\begin{aligned} f(\mathbf{z}) &= \mu(\mathbf{z}) + \sigma(\mathbf{z}) \odot \vec{\epsilon} \\ &= \begin{bmatrix} \mathbf{I}_D & \text{diag}(\vec{\epsilon}) \end{bmatrix} \begin{bmatrix} \mu(\mathbf{z}) \\ \sigma(\mathbf{z}) \end{bmatrix} = \mathbf{P}_\varepsilon \; h(\mathbf{z}), \end{aligned} \tag{7.3}$$

where $\vec{\epsilon} \sim \mathcal{N}(\vec{0}, \mathbf{I}_D)$. Here we have written the Gaussian decoder in its reparametrized form. This can be viewed as a random projection of a deterministic manifold spanned by h with projection matrix $\underline{\text{P}}_\epsilon$ (Eklund and Hauberg 2019), which can easily be given a geometry. The associated Riemannian metric,

$$\mathbf{M}(\mathbf{z}) = \mathbf{J}_\mu(\mathbf{z})^\top \mathbf{J}_\mu(\mathbf{z}) + \mathbf{J}_\sigma(\mathbf{z})^\top \mathbf{J}_\sigma(\mathbf{z}), \tag{7.4}$$

gives the shortest paths that follow the data as distances grow with the model uncertainty (Arvanitidis et al. 2018; Hauberg 2018a). An example of the shortest path $\gamma(t) \in \mathcal{Z}$

computed under this metric is shown in Fig. 7.2 and the respective curve on the corresponding expected manifold $\mu(\gamma(t)) \in \mathcal{M} \subset \mathcal{X}$.

Previous work has, thus, focused on *pulling back* the Euclidean metric from the observation space to the latent space using the reparameterization of the Gaussian decoder. This is, however, intrinsically linked with the simple reparameterization of the Gaussian, and this strategy can only extend to location-scale distributions. We propose an alternative, principled way of dealing with stochasticity by changing the focus from the observation space $\mathcal{X}$ to the parameter space $\mathcal{H}$ associated to the distribution of the decoder, leveraging the metrics defined in classical information geometry.

7.3 Information geometric latent metric

So far we have seen how we can endow the latent space $\mathcal{Z}$ with meaningful distances only when our stochastic decoders are reparameterizable and their codomain is the observation space $\mathcal{X}$. Ideally, we would like a more general framework of computing the shortest path distances for a more general class of distributions.

We first note that the codomain of a VAE decoder is the parameter space $\mathcal{H}$ of a probability density function. In particular, depending on the type of data we specify a likelihood $p(\mathbf{x}|\eta)$ with parameters $\eta \in \mathcal{H}$, which we can rewrite as $p(\mathbf{x}|\mathbf{z})$ using the mapping $h : \mathcal{Z} \to \mathcal{H}$.

With this in mind, we can ask what is a natural distance in the latent space $\mathcal{Z}$ between two infinitesimally near points $\mathbf{z}_1$ and $\mathbf{z}_2 = \mathbf{z}_1 + \epsilon$ when measured in $\mathcal{H}$. Since our latent codes map to distributions we can define the (infinitesimal) distance through the KL-divergence:

$$\text{dist}^2(\mathbf{z}_1, \mathbf{z}_2) = \text{KL}(()p(\mathbf{x}|\mathbf{z}_1),\ p(\mathbf{x}|\mathbf{z}_2)). \tag{7.5}$$

So we can define the length of a curve $\gamma : [0, 1] \to \mathcal{Z}$ as

$$\text{L}(\gamma) = \lim_{N\to\infty} \sum_{n=1}^{N-1} \text{KL}(()p(\mathbf{x}|\gamma(t_n)), p(\mathbf{x}|\gamma(t_{n+1})))^{\frac{1}{2}}, \tag{7.6}$$

and distances could be defined as before. This would satisfy our desiderata of a deterministic notion of similarity in the latent space that is applicable to wide range of decoder distributions.

This construction may seem arbitrary, but in reality it carries deeper geometric meaning. *Information geometry* (Nielsen 2020) considers families of probabilistic densities $p(\mathbf{x}|\eta)$ as represented by their parameters $\eta \in \mathcal{H}$, such that $\mathcal{H}$ is constructed as a statistical manifold equipped with the Fisher-Rao metric, which infinitesimally coincides with the KL divergence in (7.5). This is known to be a Riemannian metric over $\mathcal{H}$ that takes the following form:

$$\mathbf{I}_{\mathcal{H}}(\eta) = \int_{\mathcal{X}} [\nabla_\eta \log p(\mathbf{x}|\eta) \nabla_\eta \log p(\mathbf{x}|\eta)^\top] p(\mathbf{x}|\eta) d\mathbf{x}. \tag{7.7}$$

When the parameter space $\mathcal{H}$ is equipped with this metric, we call it a *statistical manifold*.

Definition 7.3.1
A statistical manifold consists of the parameter space $\mathcal{H}$ of a probability density function $p(\mathbf{x}|\eta)$ equipped with the Fisher-Rao information matrix $\mathbf{I}_{\mathcal{H}}(\eta)$ as a Riemannian metric.

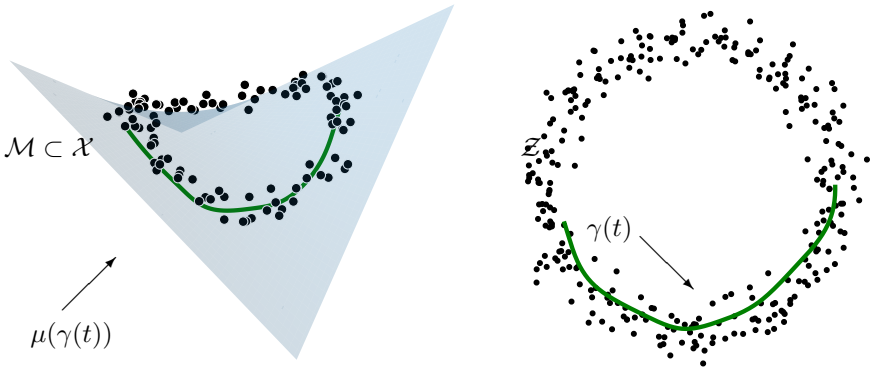


Figure 7.2: A conceptual example of a Riemannian manifold $\mathcal{M} = \mu(\mathcal{Z})$ lying in $\mathcal{X}$ and the corresponding latent space $\mathcal{Z}$, together with an associated the shortest paths.

Note that the geometry induced by the Fisher-Rao metric is predefined and can be seen as a modeling decision, since it is related to the chosen likelihood and does not change with data.

As previously mentioned, a known result in Information Geometry is that the Fisher-Rao metric coincides with the KL-divergence locally (Nielsen 2020; Amari and Nagaoka 2000):

Proposition 7.3.1
The Fisher-Rao metric is the second order approximation of the KL-divergence between perturbed distributions:

$$\mathrm{KL}(()p(\mathbf{x}|\eta), p(\mathbf{x}|\eta + \delta\eta)) = \frac{1}{2}\delta\eta^\top \mathbf{I}_\mathcal{H}(\eta)\delta\eta + o(\delta\eta^2). \tag{7.8}$$

The central idea put forward in this paper is to consider the decoder as a map $h : \mathcal{Z} \to \mathcal{H}$ instead of $f : \mathcal{Z} \to \mathcal{X}$, and let $\mathcal{H}$ be equipped with the appropriate Fisher-Rao metric. The VAE can then be interpreted as spanning a manifold $h(\mathcal{Z})$ in $\mathcal{H}$ and the latent space $\mathcal{Z}$ can be endowed with the corresponding metric. We detail this approach in the sequel.

7.3.1 The Riemannian pull-back metric

Our construction implies that the length of a latent curve $\gamma : [0, 1] \to \mathcal{Z}$ when mapped through h can be measured in the parameter space $\mathcal{H}$ using the Fisher-Rao metric therein as

$$\mathrm{L}(\gamma) = \int_0^1 \sqrt{\partial_t h(\gamma(t))^\top \mathbf{I}_\mathcal{H}(h(\gamma(t)))\partial_t h(\gamma(t))}dt, \tag{7.9}$$

with $\mathbf{M}$ the pullback metric:

Proposition 7.3.2
Let $h : \mathcal{Z} \to \mathcal{H}$ be an immersion that parametrizes the likelihood. Then, the latent space $\mathcal{Z}$ is equipped with the Riemannian pull-back metric $\mathbf{M}(\mathbf{z}) = \mathbf{J}_h^\top(\mathbf{z})\mathbf{I}_\mathcal{H}(h(\mathbf{z}))\mathbf{J}_h(\mathbf{z})$.

Proof. See appendix, Prop. C.3.1. □

Note that instead of considering the parameters $\eta \in \mathcal{H}$ of the probabilistic density function $p(\mathbf{x}|\eta)$ that approximates the data, we can consider the latent variable $\mathbf{z}$ as

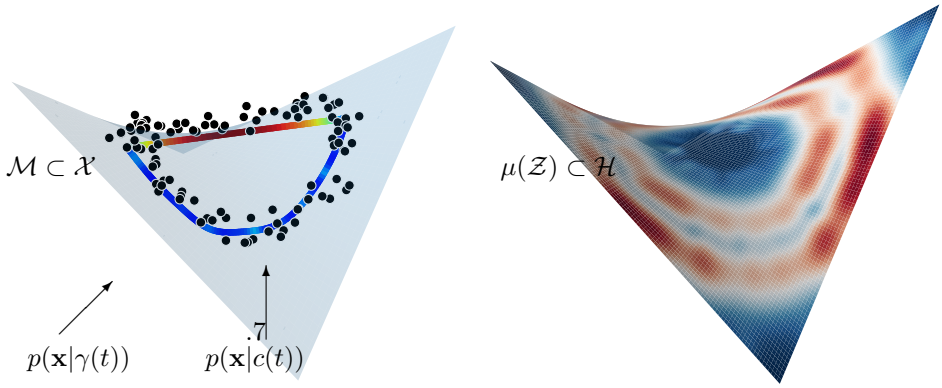


Figure 7.3: *Left*: The optimal $\gamma(t)$ under $\mathbf{M}(\mathbf{z})$ results to distributions that respect the structure of data, while the curve $c(t)$ with minimal length in $\mathcal{H}$ does not as it leaves $\mathcal{M}$. Red and green signal high and low variance respectively. *Right*: A part of the spanned manifold $h(\mathcal{Z}) = [\mu(\mathcal{Z}), \sigma(\mathcal{Z})] \in \mathcal{H}$ colored by $|\mathbf{M}(\mathbf{z})|$. Note that we design $\sigma(\mathbf{z})$ to increase far from data, which ensures that $\gamma(t)$ stays within their support.

the actual parameters of the model. This view is equivalent to the one explained above, and the corresponding pull-back metric is directly the Fisher-Rao metric endowed in the latent space $\mathcal{Z}$:

Proposition 7.3.3
The pullback metric $\mathbf{M}(\mathbf{z})$ is identical to the Fisher-Rao metric obtained over the parameter space $\mathcal{Z}$ as $\mathbf{M}(\mathbf{z}) = \int_{\mathcal{X}} \left[\nabla_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z}) \nabla_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z})^{\top}\right] p(\mathbf{x}|\mathbf{z}) d\mathbf{x}$.

Proof. See appendix, Prop. C.3.2. □

Therefore, pulling back the Fisher-Rao metric from $\mathcal{H}$ into $\mathcal{Z}$ enables us to compute length minimizing curves which are identifiable (see Sec. 7.2). The advantage of this approach is that it applies to any type of decoders and data, as the actual distance is measured over the manifold spanned by h in the parameter space $\mathcal{H}$. So the shortest paths between probability distributions move optimally on this manifold while taking the geometry of $\mathcal{H}$ into account through the Fisher-Rao metric.

Computing the shortest paths directly in $\mathcal{H}$ need not result in a sensible sequence of probability density functions $p(\mathbf{x}|\eta)$. To ensure that the shortest paths computed under our metric stay within the support of the data, we carefully design our decoder h to extrapolate to uncertain distributions outside the support of the data (see supplements for additional details).

In Fig. 7.3 we compare the shortest path $\gamma : [0, 1] \to \mathcal{Z}$ under the proposed metric $\mathbf{M}(\mathbf{z})$ against a curve $c : [0, 1] \to \mathcal{H}$ with minimal length. We consider a Gaussian likelihood with isotropic covariance. We show the resulting sequence of means for both interpolants color-coded by the corresponding variances. As expected $c(t)$ does not take into account the given data, but only respects the geometry of $\mathcal{H}$ implied by the likelihood.

7.3.2 Efficient shortest path computation

An essential task in computational geometry is to compute the shortest paths. This can be achieved by minimizing curve energy (7.2) or solving the corresponding system of ODEs (see supplementary material). The latter, however, requires inordinate computational resources, since the evaluation of the system relies on the Jacobian of the decoder and its derivatives.

Bearing in mind that the metric is an approximation of the KL divergence between perturbations (7.8), the energy is directly expressed as a sum of KL divergence terms along a discretized curve γ:

$$\mathrm{E}(\gamma) \propto \lim_{N\to\infty} \sum_{n=1}^{N-1} \mathrm{KL}(p(\mathbf{x}|\gamma(t_n)), p(\mathbf{x}|\gamma(t_{n+1}))). \tag{7.10}$$

The proof can be found in the appendix, Prop. C.1.2. A simple algorithm for computing the shortest paths is to minimize (7.10) with respect to the parameters of the curve γ. Here we represent γ as a cubic spline with fixed end-points. Then standard free-form optimization can be applied to minimize this energy.

7.3.3 Example: categorical decoders

The motivation for our approach is that, while several options for decoders exist in VAEs depending on the type of the given data, we could only capture and use the learned geometry in a principled way with Gaussian decoders. Our proposed methodology is more general.

For a constructive example, assume that $\mathbf{x}$ is a categorical variable. We can select a generalized Bernoulli likelihood $p(\mathbf{x}|\mathbf{z})$, such that $h(\mathbf{z}) = (\eta_1, \cdots, \eta_D)$ where each η_i represents the probability of x_i being 1. Thus, the parameters η lie on the unit simplex $\mathcal{H}$, and the distance under the corresponding Fisher-Rao metric between points on the simplex coincides with the spherical distance between the points $\sqrt{\eta}$ on the unit sphere,

$$\mathrm{dist}(\eta, \eta') = \arccos\left(\sqrt{\eta}^\top \sqrt{\eta'}\right). \tag{7.11}$$

We derive in detail this previously known result in the supplementary materials.

Given a curve $\gamma : [0, 1] \to \mathcal{Z}$ we can approximate the energy by using the small angle approximation $\cos\theta \approx 1 - {}^{\theta^2}/_2 \Leftrightarrow \theta^2 \approx 2 - 2\cos\theta$ to give

$$\mathrm{E}(\gamma) = \sum_{n=1}^{N-1} \left(2 - 2\sqrt{h(\gamma(t_n))}^\top \sqrt{h(\gamma(t_{n+1}))}\right), \tag{7.12}$$

for sufficiently fine discretization with $t_n = {}^n/_N$ and $t_{n+1} = {}^{n+1}/_N$. This gives a particular simple expression for the energy, which we can minimize in order to compute the shortest path.

7.3.4 Black-box random geometry

In general, we can derive suitable expressions for computing metrics and energies for families of decoders, doing so is tedious, error-prone and time-consuming. This limits the practical use of the developed theory.

Drawing inspiration from *black-box variational inference* (Ranganath et al. 2014), we propose a notion of *black-box random geometry*. Assume that we have access to a differentiable KL divergence for our choice of decoder distribution. We can then apply the methodology presented in Sec. 7.3.2 to compute the shortest paths.

In practice, modern libraries such as PyTorch (Paszke et al. 2019) have this functionality implemented for several distributions. When we do not have closed-form expression for the KL divergence, we can resort to Monte Carlo estimates thereof. More specifically, we can estimate the KL divergence by generating samples from the likelihood based

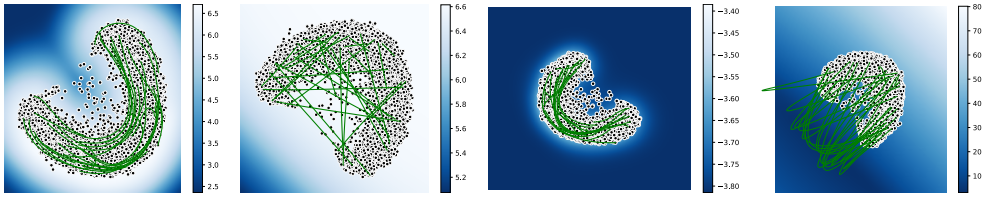


Figure 7.4: Pulling back the Euclidean and Fisher-Rao metrics with Gaussian decoders. Left to right: Euclidean pull-back with regularized uncertainty, Euclidean pull-back with a NN to model uncertainty, Fisher-Rao pull-back with regularized uncertainty, Fisher-Rao pull-back with a NN to model uncertainty.

on the re-parametrization trick, which allows us to get derivatives with automatic differentiation.

Interestingly, apart from finding the shortest path through the KL formulation, we can also approximate the actual metric tensor $\mathbf{M}(\mathbf{z})$. As we have discussed above, evaluating explicitly this metric is not a trivial task in many cases. One problem is that we need access to the Jacobian of the parametrization h, which is typically a deep neural network, so the computation is not always straightforward. Alternatively, one could use that the Fisher-Rao metric is the Hessian of the KL-divergence (7.8), but such approaches fare poorly with current tools for automatic differentiation, where higher-order derivatives are often incompatible with batching. Furthermore, the Fisher-Rao metric itself may be intractable depending on the chosen likelihood $p(\mathbf{x}|\eta)$. Nevertheless, we show that the KL formulation (7.8) allows us to approximate the latent metric as:

Proposition 7.3.4
We define perturbations vectors as $\delta e_i = \varepsilon \cdot \mathbf{e}_i$, with $\varepsilon \in \mathbb{R}_+$ a small infinitesimal quantity, and $\mathbf{e}_i$ a canonical basis vector in $\mathbb{R}^d$. For better clarity, we rename $\mathrm{KL}(p(\mathbf{x}|\mathbf{z}), p(\mathbf{x}|\mathbf{z}+\delta\mathbf{z})) = \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{z})$, and we note $\mathbf{M}_{ij} = \mathbf{M}_{ji}$ the components of $\mathbf{M}(\mathbf{z})$. We can then approximate by a system of equations the diagonal and non-diagonal elements of the metric:

$$\begin{aligned}\mathbf{M}_{ii} &\approx 2\ \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i)/\varepsilon^2 \\ \mathbf{M}_{ji} &\approx \left(\mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i+\delta\mathbf{e}_j) - \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i) - \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_j)\right)/\varepsilon^2.\end{aligned}$$

See Prop. C.3.4 in the appendix for a proof. Note that this formulation only requires h to be a smooth immersion. This is particularly useful, as the metric is used for other purposes on a Riemannian manifold and not exclusively for computing the shortest paths. For example, relying on $\mathbf{M}(\mathbf{z})$ we can compute the exponential map by solving the corresponding ODE system as an initial value problem. Assuming a fully differentiable KL divergence, then the approximated metric is also differentiable. This is all that is required for practical usage of differential geometry, and thus, we have a reasonable notion of *black-box random geometry*.

7.4 Experiments

7.4.1 Pulling back Euclidean and Fisher-Rao metric with Gaussian decoders

We start our experiments by comparing our proposed way of inducing geometry in latent spaces with the existing theory: pulling back the Euclidean metric using a stochastic Gaussian decoder (see (7.4)). We also include in this comparison the effect of regularizing the uncertainty quantification in the learned geometries. In this regularization, we use

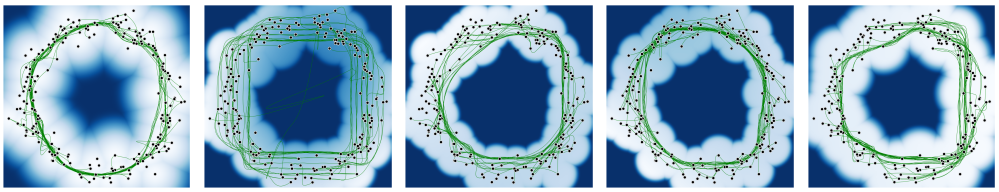


Figure 7.5: Pulling back the metric from different parameter spaces. From left to right: Normal, Bernoulli, Beta, Dirichlet and Exponential. White areas represent low entropy of the decoded distribution, while blue areas represent higher entropy. Notice that the Bernoulli latent space is darker blue (i.e. more entropic) because distributions with parameters around $1/2$ are near uniform.

transition networks (Detlefsen et al. 2019) to ensure high uncertainty outside the support of the data (see Sec. C.3.2 in the supplementary material).

In this experiment, we train four VAEs on a subset of the MNIST dataset composed of only the digits with label 1. Two of these VAEs implement a standard Gaussian decoder, and we induce a metric in the latent space by pulling the Euclidean metric back using the Jacobian of the decoder. In the other two, we consider the output of the decoder as lying in a statistical manifold and approximate the pullback of the Fisher-Rao metric by using the KL divergence locally. In each of these two sets, one of the decoders implements the uncertainty regularization described above.

Fig. 7.4 shows the latent spaces of these four decoders, illuminated by the volume measure. In each of this latent spaces, we analyze the geometry induced by the respective pullbacks by computing and plotting several shortest paths. This figure illustrates two key findings: (1) Our approach is on par with the existing literature in learning geometric structure, which can be seen by comparing the first and third latent spaces (Euclidean vs. Fisher Rao, respectively), and (2) Performing uncertainty regularization plays an instrumental role on learning a sensible geometric structure, which can be seen when comparing the first and second latent spaces (both coming from the Euclidean pullback, with and without regularization respectively), and similarly for the third and fourth.

7.4.2 The Fisher-Rao pullback metric for various distributions with toy data

For our second experiment, we induced a geometry on a known latent space (given by noisy circular data in $\mathcal{Z}_{\text{toy}} = \mathbb{R}^2$) by *pulling back* the Fisher-Rao metric from the parameter space of different distributions, showcasing the potential for computing the shortest paths efficiently, even in non-Gaussian settings. The statistical manifolds from which we pull the metric are associated with multivariate versions of the Normal, Bernoulli, Beta, Dirichlet and Exponential distributions. For this approximation to follow the support of the data we need to ensure that our mapping $\mathcal{Z}_{\text{toy}} \to \mathcal{H}$ extrapolates to high uncertainty outside our training codes (see Fig. 7.4). To do so, we perform uncertainty regularization for each one of the decoded distributions (see supplementary materials for implementation details).

In Fig. 7.5 we show the toy latent space alongside several shortest paths computed using the pullback of the Fisher-Rao metric from the statistical manifolds associated with the Gaussian, Bernoulli, Beta, Dirichlet and Exponential distributions. We parametrize the curves as cubic splines and minimize their energy using automatic differentiation (see Sec. 7.3.2). These results show that the approximated pulled-back metric induces a meaningful geometry in this latent space, which recovers the true circular structure of the data. In the case of the Bernoulli distribution, we notice that some paths fail to converge. We hypothesize that our uncertainty regularization (which decodes to the uniform

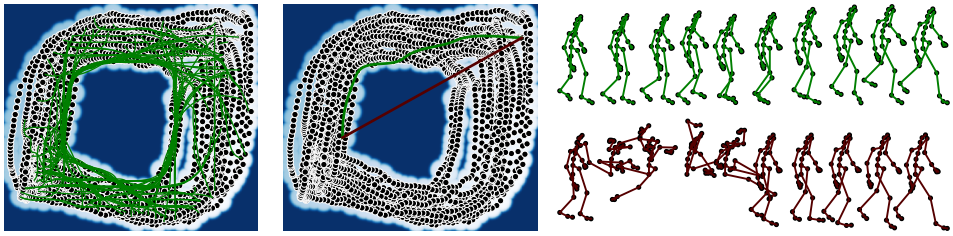


Figure 7.6: *Left:* Geodesics in the latent space of a von Mises-Fisher decoder. *Middle*: Shortest path (green) vs. linear (red). *Right*: decoding the shortest path (green) vs. the linear interpolation (red) as poses (i.e. the product of von Mises-Fisher distributions). Our path follows the trend of the data manifold, while the linear path traverses regions with no data support.

distribution outside the support) is not strong enough since Bernoulli distributions with parameters close to 1/2 are already highly entropic.

7.4.3 Motion capture data with products of von Mises-Fisher distributions

As a further demonstration of our black-box random geometry, we consider a model of human motion capture data. Here we observe a time series, where each time point represent a 'skeleton' corresponding to a human pose. As only pose, and not shape, changes over time, individual limbs on the body only change position and orientation, but not length. Each limb is then a point on a sphere in $\mathbb{R}^3$ with radius given by the limb length. Following Tournier et al. (2009) we view the skeleton representation space as a product of spheres. From this, we build a VAE where the decoder distribution is a product of von Mises-Fisher distributions. To ensure a sensible uncertainty estimates in the decoder, we enforce that the concentration parameter extrapolate to a small constant.

In this case, we do not have easily accessible Fisher-Rao metrics, so we lean on the KL formulation from Sec. 7.3.4. Since, the KL does not have a closed-form expression for the von Mises-Fisher distribution, we resort to a Monte Carlo estimate thereof. This is realizable with off-the-shelf tools (Davidson et al. 2018).

Fig. 7.6 shows the latent representation of a motion capture sequence of a person walking (Seq. `69_06` from `http://mocap.cs.cmu.edu/`) with the shortest paths superimposed. We see that our paths follow the trend of the data, and reflect the underlying periodic nature of the observed walking motion. We pick two random points in the latent space, and traverse both the shortest path and the straight line implied by a Euclidean interpretation of the latent space. As we traverse, we sample from the decoder distribution, thereby producing two new motion sequences, which appear in Fig. 7.6. As can be seen, the straight line traverses uncharted territory of the latent space and end up creating an implausible motion. This is in contrast to the shortest path, that consistently generates meaningful poses.

7.4.4 Numerical approximation of the Fisher-Rao pullback metric

Prop. 7.3.4 provide an approximation to the metric, and we test its accuracy as per (7.8). We discretize the latent space for the just-described von Mises-Fisher decoder and, for each $\mathbf{z}$ in this grid, we both approximate $\mathbf{M}(\mathbf{z})$ and compute the expected value of $\|\mathrm{KL}(p(\mathbf{x}|\mathbf{z}), p(\mathbf{x}|\mathbf{z}+\delta\mathbf{z})) - \frac{1}{2}\delta\mathbf{z}^\top\mathbf{M}(\mathbf{z})\delta\mathbf{z}\|$ for several samples of $\delta\mathbf{z}$, uniformly distributed

around the circle of radius $\varepsilon = 0.1$. Notice that we do not have a ground truth to compare against, and that this error will always be off by $o(\delta \mathbf{z}^2)$. Fig. 7.7 shows the average error, where we can see that the approximate metric is well-estimated both within and outside the support of the data. The error, however, grows at the boundaries of the support, where the distribution is changing from a concentrated von Mises-Fisher to a uniform distribution. It is worth mentioning that we observe some approximated metrics have negative determinant, showing that our numerical approximations are imprecise at the boundary. These results warrant further research on more stable ways of approximating pulled back metrics under our proposed approach.

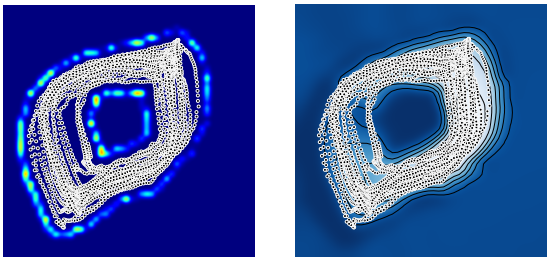


Figure 7.7: *Left*: Average error of the approximated metric in the von Mises-Fisher latent space. Darker colors indicate lower error (less than ε^2), while higher values are clear. *Right*: The LAND density well-adapts to the nonlinear structure of the latent representations due to the shortest path's behavior.

7.4.5 Statistical models on manifolds

We demonstrate the usefulness of the approximated metrics, by fitting a distribution to data in the latent space, which requires normalization according to the measure induced by the metric. In particular, we fit a locally adaptive normal distribution (LAND) (Arvanitidis et al. 2016), which extends the Gaussian distribution to learned manifolds. The probability density function is $\rho(\mathbf{z}) = C(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \exp\left(-0.5 \cdot \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{z})^\top \boldsymbol{\Gamma} \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{z})\right)$, where $\boldsymbol{\mu} \in \mathrm{R}^d$ is the mean, $\boldsymbol{\Gamma} \in \mathrm{R}^{d \times d}_{\succ 0}$ is the precision matrix and $C(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ the normalization constant. The operator $\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{z})$ returns the scaled initial velocity $\mathbf{v} = \dot{\gamma}(0) \in \mathrm{R}^d$ of the shortest connecting path with $\gamma(1) = \mathbf{z}$ and $||\mathbf{v}|| = \mathrm{Length}(\gamma)$. In Fig. 7.7 we show the LAND density on the learned latent representations under the approximated Riemannian metric from Sec. 7.4.4. Since shortest paths follow the data, so does the density ρ. See supplementary material for details.

7.4.6 Movie preferences via latent interpolants

In addition, we explored the latent space of the movie-users rating dataset MovieLens 25M (https://grouplens.org/datasets/movielens/25m/). In particular, we consider a Bernoulli VAE to model if a user has watched a movie among the 60 most popular in the dataset. Also, we considered only users who have seen less than 30 movies. The implementation and preprocessing details can be found in the supplementary material. Our VAE decodes to 60 Bernoulli parameters that are conditionally independent given the latent code $\mathbf{z}$, which state the likelihood that a given user has seen these movies. Latent codes in this space, then, can be seen as individual users with certain movie preferences.

We then computed the shortest path between two points by considering the pulled-back Fisher-Rao (see Sec. 7.3.2), and we compare against a straight line interpolation. We consider the cosine similarity of the decoded outputs. This cosine similarity measures whether two users (encoded as points in the latent space) have similar preferences

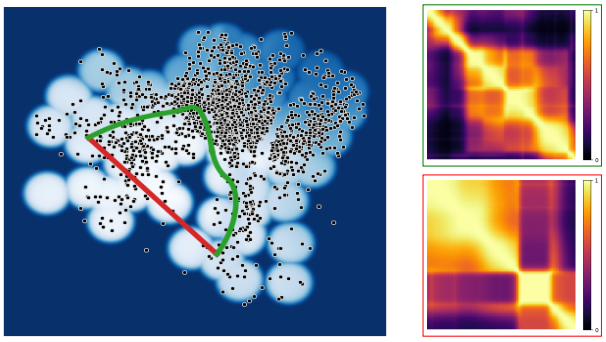


Figure 7.8: On the left, our path (green) follows users with similar preferences, as similarity is only locally high. Instead, the line (red) does not respect the learned structure resulting to users with no specific preferences. On the right, we show the cosine similarity of the decoded outputs, the upper figure for the linear path, and the bottom figure for our geodesic.

according to our model. In Fig. **??** we see that our path follows users with similar movie preferences locally, while the linear interpolation failed to capture a local notion of preference.

7.5 Related work

The literature is rich on deterministic generative models such as autoencoders (Rumelhart et al. 1986) and generative adversarial networks (Goodfellow et al. 2014),and a series of papers have investigated such deterministic decoders (Shao et al. 2018; Chen et al. 2018b; Laine 2018). However, our work is not applicable to this setting. As demonstrated in Sec. 7.4.1 stochasticity is essential to shape the latent space according to the data manifold. Hauberg (2018a) argues model uncertainty plays a role much akin to topology in classic geometry, in that it, practically, allows us to deviate from the Euclidean topology of the latent space.

Our constructions rely on information geometry and in particular Fisher-Rao metrics (Nielsen 2020). While our work is within the spirit of information geometry, it does not represent typical usage of this theory. Information geometry has been widely used in the context of optimization with *natural gradients* (Martens 2014; Martens and Grosse 2015), Markov Chain Monte Carlo methods (Girolami and Calderhead 2011) and hypothesis testing (Nielsen 2020). The key difference between natural gradients and our work is the space we wish to explore: in the case of the natural gradients, the shortest path is obtained on the space of the weights of the neural networks, while we aim to explore the latent space of a VAE. It can also be noted that Information geometry provides a rich family of alternative divergences over the here-applied KL-divergence. We did not investigate their usage in our context.

To make use of the here-developed tools, we may lean on techniques for statistics on manifolds. These provide generalizations of a long list of classic statistical algorithms (Zhang and Fletcher 2013; Hauberg 2015; Fletcher 2011). We refer the reader to Pennec (2006) for a gentle introduction to this line of research.

7.6 Conclusion and discussion

We have proposed a new approach for getting a well-defined and useful geometry in the latent space of generative models with stochastic decoders. The theory is easy to apply and readily generalize to a large family of decoder distributions. The latent geometry gives access to a series of operations on latent variables that are invariant to reparameterizations of the latent space, and therefore are not subject to a large class of identifiability issues. Such operational representations have already shown great value in applications ranging from biology (Detlefsen et al. 2020) to robotics (Scannell

et al. 2021b). We have here focused on the Fisher-Rao metric, but other geometries over distributions may apply equally well, e.g. the Wasserstein geometry may be interesting to explore.

Limitations. The largest practical hurdle with the proposed methodology, is that it only works well for decoders with well-calibrated uncertainties. That is, the decoder should yield high entropy in regions of little training data to ensure that shortest paths follow the trend of the data. This constraint is shared with existing approaches (Arvanitidis et al. 2018). Some heuristics exist (Detlefsen et al. 2019), but principled approaches are currently lacking.

Acknowledgements

This work was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). It also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research, innovation program (757360) and from a research grant (15334) from VILLUM FONDEN.

Part IV

Conclusion

Summary, discussions and future work

8

8.1 On the curvature of the loss landscape

In the paper **On the curvature of the loss landscape** (Chapter 5), we study the parameter space of the loss function as a Riemannian manifold. Since the notion of flatness is correlated with generalization in optimization, we are particularly interested in studying the curvature of the loss landscape. Our contributions are the following: (1) we derive a Riemannian metric on the parameter space of a model by considering the graph of the loss function, (2) we derive the Riemannian, Ricci and Scalar curvature of the loss landscape, (3) the scalar curvature at the minima is related to the Hessian of the loss function and is easily computed: $\mathrm{S} = \|\mathbf{H}\|_*^2 - \|\mathbf{H}\|_F^2$.

This work is motivated by the *flatness hypothesis*, which states that flat minima generalize better than sharp minima. This hypothesis is supported by empirical and theoretical evidence, but the notion of flatness is not well-defined. We know that it should be related to the Hessian of the loss function, but with an increasing number of parameters, studying the spectrum of the Hessian becomes computationally expensive. Instead, the research community looked at the norm of the Hessian as a measure of flatness. The main problem with using Frobenius norm that it can be null at a saddle point, which is not flat.

The scalar curvature, on the other hand, is a well-studied intrinsic Riemannian quantity that is derived from the Riemannian curvature, defined as the deviation of the curved space from the flat Euclidean space. In convex cases, a small scalar curvature implies a small Hessian's norm, and with a sufficiently high number of parameters, when the spectrum of the Hessian eigenvalues becomes flat (Zhang et al. 2021), and the scalar curvature reduces to the nuclear norm of the Hessian. Since many recent results that showed improved generalization and optimization using the Hessian's norm, they are also satisfied by the scalar curvature.

Yet, the scalar curvature is only a *scalar* measure of the curvature. While it might provide an intuitive enough description of the flatness, it will not fully describe the curvature of the loss landscape, as the Riemannian tensor or the Ricci tensor would. Furthermore, establishing the exact correlation between flatness and generalization remains an open problem in machine learning. This problem becomes even more challenging when we consider the nature of the optimizer, the number of data points and the way we split the data across training.

A promising direction to this work would be to study the curvature in the context of stochastic optimization. By considering one batch as a collection of subsets of the data, the Hessian would become a random matrix. The scalar curvature would transform into a random variable, whose distribution is related to batch distribution, and it would be worth investigating how it varies when the scalar curvature is null. Prior studies have also linked the Hessian norm with noise in gradient descent, raising questions about potential ties between scalar curvature and stochastic processes. Additionally, it is worth considering whether the scalar curvature plays a role in implicit model regularization.

8.2 Identifying latent distances with Finslerian geometry

In the paper **Identifying latent distances with Finslerian geometry** (Chapter 6), we compared the geodesics obtained from pulling back the expected metric with the expected geodesics. We showed that the expected length induced by a stochastic Riemannian metric is a Finsler metric. It has a closed-form expression, when the mapping is a Gaussian process. In high-dimensions, the expected Riemannian metric and the Finsler metric converge to each other.

This paper initially attempted to study the difference between the expectation of stochastic lengths and the lengths directly induced by the deterministic approximation of the stochastic metric. In some particular cases, we can think of the Riemannian lengths as the Finslerian lengths being regularized with their variance. In practice, we showed that the difference between the two lengths is negligible, which further justifies to approximate the pullback metric tensor by its expectation.

What is more interesting though is that this conclusion might stem from a broader, yet unproven, result suggesting that *the stochastic Riemannian metric, when pulled back from a high-dimensional space, converges to its expectation as the dimension of the space increases.* This claim could potentially be established by building upon existing findings in random matrix theory, or be a by-product of concentration of measure. The classical concentration of measure theorem asserts that independently and identically distributed random points tend to concentrate in a thin layer near a surface. Intuitively, if the points precisely lie on the manifold, the pullback metric becomes deterministic. If this is proved to be true, an additional direction would involve examining *how deterministic is the stochastic pullback metric* with respect to other parameters, such as the dimension of the latent space and the curvature of the manifold.

Another aspect is to consider the general strategy adopted to tackle these stochastic metrics. In this study, we compute geodesics by minimizing an expected metric. Alternatively, we could consider the geodesic equation with stochastic Christoffel symbols. This leaves us with two directions. The first one involves computing their expectation and solve the corresponding system of ordinary differential equations. Adams (2021) and Feldager (2022) have explored this approach, but unfortunately, the expected Christoffel symbols lack a straightforward expression that can be readily utilized. The second path involves directly solving a set of stochastic differential equations, a considerably more complex task.

8.3 Pulling back information geometry

In the paper **Pulling back information geometry** (Chapter 7), we pullback of the Fisher-Rao metric through a Variational AutoEncoder (VAE) decoding to several distributions. This method allows us to easily compute the geodesics in the latent space using the KL-divergence, and it has led to convincing results in practice.

One limitation of this work is that the latent space has to be constrained for the geodesics stay within the support of the data. This is a consequence of the Fisher-Rao metric being defined for the entire space of probability distributions, while the VAE only decodes for a subset of this space. We might wonder if another, more elegant, solution could solve this problem.

This work raises a more fundamental question: *How information geometry compares to optimal transport?*. Both optimal transport and information geometry provide us with Riemannian metrics to navigate the space of probability distributions. Optimal

transport focuses on determining the most efficient way to move from one distribution to another, minimizing the associated cost. In contrast, information geometry is based on the coordinate invariant properties of statistical inference. This first provides us with the Wasserstein p-distance, a Finsler metric (Agueh 2012) that reduces to a Riemannian metric when $p = 2$ (Otto 2001). The second provides us with the Fisher-Rao metric, a Riemannian metric.

Khan and Zhang (2022) nicely describes the difference between those distances as horizontal and vertical. The Wasserstein distance measures the cost of transport for moving from one distribution to another, and establish a notion of spatial distance. It is also more sensitive than the KL-divergence to the properties of the underlying space. The Fisher-Rao metric, in contrast, assumes that both distributions are in the same place but compares the height of one distribution with respect to another.

Recent research has made progress in comparing both fields from a pseudo-Riemannian geometric perspective (Wong and Yang 2022) and geometric thermodynamics (Ito 2023). However, the best approach for navigating the random data space through the space of probability distributions remains an open question.

8.4 Open problems and general questions

Finally, let us circle back to the two initial hypotheses that are necessary for the application of all the methods studied so far.

First, the **manifold hypothesis** assumes that the data space or parameter space can be represented by a low-dimensional manifold. We might wonder *if our data actually lie on a manifold*, prompting us to explore methods for verifying this hypothesis. Fefferman et al. (2016) have outlined the necessary conditions for developing an algorithm that can test the manifold hypothesis. Specifically, the manifold itself should exhibit a reasonable structure, such as being of bounded volume and possessing a minimal reach. In the presence of noisy data, the requirements become more stringent, necessitating a manifold with a sufficiently large reach and a certain number of data points (Fefferman et al. 2019). In practice, however, there is currently no algorithm that can definitively test whether the data lies on a manifold.

Second, the **immersion hypothesis** suggests that a map from a latent space to a higher-dimensional space should function as an immersion. Yet, being an immersion may not be a sufficient condition when we are also interested in the topological characteristics of the manifold. To preserve the manifold's topology, an embedding is required. This means that, when traversing a close loop the data space, may not exhibit closure through an immersion, but it will through an embedding. Additionally, it is crucial for the latent space to possess a minimal dimension in order to maintain the topological characteristics of the manifold. According to the Whitney embedding theorem, any smooth n-dimensional manifold can be smoothly embedded in Euclidean 2n-space. Although representing the data space on a 2-dimensional plane may be convenient, there is a high probability of disrupting the topology of the data manifold, which is an important consideration to bear in mind. Finally, assuming the mapping to be an immersion might also be overly stringent in practice, leading to potential scenarios where the Jacobian is not injective. This could result in problems associated with a degenerate metric, such as a null distance between distinct points or a null volume on the manifold. Further study is needed to understand the circumstances under which the pullback metric degenerates, as many results of Riemannian geometry would not hold anymore.

Appendix

On the curvature of the loss landscape

A

A.1 A primer on curvatures in Riemannian geometry

The key strength of the Riemannian geometry is to allow for calculations to be conducted independently of the choice of the coordinates. However, this flexibility results in more sophisticated computations. Specifically, as a vector moves across a manifold, its local coordinates also change. We must consider this shift, which is accomplished by including a correction factor, denoted as Γ, to the derivative of the vector. These factors Γ are known as **Christoffel symbols**.

Definition A.1.1 Christoffel symbols
Let $(\mathcal{M}, g)$ be a Riemannian manifold, and $\mathbf{u}$ and $\mathbf{v}$ two vector fields on $\mathcal{M}$. On the manifold, we need to add the **Christoffel symbols** Γ_{ij}^k to account for the variation of the local basis represented by $\mathbf{e}_i$. The covariant derivative, or **connection**, is then defined by:

$$\nabla_{\mathbf{u}}\mathbf{v} = u^i \partial_i v^j \mathbf{e}_j + u^i v^j \Gamma_{ji}^k \mathbf{e}_k,$$

with $\nabla_{\mathbf{u}}\mathbf{v} = u^i \partial_i v^j \mathbf{e}_j$ the covariant derivative of $\mathbf{v}$ along $\mathbf{u}$ in the Euclidean plane. We can further compute the Christoffel symbols based on the Riemannian metric tensor g_{ij}:

$$\Gamma_{ij}^k = \frac{1}{2} g^{kl} \left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij} \right),$$

Now, we are interested in the concept of curvature. In Riemannian geometry, the curvature is defined as the deviation of the manifold from the Euclidean plane. The principal intrinsic tool that assess the curvature of a manifold is the **Riemann curvature tensor**, denoted R. It characterises the change of the direction of a vector, when transported along an infinitesimally small closed loop. The Riemannian curvature tensor is defined the following way:

Definition A.1.2 Riemann curvature tensor
Let $(\mathcal{M}, g, \nabla)$ be a Riemannian manifold. The **Riemannian curvature tensor** is defined by:

$$\mathrm{R}(\mathbf{x}, \mathbf{y}; \mathbf{z}) = \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} \mathbf{z} - \nabla_{\mathbf{z}} \nabla_{\mathbf{y}} \mathbf{z} - \nabla_{[\mathbf{x},\mathbf{y}]} \mathbf{z},$$

for any vector fields $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathfrak{X}(\mathcal{M})$, with $[\cdot, \cdot]$ the Lie bracket. At the local basis represented by $\mathbf{e}_i$, it can be expressed in terms of indices: $\mathrm{R}_{ijk}^l = \mathbf{e}^l \mathrm{R}(\mathbf{e}_j, \mathbf{e}_k; \mathbf{e}_i)$, and in terms of the Christoffel symbols as:

$$R_{ijk}^l = \partial_i \Gamma_{jk}^l - \partial_j \Gamma_{ik}^l + \Gamma_{jk}^m \Gamma_{im}^l - \Gamma_{ik}^m \Gamma_{jm}^l$$

The Riemann curvature tensor being a fourth order tensor, it can be difficult to interpret. Instead, we can look at a scalar quantity called the **scalar curvature**, or equivalently the scalar Ricci curvature, which is a contraction of the Riemann curvature tensor.

Definition A.1.3 Scalar curvature
Let $(\mathcal{M}, g)$ be a Riemannian manifold. The **scalar curvature** is defined as:

$$\mathrm{S} = g^{ij} \mathrm{R}_{ikj}^k,$$

using the Einstein summation convention, with g^{ij} the inverse of the metric tensor g_{ij}, and R_{ikj}^k the components of the Riemannian curvature tensor.

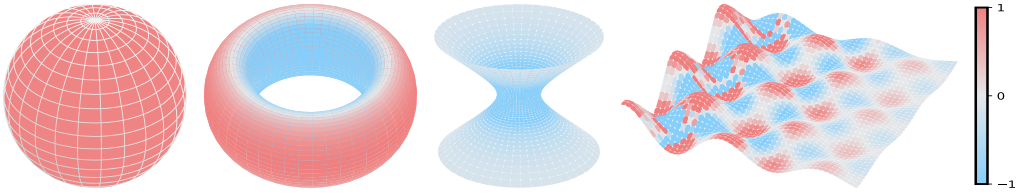


Figure A.1: The scalar curvature is plotted for different surfaces in $\mathbb{R}^3$. The sphere has always a positive scalar curvature since it is convex, the hyperboloid is always negative. In those figures, the values of the scalar curvature have been normalised between -1 and 1.

Just like the Riemannian curvature tensor and the Riemannian metric tensor, the scalar curvature is defined for every point on the manifold. The scalar curvature is null when the manifold is isometric to the Euclidean plane. It is be negative when the manifold is hyperbolic, or positive when the manifold is spherical.

By definition, the scalar curvature is an intrinsic quantity, meaning that it does not depend on the ambient space. As a consequence, the scalar curvature is equivariant under diffeomorphisms. If we map a manifold $(\mathcal{M}, g)$ to another manifold $(\mathcal{M}', g', \nabla')$ with a diffeomorphism $\varphi : \mathcal{M}' \to \mathcal{M}$, we can express the connection ∇' as the pullback of ∇: $\nabla' = d\varphi^* \nabla$. The curvature of the pullback connection is the pullback of the curvature of the original connection. In other terms: $d\varphi^* \mathrm{S}(\nabla) = \mathrm{S}(d\varphi^* \nabla)$ (Andrews and Hopper 2010, Proposition 2.59). In particular, if φ is an isometry: $\mathrm{S}(\nabla) = \mathrm{S}(\nabla')$.

A.2 Theoretical results

A.2.1 Definition of the scalar curvature and other curvature measures

Proposition A.2.1
The Christoffel symbols of the metric $\mathbf{G} = \mathbf{I}_q + \nabla_x f \nabla_x f^\top$, in the parameter space $\Omega \subset \mathbb{R}^q$ with f the loss function is given by:

$$\Gamma^i_{kl} = \frac{f_{,i} f_{,kl}}{1 + \|\nabla f\|^2}$$

Proof. We use below the Einstein sum notation, and in particular, for the scalar function f: $\partial_i \partial_j f = f_{,ij}$. The Christoffels symbols are obtained with the Riemannian metric:

$$\Gamma^i_{kl} = \frac{1}{2} g^{im} \left(g_{mk,l} + g_{ml,k} - g_{kl,m} \right)$$

Our metric is $\mathbf{G} = \mathbf{I}_q + \nabla f \ \nabla f^\top$. Using the Sherman-Morrison formula: $\mathbf{G}^{-1} = \mathbf{I}_q - \frac{\nabla f \ \nabla f^\top}{1 + \|\nabla f\|^2}$

$$\begin{aligned} g_{ij} &= \mathbf{G}_{ij} = \delta_{ij} + f_{,i} f_{,j} \\ g_{ij,k} &= f_{,ik} f_{,j} + f_{,i} f_{,jk} \\ g_{mk,l} + g_{ml,k} - g_{kl,m} &= 2 f_{,kl} f_{,m} \\ g^{im} &= \mathbf{G}^{-1}_{im} = \delta_{im} - \frac{f_{,i} f_{,m}}{1 + \|\nabla f\|^2} \end{aligned}$$

Then:

$$\Gamma^i_{kl} = \left(\delta_{im} - \frac{f_{,i}f_{,m}}{1+\|\nabla f\|^2}\right) f_{,kl}f_{,m} = f_{,kl}f_{,i} - \frac{f_{,kl}f_{,i}f^2_{,m}}{1+\|\nabla f\|^2} = \frac{f_{,i}f_{,kl}}{1+\|\nabla f\|^2}.$$

□

Lemma A.2.2
The metric tensor $\mathbf{G} = \mathbf{I}_q + \nabla f\ \nabla f^\top$ has for eigenvalues: $\{1, 1, \cdots, 1, 1+\|\nabla f\|^2\}$.

Proof. $\mathbf{G}$ is a symmetric positive definite matrix, hence it is diagonalisable and all its eigenvectors $\vec{w}$ are orthogonal. Let's note $\mathbf{v} = \nabla f$. For the eigenvector $\mathbf{v}$: $\mathbf{G}\mathbf{v} = (1+\|\mathbf{v}\|^2)\mathbf{v}$. For all the other eigenvectors, $\langle \vec{w}, \mathbf{v}\rangle = 0$ and $\mathbf{G}\vec{w} = \vec{w}$. □

Proposition A.2.3
The contraction of the Christoffel symbols for the metric $\mathbf{G} = \mathbf{I}_q + \nabla f\ \nabla f^\top$:

$$\Gamma^i_{ki} = \frac{f_{,ik}f_{,i}}{1+\|\nabla f\|^2}.$$

Proof. By definition, we have $\Gamma^i_{ki} = \partial_k \ln\sqrt{\det \mathbf{G}}$. By the previous lemma, we know that $\det G = 1+\|\nabla f\|^2 = 1+f^2_{,i}$.

$$\Gamma^i_{ki} = \partial_k \ln\sqrt{\det \mathbf{G}} = \partial_k \ln\sqrt{1+{f_{,i}}^2} = \frac{1}{2}\frac{\partial_k(1+f^2_{,i})}{1+\|\nabla f\|^2} = \frac{f_{,ik}f_{,i}}{1+\|\nabla f\|^2}.$$

Another method is to use the general expression of $\Gamma^i_{kl} = \frac{f_{,i}f_{,kl}}{1+\|\nabla f\|^2}$, and the result is obtained for $i = l$. □

Proposition A.2.4
The Riemannian curvature tensor is given by:

$$R^i_{jkm} = \beta(f_{,ik}f_{,jm} - f_{,jm}f_{,jk}) - \beta^2 f_{,i}f_{,r}(f_{,rk}f_{,im} - f_{,rm}f_{,jk})$$

Proof. The Riemannian curvature tensor is given by: $R^i_{jkm} = \partial_k\Gamma^i_{jm} - \partial_m\Gamma^i_{jk} + \Gamma^i_{rk}\Gamma^r_{jm} - \Gamma^i_{rm}\Gamma^r_{jk}$, and we have for Christoffel symbols: $\Gamma^i_{jm} = \beta f_{,i}f_{,jm}$. We note $\beta = (1+\|\nabla f\|^2)^{-1}$. We have: $\partial_k(\beta f_{,i}f_{,jm}) = \partial_k(\beta)f_{,i}f_{,jm} + \beta(f_{,ik}f_{,jm} + f_{,i}f_{,jmk})$, and $\partial_k(\beta) = -2\beta^2 f_{ka}f_a$.

$$\begin{aligned}
\partial_k\Gamma^i_{jm} &= -2\beta^2 f_{,a}f_{,ak}f_{,i}f_{,jm} + \beta(f_{,ik}f_{,jm} + f_{,i}f_{,jmk})\\
\partial_m\Gamma^i_{jk} &= -2\beta^2 f_{,a}f_{,ak}f_{,i}f_{,jm} + \beta(f_{,im}f_{,jk} + f_{,i}f_{,jkm})\\
\Gamma^i_{rk}\Gamma^r_{jm} &= \beta^2 f_{,i}f_{,rk}f_{,r}f_{,jm}\\
\Gamma^i_{rm}\Gamma^r_{jk} &= \beta^2 f_{,i}f_{,rm}f_{,r}f_{,jk}
\end{aligned}$$

□

Proposition A.2.5
The scalar curvature is given by:

$$\mathrm{S} = \beta\left(\mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H}^2)\right) + 2\beta^2\left(\nabla f^\top(\mathbf{H}^2 - \mathrm{tr}(\mathbf{H})\mathbf{H})\nabla f\right),$$

with $\mathbf{H}$ the Hessian of f.

Proof. We use $\beta^{-1} = 1 + \|\nabla f\|^2$, $\mathbf{H}$ the Hessian of f, and $\|\cdot\|_{1,1}$ the matrix norm $L_{1,1}$. The Ricci tensor is given by:

$$\begin{aligned} R_{ab} = R^i_{aib} &= \beta(f_{,ii}f_{,ab} - f_{,bi}f_{,ai}) - \beta^2 f_{,i}f_{,r}(f_{,ir}f_{,ab} - f_{,br}f_{,ai}) \\ &= \beta(\mathrm{tr}(\mathbf{H})\mathbf{H}_{ab} - \mathbf{H}^2_{ab}) - \beta^2\left((\nabla f^\top \mathbf{H}\nabla f)\mathbf{H}_{ab} - (\mathbf{H}\nabla f)_a(\mathbf{H}\nabla f)_b\right) \end{aligned}$$

The scalar is given by $g^{ab}R_{ab} = \delta_{ab}R_{ab} - \beta f_{,a}f_{,b}R_{ab}$, and we notice:

$$\begin{aligned} \mathbf{H}_{aa} &= \mathrm{tr}(\mathbf{H}) \\ f_{,a}\mathbf{H}_{ab}f_b &= \nabla f^\top \mathbf{H}\nabla f \\ (\mathbf{H}\nabla f)_a f_{,a} &= \nabla f^\top \mathbf{H}\nabla f \end{aligned}$$

Then, we have:

$$\begin{aligned} R_{ab} &= R_{aa} - \beta f_{,a}f_{,b}R_{ab} \\ R_{aa} &= \beta(\mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H})^2) - \beta^2\left((\nabla f^\top \mathbf{H}\nabla f)\,\mathrm{tr}(\mathbf{H}) - \nabla f^\top \mathbf{H}^2\nabla f\right) \\ \beta f_{,a}f_{,b}R_{ab} &= \beta^2\left(\nabla f^\top \mathbf{H}\nabla f)\,\mathrm{tr}(\mathbf{H}) - \nabla f^\top \mathbf{H}^2\nabla f\right) - \beta^3\left((\nabla f^\top \mathbf{H}\nabla f)^2 - (\nabla f^\top \mathbf{H}\nabla f)^2\right) \end{aligned}$$

Finally:

$$\mathrm{S} = \beta\left(\mathrm{tr}(\mathbf{H})^2 - \mathrm{tr}(\mathbf{H}^2)\right) - 2\beta^2\left(\nabla f^\top(\mathrm{tr}(\mathbf{H})\mathbf{H} - \mathbf{H}^2)\nabla f\right)$$

□

A.2.2 Perturbations on the weights

Proposition A.2.6
Let $\mathbf{x}_{\min}$ an extremum, $\varepsilon \ll 1$ and $\mathbf{x}$ a normalized vector. Then, minimizing the trace of the square of the Hessian is equivalent to minimizing the influence of the perturbations on the weights:

$$\|f(\mathbf{x}_{\min} + \varepsilon\mathbf{x}) - f(\mathbf{x}_{\min})\|_2^2 \leq \frac{1}{4}\varepsilon^4\,\mathrm{tr}(\mathbf{H}^2_{\min}) \tag{A.1}$$

Proof. The general Taylor expansion on f at $\mathbf{x}_{\min} + \varepsilon\mathbf{x}$, with $\varepsilon \ll 1$ is:

$$f(\mathbf{x}_{\min} + \varepsilon\mathbf{x}) = f(\mathbf{x}_{\min}) + \varepsilon\mathbf{x}^\top\mathbf{J} + \frac{\varepsilon^2}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x} + o(\varepsilon^2\|\mathbf{x}\|^2).$$

We now assume that $\mathbf{x}$ is normalized such that $\|\mathbf{x}\| = 1$. Note that, if $\mathbf{x}$ is an eigenvector of $\mathbf{H}$ then: $\mathbf{x}^\top\mathbf{H}\mathbf{x} = \mathrm{tr}(\mathbf{H})$. In general, each element of the vector is inferior to 1: $\mathbf{x}_i^2 \leq 1$ and so, $\lambda_i^2\mathbf{x}_i^4 \leq \lambda_i^2$. Furthermore, we have $\mathbf{J}(\mathbf{x}_{\min}) = 0$. Thus:

$$\|f(\mathbf{x}_{\min} + \mathbf{x}) - f(\mathbf{x}_{\min})\|_2^2 = \frac{\varepsilon^4}{4}\left(\mathbf{x}^\top\mathbf{H}\mathbf{x}\right)^2 + o(\varepsilon^4) \leq \frac{\varepsilon^4}{4}\,\mathrm{tr}(\mathbf{H}^2) + o(\varepsilon^4)$$

□

A.2.3 Curvature over minibatches

Proposition A.2.7
The Scalar curvature of the hessian of the full dataset is not equal to the expectation of the Scalar curvature over mini-batches. That is there exists a dataset, $\mathcal{D}$, and mini-batches, $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ such that:

$$R(\mathbf{H}_{\mathcal{D}}) \neq \mathbb{E}[R(\mathbf{H}_{\mathcal{B}_i})]$$

Proof. Suppose we have a dataset $\mathcal{D}$ and mini-batches $\{\mathcal{B}_1, \mathcal{B}_2\}$ such that the Hessians over the minibatches are given by:

$$\begin{bmatrix} -2 & 0 \\ 4 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 1 \\ 2 & -2 & \end{bmatrix}$$

They both have equal trace, -1, and their scalar curvatures are -2 and -6 respectively. The hessian over the full dataset is given by:

$$\begin{bmatrix} -1 & 2 \\ 6 & -1 \end{bmatrix}$$

This has the same trace as the minibatches but its scalar curvature is -22 not equal to the average of the scalar curvatures over minibatches.

□

Identifying latent distances with Finslerian geometry

B

B.1 A primer on Geometry

The main purpose of the paper is to define and compare two legitimate metrics to compute the average length between random points. Before going further, it's important to formally define the two metrics (Riemannian and Finsler metrics, respectively) which we do in sections B.1.2 and B.1.3. They are both constructed on topological manifolds, the definition of which is recalled in section B.1.1. We finally introduce the notion of random manifold in section B.1.4, which is the last notion needed to frame our problem of interest: which metric should we use to compute the average distance on a random manifold?

B.1.1 Topological and differentiable manifolds

This section aims to define core concepts in differential geometry that will be used later to define Riemannian and Finsler manifolds. Recall that two topological spaces are called homeomorphic if there is a continuous bijection between them with continuous inverse.

Definition B.1.1
A d-dimensional **topological manifold** $\mathcal{M}$ is a second-countable Hausdorff topological space such that every point has an open neighborhood homeomorphic to an open subset of $\mathbb{R}^d$.

Let $\mathcal{M}$ be a topological manifold. This means that for any $x \in \mathcal{M}$ there is an open neighborhood U_x of x and a homeomorphism $\phi_{U_x} : U_x \to \mathbb{R}^d$ onto an open subset of $\mathbb{R}^d$. Suppose that $x, y \in \mathcal{M}$ are such that $U_x \cap U_y \neq \emptyset$, let $U = U_x$, $V = U_y$ and consider the so-called coordinate change map

$$\phi_V \circ \phi^{-1}_{U|\phi_U(U\cap V)} : \phi_U(U \cap V) \to \mathbb{R}^d.$$

We call $\mathcal{M}$ together with an open cover $\{U_x\}_{x\in\mathcal{M}}$ as above a **differentiable** or **smooth** manifold if the coordinate maps are infinitely differentiable.

Beyond these technical definitions, one can imagine a differentiable manifold as a well-behaved smooth surface that possesses *locally* all the topological properties of a Euclidean space. All the manifolds in this paper are assumed to be differentiable and connected manifolds.

Definition B.1.2
We also define, for a differentiable manifold $\mathcal{M}$, the **tangent space** $\mathcal{T}_x M$ as the set of all the tangent vectors at $x \in \mathcal{M}$, and the **tangent bundle** $\mathcal{TM}$ the disjoint union of all the tangent spaces: $\mathcal{TM} = \bigcup_{x\in\mathcal{M}} \mathcal{T}_x M$.

So far, we have only defined topological and differential properties of manifolds. In order to compute geometric quantities, we need to equip those with a metric that helps us derive useful quantities such as lengths, energies and volumes. A metric is a scalar-valued function that is defined for each point on the topological manifold and takes as inputs one or two vectors (depending on the type of metric) from the tangent space at the specific point. Such a function can either be defined as a scalar product between two vectors, this is the case of a Riemannian metric or, in the case of a Finsler metric, it

is defined similarly to the norm of a vector. We will formally define these metrics and highlight their differences in the following sections.

B.1.2 Riemannian manifolds

Definition B.1.3
Let $\mathcal{M}$ be a manifold. A **Riemannian metric** is a map assigning at each point $x \in \mathcal{M}$ a scalar product $G(\cdot,\cdot) : \mathcal{T}_x M \times \mathcal{T}_x M \to \mathbb{R}$, with G a positive definite bilinear map, which is smooth with respect to x. A smooth manifold equipped with a Riemannian metric is called a **Riemannian manifold**. We usually express the metric as a symmetric positive definite matrix G, where we have for two vectors $u, v \in \mathcal{T}_x M$: $G(u,v) = \langle u, v\rangle_G = u^\top G v$. We further define the induced **norm**: $v \in \mathcal{T}_x M$, $\|v\|_G = \sqrt{G(v,v)}$.

The Riemannian metric here can either refer to the scalar product G itself, or the associated metric tensor G.

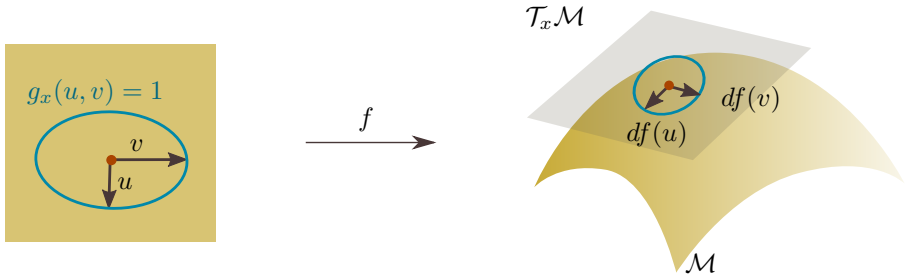


Figure B.1: f is an immersion that maps a low dimensional manifold to a high dimensional manifold $\mathcal{M}$. On $\mathcal{M}$, a tangent plane $\mathcal{T}_x\mathcal{M}$ is drawn at x. The indicatrix of the Euclidean metric is plotted in blue. When this metric is pulled-back through f, the low dimensional space is now equipped with the pullback metric g, which is a Riemannian metric by definition. The vectors $df(u)$ and $df(v)$ are called the push-forwards of the vectors u and v through f.

Definition B.1.4
We consider a curve $\gamma(t)$ and its derivative $\dot\gamma(t)$ on a Riemannian manifold $\mathcal{M}$ equipped with the metric g. Then, we define the **length of the curve**:

$$L_G(\gamma) = \int \|\dot\gamma(t))\|_G \, dt = \int \sqrt{g_t(\dot\gamma(t), \dot\gamma(t))}\, dt,$$

where $g_t = g_{\gamma(t)}$. Locally length-minimizing curves between two connecting points are called **Geodesics**.

Definition B.1.5
The **curve energy** is defined as:

$$E_G(\gamma) = \int \|\dot\gamma(t))\|_G^2 \, dt = \int g_t(\dot\gamma(t), \dot\gamma(t))\, dt.$$

There are two interesting properties to note about the length of a curve and the curve energy. First, the length is parametrization invariant: for any bijective smooth function η on the domain of γ we have that $L_G(\gamma \circ \eta) = L_G(\gamma)$. We also say the Riemannian metric gives us intrinsic coordinates to compute the length. Secondly, for a given curve γ, we have: $L_G(\gamma)^2 \leq 2E_G(\gamma)$. Because of the invariance of the curve, when we aim to minimize it, a solver can find an infinite number of solutions. On the other hand, the curve energy is convex and will lead to a unique solution. Thus, to obtain a geodesic, instead of solving the corresponding ODE equations, or directly minimizing lengths, it is easier in practice to minimize the curve energy, as a minimal energy gives a minimal length.

The Riemannian metric also provides us with an infinitesimal volume element that relates our metric G to an orthonormal basis, the same way the Jacobian determinant accommodate for a change of coordinates in the change of variables theorem.

Definition B.1.6
In local coordinates $(e^1, \cdots, e^d)$, the **volume form** of the Riemannian manifold $\mathcal{M}$, equipped with the metric tensor G, is defined as: $dV_G = V_G(x)e^1 \wedge \cdots \wedge e^d$, with:

$$V_G(x) = \sqrt{\det(G)}.$$

Remark B.1.1 The symbol $\wedge$ represents the wedge product, and it is used to manipulate differential k-forms. Here, the basis vectors $(e^1, \cdots, e^d)$ form a d-dimensional parallelepiped $(e^1 \wedge \cdots \wedge e^d)$ with unit volume.

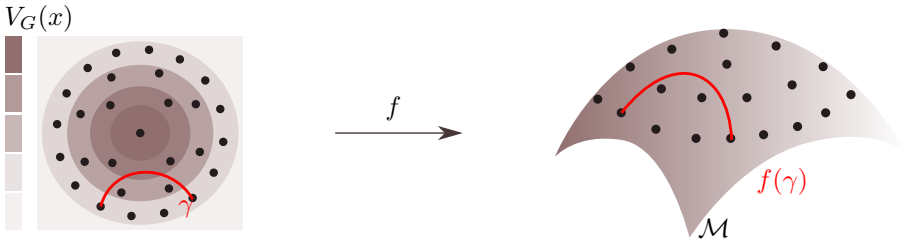


Figure B.2: Once the low dimensional manifold is equipped with a metric that captures the inherent structure of the high dimensional manifold, we can compute a geodesic γ, by minimizing the energy functional between two points. The geodesic $f(\gamma)$ will be the shortest path between two points on the manifold $\mathcal{M}$. The volume measure V_G can also be used to integrate functions over regions of the manifold, as we would do in the Euclidean space. It can also be linked to the density of the data: if the data points are uniformly distributed over the high-dimensional manifold, in the low-dimensional manifold, a low volume would correspond to a high density of data. It is a useful way to give more information about the distribution of the data.

B.1.3 Finsler manifolds

Finsler geometry is often described as an extension of Riemannian geometry, since the metric is defined in a more general way, lifting the quadratic constraint. In particular, the norm of a Riemannian metric is a Finsler metric, but the converse is not true.

Definition B.1.7
Let $F : \mathcal{T}\mathcal{M} \to \mathbb{R}_+$ be a continuous non-negative function defined on the tangent bundle $\mathcal{T}\mathcal{M}$ of a differentiable manifold M. We say that F is a **Finsler metric** if, for each point x of $\mathcal{M}$ and v on $\mathcal{T}_x M$, we have:

1. Positive homogeneity: $\forall \lambda \in \mathbb{R}_+$, $F(\lambda v) = \lambda F(v)$.
2. Smoothness: F is a C^∞ function on the slit tangent bundle $\mathcal{T}\mathcal{M} \setminus \{0\}$.
3. Strong convexity criterion: the Hessian matrix $g_{ij}(v) = \frac{1}{2}\frac{\partial^2 F^2}{\partial v^i v^j}(v)$ is positive definite for non-zero v.

A differentiable manifold $\mathcal{M}$ equipped with a Finsler metric is called a **Finsler manifold**.

Here, it is worth noting that, for a given point in the manifold, the Finsler metric is defined with only one vector in the tangent space, while the Riemannian metric is defined with two vectors. Moreover, from the previous definition, we can deduce that the metric is:

1. Positive definite: for all $x \in \mathcal{M}$ and $v \in \mathcal{T}_x M$, $F(v) \geq 0$ and $F(v) = 0$ if and only if $v = 0$.

2. Subadditive: $F(v+w) \leq F(v) + F(w)$ for all $x \in \mathcal{M}$ and $v, w \in \mathcal{T}_x M$.

We say that F is a Minkowski norm on each tangent space $\mathcal{T}_x M$. Furthermore, if F satisfies the reversibility property: $F(v) = F(-v)$, it defines a norm on $\mathcal{T}_x M$ in the usual sense.

Similarly to Riemannian geometry, lengths, energies and volumes can be defined directly from the Finsler metric:

Definition B.1.8
We consider a curve γ and its derivative $\dot{\gamma}$ on a Finsler manifold $\mathcal{M}$ equipped with the metric F. We define the **length of the curve** as follows:

$$L_F(\gamma) = \int F(\dot{\gamma}(t))dt.$$

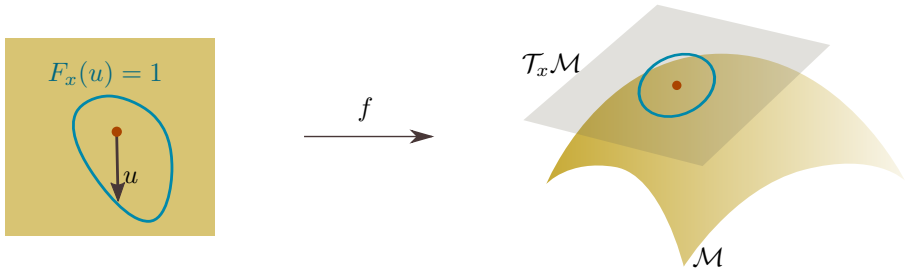


Figure B.3: f is an immersion that maps a low dimensional manifold to a high dimensional manifold $\mathcal{M}$. On $\mathcal{M}$, a tangent plane $\mathcal{T}_x\mathcal{M}$ is drawn at x. Compared to the Riemannian manifold, the Finsler indicatrix, which represents the all the vectors $u \in \mathcal{T}_x\mathcal{M}$ such that $F_x(u) = 1$, is not necessarily an ellipse. It can be asymmetric if the metric is asymmetric itself. It is always convex.

Definition B.1.9
The **curve energy** is defined as: $E_F(\gamma) = \int F(\dot{\gamma}(t))^2 dt$.

Not only are the definitions strikingly similar, they also share the same properties. The curve length is also invariant under reparameterization, and upper bounded by the curve energy. Computing geodesics on a manifold is reduced to a variational optimization problem. These propositions are proved in detail in Lemmas B.2.4 and B.2.5, in the appendix.

In Riemannian geometry, the volume measure defined by the metric is unique. In Finsler geometry, different definitions of the volume exist, and they all coincide with the Riemannian volume element when the metric is Riemannian. The most common choices of volume forms are the Busemann-Hausdorff measure and the Holmes-Thompson measure. According Wu (2011), depending on the Finsler metric and the topological manifold, some choices seem more legitimate than others. In this paper, we decided to only focus on the Busemann-Hausdorff volume, as its definition is the most commonly used and leads to easier derivations. We will later show that in high dimensions, our Finsler metric converges to a Riemannian metric, and thus, the results obtained for the Busemann-Hausdorff volume measure are also valid for the Holmes-Thomson volume measure.

Definition B.1.10
For a given point x on the manifold, we define the **Finsler indicatrix** as the set of vectors in the tangent space such that the Finsler metric is equal to one: $\{v \in \mathcal{T}_x M | F(v) = 1\}$. We denote the Euclidean unit ball in $\mathbb{R}^d$ by $\mathbb{B}^d(1)$ and for measurable subsets $S \subseteq \mathbb{R}^d$ we use $\text{vol}(S)$ to denote the standard Euclidean volume of S. In

local coordinates $(e^1, \cdots, e^d)$ on a Finsler manifold $\mathcal{M}$, the **Busemann-Hausdorff volume** form is defined as $dV_F = V_F(x)e^1 \wedge \cdots \wedge e^d$, with:

$$V_F(x) = \frac{\text{vol}(\mathbb{B}^d(1))}{\text{vol}(\{v \in \mathcal{T}_x M | F(v) < 1\})}.$$

We can interpret the volume as the ratio between the Euclidean ball, and a convex ball whose radius is defined as a unit Finsler metric. If the Finsler metric is replaced by a Riemannian metric, the volume of the indicatrix will be an ellipsoid whose semi-axis are equal to the inverse of the square root of the metric's eigenvalues. The Finsler volume then reduces to the definition of the Riemannian volume.

B.1.4 Random manifolds

So far, we have only considered deterministic data points lying on a manifold. If we consider our data to be random variables, we will need to define the associated random metric and manifold.

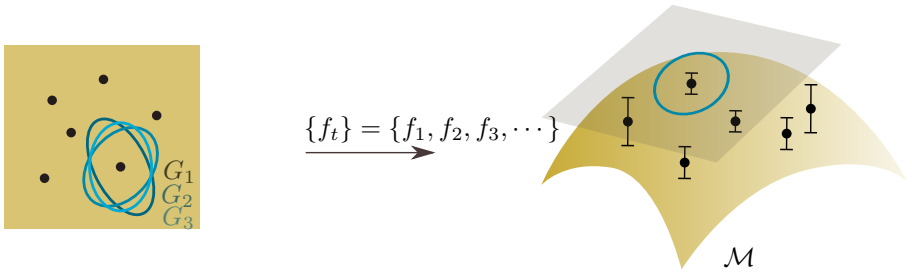


Figure B.4: Usually the immersion f would be deterministic. In the case of most generative models, where f is described by a GP-LVM, or the decoder of a VAE, the immersion is stochastic. The pullback metric is stochastic de facto.

As said previously, if we have a function $f : \mathbb{R}^q \to \mathbb{R}^D$ that parametrizes a manifold, then we can construct a Riemannian metric $G = J_f^\top J_f$, with J_f the Jacobian of the function f. In the previous cases, we assumed f to be a deterministic function, and so is the metric. We construct a stochastic Riemannian metric in the same way, with f being a stochastic process. A stochastic process $f : \mathbb{R}^q \to \mathbb{R}^D$ is a random map in the sense that samples of the process are maps from $\mathbb{R}^q$ to $\mathbb{R}^D$ (the so-called sample paths of the process).

Definition B.1.11
A stochastic process $f : \mathbb{R}^q \to \mathbb{R}^D$ is smooth if the sample paths of f are smooth. We call a smooth process f a **stochastic immersion** if the Jacobian matrix of its sample paths has full rank everywhere. We can then define the **stochastic Riemannian metric** $G = J_f^\top J_f$.

The terms *stochastic* and *random* are used interchangeably. The definition of the stochastic immersion is fairly important, as it means that its Jacobian is full rank. Since the Jacobian is full rank, the random metric G is positive definite, a necessary condition to define a Riemannian metric. Another definition of a stochastic Riemannian metric would be the following:

Definition B.1.12
A **stochastic Riemannian metric** on $\mathbb{R}^q$ is a matrix-valued random field on $\mathbb{R}^q$ whose sample paths are Riemannian metrics. A **stochastic manifold** is a differentiable manifold equipped with a stochastic Riemannian metric.

Any matrix drawn from this stochastic metric would be a proper Riemannian metric. When using the random Riemannian metric on two vectors $u, v \in \mathcal{T}_x M$, $G(u, v) = u^\top G v$ is a random variable, but both u, v are deterministic vectors. From this definition, it follows that the length, the energy and the volume are random variables.

Table B.1: Comparison of Riemannian and Finsler metrics.

Object	Riemann	Finsler
metric	$g : \mathcal{T}_x M \times \mathcal{T}_x M \to \mathbb{R}$	$F : \mathcal{T}M \to \mathbb{R}_+$
length structure	$L_G(\gamma) = \int \sqrt{g_t(\dot\gamma(t), \dot\gamma(t))}\, dt$	$L_F(\gamma) = \int F(\dot\gamma(t))\, dt$
energy structure	$E_G(\gamma) = \int g_t(\dot\gamma(t), \dot\gamma(t))\, dt$	$E_F(\gamma) = \int F(\dot\gamma(t))^2\, dt$
volume element	$V_G(x) = \sqrt{\lvert \det G \rvert}$	$V_F(x) = \mathrm{vol}(\mathbb{B}^n(1))/\mathrm{vol}(\{v \in \mathcal{T}_x M \lvert F(x, v) < 1\})$ Busemann-Hausdorff volume measure

B.2 Proofs

One of the main challenges of this paper is to find coherent notations while respecting the tradition of two geometric fields. In Riemannian geometry, we principally use a metric, noted $g_p : \mathcal{T}_p\mathcal{M} \times \mathcal{T}_p\mathcal{M} \to \mathbb{R}_+$, that is defined as an inner product and thus can induce a norm, but is not a norm. In Finsler geometry, we call interchangeably Finsler function, Finsler metric or Finsler norm, the norm traditionally noted $F : \mathcal{T}\mathcal{M} \to \mathbb{R}_+$, with $F_p(u) \coloneqq F(p, u)$ defined at a point $p \in \mathcal{M}$ for a vector $u \in \mathcal{T}_p\mathcal{M}$. We will assume that all our metric are always defined for a specific point z (or p) on our manifold $\mathcal{Z}$ (or $\mathcal{M}$), and so we will just drop this index. The following notations will be used:

Stochastic pullback metric tensor	$G = J_f^\top J_f$
Stochastic pullback metric	$\tilde g : (u, v) \to u^\top G u$
Expected Riemannian metric	$g : (u, v) \to u^\top \mathbb{E}[G] v$
Stochastic pullback induced norm	$\lVert\cdot\rVert_G : u \to \sqrt{u^\top G u}$
Expected Riemannian induced norm	$\lVert\cdot\rVert_R : u \to \sqrt{u^\top \mathbb{E}[G] u} \coloneqq \sqrt{g(u, u)}$
Finsler metric	$\lVert\cdot\rVert_F : u \to \mathbb{E}[\sqrt{u^\top G u}] \coloneqq F(u)$

B.2.1 Finslerian geometry of the expected length

In this section, we will always let $f : \mathbb{R}^q \to \mathbb{R}^D$ be a stochastic immersion, J_f its Jacobian, and $G = J_f^\top J_f$ a metric tensor. We will first prove that the function $F : \mathcal{T}M \to \mathbb{R} : v \to \mathbb{E}\left[\sqrt{v^\top G v}\right]$ is a Finsler metric. Then, for the specific case where J_f follows a non-central normal distribution, the Finsler metric F defined as the expected length follows a non-central Nakagami distribution and can be expressed in closed form.

To prove that the function F is indeed a Finsler metric, we will need to verify the criteria above, among them the strong convexity criterion is less trivial to prove than the others. It will be detailed in Lemma B.2.3. Strong convexity means that the Hessian matrix $\frac{1}{2}\mathbf{H}(F(v)^2) = \frac{1}{2}\frac{\partial^2 F^2}{\partial v^i v^j}(v)$ is strictly positive definite for non-negative v. This matrix, when F is a Finsler function, is also called the fundamental form and plays an important role in Finsler geometry. To prove the strong convexity criterion, we will need the full expression of the fundamental form, detailed in Lemma B.2.1.

Lemma B.2.1
The Hessian matrix $\frac{1}{2}\mathbf{H}(F(v)^2)$ of the function $F(v) = \mathbb{E}\left[\sqrt{v^\top G v}\right]$ is given by

$$\frac{1}{2}\mathbf{H}(F(v)^2) = \mathbb{E}\left[(v^\top G v)^{\frac{1}{2}}\right]\mathbb{E}\left[(v^\top G v)^{-\frac{1}{2}}G - (v^\top G v)^{-\frac{3}{2}}Gvv^\top G\right] + \mathbb{E}\left[(v^\top G v)^{-\frac{1}{2}}G\right]^2 vv^\top.$$

Proof. Let G be a random positive definite symmetric matrix and define $g : \mathbb{R}^q \to \mathbb{R} : v \mapsto \sqrt{v^\top G v}$, where v is considered a column vector. We would like to know the different derivatives of g with respect to v. We name by default J_g and H_g, its Jacobian and Hessian matrix. Using the chain rule, we have: $J_g = (v^\top G v)^{-\frac{1}{2}} v^\top G$ and $H_g = (v^\top G v)^{-\frac{1}{2}} G - (v^\top G v)^{-\frac{3}{2}}(G v v^\top G)$.

For the rest of the proof, we need to show that derivatives and expectation values commute.

Using the Fubini theorem, we can show that that derivatives and the expectation values commute.

For $F : \mathbb{R}^q \to \mathbb{R} : v \mapsto \mathbb{E}[\sqrt{v^\top G v}]$,

$$\mathbf{H}(F) = \mathbb{E}[H_g] = \mathbb{E}\left[(v^\top G v)^{-\frac{1}{2}}G - (v^\top G v)^{-\frac{3}{2}}Gvv^\top G\right]$$

$$\nabla F = \mathbb{E}[J_g] = \mathbb{E}[(v^\top G v)^{-\frac{1}{2}} G v].$$

We now consider the function $h : \mathbb{R}^q \to \mathbb{R} : v \mapsto \mathbb{E}[\sqrt{v^\top G v}]^2 = F(v)^2$. Using the chain rule and changing the order of expectation and derivatives, we have its Hessian

$$H_h = 2F \cdot \mathbf{H}[F] + 2\nabla F^\top \nabla F = 2\mathbb{E}[g]\mathbb{E}[H_g] + 2\mathbb{E}[J_g]^\top \mathbb{E}[J_g].$$

Finally, replacing J_g and H_g previously obtained in this expression, we conclude:

$$\frac{1}{2}H_h(x, v) = \mathbb{E}\left[(v^\top G v)^{\frac{1}{2}}\right]\mathbb{E}\left[(v^\top G v)^{-\frac{1}{2}}G - (v^\top G v)^{-\frac{3}{2}}Gvv^\top G\right] + \mathbb{E}\left[(v^\top G v)^{-\frac{1}{2}}G\right]^2 vv^\top.$$

□

Remark B.2.1 Before going further, it's important to note that $G = J_f^\top J_f$ is a random matrix that is positive definite: it is symmetric by definition and has full rank. The later statement is justified by the assumption that the stochastic process $f : \mathbb{R}^q \to \mathbb{R}^D$ is an immersion, then J_f is full rank.

Lemma B.2.2
The function $F(v) = \mathbb{E}\left[\sqrt{v^\top G v}\right]$ is:

1. positive homogeneous: $\forall \lambda \in \mathbb{R}_+$, $F(\lambda v) = \lambda F(v)$
2. smooth: $F(v)$ is a C^∞ function on the slit tangent bundle $\mathcal{TM} \setminus \{0\}$

Proof. 1) Let $\lambda \in \mathbb{R}$, then we have: $F(\lambda v) = \mathbb{E}\left[\sqrt{\lambda^2 v^\top G v}\right] = |\lambda|\left[\sqrt{v^\top G v}\right]$.

2) The multivariate function: $\mathbb{R}^q \setminus \{0\} \to \mathbb{R}_+^* : v \to v^\top G v$ is C^∞ and strictly positive, since $G = J_f^\top J_f$ is positive definite. The function $\mathbb{R}_+^* \to \mathbb{R}_+^* : x \to \sqrt{x}$ is also C^∞. Finally, $\mathbb{R}_+^* \to \mathbb{R}_+^* : x \to \mathbb{E}[x]$ is by definition differentiable. By composition, $F(v)$ is a C^∞ function on the slit tangent bundle $\mathcal{TM} \setminus \{0\}$. □

Lemma B.2.3
The function $F(v) = \mathbb{E}\left[\sqrt{v^\top G v}\right]$ satisfies the strong convexity criterion.

Proof. Proving that F satisfies the strong convexity criterion is equivalent to show that the Hessian matrix $H = \frac{1}{2}\mathbf{H}(F(v)^2)$ is strictly positive definite. Thus, we need to prove that $\forall w \in \mathbb{R}^q \backslash \{0\}, w^\top H w > 0$. According to Lemma B.2.1, because the expectation is a positive function, it's straightforward to see that $\forall w \in \mathbb{R}^q \backslash \{0\}, w^\top H w \geq 0$. The tricky part of this proof is to show that $w^\top H w > 0$. This can be obtained if one of the terms ($F \cdot \mathbf{H}(F)$ or $\nabla F^\top \nabla F$) is strictly positive.

First, let's decompose H as the sum of matrices: $H = F\mathbf{H}(F) + \nabla F^\top \nabla F$ (Lemma B.2.1), with:

$$F \cdot \mathbf{H}(F) = \mathbb{E}\left[(v^\top G v)^{\frac{1}{2}}\right] \mathbb{E}\left[(v^\top G v)^{-\frac{3}{2}} \left((v^\top G v)G - Gv(Gv)^\top\right)\right],$$

$$\nabla F^\top \nabla F = \mathbb{E}\left[(v^\top G v)^{-\frac{1}{2}} G\right]^2 v v^\top.$$

We will study two cases: when $w \in \text{span}(v)$, and when $w \notin \text{span}(v)$. We will always assume that $v \neq 0$, and so by definition: $F(v) > 0$.

Let $w \in \text{span}(v)$. We will show that $w^\top \nabla F^\top \nabla F w > 0$. We have $w = \alpha v, \alpha \in \mathbb{R}$. Because F is 1-homogeneous and using Euler theorem, we have: $\nabla F(v) v = F(v)$. Then $(\alpha v)^\top \nabla F^\top \nabla F(\alpha v) = \alpha^2 F^2$, and $\alpha^2 F(v)^2 > 0$.

Let $w \notin \text{span}(v)$. F being a scalar function, we have: $w^\top F \mathbf{H}[F] w = F w^\top \mathbf{H}[F] w$. We would like to show that: $w^\top \mathbf{H}[F] w > 0$. The strategy is the following: if we prove that the kernel of $\mathbf{H}[F]$ is equal to the span(v), then $w \notin \text{span}(v)$ is equivalent to say that $w \notin \ker(\mathbf{H}[F])$ and we can conclude that: $w^\top \mathbf{H}[F] w > 0$. Let's prove $\text{span}(v) \in \ker(\mathbf{H}(F))$. We know that $\mathbf{H}(F)v = 0$, since F is 1-homogeneous, so we have $\text{span}(v) \in \ker(\mathbf{H}(F))$. To obtain the equality, we just need to prove that the dimension of the kernel is equal to 1. Let $z \in \text{span}(v^\top G)^\top$, which is $(Gv)^T z = 0$. We have $\dim(\text{span}(v^\top M)) = 1$, and thus: $\dim(\text{span}(v^\top G)^\top) = q - 1$. Furthermore, $z^\top \mathbf{H}[F] z = z^\top \mathbb{E}\left[M(v^\top M v)^{-\frac{1}{2}}\right] z > 0$, so we can deduce that $\dim(\text{im}(\mathbf{H}[F])) = q - 1$. Using the Rank-Nullity theorem, we conclude that $\dim(\ker(\mathbf{H}(F))) = q - \dim(\text{im}(\mathbf{H}[F])) = 1$, which concludes the proof.

In conclusion, $\forall w \in \mathbb{R}^q \backslash \{0\}, w^\top \frac{1}{2}\mathbf{H}(F(v)^2) w > 0$. The function F satisfies the strong convexity criterion. □

Proposition 6.2.2
Let G be a stochastic Riemannian metric tensor. Then, the function $F_z : \mathcal{T}_z\mathcal{Z} \to \mathbb{R} : u \to \|u\|_F$ defines a Finsler metric, but it is not induced by a Riemannian metric.

Proof. Let's define F as a Riemannian metric: $F : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R} : (v_1, v_2) \to \mathbb{E}\left[\sqrt{v_1^\top G v_2}\right]$. If F were a Riemannian metric, then it would be bilinear, which is clearly not the case. Thus, F is not a Riemannian metric. According to Lemma B.2.2 and Lemma B.2.2, F is a Finsler metric. □

Proposition 6.2.3
Let f be a Gaussian process and J its Jacobian, with $J \sim \mathcal{N}(\mathbb{E}[J], \Sigma)$. The Finsler norm can be written as:

$$F_z : \mathcal{T}_z\mathcal{Z} \to \mathbb{R}_+ : \|v\|_F \coloneqq v \to \sqrt{2}\sqrt{v^\top \Sigma v} \frac{\Gamma(\frac{D}{2} + \frac{1}{2})}{\Gamma(\frac{D}{2})} {}_1F_1\left(-\frac{1}{2}, \frac{D}{2}, -\frac{\omega}{2}\right),$$

with $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J]v)$ a noncentral term, and ${}_1F_1$ the confluent hypergeometric function of the first kind.

Proof. The objective of the proof is to show that, if the Jacobian J_f follows a non-central normal distribution, then, $\forall v \in \mathbb{R}^q$, the expectation $\mathbb{E}[v^\top J_f^\top J_f v]$ will follow a non-central Nakagami distribution. This is a particular case of the derivation of moments of non-central Wishart distributions, previously shown and studied by Kent and Muirhead (1984) and Hauberg (2018b).

By hypothesis, J_f follows a non-central normal distribution: $J_f \sim \mathcal{N}(\mathbb{E}[J], I_D \otimes \Sigma)$. Then, $G = J_f^\top J_f$ follows a non-central Wishart distribution: $G \sim \mathcal{W}_d(D, \Sigma, \Sigma^{-1}\mathbb{E}[J]^\top \mathbb{E}[J])$. According to Kent and Muirhead (1984, Theorem 10.3.5.), $v^\top G v$ will also follow a non-central Wishart distribution: $v^\top G v \sim \mathcal{W}_1(D, v^\top \Sigma v, \omega)$, with: $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J]v)$.

To compute $\mathbb{E}[\sqrt{v^\top G v}]$, we shall look at the derivation of moments. Kent and Muirhead (1984, Theorem 10.3.7.) states that: if $X \sim \mathcal{W}_q(D, \Sigma, \Omega')$, with $q \leq D$, then $\mathbb{E}[(\det(X))^k] = (\det \Sigma)^k 2^{qk} \frac{\Gamma_q(\frac{D}{2}+k)}{\Gamma_q(\frac{D}{2})} {}_1F_1(-k, \frac{D}{2}, -\frac{1}{2}\Omega')$. We directly apply the theorem to our case, knowing that $v^\top G v$ is a scalar term, so $\det(v^\top G v) = v^\top G v$, $q = 1$, and $k = \frac{1}{2}$:

$$\|v\|_F \coloneqq \mathbb{E}[\sqrt{v^\top G v}] = \sqrt{2}\sqrt{v^\top \Sigma v}\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})} {}_1F_1(-\frac{1}{2}, \frac{D}{2}, -\frac{1}{2}\omega)$$

□

Lemma B.2.4
The length of a curve using a Finsler metric is invariant by reparameterization.

Proof. The proof is similar to the one obtained on a Riemannian manifold (Lee (2013), Proposition 13.25), where we make use of the homogeneity property of the Finsler metric.

Let $(\mathcal{M}, F)$ be a Finsler manifold and $\gamma : [a, b] \to \mathcal{M}$ a piece wise smooth curve segment. We call $\tilde{\gamma}$ a reparameterization of γ, such that $\tilde{\gamma} = \gamma \circ \phi$ with $\phi : [c, d] \to [a, b]$ a diffeomorphism. We want to show that $L_F(\gamma) = L_F(\tilde{\gamma})$.

$$\begin{aligned} L_F(\tilde{\gamma}) &= \int_c^d F(\dot{\tilde{\gamma}}(t))\, dt = \int_c^d F(\frac{d}{dt}(\gamma \circ \phi(t)))\, dt \\ &= \int_{\phi^{-1}(a)}^{\phi^{-1}(b)} |\dot{\phi}(t)| F(\dot{\gamma} \circ \phi(t))\, dt = \int_a^b F(\dot{\gamma}(t))\, dt = L_F(\gamma) \end{aligned}$$

□

Lemma B.2.5
If a curve globally minimizes its energy on a Finsler manifold, then it also globally minimizes its length and the Finsler function F of the velocity vector along the curve is constant.

Proof. The curve energy and the curve length are defined as: $E_F(\gamma) = \int_0^1 F^2(\dot{\gamma}(t))dt$ and $L_F(\gamma) = \int_0^1 F(\dot{\gamma}(t))dt$, with $\gamma : [0, 1] \to \mathbb{R}^d$. Let's define f and g two real-valued functions

such that: $f : \mathbb{R} \to \mathbb{R} : t \mapsto F(\dot{\gamma}(t))$ and $g : \mathbb{R} \to \mathbb{R} : t \mapsto 1$. Applying Cauchy-Schwartz inequality, we directly obtain:

$$\left(\int_0^1 F(\dot{\gamma}(t))dt\right)^2 \leq \int_0^1 F(\dot{\gamma}(t))^2 dt \cdot \int_0^1 1^2 dt, \quad \text{which means:} \quad L_F(\gamma)^2 \leq E_F(\gamma).$$

The equality is obtained exactly when the functions f and g are proportional, hence, when the Finsler function is constant. □

B.2.2 Comparison of Riemannian and Finsler metrics

We have defined both a Riemannian ($g : (v_1, v_2) \to v_1^\top \mathbb{E}[G] v_2$) and a Finsler ($F : (x, v) \to \mathbb{E}[\sqrt{v^\top G v}]$) metric, in the hope to compute the average length between two points on a random manifold created by the random field f: $G = J_f^\top J_f$. The main idea of this section is to better compare those two metrics and in what extend they differ in terms of length, energy and volume. From now on, $f : \mathbb{R}^q \to \mathbb{R}^D$ will always be defined as a stochastic non-central Gaussian process. Its Jacobian J_f also follows a non-central Gaussian distribution, $G = J_f^\top J_f$ a non-central Wishart distribution, and $F : (x, v) = \mathbb{E}[\sqrt{v^\top G v}]$ a non-central Nakagami distribution (Proposition 6.2.3). The Finsler metric can be written in closed form.

In section B.2.2, we will see that the Finsler metric is upper and lower bounded by two Riemannian tensors (Proposition 6.3.1), and we can deduce an upper and lower bound for the length, the energy and the volume (Corollary 6.3.2). Then, in section B.2.2, we will show that the relative difference between the Finsler norm and the Riemannian induced norm is always positive and upper bounded a term that is inversely proportional to the number of dimensions D (Proposition 6.3.4). Similarly, we will deduce the same for the length, the energy and the volume (Corollary 6.3.5). From this last results, we can directly conclude in section B.2.2 that both metrics are equal in high dimensions (Corollary 6.3.7). A possible interpretation is that in high dimensions the data distribution obtained on those manifolds becomes more and more concentrated around the mean, reducing the variance term to zero. The manifold becoming deterministic, both metrics become equal.

Remark B.2.2 Most of the following proofs will be a bit technical, as they rely on the closed form expression of the non-central Nakagami distribution. Once proving the main propositions, obtaining the corollaries is straightforward. While we do not have closed form expression of the indicatrix, we will show that it's a monotonous function which can upper and lower bounded.

Bounds on the Finsler metric

Proposition 6.3.1
We define $\alpha = 2\left(\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})}\right)^2$. The Finsler norm: $\|\cdot\|_F$ is bounded by two norms, $\|\cdot\|_{\alpha\Sigma}$ and $\|\cdot\|_R$, induced by the two respective Riemannian metric tensors: the covariance tensor $\alpha\Sigma_z$ and the expected metric tensor $\mathbb{E}[G_z]$.

$$\forall (z, v) \in \mathcal{Z} \times \mathcal{T}_z Z : \ \|v\|_{\alpha\Sigma} \leq \|v\|_F \leq \|v\|_R$$

Proof. Let's first recall that the Finsler function can be written as:

$$\|v\|_F \coloneqq F(v) = \sqrt{2}\sqrt{v^\top \Sigma v}\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})} \cdot {}_1F_1(-\frac{1}{2}, \frac{D}{2}, -\frac{1}{2}\omega).$$

The confluent hypergeometric function is defined as: ${}_1F_1(a,b,z) = \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k} \frac{z^k}{k!}$, with $(a)_k$ and $(b)_k$ being the Pochhammer symbols. Note that, despite their confusing notation, they are defined as rising factorials. By definition, we have: $\frac{(a)_k}{(b)_k} = \frac{\Gamma(a+k)}{\Gamma(b+k)} \frac{\Gamma(b)}{\Gamma(a)}$. We can use the Kummer transformation to obtain:

${}_1F_1(a,b,-z) = e^{-z}\,{}_1F_1(b-a,b,z)$. Replacing $a = -\frac{1}{2}$, $b = \frac{D}{2}$ and $z = \frac{1}{2}\omega$, we finally get:

$$F(v) = \sqrt{2}\sqrt{v^\top \Sigma v} \cdot e^{-z} \sum_{k=0}^{\infty} \frac{\Gamma(\frac{D}{2} + \frac{1}{2} + k)}{\Gamma(\frac{D}{2} + k)} \frac{z^k}{k!}.$$

1) Let's show that: $\forall v \in \mathcal{T}_x M :\ \sqrt{v^\top \alpha \Sigma v} \leq F(v)$, with $\alpha = 2\left(\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})}\right)^2$.

The Pochhammer symbol is defined as $(x)_k = x(x+1)\dots(x+k-1) = \frac{\Gamma(x+k)}{\Gamma(x)}$. For $x \in \mathbb{R}_+^*$, we have: $(x)_k \leq (x+\frac{1}{2})_k$. Thus, $\frac{\Gamma(\frac{D}{2}+k)}{\Gamma(\frac{D}{2})} \leq \frac{\Gamma(\frac{D}{2}+\frac{1}{2}+k)}{\Gamma(\frac{D}{2}+\frac{1}{2})}$. The Gamma function being strictly positive on $\mathbb{R}_+$, we obtain:

$$\begin{aligned}
\frac{\Gamma(\frac{D}{2} + \frac{1}{2})}{\Gamma(\frac{D}{2})} &\leq \frac{\Gamma(\frac{D}{2} + \frac{1}{2} + k)}{\Gamma(\frac{D}{2} + k)} \\
\sqrt{2}\sqrt{v^\top \Sigma v}\, \frac{\Gamma(\frac{D}{2} + \frac{1}{2})}{\Gamma(\frac{D}{2})} \cdot e^{-z} \sum_{k=0}^{\infty} \frac{z^k}{k!} &\leq \sqrt{2}\sqrt{v^\top \Sigma v} \cdot e^{-z} \sum_{k=0}^{\infty} \frac{\Gamma(\frac{D}{2} + \frac{1}{2} + k)}{\Gamma(\frac{D}{2} + k)} \frac{z^k}{k!} \\
\sqrt{2}\frac{\Gamma(\frac{D}{2} + \frac{1}{2})}{\Gamma(\frac{D}{2})}\sqrt{v^\top \Sigma v} &\leq \sqrt{2}\sqrt{v^\top \Sigma v} \cdot e^{-z} \sum_{k=0}^{\infty} \frac{\Gamma(\frac{D}{2} + \frac{1}{2} + k)}{\Gamma(\frac{D}{2} + k)} \frac{z^k}{k!} \\
\sqrt{v^\top \alpha \Sigma v} &\leq F(v).
\end{aligned}$$

2) Let's show that: $\forall v \in \mathcal{T}_x M :\ F(v) \leq \sqrt{v^\top \mathbb{E}[G] v}$.

Wendel (1948) proved: $\frac{\Gamma(x+y)}{\Gamma(x)} \leq x^y$, for $x > 0$ and $y \in [0,1]$. With $x = \frac{D}{2} + k$, $y = \frac{1}{2}$, we obtained $\frac{\Gamma(\frac{D}{2}+\frac{1}{2}+k)}{\Gamma(\frac{D}{2}+k)} \leq \sqrt{\frac{D}{2} + k}$, which leads to: $F(v) \leq \sqrt{2v^\top \Sigma v} \cdot e^{-z} \sum_{k=0}^{\infty} \sqrt{\frac{D}{2} + k} \frac{z^k}{k!}$.

Furthermore, $\sum_{k=0}^{\infty} e^{-z} \frac{z^k}{k!} = 1$ and the function $x \to \sqrt{\frac{D}{2} + x}$ is concave. Then by Jensen's inequality: $e^{-z} \sum_{k=0}^{\infty} \sqrt{\frac{D}{2} + k} \frac{z^k}{k!} \leq \sqrt{\frac{D}{2} + e^{-z} \sum_{k=0}^{\infty} \frac{z^k}{k!} k}$. Knowing that $\sum_{k=0}^{\infty} \frac{z^k}{k!} = z e^z$, we have: $e^{-z} \sum_{k=0}^{\infty} \sqrt{\frac{D}{2} + k} \frac{z^k}{k!} \leq \sqrt{\frac{D}{2} + z}$.
And with $z = \frac{\Omega}{2}$, we obtain: $F(v) \leq \sqrt{v^\top \Sigma (D + \Omega) v}$.

From Kent and Muirhead (1984, p. 442), the expectation of a non-central Wishart distribution ($G \sim \mathcal{W}_q(D, \Sigma, \Omega)$) is: $\mathbb{E}[G] = D\Sigma + \Sigma\Omega$. This finally leads to:

$$F(v) \leq \sqrt{v^\top \mathbb{E}[G] v}.$$

□

Remark B.2.3 As a side note, the second part of the inequality $F(v) \leq \sqrt{v^\top \mathbb{E}[G] v}$ can be obtained using directly Proposition 6.3.3.

Corollary 6.3.2
The length, the energy and the Busemann-Hausdorff volume of the Finsler metric are bounded respectively by the Riemannian length, energy and volume of the covariance

tensor $\alpha\Sigma$ (noted $L_{\alpha\Sigma}, E_{\alpha\Sigma}, V_{\alpha\Sigma}$) and the expected metric $\mathbb{E}[G]$ (noted L_R, E_R, V_R):

$$\begin{aligned} \forall z \in \mathcal{Z},\ L_{\alpha\Sigma}(z) &\leq L_F(z) \leq L_R(z) \\ E_{\alpha\Sigma}(z) &\leq E_F(z) \leq E_R(z) \\ V_{\alpha\Sigma}(z) &\leq V_F(z) \leq V_R(z) \end{aligned}$$

Proof. From Proposition 6.3.1, we have $\forall (x, v) \in \mathcal{M} \times \mathcal{T}_x M :\ \sqrt{h(v)} \leq F(v) \leq \sqrt{g(v)}$, with $h : v \to v^\top \alpha\Sigma v$ and $g : v \to v^\top \mathbb{E}[G] v$ Riemannian metrics. We also define the parametric curve: $\forall t \in \mathbb{R}, \gamma(t) = x$ and $\dot{\gamma}(t) = v$.

1) Let's show that $L_\Sigma(x) \leq L_F(x) \leq L_R(x)$. Because of the monotonicity of the Lebesgue integrals, we directly have: $\int \sqrt{h(\dot{\gamma}(t))}dt \leq \int F(\dot{\gamma}(t))dt \leq \int \sqrt{g(\dot{\gamma}(t))}dt$.

2) Let's show that $E_\Sigma(x) \leq E_F(x) \leq E_R(x)$. Since all the functions are positive, we can raise them to the power two, and again, with the monotonicity of the Lebesgue integrals, we have: $\int h(\dot{\gamma}(t))dt \leq \int F^2(\gamma(t), \dot{\gamma}(t))dt \leq \int \sqrt{g(\dot{\gamma}(t))}dt$.

3) Let's show that $V_\Sigma(x) \leq V_F(x) \leq V_R(x)$. We write the vectors $v \in \mathcal{T}_x M$ in hyperspherical coordinates: $v = re$, with $r = \|v\|$ the radial distance and e the angular coordinates. With $v = re$, we have: $r \cdot \sqrt{h(e)} \leq r \cdot F(e) \leq r \cdot g(e) \iff \sqrt{h(e)}^{-1} \geq F(e)^{-1} \geq \sqrt{g(e)}^{-1}$.

We want to identify an inequality between the indicatrices, noted $\text{vol}(I_h)$, $\text{vol}(I_g)$, $\text{vol}(I_F)$, formed by the functions h, g and F. Let's define: $r_g\sqrt{h(e)} = r_h\sqrt{g(e)} = r_F F(e) = 1$. For every angular coordinate e, we obtain: $r_h \geq r_F \geq r_g$. Intuitively, this means that the Finsler indicatrix will always be bounded by the indicatrices formed by h and g. The Busemann-Hausdorff volume of a function f is defined as: $\sigma_B(f) = \text{vol}(\mathbb{B}^n(1))/\text{vol}(I_f)$, with $\text{vol}(\mathbb{B}^n(1))$ the volume of the unit ball and $\text{vol}(I_f)$ the volume of the indicatrix formed by f. The previous inequality and the definition of the Busemann-Hausdorff volume implies that: $\text{vol}(I_h) \geq \text{vol}(I_F) \geq \text{vol}(I_g) \Rightarrow \sigma_B(h) \leq \sigma_B(F) \leq \sigma_B(g)$. The functions g and h being Riemannian, we have: $\sigma_B(h) = \sqrt{det(\alpha\Sigma)}$ and $\sigma_B(g) = \sqrt{det(\mathbb{E}[G])}$, which concludes the proof.

□

Relative bounds between the Finsler and the Riemannian metric

Proposition 6.3.3
Let f be a stochastic immersion. f induces the stochastic norm $\|\cdot\|_G$, defined in Section 6.2. The relative difference between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$0 \leq \frac{\|v\|_R - \|v\|_F}{\|v\|_R} \leq \frac{\text{Var}\left[\|v\|_G^2\right]}{2\mathbb{E}\left[\|v\|_G^2\right]^2}.$$

Proof. We will directly use a sharper version of Jensen's inequality obtained by Liao and Berg (2019): Let X be a one-dimensional random variable with mean μ and $P(X \in (a, b)) = 1$, where $-\infty \leq a \leq b \leq +\infty$. Let ϕ a twice derivable function on (a, b). We further define: $h(x, \mu) = \frac{\phi(x) - \phi(\mu)}{(x-\mu)^2} - \frac{\phi'(\mu)}{x-\mu}$. Then:

$$\inf_{x \in (a,b)} \{h(x, \mu)\} \text{Var}[X] \leq \mathbb{E}[\phi(x)] - \phi(\mathbb{E}[x]) \leq \sup_{x \in (a,b)} \{h(x, \mu)\} \text{Var}[X].$$

In our case, we will choose $\phi : z \to \sqrt{z}$ with z a one-dimensional random variable defined as $z = v^\top G v$. $a = 0$, $b = +\infty$ and $\mu = \mathbb{E}[z]$. $h(z, \mu) = (\sqrt{z} - \sqrt{\mu})(z-\mu)^{-2} - (2(z-\mu)\sqrt{\mu})^{-1}$. Because its first derivative ϕ' is convex, the function $x \to h(x, \mu)$ is monotonically increasing. Thus:

$$\inf_{z\in(0,+\infty)}\{h(x,\mu)\} = \lim_{z\to 0} = -\frac{\sqrt{\mu}}{2\mu^2} \quad \text{and} \quad \sup_{z\in(0,+\infty)}\{h(x,\mu)\} = \lim_{z\to+\infty} = 0.$$

It finally gives:

$$-\frac{\sqrt{\mu}}{2\mu^2}\mathrm{Var}[z] \leq \mathbb{E}[\sqrt{z}] - \sqrt{\mathbb{E}[z]} \leq 0.$$

Replacing $\|v\|_F \coloneqq F(v) = \mathbb{E}[\sqrt{z}]$ and $\|v\|_R \coloneqq \sqrt{g(v)} = \sqrt{\mathbb{E}[z]} = \sqrt{\mu}$ concludes the proof. □

Lemma B.2.6
Let $z \sim \mathcal{W}_1(D, \sigma, \Omega)$ following a one-dimensional non-central Wishart distribution. Then:

$$\frac{\mathrm{Var}[z]}{2\mathbb{E}[z]^2} = \frac{1}{D+\Omega} + \frac{\Omega}{(D+\Omega)^2}$$

Proof. Kent and Muirhead (1984, Theorem 10.3.7.) states that if $z \sim \mathcal{W}_1(D, \sigma, \omega)$ then $\mathbb{E}[z^k] = \sigma^k 2^k \frac{\Gamma(\frac{D}{2}+k)}{\Gamma(\frac{D}{2})} {}_1F_1(-k, \frac{D}{2}, -\frac{1}{2}\Omega)$. In particular, for $k = 1$ and $k = 2$, we have ${}_1F_1(-1, b, c) = 1 - \frac{c}{b}$ and ${}_1F_1(-2, b, c) = 1 - \frac{2c}{b} + \frac{c^2}{b(b+1)}$. We also have $\frac{\Gamma(\frac{D}{2}+1)}{\Gamma(\frac{D}{2})} = \frac{D}{2}$ and $\frac{\Gamma(\frac{D}{2}+2)}{\Gamma(\frac{D}{2})} = \frac{D}{2}\left(\frac{D}{2}+1\right)$, which leads to: $\mathbb{E}[z] = \sigma(D+\Omega)$ and $\mathbb{E}[z^2] = \sigma^2(2\omega + 2(D+\omega) + (D+\omega)^2)$. Finally, we conclude:

$$\frac{\mathrm{Var}[z]}{\mathbb{E}[z]^2} = \frac{\mathbb{E}[z^2]}{\mathbb{E}[z]^2} - 1 = \frac{2\omega}{(D+\omega)^2} + \frac{2}{D+\omega}.$$

□

Proposition 6.3.4
Let f be a Gaussian process. We note $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J] v)$, with J the Jacobian of f, and Σ the covariance matrix of J.
The relative ratio between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$0 \leq \frac{\|v\|_R - \|v\|_F}{\|v\|_R} \leq \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2}.$$

Proof. The result is directly obtained using Proposition 6.3.3 and Lemma B.2.6. □

Corollary 6.3.5
When f is a Gaussian Process, the relative ratio between the length, the energy and the volume of the Finsler norm (noted L_F, E_F, V_F) and the Riemannian norm (noted

$L_R, E_R, V_R)$ is:

$$
\begin{aligned}
0 \leq & \frac{L_R(z) - L_F(z)}{L_R(z)} \leq \max_{v \in \mathcal{T}_z \mathcal{Z}} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\} \\
0 \leq & \frac{E_R(z) - E_F(z)}{E_R(z)} \leq \max_{v \in \mathcal{T}_z \mathcal{Z}} \left\{ \frac{2}{D+\omega} + \frac{1+2\omega}{(D+\omega)^2} + \frac{2\omega}{(D+\omega)^3} + \frac{\omega^2}{(D+\omega)^4} \right\} \\
0 \leq & \frac{V_R(z) - V_F(z)}{V_R(z)} \leq 1 - \left(1 - \max_{v \in \mathcal{T}_z \mathcal{Z}} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\} \right)^q
\end{aligned}
$$

Proof. Let's call $M = \max_{v \in \mathcal{T}_x M} \{ \frac{\omega}{(D+\omega)^2} + \frac{1}{D+\omega} \}$.

From Proposition 6.3.4, we have:

$$
0 \leq \|v\|_R - \|v\|_F \leq M \ \|v\|_R \,, \quad \text{with} \quad M = \max_{v \in \mathcal{T}_x M} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\}
$$

1) By the monotonicity of the Lebesgue integral, we can directly integrate the previous norms along a curve γ, which immediately leads to: $0 \leq L_R(x) - L_F(x) \leq M L_R(x)$.

2) Since all the functions are positive: $0 \leq \|v\|_F \leq \|v\|_R \leq M \|v\|_R + \|v\|_F$ leads to: $\|v\|_F^2 \leq \|v\|_R^2 \leq M^2 \|v\|_R^2 + 2M \|v\|_F \|v\|_R + \|v\|_F^2$, and replacing $\|v\|_F \leq \|v\|_R$ in the right-hand term: $\|v\|_F^2 \leq \|v\|_R^2 \leq (M^2 + 2M) \|v\|_R^2 + \|v\|_F^2$, and finally: $0 \leq \|v\|_R^2 - \|v\|_F^2 \leq (M^2 + 2M) \|v\|_R^2$. Again, by continuity of the Lebesgue integral, we directly obtain: $0 \leq E_R(x) - E_F(x) \leq (M^2 + 2M) E_R(x)$.

3) In order to compare the volume between the Finsler and the Riemannian metric, we need to compare the volume of their indicatrices, noted: $\text{vol}(I_g)$ and $\text{vol}(I_F)$ respectively. We write the vectors $v \in \mathcal{T}_x M$ in hypershperical coordinates, with $v = re$, $r = \|v\|$ the radial distance and e the angular coordinates. The volume of the indicatrices obtained in dimension q (dimension of the latent space) can be written as: $d^q V = r^{q-1} dr d\Phi$, with Φ defining the different angles. We will note r_F and r_g the radial distances of the Finsler and Riemann metrics such that: $\|v\|_F = r_F \|e\|_F = 1$ and $\|v\|_R = r_g \|e\|_R = 1$ obtained for a specific angle e.

$$
\begin{aligned}
\text{vol}(I_F) - \text{vol}(I_g) &= \int_\Phi \left(\int_0^{r_f} r^{q-1} dr - \int_0^{r_g} r^{q-1} dr \right) d\Phi \\
&= \int_\Phi \frac{r_f^q}{q} \left(1 - \left(\frac{r_g}{r_f} \right)^q \right) d\Phi \\
&\leq \int_\Phi \frac{r_f^q}{q} d\Phi \cdot \left(1 - \left(\frac{r_g}{r_f} \right)^q \right)
\end{aligned}
$$

and by definition: $\text{vol}(I_F) = \int_\Phi (r_f^q / q) d\Phi$. Furthermore, for a specific angle e, we have: $r_g / r_F = \sqrt{g(e)} / F(e) \geq 1 - M$, from Proposition 6.3.4. We have:

$$
0 \leq \frac{\text{vol}(I_F) - \text{vol}(I_g)}{\text{vol}(I_F)} \leq \cdot 1 - \left(\frac{r_g}{r_f} \right)^q \leq 1 - (1 - M)^q \,,
$$

and by the definition of the Busemann Hausdorff volume: $\frac{V_F(x) - V_G(x)}{V_F(x)} = \frac{\text{vol}(I_F) - \text{vol}(I_g)}{\text{vol}(I_F)}$, we conclude the proof.

□

Implications in High Dimensions

In this section, we want to show that the difference between the Finsler norm and the Riemannian induced norm, as well as their respective functionals, tend to zero at a rate of $\mathcal{O}(\frac{1}{D})$. We need to be sure that ω doesn't grow faster than D, in other terms: $\omega = \mathcal{O}(D)$. This can be obtained if we assume that every element of the expectation of Jacobian is upper bounded ($\exists m \in \mathbb{R}_+^*, \forall i, j\ \mathbb{E}[J_{ij}] \leq m$). This happens in at least two cases: (1) $\mathbb{E}[f]$ is somehow Lipschitz continuous; or (2) if f is a Gaussian Process and its covariance is upper bounded. The latter case happens when the process is defined over a bounded domain.

Lemma B.2.7
Our Finsler metric $v \to \mathbb{E}[\sqrt{v^\top G v}]$ is defined with $v^\top G v \sim \mathcal{W}_1(D, v^\top \Sigma v, \omega)$, and $\omega = (v^\top \Sigma v)^{-1}(v^\top \mathbb{E}[J]^\top \mathbb{E}[J] v)$.

If the Finsler manifold is bounded, then: $\omega \leq DM$, with $M \in \mathbb{R}_+$.

Proof. By definition, Σ does not depend on D. We assume the manifold is bounded, which means that every element of the expected Jacobian is upper bounded: $\mathbb{E}[J]_{ij} \leq m$, with $m \in \mathbb{R}_+^*$. We call $\sigma = v^\top \Sigma v \in \mathbb{R}_+^*$.

We have:

$$\omega = \sigma^{-1} \sum_{k=1}^{D} \sum_{i=1}^{q} \sum_{j=1}^{q} v_i \mathbb{E}[J]_{ki} \mathbb{E}[J]_{kj} v_j \leq \sigma^{-1} \sum_{k=1}^{D} m^2 \left\|v\right\|^2 \leq DM,$$

with $M = \sigma^{-1} m^2 \left\|v\right\|^2 \in \mathbb{R}_+^*$, and M does not depend on D. □

Corollary 6.3.6
Let f be a Gaussian Process. In high dimensions, we have:

$$\frac{L_R(z) - L_F(z)}{L_R(z)} = \mathcal{O}\left(\frac{1}{D}\right), \quad \frac{E_R(z) - E_F(z)}{E_R(z)} = \mathcal{O}\left(\frac{1}{D}\right),$$
$$\text{and} \quad \frac{V_R(z) - V_F(z)}{V_R(z)} = \mathcal{O}\left(\frac{q}{D}\right).$$

When D converges toward infinity: $L_R \underset{+\infty}{\sim} L_F$, $E_R \underset{+\infty}{\sim} E_F$ and $V_R \underset{+\infty}{\sim} V_F$.

Proof. From Corollary 6.3.5, we directly obtained the results in high dimensions.

We assume that our latent space is bounded, then by B.2.7, we have: $0 \leq \omega \leq MD$, with $M \in \mathbb{R}+$.

For the length, we have:

$$\begin{aligned}\frac{L_G(x) - L_F(x)}{L_G(x)} &\leq \max_{v \in \mathcal{T}_x M} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\} \\ &\leq \frac{1+M}{D}\end{aligned}$$

For the energy functional, we have:

$$\begin{aligned}\frac{E_G(x) - E_F(x)}{E_G(x)} &\leq \max_{v \in \mathcal{T}_x M} \left\{ \frac{2}{D+\omega} + \frac{1+2\omega}{(D+\omega)^2} + \frac{2\omega}{(D+\omega)^3} + \frac{\omega^2}{(D+\omega)^4} \right\} \\ &\leq \frac{2+2M}{D} + \frac{1+2M+M^2}{D^2} \\ \limsup_{D\to\infty} D \times \frac{E_G(x) - E_F(x)}{E_G(x)} &\leq \limsup_{D\to\infty} 2(1+M) + \frac{M^2+2M+1}{D^2} \to 2(1+M)\end{aligned}$$

For the volume, we have:

$$\begin{aligned}\frac{V_G(x) - V_F(x)}{V_G(x)} &\leq 1 - \left(1 - \max_{v \in \mathcal{T}_x M} \left\{ \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \right\}\right)^q \\ &\leq 1 - \left(1 - \frac{1+M}{D}\right)^q\end{aligned}$$

Using Taylor series expansion, when $x \sim 0$, we have: $1 - (1-x)^q = qx + o(x^2)$. Let's call $\varepsilon \ll 1$, and rewrite the Taylor series:

$$\begin{aligned}\frac{V_G(x) - V_F(x)}{V_G(x)} &\leq q\frac{1+M}{D} + \varepsilon q \frac{1+M}{D} \\ \limsup_{D\to\infty} \frac{D}{q} \times \frac{V_G(x) - V_F(x)}{V_G(x)} &\leq (1+M)(1+\varepsilon)\end{aligned}$$

The difference between the functionals can converge to zero if they are similar in high dimensions, or if they all diverge to infinity. This latter case does not happen as we assume the latent manifold being bounded, and so the metrics are then finite, which concludes the proof.

□

Corollary 6.3.7
Let f be a Gaussian Process. In high dimensions, the relative ratio between the Finsler norm $\|\cdot\|_F$ and the Riemannian norm $\|\cdot\|_R$ is:

$$\frac{\|v\|_R - \|v\|_F}{\|v\|_R} = \mathcal{O}\left(\frac{1}{D}\right)$$

And, when D converges toward infinity: $\forall v \in \mathcal{T}_z\mathcal{Z},\ \|v\|_R \underset{+\infty}{\sim} \|v\|_F$.

Proof. Similar to the 6.3.6, assuming that our latent space is bounded, from B.2.7, we have $0 \leq \omega \leq MD$. From 6.3.4, we deduce:

$$\begin{aligned}0 \leq \frac{\|v\|_R - \|v\|_F}{\|v\|_R} &\leq \frac{1}{D+\omega} + \frac{\omega}{(D+\omega)^2} \\ &\leq \frac{1+M}{D}\end{aligned}$$

In a bounded manifold, the metric are finite. We can deduce that they converge to each other in high dimensions. □

B.3 Computations

B.3.1 Computing geodesics with Stochman and minimizing the curve energy functionals

An essential task is to compute the shortest paths, or geodesics, between data points in the latent space. Those shortest paths can be obtained in two ways: either by solving a corresponding system of ODEs, or by minimizing the curve energy in the latent space. The former being computationally expensive, we favor the second approach which consists in optimizing a parametrized spline on the manifold. This method is already implemented in Stochman Detlefsen et al. (2021), where we can easily optimize splines by redefining the curve energy function of a manifold class.

We need two curve-energy functionals: one for the expected Riemannian metric and one for the Finsler metric.

Curve energy for the Riemannian metric

We know that the stochastic metric tensor G_t defined on a point t follows a non-central Wishart distribution. Thus, we can compute its expectation $\mathbb{E}[G_t]$ knowing the Jacobian covariance and expectation: $\mathbb{E}[J_t]$ and Σ. The next Section B.3.2 explains how to compute those quantities.

Assuming the spline is discretized into N points, we can compute the curve energy with:

$$E_G(\gamma(t)) = \int_0^1 \dot{\gamma}(t)^\top \mathbb{E}[G_t]\dot{\gamma}(t)\, dt \approx \sum_{i=1}^{N} \dot{\gamma}_i^\top \left(\mathbb{E}[J_i]^T \mathbb{E}[J_i] + D\Sigma_i\right) \dot{\gamma}_i.$$

Curve energy for the Finsler metric

In order to compute of the curve energy $\mathcal{E}(\gamma)$, we must first derive the expectation $\mathbb{E}[J_t]$ and covariance Σ of the Jacobian of f, which should follow a normal distribution: $J_i = \partial f_i \sim \mathcal{N}(\mathbb{E}[J], \Sigma)$. We assume the points ∂f_i are independent samples with the same variance drawn from a normal distribution. We can then compute the Finsler metric which follows a non-central Nakagami distribution (See Proposition 6.2.2):

$$\mathcal{E}(\gamma(t)) = \int_0^1 F(t, \dot{\gamma}(t))^2\, dt \approx \sum_{i=1}^{N} 2\dot{\gamma}_i^\top \Sigma_i \dot{\gamma}_i \left(\frac{\Gamma(\frac{D}{2}+\frac{1}{2})}{\Gamma(\frac{D}{2})}\right)^2 {}_1F_1\left(-\frac{1}{2}, \frac{D}{2}, -\frac{\omega_i}{2}\right)^2,$$

with ${}_1F_1$ the confluent hypergeometric function of the first kind and $\omega_i = (\dot{\gamma}_i^\top \Sigma_i \dot{\gamma}_i)^{-1}(\dot{\gamma}_i^\top \Omega_i \dot{\gamma}_i)$ and $\Omega_i = \Sigma_i^{-1}\mathbb{E}[J_i]^\top \mathbb{E}[J_i]$.

This function has been implemented in PyTorch using the known gradients for the hypergeometric function: $\frac{\partial}{\partial x}{}_1F_1(a,b,x) = \frac{a}{b}{}_1F_1(a+1, b+1, x)$.

B.3.2 Accessing the posterior derivatives

We assume that the probabilistic mapping f from the latent variables X to the observational variables Y follows a normal distribution. We would like to obtain the posterior kernel Σ_* and expectation μ_* such that $p(\partial_t f | Y, X) \sim \mathcal{N}(\mu_*, \Sigma_*)$.

Furthermore, we make the hypothesis the observed variables are modelled with a Gaussian noise ϵ whose variance is the same in every dimension. In particular, for the n^{th} latent (x) and observed (y) variable in the j^{th} dimension: $y_{n,j} = f_j(x_{n,:}) + \epsilon_n$. Thus, the output variables have the same variance, and the posterior kernel Σ_* is then isotropic with respect to the output dimensions: $\Sigma_* = \sigma_*^2 \cdot I_D$.

There are two ways of obtaining the posterior variance and expectation:

- We use the Gaussian processes to predict the derivative $(\partial_c f)$ of the mapping function f, and we multiply the obtained posterior kernel by the curve derivative $(\partial_t c)$, following the chain rule: $\frac{df(c(t))}{dt} = \frac{df}{dc} \cdot \frac{dc}{dt}$ (Section: B.3.2)
- We discretize the derivative of the mapping function as the difference of this function evaluated at two close points. We use a linear operation to obtain the posterior variance and expectation: $\partial_t f(c(t)) \sim f(c(t_{i+1})) - f(c(t_i))$. (Section: B.3.2)

Closed-form expressions

We assume that f is a Gaussian process. Hence, because the differentiation is a linear operation, the derivative of a Gaussian Process is also a Gaussian Process Rasmussen and Williams (2005).

The data $Y \in \mathcal{R}^{N \times D}$ follows a normal distribution, so we can infer the partial derivative of one data point $(J^T)_{ji} = \frac{\partial y_i}{\partial x_j}$, with $i = 1 \dots D$ and $j = 1 \dots d$. We have:

$$\begin{bmatrix} Y \\ (J)^\top \end{bmatrix} = \prod_{i=1}^{D} \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \mu_{\partial y} \end{bmatrix}, \begin{bmatrix} K(x,x) & \partial K(x,x_*) \\ \partial K(x_*,x) & \partial^2 K(x_*,x_*) \end{bmatrix} \right),$$

and $J^\top$ can be predicted:

$$p(J^\top | Y, X) = \prod_{i=1}^{D} \mathcal{N}\left(\mu_*, \Sigma_*\right),$$

with:

$$\begin{aligned} \mu_* &= \partial K(x_*,x) \cdot K(x,x)^{-1} \cdot (y - \mu_y) + \mu_{\partial y} \\ \Sigma_* &= \partial^2 K(x_*,x_*) - \partial K(x_*,x) \cdot K(x,x)^{-1} \cdot \partial K(x,x_*). \end{aligned}$$

Finally, $\partial_t f$ is obtained:

$$p(\partial_t f(c(t)) | f(x), x) = \prod_{i=1}^{D} \mathcal{N}\left(\dot{c}\mu_*, \dot{c}^\top \Sigma_* \dot{c} \cdot I_D\right).$$

Discretization

One can notice that: $\partial_t f(c(t)) \sim f(c(t_{i+1})) - f(c(t_i))$. We know that $f(c(t_{i+1}))$ and $f(c(t_i))$ both follows a normal distribution.

$$\begin{bmatrix} f(c(t_i)) \\ f(c(t_{i+1})) \end{bmatrix} = \prod_{j=1}^{D} \mathcal{N}\left(\begin{bmatrix} \mu_i \\ \mu_{i+1} \end{bmatrix}, \begin{bmatrix} \sigma_{ii}^2 & \sigma_{i,i+1}^2 \\ \sigma_{i+1,i}^2 & \sigma_{i+1,i+1}^2 \end{bmatrix} \right).$$

If $Y = AX$ affine transformation of a multivariate Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$, then Y is also a multivariate Gaussian with: $Y \sim \mathcal{N}(A\mu, A^T\sigma^2 A)$. In our case, we choose $A^T = [-1, 1]$. We have:

$$f(c(t_{i+1})) - f(c(t_i)) \sim \mathcal{N}(\mu_*, \sigma_*^2 \cdot I_D),$$

with:

$$\mu_* = \mu_{i+1} - \mu i$$
$$\sigma_*^2 = \sigma_{ii}^2 + \sigma_{i+1,i+1}^2 - 2\sigma_{i,i+1}^2$$

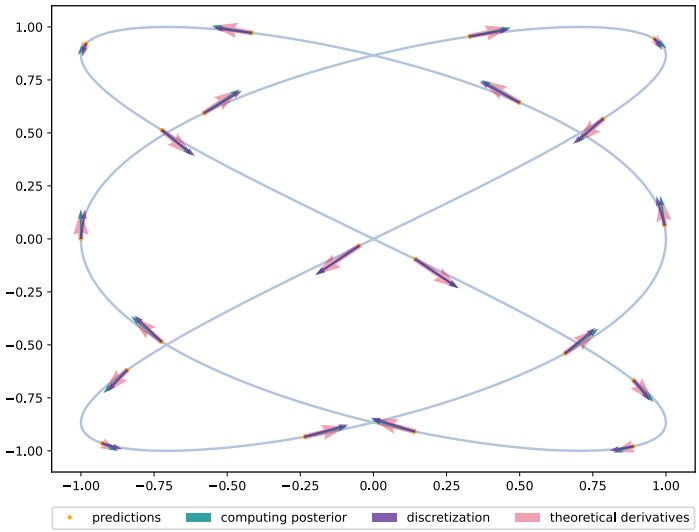


Figure B.5: Illustration of the derivatives obtained with a trained GP on a simple parametrized function: both methods give the correct derivatives if enough points are sampled.

B.4 Experiments

B.4.1 Datasets

Font data

The dataset represents 46 different fonts for each letter (upper and lower case) whose contour is parametrized by a spline (or two splines, depending on the letter used) obtained from at least 500 points Campbell and Kautz (2014).

In our case, we choose to learn the manifold of the letter **f**. The dataset is composed of 46 different fonts, each letter being drawn by 1024 points. We reduce this number from 1024 to 256 by sampling one point every 4. The dimension of the observational space is then 256.

qPCR

The qPCR data, gathered from Guo et al. (2010), was used to illustrate the training of a GPLVM in Pyro Pyro (2022) and is available at the Open Data Science repository Ahmed et al. (2019). It consists of 437 single-cell qPCR data for which the expression of 48 genes has been measured during 10 different cell stages. We then have 437 data

points, 48 observations, and 10 classes. Before training the GP-LVM, the data is grouped by the capture time, as illustrated in the Pyro documentation.

Pinwheel on a sphere

A pinwheel in 2-dimension is created and then projected onto a sphere using a stereographic projection method. The final dataset is composed of 1000 points with their coordinates in 3-dimensions.

B.4.2 GP-LVM training

We learn our two-dimensional latent space by training a GP-LVM Lawrence (2003) with Pyro Bingham et al. (2019). The Gaussian Process used is a Sparse GP, defined with a kernel (RBF, or Matern) composed of a one-dimensional length scale and variance. The parameters are learned with the Adam optimizer Kingma and Ba (2014). The number of steps and the initialization of the latent space vary with the dataset.

Table B.2: Description of the datasets trained with a GP-LVM.

Datasets	pinwheel	font data	qPCR
Number of data points	500	46	437
Number of observations	3	256	48
initialization	PCA	PCA	custom
kernel	RBF	Matern52	Matern52
steps	17000	5000	5000
learning rate	1e-3	1e-4	1e-4
lengthscale	0.24	0.88	0.15
variance	0.95	0.30	0.75
noise	1e-4	1e-3	1e-3

B.4.3 Computing indicatrices

An indicatrix of a function g at a point x is defined such that: $v \in \mathcal{T}_x M | g_x(v) < 1$. In other terms, the indicatrix is the representation of a unit ball in our latent space. If we use a Euclidean metric, our indicatrix in our 2-dimensional latent space would be a unit ball, as we need to solve: $v \in \mathcal{T}_x M$, $\|v\| < 1$. For a Riemannian metric, our indicatrix is necessarily an ellipse, whose semi axis are the square-roots of the eigenvalues of the metric tensor G:: $v \in \mathcal{T}_x M$, $v^\top G v < 1$. For our Finsler metric, we don't have an analytical solution, and so it's difficult to predict the shape of the convex polygon.

In this paper, the indicatrices are drawn the following way: for a single point in our latent space, we compute the value of $v^\top G v$ and $F(x, v)$ for v varying over the space. We then extract the contour when $v^\top G v$ and $F(x, v)$ are equal to 1. Computing the area of the indicatrices will be used in the section B.4.4 to compute the volume measures.

B.4.4 Computing the volume forms

For the figures used in this paper, by default, the background of the latent space represents the volume measure of the expected Riemannian metric ($V_G = \sqrt{\mathbb{E}[G]}$) on a logarithm scale. In figure 6.3, the volume measure of the Finsler metric is also computed.

Finsler metric

To compute the volume measure of our Finsler metric, we choose the Busemann-Hausdorff definition, which is the ratio of a unit ball over the volume of its indicatrix: $\mathcal{V} = \text{vol}(\mathbb{B}^n(1))/\text{vol}(\{v \in \mathcal{T}_x M | F(x,v) < 1\})$. While our Finsler function has an analytical form, its expression doesn't allow to directly solve the equation: $v \in \mathcal{T}_x M, F(x,v) < 1$. Instead, we approximate its indicatrix as describe in B.4.3, using a contour plot and extracting the paths vertices. We can then compute the area of the obtained polygon, and divide with the volume of a unit ball: $\text{vol}(\mathbb{B}^2(1)) = \pi$.

The volume measure can then be computed for each point over a grid (32 x 32, in figure 6.3), and we interpolate all the other points. Note that this method can only be used when our latent space is of dimension 2.

Expected Riemannian metric

There are two ways to compute the volume measure of the expected Riemannian metric. One way is to directly use the metric tensor: $V_G = \sqrt{\mathbb{E}[G]}$. Another one is to remember that any Riemannian metric is a Finsler metric, and thus, the Busemann-Hausdorff definition also applied for our metric: $V_G = \text{vol}(\mathbb{B}^n(1))/\text{vol}(\{v \in \mathcal{T}_x M | v^\top \mathbb{E}[G] v < 1\})$. Solving $v^\top \mathbb{E}[G] v < 1$ for $v \in \mathcal{T}_x M$ is equivalent to solving the area of an ellipse.

For the first method, we can either sample multiple times the metric, which is computationally expensive, or use the fact that our metric tensor is a non-central Wishart matrix: $G = J^\top J \sim \mathcal{W}_q(D, \Sigma, \Sigma^{-1}\mathbb{E}[J]^\top \mathbb{E}[J])$, with Σ the covariance of the Jacobian J and D the dimension of the observational space. In this case, its expectation is: $\mathbb{E}[G] = \mathbb{E}[J]^\top \mathbb{E}[J] + D\Sigma$. We can access the derivatives of the function f (detailed in section B.3.2), and compute both quantities $\mathbb{E}[J]$ and Σ needed to estimate the expected metric and its determinant.

For the second method, we can compute the area of the ellipse in the same way we compute the Finsler volume measure.

B.4.5 Experiments when increasing the number of dimensions

In Figure 6.4, we computed the volume ratio and draw indicatrices while varying the number of dimensions, to illustrate that both our Finsler metric and the expected Riemannian metric seem to converge when D increases.

The main issue with this experiment is to vary only one factor, the number of dimensions D, while keeping the other factors unchanged. This is difficult for two reasons: 1) Even with a very low dimensional observational, both metrics are already very similar to each other. It would be difficult to illustrate a convergence while increasing the number of dimensions. 2) the function f needs to be learned again each time we increase the number of dimensions of the observational space, and the parameters of the Gaussian Process will change too.

Instead, we try to illustrate our results by computing empirically the stochastic metric tensor $G = J^\top J$, using its Jacobian $J \sim \mathcal{N}(\mathbb{E}[J], \Sigma)$, a $D \times q$ matrix. The number of dimensions is modified by simply truncatenating the Jacobian J. In Figure 6.4, the volume ratio is computed for 12 Jacobians obtained with different random parameters $\mathbb{E}[J]$ and Σ. The Finsler and Riemannian indicatrices (lower right) are drawn for only one Jacobian selected randomly.

C Pulling back information geometry

C.1 Additional details for information geometry

In this section we provide additional information regarding information geometry. We note that many of these propositions are already known in the literature, however, we include them for completion and for the paper to be standalone.

The Fisher-Rao metric is positive definite only if it is non-singular, and then, defines a Riemannian metric (Nielsen 2020). In this paper, we assume that the observation $\mathbf{x} \in \mathcal{X}$ is a random variable following a probability distribution $p(\mathbf{x})$ such that $\mathbf{x} \sim p(\mathbf{x}|\eta)$, and any smooth changes of the parameter η would alter the observation $\mathbf{x}$. This way, the Fisher-Rao metric used in our paper is non-singular and the statistical manifold $\mathcal{H}$ is a Riemannian manifold.

A known result in *information geometry* (Nielsen 2020; Amari and Nagaoka 2000) is that the Fisher-Rao metric is the first order approximation of the KL-divergence, as recall in Proposition C.1.1. Using this fact, we can define the Fisher-Rao distance and energy in function of the KL-divergence, leading to Proposition C.1.2.

Proposition C.1.1
The Fisher-Rao metric is the first order approximation of the KL-divergence between perturbed distributions:

$$\mathrm{KL}(p(\mathbf{x}|\eta),\ p(\mathbf{x}|\eta+\delta\eta)) = \frac{1}{2}\delta\eta^\top \mathbf{I}_{\mathcal{H}}(\eta)\delta\eta + o(\delta\eta^2),$$

with $\mathbf{I}_{\mathcal{H}}(\eta) = \int p(\mathbf{x}|\eta)\left[\nabla_\eta \log p(\mathbf{x}|\eta) \nabla_\eta \log p(\mathbf{x}|\eta)^\top\right] d\mathbf{x}$.

Proof. Let's decompose $\log p(\mathbf{x}|\eta+\delta\eta)$ using the Taylor expansion:

$$\log p(\mathbf{x}|\eta+\delta\eta) = \log p(\mathbf{x}|\eta) + \nabla_\eta \log p(\mathbf{x}|\eta)^\top \delta\eta + \frac{1}{2}\delta\eta^\top \mathbf{H}_\eta\left[\log p(\mathbf{x}|\eta)\right]\delta\eta + o(\delta\eta^2),$$

where the Hessian is $\mathbf{H}_\eta\left[\log p(\mathbf{x}|\eta)\right] = \frac{\mathbf{H}_\eta[p(\mathbf{x}|\eta)]}{p(\mathbf{x}|\eta)} - \nabla_\eta \log p(\mathbf{x}|\eta)\nabla_\eta \log p(\mathbf{x}|\eta)^\top$ and the $\nabla_\eta \log p(\mathbf{x}|\eta) = \frac{\nabla_\eta p(\mathbf{x}|\eta)}{p(\mathbf{x}|\eta)}$.

Also, $\int \nabla_\eta p(\mathbf{x}|\eta) d\mathbf{x} = \nabla_\eta \int p(\mathbf{x}|\eta) d\mathbf{x} = 0$ and $\int \mathbf{H}_\eta[p(\mathbf{x}|\eta)] d\mathbf{x} = \mathbf{H}_\eta[\int p(\mathbf{x}|\eta) d\mathbf{x}] = 0$.

Replacing all those expressions to the first equation finally gives:

$$\begin{aligned}
\mathrm{KL}(p(\mathbf{x}|\eta),\, p(\mathbf{x}|\eta+\delta\eta)) &= \int p(\mathbf{x}|\eta)\log p(\mathbf{x}|\eta) d\mathbf{x} - \int p(\mathbf{x}|\eta)\log p(\mathbf{x}|\eta+\delta\eta) d\mathbf{x} \\
&= -\int p(\mathbf{x}|\eta)\left(\nabla_\eta \log p(\mathbf{x}|\eta)^\top \delta\eta + \frac{1}{2}\delta\eta^\top \mathbf{H}_\eta[\log p(\mathbf{x}|\eta)]\delta\eta + o(\delta\eta^2)\right) d\mathbf{x} \\
&= \frac{1}{2}\delta\eta^\top \left[\int p(\mathbf{x}|\eta)\left[\nabla_\eta \log p(\mathbf{x}|\eta)\nabla_\eta \log p(\mathbf{x}|\eta)^\top\right] d\mathbf{x}\right]\delta\eta + o(\delta\eta^2).
\end{aligned}$$

□

Definition C.1.1
We consider a curve $\gamma(t)$ and its derivative $\dot{\gamma}(t)$ on the statistical manifold such that, $\forall t \in [0,1], \gamma(t) = \eta_t \in \mathcal{H}$. The manifold is equipped with the Fisher-Rao metric. The

length and the energy functionals are defined with respect to the metric $\mathbf{I}_{\mathcal{H}}(\eta)$:

$$\text{Length}(\gamma) = \int_0^1 \sqrt{\dot\gamma(t)^\top \mathbf{I}_{\mathcal{H}}(\eta)\dot\gamma(t)}dt \quad \text{and} \quad \text{Energy}(\gamma) = \int_0^1 \dot\gamma(t)^\top \mathbf{I}_{\mathcal{H}}(\eta)\dot\gamma(t)dt.$$

Locally length-minimizing curves between two connecting points are called geodesics. These can be found by minimizing the energy using the Euler-Lagrange equations which gives the following system of 2nd order nonlinear ordinary differential equations (ODEs) (Arvanitidis et al. 2018)

$$\ddot\gamma(t) = -\frac{1}{2}\mathbf{I}_{\mathcal{H}}^{-1}(\gamma(t))\Big[2(\dot\gamma(t)^\top \otimes \mathbf{I}_d)\frac{\partial \text{vec}[\mathbf{I}_{\mathcal{H}}(\gamma(t))]}{\partial\gamma(t)}\dot\gamma(t) - \frac{\partial \text{vec}[\mathbf{I}_{\mathcal{H}}(\gamma(t))]}{\partial\gamma(t)}^\top (\dot\gamma(t)\otimes\dot\gamma(t))\Big]. \tag{C.1}$$

Proposition C.1.2
The KL-divergence between two close elements of the curve γ is defined as: $\text{KL}(p_t,\ p_{t+\delta t}) = \text{KL}(p(\mathbf{x}|\gamma(t)),\ p(\mathbf{x}|\gamma(t+\delta t)))$. The length and the energy functionals can be approximated with respect to this KL-divergence:

$$\text{Length}(\gamma) \approx \sqrt{2}\sum_{t=1}^T \sqrt{\text{KL}(p_t, p_{t+\delta t})} \quad \text{and} \quad \text{Energy}(\gamma) \approx \tfrac{2}{\delta t}\sum_{t=1}^T \text{KL}(p_t, p_{t+\delta t})$$

Proof. On the statistical manifold, we have $\gamma(t+\delta t) = \gamma(t) + \delta t\dot\gamma(t)$. The KL-divergence between perturbed distributions can be defined as: $\text{KL}(p_t, p_{t+\delta t}) = \text{KL}(p(\mathbf{x}|\gamma(t)), p(\mathbf{x}|\gamma(t+\delta t))) = \text{KL}(p(\mathbf{x}|\eta_t), p(\mathbf{x}|\eta_t + \delta\eta_t))$, with $\eta_t = \gamma(t)$ and $\delta\eta_t = \delta t\ \dot\gamma(t)$. Then, we obtain:

$$\text{KL}(p_t, p_{t+\delta t}) = \frac{1}{2}\delta t^2\ \dot\gamma(t)^\top \mathbf{I}_{\mathcal{H}}(\eta_t)\dot\gamma(t) + o(\delta t^2).$$

The length and energy terms appear in the following equations:

$$\int_0^1 \text{KL}(p_t, p_{t+\delta t})dt = \frac{\delta t^2}{2}\int_0^1 \dot\gamma(t)^\top \mathbf{I}_{\mathcal{H}}(\eta_t)\dot\gamma(t)dt + o(\delta t^2) = \frac{\delta t^2}{2}\ \text{Energy}(\gamma) + o(\delta t^2),$$
$$\int_0^1 \sqrt{\text{KL}(p_t, p_{t+\delta t})}dt = \frac{\delta t}{\sqrt{2}}\int_0^1 \sqrt{\dot\gamma(t)^\top \mathbf{I}_{\mathcal{H}}(\eta_t)\dot\gamma(t)}dt + o(\delta t^2) = \frac{\delta t}{\sqrt{2}}\ \text{Length}(\gamma) + o(\delta t^2).$$

If we want approximate any continuous function f with a discrete sequence, by partitioning it in T small segments, such that: $\delta t \approx \frac{1}{T}$, we have: $\int_0^1 f(t)dt \approx \sum_{t=1}^T f(t)\delta t$, which in our case gives:

$$\text{Length}(\gamma) \approx \sqrt{2}\sum_{t=1}^T \sqrt{\text{KL}(p_t, p_{t+\delta t})} \quad \text{and} \quad \text{Energy}(\gamma) \approx \tfrac{2}{\delta t}\sum_{t=1}^T \text{KL}(p_t, p_{t+\delta t}).$$

□

Table C.1: List of distributions

Distributions	Probability density functions	Parameters	Fisher-Rao matrix
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}$	μ,σ^2	$\mathbf{I}_{\mathcal{N}}(\mu,\sigma^2)$
Bernoulli	$\theta^{\mathbf{x}}(1-\theta)^{1-\mathbf{x}}$	θ	$\mathbf{I}_{\mathcal{B}}(\theta)$
Categorical	$\prod_{k=1}^{K}\theta_k^{\mathbf{x}_k}$	$\theta_1,\dots,\theta_K$	$\mathbf{I}_{\mathcal{C}}(\theta_1,\dots,\theta_K)$
Gamma	$\frac{\beta^\alpha\mathbf{x}^{\alpha-1}e^{-\beta\mathbf{x}}}{\Gamma(\alpha)}$	α,β	$\mathbf{I}_{\mathcal{G}}(\alpha,\beta)$
von Mises-Fisher, for $\mathbb{S}^2$	$\frac{\kappa}{4\pi\sinh\kappa}\exp(\kappa\mu^\top\mathbf{x})$	κ,μ	$\mathbf{I}_{\mathcal{S}}(\kappa,\mu)$
Beta	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\mathbf{x}^{\alpha-1}(1-\mathbf{x})^{\beta-1}$	α,β	$\mathbf{I}_{\mathcal{B}}(\alpha,\beta)$

C.1.1 The Fisher-Rao metric for several distributions

With the notations of Table C.1, the Fisher-Rao matrices of the univariate Normal, Bernoulli and Categorical are:

$$\mathbf{I}_{\mathcal{N}}(\mu,\sigma^2)=\begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^2}\end{pmatrix},\quad \mathbf{I}_{\mathcal{B}}(\theta)=\frac{1}{\theta(1-\theta)},\quad \mathbf{I}_{\mathcal{C}}(\theta_1,\dots,\theta_K)=\begin{pmatrix}1/\theta_1 & 0 & \dots & 0\\ 0 & 1/\theta_2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1/\theta_K\end{pmatrix}$$

In addition, the Fisher-Rao matrices of the Gamma, Von Mises-Fisher and the Beta distributions are:

$$\begin{aligned}
\mathbf{I}_{\mathcal{G}}(\alpha,\beta)&=\begin{pmatrix}\frac{\alpha}{\beta^2} & -\frac{1}{\beta}\\ -\frac{1}{\beta} & \Psi_1(\alpha)\end{pmatrix},\\
\mathbf{I}_{\mathcal{S}}(\kappa,\mu)&=\begin{pmatrix}\kappa K(\kappa)(\mathbf{I}-3\mu\mu^\top)+\kappa^2\mu\mu^\top & (\kappa K(\kappa)^2-\frac{2}{k}K(\kappa)+1)\mu\\ (\kappa K(\kappa)^2-\frac{2}{k}K(\kappa)+1)\mu^\top & 3K(\kappa)^2-\frac{2}{\kappa}K(\kappa)+1\end{pmatrix},\\
\mathbf{I}_{\mathcal{B}}(\alpha,\beta)&=\begin{pmatrix}\Psi_1(\alpha)-\Psi_1(\alpha+\beta) & -\Psi_1(\alpha+\beta)\\ -\Psi_1(\alpha+\beta) & \Psi_1(\beta)-\Psi_1(\alpha+\beta)\end{pmatrix},
\end{aligned}$$

with $\Psi_1(\alpha)=\frac{\partial^2\ln\Gamma(\alpha)}{\partial\alpha}$ the trigamma function, and $K(\kappa)=\coth\kappa-\frac{1}{\kappa}$.

Proof. The univariate Normal, Bernoulli and Categorical have already been studied by Tomczak (2012), and the Beta distribution by Brigant and Puechmorel (2019). We will then focus our proof on the Gamma and the von Mises-Fisher distributions.

In order to bypass unnecessary details, we will use the following notations, we redefine the Fisher-Rao as: $\mathbf{I}(\eta)=\mathbb{E}_x[g(\eta,x)g(\eta,x)^\top]$, with $g(\eta,x)=\nabla_\eta\ln p(x|\eta)$ the Fisher score. We call $G=g(\eta,x)g(\eta,x)^\top$, and G_{ij} the matrix elements.

Gamma distribution:

We have $p(x|\alpha,\beta)=\Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1}e^{-\beta x}$, which leads to:

$$\begin{aligned}
\ln p(x|\alpha,\beta)&=-\ln\Gamma(\alpha)+\alpha\ln\beta+(\alpha-1)\ln x-\beta x,\\
\frac{\partial\ln p}{\partial\alpha}&=-\Psi(\alpha)+\ln\beta+\ln x,\\
\frac{\partial\ln p}{\partial\beta}&=\frac{\alpha}{\beta}-x.
\end{aligned}$$

Then:

$$G_{11} = \left(\frac{\partial \ln p}{\partial \alpha}\right)^2 = (\Psi_0(\alpha) + \ln \beta)^2 + 2(\Psi(\alpha) + \ln \beta) lnx + \ln^2 x,$$
$$G_{22} = \left(\frac{\partial \ln p}{\partial \beta}\right)^2 = \left(\frac{\alpha}{\beta}\right)^2 - 2\frac{\alpha}{\beta}x + x^2,$$
$$G_{12} = G_{21} = \frac{\partial \ln p}{\partial \alpha} \cdot \frac{\partial \ln p}{\partial \beta} = (\Psi(\alpha) + \ln \beta)\left(\frac{\alpha}{\beta} - x\right) + \frac{\alpha}{\beta}\ln x - x \ln x.$$

We know that $\mathbb{E}[x] = \frac{\alpha}{\beta}$. We can compute, using your favorite symbolic computation software, the following moments:

$$\mathbb{E}[\ln x] = -\ln \beta + \Psi(\alpha)$$
$$\mathbb{E}[x \ln x] = \frac{\alpha}{\beta}(\Psi(\alpha + 1) - \ln \beta)$$
$$\mathbb{E}[\ln^2 x] = (\ln \beta - \Psi(\alpha))^2 + \Psi_1(\alpha)$$

Replacing the moments for the following equations: $\mathbb{E}[G_{11}]$, $\mathbb{E}[G_{22}]$ and $\mathbb{E}[G_{12}]$ will finally give the Fisher-Rao matrix.

Von Mises-Fisher distribution, for $\mathbb{S}^2$:

We have $p(\mathbf{x}|\mu, \kappa) = C_3(\kappa)\exp(\kappa\mu^\top \mathbf{x})$, with $C_3(\kappa) = \kappa(4\pi \sinh \kappa)^{-1}$. Here, μ is a 3-dimensional vector with $\|\mu\| = 1$.

$$\ln p(\mathbf{x}|\mu, \kappa) = \ln \kappa - \ln 4\pi - \ln \sinh(\kappa) + \kappa\mu^\top \mathbf{x}$$
$$\nabla_\mu \ln p = \kappa \mathbf{x}$$
$$\frac{\partial \ln p}{\partial \kappa} = \kappa^{-1} - \coth(\kappa) + \mu^\top \mathbf{x}.$$

Here, the Fisher-Rao matrix $\mathbf{I}_\mathcal{S}$ will be composed of block matrices, such that: $\mathbf{I}_\mathcal{S} = \mathbb{E}[G]$, with G_{11} a 3×3-matrix, G_{22} a scalar, and $G_{12} = G_{21}^\top$ a 3-dimensional vector.

$$G_{11} = \nabla_\mu \ln p \nabla_\mu \ln p^\top = \kappa^2 \mathbf{x}\mathbf{x}^\top$$
$$G_{22} = \left(\frac{\partial \ln p}{\partial \kappa}\right)^2 = K(\kappa)^2 + 2K(\kappa)\mu^\top \mathbf{x} + (\mu^\top \mathbf{x})^2$$
$$G_{12} = G_{21}^\top = \frac{\partial \ln p}{\partial \kappa} \cdot \nabla_\mu \ln p = \left(K(\kappa) + \mu^\top \mathbf{x}\right)\kappa \mathbf{x},$$

with $K(\kappa) = \coth(\kappa) - \frac{1}{\kappa}$.

We know from Hillen et al. (2016) that the mean and variance of the von Mises-Fisher distribution in the 3-dimensional case is: $\mathbb{E}[\mathbf{x}] = K(\kappa)\mu$ and $\mathrm{Var}[\mathbf{x}] = \frac{1}{\kappa}K(\kappa)\mathbf{I} + (1 - \frac{\coth(\kappa)}{\kappa} + \frac{2}{\kappa^2} - \coth^2(\kappa))\mu\mu^\top$. We can then deduce the following meaningful moments:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathrm{Var}[\mathbf{x}] + \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top = \left(1 - \frac{3}{\kappa}K(\kappa)\right)\mu\mu^\top + \frac{1}{\kappa}K(\kappa)\mathbf{I},$$
$$\mathbb{E}[\mu^\top \mathbf{x}] = \mu^\top \mathbb{E}[\mathbf{x}] = K(\kappa)\mu^\top \mu = K(\kappa)$$
$$\mathbb{E}[(\mu^\top \mathbf{x})^2] = \mu^\top \mathrm{Var}[\mathbf{x}]\mu + \mathbb{E}[\mu^\top \mathbf{x}]^2 = 1 - \frac{2}{\kappa}K(\kappa),$$
$$\mathbb{E}[\mu^\top \mathbf{x}\mathbf{x}] = \mathbb{E}[\mu \mathbf{x}\mathbf{x}^\top] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mu = \left(1 - \frac{3}{\kappa}K(\kappa)\right)\mu + \frac{1}{\kappa}K(\kappa)\mu.$$

Replacing those moments in the following expressions: $\mathbb{E}[G_{11}]$, $\mathbb{E}[G_{22}]$, $\mathbb{E}[G_{12}]$ directly gives the Fisher-Rao metric. □

C.2 Curve energy approximation for categorical data

In this section we present the details of the example in Section 7.3.3. In particular, we the steps to derive an approximation to the energy of a latent curve in closed form, which is suitable for applying automatic differentiation. This is particularly useful for our setting, since it allows us to consider our framework as a Black Box Random Geometry processing toolbox.

Let a random variable $\mathbf{x} \in \mathrm{R}^D$ that follows a generalized Bernoulli likelihood $p(\mathbf{x}|\eta)$, so the vector $\mathbf{x} \in \mathrm{R}^D$ is of the form $\mathbf{x} = (0, \cdots, 1, \cdots, 0)$ with $\sum_i x_i = 1$. The parameters $\eta \in \mathrm{R}^D$ are given as $\eta = h(\mathbf{z})$, with $\eta_i \geq 0 \ \forall i$ and $\sum_i \eta_i = 1$, so we know that the parameters lie on the unit simplex. Actually, they represent the probability the corresponding dimension to be 1 on a random draw. Also, the $p(\mathbf{x}|\mathbf{z}) = \eta_1^{[x_1]} \cdots \eta_D^{[x_D]}$, where $[x_i] = 1$ if $x_i = 1$ else $[x_i] = 0$ which can be seen as an indicator function. The $\log p(\mathbf{x}|\eta) = \sum_i [x_i] \log(\eta_i)$ and $\nabla_\eta \log p(\mathbf{x}|\eta) = \left(\frac{[x_1]}{\eta_1}, \ldots, \frac{[x_D]}{\eta_D} \right)$. Due to the outer product we have to compute the following expectations

$$\mathbb{E}_{\mathbf{x}} \left[\frac{[x_i]}{\eta_i} \frac{[x_j]}{\eta_j} \right] = 0, \quad \text{if} \quad i \neq j, \tag{C.2}$$

$$\mathbb{E}_{\mathbf{x}} \left[\left(\frac{[x_i]}{\eta_i} \right)^2 \right] = \frac{1}{\eta_i}, \quad \text{if} \quad i = j, \tag{C.3}$$

because the $[x_i]$ and $[x_j]$ cannot be 1 on the same time, while the $\mathbb{E}_{\mathbf{x}}[[x_i]^2] = \eta_i$ as it shows the number of times $x_i = 1$. So the Fisher-Rao metric of $\mathcal{H}$ is equal to $\mathbf{I}_{\mathcal{H}}(\eta) = \mathrm{diag}\,(1/\eta_1, \ldots, 1/\eta_D)$. Note that the shortest paths between two distributions must be on the unit simplex in $\mathcal{H}$, while on the same time respecting the geometry of the Fisher-Rao metric.

We can easily parametrize the unit simplex by $[\eta_1, \ldots, \eta_{D-1}, \widetilde{\eta}_D]$ with

$$\widetilde{\eta}_D(\eta_1, \ldots, \eta_{D-1}) = 1 - \sum_{i=1}^{D-1} \eta_i. \tag{C.4}$$

This allows to pullback the Fisher-Rao metric in the latent space $[\eta_1, \ldots, \eta_{D-1}]$ as we have described in this paper. Intuitively, the $\mathbf{z} = [\eta_1, \ldots, \eta_{D-1}]$ and the function h is the parametrization of the simplex. Hence, we are able to compute the shortest path using the induced metric.

However, there is a simpler way to compute this path. We know that the element-wise square root of the parameters η gives a point on the positive orthonant of the unit sphere as $y_i = \sqrt{\eta_i} \ \Rightarrow \ \sum_i y_i^2 = \sum_i \sqrt{\eta_i}^2 = 1$. We also know that the shortest path on a sphere is the great-circle. Therefore, the distance between two distributions parametrized by η and η' on the unit simplex in $\mathcal{H}$, can be equivalently measured using the great-circle distance between their square roots as

$$\mathrm{dist}(\eta, \eta') = \arccos \sqrt{\eta}^\top \sqrt{\eta'}. \tag{C.5}$$

In this way, we can approximate the energy of a curve $c(t)$ in the latent space as follows

$$\begin{aligned} \text{Energy}[c] &\approx \sum_{n=1}^{N-1} \mathrm{dist}^2(h(c(n/N)), h(c(n+1/N))) = \sum_{n=1}^{N-1} \arccos^2 \sqrt{h(c(n/N))}^\top \sqrt{h(c(n+1/N))} \\ &= \sum_{n=1}^{N-1} \left(2 - 2\sqrt{h(c(n/N))}^\top \sqrt{h(c(n+1/N))} \right), \end{aligned} \tag{C.6}$$

where we used at the last step the small angle approximation $\cos\theta \approx 1 - \frac{\theta^2}{2} \Leftrightarrow \theta^2 \approx 2 - 2\cos\theta$. Note that this formulation is suitable for our proposed method to compute the shortest paths (see Section 7.3.2).

The derivation above represents the conceptual strategy, while in general we proposed to use the KL divergence approximation result (7.8) in place of the great-circle distance. Intuitively, when the KL divergence has an analytic solution, we can derive an analogous energy approximation. Even if the solution of the KL is intractable, we can still use our approach as long as we can estimate the KL using Monte Carlo and propagate the gradient through the samples using a re-parametrization scheme or a score function estimator.

C.3 Information geometry in generative modeling

In this section we present the additional technical information related to the pullback Fisher-Rao metric in the latent space of a VAE.

C.3.1 Details for the pullback metric in the latent space

We call h the non-linear function, typically parametrized as deep neural networks, that maps the variables from the latent space $\mathcal{Z}$ to the parameter space $\mathcal{H}$, such that: $h(\mathbf{z}) = \eta$, with $\mathbf{z} \in \mathcal{Z}$ and $\eta \in \mathcal{H}$. Furthermore, the data $\mathbf{x} \in \mathcal{X}$ is reconstructed such that it follows a specific distribution: $\mathbf{x} \sim p(\mathbf{x}|\eta)$, with $p(\mathbf{x}|\eta)$ being for instance a Bernoulli or Gaussian distribution. The parameter space $\mathcal{H}$ is a statistical manifold equipped with Fisher-Rao metric: $\mathbf{I}_{\mathcal{H}}(\eta) \triangleq \int p(\mathbf{x}|\eta) \left[\nabla_\eta \log p(\mathbf{x}|\eta) \nabla_\eta \log p(\mathbf{x}|\eta)^\top\right] d\mathbf{x}$. We denote by $\mathbf{J}_h$ the Jacobian of h.

Proposition C.3.1
The latent space $\mathcal{Z}$ is equipped with the Riemannian pullback metric tensor:

$$\mathbf{M}(\mathbf{z}) \triangleq \mathbf{J}_h(\mathbf{z})^\top \mathbf{I}_{\mathcal{H}}(h(\mathbf{z})) \mathbf{J}_h(\mathbf{z}).$$

Proof. The parameter space is a statistical manifold equipped with the Fisher-Rao metric $\mathbf{I}_{\mathcal{H}}(\eta)$, thus the scalar product at η between two vectors $d\eta_1, d\eta_2 \in \mathcal{H}$ is: $\langle d\eta_1, d\eta_2 \rangle_{\mathbf{I}_{\mathcal{H}}(\eta)} = d\eta_1^\top \mathbf{I}_{\mathcal{H}}(\eta) d\eta_2$. For two vectors $d\mathbf{z}_1, d\mathbf{z}_2 \in \mathcal{Z}$, we have at $\eta = f(\mathbf{z})$ that: $\langle d\eta_1, d\eta_2 \rangle_{\mathbf{I}_{\mathcal{H}}(\eta)} = \langle \mathbf{J}_h(\mathbf{z}) d\mathbf{z}_1, \mathbf{J}_h(\mathbf{z}) d\mathbf{z}_2 \rangle_{\mathbf{I}_{\mathcal{H}}(\eta)} = d\mathbf{z}_1^\top (\mathbf{J}_h(\mathbf{z})^\top \mathbf{I}_{\mathcal{H}}(h(\mathbf{z})) \mathbf{J}_h(\mathbf{z})) d\mathbf{z}_2$.

$\mathbf{I}_{\mathcal{H}}(h(\mathbf{z}))$ is a Riemannian metric tensor by definition, and it is then positive definite. Furthermore, $h : \mathcal{Z} \to \mathcal{H}$ is a smooth immersion, and so $\mathbf{J}_h(\mathbf{z})$ is full-rank. It follows that $\mathbf{J}_h(\mathbf{z})^\top \mathbf{I}_{\mathcal{H}}(h(\mathbf{z})) \mathbf{J}_h(\mathbf{z})$ is positive definite. Hence, $\mathbf{M}(\mathbf{z})$ is a Riemannian metric tensor. □

Proposition C.3.2
Our pullback metric $\mathbf{M}(\mathbf{z})$ is actually equal to the Fisher-Rao metric obtained over the parameter space $\mathcal{Z}$:

$$\mathbf{M}(\mathbf{z}) = \mathbf{I}_{\mathcal{Z}}(\mathbf{z}) \triangleq \int p(\mathbf{x}|\mathbf{z}) \left[\nabla_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z}) \nabla_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z})^\top\right] d\mathbf{x}$$

Proof. We will show that $\mathbf{I}_{\mathcal{Z}}(\mathbf{z}) = \mathbf{J}_f(\mathbf{z})^\top \mathbf{I}_{\mathcal{H}}(\eta)\mathbf{J}_f(\mathbf{z})$. Let's consider the definition of the Fisher-Rao metric in $\mathcal{Z}$:

$$\begin{aligned}
\mathbf{I}_{\mathcal{Z}}(\mathbf{z}) &= \int \nabla_{\mathbf{z}} \log p(\mathbf{x} \mid \mathbf{z}) \cdot \nabla_{\mathbf{z}} \log p(\mathbf{x} \mid \mathbf{z})^\top p(\mathbf{x} \mid \mathbf{z}) d\mathbf{x} && \text{(C.7)} \\
&= \int \mathbf{J}_f(\mathbf{z})^\top \nabla_\eta \log p(\mathbf{x}|\eta) \nabla_\eta \log p(\mathbf{x}|\eta)^\top \mathbf{J}_f(\mathbf{z}) p(\mathbf{x}|\eta) d\mathbf{x} && \text{(C.8)} \\
&= \mathbf{J}_f(\mathbf{z})^\top \left[\int_{\mathcal{X}} \nabla_\eta \log p(\mathbf{x}|\eta) \nabla_\eta \log p(\mathbf{x}|\eta)^\top p(\mathbf{x}|\eta) d\mathbf{x} \right] \mathbf{J}_f(\mathbf{z}) && \text{(C.9)} \\
&= \mathbf{J}_f(\mathbf{z})^\top \mathbf{I}_{\mathcal{H}}(f(\mathbf{z})) \mathbf{J}_f(\mathbf{z}) = \mathbf{M}(\mathbf{z})
\end{aligned}$$

where we use the fact that $\eta = f(\mathbf{z})$ so the $\nabla_{\mathbf{z}} \log p(\mathbf{x}|f(\mathbf{z})) = \mathbf{J}_f(\mathbf{z})^\top \cdot \nabla_\eta \log p(\mathbf{x}|\eta)$

The same argument can be proved as follows:

$$\begin{aligned}
&\langle d\eta, \mathbf{I}_{\mathcal{H}}(\eta) d\eta \rangle = \langle \mathbf{J}_f(\mathbf{z}) d\mathbf{z}, \mathbf{I}_{\mathcal{H}}(f(\mathbf{z})) \mathbf{J}_f(\mathbf{z}) d\mathbf{z} \rangle && \text{(C.10)} \\
&= \langle \mathbf{J}_f(\mathbf{z}) d\mathbf{z}, \int \nabla_\eta \log p(\mathbf{x}|\eta) \nabla_\eta \log p(\mathbf{x}|\eta)^\top p(\mathbf{x}|\eta) d\mathbf{x}\ \mathbf{J}_f(\mathbf{z}) d\mathbf{z} \rangle && \text{(C.11)} \\
&= \langle d\mathbf{z}, \int \mathbf{J}_f(\mathbf{z})^\top \cdot \nabla_\eta \log p(\mathbf{x}|\eta)\ \nabla_\eta \log p(\mathbf{x}|\eta)^\top \cdot \mathbf{J}_f(\mathbf{z}) p(\mathbf{x}|\eta) d\mathbf{x}\ d\mathbf{z} \rangle && \text{(C.12)} \\
&= \langle d\mathbf{z}, \int \nabla_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z}) \nabla_{\mathbf{z}} \log p(\mathbf{x}|\mathbf{z})^\top p(\mathbf{x}|\mathbf{z}) d\mathbf{x}\ d\mathbf{z} \rangle = \langle d\mathbf{z}, \mathbf{I}_{\mathcal{Z}}(\mathbf{z}) d\mathbf{z} \rangle && \text{(C.13)}
\end{aligned}$$

□

In section C.1.1, we have seen how to derive a closed-form expression of the Fisher-Rao metric for a one-dimensional observation x that follows a specific distribution. In practice, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$ is a multidimensional variable where each dimension represents, for instance, a pixel when working with images or a feature when working with tabular data. Each feature, x_i with $i = 1 \cdots D$, is obtained for a specific set of parameters $\{\eta_i\}$. We assume that the features follow the same distribution $\mathcal{D}$, such that: $x_i \sim p(x_i|\eta_i)$, and $p(\mathbf{x}|\eta) = \prod_{i=1}^D p(x_i|\eta_i)$.

Proposition C.3.3
If the features follow the same distribution $\mathcal{D}$, such that: $x_i \sim p(x_i|\eta_i)$ and $p(\mathbf{x}|\eta) = \prod_{i=1}^D p(x_i|\eta_i)$, then the Fisher-Rao metric $\mathbf{I}_{\mathcal{H}}(\eta)$ is a block matrix where the diagonal terms are the Fisher-Rao matrices $\mathbf{I}_{\mathcal{H},i}$ obtained for each data feature x_i:

$$\mathbf{I}_{\mathcal{H}}(\eta) = \begin{pmatrix} \mathbf{I}_{\mathcal{H},1} & 0 & \dots & 0 \\ 0 & \mathbf{I}_{\mathcal{H},2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{I}_{\mathcal{H},D} \end{pmatrix}$$

Proof. We have $x_i \sim p(x_i|\eta_i)$ and $\mathbf{I}_{\mathcal{H},i} = \int p(x_i|\eta_i) \left[\nabla_{\eta_i} \log p(x_i|\eta_i) \nabla_{\eta_i} \log p(x_i|\eta_i)^\top \right] dx_i$.
Also, we assumed: $p(\mathbf{x}|\eta) = \prod_{i=1}^D p(x_i|\eta_i)$. We then have: $\log p(\mathbf{x}|\eta) = \sum_{i=1}^D \log p(x_i|\eta_i)$,
and the Fisher score: $\nabla_\eta \log p(\mathbf{x}|\eta) = \nabla_\eta \sum_{i=1}^D \log p(x_i|\eta_i) = [\nabla\eta_1 \ln p(x_1|\eta_1), \dots, \nabla\eta_D \ln p(x_1|\eta_D)]^\top$.

The matrix $\mathbf{I}_{\mathcal{H}}(\eta)$ is thus a $D \times D$ block matrix, where the (i, j)-block element is:

$$I_{ij} = \int p(x_i|\eta_i) \left[\nabla_{\eta_i} \log p(x_i|\eta_i) \nabla_{\eta_i} \log p(x_j|\eta_j)^\top \right] dx_i.$$

Let's note that:

$$\int p(x_i|\eta_i)\nabla_{\eta_i}\log p(x_i|\eta_i)dx_i = \int p(x_i|\eta_i)\frac{\nabla_{\eta_i}p(x_i|\eta_i)}{p(x_i|\eta_i)}dx_i = \nabla_{\eta_i}\int p(x_i|\eta_i)dx_i = 0.$$

When $i = j$, we have $\mathbf{I}_{ii} = \mathbf{I}_{\mathcal{H},i}$, with $\mathbf{I}_{\mathcal{H},i}$ being the Fisher-Rao metric obtained for: $x_i \sim p(x_i|\eta_i)$.
When $i \neq j$, we have: $\mathbf{I}_{ij} = \nabla\log p(x_j|\eta_j)^\top \int p(x_i|\eta_i)\nabla_{\eta_i}\log p(x_i|\eta_i)dx_i = 0$. □

Then, for example, if we are dealing with binary images, and make the assumption that each pixel x_i follows a Bernoulli distribution: $p(x_i|\eta_i) = \eta^{x_i}(1-\eta_i)^{1-x_i}$, then according to Section C.1.1 and Proposition C.3.3, the Fisher-Rao matrix that endows the parameter space $\mathcal{H}$ is:

$$\mathbf{I}_{\mathcal{H}}(\eta) = \begin{pmatrix} \frac{1}{\eta_1(1-\eta_1)} & 0 & \dots & 0 \\ 0 & \frac{1}{\eta_2(1-\eta_2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\eta_D(1-\eta_D)} \end{pmatrix}.$$

We have seen that in theory, we can obtain a close form expression for the pullback metric, if the probability distribution is known. In practice, we can directly infer the metric using the approximation of the KL-divergence.

Proposition C.3.4
We define perturbations vectors as: $\delta e_i = \varepsilon \cdot \mathbf{e}_i$, with $\varepsilon \in \mathbb{R}_+$ a small infinitesimal quantity, and $(\mathbf{e}_i)$ a canonical basis vector in $\mathbb{R}^d$. For better clarity, we rename $\mathrm{KL}(p(\mathbf{x}|\mathbf{z}), p(\mathbf{x}|\mathbf{z}+\delta\mathbf{z})) = \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{z})$, and we note $\mathbf{M}_{ij}$ the components of $\mathbf{M}(\mathbf{z})$. We can then approximate by a system of equations the diagonal and non-diagonal elements of the metric:

$$\begin{aligned} \mathbf{M}_{ii} &\approx 2\ \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i)/\varepsilon^2 \\ \mathbf{M}_{ij} = \mathbf{M}_{ji} &\approx \left(\mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i + \delta\mathbf{e}_j) - \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i) - \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_j)\right)/\varepsilon^2. \end{aligned}$$

Proof. From Proposition C.1.1, we know that:

$$\mathrm{KL}_{\mathbf{z}}(\delta\mathbf{z}) = \frac{1}{2}\delta\mathbf{z}^\top\mathbf{M}(\mathbf{z})\delta\mathbf{z} + o(\delta\mathbf{z}^2).$$

Let's take $\delta e_i = \varepsilon \cdot \mathbf{e}_i$. On one hand, we have: $\delta e_i^\top\mathbf{M}(\mathbf{z})\delta e_i = \varepsilon^2\mathbf{M}_{ii}$. On the second hand, we also have: $\delta e_i^\top\mathbf{M}(\mathbf{z})\delta e_i \approx 2\mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i)$, which gives us the equation to infer the diagonal elements of the metric.

Now, let's take $\delta e_i + \delta e_j = \varepsilon \cdot (\mathbf{e}_i + \mathbf{e}_j)$. Then, we have: $(\delta e_i + \delta e_j)^\top\mathbf{M}(\mathbf{z})(\delta e_i + \delta e_j) = \varepsilon^2(\mathbf{M}_{ii} + \mathbf{M}_{jj} + \mathbf{M}_{ij} + \mathbf{M}_{ji})$. We also know that $\mathbf{M}_{ji} = \mathbf{M}_{ij}$. Again, we also have: $(\delta e_i + \delta e_j)^\top\mathbf{M}(\mathbf{z})(\delta e_i + \delta e_j) \approx 2\mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i + \delta\mathbf{e}_j)$.

We can replace the terms $\mathbf{M}_{ii}$ and $\mathbf{M}_{jj}$ in the equation obtained above with the KL-divergence for the diagonal terms. Which finally gives us:

$$\mathbf{M}_{ij} = \mathbf{M}_{ji} \approx \left(\mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i + \delta\mathbf{e}_j) - \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_i) - \mathrm{KL}_{\mathbf{z}}(\delta\mathbf{e}_j)\right)/\varepsilon^2.$$

□

Table C.2: This table shows the implementations of the decoder and extrapolate functions in Eq. (C.15) for all the distributions studied in our second experiment (see Sec. 7.4.2). Here we represent a randomly initialized neural network with f_i, where i represents the size of the co-domain. For example, in the case of the Dirichlet distribution, we use a randomly initialized neural network to compute the parameters α of the distribution and, since these have to be positive, we pass the output of this network through a Softplus activation; moreover, since the Dirichlet distribution is approximately uniform when all its parameters equal 1, our extrapolation mechanism consists of replacing the output of the network with a constant vector of ones.

Distribution	$h\colon \mathcal{Z}_{\text{toy}} \to \mathcal{H}$	Extrapolation mechanism
Normal	$\mu(\mathbf{z}) = 10 \cdot f_3(\mathbf{z}), \quad \sigma(\mathbf{z}) = 10 \cdot \text{Softplus}(f_3(\mathbf{z}))$	$\sigma(\mathbf{z}) \to \infty$
Bernoulli	$p(\mathbf{z}) = \text{Sigmoid}(f_{15}(\mathbf{z}))$	$p(\mathbf{z}) = 1/2$
Beta	$\alpha(\mathbf{z}) = 10 \cdot \text{Softplus}(f_3(\mathbf{z})), \quad \beta(\mathbf{z}) = 10 \cdot \text{Softplus}(f_3(\mathbf{z}))$	$(\alpha(\mathbf{z}), \beta(\mathbf{z})) = (1, 1)$
Dirichlet	$\alpha(\mathbf{z}) = \text{Softplus}(f_3(\mathbf{z}))$	$\alpha(\mathbf{z}) = 1$
Exponential	$\lambda(\mathbf{z}) = \text{Softplus}(f_3(\mathbf{z}))$	$\lambda(\mathbf{z}) \to 0$

C.3.2 Uncertainty quantification and regularization

As discussed in the main text, we carefully design our mappings from latent space to parameter space such that they model the training codes according to the learned decoders, and extrapolate to uncertainty outside the support of the data. This, we refer to as **uncertainty regularization**. In this section we explain it in detail. The core idea of this uncertainty regularization is imposing a "slider" that forces the distribution $p(\mathbf{x}|\mathbf{z})$ to change when $\mathbf{z}$ is far from the training latent codes. For this, we use a combination of KMeans and the sigmoid activation function.

We start by encoding our training data, arriving at a set of latent codes $\{\mathbf{z}_n\}_{n=1}^N \subseteq \mathcal{Z}$. We then train KMeans($k$) on these latent codes (where k is a hyperparameter that we tweak manually), arriving at k cluster centers $\{\underline{\mathrm{c}}_j\}_{j=1}^k$. These cluster centers serve as a proxy for "closeness" to the data: we know that a latent code $\mathbf{z} \in \mathcal{Z}$ is near the support if $D(\mathbf{z}) := \min_j \left\{\|\mathbf{z} - \underline{\mathrm{c}}_j\|^2\right\}$ is close to 0.

The next step in our regularization process is to reweight our decoded distributions such that we decode to high uncertainty when $D(\mathbf{z})$ is large, and we decode to our learned distributions when $D(\mathbf{z}) \approx 0$. This mapping from $[0, \infty) \to (0, 1)$ can be constructed using a modified sigmoid function Detlefsen et al. (2020) and Detlefsen et al. (2019), consider indeed

$$\tilde{\sigma}_\beta(d) = \text{Sigmoid}\left(\frac{d - c \cdot \text{Softplus}(\beta)}{\text{Softplus}(\beta)}\right), \tag{C.14}$$

where $\beta \in \mathbb{R}$ is another hyperparameter that we manually tweak, and $c \approx 7$.

With this translated sigmoid, we have that $\tilde{\sigma}_\beta(D(\mathbf{z}))$ is close to 0 when $\mathbf{z}$ is close to the support of the data (i.e. close to the cluster centers), and it converges to 1 when $D(\mathbf{z}) \to \infty$. $\tilde{\sigma}_\beta(D(\mathbf{z}))$ serves, then, as a slider that indicates closeness to the training codes. This reweighting takes the following form:

$$\text{reweight}(\mathbf{z}) = (1 - \tilde{\sigma}_\beta(D(\mathbf{z})))h(\mathbf{z}) + \tilde{\sigma}_\beta(D(\mathbf{z}))\,\text{extrapolate}(\mathbf{z}), \tag{C.15}$$

where $h(\mathbf{z}) = \eta \in \mathcal{H}$ represents our learned networks in parameter space, and extrapolate($\mathbf{z}$) returns the parameters of the distribution that maximize uncertainty (e.g. $\sigma \to \infty$ in the case of an isotropic Gaussian, $p \to 1/2$ in the case of a Bernoulli, and $\kappa \to 0$ in the case of the von Mises-Fisher).

For the particular case of the experiment in which we pull back the Fisher-Rao metric from the parameter space of several distributions (see 7.4.2), Table C.2 provides the exact extrapolation mechanisms and implementations of $h(\mathbf{z})$.

C.4 Details for our implementation and experiments

In this section we present the technical details that we used in our implementation and experiments. We are currently implementing an open-source version of our code here.

C.4.1 What we mean when we say black-box random geometry

Before we dive into the specific details of our experiments, it is worth noting that they were all made using the same *interface*. This is precisely what we mean when we say that our results open the doors for black-box random geometry: We can define a `curve energy` method that is agnostic to the distribution our models decode to.

To hammer this point home, consider the following interface, written in Python:

```
1  class StatisticalManifold:
2      def __init__(self, model: torch.nn.Module):
3          # A model with regularized uncertainty (see Uncertainty
       Quantification)
4          self.model = model
5          assert "decode" in dir(model)
6
7      def curve_energy(self, curve: CubicSpline) -> torch.Tensor:
8          # An energy function that can be minimized using
       autodifferentiation.
9
10         dt = (curve[1] - curve[0])
11         dist1 = self.model.decode(curve[:-1])
12         dist2 = self.model.decode(curve[1:])
13         kl = kl_divergence(dist1, dist2)
14         energy = kl.sum() * (2 * dt ** -1)
15
16         return energy
```

Notice that the user need only provide a `model` that implements a `decode` function which is expected to return a distribution with proper uncertainty estimates (as described in Sec. C.3.2). Line 14 is a direct implementation of our derived expression for the energy (see Prop. C.1.2). Most distributions of interest are available in the Torch submodule `torch.distributions`, and similar implementations could be done for other frameworks.

C.4.2 Shortest path approximation with cubic splines

As we described in the main paper, we use an approximate solution for the shortest paths based on cubic splines. Let a cubic spline $c_{\boldsymbol{\psi}}(t) = [1, t, t^2, t^3]^\top [\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3]$ with parameters $\boldsymbol{\psi}_i \in \mathrm{R}^{d\times 1}$. Also, in our implementation the actual curve is a piece wise cubic spline, and we optimize the K control points $\underline{c}_k$ as well. We optimize the parameters using the approximation of the curve energy $\{\boldsymbol{\psi}_k^*, \underline{c}_k^*\}_{k=1}^K = \operatorname{argmin}_{\boldsymbol{\psi}} \text{Energy}[c_{\boldsymbol{\psi}}]$. In general, we can use Prop. C.1.2 as long as we can propagate the gradient through the KL or as in (C.6) if an explicit closed form solution exists. In this case, we are able to use automatic differentiation for the optimization of the parameters (as discussed in Sec. C.4.1).

In practical terms, we compute these shortest paths by creating a uniform grid in latent space and computing, only once, the curve energy for the edges of this grid. After this expensive computation (which only needs to be performed once) we can use shortest-paths algorithms in graphs to create a suitable initialization of the geodesic. We fit a cubic spline to this initialization and then optimize its parameters further.

Table C.3: This table shows the Variational Autoencoder used in our first experiment (see Sec. 7.4.1). The network for approximating the standard deviation σ leverages ideas from Arvanitidis et al. (2018), in which an RBF network is trained on latent codes using centers positioned through KMeans. The operation PosLinear(a, b) represents the usual Linear transformation with a inputs and b outputs, but considering only positive weights. To compare between having and not having uncertainty regularization, we use two different approximations of the standard deviation in the decoder: σ_{UR} when performing meaningful uncertainty quantification, and $\sigma_{\text{no UR}}$ otherwise.

Pulling back the Euclidean vs. Fisher-Rao (Sec. 7.4.1)

Module	**MLP**
	Encoder
μ	Linear(728, 2)
	Decoder
μ	Linear(2, 728)
σ_{UR}	RBF(), PosLinear(500, 1), Reciprocal(), PosLinear(1, 728)
$\sigma_{\text{no UR}}$	Linear(2, 728), Softplus()
Optimizer	Adam ($\alpha = 1 \times 10^{-5}$)
Batch size	32

C.4.3 Models used

In this section we describe, in detail, the models that we used for our experiments (see Sec.7.4). All the networks that we used are Multi-Layer Perceptrons implemented in PyTorch.

First, Table C.3 shows the Variational Autoencoder implemented for the experiment described in Sec. 7.4.1. In the computations *without* uncertainty regularization, we used a simpler model for the uncertainty quantification (namely, a single Linear layer, followed by a Softplus activation). For our second experiment involving a toy latent space, we also provide the implementation of the respective MLPs in Table C.4. Finally, Table C.5 and C.6 respectively represents the VAE trained for the experiments related to motion capture (Sec. 7.4.3) and movie rating (Sec. 7.4.6). For the motion capture experiments, we are training a VAE that decodes to a von Mises Fisher distribution, and for the movie rating experiments, we decode to a Bernoulli distribution.

All of these VAEs were trained by maximizing the Evidence Lower Bound with different values for KL annealing which can be read from the different tables. For example, Table C.5 shows that the KL annealing constant was chosen to be 0.01.

C.4.4 Metric approximation and KL by sampling

When visualizing our latent space as a statistical manifold, we can obtain a direct approximation of the metric using the KL-divergence between two close distributions (Proposition 7.3.4). We will show here, in simple cases, how our metric approximation compares to closed-form expressions.

In the following experiment, our statistical manifold is the parameter space of known distributions (Beta and Normal). Their Fisher-Rao matrices are well-known (Sec. C.1.1), and we approximate them by computing the KL-divergence of sampled distributions. We call $\mathbf{M}_t$ the theoretical metric and $\mathbf{M}_a$ the approximated metric, and we note $\varepsilon_r = \frac{\|\mathbf{M}_t - \mathbf{M}_a\|}{\|\mathbf{M}_t\|}$ the relative error between the theoretical and approximated matrices. Here, $\|\cdot\|$ denotes the Frobenius norm. For the Normal distribution, we empirically obtain: $\varepsilon_r = 5.32 \cdot 10^{-4} \pm 9.63 \cdot 10^{-4}$, and for the Beta distribution, we have: $\varepsilon_r = 1.73 \cdot 10^{-5} \pm 1.17 \cdot 10^{-5}$.

Table C.4: This table describes the neural networks used for the experiment presented in Sec. 7.4.2. Following the notation of PyTorch, Linear(a, b) represents an MLP layer with a input nodes and b output nodes. In each of these networks, we implement the reweighting operation described in Sec. C.3.2, and we describe the β hyperparameter present in the modified sigmoid function (Eq. (C.14)). These networks were not trained in any way, and they were initialized using the provided seed.

Toy latent spaces (Sec. 7.4.2)

Distribution	Module	MLP	Seed for randomness	β in $\tilde{\sigma}_\beta$
Normal	μ σ	Linear(2,3) Linear(2,3), Softplus()	1	-2.5
Bernoulli	p	Linear(2,15), Sigmoid()	1	-3.5
Beta	α β	Linear(2,3), Softplus() Linear(2,3), Softplus()	1	-4.0
Dirichlet	α	Linear(2,3), Softplus()	17	-4.0
Exponential	λ	Linear(2,3), Softplus()	17	-4.0

C.4.5 Computational complexity

Proposition C.1.1 shows the system of equations required to approximate the pullback metric in the latent space. Each KL operation requires 2 forward passes from the decoder to compute, so first we establish the lower bound on the time complexity of the decoder forward pass. Ignoring all activation function related operations, for an MLP with H hidden layers, N-dimensional network output, K-dimensional hidden layer output and single M-dimensional vector input, this lower bound is:

$$\Omega\left(MK_1 + K_H N + \sum_{i=1}^{H-1} M_i M_{i+1}\right) \tag{C.16}$$

For each diagonal element $\mathbf{M}_{ii}$ of the metric tensor we need to compute a single KL divergence, which will require two forward passes through the decoder network giving us a (lower bounded) time complexity of $\Omega\left[2\left(MK_1 + K_H N + \sum_{i=1}^{H-1} M_i M_{i+1}\right)\right]$ for each element. For the off-diagonal elements we will need to compute the KL three times which corresponds to six forward passes through the decoder network, which yields a (lower bounded) time complexity $\Omega\left[6\left(MK_1 + K_H N + \sum_{i=1}^{H-1} M_i M_{i+1}\right)\right]$ per element.

C.4.6 Information for the movie preferences experiment

For this experiment we used the MovieLens 25M dataset (https://grouplens.org/datasets/movielens/25m/). Each cell of the data matrix represents the rating of a user (row) from 1 to 5 for the corresponding movie (column). In order to fit a Bernoulli VAE we considered the matrix as binary i.e. if a user has seen a movie (1) or not (0). We then selected the 60 most popular movies, as well as, 10000 users who have seen between 2 and 30 of these movies. We also verified that all the movies have been seen from at least 600 users. In this way we reduced the size of the dataset, obtaining a realistic scenario where: 1) some movies are more popular than the others, and 2) we do not include users that have seen 0 or almost all the movies. We show in Fig. C.1 the number of views for each movie and the number of movies each user has seen. In Table C.6 we present the details for the Bernoulli VAE.

Table C.5: This table shows the Variational Autoencoder used in our last two experiments (see Sec. 7.4.3, 7.4.4). Our motion capture data tracked 26 different bones, and thus we decode to a product of 26 different von Mises-Fisher distributions.

Decoding to a von Mises-Fisher Distribution (Sec 7.4.3, 7.4.4, 7.4.5)

Module	MLP
	Encoder (Normal dist.)
μ	Linear(3×26, 90), Linear(90, 2)
σ	Linear(3×26, 90), Linear(90, 2), Softplus()
	Decoder (vMF dist.)
μ	Linear(2, 90), Linear(90, 3×26), Linear(3×26, 3×26)
κ	Linear(2, 90), Linear(90, 3×26), Linear(3×26, 26), Softplus()
Optimizer	Adam ($\alpha = 1 \times 10^{-3}$)
Batch size	16
β in $\tilde{\sigma}_\beta$	-5.5
KL annealing	0.01
Extrapolation mechanism	$\kappa \to 0.1$

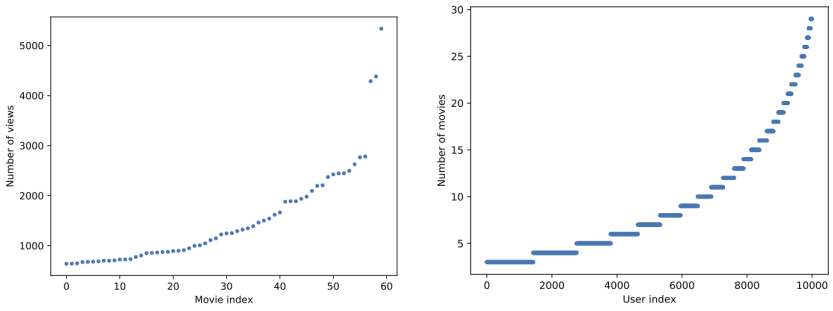


Figure C.1: The numbers of views for the movies and the users.

C.4.7 Information for fitting the LAND model

The locally adaptive normal distribution (LAND) (Arvanitidis et al. 2016) is the extension of the normal distribution on Riemannian manifolds learned from data. Pennec (2006) first derived this distribution on predefined manifolds as the sphere and also showed that it is the maximum entropy distribution given a mean and a precision matrix. The flexibility of this probability density relies on the shortest paths. However, the computational demand to fit this model is relatively high, especially in our case, since we need to use an approximation scheme to find the shortest paths.

In particular, we compute the *logarithmic map* $\mathbf{v} = \text{Log}_{\mathbf{x}}(\mathbf{y})$ by first finding the shortest path between $\mathbf{x}$ and $\mathbf{y}$, and then, rescaling the initial velocity as $\mathbf{v} = \frac{\dot{c}(0)}{||\dot{c}(0)||}\text{Length}(c)$, which ensures that $||\mathbf{v}|| = \text{Length}(c)$. In addition, for the estimation of the normalization constant we use the *exponential map* $\text{Exp}_{\mathbf{x}}(\mathbf{v}) = c_{\mathbf{v}}(t)$, which is the inverse operator that generates the shortest path with $c(1) = \mathbf{y}$ taking the rescaled initial velocity $\mathbf{v}$ as input. Also, we should be able to evaluate the metric. While the logarithmic map can be approximated using our approach (Section C.4.2), for the exponential map we need to solve the ODEs system (C.1) as an initial value problem (IVP). Note that we fit the LAND using gradient descent, which implies that the computation of these operators is the main computational bottleneck.

We provided a method in Proposition 7.3.4, which enables us to approximate the pullback metric in the latent space of a generative model using the corresponding KL divergence. Even if this is a sensible approach, in practice, the computational cost is relatively high

Table C.6: This table shows the Variational Autoencoder used in the movie rating experiment (see Sec. 7.4.6). The MovieLens 25M dataset has been preprocessed such that it is composed of 10000 users rating if they have seen some of 60 selected movies. We only select users that have seen more than two movies and less than 30 movies, to avoid outliers and aim for a more realistic scenario. We used the same extrapolation mechanism described in the toy experiments for the Bernoulli: having the probits be 1/2 (see Sec. C.3.2).

Decoding to a Bernoulli Distribution (Sec 7.4.6)

Module	MLP
	Encoder (Normal dist.)
μ	Linear(60, 16), Tanh(), Linear(16, 16), Tanh(), Linear(16, 2)
σ	Linear(60, 16), Tanh(), Linear(16, 16), Tanh(), Linear(16, 2), Softplus()
	Decoder (Bernoulli dist.)
p	Linear(2, 16), Tanh(), Linear(16, 16), Tanh(), Linear(16, 60), Sigmoid()
Optimizer	Adam ($\alpha = 1 \times 10^{-3}$, $\omega = 1 \times 10^{-7}$)
Batch size	256
β in $\tilde{\sigma}_\beta$	-3.0
KL annealing	0.01
Extrapolation mechanism	$p \to 1/2$

as we might need to estimate the KL using Monte Carlo. For example, this is the case when the likelihood is the von Mises-Fisher. This further implies that fitting the LAND using this approach is prohibited due to the computational cost. Especially, since we need to evaluate many times the metric and its derivative for the computation of each exponential map. Hence, in order to fit the LAND efficiently, we used the following approximation based on Hauberg et al. (2012).

First we construct a uniformly spaced grid in the latent space. Then, we evaluate the metric using Proposition 7.3.4 for each point on the grid getting a set $\{\mathbf{z}_s, \mathbf{M}_s\}_{s=1}^S$ of metric tensors. Thus, we can estimate the metric at any point $\mathbf{z}$ as

$$\mathbf{M}(\mathbf{z}) = \sum_{s=1}^{S} \widetilde{w}_s(\mathbf{z})\mathbf{M}_s, \quad \text{with} \quad \widetilde{w}_s(\mathbf{z}) = \frac{w_s(\mathbf{z})}{\sum_{j=1}^{S} w_s(\mathbf{z})} \quad \text{and} \quad w_s(\mathbf{z}) = \exp\left(-\frac{||\mathbf{z}_s - \mathbf{z}||^2}{2\sigma^2}\right) \tag{C.17}$$

where $\sigma > 0$ the bandwidth parameter. This is by definition a Riemannian metric as a weighted sum of Riemannian metrics with a smooth weighting function. In this way, we can approximate the pullback of the Fisher-Rao metric in the latent space $\mathcal{Z}$ in order to perform the necessary computations more efficiently.

Bibliography

Adams, Larin Cole (2021). 'Gaussian Process Manifold Learning'. PhD thesis. Vanderbilt University.

Agueh, Martial (2012). 'Finsler structure in the p-Wasserstein space and gradient flows'. In: *Comptes Rendus Mathematique* 350.1-2, pp. 35–40.

Ahmed, Sumon, Magnus Rattray, and Alexis Boukouvalas (2019). 'GrandPrix: scaling up the Bayesian GPLVM for single-cell data'. In: *Bioinformatics* 35.1, pp. 47–54.

Amari, Shun-ichi and Hiroshi Nagaoka (2000). *Methods of information geometry*. Vol. 191. American Mathematical Soc.

Andrews, Ben and Christopher Hopper (2010). *The Ricci flow in Riemannian geometry: a complete proof of the differentiable 1/4-pinching sphere theorem*. springer.

Antonelli, Peter L and Radu Miron (2013). *Lagrange and Finsler Geometry: Applications to Physics and Biology*. Vol. 76. Springer Science & Business Media.

Arvanitidis, Georgios, Lars Kai Hansen, and Søren Hauberg (2016). 'A Locally Adaptive Normal Distribution'. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

– (2018). 'Latent Space Oddity: on the Curvature of Deep Generative Models'. In: *International Conference on Learning Representations (ICLR)*.

Arvanitidis, Georgios, Søren Hauberg, and Bernhard Schölkopf (2021). 'Geometrically Enriched Latent Spaces'. In: *Artificial Intelligence and Statistics (AISTATS)*.

Arvanitidis, Georgios et al. (2019). 'Fast and Robust Shortest Paths on Manifolds Learned from Data'. In: *Artificial Intelligence and Statistics (AISTATS)*.

Arvanitidis, Georgios et al. (2022). 'Pulling back information geometry'. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. Equal contribution from all authors. PMLR, pp. 4872–4894. URL: https://proceedings.mlr.press/v151/arvanitidis22b.html.

Ay, Nihat et al. (2015). 'Information geometry and sufficient statistics'. In: *Probability Theory and Related Fields* 162, pp. 327–364.

Bao, David, Colleen Robles, and Zhongmin Shen (2004). 'Zermelo navigation on Riemannian manifolds'. In: *Journal of Differential Geometry* 66.3, pp. 377–435.

Bauer, Martin, Martins Bruveris, and Peter W Michor (2016). 'Uniqueness of the Fisher–Rao metric on the space of smooth densities'. In: *Bulletin of the London Mathematical Society* 48.3, pp. 499–506.

Beik-Mohammadi, Hadi et al. (2021). 'Learning riemannian manifolds for geodesic motion skills'. In: *arXiv preprint arXiv:2106.04315*.

Bingham, Eli et al. (2019). 'Pyro: Deep universal probabilistic programming'. In: *The Journal of Machine Learning Research* 20.1, pp. 973–978.

Bourguin, Solesne and Thanh Dang (2022). 'High-dimensional regimes of non-stationary Gaussian correlated Wishart matrices'. In: *Random Matrices: Theory and Applications* 11.01, p. 2250006.

Brigant, Alice Le and Stéphane Puechmorel (2019). *The Fisher-Rao geometry of beta distributions applied to the study of canonical moments*. arXiv: 1904.08247 [math.ST].

Buntine, Wray L (1991). 'Bayesian backpropagation'. In: *Complex systems* 5, pp. 603–643.

Campbell, Neill D. F. and Jan Kautz (2014). 'Learning a Manifold of Fonts'. In: *ACM Trans. Graph.* 33.4. ISSN: 0730-0301. DOI: 10.1145/2601097.2601212. URL: https://doi.org/10.1145/2601097.2601212.

Cartan, Élie (1926). 'Les groupes d'holonomie des espaces généralisés'. In: *Acta Mathematica* 48, pp. 1–42.

Chaudhari, Pratik et al. (2019). 'Entropy-sgd: Biasing gradient descent into wide valleys'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124018.

Chen, Nutan et al. (2018a). *Active Learning based on Data Uncertainty and Model Sensitivity*. arXiv: 1808.02026 [stat.ML].

Chen, Nutan et al. (2018b). 'Metrics for deep generative models'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1540–1550.

Chentsov, Nikolai Nikolaevich (1982). 'Statiscal decision rules and optimal inference'. In: *Monog* 53.

Chern, Shiing-Shen (1996). 'Finsler geometry is just Riemannian geometry without the quadratic equation'. In: *Notices of the American Mathematical Society* 43.9, pp. 959–963.

Chern, Shiing-Shen and Zhongmin Shen (2005). *Riemann-finsler geometry*. Vol. 6. World Scientific Publishing Company.

Cramér, Harald (1946). *Mathematical methods of statistics*. Vol. 43. Princeton university press.

Davidson, Tim R. et al. (2018). 'Hyperspherical Variational Auto-Encoders'. In: *34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

Detlefsen, Nicki S, Martin Jørgensen, and Søren Hauberg (2019). 'Reliable training and estimation of variance networks'. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Detlefsen, Nicki S. et al. (2021). 'StochMan'. In:

Detlefsen, Nicki Skafte, Søren Hauberg, and Wouter Boomsma (2020). *What is a meaningful representation of protein sequences?* arXiv: 2012.02679 [q-bio.BM].

– (2022). 'Learning meaningful representations of protein sequences'. In: *Nature communications* 13.1, p. 1914.

Dinh, Laurent et al. (2017). 'Sharp minima can generalize for deep nets'. In: *International Conference on Machine Learning*. PMLR, pp. 1019–1028.

Eklund, David and Søren Hauberg (2019). 'Expected path length on random manifolds'. In: *arXiv preprint arXiv:1908.07377*.

Fefferman, Charles, Sanjoy Mitter, and Hariharan Narayanan (2016). 'Testing the manifold hypothesis'. In: *Journal of the American Mathematical Society* 29.4, pp. 983–1049.

Fefferman, Charles et al. (2019). 'Fitting a manifold of large reach to noisy data'. In: *arXiv preprint arXiv:1910.05084*.

Feldager, Cilie Werner (2022). 'Statistics under stochastic metrics'. PhD thesis. Technical University of Denmark.

Felice, Domenico and Nihat Ay (2021). 'Towards a canonical divergence within information geometry'. In: *Information geometry* 4.1, pp. 65–130.

Finsler, Paul (1918). *Ueber kurven und Flächen in allgemeinen Räumen*. Philos. Fak., Georg-August-Univ.

Fisher, Ronald A (1922). 'On the mathematical foundations of theoretical statistics'. In: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604, pp. 309–368.

Fletcher, Thomas (2011). 'Geodesic regression on Riemannian manifolds'. In: *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pp. 75–86.

Gallot, Sylvestre, Dominique Hulin, and Jacques Lafontaine (2004). 'Riemannian metrics'. In: Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 51–127.

Gallot, Sylvestre, Dominique Hulin, Jacques Lafontaine, et al. (1990). *Riemannian geometry*. Vol. 2. Springer.

Gauss, Carl Friedrich (1827). *Disquisitiones generales circa superficies curvas*. Vol. 1. Typis Dieterichianis.

Ghorbani, Behrooz, Shankar Krishnan, and Ying Xiao (2019). 'An investigation into neural net optimization via hessian eigenvalue density'. In: *International Conference on Machine Learning*. PMLR, pp. 2232–2241.

Girolami, Mark and Ben Calderhead (2011). 'Riemann manifold Langevin and Hamiltonian Monte Carlo methods'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2, pp. 123–214.

González-Duque, Miguel et al. (2022). 'Mario Plays on a Manifold: Generating Functional Content in Latent Space through Differential Geometry'. In: *2022 IEEE Conference on Games (CoG)*. IEEE, pp. 385–392.

Goodfellow, Ian et al. (2014). 'Generative Adversarial Nets'. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Guo, Guoji et al. (2010). 'Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst'. In: *Developmental Cell* 18.4, pp. 675–685. ISSN: 1534-5807. DOI: https://doi.org/10.1016/j.devcel.2010.02.012. URL: https://www.sciencedirect.com/science/article/pii/S1534580710001103.

Hauberg, Søren (2015). 'Principal Curves on Riemannian Manifolds'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

– (2018a). 'Only Bayes should learn a manifold (on the estimation of differential geometric structure from data)'. In: *arXiv preprint arXiv:1806.04994*.

– (2018b). *The non-central Nakagami distribution*. Tech. rep. Technical University of Denmark. URL: http://www2.compute.dtu.dk/~sohau/papers/nakagami2018/nakagami.pdf.

Hauberg, Søren, Oren Freifeld, and Michael J. Black (2012). 'A Geometric Take on Metric Learning'. In: *Advances in Neural Information Processing Systems (NeurIPS) 25*.

Hauberg, Søren (2019). *Only Bayes should learn a manifold (on the estimation of differential geometric structure from data)*. arXiv: 1806.04994 [stat.ML].

Hennig, Philipp and Søren Hauberg (2014). 'Probabilistic Solutions to Differential Equations and their Application to Riemannian Statistics'. In: *Proceedings of the 17th international Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 33.

Hilbert, David et al. (1900). 'Mathematical problems'. In: *Bulletin-American Mathematical Society* 37.4, pp. 407–436.

Hillen, Thomas et al. (Dec. 2016). 'Moments of von Mises and Fisher distributions and applications'. In: *Mathematical Biosciences and Engineering* 14, pp. 673–694. DOI: 10.3934/mbe.2017038.

Hinton, Geoffrey E and Drew Van Camp (1993). 'Keeping the neural networks simple by minimizing the description length of the weights'. In: *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). 'Flat minima'. In: *Neural computation* 9.1, pp. 1–42.

Imaizumi, Masaaki and Johannes Schmidt-Hieber (2022). 'On generalization bounds for deep networks based on loss surface implicit regularization'. In: *IEEE Transactions on Information Theory* 69.2, pp. 1203–1223.

Ito, Sosuke (2023). 'Geometric thermodynamics for the Fokker–Planck equation: stochastic thermodynamic links between information geometry and optimal transport'. In: *Information Geometry*, pp. 1–42.

Itoh, Mitsuhiro and Hiroyasu Satoh (2022). 'Information geometry of the space of probability measure and barycenter maps'. In: *arXiv preprint arXiv:2208.11861*.

Izmailov, Pavel et al. (2018). 'Averaging weights leads to wider optima and better generalization'. In: *arXiv preprint arXiv:1803.05407*.

Jastrzebski, Stanislaw et al. (2017). 'Three factors influencing minima in sgd'. In: *arXiv preprint arXiv:1711.04623*.

Javaloyes, Miguel Angel and Miguel Sánchez (2011). 'On the definition and examples of Finsler metrics'. In: *arXiv preprint arXiv:1111.5066*.

Jørgensen, Martin and Soren Hauberg (2021). 'Isometric gaussian process latent variable model for dissimilarity data'. In: *International Conference on Machine Learning*. PMLR, pp. 5127–5136.

Ju, Haotian et al. (2023). 'Generalization in Graph Neural Networks: Improved PAC-Bayesian Bounds on Graph Diffusion'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 6314–6341.

Kent, John T. and R. J. Muirhead (1984). 'Aspects of Multivariate Statistical Theory.' In: *The Statistician*. ISSN: 00390526. DOI: 10.2307/2987858.

Keskar, Nitish Shirish et al. (2016). 'On large-batch training for deep learning: Generalization gap and sharp minima'. In: *arXiv preprint arXiv:1609.04836*.

Khan, Gabriel and Jun Zhang (2022). 'When optimal transport meets information geometry'. In: *Information Geometry* 5.1, pp. 47–78.

Kingma, Diederik P and Jimmy Ba (2014). 'Adam: A method for stochastic optimization'. In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). 'Auto-encoding variational bayes'. In: *arXiv preprint arXiv:1312.6114*.

Kingma, Diederik P, Max Welling, et al. (2019). 'An introduction to variational autoencoders'. In: *Foundations and Trends® in Machine Learning* 12.4, pp. 307–392.

Kühnel, Wolfgang (2015). *Differential geometry*. Vol. 77. Theorem 6.5, page 247. American Mathematical Soc.

Kullback, Solomon and Richard A Leibler (1951). 'On information and sufficiency'. In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Laine, Samuli (2018). 'Feature-based metrics for exploring the latent space of generative models'. In:

Lawrence, Neil (2003). 'Gaussian process latent variable models for visualisation of high dimensional data'. In: *Advances in neural information processing systems* 16.

LeCun, Yann (1998). 'The MNIST database of handwritten digits'. In: *http://yann. lecun. com/exdb/mnist/*.

Lee, John M (2013). 'Smooth manifolds'. In: *Introduction to smooth manifolds*. Springer, pp. 1–31.

– (2018). *Introduction to Riemannian manifolds*. Vol. 2. Springer.

Liao, J. G. and Arthur Berg (2019). 'Sharpening Jensen's Inequality'. In: *The American Statistician* 73.3, pp. 278–281. DOI: 10.1080/00031305.2017.1419145.

Lopez, Federico et al. (2021). 'Symmetric Spaces for Graph Embeddings: A Finsler-Riemannian Approach'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 7090–7101. URL: https://proceedings.mlr.press/v139/lopez21a.html.

Lovas, Rezso L (2007). 'A note on Finsler-Minkowski norms'. In: *Houston J. Math* 33.3, pp. 701–707.

Loveridge, Lee C (2004). 'Physical and geometric interpretations of the Riemann tensor, Ricci tensor, and scalar curvature'. In: *arXiv preprint gr-qc/0401099*.

Mandt, Stephan, Matthew D Hoffman, and David M Blei (2017). 'Stochastic gradient descent as approximate bayesian inference'. In: *arXiv preprint arXiv:1704.04289*.

Markvorsen, Steen (2016). 'A Finsler geodesic spray paradigm for wildfire spread modelling'. In: *Nonlinear Analysis: Real World Applications* 28, pp. 208–228.

Martens, James (2014). 'New insights and perspectives on the natural gradient method'. In: *arXiv preprint arXiv:1412.1193*.

Martens, James and Roger Grosse (2015). 'Optimizing neural networks with kronecker-factored approximate curvature'. In: *International conference on machine learning*. PMLR, pp. 2408–2417.

Moreno-Muñoz, Pablo, Cilie Feldager, and Søren Hauberg (2022). 'Revisiting Active Sets for Gaussian Process Decoders'. In: *Advances in Neural Information Processing Systems* 35, pp. 6603–6614.

Nakagami, Minoru (1960). 'The m-distribution—A general formula of intensity distribution of rapid fading'. In: *Statistical methods in radio wave propagation*. Elsevier, pp. 3–36.

Neyshabur, Behnam et al. (2017). 'Geometry of optimization and implicit regularization in deep learning'. In: *arXiv preprint arXiv:1705.03071*.

Nielsen, Frank (2020). 'An elementary introduction to information geometry'. In: *Entropy* 22.10, p. 1100.

Noack, Marcus M and James A Sethian (2022). 'Advanced stationary and nonstationary kernel designs for domain-aware gaussian processes'. In: *Communications in Applied Mathematics and Computational Science* 17.1, pp. 131–156.

Nourdin, Ivan and Guangqu Zheng (2018). 'Asymptotic behavior of large Gaussian correlated Wishart matrices'. In: *Journal of Theoretical Probability*, pp. 1–30.

Orvieto, Antonio et al. (2022). 'Anticorrelated noise injection for improved generalization'. In: *International Conference on Machine Learning*. PMLR, pp. 17094–17116.

Otto, Felix (2001). 'The geometry of dissipative evolution equations: the porous medium equation'. In:

Paszke, Adam et al. (2019). 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Pennec, Xavier (2006). 'Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements'. In: *Journal of Mathematical Imaging and Vision*.

Pittorino, Fabrizio et al. (2021). 'Entropic gradient descent algorithms and wide flat minima'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12, p. 124015.

Pouplin, Alison et al. (2022). 'Identifying latent distances with Finslerian geometry'. In: *arXiv preprint arXiv:2212.10010*.

Pyro (2022). *Gaussian Process Latent Variable Model*. URL: https://pyro.ai/examples/gplvm.html (visited on 08/01/2022).

Randers, Gunnar (1941). 'On an asymmetrical metric in the four-space of general relativity'. In: *Physical Review* 59.2, p. 195.

Ranganath, Rajesh, Sean Gerrish, and David Blei (2014). 'Black Box Variational Inference'. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Rao, C Radhakrishna (1945). 'Information and the accuracy attainable in the estimation of statistical parameters'. In: *Breakthroughs in Statistics: Foundations and basic theory*. Springer, pp. 235–247.

Rasmussen, Carl Edward and Christopher K. I. Williams (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press. ISBN: 026218253X.

Ratliff, Nathan D. et al. (2021). 'Generalized Nonlinear and Finsler Geometry for Robotics'. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10206–10212. DOI: 10.1109/ICRA48506.2021.9561543.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). 'Stochastic backpropagation and approximate inference in deep generative models'. In: *International conference on machine learning*. PMLR, pp. 1278–1286.

Ricci, MMG and Tullio Levi-Civita (1900). 'Méthodes de calcul différentiel absolu et leurs applications'. In: *Mathematische Annalen* 54.1-2, pp. 125–201.

Riemann, Bernhard (1867). *Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse*. G. F. Teubner.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). 'Learning Representations by Back-propagating Errors'. In: *Nature* 323.

Scannell, Aidan, Carl Henrik Ek, and Arthur Richards (2021a). 'Trajectory optimisation in learned multimodal dynamical systems via latent-ode collocation'. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 12745–12751.

– (2021b). 'Trajectory Optimisation in Learned Multimodal Dynamical Systems Via Latent-ODE Collocation'. In: *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE.

Seong, Sihyeon et al. (2018). 'Towards Flatter Loss Surface via Nonmonotonic Learning Rate Scheduling.' In: *UAI*, pp. 1020–1030.

Shao, Hang, Abhishek Kumar, and P Thomas Fletcher (2018). 'The riemannian geometry of deep generative models'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 315–323.

Shen, Yi-Bing and Zhongmin Shen (2016). *Introduction to modern Finsler geometry*. World Scientific Publishing Company.

Tomczak, Jakub (2012). *Fisher information matrix for Gaussian and categorical distributions*. https://www.ii.pwr.edu.pl/~tomczak/PDF/[JMT]Fisher_inf.pdf. Online; accessed 17 Mai 2021.

Tosi, Alessandra et al. (2014). 'Metrics for probabilistic geometries'. In: *arXiv preprint arXiv:1411.7432*.

Tournier, Maxime et al. (2009). 'Motion compression using principal geodesics analysis'. In: *Computer Graphics Forum*. Vol. 28. 2. Wiley Online Library, pp. 355–364.

Uhlenbeck, George E and Leonard S Ornstein (1930). 'On the theory of the Brownian motion'. In: *Physical review* 36.5, p. 823.

Wei, Mingwei and David J Schwab (2019). 'How noise affects the hessian spectrum in overparameterized neural networks'. In: *arXiv preprint arXiv:1910.00195*.

Wendel, J. G. (1948). 'Note on the Gamma Function'. In: *The American Mathematical Monthly* 55.9, pp. 563–564. ISSN: 00029890, 19300972. URL: http://www.jstor.org/stable/2304460.

Wong, Ting-Kam Leonard and Jiaowen Yang (2022). 'Pseudo-Riemannian geometry encodes information geometry in optimal transport'. In: *Information Geometry* 5.1, pp. 131–159.

Wu, Bing Ye (2011). 'Volume form and its applications in Finsler geometry'. In: *Publ. Math. Debrecen* 78.3-4, pp. 723–741.

Xiao, Han, Kashif Rasul, and Roland Vollgraf (2017). 'Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms'. In: *arXiv preprint arXiv:1708.07747*.

Yang, Tao et al. (2018). 'Geodesic Clustering in Deep Generative Models'. In: *arXiv preprint*.

Yi, Mingyang et al. (2019). 'Positively scale-invariant flatness of relu neural networks'. In: *arXiv preprint arXiv:1903.02237*.

Zermelo, Ernst (1931). 'Über das Navigationsproblem bei ruhender oder veränderlicher Windverteilung'. In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 11.2, pp. 114–124.

Zhang, Miaomiao and Tom Fletcher (2013). 'Probabilistic principal geodesic analysis'. In: *Advances in Neural Information Processing Systems* 26, pp. 1178–1186.

Zhang, Shuofeng et al. (2021). 'Why flatness does and does not correlate with generalization for deep neural networks'. In: *arXiv preprint arXiv:2103.06219*.

Zhu, Zhanxing et al. (2018). 'The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects'. In: *arXiv preprint arXiv:1803.00195*.