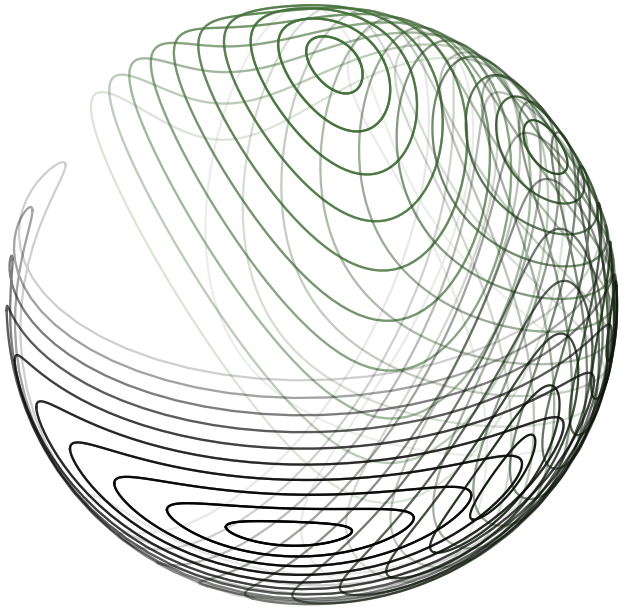
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



PhD thesis

Anton Mallasto

Geometric Methods in Probabilistic Modelling

Department of Computer Science

Advisors: Aasa Feragen, Tom Dela Haije and Mads Nielsen

Handed in: October 31, 2019

University of Copenhagen
Faculty of Science
Department of Computer Science

Geometric Methods in Probabilistic Modelling

Anton Mallasto

Supervisors: Aasa Feragen
Tom Dela Haije
Mads Nielsen

This thesis has been submitted to the PhD School of The Faculty of Science,
University of Copenhagen, on October 31, 2019

Abstract

During the past decade, machine learning has established itself as the foundation of artificial intelligence, viewing learning as a statistical task that can be well quantified. An increasingly popular toolkit in machine learning is provided by geometry, which is utilized in two main areas as we describe below: (i) studying the geometry induced by machine learning models; (ii) designing machine learning models that respect specified geometric properties of data.

First, most machine learning approaches quantify learning objectives by minimizing a loss function between a *model* and given data. The model can be *deterministic* or *probabilistic*, outputting *point predictions* or *stochastic predictions*, respectively. In contrast to deterministic models, probabilistic modelling allows us to carry out *uncertainty quantification*, which helps assess whether a prediction is trustworthy or not. The loss function often describes some kind of geometric similarity between the model and the data, and thus in the case of probabilistic modelling requires studying the geometry of probability measures. Two main approaches exist for this, *information geometry* and *optimal transport*, of which the latter is the main focus of this thesis.

Second, all data has structure. Sometimes, the data structure is not very restrictive, e.g., when the data lives in a vector space, in which case any appropriate vector could *a priori* be a data point. However, depending on the application, the data structure could be more restrictive. For example, if we are interested in location data on earth, then all possible data points have to lie approximately on a sphere. More generally, the data structure might be described as living on a *Riemannian manifold*. This restricts us to models that take the known geometry into account, which can be enforced by the machinery of *Riemannian geometry*, instead of linear algebra in the vector space case.

This thesis studies both realisations of geometry in machine learning. We start by considering *Gaussian processes*; on one hand, we study the optimal transport geometry between Gaussian processes, which can be utilized to define statistics with Gaussian process covariates. On the other hand, we generalize Gaussian process regression to Riemannian manifolds, thus allowing Gaussian processes to be used in statistics with variates living on manifolds. Finally, we also consider the *natural gradient* on statistical manifolds, and study optimal transport in relation to *Wasserstein generative adversarial networks*.

Resumé

Igennem det sidste årti har maskinlæring (machine learning) etableret sig som fundamentet for kunstig intelligens, hvor læring er et statistisk begreb der kan kvantificeres. Geometri er et remedie hvis popularitet er voksende inden for maskinlæring, hvilket bliver udnyttet inden for to hovedområder, som beskrives nedenfor: (i) studiet af geometrien induceret af maskinlæringsmodeller; (ii) designe maskinlæringsmodeller der respekterer specificerede geometriske egenskaber ved data.

For det første kvantificerer de fleste maskinlæringsmetoder læringsmål ved at minimere en tabsfunktion mellem en model og givent data. Modellen kan enten være deterministisk eller baseret på sandsynligheder, med henholdsvis punkt prædiktioner eller prædiktioner med støj. Sandsynlighedsbaserede modeller tillader os, i modsætning til deterministiske modeller, at udføre usikkerhedskvantificering som hjælper med at klassificere en prædiktion som troværdig eller ej. Tabsfunktionen beskriver ofte en form for geometrisk lighed mellem modellen og data og derfor, i tilfældet af sandsynlighedsbaserede modeller, er det nødvendigt at studere geometrien af sandsynlighedsmål. Der eksisterer to overordnede metoder til dette studie, informationsgeometri og optimal transport, af hvilke den sidste metode er afhandlingens hovedfokus.

For det andet har alt data struktur. Nogle gange er data strukturen ikke særlig restriktiv, for eksempel når data lever i et vektorrum. I et sådanne tilfælde vil enhver passende vektor a priori være et datapunkt. Dog, alt afhængigt af anvendelsen, kan datastrukturen være mere restriktiv. Hvis vi for eksempel er interesserede i placeringsdata på jorden, da vil alle de mulige datapunkter approksimativt ligge på en kugleoverflade. Mere generelt vil datastrukturen muligvis beskrives som tilhørende en Riemannsk mangfoldighed. Dette restringerer os til modeller der tager den kendte geometri i betragtning, hvilket kan håndhæves af Riemannsk geometri, i stedet for lineær algebra i tilfældet af et vektorrum.

Denne afhandling studerer begge realisationer af geometri i maskinlæring. Vi starter med at betragte Gaussiske processer; på den ene side studerer vi geometrien af optimal transport mellem Gaussiske processer, som kan udnyttes til at definere statistik med kovariater fra Gaussiske processer. På den anden side generaliserer vi Gaussisk process regression til Riemannske mangfoldigheder og derved tillade Gaussiske processer at blive benyttet i statistik med variater liggende på manfoldigheder. Sluttelig betragter vi også den naturlige gradient på statistiske mangfoldigheder og studerer optimal transport i relation til Wasserstein generative adversarial networks.

Acknowledgements

The three years I have spent working on this thesis have been truly formative, not just by teaching me how to be a part of the academic community, but also by maturing me as a person. The plethora of possibilities I have been given to meet brilliant minds all over the world has left me with a deep sense of gratitude.

I would like to start by giving thanks to my main supervisor, Aasa, and especially, to her patience with me. She taught me more lessons in science than I can count (so maybe I wasn't awake during the counting lessons), not just on the substance part, but even more importantly, on how to communicate my work to others. On top of being a mentor, you are also a very caring friend, always providing help when needed, even outside the academic world. On the note of thanking supervisors, I would also like to thank Mads for fruitful discussions that helped me when starting this project. Finally, I want to thank Tom for being a supervisor disguised as a friend, always being available for help, especially when it came to correcting my spelling mistakes.

This thesis was made possible by Center for Stochastic Geometry and Advanced Bioimaging (CSGB), not just by funding, but also by letting me be a part of a diverse research community, that provided perspective to my research.

A huge thanks goes to all my lovely colleagues and friends here in Copenhagen, especially to the ones who have tolerated my jokes of subpar quality for all this time. It is thanks to you, that I have been happy to call Copenhagen a home. At this point, I might have to apologize to my hosts for picking on your language and culture that much, but on the other hand, maybe I will skip that part, as a retaliation for going to miss being here more than I am willing to acknowledge.

My friends and family in Finland also deserve their thanks. I am very thankful to Miika and Maura for always cheering me up, and for their company that was the best remedy for the homesickness that developed during my stay abroad. Last but not least, such an adventure was made possible to be taken on carefree, thanks to the support and encouragement from my family.

Contents

Chapter 1

Introduction and Background

The thesis at hand is formatted as a compilation of the articles written during the author's enrollment in the PhD School of Science at University of Copenhagen. The structure is as follows: in this chapter, we begin by reviewing the essential mathematical background needed in the articles, which are presented in Appendices A–G.

The aim of the background is to provide the grounds for discussing the interplay between probabilistic modelling and geometry. There are two important ways that probability and geometry come together: we can either study the geometry of the space of probability measures, in order to define statistical tools for populations of uncertain data-objects represented by these measures; or we can study probabilistic models on data with geometric structure we wish to respect in a statistical pipeline.

Section 1.1 starts with introducing the most common probabilistic objects considered in this thesis – Gaussian processes (GPs) – which are considered in the work presented in Appendices A–C, where in Appendix A the *optimal transport* (OT) geometry between GPs is considered, and in Appendices B–C, GPs are generalized to Riemannian manifolds to be used as probabilistic models in learning tasks involving geometric data.

In Section 1.2, we discuss the OT framework, which geometrizes the space of probability measures. OT carries out this task by extending the geometry of the sample space, resulting from a cost function between two samples, to the space of probabilities, by computing a transportation plan that minimizes the total cost of transporting one measure to another. This is in contrast to *information geometry*, another popular approach to geometrizing the space of probability measures, which studies *divergences* that only considers the difference in mass assignment between two measures. There are two definite pros of OT over information geometry: the sample space geometry is taken into account, and actual distance metrics can be derived between probabilities. In addition to Appendix A, OT is considered in Appendix D, where an OT distance between *wrapped Gaussian distributions* is derived, and Appendices F–G, where OT defines a loss function in generative modelling, specifically in *Wasserstein generative adversial networks*. Summarizing, the applications of OT in these works fall into two categories: defining statistical machinery for data objects that are intrinsically noisy; or

using OT to define a loss function for learning tasks concerning probabilistic models.

In special cases, the geometry induced by optimal transport can be studied through the lens of *Riemannian geometry*. Not only that, the data objects considered might naturally live on a Riemannian manifold, in which case Riemannian geometry, which we discuss in Section 1.3, provides a standard toolkit for generalizing statistical methods from Euclidean spaces to be used with the geometrical data. Riemannian geometry can be found in all the works included in this thesis; In Appendices A–B, the data objects considered naturally live on a Riemannian manifold, and in the rest of the work, the geometry induced by OT is studied through the lens of Riemannian geometry.

After reviewing these basics, we discuss the nature of geometric data in more detail in Section 1.4. Especially, we focus on how we can tackle uncertainty quantification, by viewing uncertain data objects as elements of the manifold of probability measures. Finally, we consider the geometry in learning probabilistic models by first discussing *statistical manifolds* in Section 1.5, for which we consider generative modelling as a use case. Our aim in the use of statistical manifolds is to pull the geometry of probability measures to the space of parameters of the probabilistic models, which we take an advantage of in Appendix E, where the geometry is used to improve optimization in the learning of a model.

1.1 Gaussian Processes

The following discussion on Gaussian processes is loosely based on [56], which provides a thorough exposition for the use of GPs in machine learning.

A random variable X taking values in $\mathbb{R}^n$ has an n-dimensional multivariate Gaussian law with mean μ and covariance matrix K, if

$$\mathbb{P}\{X = x\} = ((2\pi)^n \det(K))^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)\right), \tag{1.1}$$

which we denote by $X \sim \mathcal{N}(\mu, K)$, and write $\mathcal{N}(x|\mu, K) = \mathbb{P}\{X = x\}$. Given two Gaussian random variables $X_i \sim \mathcal{N}(\mu_i, K_i)$, $i = 1, 2$, we say that they have a *joint Gaussian distribution*, if we can write

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} K_1 & K_{12} \\ K_{12}^T & K_2 \end{bmatrix}\right), \tag{1.2}$$

for some matrix K_{12}.

A *Gaussian process* (GP) can be viewed as an infinite-dimensional Gaussian distribution, defined as a collection f of random variables, such that any finite subcollection $(f(\omega_i))_{i=1}^N$ has a joint Gaussian distribution, where $\omega_i \in \Omega \subset \mathbb{R}^l$,

and Ω is the *index set*. A GP is entirely characterized by the pair

$$m(\omega) = \mathbb{E}\left[f(\omega)\right], \tag{1.3}$$
$$k(\omega, \omega') = \mathbb{E}\left[(f(\omega) - m(\omega))(f(\omega') - m(\omega'))^T\right], \tag{1.4}$$

where m and k are called the *mean function* and *covariance function*, respectively. We denote such a GP by $f \sim \mathcal{GP}(m, k)$. It follows from the definition that the covariance function (*kernel*) k is symmetric and positive semi-definite.

The importance of Gaussian processes for machine learning comes from our ability to express their conditional distributions in closed form. This fits the Bayesian statistics framework, where models are learned by conditioning a *prior distribution* on observed data, resulting in the *posterior* conditional distribution. As an example, we consider a regression task utilizing GPs. Let $\boldsymbol{D} = \{(x_i, y_i) \mid x_i \in \boldsymbol{x} \subset \mathbb{R}^l,\ y_i \in \boldsymbol{y} \subset \mathbb{R}^n,\ i = 1, ..., N\}$ be the training data, and assume the model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, ..., N \tag{1.5}$$

holds, where f is a relation between the *independent* variable x_i and the dependent variable y_i. Furthermore, ε_i is independent, identically distributed noise, given by $\varepsilon_i \sim \mathcal{N}(0, \varepsilon_{\text{err}}^2 I)$, where ε_{err} gives the variance of the noise model. Then, assuming a prior distribution $f \sim \mathcal{GP}(0, k)$, we can compute the *predictive distribution* p for the outputs $\boldsymbol{y}_*$ at the inputs $\boldsymbol{x}_*$, given in vector form, by

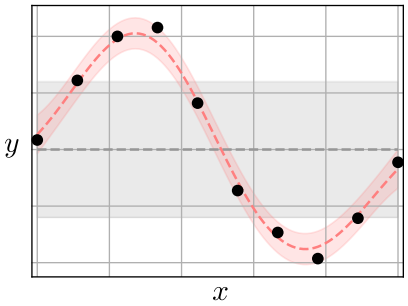


FIGURE 1.1: Gaussian process regression, with the prior distribution for f given in gray, which is then conditioned on the black data points to yield the red posterior distribution. The mean functions are given in dashed lines, and the variance is illustrated with shaded color.

$$p(\boldsymbol{y}_*|\boldsymbol{D}, \boldsymbol{x}_*) = \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*), \tag{1.6}$$
$$\boldsymbol{\mu}_* = \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{y}, \tag{1.7}$$
$$\Sigma_* = \boldsymbol{k}_{**} - \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{k}_*, \tag{1.8}$$

where we use the notation $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}, \boldsymbol{x}_*)$, $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ and K_{err} is the measurement error variance. In the notation above, the functions f and k are applied elementwise on the vectors $\boldsymbol{x}, \boldsymbol{x}_*$. This is known as *Gaussian process regression*, which is illustrated in Fig. 1.1.

Typically in model selection, the kernel k is picked from a parametric family $\{k_\theta | \theta \in \Theta\}$ of covariance functions, such as the *radial basis function* (RBF) kernels, of which a popular choice is the *Gaussian kernel*

$$k_{\sigma^2,\lambda}(x, y) = \sigma^2 \exp\left(-\frac{\|x - y\|^2}{2\lambda}\right), \ \sigma^2, \lambda > 0, \tag{1.9}$$

choosing the parameters $(\sigma^2, \sigma^2_{\text{err}}, \lambda)$ so that the *marginal likelihood* $\mathbb{P}\{\boldsymbol{y}|(\sigma^2, \sigma^2_{\text{err}}, \lambda)\}$ is maximized.

1.2 Optimal Transport

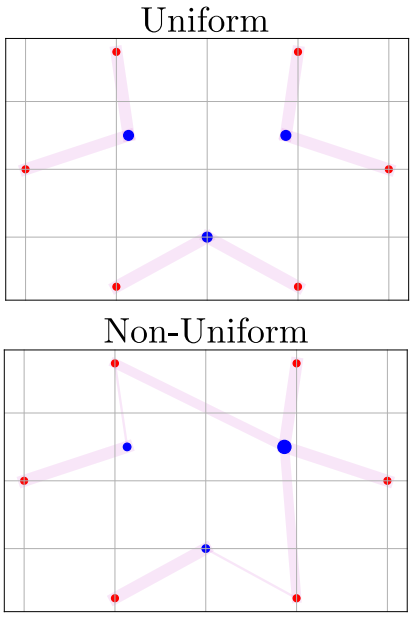


FIGURE 1.2: Optimal transport between the red and blue measures with assignments in magenta. Note how the mass distribution crucially affects the way mass is transported. In the figure above, both measures have uniform distributions, whereas in the figure below, the blue measure has non-uniform distribution.

Optimal transport forms an intriguing subfield of mathematics that concerns the geometry of probability distributions. The rich structure induced comes as quite a surprise, as the essence of optimal transport stems from a very elementary question: how to transport goods produced at given factories in the most economical way to satisfy the demand at given outlets? The answering the question requires specifying three factors: the distribution of production over the factories, the distribution of demand over the outlets, and the cost of transportation. An excellent resource of OT is given by Villani [62], and a reference for its computational aspects can be found in [53].

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces, that is, separable and complete metric spaces. Given a continuous and lower-bounded *cost function* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the optimal transport problem between two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ is defined as

$$\text{OT}_c(\mu, \nu) = \min_{\gamma \in \text{ADM}(\mu,\nu)} \mathbb{E}_\gamma[c], \tag{1.10}$$

where $\mathbb{E}_\mu[f] = \int_{\mathcal{X}} f(x) d\mu(x)$ is the expectation of a measurable function f with respect to μ, and the set $\text{ADM}(\mu, \nu)$ is defined to be the set of joint probability distributions of μ and ν. A γ minimizing (1.10) is called a *transport plan*.

Dual problem. (1.10) can be viewed as a linear program, as the objective is linear, and the constraints are given by projections of the transport plan γ onto its two marginals. As such, the program, admits a *dual formulation*. Denote by $L^1(\mu) = \{f : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}_\mu[f] < \infty\}$ the set of measurable functions of μ that have finite expectations under μ, and by $\text{ADM}(c)$ the set of *admissible pairs* $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ that for any $x, y \in \mathcal{X} \times \mathcal{Y}$ satisfy

$$\varphi(x) + \psi(y) \leq c(x, y). \tag{1.11}$$

Then, the following duality holds [62, Sec. 5]

$$\mathrm{OT}_c(\mu,\nu) = \sup_{(\varphi,\psi)\in \mathrm{ADM}(c)} \left\{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\psi]\right\}. \tag{1.12}$$

When the supremum is attained, the optimal φ^*, ψ^* in (1.12) are called *Kantorovich potentials*, which, in particular, satisfy $\varphi^*(x) + \psi^*(y) = c(x,y)$ for any $(x,y) \in \mathrm{Supp}(\gamma^*)$, where γ^* solves (1.10). Given φ, we can obtain ψ satisfying (1.12) through the *c-transform* of φ,

$$\varphi^c : \mathcal{Y} \to \mathbb{R}, \quad y \mapsto \inf_{x\in\mathcal{X}} \left\{c(x,y) - \varphi(x)\right\}. \tag{1.13}$$

Moreover, the Kantorovich potentials satisfy $\psi^* = (\varphi^*)^c$, and therefore (1.12) can be written as [62, Thm. 5.9]

$$\mathrm{OT}_c(\mu,\nu) = \max_{(\varphi,\varphi^c)\in \mathrm{ADM}(c)} \left\{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\varphi^c]\right\}. \tag{1.14}$$

In other words, the $\mathrm{ADM}(c)$ constraints given in (1.11) can be enforced with the c-transform, and reduces the optimization in (1.14) to be carried over a single function.

Monge formulation. Optimal transport was originally introduced by the french mathematician Gaspard Monge in his work *Mémoire sur la théorie des déblais et des remblais* [46]. The formulation by Monge differs from (1.10), which resulted from the soviet mathematician Leonid Kantorovich relaxing Monge's formulation. This relaxation was the birth of *linear programming*, which subsequently brought Kantorovich the Nobel prize in economics.

Instead of a *transport plan* representing a joint distribution between the two margins μ and ν, Monge considered a *transport map* T^* satisfying

$$T^* = \arg\min_T \int c(x, T(x)) d\mu(x), \tag{1.15}$$

where the T *pushes forward* μ to ν, denoted by

$$\nu = T_\# \mu(A) := \mu(T^{-1}(A)), \tag{1.16}$$

for any measurable set $A \subset \mathcal{Y}$. Assuming μ or ν is absolutely continuous, the two formulations 1.10 and 1.15 are equivalent. As the Monge formulation is more restrictive, a minimizer is not guaranteed. However, when a minimizing map exists, the two formulations are equivalent. In practice, having access to a transport map is appreciated, as then no *mass splitting* occurs when interpolating between two measures under the OT framework.

Wasserstein Metric. Consider the case when $\mathcal{X} = \mathcal{Y}$ and $c(x,y) = d^p_{\mathcal{X}}(x,y)$, $p \geq 1$, where we refer to $d_{\mathcal{X}}$ as the *ground metric*. Then, the optimal transport

problem (1.10) defines the *p-Wasserstein metric* $W_p(\mu,\nu) := \mathrm{OT}_{d_{\mathcal{X}}^p}(\mu,\nu)^{\frac{1}{p}}$ on

$$\mathcal{P}^p_{d_{\mathcal{X}}}(X) = \left\{ \mu \in \mathcal{P}(X) \,\middle|\, \int d^p_{\mathcal{X}}(x_0,x) d\mu(x) < \infty \right\}, \quad \text{for some } x_0 \in \mathcal{X}, \tag{1.17}$$

that is the space of probability measures with finite p-moments. It can be shown that $(\mathcal{P}^p_{d_{\mathcal{X}}}(\mathcal{X}), W_p)$ forms a complete, separable metric space [62, Sec. 6, Thm 6.16]. Specifically, in the case $p = 2$, the metric space $(\mathcal{P}^2_{d_{\mathcal{X}}}(\mathcal{X}), W_2)$ can be endowed with a weak Riemannian metric inducing W_2, which we will describe in detail in Sec. 1.3.1. When $p \neq 2$, an associated Finsler structure can be found, instead, as shown by Agueh [3].

1.3 Riemannian Geometry

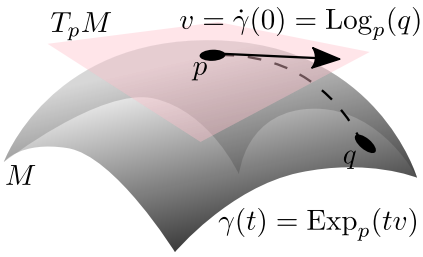


FIGURE 1.3: Illustration of the logarithmic and exponential map on a Riemannian manifold M. The geodesic associated with (p, v) is given by the dashed curve.

As mentioned above, the principal geometry considered in this thesis is the *Riemannian geometry* [16, 39], which studies *smooth manifolds* endowed with a *Riemannian metric*. The metric provides a smoothly changing local inner product, which allows us to compare tangent vectors in a fixed tangent space to each other. Furthermore, the metric induces a distance function on the manifold itself, telling us how far apart points are. A *Riemannian manifold* is a smooth n-dimensional manifold $\mathcal{M}$ with a Riemannian metric $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$, which gives an inner product at each tangent space $T_p\mathcal{M}$, $p \in \mathcal{M}$. A frame $E(p) = (e_1(p), ..., e_n(p))$ is an orthonormal basis for $T_p\mathcal{M}$, that varies smoothly with p. Such a frame allows us to define the *Riemannian metric tensor* G_p as $g_p(u,v) = u^T G_p v$, for any $u, v \in T_p\mathcal{M}$. The inner product also allows us to define a *norm* at each tangent space $T_p\mathcal{M}$ by

$$\|v\|_p^2 = g_p(v,v)^2, \; v \in T_p\mathcal{M}. \tag{1.18}$$

The Riemannian metric also induces a metric distance function d between points $p, q \in \mathcal{M}$ by

$$d_{\mathcal{M}}(p,q) = \inf_{\gamma} \left\{ \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)} dt \right\}, \tag{1.19}$$

where $\gamma \in C^1([0,1], \mathcal{M})$, satisfying

$$\begin{aligned} \gamma(0) &= p, \\ \gamma(1) &= q, \\ d_{\mathcal{M}}(\gamma(s), \gamma(t)) &= |s - t|, \end{aligned} \tag{1.20}$$

and $\dot{\gamma}(t) = \frac{d}{dt}\gamma(t)$. A curve γ achieving the infimum is called a *geodesic* between p and q. In general, a geodesic is a curve on $\mathcal{M}$ that locally minimizes the distance between its points, which generalizes straight lines to Riemannian manifolds.

A relation between geodesics and elements of the tangent bundle is given as follows: each element (p, v) in the tangent bundle $T\mathcal{M} = \bigcup_{p\in\mathcal{M}} (p \times T_p\mathcal{M})$ defines a geodesic γ on $\mathcal{M}$, so that $\gamma(0) = p$ and $\dot{\gamma}(0) = v$. This connection allows us to define the basic algebraic tools needed for defining machine learning on manifolds: the exponential and logarithmic maps, which generalizes addition and subtraction, respectively, to Riemannian manifolds. The *exponential map* is defined as

$$\begin{aligned} \mathrm{Exp} : T\mathcal{M} &\to \mathcal{M}, \\ (p, v) &\mapsto \mathrm{Exp}_p(v) = \gamma(1), \end{aligned} \tag{1.21}$$

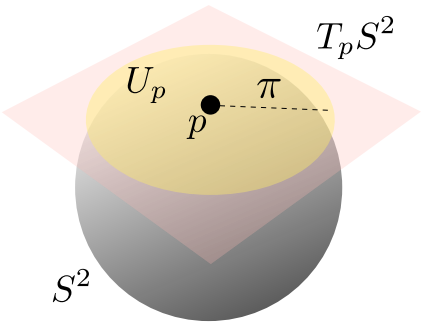


FIGURE 1.4: At any point p on the 2-sphere S^2, the injectivity radius is π, with area of injectivity consisting of all but the antipodal points point.

where γ is the geodesic corresponding to (p, v). The exponential map Exp_p at p is a diffeomorphism between a neighborhood $0 \in U_p \subset T_p\mathcal{M}$ and neighbourhood $p \in V_p \subset \mathcal{M}$, which is chosen in a maximal way, so if $V_p \subsetneq V_p'$, then a diffeomorphism between V_p' and a neighborhood in the tangent space cannot be defined anymore. We also call V_p the *area of injectivity* at p, and the minimum distance from p to the boundary of a maximal V_p is the *injectivity radius of* Exp_p. The inverse of the exponential map is the *logarithmic map*

$$\mathrm{Log}_p : V_p \to T_p\mathcal{M}, \qquad q \mapsto \mathrm{Log}_p(q) = \dot{\gamma}(0), \tag{1.22}$$

which we illustrate together with the exponential map in Fig. 1.3.

Related to the area of injectivity, is the *cut-locus*. The *cut-locus of p in $T_p\mathcal{M}$* is defined as the set of elements $v \in T_p\mathcal{M}$, so that $\gamma(t) = \mathrm{Exp}_p(tv)$ is a minimizing geodesic for $t \in [0, 1]$, but fails to be minimizing for $t > 1$. The *cut-locus of p in $\mathcal{M}$* is the image of the cut-locus in $T_p\mathcal{M}$ under the exponential map, denoted by $\mathcal{C}_p$. The manifolds with non-positive curvature form an important class of manifolds with infinite injectivity radius, which also translates into them having an empty cut-locus $\mathcal{C}_p$ for every $p \in \mathcal{M}$.

Using a frame $E(p) = (e_1(p), ..., e_n(p))$, that induces a basis on the tangent space $T_p\mathcal{M}$, yields a chart on V_p, called the *normal coordinate chart*, through the exponential map as

$$x \mapsto \mathrm{Exp}_p\left(\sum_{i=1}^{n} x^i e_i(p)\right). \tag{1.23}$$

The normal coordinate chart is commonly used in applications. One reason for this is its *radial isometry*, resulting from Gauss's lemma, which gives us the

identity

$$d^2_{\mathcal{M}}(p, \mathrm{Exp}_p(v)) = \|v\|^2_p. \tag{1.24}$$

Euclidean space as a Riemannian manifold. Riemannian geometry generalizes Euclidean geometry, as the Euclidean space $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ forms a Riemannian manifold. First, note that $\mathbb{R}^n$ is homeomorphic to itself, an open Euclidean space, and thus forms a smooth manifold. Secondly, the inner product

$$\langle v, u \rangle_p = v^T u, \quad v, u \in T_p\mathbb{R}^n, \tag{1.25}$$

varies smoothly with any $p \in \mathbb{R}^n$. Under this metric, the exponential map is given by

$$\mathrm{Exp}^{\mathrm{Euc}}_x(v) = p + v, \tag{1.26}$$

and the logarithmic map by

$$\mathrm{Log}^{\mathrm{Euc}}_p(q) = q - p. \tag{1.27}$$

1.3.1 Weak Riemannian Structure of the 2-Wasserstein metric

In Section 1.2 we mentioned how the metric space $\left(\mathcal{P}^2_{d_{\mathcal{X}}}(\mathcal{X}), W_2\right)$ is induced by a weak Riemannian structure (weak, as $\mathcal{P}^2_{d_{\mathcal{X}}}(\mathcal{X})$ does not form a smooth manifold). This viewpoint is commonly referred to as *Otto calculus* [48], which we will describe now in more detail based on [8, Sec. 2.3.2] and [17]. For simplicity, we look at $\mathrm{AC}(\mathbb{R}^n)$, the space of absolutely continuous distributions on $(\mathbb{R}^n, d_{\mathrm{Euc}})$, where d_{Euc} is the standard Euclidean distance. Then, what we mean by the Riemannian structure, stems from the following *Benamou–Brenier* [12] formulation (also known as the dynamic formulation) of optimal transport, which states

$$W_2^2(\mu, \nu) = \inf_{(\mu_t, \Phi_t)} \left\{ \int_0^1 \|\nabla \Phi_t\|^2_{\mu_t} dt \right\}, \tag{1.28}$$

where $\|\nabla \Phi_t\|^2_\mu = \int \|\nabla \Phi_t(x)\|^2_{\mathrm{Euc}} d\mu(x)$, and $(\mu_t, \Phi_t) \in \mathrm{AC}(\mathbb{R}^n) \times C^\infty(\mathbb{R}^n)$ is a weak solution to the *continuity equation*

$$\frac{d}{dt}\mu_t + \nabla \cdot (\mu_t \nabla \Phi_t) = 0, \quad t \in [0, 1], \quad \mu_0 = \mu, \quad \mu_1 = \nu. \tag{1.29}$$

In (1.29), $\frac{d}{dt}\mu_t$ gives the rate of change of mass at each point at time t, and Φ_t is a potential field, such that $\nabla \Phi_t$ is a velocity field giving the transfer of mass at each point. At each point $\mu \in \mathrm{AC}(\mathbb{R}^n)$, the tangent space is given by

$$T_\mu \mathrm{AC}(\mathbb{R}^n) = \left\{ v \in C^\infty(\mathbb{R}^n) \ : \ \int v(x) dx = 0 \right\}, \tag{1.30}$$

and so $\frac{d}{dt}\mu_t \in T_\mu \mathrm{AC}(\mathbb{R}^n)$. Now, as a shorthand notation, when (μ, Φ) is fixed, we denote by $V_{(\mu,\nabla\Phi)}$ the tangent space element solving

$$V_{(\mu,\nabla\Phi)} + \nabla \cdot (\mu \nabla \Phi) = 0. \tag{1.31}$$

Then, the continuity equation gives an isomorphism between the space of smooth real-valued functions on $\mathbb{R}^n$, modulo additive constants, and the tangent space, by

$$C^\infty(\mathbb{R}^n)/\mathbb{R} \to T_\mu \mathrm{AC}(\mathbb{R}^n), \quad \Phi \mapsto V_{(\mu,\nabla\Phi)}. \tag{1.32}$$

Due to this isomorphism, we denote $T^*_\mu \mathrm{AC}(\mathbb{R}^n) = C^\infty(\mathbb{R}^n)/\mathbb{R}$.

Now, (1.28) can be seen as the distance induced by the Riemannian metric

$$g_\mu(v_1, v_2) = \int \langle \nabla \Phi_{v_1}, \nabla \Phi_{v_2} \rangle d\mu, \tag{1.33}$$

where Φ_{v_i} solves the continuity equation

$$v_i + \nabla \cdot (\mu \nabla \Phi_{v_i}) = 0, \ i = 1, 2, \tag{1.34}$$

such that $\nabla \Phi_{v_i}$ is of minimal norm of all the possible solutions to (1.34). This induction can be seen, as

$$\begin{aligned} W_2^2(\mu, \nu) &= \inf_{(\mu_t, \Phi_t)} \left\{ \int_0^1 \|\nabla \Phi_t\|^2_{\mu_t} dt \right\} \\ &= \inf_{\mu_t} \left\{ \int_0^1 g_{\mu_t} \left(\nabla \Phi_{v_t}, \nabla \Phi_{v_t} \right) dt \right\}, \end{aligned} \tag{1.35}$$

where (μ_t, Φ_t) has to satisfy (1.29), and $v_t = \frac{d}{dt}\mu_t$.

Let Id denote the identity map on $\mathbb{R}^n$. Then, under this metric, the exponential map is defined as

$$\mathrm{Exp}_\mu(v) = (\mathrm{Id} + \nabla \Phi_v)_{\#} \mu, \tag{1.36}$$

and the logarithmic map is given by

$$\mathrm{Log}_\mu(\nu) = V_{(\mu, T^\nu_\mu - \mathrm{Id})} \tag{1.37}$$

where T^ν_μ is the transport map from μ to ν (which exists as the measures are absolutely continuous).

Remark 1. As can be seen in expressions (1.33), (1.36) and (1.37), the metric structure acts on elements of $T_\mu \mathrm{AC}(\mathbb{R}^N)$ through the gradients of elements of $T^*_\mu \mathrm{AC}(\mathbb{R}^N)$. Due to this, another convention is to define the L^2 closure of

$$\left\{ \nabla \Phi \ : \ \Phi \in T^*_\mu \mathrm{AC}(\mathbb{R}^N) \right\}, \tag{1.38}$$

as the tangent space at μ.

Gaussian example. Denote by $\mathcal{N}(\mathbb{R}^n)$ the space of non-degenerate multivariate Gaussian distributions on $\mathbb{R}^n$. Its 2-Wasserstein geometry as a Riemannian manifold was studied in [42, 61], which we will summarize below. After the

summary, we also provide a novel derivation as an example of the Otto calculus discussed above.

The 2-Wasserstein metric between $X_i \sim \mathcal{N}(\mu_i, K_i) \in \mathcal{N}(\mathbb{R}^n)$, $i = 1, 2$, is given by

$$W_2^2(X_1, X_2) = \|\mu_1 - \mu_2\|_{\text{Euc}}^2 + \text{Tr}K_1 + \text{Tr}K_2 - 2\text{Tr}\left(K_1 K_2\right)^{\frac{1}{2}}, \tag{1.39}$$

where $\text{Tr}K = \sum_{i=1}^n K_{ii}$ denotes the *trace* of K. Based on (1.39), we have the isometry

$$\mathcal{N}(\mathbb{R}^n) \simeq L^2(\mathbb{R}^n) \times \mathcal{N}_0(\mathbb{R}^n) \tag{1.40}$$

where $\mathcal{N}_0(\mathbb{R}^n)$ is the space of centered Gaussian distributions. For simplicity, we can view the tangent space of $\mathcal{N}_0(\mathbb{R}^n)$ (seen as the space of symmetric positive definite matrices) as the space of symmetric matrices $\text{Sym}(\mathbb{R}^n)$, that is,

$$T_\mu \mathcal{N}_0(\mathbb{R}^n) \simeq \text{Sym}(\mathbb{R}^n), \quad \mu \in \mathcal{N}_0(\mathbb{R}^n). \tag{1.41}$$

Then, the distance (1.39) is induced by

$$g_{\mathcal{N}(0,K)}(V, U) = \text{Tr}(v_{(K,V)} K v_{(K,U)}), \tag{1.42}$$

and the exponential and logarithmic maps are given by

$$\begin{aligned} \text{Log}_{K_1}(K_2) &= K_1 T_1^2 + T_1^2 K_1 - I, \\ \text{Exp}_K(V) &= (I + v_{(K,V)}) K (I + v_{(K,V)}), \end{aligned} \tag{1.43}$$

where $v_{(K,V)}$ is the unique symmetric matrix solving the *Sylvester equation*

$$V = K v_{(K,V)} + v_{(K,V)} K, \tag{1.44}$$

and T_1^2 is the Monge transport map between $\mathcal{N}(0, K_1)$ and $\mathcal{N}(0, K_2)$, given by

$$T_1^2 = K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}}\right)^{\frac{1}{2}} K_1^{-\frac{1}{2}}, \tag{1.45}$$

Now, we will derive this structure. Note, that $\mathcal{N}_0(\mathbb{R}^n)$ is isomorphic to $\text{PD}(\mathbb{R}^n)$, the space of n-by-n symmetric positive definite matrices, through

$$\varphi \colon K \mapsto \mathcal{N}(0, K). \tag{1.46}$$

This isomorphism induces on $\text{PD}(\mathbb{R}^n)$ the metric

$$W_2^2(K_1, K_2) = \text{Tr}K_1 + \text{Tr}K_2 - 2\text{Tr}\left(K_1 K_2\right)^{\frac{1}{2}}. \tag{1.47}$$

As the Euclidean geometry of $L^2(\mathbb{R}^n)$ is trivial, we focus on the geometry of $(\text{PD}(\mathbb{R}^n), W_2)$.

First, recall that $\text{PD}(\mathbb{R}^n)$ forms an open cone of $\text{Sym}(\mathbb{R}^n)$, which is a Euclidean space. An immediate consequence is that $\text{PD}(\mathbb{R}^n)$ is a smooth manifold, and

shares the same tangent space with $\mathrm{Sym}(\mathbb{R}^n)$, i.e.,

$$T_K\mathrm{PD}(\mathbb{R}^n) = T_K\mathrm{Sym}(\mathbb{R}^n) = \mathrm{Sym}(\mathbb{R}^n), \quad \forall K \in \mathrm{PD}(\mathbb{R}^n). \tag{1.48}$$

Now, we wish to pullback the Riemannian metric $g^{\mathcal{N}_0(\mathbb{R}^n)}$ inducing W_2 on $\mathcal{N}_0(\mathbb{R}^n)$ to $\mathrm{PD}(\mathbb{R}^n)$. To do this, we define

$$g_K^{\mathrm{PD}(\mathbb{R}^n)}(V,U) = g_{\mathcal{N}(0,K)}^{\mathcal{N}_0(\mathbb{R}^n)}\left([D_K\varphi]\,V, [D_K\varphi]\,U\right), \tag{1.49}$$

where $D_K\varphi$ is the differential of φ at K, acting on any $V \in T_K\mathrm{PD}(\mathbb{R}^n)$ by

$$\left([D_K\varphi]\,V\right)(x) = \frac{1}{2}\left(-\mathrm{Tr}\left(K^{-1}V\right) + x^TK^{-1}VK^{-1}x\right)\mathcal{N}(x \mid 0, K), \tag{1.50}$$

Then, we find that

$$([D_K\varphi]V)(x) + \nabla \cdot \left(\mathcal{N}(x \mid 0, K)v_{(K,V)}(x)\right) = 0, \quad \forall x \in \mathbb{R}^n, \tag{1.51}$$

is solved by the linear map $v_{(K,V)}$ with the associated matrix being the unique symmetric element solving (1.44).

Thus, using (1.33), we can write (1.49) as

$$\begin{aligned} g_K^{\mathrm{PD}(\mathbb{R}^n)}(V,U) &= \int_{\mathbb{R}^d} \langle v_{(K,V)}, v_{(K,U)}\rangle \mathcal{N}(x \mid 0, K)dx \\ &= \mathrm{Tr}\left(v_{(K,V)}Kv_{(K,U)}\right). \end{aligned} \tag{1.52}$$

Now, for the exponential and logarithmic maps, recall the Monge transport map between $\mathcal{N}(0,K_1)$ and $\mathcal{N}(0,K_2)$ given in (1.45), and that pushing $\mathcal{N}(0,K)$ forward by a linear map A results in

$$A_{\#}\mathcal{N}(0,K) = \mathcal{N}(0, AKA^T). \tag{1.53}$$

This allows us to write the exponential map in (1.36) as

$$\mathrm{Exp}_K(V) = (I + v_{(K,V)})K(I + v_{(K,V)}), \tag{1.54}$$

and the logarithmic map in (1.37) as

$$\mathrm{Log}_{K_1}(K_2) = K_1T_1^2 + T_1^2K_1 - I, \quad K_1, K_2 \in \mathrm{PD}(\mathbb{R}^n). \tag{1.55}$$

Remark 2. We could as well have associated a Gaussian distribution with its precision matrix, i.e.,

$$\varphi : K \mapsto \mathcal{N}(0, K^{-1}), \quad K \in \mathrm{PD}(\mathbb{R}^n). \tag{1.56}$$

In this case, the resulting metric, following from a similar derivation as above, would be

$$g_K^{\mathrm{PD}(\mathbb{R}^n)}(V,U) = \mathrm{Tr}\left(v_{(K,-V)}K^{-1}v_{(K,-U)}\right). \tag{1.57}$$

Remark 3. The derivation in [61] follows the convention mentioned in Remark 1, and thus the metric provided there is given by

$$g_K(V,U) = \mathrm{Tr}(VKU), \quad K \in \mathrm{PD}(\mathbb{R}^n),\ V,U \in \mathrm{Sym}(\mathbb{R}^n). \tag{1.58}$$

1.4 Statistics of Geometric Data

Much of statistics is based on the premise of the data living in a *Euclidean space* $\mathbb{R}^n$, which allows the utilization of its vector space properties of being closed under addition and scalar multiplication. In modern data analysis, a plethora of examples of data with non-Euclidean structure exists. Essentially what this means, is that the data cannot be well represented by such a vector in $\mathbb{R}^n$. In many such cases, the data instead lives on a smooth manifold $\mathcal{M}$, which then consists of all the plausible data points. Such examples include *shapes* (that are invariant with respect to translation, scaling and rotation), *diffusion tensors* in magnetic resonance imaging (MRI) (that can be modelled as positive semi-definite matrices), and *directional data* (that lives on the unit sphere S^{n-1}).

On top of the spatial constraints provided by $\mathcal{M}$, we often want to relate elements of $\mathcal{M}$ to each other in a way that further allows us to define certain statistical tools. These include for example the definition of the mean of a population, and the ability to interpolate between data points, and this is typically achieved through the definition of a metric d on $\mathcal{M}$. The chosen metric has a significant impact on the resulting statistics, and is thus an important part of the modelling choices done prior to statistical analysis. Notably, one might want to consider statistical models that are invariant with respect to given transformations, which can be incorporated via an invariant metric. For example, the distance between two shapes should remain invariant with respect to translations, scalings and rotations. In this thesis, we will consider metrics d induced by Riemannian metrics g, and so our data space forms a Riemannian manifold $(\mathcal{M}, g)$. Statistics on manifolds can be divided into *extrinsic* and *intrinsic* approaches.

1.4.1 Extrinsic Statistics on Manifolds

According to the Whitney embedding theorem, any n-dimensional smooth manifold can be smoothly embedded into $\mathbb{R}^{2n}$, which allows the manifold to be considered in a global coordinate system. This then allows one to utilize Euclidean statistical analysis in the *ambient space*, the result of which can then be projected back onto the embedded manifold, by mapping to the nearest point on the manifold. This is called the *extrinsic* approach. The extrinsic approach is straightforward and computationally favorable, however, as describe above, it fails to take the *intrinsic* geometry into account. This can be fixed, by considering the Nash embedding theorem, which states that any Riemannian manifold can be *isometrically* embedded into $\mathbb{R}^m$, for some $m \leq \frac{1}{2}n(n+1)(3n+11)$. Unfortunately, computing the Nash embedding is usually infeasible in practice.

The extrinsic approach [10, 13] is straightforward and computationally favorable [14], however, it fails to take the *intrinsic* geometry into account, as the Riemannian metric is entirely ignored, and as the Euclidean models do not take the manifold geometry into account. For example, consider a population of elements $\{p_i\}_{i=1}^N$ on a Riemannian manifold $\mathcal{M}$ with an embedding $\psi : \mathcal{M} \to \mathbb{R}^{2n}$. Then, one could compute the mean $\bar{p}$ of this population as

$$\bar{p} = \psi^{-1}\left(\pi_{\psi(\mathcal{M})}\left(\frac{1}{N}\sum_{i=1}^{N}\psi(p_i)\right)\right), \tag{1.59}$$

where $\pi_{\psi(\mathcal{M})} \colon \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is the projection onto the embedded manifold, given by

$$\pi_{\psi(\mathcal{M})} : x \mapsto \underset{y\in\psi(\mathcal{M})}{\arg\min}\, d^2_{\text{Euc}}(x, y). \tag{1.60}$$

To further illustrate this example, assume $\mathcal{M} = S^{n-1}$, then the embedded mean would be given by

$$\psi(\bar{p}) = \frac{\sum_{i=1}^{N}\psi(p_i)}{\|\sum_{i=1}^{N}\psi(p_i)\|}. \tag{1.61}$$

As can be seen, this approach fails to take into account the metric g on $\mathcal{M}$, unless the embedding is *isometric*. However, in practice, computing an isometric embedding is highly non-trivial, and therefore not considered here.

1.4.2 Intrinsic Statistics on Manifolds

Intrinsic statistics on manifolds [51] is defined independently of charts (coordinate systems) using the machinery of differential geometry. In practice, the normal coordinate charts given by the exponential map are commonly used for computing the statistics. This approach allows one to completely ignore the ambient space, as the statistical models only see the manifold, nothing else. Furthermore, the intrinsic approach allows us to employ the chosen Riemannian metric in the design of models. In this thesis, we focus on the intrinsic approach.

Generalizing statistics from the linear world to Riemannian manifolds poses many challenges, due to nonlinearity and changes in topology. However, a principled approach exists, where the Euclidean geometry in statistical machinery is replaced with the Riemannian geometry. Below, we demonstrate this in a couple of examples, to give an overview. For the interested reader, more details about regression on manifolds can be found in [11, 23, 34], submanifold learning in [24, 33, 35], and probabilistic methods in [25, 55, 58].

Population mean. Consider the mean element of a population of elements $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$ in a Euclidean space, which is defined as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.62}$$

Note, how in the definition scalar multiplication and addition are used, which are not available in the Riemannian setting. Instead, we want to express $\bar{x}$ geometrically, which can be done by noticing that $\bar{x}$ is the unique element satisfying

$$\bar{x} = \arg\min_{x \in \mathbb{R}^n} \sum_{i=1}^{N} d^2_{\text{Euc}}(x, x_i), \tag{1.63}$$

where $d_{\text{Euc}}(x, x_i) = \|x - x_i\|$ is the standard Euclidean distance in $\mathbb{R}^n$. Now consider a population of elements $\{p_i\}_{i=1}^N \subset \mathcal{M}$ on a Riemannian manifold. Substituting the Euclidean distance d_{Euc} with the distance $d_{\mathcal{M}}$ on the Riemannian manifold in (1.63) results in the definition of the *Fréchet mean* [26, 31] $\bar{p}$, which satisfies

$$\bar{p} \in \arg\min_{p \in \mathcal{M}} \sum_{i=1}^{N} d^2_{\mathcal{M}}(p, p_i). \tag{1.64}$$

Note that in the Riemannian case, the minimizers need not be unique, although it is unique in special cases such as non-positively curved manifolds.

In practice, we can numerically approximate the Fréchet mean by gradient descent, using the identity

$$\nabla_p d^2_{\mathcal{M}}(p, q) = -2\text{Log}_p(q). \tag{1.65}$$

Where in the Euclidean case the gradient descent updates are given by

$$x \leftarrow x - \lambda v, \tag{1.66}$$

where v is the gradient and λ the learning rate, in the Riemannian case one needs to replace the addition with the Riemannian exponential

$$p \leftarrow \text{Exp}_p(-\lambda v). \tag{1.67}$$

Residuals. A common way to define an error function in statistics is through the residuals, which give the discrepancy between a model and observations: given a data point (x_i, y_i), where $y_i \in \mathbb{R}^n$, and a predictive model f, the residual is defined as $r_{\text{Euc}}(x_i) = y_i - f(x_i)$. In the Riemannian case such a subtraction is not possible, however, we are able to substitute addition and subtraction with other operations, as we did the in the gradient descent update. For a data point (x_i, p_i) with $p_i \in \mathcal{M}$, we can define the Riemannian residual as $r_{\mathcal{M}}(x_i) = \text{Log}_{f(x_i)}(p_i)$ [23]. This generalization is actually quite straightforward: the Euclidean space $\mathbb{R}^n$ with the Euclidean metric $\langle \cdot, \cdot \rangle$ also forms a Riemannian manifold as detailed in Section 1.3, and in particular, $r_{\text{Euc}}(x_i) = \text{Log}^{\text{Euc}}_{f(x_i)}(y)$.

Gaussian distribution. As an example where the generalization process is not as straightforward, we consider the Gaussian distribution. The difficulty arises due to the Gaussian distribution having many equivalent characterizations in the Euclidean domain, that do not end up being equivalent in the Riemannian case. To name a few, we list generalizations of these characterizations onto Riemannian manifolds

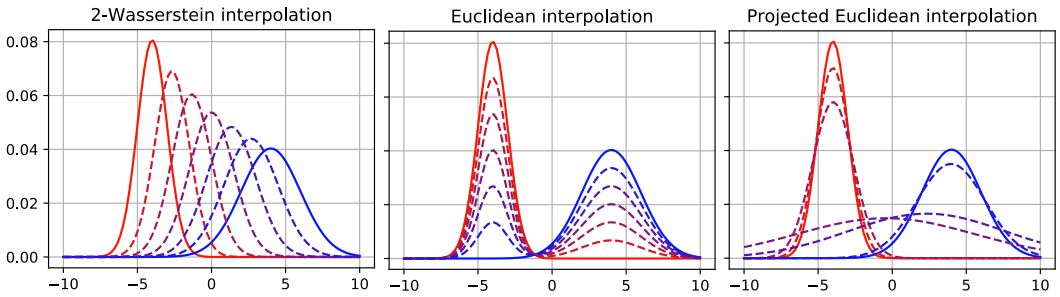


FIGURE 1.5: Interpolations (dashed) between two univariate Gaussians (red and blue) in the different geometries. The means are given by the middlemost interpolants.

1. The distribution with fixed mean μ and covariance matrix K maximizing entropy is [50].
2. The istropic distribution with a density satisfying [57]

$$p(x) \propto \exp\left(-\frac{1}{2\sigma^2} d^2_{\mathcal{M}}(\mu, x)\right). \tag{1.68}$$

3. The density is given by the heat kernel, which is the smallest positive solution to the heat equation [21]

$$\frac{\partial f}{\partial t} - \Delta f = 0, \tag{1.69}$$

   where Δ is the Laplace–Beltrami operator on the manifold (only provides an isotropic distribution).
4. The push-forward with respect to Exp_μ of $\mathcal{N}(0, K)$ defined in the tangent space $T_\mu\mathcal{M}$ [43]

$$\left(\mathrm{Exp}_\mu\right)_{\#} \mathcal{N}(0, K). \tag{1.70}$$

We consider 4., which is referred to as the *wrapped Gaussian distribution* (WGD). As the name suggests, WGDs might wrap mass around a compact manifold, in which case analytical evaluation of the density function becomes too difficult in practice. However, it is the only distribution of the above generalizations that can be conditioned with ease, owing to its analytical expression, which is used in the articles presented in Appendices B and C.

Note that, we can also characterize the population mean as the maximum likelihood estimator for the mean parameter of a Gaussian distribution. This is no longer true in the WGD case.

1.4.3 Comparative Example

We further illustrate the differences between the extrinsic and intrinsic approaches in the context of this thesis: By considering populations of univariate Gaussian distributions. Ideally, in the extrinsic approach, one would embed the space of Gaussians in a way that respects the geometry through *Nash embeddings*, however, in practice it is usually infeasible, as is the case here. Therefore, we consider the embedding of the measures through their densities into $L^2(\mathbb{R})$. On the other hand, intrinsic statistics requires defining a suitable metric, which in this example we choose to be the 2-Wasserstein metric discussed in Section 1.3.1.

Euclidean geometry. Absolutely continuous probability measures on $\mathbb{R}$ can be embedded in the Euclidean space (also a Hilbert space) $L^2(\mathbb{R})$ through their density functions, which we denote by

$$\psi\colon \mathrm{AC}(\mathbb{R}) \to L^2(\mathbb{R}), \quad \psi\colon \nu \mapsto \rho_\nu \tag{1.71}$$

With the inherited geometry, the distance between the embeddings of $\nu_1 = \mathcal{N}(m_1, \sigma_1^2)$ and $\nu_2 = \mathcal{N}(m_2, \sigma_2^2)$ is given by

$$d_{\mathrm{Euc}}(\psi(\nu_1), \psi(\nu_2)) = \left(\int_{\mathbb{R}} \left(\rho_{\nu_1}(x) - \rho_{\nu_2}(x)\right)^2 dx \right)^{\frac{1}{2}} \tag{1.72}$$

realised by the geodesic

$$\gamma : t \mapsto (1-t)\rho_{\nu_1} + t\rho_{\nu_2}, \quad t \in [0,1], \tag{1.73}$$

where the sum of functions is defined by $(f+g)(x) = f(x) + g(x)$. Note how for $t \in (0,1)$ the geodesic γ exits the manifold of Gaussian densities, returning a multimodal distribution, instead. Furthermore, the embedded mean $\bar{\rho}$ of a population of univariate Gaussians $\{\mathcal{N}(\mu_i, \sigma_i^2)\}_{i=1}^N$ is given by

$$\bar{\rho} = \frac{1}{N} \sum_{i=1}^{N} \rho_{\nu_i}, \tag{1.74}$$

which can then be projected back onto the space of Gaussians as $\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$, where

$$(\bar{\mu}, \bar{\sigma}^2) = \operatorname*{arg\,min}_{(\mu,\sigma^2)} \int_{\mathbb{R}} \left(\bar{\rho}(x) - \mathcal{N}(x|\mu, \sigma^2)\right)^2 dx. \tag{1.75}$$

Wasserstein geometry. Now, we view the space of univariate Gaussians as a finite-dimensional submanifold of the metric space $(\mathcal{P}_2(\mathbb{R}), W_2)$. From (1.39) we get the distance between two univariate Gaussians as

$$W_2(\nu_1, \nu_2) = \left(|\mu_1 - \mu_2|^2 + \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\right)^{\frac{1}{2}}, \tag{1.76}$$

realised by the geodesic, given by (1.36) and (1.37)

$$\gamma\colon t \mapsto \mathcal{N}\left((1-t)\mu_1 + t\mu_2, ((1-t)\sigma_1 + t\sigma_2)^2\right). \tag{1.77}$$

Finally, the Fréchet mean $\bar{\nu}$ of the population $\{\mathcal{N}(\mu_i, \sigma_i^2)\}_{i=1}^N$ minimizing (1.64) is given by

$$\bar{\nu} = \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^N \mu_i, \left(\frac{1}{N}\sum_{i=1}^N \sigma_i\right)^2\right), \tag{1.78}$$

Comparison. The means and geodesic interpolations between two Gaussians densities are given in Fig. 1.5, where the means are given by the middlemost interpolants. In the 2-Wasserstein case, the mean and variance parameters are independent in the geometry, and so the mean parameter of the Fréchet mean only depends on the mean parameters of the population, likewise, the covariance only depends on the covariance matrices of the population. In the Euclidean case, the interpolations and means generally fail to be Gaussian, as the resultants are multimodal. Projecting the resulting densities onto the space of Gaussians returns a density with high variance that tries to cover all the modalities.

1.5 Statistical Manifolds

Statistical manifolds are the central objects studied in *information geometry* [7]. They consist of a parametrized family of probability distributions and a *divergence function* on the space of probabilities, which quantifies the dissimilarity between two elements. A common choice is the *Kullback–Leibler* (KL) divergence, but we could also choose the p-Wasserstein metric, for example, leading to *Wasserstein information geometry* [40]. The parametrization can be viewed as a chart, and thus a statistical manifold is a submanifold, in a weak sense, of the space of probability measures.

From the practical point of view, relating the geometry endowed on the space of probabilities to the parameter space helps us in optimization problems related to learning tasks. This is because changes in the parameters would then correspond to appropriate changes in the probability measures.

We define a statistical manifold as a triple $(\mathcal{X}, \Theta, \rho)$, where $\mathcal{X}$ is the *sample space*, $\Theta \subseteq \mathbb{R}^n$ the *parameter space*, and

$$\rho\colon \Theta \to \mathcal{P}(\mathcal{X}), \qquad \rho\colon \theta \mapsto \rho_\theta \tag{1.79}$$

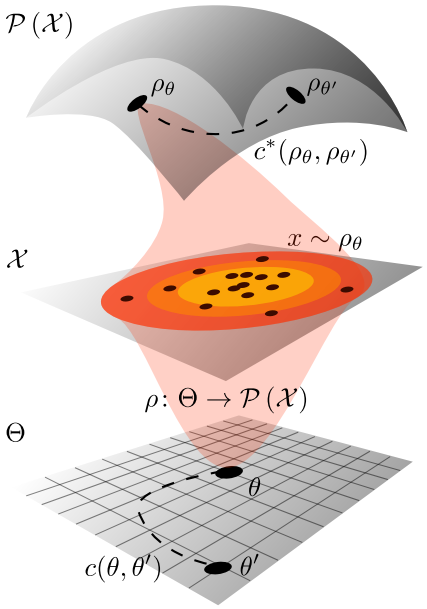


FIGURE 1.6: Illustration of a statistical manifold $(\mathcal{X}, \Theta, \rho)$, with the divergence measure c^* and its pull-back c on Θ.

associates each parameter to a probability measure in $\mathcal{P}(\mathcal{X})$. By abuse of language, we also call Θ a statistical manifold. We equip Θ with a geometry, by choosing a *divergence measure*

$$\begin{aligned} &c^*\colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0} \\ &c^*(\rho_1, \rho_2) = 0 \iff \rho_1 = \rho_2, \end{aligned} \tag{1.80}$$

on the probability space $\mathcal{P}(\mathcal{X})$. This divergence can then be *pulled back* to the statistical manifold Θ, by defining

$$c\colon \Theta \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}, \qquad c\colon (\theta_1, \rho) \mapsto c^*(\rho_{\theta_1}, \rho). \tag{1.81}$$

By abuse of notation, we write $c(\theta_1, \rho_{\theta_2}) = c(\theta_1, \theta_2)$. As a technical condition, we require $\theta \mapsto c(\theta, \theta_2)$ to be C^2 whenever $\theta \neq \theta_2$. A natural context where such a similarity measure arises is when learning a probabilistic model. See Fig. 1.6 for an illustration.

The given divergence measure c^* can be approximated by a Riemannian metric through a local second order approximation,

$$G^*_p = \nabla^2_{q=p} c^*(q, p), \tag{1.82}$$

called the *Riemannian metric associated with the divergence* c^*, whose pull-back Riemannian metric tensor on Θ is then given by

$$G_\theta = J_\theta^T \nabla^2_{\eta=\theta} c(\eta, \theta) J_\theta, \tag{1.83}$$

where J_θ stands for the Jacobian of the chart

$$J_\theta = \frac{\partial}{\partial \theta} \rho_\theta \tag{1.84}$$

A practical application of the statistical manifold is studied in the paper presented in Appendix E, where the pull-back metric tensor is used for *natural gradient descent* in the parameter space.

1.5.1 Wasserstein Generative Adversial Networks

As an example of a statistical manifold, we consider *generative adversial networks* (GANs) in generative modelling. In this context, the role of ρ is played by the *model* distribution μ_m, defined through a *generator*. Then, we wish to learn the parameters of the generator, so that a divergence on $\mathcal{P}(\mathcal{X})$ between the model and given *target* data distribution μ_t is minimized. Especially, after a general formulation, we will consider *Wasserstein GANs*, where the divergence is given by the 1-Wasserstein distance. GANs have found numerous applications, e.g., in medical imaging [36], inverse problems [2, 41], and 3D object generation [27, 63].

The model distribution is defined by pushing a *source* distribution μ_s forward by a parametrized map $g_{\omega'}$, the *generator*, with parameter ω', yielding

$$\mu_m = (g_{\omega'})_{\#}\, \mu_s. \tag{1.85}$$

The generator is typically given by a neural network, such as a *multilayer perceptron* (MLP), or a *convolutional neural network* (ConvNet). The source distribution μ_s is usually chosen to be a standard Gaussian in a low dimensional space, for example, $\mu_s = \mathcal{N}(0, I) \in \mathcal{P}(\mathbb{R}^{100})$. The low dimensionality is commonly justified by the *manifold hypothesis* [22] in machine learning, which assumes that the data lies on a low dimensional submanifold of the ambient data space (such as $\mathbb{R}^{3\times 64\times 64}$ for 64×64 color images).

GANs. The type of GAN is determined by the divergence c^* minimized on the space of probabilities. Choosing c^*, the objective becomes

$$\min_{\omega'} c^* \left((g_{\omega'})_{\#}\, \mu_s, \mu_t \right). \tag{1.86}$$

The original formulation by Goodfellow et al. [30] minimizes an approximation to the *Jensen–Shannon* (JS) divergence, given by

$$\mathrm{JS}(\nu \| \mu) \approx \max_{\omega} \left\{ \mathbb{E}_{x\sim\mu}[\log(\varphi_\omega(x))] + \mathbb{E}_{y\sim\nu}[\log(1-\varphi_\omega(y))] \right\}, \tag{1.87}$$

where, the function

$$\varphi_\omega : \mathcal{X} \to [0,1], \tag{1.88}$$

is the *discriminator* with parameters ω. In this formulation, the discriminator plays the *adversial part*, assigning a probability describing whether an element of the sample space comes from the target μ_t, or not. Putting (1.86) and (1.87) together yields the minimax objective

$$\min_{\omega'} \max_{\omega} \left\{ \mathbb{E}_{x\sim\mu_t} \left[\log(\varphi_\omega(x))\right] + \mathbb{E}_{z\sim\mu_s} \left[\log(1-\varphi_\omega(g_{\omega'}(z)))\right] \right\}. \tag{1.89}$$

In practice, this objective is solved by Monte Carlo methods and stochastic gradient descent: at each iteration we sample minibatches $\{x_i\}_{i=1}^N$ and $\{z_i\}_{i=1}^N$ from the target μ_t and source μ_s, respectively. The GAN pipeline is illustrated in Fig. 1.7.

Wasserstein GANs. An alternative to (1.87) is given by the 1-Wasserstein metric, and especially its dual formulation, introduced in Section 1.2, which can be written as

$$W_1(\mu,\nu) = \max_{(\varphi,\varphi^c))\in \mathrm{ADM}(c} \{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\varphi^c]\}, \tag{1.90}$$

where $c = d_{\mathrm{Euc}}$. In the special case where φ is 1-Lipschitz and $c = d$ for any metric distance d, the c-transform is given by

$$\varphi^c = -\varphi. \tag{1.91}$$

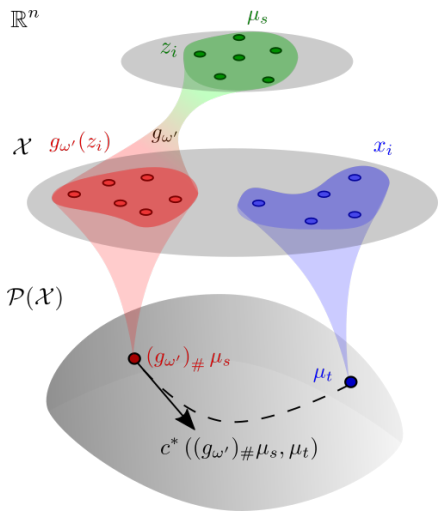


FIGURE 1.7: Illustration of the GAN setting, where the similarity measure c^* is minimized between the model $(g_{\omega'})_{\#}\,\mu_s$ and μ_t.

Additionally, one can show that any Kantorovich potential φ maximizing (1.90) is indeed a 1-Lipschitz function, and therefore

$$W_1(\mu,\nu) = \max_{\varphi \text{ is 1-Lipschitz}} \{\mathbb{E}_\mu[\varphi] - \mathbb{E}_\nu[\varphi]\}. \tag{1.92}$$

Letting $c^* = W_1$, and approximating the Kantorovich potential by a neural network

$$\varphi_\omega : \mathcal{X} \to \mathbb{R}, \tag{1.93}$$

yields the *Wasserstein GAN* (WGAN) objective

$$\min_{\omega'} \max_{\omega} \left\{\mathbb{E}_{x\sim\mu_t}\left[\varphi_\omega(x)\right] - \mathbb{E}_{z\sim\mu_s}\left[\varphi_\omega(g_{\omega'}(z))\right]\right\}. \tag{1.94}$$

This time, φ_ω is called a critic and not a discriminator, as it can assign any value from $\mathbb{R}$ to the samples. Note that one has to take care of enforcing the 1-Lipschitzness of the critic, which poses to be the main implementational difficulty of WGANs. In Appendix F we consider not enforcing the 1-Lipschitzness, but instead computing a discrete version of the c-transform. In Appendix G, we compare the ability of different methodologies enforcing these constraints in approximating and estimating the optimal transport quantity.

Chapter 2

Summary and Future Work

This chapter is devoted to summarizing the work presented in the appendices, and to discussing possible future directions for the research carried out during this PhD project. The summary is kept brief, as more detailed summaries will be given in each appendix.

In Chapter 1, we introduced the background required to describe the interplay between geometry and probabilistic modelling that is present in the work in Appendices A–G. This interplay can naturally be divided into two categories: modelling of geometric data, and geometry of probabilistic models. Below, we will discuss the relation of these two categories to the articles in the appendices, and how they relate to uncertainty quantification.

Modelling geometric data. As described in Section 1.4, data that naturally posses geometric structure can be studied with models, that are constrained to respect the geometric structure. As examples of such data, we mentioned shapes, diffusion tensors and directional data. On top of such deterministic objects, data that can be represented as probability measures can also be endowed with such geometric structure, as is one of the key takeaways of Chapter 1.

In Appendix A, we consider the geometrization of GPs under the OT framework. This is done in order to utilize the toolkit of geometric statistics on stochastic curves, which can be represented as GPs. The main motivation of this, is to incorporate the uncertainty of the curves to the statistical pipeline. Otherwise, the resulting statistics would be blind to this intrinsic uncertainty, preventing us from carrying out accurate uncertainty quantification.

In Appendices B and C, we focus on general methodology for statistics on manifolds. Particularly, we generalize GPs onto manifolds by *wrapped GPs* (WGPs), which are then utilized in non-parametric regression on geometric data in Appendix B, and for unsupervised learning of submanifolds formed by geometric data in Appendix C. This provides us with non-parametric, non-linear models for statistics on manifolds, whose conditional distributions provide uncertainty estimates.

Data objects living on Riemannian manifolds can also be stochastic, i.e., they can be represented as random points on the given manifold. In Appendix D,

we consider the OT geometry of wrapped Gaussian distributions, which can be utilized in similar vain to Appendix A.

Future work in this direction includes studying the OT geometry for more complicated representations of random curves than GPs. Furthermore, we utilized WGPs to generalize only two popular GP tools to manifolds, that is, GP regression and Gaussian process latent variable models (GPLVMs). We could further study the extension of GP statistics on the Euclidean space to manifolds.

Geometry of probabilistic models. Learning probabilistic models can often be cast as minimizing a geometric quantity between the model and a given data distribution, which we can cast into the context of statistical manifolds presented in Section 1.5. This allows us to relate the geometry of the model to its parameter space. On the other hand, computing the geometric quantity is often non-trivial, making studying its approximations worthwhile.

In Appendix E we study the natural gradient for an arbitrary divergence, which results from preconditioning the Euclidean gradient with the pull-back of the Riemannian metric associated with the divergence on the space of probabilities. The natural gradient can be utilized when learning the model as part of optimizing the model's parameters via *natural gradient descent*. This accelerates and smoothens the optimization process, as changes in the parameters are not allowed to make large deviations in the model, when measured with respect to the geometry of the space of probabilities. Due to computational bottlenecks, the natural gradient has not been widely considered in deep learning. Future work could entail *Monte Carlo* methods that make natural gradient compatible with deep learning, in the spirit of [65].

Wasserstein GANs are studied in Appendices F and G. In particular, approximating the 1-Wasserstein distance, with the use of discriminator neural networks, is considered. Appendix F studies the use of the c-transform in enforcing the constraints associated with the dual formulation of the 1-Wasserstein distance, whereas in Appendix G, we consider the ability of different heuristic schemes to approximate and estimate the 1-Wasserstein distance. This work can naturally be continued, by considering in more detail, how the different schemes affect the resulting Kantorovich potentials. On the other hand, it would be interesting to have results stating, how complicated Kantorovich potentials are required between given families of distributions. Finally, applying the methodology of WGANs in the multimarginal OT setting [28, 49] would be an interesting new direction.

Bibliography

[1] Adler, J., Lunz, S.: Banach wasserstein gan. In: Advances in Neural Information Processing Systems. pp. 6754–6763 (2018)

[2] Adler, J., Öktem, O.: Deep bayesian inversion. arXiv preprint arXiv:1811.05910 (2018)

[3] Agueh, M.: Finsler structure in the p-Wasserstein space and gradient flows. Comptes Rendus Mathematique **350**(1-2), 35–40 (2012)

[4] Amari, S.I.: Natural gradient works efficiently in learning. Neural computation **10**(2), 251–276 (1998)

[5] Amari, S.i.: Divergence function, information monotonicity and information geometry. In: Workshop on Information Theoretic Methods in Science and Engineering (WITMSE). Citeseer (2009)

[6] Amari, S.i.: Information geometry and its applications, vol. 194. Springer (2016)

[7] Amari, S.i., Nagaoka, H.: Methods of information geometry, vol. 191. American Mathematical Soc. (2007)

[8] Ambrosio, L., Gigli, N.: A user's guide to optimal transport. In: Modelling and optimisation of flows on networks, pp. 1–155. Springer (2013)

[9] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint arXiv:1701.07875 (2017)

[10] Bandulasiri, A., Bhattacharya, R.N., Patrangenaru, V.: Nonparametric inference for extrinsic means on size-and-(reflection)-shape manifolds with applications in medical imaging. Journal of Multivariate Analysis **100**(9), 1867–1882 (2009)

[11] Banerjee, M., Chakraborty, R., Ofori, E., Vaillancourt, D., Vemuri, B.C.: Nonlinear regression on Riemannian manifolds and its applications to neuroimage analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 719–727. Springer (2015)

[12] Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik **84**(3), 375–393 (2000)

[13] Bhattacharya, R., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds. The Annals of Statistics **31**(1), 1–29 (2003)

[14] Bhattacharya, R.N., Ellingson, L., Liu, X., Patrangenaru, V., Crane, M.: Extrinsic analysis on manifolds is computationally faster than intrinsic analysis with applications to quality control by machine vision. Applied Stochastic Models in Business and Industry **28**(3), 222–235 (2012)

[15] Bulygina, O., Razuvaev, V.: Daily temperature and precipitation data for 518 russian meteorological stations (1881-2010). Tech. rep., Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National ... (2012)

[16] do Carmo, M.P.: Riemannian geometry. Birkhäuser (1992)

[17] Chen, Y., Li, W.: Natural gradient in Wasserstein statistical manifold. arXiv preprint arXiv:1805.08380 (2018)

[18] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in neural information processing systems. pp. 2292–2300 (2013)

[19] Dowson, D., Landau, B.: The Fréchet distance between multivariate normal distributions. Journal of multivariate analysis **12**(3), 450–455 (1982)

[20] Dukler, Y., Li, W., Lin, A., Montufar, G.: Wasserstein of Wasserstein loss for learning generative models. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 1716–1725. PMLR (09–15 Jun 2019)

[21] Émery, M.: Stochastic calculus in manifolds. Springer Science & Business Media (2012)

[22] Fefferman, C., Mitter, S., Narayanan, H.: Testing the manifold hypothesis. Journal of the American Mathematical Society **29**(4), 983–1049 (2016)

[23] Fletcher, P.T.: Geodesic regression and the theory of least squares on Riemannian manifolds. International journal of computer vision **105**(2), 171–185 (2013)

[24] Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE transactions on medical imaging **23**(8), 995–1005 (2004)

[25] Fletcher, P.T., Zhang, M.: Probabilistic geodesic models for regression and dimensionality reduction on Riemannian manifolds. In: Riemannian Computing in Computer Vision, pp. 101–121. Springer (2016)

[26] Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. In: Annales de l'institut Henri Poincaré. vol. 10, pp. 215–310 (1948)

[27] Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: 2017 International Conference on 3D Vision (3DV). pp. 402–411. IEEE (2017)

[28] Gerolin, A., Kausamo, A., Rajala, T.: Duality theory for multi-marginal optimal transport with repulsive costs in metric spaces. ESAIM: Control, Optimisation and Calculus of Variations **25**, 62 (2019)

[29] Givens, C.R., Shortt, R.M., et al.: A class of Wasserstein metrics for probability distributions. The Michigan Mathematical Journal **31**(2), 231–240 (1984)

[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)

[31] Grove, K., Karcher, H.: How to conjugate c^1-close group actions. Mathematische Zeitschrift **132**(1), 11–20 (1973)

[32] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in neural information processing systems. pp. 5767–5777 (2017)

[33] Hauberg, S., Schober, M., Liptrot, M., Hennig, P., Feragen, A.: A random Riemannian metric for probabilistic shortest-path tractography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 597–604. Springer (2015)

[34] Hinkle, J., Muralidharan, P., Fletcher, P.T., Joshi, S.: Polynomial regression on Riemannian manifolds. In: European Conference on Computer Vision. pp. 1–14. Springer (2012)

[35] Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: Geodesic pca for Riemannian manifolds modulo isometric Lie group actions. Statistica Sinica pp. 1–58 (2010)

[36] Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. arXiv preprint arXiv:1809.06222 (2018)

[37] Knott, M., Smith, C.S.: On the optimal mapping of distributions. Journal of Optimization Theory and Applications **43**(1), 39–49 (1984)

[38] Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. In: Advances in neural information processing systems. pp. 329–336 (2004)

[39] Lee, J.M.: Riemannian manifolds: an introduction to curvature, vol. 176. Springer Science & Business Media (2006)

[40] Li, W., Montúfar, G.: Natural gradient via optimal transport. Information Geometry **1**(2), 181–214 (2018)

[41] Lunz, S., Öktem, O., Schönlieb, C.B.: Adversarial regularizers in inverse problems. In: Advances in Neural Information Processing Systems. pp. 8507–8516 (2018)

[42] Malagò, L., Montrucchio, L., Pistone, G.: Wasserstein Riemannian geometry of Gaussian densities. Information Geometry **1**(2), 137–179 (2018)

[43] Mardia, K.V., Jupp, P.E.: Directional statistics, vol. 494. John Wiley & Sons (2009)

[44] Masarotto, V., Panaretos, V.M., Zemel, Y.: Procrustes metrics on covariance operators and optimal transportation of gaussian processes. Sankhya A **81**(1), 172–213 (2019)

[45] Minh, H.Q., Murino, V.: From covariance matrices to covariance operators: Data representation from finite to infinite-dimensional settings. In: Algorithmic Advances in Riemannian Geometry and Applications, pp. 115–143. Springer (2016)

[46] Monge, G.: Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris (1781)

[47] Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra and its Applications **48**, 257–263 (1982)

[48] Otto, F.: The geometry of dissipative evolution equations: The porous medium equation. Communications in Partial Differential Equations **26**(1-2), 101–174 (2001). https://doi.org/10.1081/PDE-100002243

[49] Pass, B.: Multi-marginal optimal transport: theory and applications. ESAIM: Mathematical Modelling and Numerical Analysis **49**(6), 1771–1790 (2015)

[50] Pennec, X.: Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision **25**(1), 127 (2006)

[51] Pennec, X., Sommer, S., Fletcher, T.: Riemannian Geometric Statistics in Medical Image Analysis. Elsevier (2020). https://doi.org/10.1016/C2017-0-01561-6

[52] Pennec, X., et al.: Barycentric subspace analysis on manifolds. The Annals of Statistics **46**(6A), 2711–2746 (2018)

[53] Peyré, G., Cuturi, M., et al.: Computational optimal transport. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)

[54] Pigoli, D., Aston, J.A., Dryden, I.L., Secchi, P.: Distances and inference for covariance operators. Biometrika **101**(2), 409–422 (2014)

[55] Pigoli, D., Menafoglio, A., Secchi, P.: Kriging prediction for manifold-valued random fields. Journal of Multivariate Analysis **145**, 117–131 (2016)

[56] Rasmussen, C.E.: Gaussian processes in machine learning. In: Summer School on Machine Learning. pp. 63–71. Springer (2003)

[57] Said, S., Bombrun, L., Berthoumieu, Y., Manton, J.H.: Riemannian Gaussian distributions on the space of symmetric positive definite matrices. IEEE Transactions on Information Theory **63**(4), 2153–2170 (2017)

[58] Schiratti, J.B., Allassonnière, S., Colliot, O., Durrleman, S.: A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. Journal of Machine Learning Research **18**(133), 1–33 (2017)

[59] Schober, M., Kasenburg, N., Feragen, A., Hennig, P., Hauberg, S.: Probabilistic shortest path tractography in DTI using Gaussian process ODE solvers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 265–272. Springer (2014)

[60] Sommer, S., Lauze, F., Hauberg, S., Nielsen, M.: Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In: European conference on computer vision. pp. 43–56. Springer (2010)

[61] Takatsu, A., et al.: Wasserstein geometry of Gaussian measures. Osaka Journal of Mathematics **48**(4), 1005–1026 (2011)

[62] Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008)

[63] Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in neural information processing systems. pp. 82–90 (2016)

[64] Zhang, M., Fletcher, T.: Probabilistic principal geodesic analysis. In: Advances in Neural Information Processing Systems. pp. 1178–1186 (2013)

[65] Zhao, J., Li, W.: Wasserstein information matrix. arXiv preprint arXiv:1910.11248 (2019)

List of Publications

Appendix A Anton Mallasto and Aasa Feragen. "Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes." Advances in Neural Information Processing Systems. 2017.

Appendix B Anton Mallasto and Aasa Feragen. "Wrapped Gaussian process regression on Riemannian manifolds." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Appendix C Anton Mallasto, Søren Hauberg, and Aasa Feragen. "Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models." Proceedings of Machine Learning Research. 2019.

Appendix D Anton Mallasto and Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions." 38^{th} International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. 2018.

Appendix E Anton Mallasto, Tom Dela Haije, and Aasa Feragen. "A Formalization of The Natural Gradient Method for General Similarity Measures." Proceedings of 4^{th} International Conference on Geometric Science of Information. 2019.

Appendix F Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen "(q, p)-Wasserstein GANs: Comparing Ground Metrics for Wasserstein GANs." arXiv preprint arXiv:1902.03642. 2019.

Appendix G Anton Mallasto, Guido Montúfar, and Augusto Gerolin. "How Well Do WGANs Estimate the Wasserstein Metric?" arXiv preprint arXiv:1910.03875. 2019

Appendix A

Learning from Uncertain Curves: The 2-Wasserstein Metric for Gaussian Processes

---

The following chapter presents (up to formatting) the article

> Anton Mallasto, and Aasa Feragen. "Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes." Advances in Neural Information Processing Systems (NeurIPS). 2017.

This work studies the 2-Wasserstein geometry of GPs and covariance operators, in particular, how the metric between GPs relates to the metric between their finite-dimensional marginals in the sense of continuity. The framework is derived in order to carry out simple statistics such as interpolation and computing Fréchet means for a population of GPs, that can be used to model curves with uncertainty. The continuity results justify practical computations that have to rely on finite-dimensional marginals. The framework was demonstrated on GPs modelling yearly temperature variation [15], and GPs representing uncertain estimates of trajectories in diffusion MRI [33, 59].

The work can be seen as a natural continuation of the study of the 2-Wasserstein metric between Gaussian distributions [19, 29, 37, 47], whose related Riemannian structure has been explored in [42, 61]. On the other hand, it continues the geometric study of covariance operators, see [45, 54]. Briefly after the publication of this work, another article with similar contribution to ours appeared, with a more general exposition [44].

Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes

Anton Mallasto Aasa Feragen

Department of Computer Science, University of Copenhagen

Abstract

We introduce a novel framework for statistical analysis of populations of non-degenerate Gaussian processes (GPs), which are natural representations of uncertain curves. This allows inherent variation or uncertainty in function-valued data to be properly incorporated in the population analysis. Using the 2-Wasserstein metric we geometrize the space of GPs with L^2 mean and covariance functions over compact index spaces. We prove uniqueness of the barycenter of a population of GPs, as well as convergence of the metric and the barycenter of their finite-dimensional counterparts. This justifies practical computations. Finally, we demonstrate our framework through experimental validation on GP datasets representing brain connectivity and climate development. A MATLAB library for relevant computations will be published at `https://sites.google.com/view/antonmallasto/software`.

1 Introduction

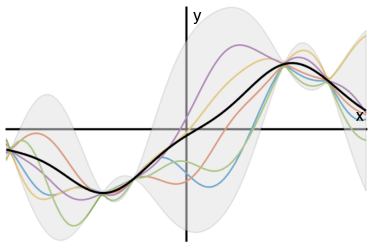


Figure 1: An illustration of a GP, with mean function (*in black*) and confidence bound (*in grey*). The colorful curves are sample paths of this GP.

Gaussian processes (GPs, see Fig. 1) are the counterparts of Gaussian distributions (GDs) over functions, making GPs natural objects to model uncertainty in estimated functions. With the rise of GP modelling and probabilistic numerics, GPs are increasingly used to model uncertainty in function-valued data such as segmentation boundaries [17, 19, 30], image registration [38] or time series [28]. Centered GPs, or covariance operators, appear as image features in computer vision [12, 16, 25, 26] and as features of phonetic language structure [23]. A natural next step is therefore to analyze populations of GPs, where performance depends crucially on proper incorporation of inherent uncertainty or variation. This

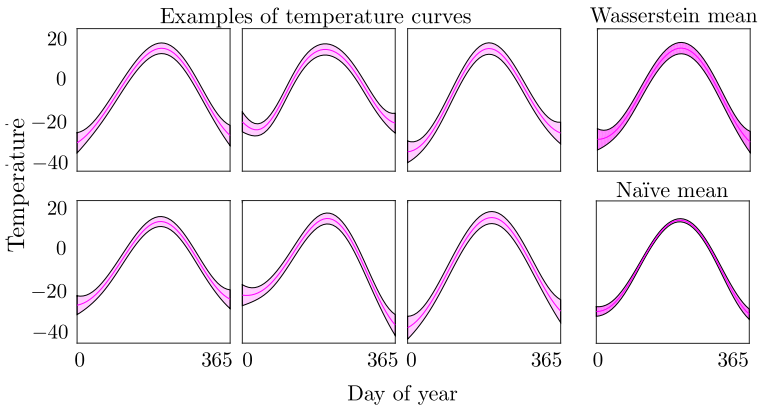


Figure 2: **Left:** Example GPs describing the daily minimum temperatures in a Siberian city (see Sec. 4). **Right top:** The mean GP temperature curve, computed as a Wasserstein barycenter. Note that the inherent variability in the daily temperature is realistically preserved, in contrast with the naïve approach. **Right bottom:** A naïve estimation of the mean and standard deviation of the daily temperature, obtained by taking the day-by-day mean and standard deviation of the temperature. All figures show a 95% confidence interval.

paper contributes a principled framework for population analysis of GPs based on Wasserstein, a.k.a. earth mover's, distances.

The importance of incorporating uncertainty into population analysis is emphasized by the example in Fig. 2, where each data point is a GP representing the minimal temperature in the Siberian city Vanavara over the course of one year [9,34]. A naïve way to compute its average temperature curve is to compute the per-day mean and standard deviation of the yearly GP mean curves. This is shown in the bottom right plot, and it is clear that the temperature variation is grossly underestimated, especially in the summer season. The top right figure shows the mean GP obtained with our proposed framework, which preserves a far more accurate representation of the natural temperature variation.

We propose analyzing populations of GPs by geometrizing the space of GPs through the *Wasserstein distance*, which yields a metric between probability measures with rich geometric properties. **We contribute** i) closed-form solutions for arbitrarily good approximation of the Wasserstein distance by showing that the 2-Wasserstein distance between two finite-dimensional GP representations converges to the 2-Wasserstein distance of the two GPs; and ii) a characterization of a non-degenerate barycenter of a population of GPs, and a proof that such a barycenter is unique, and can be approximated by its finite-dimensional counterpart.

We evaluate the Wasserstein distance in two applications. First, we illustrate the use of the Wasserstein distance for processing of uncertain white-matter trajectories in the brain segmented from noisy diffusion-weighted imaging (DWI) data using *tractography*. It is well known that the noise level and the low resolution of DWI images result in unreliable trajectories (*tracts*) [24]. This is problematic as the estimated tracts are e.g. used for surgical planning [8].

Recent work [17, 30] utilizes probabilistic numerics [29] to return *uncertain* tracts represented as GPs. We utilize the Wassertein distance to incorporate the estimated uncertainty into typical DWI analysis tools such as tract clustering [37] and visualization. Our second study quantifies recent climate development based on data from Russian meteorological stations using permutation testing on population barycenters, and supplies interpretability of the climate development using GP-valued kernel regression.

Related work. Multiple frameworks exist for comparing Gaussian distributions (GDs) represented by their covariance matrices, including the Frobenius, Fisher-Rao (affine-invariant), log-Euclidean and Wasserstein metrics. Particularly relevant to our work is the 2-Wasserstein metric on GDs, whose Riemannian geometry is studied in [33], and whose barycenters are well understood [1,4].

A body of work exists on generalizing the aforementioned metrics to the infinite-dimensional covariance operators. As pointed out in [23], extending the affine-invariant and Log-Euclidean metrics is problematic as covariance operators are not compatible with logarithmic maps and their inverses are unbounded. These problems are avoided in [25, 26] by regularizing the covariance operators, but unfortunately, this also alters the data in a non-unique way. The Procrustes metric from [23] avoids this, but as it is, only defines a metric between covariance operators.

The 2-Wasserstein metric, on the other hand, generalizes naturally from GDs to GPs, does not require regularization, and can be arbitrarily well approximated by a closed form expression, making the computations cheap. Moreover, the theory of optimal transport [5, 6, 36] shows that the Wasserstein metric yields a rich geometry, which is further demonstrated by the previous work on GDs [33].

After this work was presented in NIPS, a preprint appeared [20] which also studies convergence results and barycenters of GPs in the Wasserstein geometry, in a more general setting.

Structure. Prior to introducing the Wasserstein distance between GPs, we review GPs, their Hilbert space covariance operators and the corresponding Gaussian measures in Sec. 2. In Sec. 3 we introduce the Wasserstein metric and its barycenters for GPs and prove convergence properties of the metric and barycenters, when GPs are approximated by finite-dimensional GDs. Experimental validation is found in Sec. 4, followed by discussion and conclusion in Sec. 5.

2 Prerequisites

Gaussian processes and measures. A *Gaussian process* (GP) f is a collection of random variables, such that any finite restriction of its values $(f(x_i))_{i=1}^N$

has a joint Gaussian distribution, where $x_i \in X$, and X is the *index set*. A GP is entirely characterized by the pair

$$m(x) = \mathbb{E}\left[f(x)\right], \; k(x,x') = \mathbb{E}\left[(f(x) - m(x))(f(x') - m(x'))\right] \; , \tag{1}$$

where m and k are called the *mean function* and *covariance function*, respectively. We use the notation $f \sim \mathcal{GP}(m,k)$ for a GP f with mean function m and covariance function k. It follows from the definition that the covariance function k is symmetric and positive semidefinite. We say that f is *non-degenerate*, if k is strictly positive definite. We will assume the GPs used to be non-degenerate.

GPs relate closely to *Gaussian measures* on Hilbert spaces. Given probability spaces (X, Σ_X, μ) and (Y, Σ_Y, ν), we say that the measure ν is a *push-forward* of μ if $\nu(A) = \mu(T^{-1}(A))$ for a measurable $T\colon X \to Y$ and any $A \in \Sigma_Y$. Denote this by $T_{\#}\mu = \nu$. A Borel measure μ on a separable Hilbert space $\mathcal{H}$ is a *Gaussian measure*, if its push-forward with respect to any non-zero continuous element of the dual space of $\mathcal{H}$ is a non-degenerate Gaussian measure on $\mathbb{R}$ (i.e., the push-forward gives a univariate Gaussian distribution). A Borel-measurable set B is a *Gaussian null set*, if $\mu(B) = 0$ for any Gaussian measure μ on X. A measure ν on $\mathcal{H}$ is *regular* if $\nu(B) = 0$ for any Gaussian null set B. Note that regular Gaussian measures correspond to non-degenerate GPs.

Covariance operators. Denote by $L^2(X)$ the space of L^2-integrable functions from X to $\mathbb{R}$. The covariance function k has an associated integral operator $K\colon L^2(X) \to L^2(X)$ defined by

$$[K\phi](x) = \int_X k(x,s)\phi(s)ds, \; \forall \phi \in L^2(X) \; , \tag{2}$$

called the *covariance operator* associated with k. As a by-product of the 2-Wasserstein metric on centered GPs, we get a metric on covariance operators. The operator K is Hilbert-Schmidt, self-adjoint, compact, positive, and of trace class, and the space of such covariance operators is a convex space. Furthermore, the assignment $k \mapsto K$ from $L^2(X \times X)$ is an isometric isomorphism onto the space of Hilbert-Schmidt operators on $L^2(X)$ [7, Prop. 2.8.6]. This justifies us to write both $f \sim \mathcal{GP}(m,K)$ and $f \sim \mathcal{GP}(m,k)$.

Trace of an operator. The Wasserstein distance between GPs admits an analytical formula using traces of their covariance operators, as we will see below. Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a *separable* Hilbert space with the orthonormal basis $\{e_k\}_{k=1}^{\infty}$. Then the *trace* of a bounded linear operator T on $\mathcal{H}$ is given by

$$\text{Tr } T := \sum_{k=1}^{\infty} \langle Te_k, e_k \rangle \; , \tag{3}$$

which is absolutely convergent and independent of the choice of the basis if $\text{Tr } (T^*T)^{\frac{1}{2}} < \infty$, where T^* denotes the adjoint operator of T and $T^{\frac{1}{2}}$ is the square-root of T. In this case T is called a *trace class operator*. For positive self-adjoint operators, the trace is the sum of their eigenvalues.

The Wasserstein metric. The *Wasserstein metric* on probability measures derives from the optimal transport problem introduced by Monge and made rigorous by Kantorovich. The p-Wasserstein distance describes the minimal cost of transporting the unit mass of one probability measure into the unit mass of another probability measure, when the cost is given by a L^p distance [5, 6, 36].

Let (M, d) be a Polish space (complete and separable metric space) and denote by $\mathcal{P}_p(M)$ the set of all probability measures μ on M satisfying $\int_M d^p(x, x_0)d\mu(x) < \infty$ for some $x_0 \in M$. The p-Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{P}_p(M)$ is given by

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma[\mu,\nu]} \int_{M\times M} d^p(x_1, x_2)d\gamma(x_1, x_2) \right)^{\frac{1}{p}}, \ (x_1, x_2) \in M \times M, \quad (4)$$

where $\Gamma[\mu, \nu]$ is the set of joint measures on $M \times M$ with marginals μ and ν. Defined as above, W_p satisfies the properties of a metric. Furhermore, a minimizer in (4) is always achieved.

3 The Wasserstein metric for GPs

We will now study the Wasserstein metric with $p = 2$ between GPs. For GDs, this has been studied in [11, 14, 18, 22, 33].

From now on, assume that all GPs $f \sim \mathcal{GP}(m, k)$ are indexed over a compact $X \subset \mathbb{R}^n$ so that $\mathcal{H} := L^2(X)$ is separable. Furthermore, we assume $m \in L^2(X)$, $k \in L^2(X \times X)$, so that observations of f live almost surely in $\mathcal{H}$. Let $f_1 \sim \mathcal{GP}(m_1, k_1)$ and $f_2 \sim \mathcal{GP}(m_2, k_2)$ be GPs with associated covariance operators K_1 and K_2 , respectively. As the sample paths of f_1 and f_2 are in $\mathcal{H}$, they induce Gaussian measures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathcal{H})$ on $\mathcal{H}$, as there is a 1-1 correspondence between GPs having sample paths almost surely on a $L^2(X)$ space and Gaussian measures on $L^2(X)$ [27].

The 2-Wasserstein metric between the Gaussian measures μ_1, μ_2 is given by [13]

$$W_2^2(\mu_1, \mu_2) = d_2^2(m_1, m_2) + \text{Tr } (K_1 + K_2 - 2(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}})^{\frac{1}{2}}), \quad (5)$$

where d_2 is the canonical metric on $L^2(X)$. Using this, we get the following definition

Definition 1. *Let f_1, f_2 be GPs as above, and the induced Gaussian measures of f_1 and f_2 be μ_1 and μ_2, respectively. Then, their squared 2-Wasserstein distance is given by*

$$W_2^2(f_1, f_2) := W_2^2(\mu_1, \mu_2) = d_2^2(m_1, m_2) + Tr\ (K_1 + K_2 - 2(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}})^{\frac{1}{2}}) \ .$$

Remark 2. *Note that the case $m_1 = m_2 = 0$ defines a metric for the covariance operators K_1, K_2, as (5) shows that the space of GPs is isometric to the cartesian product of $L^2(X)$ and the covariance operators. We will denote this metric by $W_2^2(K_1, K_2)$. Furthermore, as GDs are just a subset of GPs, W_2^2 yields also the 2-Wasserstein metric between GDs studied in [11, 14, 18, 22, 33].*

Barycenters of Gaussian processes. Next, we define and study barycenters of populations of GPs, in a similar fashion as the GD case in [1].

Given a population $\{\mu_i\}_{i=1}^N \subset \mathcal{P}_2(\mathcal{H})$ and weights $\{\xi_i \geq 0\}_{i=1}^N$ with $\sum_{i=1}^N \xi_i = 1$, and $\mathcal{H}$ a separable Hilbert space, the solution $\bar{\mu}$ of the problem

$$(\mathcal{P}) \quad \inf_{\mu \in \mathcal{P}_2(\mathcal{H})} \sum_{i=1}^N \xi_i W_2^2(\mu_i, \mu),$$

is the *barycenter* of the population $\{\mu_i\}_{i=1}^N$ with *barycentric coordinates* $\{\xi_i\}_{i=1}^N$. The barycenter for GPs is defined to be the barycenter of the associated Gaussian measures.

Remark 3. *The following theorems require the assumption that the barycenter is non-degenerate; it is still a conjecture that the barycenter of non-degenerate GPs is nondegenerate [20], but this holds in the finite-dimensional case of GDs.*

We now state the main theorem of this section, which follows from Prop. 5 and Prop. 6 below.

Theorem 4. *Let $\{f_i\}_{i=1}^N$ be a population of GPs with $f_i \sim \mathcal{GP}(m_i, K_i)$, then there exists a unique barycenter $\bar{f} \sim \mathcal{GP}(\bar{m}, \bar{K})$ with barycentric coordinates $(\xi_i)_{i=1}^N$. If $\bar{f}$ is non-degenerate, then $\bar{m}$ and $\bar{K}$ satisfy*

$$\bar{m} = \sum_{i=1}^N \xi_i m_i, \quad \sum_{i=1}^N \xi_i \left(\bar{K}^{\frac{1}{2}} K_i \bar{K}^{\frac{1}{2}} \right)^{\frac{1}{2}} = \bar{K}.$$

Proposition 5. *Let $\{\mu_i\}_{i=1}^N \subset \mathcal{P}_2(\mathcal{H})$ and $\bar{\mu}$ be a barycenter with barycentric coordinates $(\xi_i)_{i=1}^N$. Assume μ_i is regular for some i, then $\bar{\mu}$ is the unique minimizer of $(\mathcal{P})$.*

Proof. We first show that the map $\nu \mapsto W_2^2(\mu, \nu)$ is convex, and strictly convex if μ is a regular measure. To see this, let $\nu_i \in \mathcal{P}_2(\mathcal{H})$ and $\gamma_i^* \in \Gamma[\mu, \nu_i]$ be the optimal transport plans between μ and ν_i for $i = 1, 2$, then $\lambda\gamma_1^* + (1-\lambda)\gamma_2^* \in \Gamma[\mu, \lambda\nu_1 + (1-\lambda)\nu_2]$ for $\lambda \in [0, 1]$. Therefore

$$\begin{aligned} W_2^2(\mu, \lambda\nu_1 + (1-\lambda)\nu_2) &= \inf_{\gamma \in \Gamma[\mu, \lambda\nu_1 + (1-\lambda)\nu_2]} \int_{\mathcal{H}\times\mathcal{H}} d^2(x, y) d\gamma \\ &\leq \int_{\mathcal{H}\times\mathcal{H}} d^2(x, y) d(\lambda\gamma_1^* + (1-\lambda)\gamma_2^*) \\ &= \lambda W_2^2(\mu, \nu_1) + (1-\lambda) W_2^2(\mu, \nu_2), \end{aligned}$$

which gives convexity. Note that for $\lambda \in]0, 1[$, the transport plan $\lambda\gamma_1^* + (1-\lambda)\gamma_2^*$ splits mass. Therefore it cannot be the unique optimal plan between μ and $(1-t)\nu_1 + t\nu_2$. As μ is regular, the optimal plan does not split mass, as it is induced by a map [3, Thm. 6.2.10], so we have strict convexity. From this follows the strict convexity of the object function in $(\mathcal{P})$. □

Next we characterize the barycenter, assuming it is non-degenerate, in the spirit of the finite-dinemsional case in [1, Thm. 6.1].

Proposition 6. *Let $\{f_i\}_{i=1}^N$ be a population of centered GPs, $f_i \sim \mathcal{GP}(0, K_i)$. Then $(\mathcal{P})$ has a unique solution $\bar{f} \sim \mathcal{GP}(0, \bar{K})$. If $\bar{f}$ is non-degenerate, then $\bar{K}$ is the unique bounded self-adjoint positive linear operator satisfying*

$$\sum_{i=1}^{N} \xi_i \left(K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \right)^{\frac{1}{2}} = K. \tag{6}$$

Proof. Existence can be shown following the proof for the finite dimensional case [1, Prop. 4.2], which uses *multimarginal optimal transport*; this appears in the preprint [20, Cor. 9]. For the characterization, assume $\bar{f}$ to be non-degenerate, and let

$$\mathrm{BC}(f) = \sum_{i=1}^{N} \xi_i W_2^2(f_i, f),$$

be the barycentric expression, and assume that the minimizer $\bar{f}$ of BC is non-degenerate. Let $0 < \lambda_1, \lambda_2, ...$ be the eigenvalues of $\bar{K}$ with eigenfunctions $e_1, e_2, ...$ Then, by [10, Prop. 2.2.] the transport map between $\bar{f}$ and f_k is given by

$$T_k(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{\langle x, e_j \rangle \langle (\bar{K}^{\frac{1}{2}} K_k \bar{K}^{\frac{1}{2}})^{\frac{1}{2}} e_j, e_i \rangle}{\lambda_i^{\frac{1}{2}} \lambda_j^{\frac{1}{2}}} e_i(x) \ . \tag{7}$$

Using [6, Thm. 8.4.7], we can write the gradient of the barycentric expression. We furthermore know that the expression is strictly convex, thus the gradient at $\bar{f}$ equals zero if and only if $\bar{f}$ is the minimizer. Now let Id be the identity operator, then

$$\nabla \mathrm{BC}(\bar{f}) = \sum_{i=1}^{N} (T_k - \mathrm{Id}\ ) = 0,$$

substituting in (7), we get

$$\sum_{i=1}^{N} \xi_i \left(K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \right)^{\frac{1}{2}} = K.$$

□

Proof of Theorem 4. Use Prop. 6, the properties of a barycenter in a Hilbert space, and that the space of GPs is isometric to the cartesian product of $L^2(X)$ and the covariance operators. □

Remark 7. *For the practical computations of barycenters of GDs approximating GPs, to be discussed below, a fixed-point iteration scheme with a guarantee of convergence exists [4, Thm. 4.2].*

Convergence properties. Now, we show that the 2-Wasserstein metric for GPs can be approximated arbitrarily well by the 2-Wasserstein metric for GDs. This is important, as in real-life we observe finite-dimensional representations of the covariance operators.

Let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis for $L^2(X)$. Then we define the GDs given by restrictions m_{in} and K_{in} of m_i and K_i, $i = 1, 2$, on $V_n = \text{span}(e_1, ..., e_n)$ by

$$m_{in}(x) = \sum_{k=1}^{n} \langle m_i, e_k \rangle e_k(x), \;\; K_{in}\phi = \sum_{k=1}^{n} \langle \phi, e_k \rangle K_i e_k, \;\; \forall \phi \in V_n, \;\; \forall x \in X \;, \quad (8)$$

and prove the following:

Theorem 8. *The 2-Wasserstein metric between GDs on finite samples converges to the Wasserstein metric between GPs, that is, if $f_{in} \sim \mathcal{N}(m_{in}, K_{in})$, $f_i \sim \mathcal{GP}(m_i, K_i)$ for $i = 1, 2$, then*

$$\lim_{n\to\infty} W_2^2(f_{1n}, f_{2n}) = W_2^2(f_1, f_2).$$

By the same argument, it also follows that $W_2^2(\cdot, \cdot)$ is continuous in both arguments in operator norm topology.

Proof. $K_{in} \to K_i$ in operator norm as $n \to \infty$. Because taking a sum, product and square-root of operators are all continuous with respect to the operator norm, it follows that

$$K_{1n} + K_{2n} - 2(K_{1n}^{\frac{1}{2}} K_{2n} K_{1n}^{\frac{1}{2}})^{\frac{1}{2}} \to K_1 + K_2 - 2(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}})^{\frac{1}{2}}.$$

Note that for any sequence $A_n \to A$ with convergence in operator norm, we have

$$|\text{Tr } A - \text{Tr } A_n| \leq \sum_{k=1}^{\infty} |\langle (A - A_n)e_k, e_k \rangle| \overset{\text{Cauchy-Schwarz}}{\leq} \sum_{k=1}^{\infty} \|(A - A_n)e_k\|_{L^2} \overset{\text{MCT}}{\to} 0 \,, \quad (9)$$

as $\lim_{n\to\infty} \sup_{v \in L^2_\omega(X)} \|(A - A_n)v\|_{L^2} = 0$ due to the convergence in operator norm. Here MCT stands for the monotone convergence theorem. Thus we have

$$\begin{aligned} W_2^2(f_{1n}, f_{2n}) &= d_2^2(m_{1n}, m_{2n}) + \text{Tr } (K_{1n} + K_{2n} - 2(K_{1n}^{\frac{1}{2}} K_{2n} K_{1n}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &\overset{n\to\infty}{\to} d_2^2(m_1, m_2) + \text{Tr } (K_1 + K_2 - 2(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= W_2^2(f_1, f_2). \end{aligned}$$

□

The importance of Proposition 8 is that it justifies computations of distances using finite representations of GPs as approximations for the infinite-dimensional case.

Next, assuming the barycenter is non-degenerate, we show that we can also approximate the barycenter of a population of GPs by computing the barycenters of populations of GDs converging to these GPs. In the degenerate case, see [20, Thm. 11].

Theorem 9. *Assuming the barycenter of a population of GPs is non-degenerate, then it varies continuously, that is, the map* $(f_1, ..., f_N) \mapsto \bar{f}$ *is continuous in the operator norm. Especially, this implies that the barycenter* $\bar{f}_n$ *of the finite-dimensional restrictions* $\{f_{in}\}_{i=1}^N$ *converges to* $\bar{f}$*.*

First, we show that if $f_i \sim \mathcal{GP}(m_i, K_i)$ and $\bar{f} = \mathcal{GP}(\bar{m}, \bar{K})$, then that the map $(K_1, ..., K_N) \mapsto \bar{K}$ is continuous. Continuity of $(m_1, ..., m_N) \mapsto \bar{m}$ is clear.

Let K be a covariance operator, denote its maximal eigenvalue by $\lambda_{\max}(K)$. Note that this map is well-defined, as K is also bounded, normal operator, thus $\lambda_{\max}(K) = \|K\|_{op} < \infty$ holds. Now let $\mathbf{a} = (K_1, ..., K_N)$ be a population of covariance operators, denote i^{th} as $\mathbf{a}(i) = K_i$, then define the continuous function β and correspondence (a set valued map) Φ as follows

$$\beta : \mathbf{a} \mapsto \left(\sum_{i=1}^{N} \xi_i \sqrt{\lambda_{max}(\mathbf{a}(i))} \right)^2, \ \Phi : \mathbf{a} \mapsto K_{\beta(\mathbf{a})} = \{K \in \mathrm{HS}(\mathcal{H}) \mid \beta(\mathbf{a}) I \geq K \geq 0\}.$$

Then the fixed point of (6) can be found in $\Phi(\mathbf{a})$, as the map

$$F(K) = \sum_{i=1}^{N} \xi_i \left(K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \right)^{\frac{1}{2}},$$

is a compact operator, $\Phi(\mathbf{a})$ is bounded, and so the closure of $F(\Phi(\mathbf{a}))$ is compact. Furthermore, do note that F is a map from $\Phi(\mathbf{a})$ to itself, so by Schauder's fixed point theorem, there exists a fixed point.

Now, we want to show that this correspondence is continuous in order to put the Maximum theorem to use. A correspondence $\Phi : A \to B$ is *upper hemi-continuous* at $a \in A$, if all convergent sequences $(a_n) \in A$, $(b_n) \in \Phi(a_n)$ satisfy $\lim_{n\to\infty} b_n = b$, $\lim_{n\to\infty} a_n = a$ and $b \in \Phi(a)$. The correspondence is *lower hemi-continuous* at $a \in A$, if for all convergent sequences $a_n \to a$ in A and any $b \in \Phi(a)$, there is a subsequence a_{n_k}, so that we have a sequence $b_k \in \Phi(a_{n_k})$ which satisfies $b_k \to b$. If the correspondence is both upper and lower hemi-continuous, we say that it is *continuous*. For more about the Maximum theorem and hemi-continuity, see [2].

Lemma 10. *The correspondence* $\Phi : \mathbf{a} \mapsto K_{\beta(\mathbf{a})}$ *is continuous as correspondence.*

Proof. First, we show the correspondence is lower hemi-continuous. Let $(\mathbf{a}_n)_{n=1}^{\infty}$ be a sequence of populations of covariance operators of size N, that converges $\mathbf{a}_n \to \mathbf{a}$. Use the shorthand notation $\beta_n := \beta(\mathbf{a}_n)$, then $\beta_n \to \beta_\infty := \beta(\mathbf{a})$, and let $\mathbf{b} \in \Phi(\mathbf{a}) = K_{\beta_\infty}$.

Pick subsequence $(\mathbf{a}_{n_k})_{k=1}^{\infty}$ so that $(\beta_{n_k})_{k=1}^{\infty}$ is increasing or decreasing. If it was decreasing, then $K_{\beta_\infty} \subseteq K_{\beta_{n_k}}$ for every n_k. Thus the proof would be finished by choosing $\mathbf{b}_k = \mathbf{b}$ for every k. Hence assume the sequence is increasing, so that $K_{\beta_{n_k}} \subseteq K_{\beta_{n_{k+1}}}$. Now let $\gamma(t) = (1-t)\mathbf{b}_1 + t\mathbf{b}$, where $\mathbf{b}_1 \in K_{\beta_1}$, and let t_{n_k} be the solution to $(1-t)\beta_1 + t\beta_\infty = \beta_{n_k}$, then $\mathbf{b}_k := \gamma(t_{n_k}) \in K_{\beta_{n_k}}$ and $\mathbf{b}_k \to \mathbf{b}$.

For upper hemicontinuity, assume that $\mathbf{a}_n \to \mathbf{a}$, $\mathbf{b}_n \in K_{\beta_n}$ and that $\mathbf{b}_n \to \mathbf{b}$. Then using the definition of Φ, we get the positive sequence $\langle (\beta_n I - \mathbf{b}_n)x, x \rangle \geq 0$

indexed by n, then by continuity and the positivity of this sequence it follows that

$$0 \leq \lim_{n\to\infty} \langle(\beta_n I - \mathbf{b}_n)x, x\rangle = \langle(\beta_\infty I - \mathbf{b})x, x\rangle.$$

One can check the criterion $\mathbf{b} \geq 0$ similarly, and so we are done. □

Proof of Theorem 9. Now let $\mathbf{a} = (K_1, ..., K_n)$, $\mathbf{f}(K, \mathbf{a}) := \sum_{i=1}^N \xi_i W_2^2(K, K_i)$ and $F(K) := \sum_{i=1}^N \xi_i (K^{\frac{1}{2}} K_i K^{\frac{1}{2}})^{\frac{1}{2}}$, then the unique minimizer $\bar{K}$ of $\mathbf{f}$ is the fixed point of F. Furthermore, the closure $\mathrm{cl}(F(K_{\beta(\mathbf{a})}))$ is compact, $\mathbf{a} \mapsto \mathrm{cl}(F(K_{\beta(\mathbf{a})}))$ is a continuous correspondence as the closure of composition of two continuous correspondence. Additionally, we know that $\bar{K} \in \mathrm{cl}(F(K_{\beta(a)}))$, so applying the maximum theorem, we have shown that the barycenter of a population of covariance operators varies continuously, i.e. the map $(K_1, ..., K_N) \mapsto \bar{K}$ is continuous, finishing the proof. □

4 Experiments

We illustrate the utility of the Wasserstein metric in two different applications: Processing of uncertain white-matter tracts estimated from DWI, and analysis of climate development via temperature curve GPs.

Experimental setup. The white-matter tract GPs are estimated for a single subject from the Human Connectome Project [15, 32, 35], using probabilistic shortest-path tractography [17]. See the supplementary material for details on the data and its preprocessing. From daily minimum temperatures measured at a set of 30 randomly sampled Russian metereological stations [9, 34], GP regression was used to estimate a GP temperature curve per year and station for the period $1940 - 2009$ using maximum likelihood parameters. All code for computing Wasserstein distances and barycenters was implemented in MATLAB and ran on a laptop with 2,7 GHz Intel Core i5 processor and 8 GB 1867 MHz DDR3 memory. On the temperature GP curves (represented by 50 samples), the average runtime of the 2-Wasserstein distance computation was 0.048 ± 0.014 seconds (estimated from 1000 pairwise distance computations), and the average runtime of the 2-Wasserstein barycenter of a sample of size 10 was 0.69 ± 0.11 seconds (estimated from 200 samples).

White-matter tract processing. The *inferior longitudinal fasiculus* is a white-matter bundle which splits into two separate bundles. Fig. 3 (top) shows the results of agglomerative hierarchical clustering of the GP tracts using average Wasserstein distance. The per-cluster Wasserstein barycenter can be used to represent the tracts; its overlap with the individual GP mean curves is shown in Fig. 3 (bottom).

The individual GP tracts are visualized via their mean curves, but they are in fact a population of GPs. To confirm that the two clusters are indeed different also

when the covariance function is taken into account, we perform a permutation test for difference between per-cluster Wasserstein barycenters, and already with 50 permutations we observe a p-value of $p = 0.0196$, confirming that the two clusters are significantly different at a 5% significance level.

Quantifying climate change. Using the Wasserstein barycenters we perform nonparametric kernel regression to visualize how yearly temperature curves evolve with time, based on the Russian yearly temperature GPs. Fig. 4 shows snapshots from this evolution, and a continuous movie version `climate.avi` is found in the supplementary material. The regressed evolution indicates an increase in overall temperature as we reach the final year 2009. To quantify this observation, we perform a permutation test using the Wasserstein distance between population Wasserstein barycenters to compare the final 10 years 2000-2009 with the years 1940-1999. Using 50 permutations we obtain a p-value of 0.0392, giving significant difference in temperature curves at a 95% confidence level.

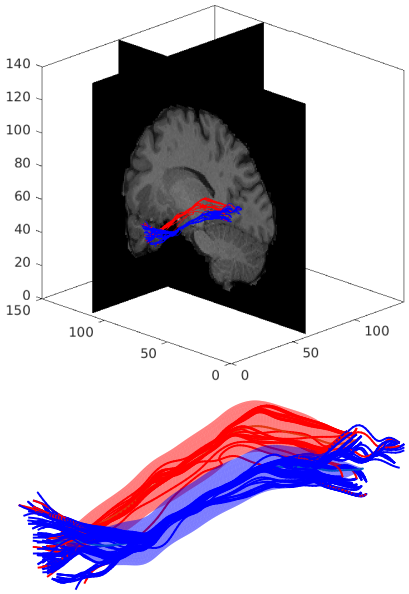


Figure 3: **Top:** The mean functions of the individual GPs, colored by cluster membership, in the context of the corresponding T1-weighted MRI slices. **Bottom:** The tract GP mean functions and the cluster mean GPs with 95% confidence bounds.

Significance. Note that the state-of-the-art in tract analysis as well as in functional data analysis would be to ignore the covariance of the estimated curves and treat the mean curves as observations. We contribute a framework to incorporate the uncertainty into the population analysis – but why would we want to retain uncertainty? In the white-matter tracts, the GP covariance represents spatial uncertainty in the estimated curve trajectory. The individual GPs represent connections between different endpoints. Thus, they do not represent observations of the exact same trajectory, but rather of distinct, nearby trajectories. It is common in diffusion MRI to represent such sets of estimated trajectories by a few prototype trajectories for visualization and comparative analysis; we obtain prototypes through the Wasserstein barycenter. To correctly interpret the spatial uncertainty, e.g. for a brain surgeon [8], it is crucial that the covariance of the prototype GP represents the covariances of the individual GPs, and not smaller. If you wanted to reduce uncertainty by increasing sample size, you would need more images, not more curves – because the noise is in the image. But more images are not usually available. In the climate data, the GP covariance models natural temperature variation, *not* measurement noise.

Increasing the sample size decreases the error of the temperature distribution, but should not decrease this natural variation (i.e. the covariance).

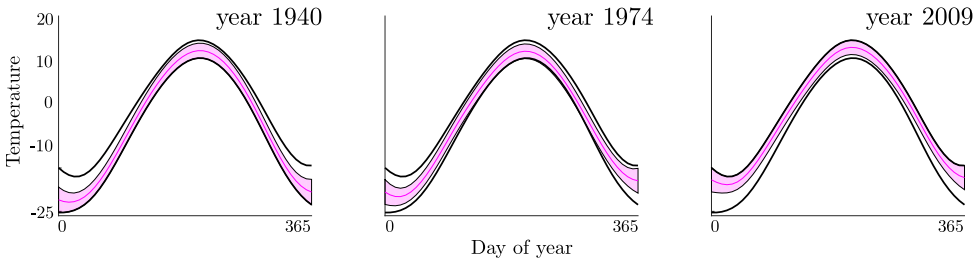


Figure 4: Snapshots from the kernel regression giving yearly temperature curves 1940-2009. We observe an apparent temperature increase which is confirmed by the permutation test.

5 Discussion and future work

We have shown that the Wasserstein metric for GPs is both theoretically and computationally well-founded for statistics on GPs: It defines unique barycenters, and allows efficient computations through finite-dimensional representations. We have illustrated its use in two different applications: Processing of uncertain estimates of white-matter trajectories in the brain, and analysis of climate development via GP representations of temperature curves. We have seen that the metric itself is discriminative for clustering and permutation testing, and we have seen how the GP barycenters allow truthful interpretation of uncertainty in the white matter tracts and of variation in the temperature curves.

Future work includes more complex learning algorithms, starting with preprocessing tools such as PCA [31], and moving on to supervised predictive models. This includes a better understanding of the potentially Riemannian structure of the infinite-dimensional Wasserstein space, which would enable us to draw on existing results for learning with manifold-valued data [21].

The Wasserstein distance allows the inherent uncertainty in the estimated GP data points to be appropriately accounted for in every step of the analysis, giving truthful analysis and subsequent interpretation. This is particularly important in applications where uncertainty or variation is crucial: Variation in temperature is an important feature in climate change, and while estimated white-matter trajectories are known to be unreliable, they are used in surgical planning, making uncertainty about their trajectories a highly relevant parameter.

6 Acknowledgements

This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Data were provided

[in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The authors would also like to thank Mads Nielsen for valuable discussions and supervision. Finally, the authors would like to thank Victor Panaretos for valuable discussions and, in particular, for pointing out an error in an earlier version of the manuscript.

References

[1] Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis **43**(2), 904–924 (2011)

[2] Aliprantis, C., Border, K.: Infinite dimensional analysis: a hitchhiker's guide. Studies in Economic Theory **4** (1999)

[3] Álvarez-Esteban, P., Del Barrio, E., Cuesta-Albertos, J., Matrán, C., et al.: Uniqueness and approximate computation of optimal incomplete transportation plans. In: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. vol. 47, pp. 358–375. Institut Henri Poincaré (2011)

[4] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J., Matrán, C.: A fixed-point approach to barycenters in Wasserstein space. Journal of Mathematical Analysis and Applications **441**(2), 744–762 (2016)

[5] Ambrosio, L., Gigli, N.: A user's guide to optimal transport. In: Modelling and optimisation of flows on networks, pp. 1–155. Springer (2013)

[6] Ambrosio, L., Gigli, N., Savaré, G.: Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media (2008)

[7] Arveson, W.: A short course on spectral theory, vol. 209. Springer Science & Business Media (2006)

[8] Berman, J.: Diffusion MR tractography as a tool for surgical planning. Magnetic resonance imaging clinics of North America **17**(2), 205–214 (2009)

[9] Bulygina, O., Razuvaev, V.: Daily temperature and precipitation data for 518 russian meteorological stations. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee (2012)

[10] Cuesta-Albertos, J., Matrán-Bea, C., Tuero-Diaz, A.: On lower bounds for the l^2-Wasserstein metric in a Hilbert space. Journal of Theoretical Probability **9**(2), 263–283 (1996)

[11] Dowson, D., Landau, B.: The Fréchet distance between multivariate normal distributions. Journal of multivariate analysis **12**(3), 450–455 (1982)

[12] Faraki, M., Harandi, M.T., Porikli, F.: Approximate infinite-dimensional region covariance descriptors for image classification. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. pp. 1364–1368. IEEE (2015)

[13] Gelbrich, M.: On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. Mathematische Nachrichten **147**(1), 185–203 (1990)

[14] Givens, C.R., Shortt, R.M., et al.: A class of Wasserstein metrics for probability distributions. The Michigan Mathematical Journal **31**(2), 231–240 (1984)

[15] Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the Human Connectome project. Neuroimage **80**, 105–124 (2013)

[16] Harandi, M., Salzmann, M., Porikli, F.: Bregman divergences for infinite dimensional covariance matrices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1003–1010 (2014)

[17] Hauberg, S., Schober, M., Liptrot, M., Hennig, P., Feragen, A.: A random Riemannian metric for probabilistic shortest-path tractography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 597–604. Springer (2015)

[18] Knott, M., Smith, C.S.: On the optimal mapping of distributions. Journal of Optimization Theory and Applications **43**(1), 39–49 (1984)

[19] Lê, M., Unkelbach, J., Ayache, N., Delingette, H.: GPSSI: Gaussian process for sampling segmentations of images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 38–46. Springer (2015)

[20] Masarotto, V., Panaretos, V.M., Zemel, Y.: Procrustes metrics on covariance operators and optimal transportation of gaussian processes. arXiv preprint arXiv:1801.01990 (2018)

[21] Masci, J., Boscaini, D., Bronstein, M., Vandergheynst, P.: Geodesic convolutional neural networks on Riemannian manifolds. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 37–45 (2015)

[22] Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra and its Applications **48**, 257–263 (1982)

[23] Pigoli, D., Aston, J.A., Dryden, I.L., Secchi, P.: Distances and inference for covariance operators. Biometrika **101**(2), 409–422 (2014)

[24] Pujol, S., Wells, W., Pierpaoli, C., Brun, C., Gee, J., Cheng, G., Vemuri, B., Commowick, O., Prima, S., Stamm, A., et al.: The DTI challenge:

toward standardized evaluation of diffusion tensor imaging tractography for neurosurgery. Journal of Neuroimaging **25**(6), 875–882 (2015)

[25] Quang, M.H., Murino, V.: From covariance matrices to covariance operators: Data representation from finite to infinite-dimensional settings. In: Algorithmic Advances in Riemannian Geometry and Applications, pp. 115–143. Springer (2016)

[26] Quang, M.H., San Biagio, M., Murino, V.: Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In: Advances in Neural Information Processing Systems. pp. 388–396 (2014)

[27] Rajput, B.S.: Gaussian measures on Lp spaces, $1 \leq p < \infty$. Journal of Multivariate Analysis **2**(4), 382–403 (1972)

[28] Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., Aigrain, S.: Gaussian processes for time-series modelling. Phil. Trans. R. Soc. A **371**(1984), 20110550 (2013)

[29] Schober, M., Duvenaud, D.K., Hennig, P.: Probabilistic ODE solvers with Runge-Kutta means. In: Advances in neural information processing systems. pp. 739–747 (2014)

[30] Schober, M., Kasenburg, N., Feragen, A., Hennig, P., Hauberg, S.: Probabilistic shortest path tractography in DTI using Gaussian Process ODE solvers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 265–272. Springer (2014)

[31] Seguy, V., Cuturi, M.: Principal geodesic analysis for probability measures under the optimal transport metric. In: Advances in Neural Information Processing Systems. pp. 3312–3320 (2015)

[32] Sotiropoulos, S., Moeller, S., Jbabdi, S., Xu, J., Andersson, J., Auerbach, E., Yacoub, E., Feinberg, D., Setsompop, K., Wald, L., et al.: Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE. Magnetic resonance in medicine **70**(6), 1682–1689 (2013)

[33] Takatsu, A., et al.: Wasserstein geometry of Gaussian measures. Osaka Journal of Mathematics **48**(4), 1005–1026 (2011)

[34] Tatusko, R., Mirabito, J.A.: Cooperation in climate research: An evaluation of the activities conducted under the US-USSR agreement for environmental protection since 1974. National Climate Program Office (1990)

[35] Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn Human Connectome project: an overview. Neuroimage **80**, 62–79 (2013)

[36] Villani, C.: Topics in optimal transportation. No. 58, American Mathematical Soc. (2003)

[37] Wassermann, D., Bloy, L., Kanterakis, E., Verma, R., Deriche, R.: Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers. NeuroImage **51**(1), 228–241 (2010)

[38] Yang, X., Niethammer, M.: Uncertainty quantification for LDDMM using a low-rank Hessian approximation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 289–296. Springer (2015)

Appendix B

Wrapped Gaussian Process Regression on Riemannian Manifolds

The following chapter presents (up to formatting) the article

> Anton Mallasto, and Aasa Feragen. "Wrapped Gaussian process regression on Riemannian manifolds." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

This work introduces wrapped Gaussian processes (WGPs), which generalize GPs to a Riemannian manifold $\mathcal{M}$ as a collection of random points on the manifold, such that any finite subcollection of size N forms a WGD on the product manifold $\mathcal{M}^N$. An explicit expression for the conditional distribution of a WGD is derived, which is then utilized in supervised learning through WGP regression. The method is demonstrated on data on the 2-sphere, on a DTI dataset lying in the space of symmetric positive definite matrices, and landmark shape data living in the Kendall shape space.

The fundamental contribution of the work is the generalization of GPs to manifolds as WGPs, as this opens the door for the rich statistical machinery of GPs, providing flexible non-parametric, non-linear probabilistic models, to be applied in manifold valued statistics.

On the other hand, the particular application in regression continues in the tradition of generalizing regression in Euclidean spaces to manifolds, which includes methods such as the geodesic regression [23] and its probabilistic counterpart [25]. Other relevant non-geodesic work includes kernel based methods [11], and especially the local kriging method introduced in [55], which closely relates to the WGP regression method introduced here.

Wrapped Gaussian Process Regression on Riemannian Manifolds

Anton Mallasto Aasa Feragen

Department of Computer Science, University of Copenhagen

Abstract

Gaussian process (GP) regression is a powerful tool in non-parametric regression providing uncertainty estimates. However, it is limited to data in vector spaces. In fields such as shape analysis and diffusion tensor imaging, the data often lies on a manifold, making GP regression non-viable, as the resulting predictive distribution does not live in the correct geometric space. We tackle the problem by defining wrapped Gaussian processes (WGPs) on Riemannian manifolds, using the probabilistic setting to generalize GP regression to the context of manifold-valued targets. The method is validated empirically on diffusion weighted imaging (DWI) data, directional data on the sphere and in the Kendall shape space, endorsing WGP regression as an efficient and flexible tool for manifold-valued regression.

1 Introduction

Regressing functions from Euclidean training data $\{(x_i, y_i)\}_{i=1}^N$ is well studied. Manifold-valued y_i, on the other hand, pose difficulties due to the lack of the vector space structure: Euclidean statistics do not respect the intrinsic structure of manifold-valued data, and the product of inference might not belong to the object category of the data. For example, see Fig. 1, where Gaussian process regression escapes the 2-sphere.

Sometimes the data observed is uncertain. In this case, it is favorable to estimate a distribution over possible regressed functions, yielding uncertainty estimates of the resulting inference. Gaussian process (GP) regression achieves this in a tractable manner. Furthermore, GP regression is an example of Bayesian inference, where it is possible to incorporate prior knowledge to aid the inference. These qualitative properties motivate us to generalize GP regression to Riemannian manifolds.

Related work. Fletcher [5] generalized linear regression to handle manifold-valued data with real covariates by *geodesic regression*; this was later extended

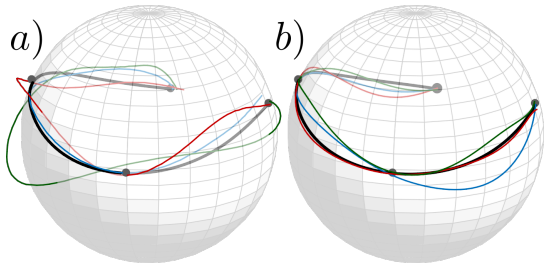


Figure 1: **Why geometrically intrinsic regression is important.** Consider data points (black) on the 2-sphere. In a), we apply ordinary GP regression. The black curve is the prediction and the colorful curves are samples from the predictive distribution, which clearly escape the sphere. In b), we visualize the result using WGP regression, which respects the geometrical constraints of the data.

Method	Non-geod.	Priors	Uncert.	Global
Geod. reg.[5, 16, 23]	No	No	No	Yes
Poly. reg.[9]	Yes	No	No	Yes
Mani. Kriging [23]	Yes	Yes	Yes	No
Kernel reg. [2, 3]	Yes	Yes	No	Yes
Stoch. dev. [17]	No	No	Yes	Yes
Hong et al.[10]	No	Yes	Yes	Yes
WGP reg.	Yes	Yes	Yes	Yes

Table 1: Qualitative comparison of manifold regression models mentioned in this paper. *Global* means, that regression is not carried out in a single tangent space, uncert. is short for uncertainty and Non-geod short for non-geodesic.

to include multi-dimensional covariates [16]. Prior work also consider uncertainty estimates for geodesic regression; by a Kalman filter approach [10] and by stochastic development [17].

Manifold-valued data, however, does not always follow a geodesic trend. Approaches for this non-geodesic setting include kernel-based approaches [2, 3] and a generalization of polynomial regression [9]. Unfortunately, these models do not provide uncertainty estimates.

Improving on this, Pigoli et al. [23] consider a kriging (GP regression) method. The method uses multivariate geodesic regression to form a reference coordinate system, which is used to compute residuals of the manifold-valued data points. Regular GP regression is then applied on the residuals and the result is mapped back onto the manifold. The procedure, however, depends heavily on the localization of the problem to a single tangent space, and does not offer an intrinsic probabilistic interpretation. Relying on WGPs, our method offers interpretability, and the prior basepoint function used in WGP regression allows avoiding being too local. Furthermore, the kriging method in [23] took advantage of the geodesic submanifold regression to initialize a reference coordinate system. Our method, enables one to take advantage of more general priors, including the use of geodesic submanifold regression.

Steinke and Hein [27] consider the problem of approximating a function between

manifolds via minimizing regularized empirical risk. In this setting, also the independent variables are manifold-valued. The WGP regression proposed in this paper can be extended to this setting, as long as a kernel can be defined on the domain, carrying on all the advantages of WGPs mentioned.

Wrapped Gaussian processes appear in directional statistics [13], where a wrapped normal distribution is defined on a 1-sphere S^1, which is then generalized to a multivariate version, and this is then used to define a WGP. This is a special case of our setting, when the manifold is chosen to be the torus $S^1 \times S^1 \times ... \times S^1$.

The contribution can be summarized as follows: We generalize GPs to Riemannian manifolds as wrapped Gaussian procesess (WGPs), and provide a novel framework for non-parametric regression with uncertainty estimates using WGP regression. We demonstrate the method in Section 5 on the 2-sphere by considering a toy example and orientations of the left femur of a walking person, on the manifold of symmetric positive definite matrices for DTI upsampling, and on Kendall shape space, using a data set of Corpus Callosum shapes. The method is analytically tractable for manifolds with infinite injectivity radius, such as manifolds with non-positive curvature. Otherwise, we suggest the approximation in Remark 2. Computationally, the method is relatively cheap, as the only addition compared to GP regression is a single application of the logarithmic map per data point and single exponential map per predicted point.

2 Preliminaries

We briefly summarize the mathematical prerequisities needed. First, we recall how GPs are used in non-parametric regression in the Euclidean case, after which we turn to basic concepts in Riemannian geometry and briefly discuss geodesic submanifold regression.

2.1 Gaussian process regression

Denote by $\mathcal{N}(\mu, \Sigma)$ the multivariate Gaussian distribution with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, and write the probability density function p as $p(v) = \mathcal{N}(v|\mu, \Sigma)$ for $v \in \mathbb{R}^n$.

A *Gaussian process* (GP) [24] is a collection f of random variables, such that any finite subcollection $(f(\omega_i))_{i=1}^N$ has a joint Gaussian distribution, where $\omega_i \in \Omega \subset \mathbb{R}^l$, and Ω is the *index set*. A GP is entirely characterized by the pair

$$m(\omega) = \mathbb{E}\left[f(\omega)\right], \tag{1}$$

$$k(\omega, \omega') = \mathbb{E}\left[(f(\omega) - m(\omega))(f(\omega') - m(\omega'))^T\right], \tag{2}$$

where m and k are called the *mean function* and *covariance function*, respectively. We denote such a GP by $f \sim \mathcal{GP}(m, k)$. It follows from the defini-

tion that the covariance function (*kernel*) k is symmetric and positive semidefinite.

Let $\boldsymbol{D} = \{(x_i, y_i) \mid x_i \in \boldsymbol{x} \subset \mathbb{R}^l,\ y_i \in \boldsymbol{y} \subset \mathbb{R}^n\}$ be the training data. The GP predictive distribution for outputs $\boldsymbol{y}_*$ at the test inputs $\boldsymbol{x}_*$, given in vector form, is

$$p(\boldsymbol{y}_*|\boldsymbol{D}, \boldsymbol{x}_*) = \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*), \tag{3}$$

$$\boldsymbol{\mu}_* = \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{y}, \tag{4}$$

$$\Sigma_* = \boldsymbol{k}_{**} - \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{k}_*, \tag{5}$$

where, given a kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we use the notation $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}, \boldsymbol{x}_*)$, $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ and K_{err} is the measurement error variance. In the notation above, the function and k is applied elementwise on the vectors $\boldsymbol{x}, \boldsymbol{x}_*$.

Typically in model selection, the kernel k is picked from a parametric family $\{k_\theta|\theta \in \Theta\}$ of covariance functions, such as the *radial basis function* (RBF) kernels

$$k_{\sigma^2,\lambda}(x, y) = \sigma^2 \exp\left(-\frac{\|x - y\|^2}{2\lambda}\right),\ \sigma^2, \lambda > 0, \tag{6}$$

choosing the parameters (σ^2, λ) so that the *marginal likelihood* $\mathbb{P}\{\boldsymbol{y}|(\sigma^2, \lambda)\}$ is maximized.

2.2 Riemannian geometry

To fix notation, we briefly present the essentials of Riemannian geometry. For a thorough presentation, see [4]. A *Riemannian manifold* is a smooth manifold M with a smoothly varying inner product $g_p(\cdot, \cdot)$ (we will often use the notation $\langle \cdot, \cdot \rangle_p$) on the tangent space T_pM at each $p \in M$, called a *Riemannian metric*, inducing the distance function d between points on the M. Each element (p, v) in the tangent bundle $TM = \bigcup_{p \in M} (p \times T_pM)$ defines a geodesic γ (a curve locally minimizing distance between two points) on M, so that $\gamma(0) = p$ and $\frac{d}{dt}\gamma(t) \mid_{t=0} = v$. The *exponential map* $\text{Exp} : TM \to M$ given by $(p, v) \mapsto \text{Exp}_p(v) = \gamma(1)$, where γ is the geodesic corresponding to (p, v). The exponential map Exp_p at p is a diffeomorphism between a neighborhood $0 \in U \subset T_pM$ and neighbourhood $p \in V \subset M$, which is chosen in a maximal way, so if $V \subsetneq V'$, then a diffeomorphism between V' and a neighborhood in the tangent space cannot be defined anymore. We also call V the *area of injectivity*.

We can define the inverse map $\text{Log}_p : V \to T_pM$, characterized by

$$\text{Exp}_p(\text{Log}_p(p')) = p'. \tag{7}$$

Outside of V, we use $\text{Log}_p(p')$ to denote a smallest $v \in T_pM$ chosen in a *measurable*, consistent way. We call the the minimun distance from p to the boundary of a maximal V the *injectivity radius* of Exp_p and the complement of V in M

the *cut-locus* at p denoted by $\mathcal{C}_p$. The manifolds with non-positive curvature form an important class of manifolds with infinite injectivity radius, that is, they have an empty cut-locus $\mathcal{C}_p$ for every $p \in M$.

Let M_i be Riemannian manifolds with metrics g_i, exponential maps Exp^i and logarithmic maps Log^i for $i = 1, 2$. Then $M = M_1 \times M_2$ turns into a Riemannian manifold when endowed with the metric $g = g_1 + g_2$, which has the component-wise computed exponential map $\mathrm{Exp}_{(p_1,p_2)}((v_1, v_2)) = \big(\mathrm{Exp}^1_{p_1}(v_1), \mathrm{Exp}^2_{p_2}(v_2)\big)$, akin to the logarithmic map Log on the product manifold.

2.2.1 Probabilistic notions

Let X be a random point on a Riemannian manifold M, the set

$$\mathbb{E}[X] := \left\{ p \mid p \in \arg\min_{q \in M} (\mathbb{E}[d(q, X)^2]) \right\}. \tag{8}$$

is called the *Fréchet means* of X. If there is a unique mean $\bar{p}$, then by abuse of notation we write $\mathbb{E}[X] = \bar{p}$. Given a data set $\boldsymbol{p} = \{p_i \in M\}_{i=1}^N$, an *empirical Fréchet mean* is a minimizer of the quantity

$$\min_{q \in M} \sum_{i=1}^{N} d(q, p_i)^2. \tag{9}$$

The set of empirical Fréchet means is denoted by $\mathbb{E}[\boldsymbol{p}]$.

Given two probability spaces $(\mathcal{X}_i, \mathcal{S}_i, \nu_i)$ for $i = 1, 2$ and a measurable map $F : \mathcal{X}_1 \to \mathcal{X}_2$, we say that the measure ν_2 is the push-forward of the measure ν_1 with respect to F, if $\nu_2(A) = \nu_1(F^{-1}(A))$ for every A in the sigma-algebra $\mathcal{S}_2$. We denote this by $\nu_2 = F_{\#}\nu_1$.

For more about intrinsic statistics on manifolds, see [21].

2.2.2 Geodesic submanifold regression

Geodesic regression on a Riemannian manifold M was introduced by Fletcher [5]. It is a generalization of linear regression, that seeks the geodesic parametrized by $(p, v) \in TM$ that minimizes the quantity

$$E(p, v) = \frac{1}{2} \sum_{i=1}^{N} d(\mathrm{Exp}_p(t_i v), p_i)^2, \tag{10}$$

given the training data $(t_i, p_i) \in \mathbb{R} \times M$ for $i = 1, ..., N$.

This framework has been generalized to deal with more covariates [16]; assume we are given data $(x_i, p_i) \in \mathbb{R}^l \times M$ for $i = 1, ..., N$. Then, we want to solve for the submanifold γ parametrized by $(p, v_1, ..., v_l)$ that minimizes

$$E(p, v_1, ..., v_l) = \frac{1}{2} \sum_{i=1}^{N} d\left(\mathrm{Exp}_p \left(\sum_{j=1}^{l} x_i(j) v_j \right), p_i \right)^2. \tag{11}$$

This is analogous to fitting a hyperplane in the Euclidean case. Another generalization for multiple independent variables was carried out in [23]. Later on in this work, we propose a way to construct priors for the GP regression on manifolds by regressing a geodesic model.

Tangent space geodesic regression is a Naïve generalization of linear regression, achieved by linearizing the space by picking $p \in M$, transforming the data set $(x_i, p_i) \in \mathbb{R}^l \times M$ for $i = 1, ..., N$ into images of the Riemannian logarithmic map at p. Then, one can carry out linear regression in the tangent space and map the result onto the manifold using the exponential map, yielding a quick approximation of geodesic submanifold regression.

3 Wrapped Gaussian processes

We are now ready to introduce *wrapped Gaussian distributions* (WGDs), computing the conditional distribution of two jointly WGD random points on the manifold. This is an essential part of wrapped Gaussian process (WGP) regression on manifolds introduced in the next chapter, alike in the Euclidean case. In this chapter we also introduce WGPs in a formal way, without studying their properties further.

3.1 Wrapped Gaussian distributions

Wrapped Gaussian distributions (WGDs) originated in directional statistics [18]. There exist multiple different ways of generalizing Gaussian distributions to manifolds. For example, Sommer [25] uses an instrinsic, anisotropic diffusion process for the generalization. Pennec [20], on the other hand, generalizes the Gaussian as the distribution maximizing entropy with a fixed mean and covariance. WGDs rely on linearizing the manifold through a wrapping function, in our case the Riemannian exponential map.

Let (M, d) be an n-dimensional Riemannian manifold. We say that a random point X on M follows a *wrapped Gaussian distribution* (WGD), if for some $\mu \in M$ and symmetric positive definite matrix $K \in \mathbb{R}^{n \times n}$

$$X \sim \left(\mathrm{Exp}_\mu\right)_\# \left(\mathcal{N}(0, K)\right), \tag{12}$$

denoted by $X \sim \mathcal{N}_M(\mu, K)$. To sample from this distribution, draw v from $\mathcal{N}(0, K)$ and map the sample to the manifold by $\mathrm{Exp}_\mu(v)$. Now, define *the basepoint* and *tangent space covariance* of X as

$$\mu_{\mathcal{N}_M}(X) := \mu, \ \mathrm{Cov}_{\mathcal{N}_M}(X) := K. \tag{13}$$

In the case of infinite injectivity radius $\mu_{\mathcal{N}_M}(X) \in \mathbb{E}[X]$, but not in general [19, Prop. 2.11]. The random points $X_i \sim \mathcal{N}_{M_i}(\mu_i, K_i)$, $i = 1, 2$, are *jointly WGD*,

if the random point (X_1, X_2) on $M_1 \times M_2$ is WGD, that is,

$$(X_1, X_2) \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}K_1 & K_{12}\\ K_{21} & K_2\end{pmatrix}\right), \tag{14}$$

for some matrix $K_{12} = K_{21}^T$.

We now compute the conditional distribution of two jointly WGD random points, which is the core of WGP regression in Section 4.

Theorem 1. *Assume X_1, X_2 are jointly WGD as in* (14)*, then we have the conditional distribution*

$$X_1|(X_2 = p_2) \sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#}\left(\sum_{v\in A}\lambda_v \mathcal{N}\left(\mu_v, K_v\right)\right), \tag{15}$$

where

$$\begin{aligned}
\mu_v &= K_{12}K_2^{-1}v,\\
K_v &= K_1 - K_{12}K_2^{-1}K_{12}^T,\\
\lambda_v &= \frac{\mathcal{N}(v|\mathbf{0}, K_2)}{\mathbb{P}\{A\}},\\
A &= \{v \in T_{\mu_2}M \mid \mathrm{Exp}_{\mu_2}(v) = p_2\},\\
\mathbb{P}\{A\} &= \sum_{v\in A}\mathcal{N}(v|\mathbf{0}, K_2).
\end{aligned} \tag{16}$$

Proof. Pick $p_1 \in M$. Let $B = \mathrm{Exp}_{\mu_1}^{-1}(p_1)$ be the preimage of p_1 in $T_{\mu_1}M$, similarly $A = \mathrm{Exp}_{\mu_2}^{-1}(p_2)$ as above for p_2, and furthermore K be the tangent space covariance of (X_1, X_2) given in (14), then

$$\begin{aligned}
&\mathbb{P}\{X_1 = p_1|(X_2 = p_2)\}\\
=&\frac{\mathbb{P}\{u \in B, v \in A\}}{\mathbb{P}\{v \in A\}}\\
=&\sum_{v\in A, u\in B}\frac{\mathcal{N}(v|\mathbf{0}, K_2)}{\mathbb{P}\{A\}}\frac{\mathcal{N}\left((u,v)|\mathbf{0}, K\right)}{\mathcal{N}(v|\mathbf{0}, K_2)}\\
=&\sum_{v\in A, u\in B}\lambda_v\mathcal{N}\left(u|\mu_v, K_v\right)\\
=&\mathbb{P}\{Z = p_1\},
\end{aligned} \tag{17}$$

where $Z \sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#}\left(\sum_{v\in A}\lambda_v\mathcal{N}\left(\mu_v, K_v\right)\right)$, and $\mathcal{N}(u|\mu_v, K_v)$ is the predictive distribution calculated as in the Euclidean case in (3). □

Remark 2. *If the injectivity radius of the exponential map is infinite, then*

$$\begin{aligned}
&X_1|(X_2 = p_2)\\
&\quad\sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#}\left(\mathcal{N}\left(\mu_{\mathrm{Log}_{\mu_2}(p_2)}, K_{\mathrm{Log}_{\mu_2}(p_2)}\right)\right),
\end{aligned} \tag{18}$$

following the notation in (16)*. Furthermore, if the probability mass on the area of injectivity of the exponential map is large enough, we can use this expression*

as a reasonable approximation for the predictive distribution, as the Gaussian mixture distribution in the tangent space can be well approximated by a single Gaussian.

3.2 Wrapped Gaussian processes

A collection f of random points on a manifold M indexed over a set Ω is a *wrapped Gaussian process* (WGP), if every finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly WGD on M^N. We define

$$m(\omega) := \mu_{\mathcal{N}_M}(f(\omega)) \tag{19}$$
$$k(\omega, \omega') := \mathrm{Cov}_{\mathcal{N}_M}(f(\omega), f(\omega')), \tag{20}$$

called the *basepoint function* (BPF) and *tangent space covariance function* (TSCF) of f, respectively. The restriction we have on Ω, is being able to define a kernel on it.

A WGP f can be viewed as a WGD on the possibly infinite-dimensional product manifold $M^{|\Omega|}$. To elaborate, formally one can state

$$f \sim (\mathrm{Exp}_m)_\#(\mathcal{GP}(0, k)). \tag{21}$$

The difference is, that the tangent space distribution is a GP instead of a GD. The WGP is entirely characterized by the pair (m, k), similar to the Euclidean case. Therefore, we introduce the notation $f \sim \mathcal{GP}_M(m, k)$.

4 Gaussian process inference on manifolds

In the following, we discuss two different methods of GP regression on a Riemannian manifold M with infinite injectivity radius (or using the approximation in Remark 2), given the noise-free training data

$$\boldsymbol{D}_M = \{(x_i, p_i) \mid x_i \in \mathbb{R}^l,\ p_i \in M,\ i = 1, ..., N\}. \tag{22}$$

For shorthand notation, we denote $\boldsymbol{x} = (x_i)_{i=1}^N$ and $\boldsymbol{p} = (p_i)_{i=1}^N$. Additionally, $\boldsymbol{x}_*$ is used for the test inputs, and $\boldsymbol{p}_*$ for the test outputs. Later, we remark that the first approach is actually a special case of the latter one, see Fig. 2.

4.1 Naïve tangent space approach

Choose $p \in M$ (typically $p \in \mathbb{E}[\boldsymbol{p}]$), and transfrom the training data $\boldsymbol{D}_M$ into $\boldsymbol{D}_{T_pM}$ by

$$\boldsymbol{D}_{T_pM} = (\boldsymbol{x}, \boldsymbol{y}) := \{(x_i, y_i) \mid y_i = \mathrm{Log}_p(p_i)\}, \tag{23}$$

see Fig. 2 a). As $\boldsymbol{D}_{T_pM} \subset \mathbb{R}^l \times T_pM$ now lives in a Euclidean space, fit a GP $f_{euc} \sim \mathcal{GP}(m_{euc}, k_{euc})$ to the data using GP regression, resulting in the

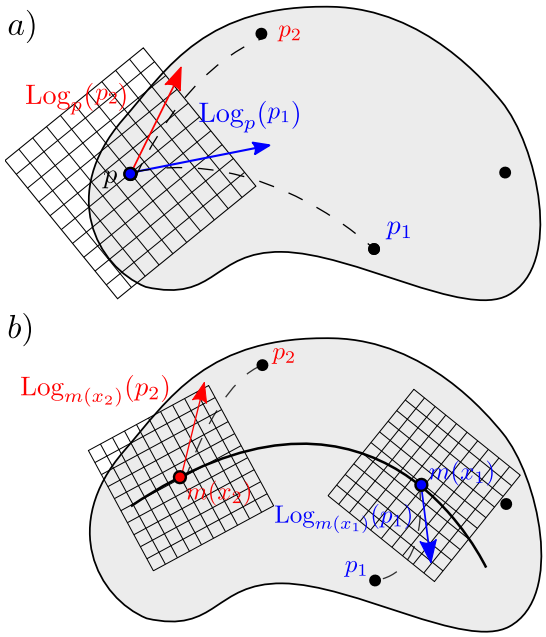


Figure 2: a) Tangent space GP data transformation. Data point p_i (in black) is transformed into $\mathrm{Log}_p(p_i) \in T_pM$. This can be seen as a special case of WGP regression, with a fixed prior BPF $m(x) = p$. In b), the data transformation is visualized with a more general prior BPF m (black curve).

predictive distribution $\boldsymbol{y}_*|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)$. Then, reversing the previous data transformation, we can map the random vector to a random point $\boldsymbol{p}_*|\boldsymbol{p}$ on the manifold M, resulting in

$$\boldsymbol{p}_*|\boldsymbol{p} = \mathrm{Exp}_p(\boldsymbol{y}_*) \sim \left(\mathrm{Exp}_p\right)_{\#} \left(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)\right). \tag{24}$$

4.2 Wrapped Gaussian process regression

Now we generalize GP regression inside a probabilistic framework, relying on the results presented in Section 3, by assuming a WGP prior $f_{prior} \sim \mathcal{GP}_M(m, k)$. According to the prior, the joint distribution between the training outputs $\boldsymbol{p}$ and test outputs $\boldsymbol{p}_*$ at $\boldsymbol{x}_*$ is given by

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2} \left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} \end{pmatrix} \right), \tag{25}$$

where $\boldsymbol{m} = m(\boldsymbol{x})$, $\boldsymbol{m}_* = m(\boldsymbol{x}_*)$, $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}_*, \boldsymbol{x})$, and $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$. Therefore, by Theorem 1 and using the approximation in Remark 2 (if necessary)

$$\begin{aligned} \boldsymbol{p}_*|\boldsymbol{p} &\sim \left(\mathrm{Exp}_{\boldsymbol{m}_*}\right)_{\#} \left(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)\right), \\ \boldsymbol{\mu}_* &= \boldsymbol{k}_* \boldsymbol{k}^{-1} \mathrm{Log}_{\boldsymbol{m}} \boldsymbol{p}, \\ \Sigma_* &= \boldsymbol{k}_{**} - \boldsymbol{k}_* \boldsymbol{k}^{-1} \boldsymbol{k}_*^T. \end{aligned} \tag{26}$$

The predictive distribution $\boldsymbol{p}_*|\boldsymbol{p}$ is not necessarily WGD, as $\boldsymbol{\mu}_*$ might be non-zero. The distribution can be sampled from, but computing exactly quantities

such as $\mathbb{E}[\boldsymbol{p}_*|\boldsymbol{p}]$ is not trivial. As in [8, Sect. 3.1.1], the distribution can be approximated via Riemannian unscented transform or by using a WGD with the basepoint at $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ and parallel transporting the tangent space covariance to this point along the geodesic $\gamma(t) = \mathrm{Exp}_{\boldsymbol{m}_*}(t\boldsymbol{\mu}_*)$.

Remark 3. $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ *is not necessarily a Fréchet mean of* $\boldsymbol{p}_*|\boldsymbol{p}$*. However, it is the* maximum a posteriori *(MAP) estimate. For this reason, we will use* $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ *as a point prediction in Section 5.*

4.2.1 Choosing a prior

The prior WGP $f_{prior} \sim \mathcal{GP}_M(m, k)$ indexed over Ω is chosen by picking a kernel k on Ω to be the TSCF, and picking a BPF m so that p and $m(x_i)$ live in the same connected component of M for every data-point (x_i, p_i).

In Section 5, two kinds of prior BPFs are used. The first BPF m_1 is a generalization of a centered GP, given by $m_1(\omega) = \bar{p}$, for all $x \in \Omega$ and a $\bar{p} \in \mathbb{E}[\boldsymbol{p}]$. The second kind m_2, uses a previous regression (such as geodesic submanifold regression) γ on the dataset $\boldsymbol{D}_M$. That is, $m_2(\omega) = \gamma(\omega)$ for all $\omega \in \Omega$. For computational reasons, we only consider TSCFs that assume each tangent space coordinate independent, resulting in the *diagonal RBF* kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{diag}(k_1(\boldsymbol{x}, \boldsymbol{x}'), k_2(\boldsymbol{x}, \boldsymbol{x}'), ..., k_n(\boldsymbol{x}, \boldsymbol{x}')), \tag{27}$$

where each k_i are chosen to be RBF kernels, $\mathrm{diag}(A, B)$ is a block-diagonal matrix with blocks A and B, $\boldsymbol{x}, \boldsymbol{x'} \subset \Omega$, and n is the dimension of M. The diagonal RBF yields uncertainty estimates, but not a generative model, as this would need covariance between coordinates.

Optimizing hyperparameters. We choose the TSCF from a parametric family of kernels $\{k_\theta\}_{\theta \in \Theta}$ maximizing the *marginal likelihood*, as in the Euclidean case. In the setting of WGPs, the marginal likelihood becomes

$$\mathbb{P}\{\boldsymbol{p}|\theta\} = \sum_{v \in \mathrm{Exp}_{\boldsymbol{m}}^{-1}(\boldsymbol{p})} \mathcal{N}(v|\boldsymbol{0}, K_\theta), \tag{28}$$

where $K_\theta = k_\theta(\boldsymbol{x}, \boldsymbol{x})$. To improve the approximation discussed in Remark 2, we propose to maximize the quantity

$$\mathbb{P}\{\boldsymbol{p}|\theta\} \approx \mathcal{N}\left(\mathrm{Log}_{\boldsymbol{m}}(\boldsymbol{y})|\boldsymbol{0}, K_\theta\right), \tag{29}$$

as maximizing this quantity increases the probability mass given by the prior distribution to the area of injectivity. The diagonal RBF kernel (Eq. (27)) can be optimized by choosing each k_i to maximize the marginal likelihood of the respective tangent space coordinate independently. That is, k_i is chosen to maximize the marginal likelihood of the data set $\left\{\left(x_j, \pi_i\left(\mathrm{Log}_{m(x_j)}(p_j)\right)\right)\right\}_{j=1}^N$, where π_i is the projection onto the ith component.

A part of engineering the kernel is to pick a frame for the manifold. A frame is a smooth map $\rho : M \to \mathbb{R}^{n \times n}$, so that the columns of $\rho(p)$ form an orthonormal

basis for T_pM. This way, there is a relation between tangent vectors in different tangent spaces, and so the covariance becomes meaningful.

The WGP regression process is summarized in Alg. 4.

Algorithm 4 (WGP regression.)**.** *The following describes step-by-step how to carry out WGP regression.*

Input *Manifold-valued training data* $\boldsymbol{D}_M = \{(x_i, p_i)\}_{i=1}^n$.

Output *Predictive distribution for* $\boldsymbol{p}_*|\boldsymbol{p}$ *at* $\boldsymbol{x}_*$.

i. Choose a prior BPF m.

ii. Transform $\boldsymbol{D}_{T_{\boldsymbol{m}}M} \leftarrow \{(x_i, \mathrm{Log}_{m(x_i)}(p_i))\}_{i=1}^N$.

iii. Choose a prior TSCF k *from a parametric family by optimizing the hyperparameters.*

iv. Using GP prior $\mathcal{GP}(0, k)$*, carry out Euclidean GP regression for the transformed data* $\boldsymbol{D}_{T_{\boldsymbol{m}}M}$*, yielding the mean and covariance* $(\boldsymbol{\mu}_*, \Sigma_*)$.

vi. End with the predictive distribution
$\boldsymbol{p}_*|\boldsymbol{p} \sim (\mathrm{Exp}_{\boldsymbol{m}_*})_\#(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*))$

4.2.2 Observations with noise

A difficulty arises, when introducing a noise model on our observations. In the Euclidean case, a popular noise model on the observations (x_i, p_i) is given by $p_i = f(x_i) + \epsilon$, where f is the function we approximate and $\epsilon \sim \mathcal{N}(0, K_{\mathrm{err}})$ is the noise term. In [5], this model is generalized to the manifold setting implicitely as

$$p_i = \mathrm{Exp}_{f(x_i)}(\epsilon), \tag{30}$$

which is also supported by the central limit theorem provided in [15]. However, this makes the WGP analytically intractable. To allow computations, we propose the error model $\mathrm{Log}_{m(x_i)}(p_i) = \mathrm{Log}_{m(x_i)}(f(x_i)) + \epsilon$, that is, the error lives in the tangent space of the prior mean at x_i. This can be viewed as a first order approximation of (30) around $m(x_i)$. Introduction of this error changes the regression procedure only slighty; the joint distribution of $\boldsymbol{p}$ and $\boldsymbol{p}_*$ changes into

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} + K_{\mathrm{err}} \end{pmatrix} \right). \tag{31}$$

Rest of the computations are then carried out similarly, with the replacement of $\boldsymbol{k}$ with $\boldsymbol{k} + K_{\mathrm{err}}$ everywhere.

5 Experiments

We demonstrate WGP regression on three manifolds. First, we visualize our algorithm on the 2-sphere using both an illustrative toy dataset and fitting a

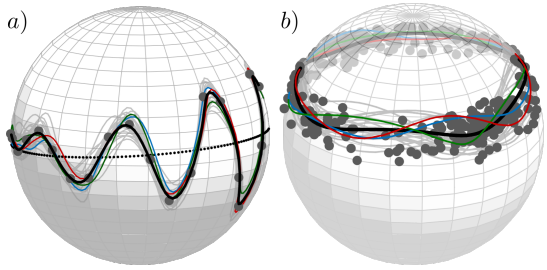


Figure 3: Depicted in a) is WGP regression using a prior BPF given by geodesic regression (dotted black) on a toy data set (grey dots) on S^2. The predictive distribution is visualized using the MAP estimate (black line, see Remark 3) and 20 samples from the distribution (in gray) with three samples emphasized (in red, green and blue). In b), a motion capture dataset of the orientation of the left *femur* of a walking person. The independent variables were estimated by *principal curve* analysis, and a WGP was fitted.

WGP to motion capture data of the left *femur* of a person walking in a circular pattern. Next, we illustrate DTI upsampling with uncertainty estimates as a tensor field prediction task on a single DTI slice living on the manifold of symmetric and positive definite matrices, and finally we study the effect of age on the shape of Corpus Callosum in Kendall's shape space.

5.1 Data on 2-sphere

As a sanity check, we first visualize our method on a toy dataset on the 2-sphere seen as a Riemannian manifold with the Riemannian metric induced by the Euclidean metric on $\mathbb{R}^3$. This manifold has a finite injectivity radius, thus the approximation presented in Remark 2 is used. A regressed geodesic γ is used as the prior BPF (Sec.3.2), and a diagonal RBF kernel (as in Eq. (27)) with optimized hyperparameters is chosen as the prior TSCF. See Fig. 3 a).

Next, we consider motion capture data of the orientation of the left *femur* of a person walking in a circular pattern [12, 11, 7]. This data naturally lives on S^2 and is periodic. We estimate the periodic independent variables of the data by computing its principal curve as described in [7]. Then, we fit a WGP using Fréchet mean BPF and the TSCF is chosen to be diagonal with the periodic kernel k given by

$$k(t,t') = \sigma^2 \exp\left(-\frac{2\sin^2(|t-t'|/2)}{l^2}\right), \tag{32}$$

where the hyperparameters σ^2 and l^2 are optimized as described in Sect. 4.2.1. Note that the Fréchet mean BPF was used, as the data is not geodesic in trend. The resulting WGP is depicted in Fig. 3 b).

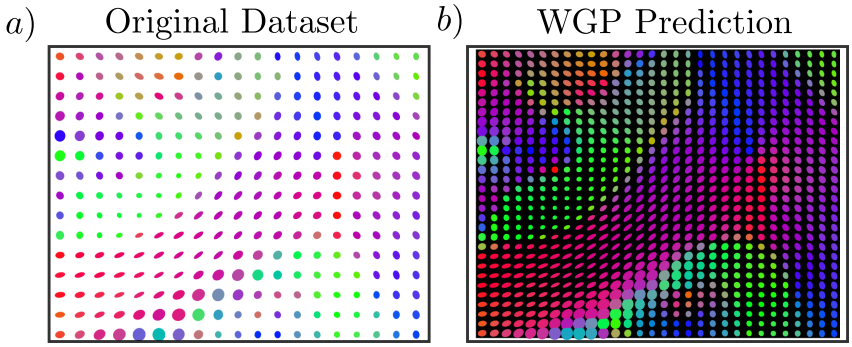


Figure 4: Upsampling DTI tensor field by WGP regression. Colors depict the direction of the principal eigenvector of the respective tensor. a)Tthe slice shown as a tensor field, b) MAP estimate of the predictive distribution of WGP regression on the original data set with uncertainty visualized below (white indicates maximum relative error, black indicates no error). The relative error is computed by dividing by the maximal error over the experiment here and in Fig. 5 c) and e).

5.2 Diffusion tensor imaging data

We consider a patch of estimated voxel-wise DTI tensors from a coronal slice of an HCP subject [6, 26, 28]. The tensors reside on the manifold $\mathbb{R}^2 \times \mathrm{PD}(3)$, where $\mathrm{PD}(n)$ is the set of $n \times n$ positive definite matrices. When endowed with the affine-invariant metric [22], $\mathrm{PD}(n)$ forms a Riemannian manifold of non-positive curvature, meaning we can perform exact WGP regression with values in $\mathrm{PD}(n)$. The data set consists of 15×19 tensors (elements of $\mathrm{PD}(3)$) with isotropic spacing, see Fig. 4 a). DTI upsampling is performed as an interpolation task on a 30×30 grid, fitting a WGP to the data and estimating up-sampled values using the estimated WGP. As a measure of uncertainty of the result, we calculate the sum of variances of each tangent space coordinate at the interpolated points; this is visualized as a background intensity in Fig. 4 b).

To illustrate the flexibility of WGP regression, we perform a second upsampling experiment, where we randomly subsample only a fifth of the original DTI tensors, see Fig. 5 a). In Fig. 5 c) is shown the corresponding MAP estimate of the predictive distribution (see Remark 3), where empirical Fréchet mean was used as the prior BPF (Fig. 5 b)) and diagonal RBFs with optimized hyperparameters as the prior TSCFs. Finally, to illustrate the effect of the choice of prior BPF, a final experiment used the result of geodesic submanifold regression as the prior BPF, see Fig. 5 d), e).

Note that the tensor field can be reconstructed well even from just 20% of the data, although with increased uncertainty, as can be seen when comparing Figs. 5 c), e) to Fig. 4 a). The predictive WGPs in Figs. 5 c) and e) do not differ vastly, although different BPFs were used. They yield a different result in the upper-left corner area, where the subsampled dataset is not dense, hence the

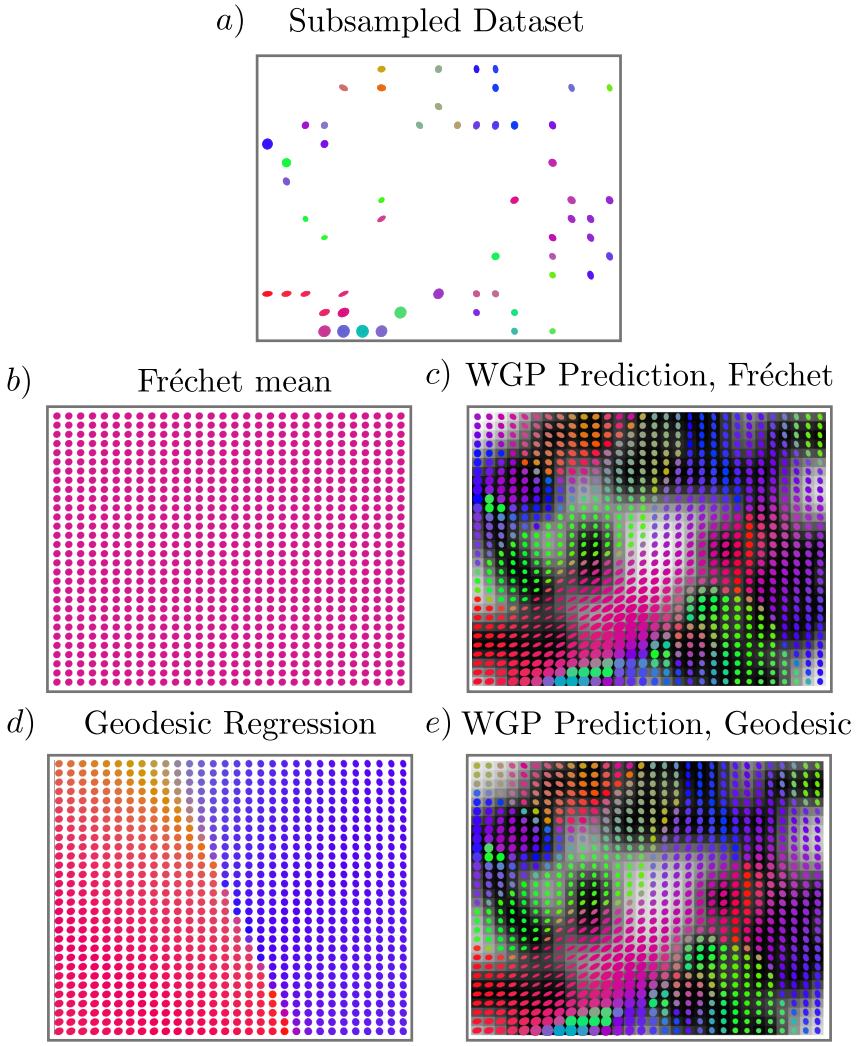


Figure 5: Upsampling DTI tensor field by WGP regression. This time, we carry out the regression on a subsampled tensor field (shown in a)), where only 20% of the elements of the original tensor field (see Fig. 4 a)) are present. We carry out the regression using two different prior WGP BPFs. In b), the first prior BPF using the Fréchet mean is shown and the corresponding predictive WGP is visualized in c), using the MAP estimate to plot the tensors. The second prior BPF is given by geodesic regression, shown in d), with the corresponding predictive WGP in e). For color descriptions, refer to the caption of Fig. 4. The uncertainty fields in c) and e) have similar shapes, but the magnitudes differ.

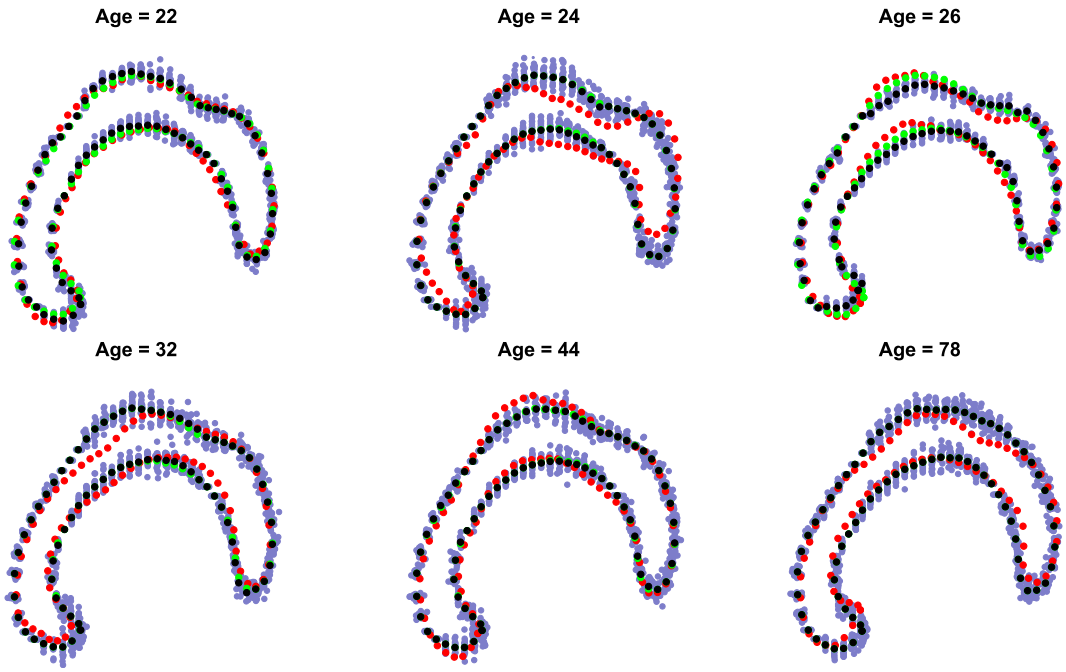


Figure 6: WGP regression applied to a population of Corpus Callosum shapes labeled by age. Red shapes are data points from the test set, not used for training. In black, the MAP estimates of the predictive distributions, in green values of the prior BPF at corresponding ages. Drawn in blue are 20 samples from the predictive distribution.

regressed result is influenced by the prior BPF. In the middle, where we also lack information, the resulting tensor fields look similar. The error structures are very similar, seen in Figs. 5c), e). This can be explained by the optimized prior hyperparameters of the TSCFs being similar in both cases (the residuals do not affect the posterior covariance other than through hyperparameter optimization).

5.3 Corpus Callosum data

Next, we turn to a dataset of landmark representations of Corpus Callosum (CC) shapes [5]. A landmark representation is a set of k points in $\mathbb{R}^2$, so that length, translation and rotation factors have been quotiented out, resulting in a point in the *Kendall's shape space* [14]. The dataset consists of 65 shapes, of which we pick randomly 6 to be the test set, the rest are used for training.

Results are presented in Fig. 6. A tangent space geodesic regression is used as the prior BPF, and a diagonal RBF kernel with optimized hyperparameters is used as the prior TSCF. As the CC shapes vary considerably even in the same age group, the WGP predictive mean does not yield notable gains on the tangent space geodesic regression used as prior BPF. However, it provides uncertainty estimates of the shape. Notably, the results imply that aging brings about wider variation in the upper-right part of the CC.

6 Conclusion and discussion

This paper introduced WGP regression on Riemannian manifolds in a novel Bayesian inference framework relying on WGPs, defined via WGDs. Then, the conditional distribution of two jointly WGD random points was computed for WGP regression. We demonstrated the method on three manifolds; on the 2-sphere using a toy data set and motion capture data of the femur of a walking person, tensor data originating from DTI and on a set of Corpus Callosum shapes. The results of the experiments imply that WGP regression can be used effectively on Riemannian manifolds, providing meaningful uncertainty estimates.

This being the first step, there are still open questions; how to engineer prior distributions efficiently, and how to treat the predictive distribution? The predictive distribution admits an explicit expression, but the prediction is not a WGP anymore. Therefore, we do not have same closure properties of the family of distributions as in the Euclidean case. This leaves open the question, whether one should consider other generalizations of GDs than the wrapped one when carrying out GP regression on manifolds?

We suggested an approximation in Remark 2, not quantifying how reliable it is in the case of non-infinite injectivity radius. In practice the approximation seems plausible (see Fig. 3), but should be studied in more detail. Furthermore, it is of interest, in which cases the computations can be carried out analytically, when the injectivity radius is non-infinite.

The central limit theorem presented in [15] suggests to use WGD distributed error terms, but this poses the difficulty of incorporating the noise term into the prior, when the noise term might live in a different tangent space. The workaround used in this paper was to approximate this error term linearly in the tangent space of the prior BPF, however, other models should also be considered.

Finally, GP regression could be generalized to a broader family of spaces than Riemannian manifolds. In WGP regression, the key is having a wrapping function from a model vector space onto the manifold. For example, another context where such structure appears, is the weak Riemannian structure of the space of probability measures under the Wasserstein metric [1].

7 Acknowledgements

This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at

Washington University. The authors would also like to thank Tom Dela Haije and Søren Hauberg for fruitful discussions and feedback.

References

[1] Ambrosio, L., Gigli, N.: A user's guide to optimal transport. In: Modelling and optimisation of flows on networks, pp. 1–155. Springer (2013)

[2] Banerjee, M., Chakraborty, R., Ofori, E., Vaillancourt, D., Vemuri, B.C.: Nonlinear regression on Riemannian manifolds and its applications to Neuro-image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 719–727. Springer (2015)

[3] Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S.: Population shape regression from random design data. International journal of computer vision **90**(2), 255–266 (2010)

[4] Do Carmo, M.P., Flaherty Francis, J.: Riemannian geometry, vol. 115. Birkhäuser Boston (1992)

[5] Fletcher, P.T.: Geodesic regression and the theory of least squares on Riemannian manifolds. International journal of computer vision **105**(2), 171–185 (2013)

[6] Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage **80**, 105–124 (2013)

[7] Hauberg, S.: Principal curves on riemannian manifolds. IEEE transactions on pattern analysis and machine intelligence **38**(9), 1915–1921 (2016)

[8] Hauberg, S., Lauze, F., Pedersen, K.S.: Unscented Kalman filtering on Riemannian manifolds. Journal of mathematical imaging and vision **46**(1), 103–120 (2013)

[9] Hinkle, J., Muralidharan, P., Fletcher, P.T., Joshi, S.: Polynomial regression on Riemannian manifolds. In: European Conference on Computer Vision. pp. 1–14. Springer (2012)

[10] Hong, Y., Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Regression uncertainty on the grassmannian. In: Artificial Intelligence and Statistics. pp. 785–793 (2017)

[11] Ionescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2220–2227. IEEE (2011)

[12] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural envi-

ronments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2014)

[13] Jona-Lasinio, G., Gelfand, A., Jona-Lasinio, M., et al.: Spatial analysis of wave direction data using wrapped gaussian processes. The Annals of Applied Statistics **6**(4), 1478–1498 (2012)

[14] Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London Mathematical Society **16**(2), 81–121 (1984)

[15] Kendall, W.S., Le, H., et al.: Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. Brazilian Journal of Probability and Statistics **25**(3), 323–352 (2011)

[16] Kim, H.J., Adluru, N., Collins, M.D., Chung, M.K., Bendlin, B.B., Johnson, S.C., Davidson, R.J., Singh, V.: Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2705–2712 (2014)

[17] Kühnel, L., Sommer, S.: Stochastic development regression on non-linear manifolds. In: International Conference on Information Processing in Medical Imaging. pp. 53–64. Springer (2017)

[18] Mardia, K.V., Jupp, P.E.: Directional statistics, vol. 494. John Wiley & Sons (2009)

[19] Oller, J., Corcuera, J.M., et al.: Intrinsic analysis of statistical estimation. The Annals of Statistics **23**(5), 1562–1581 (1995)

[20] Pennec, X.: Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. In: NSIP. pp. 194–198 (1999)

[21] Pennec, X.: Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision **25**(1), 127–154 (2006)

[22] Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. International Journal of Computer Vision **66**(1), 41–66 (2006)

[23] Pigoli, D., Menafoglio, A., Secchi, P.: Kriging prediction for manifold-valued random fields. Journal of Multivariate Analysis **145**, 117–131 (2016)

[24] Rasmussen, C.E., Williams, C.K.: Gaussian processes for machine learning, vol. 1. MIT press Cambridge (2006)

[25] Sommer, S.: Anisotropic distributions on manifolds: template estimation and most probable paths. In: International Conference on Information Processing in Medical Imaging. pp. 193–204. Springer (2015)

[26] Sotiropoulos, S., Moeller, S., Jbabdi, S., Xu, J., Andersson, J., Auerbach, E., Yacoub, E., Feinberg, D., Setsompop, K., Wald, L., et al.: Effects of image reconstruction on fiber orientation mapping from multichannel

diffusion MRI: reducing the noise floor using SENSE. Magnetic resonance in medicine **70**(6), 1682–1689 (2013)

[27] Steinke, F., Hein, M.: Non-parametric regression between manifolds. In: Advances in Neural Information Processing Systems. pp. 1561–1568 (2009)

[28] Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The WU-Minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)

Appendix C

Probabilistic Riemannian Submanifold Learning with Wrapped Gaussian Process Latent Variable Models

The following chapter presents (up to formatting) the article

> Anton Mallasto, Søren Hauberg, and Aasa Feragen. "Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models." Proceedings of Machine Learning Research (AISTATS). 2019.

This work applies the WGP framework in Appendix B to the unsupervised learning setting. This is done through generalizing *Gaussian process latent variable models* (GPLVMs) [38] to Riemannian manifolds with the use of WGPs. Doing so, allows learning stochastic submanifolds of data lying on the given Riemannian manifold. The method is demonstrated on directional data lying on the 2-sphere, landmark representations of *diatom* organisms living in the Kendall shape space, on DTI diffusion tensors living in the space of 3-by-3 symmetric positive matrices, and on 20-by-20 covariance matrices between crypto-currency prices.

This work falls into the category of *submanifold learning*, where one takes into account differential geometric constraints on given data, and then learns a submanifold of data living in the constrained *ambient* Riemannian manifold. Work in this direction includes *principal geodesic analysis* [24,60], its probabilistic counterpart [64], and non-geodesic approaches such as *barycentric subspace analysis* [52]. Our work extends the literature, by allowing for non-geodesic, non-parametric learning of the submanifolds, and provides probabilistic results, whose uncertainty can be quantified.

Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models

Anton Mallasto[1] Søren Hauberg[2] Aasa Feragen[1]

[1]Department of Computer Science,University of Copenhagen
[2]DTU Compute, Technical University of Denmark

Abstract

Latent variable models (LVMs) learn probabilistic models of data manifolds lying in an *ambient* Euclidean space. In a number of applications, a priori known spatial constraints can shrink the ambient space into a considerably smaller manifold. Additionally, in these applications the Euclidean geometry might induce a suboptimal similarity measure, which could be improved by choosing a different metric. Euclidean models ignore such information and assign probability mass to data points that can never appear as data, and vastly different likelihoods to points that are similar under the desired metric. We propose the wrapped Gaussian process latent variable model (WGPLVM), that extends Gaussian process latent variable models to take values strictly on a given ambient Riemannian manifold, making the model blind to impossible data points. This allows non-linear, probabilistic inference of low-dimensional Riemannian submanifolds from data. Our evaluation on diverse datasets show that we improve performance on several tasks, including encoding, visualization and uncertainty quantification.

1 Introduction

Unsupervised learning aims at modelling structure in unlabeled data, such as its geometry. Sometimes, information on this geometry is available through spatial constraints or a non-Euclidean metric, e.g. the data lives on a Riemannian manifold. Incorporating the known Riemannian manifold in a probabilistic model should improve model fit, and save us from learning what we already know. In this work, we study a probabilistic latent variable model that takes the geometry into account.

Where do manifolds come from? Data points on a sphere are forced to have norm one, covariance matrices are symmetric and positive definite, and shapes do not depend on scale, rotation or placement. Enforcing such constraints or

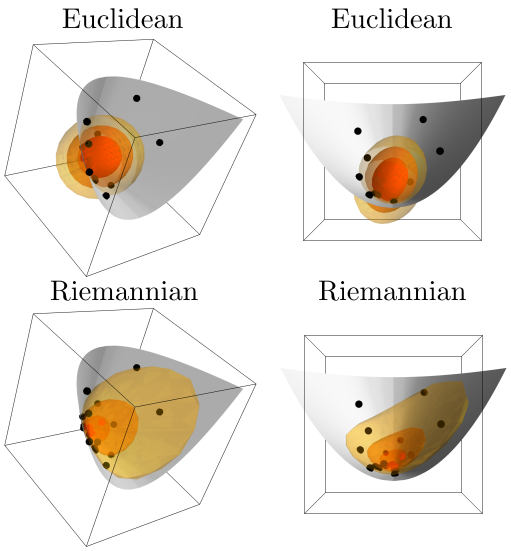


Figure 1: The ambient manifold $SPD(2)$ is the open subset *on the inside* of the visualized grey cone in the ambient Euclidean space $\mathbb{R}^3$. **Top row:** A Euclidean Gaussian distribution fitted to a set of $SPD(2)$ matrices (black dots) escapes *outside* of $SPD(2)$. **Bottom row:** The Riemannian Log-Euclidean metric yields a wrapped Gaussian distribution that remains inside $SPD(2)$, providing a better fit to the data. The colored trust regions are confidence regions of the (W)GDs.

invariances, one replaces the ambient Euclidean space by an ambient *manifold*. The *ambient space* refers to the set of all those points, which the model views as possible data points. The constraints alter the shortest paths between data objects, giving rise to a *Riemannian metric*. Riemannian metrics can also be imposed by modelling choices; closeness under the Euclidean metric does not always express desired similarity of data objects. These metrics can be learned from data (Hauberg et al., 2012) or imposed based on domain knowledge (Arsigny et al., 2006).

Euclidean probabilistic models on manifold data assign probability mass to impossible data points under spatial constraints. Furthermore, points that are similar under the chosen non-Euclidean metric can be assigned very different likelihoods, which can cause a poor fit to the data. Both issues affect especially the uncertainty estimates. These issues can be avoided by exploiting the Riemannian geometry in the model. Fig. 1 shows points in $SPD(2)$, the space of 2×2 symmetric positive-definite matrices, with fitted Euclidean and Riemannian models. The points outside the cone are not $SPD(2)$ matrices. Under the Log-Euclidean metric, which generalizes the log transform to matrices, elements on the boundary (in gray) lie infinitely far from interior points. The metrics, and hence the induced models, are vastly different. This results in the Riemannian model with an improved model fit.

Contributions. Motivated by these observations, we introduce the *wrapped Gaussian process latent variable model* (WGPLVM). This extends the Gaussian process latent variable model (GPLVM) to data on Riemannian manifolds

by employing *wrapped Gaussian processes* (WGPs). Like the GPLVM, the WGPLVM defines a probabilistic model between elements in a lower dimensional *latent space* and the data, providing uncertainty estimates. As WGPs take values strictly on a given Riemannian manifold, the WGPLVM enforces known constraints and invariances, and accounts for modelling choices concerning the metric.

We demonstrate the WGPLVM on several different manifolds and tasks. We show that our method provides more efficient encoding of the original data compared to the Euclidean GPLVM, provides superior uncertainty estimates and better captures trends in the data, resulting in improved visualization results.

Related Literature. First, we discuss methods in *manifold learning*, which view data points as elements of a Euclidean space. Then, we discuss related work in *submanifold learning*, that works strictly on Riemannian manifolds. Note that some manifold learning methods can impose known geometry on the latent space. Models relying on kernels (e.g. the GPLVM and WGPLVM) can encode such structure on the latent space (Lin et al., 2017). This is different from imposing geometric constraints on the data space.

Manifold learning infers a low-dimensional manifold that captures the trend of given data. Classical algorithms (Belkin and Niyogi, 2003; Roweis and Saul, 2000; Tenenbaum et al., 2000) learn a low distortion projection from a data submanifold of the original, Euclidean ambient space, onto a low-dimensional Euclidean space. Latent variable models (LVMs) (Goodfellow et al., 2014; Kingma and Welling, 2014; Lawrence, 2005) learn the reverse *latent embedding* from the latent space into the ambient space, associating each point in the latent space with an ambient space point. In the well-known *Gaussian process latent variable model* (GPLVM) (Lawrence, 2005), the latent embedding is a Gaussian process (GP) over the latent space, and hence learns not only a manifold embedding into $\mathbb{R}^n$, but also a model of its uncertainty. GPLVMs have inspired other LVMs (Lawrence and Moore, 2007; Titsias and Lawrence, 2010; Urtasun and Darrell, 2007), that all rely on Euclidean geometry. Urtasun et al. (2008) consider topologically constrained LVMs and Varol et al. (2012) consider GPLVMs with spatial constraints, where the constraints are enforced through slack variables and local linearization. Our method works intrinsically on the specific Riemannian manifold, taking the topology, spatial constraints and the Riemannian metric into account. Thus the WGPLVM falls into the category of submanifold learning.

Submanifold learning algorithms, illustrated in Fig. 2, aim to infer a model φ from a latent space L to a submanifold M (dashed red) of a known *ambient manifold* $\mathcal{M}$ of points that satisfy the constraints. The map φ associates the data $p_i \in \mathcal{M}$ (dark grey) with latent variables $x_i \in L$ (blue). *Principal geodesic analysis* (PGA) (Fletcher et al., 2004; Huckemann et al., 2010) estimates geodesic submanifolds, *Riemannian principal curves* (Hauberg, 2016) and *barycentric subspaces* (Pennec, 2015) estimate less constrained submanifolds. *Probabilistic PGA* (Zhang and Fletcher, 2013) introduces uncertainty by estimating probabilis-

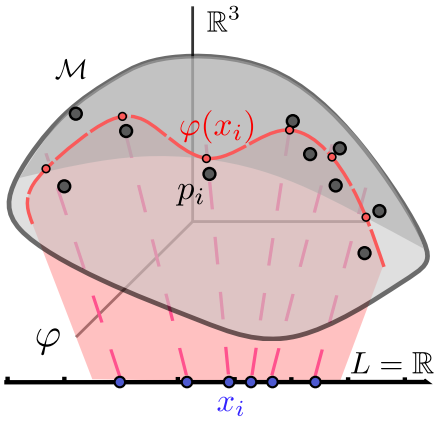


Figure 2: Illustration of submanifold learning.

tic geodesic subspaces. The WGPLVM contributes non-geodesic, probabilistic learning of the submanifold from a prior model, allowing considerable flexibility compared to previous models.

Examples of manifold valued data include directional statistics, which consider spherical data (Mardia and Jupp, 2009; Urtasun et al., 2006), covariance matrices as data objects in economics and computer vision (Tuzel et al., 2006; Wilson and Ghahramani, 2011) and in diffusion MRI or materials science (Batchelor et al., 2005; Fletcher and Joshi, 2004), and statistics of shape, which is of fundamental interest in computer vision (Freifeld and Black, 2012; Kendall, 1984). In each example, the common approach is to incorporate the Riemannian structure in the statistical analysis.

2 Preliminaries

This section introduces the necessary preliminaries and notation. We first review Gaussian processes (GPs) and the Gaussian process latent variable model (GPLVM) (Lawrence, 2004). Next, we summarize the necessary concepts from Riemannian geometry. Subsequently, we review the wrapped Gaussian processes (WGPs) introduced by Mallasto and Feragen (2018), which form the cornerstone of the present work.

Gaussian processes. Let $\mathcal{N}(\mu, \Sigma)$ denote a multivariate Gaussian distribution (GD) with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, and write the associated probability density function as $\mathcal{N}(v|\mu, \Sigma)$ for $v \in \mathbb{R}^d$. A *Gaussian process* (GP) is a collection f of random variables, so that any finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly Gaussian, where $\omega_i \in \Omega$ are elements of the *index set*. Any GP f is uniquely characterized by

$$\begin{aligned} m(\omega) &= \mathbb{E}\left[f(\omega)\right], \\ k(\omega, \omega') &= \mathbb{E}\left[(f(\omega) - m(\omega))(f(\omega') - m(\omega'))^T\right], \end{aligned} \tag{1}$$

called the *mean function* m and *covariance function* k, denoted $f \sim \mathcal{GP}(m, k)$. For more about GPs and their applications, see Rasmussen (2004).

Gaussian process latent variable model. The Gaussian process latent variable model (GPLVM) is a GP-based dimensionality reduction technique, which aims to learn a probabilistic model relating elements in the low dimensional *latent space* $L \subseteq \mathbb{R}^{n'}$ to observed data $Y = \{y_i\}_{i=1}^N \subset \mathbb{R}^n$, with $n' < n$. The model approximates the manifold that Y lives on. The probabilistic model is computed by choosing a prior GP $f \sim \mathcal{GP}(m, k_\theta)$ with hyper-parameters $\theta \in \Theta$. The hyper-parameters are optimized with the *latent variables* $X = \{x_i\}_{i=1}^N \in L$ to maximize the log-likelihood

$$\begin{aligned} \log(\mathbb{P}(Y|X,\theta)) = &- \frac{nN}{2}\ln(2\pi) - \frac{n}{2}\ln|K_{X,\theta}| \\ &- \frac{1}{2}\mathrm{Tr}\left(K_{X,\theta}^{-1} Y Y^T\right), \end{aligned} \tag{2}$$

where $(K_{X,\theta})_{ij} = k_\theta(x_i, x_j)$, and X, Y denote the corresponding data matrices. Finally, we condition the optimal prior f on the chosen latent variables X and data Y, to yield the predictive distribution of the model. Note that any prediction $f(x)$ has support in the whole $\mathbb{R}^n$, thus ignoring any constraints or invariances.

In differential geometric terms, a GPLVM can be viewed to learn a stochastic *chart* for the approximate manifold on which the dataset Y lives.

Riemannian geometry. A *Riemannian manifold* is a *smooth manifold* M with a *Riemannian metric*, i.e. a smoothly varying inner product $g_p(\cdot,\cdot)$ on the tangent space T_pM at each $p \in M$, which induces a distance function d_M on M. Each (p, v) in the *tangent bundle* $TM = \bigcup_{p\in M}(\{p\} \times T_pM)$ defines a *geodesic* γ (locally shortest path) on M, so that $\gamma(0) = p$ and $\dot\gamma(0) = v$.

The Riemannian *exponential map* $\mathrm{Exp}\colon TM \to M$ is given by $(p, v) \mapsto \mathrm{Exp}_p(v) = \gamma(1)$, where γ is the geodesic corresponding to (p, v). The exponential Exp_p at p is a diffeomorphism between a neighborhood $0 \in U_p \subset T_pM$ and a neighborhood $p \in V_p \subset M$, which is chosen in a maximal way to preserve injectivity. The *logarithmic map* $\mathrm{Log}_p\colon V_p \to T_pM$ is characterized by the identity $\mathrm{Exp}_p(\mathrm{Log}_p(p')) = p'$. Outside of V_p, we use $\mathrm{Log}_p(p')$ to denote $v \in \mathrm{Exp}_p^{-1}(p')$ with a minimal norm, chosen in a *measurable* way. The complement of V_p in M is called the *cut-locus* at p, where unique geodesics cannot be defined. Multiple useful manifolds have empty cut-locus, so that $V_p = M$, including manifolds with non-positive curvature as well as the space of positive-definite symmetric matrices used below.

Let $\mathrm{Exp}_p(v) = q$ and $\gamma(t) = \mathrm{Exp}_p(tv)$. The differential $D_p\mathrm{Log}_p(q)$ (in some coordinate chart) is given by (see supplementary material for (Pennec, 2016))

$$D_p\mathrm{Log}_p(q) = (J_0(1))^{-1} J_1(1), \tag{3}$$

where J_i are Jacobi fields solving the linear ordinary differential equation

$$\ddot J_i(t) + R(t)J_i(t) = 0, \tag{4}$$

with initial conditions $J_0(0) = 0$, $\dot J_0(0) = I_n$, and $J_1(0) = I_n$, $\dot J_1(0) = 0$. Here $R(t)$ is given by $R_{ij} = \langle \mathrm{Riem}_{\gamma(t)}(\dot\gamma(t), e_i(t))\dot\gamma(t), e_j(t)\rangle_{\gamma(t)}$ and $(e_1(t), ..., e_n(t))$ is

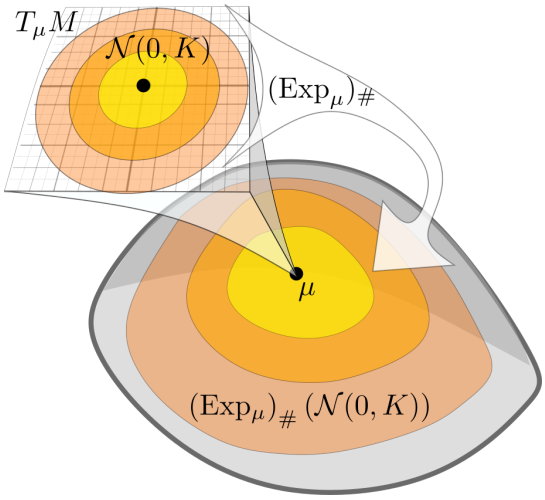


Figure 3: WGDs defined as a Gaussian $\mathcal{N}(0, K)$ in the tangent space $T_\mu M$ over the basepoint μ, which is pushed forward by the exponential map Exp_μ to M.

an orthonormal basis for $T_{\gamma(t)}M$, defined by $e_1(0) = \frac{v}{\|v\|_2}$ and each $e_j(t)$ evolves through parallel transportation. Furthermore, $\mathrm{Riem}_{\gamma t}$ denotes the curvature tensor and I_n is the n-by-n identity matrix, where n is the dimension of the manifold. For a thorough exposition in Riemannian geometry, see (Do Carmo, 1992).

Let M_i be Riemannian manifolds with metrics g_i, exponential maps Exp^i and logarithmic maps Log^i for $i = 1, 2$. Then $M = M_1 \times M_2$ turns into a Riemannian manifold when endowed with the metric $g = g_1 + g_2$, which has the component-wise computed exponential map $\mathrm{Exp}_{(p_1,p_2)}((v_1, v_2)) = \left(\mathrm{Exp}^1_{p_1}(v_1), \mathrm{Exp}^2_{p_2}(v_2)\right)$. The logarithmic map Log on the product manifold is defined likewise.

Wrapped Gaussian distributions. Let (M, g) be an n-dimensional geodesically complete Riemannian manifold. Let ν be a measure on X and $f\colon X \to Y$ be a measurable map. We define the *push-forward* as $f_\# \nu(A) := \nu(f^{-1}(A))$ for any measurable set A in Y. A random point X on M follows a *wrapped Gaussian distribution* (WGD), if for some $\mu \in M$ and a symmetric positive definite matrix $K \in \mathbb{R}^{n\times n}$

$$X \sim \left(\mathrm{Exp}_\mu\right)_\# \left(\mathcal{N}(0, K)\right), \tag{5}$$

denoted by $X \sim \mathcal{N}_M(\mu, K)$. The WGD is thus defined by a GD $\mathcal{N}(0, K)$ in the tangent space $T_\mu M$, that is pushed-forward onto M by the exponential map Exp_μ (see Fig. 3). We call $\mu =: \mu_{\mathcal{N}_M}(X)$ the *basepoint* of X, and $K =: \mathrm{Cov}_{\mathcal{N}_M}(X)$ the *tangent space covariance*.

Two random points $X_i \sim \mathcal{N}_{M_i}(\mu_i, K_i)$, $i = 1, 2$ are *jointly WGD*, if (X_1, X_2) is a WGD on the product manifold $M_1 \times M_2$, given by

$$(X_1, X_2) \sim \mathcal{N}_{M_1\times M_2}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}K_1 & K_{12}\\ K_{21} & K_2\end{pmatrix}\right), \tag{6}$$

for some matrix $K_{12} = K_{21}^T$. Then, X_1 can be conditioned on X_2, resulting in a

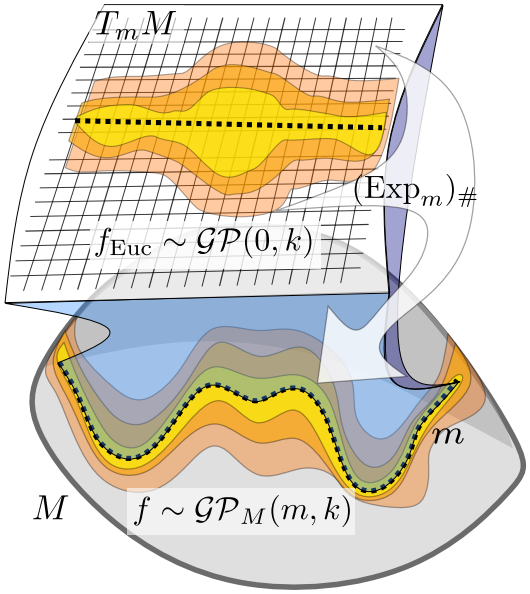


Figure 4: A WGP f can be viewed as defining a GP f_{Euc} in the tangent spaces $T_m M \subset M$ over the basepoint function, so that each marginal $f(x_i)$ is pushed-forward onto M by $(\mathrm{Exp}_{m(x_i)})_\#(f(x_i))$.

push-forward of a Gaussian mixture in $T_{\mu_1} M_1$ by the exponential map

$$X_1 | (X_2 = p_2) \sim \left(\mathrm{Exp}_{\mu_1}\right)_\# \left(\sum_{v \in A} \lambda_v \mathcal{N}\left(\mu_v, K_v\right) \right), \tag{7}$$

where $A = \{v \in T_{\mu_2} M \mid \mathrm{Exp}_{\mu_2}(v) = p_2\}$ is the preimage of p_2. The means and covariance matrices of the Gaussian mixture components are given by

$$\mu_v = K_{12} K_2^{-1} v, \qquad K_v = K_1 - K_{12} K_2^{-1} K_{12}^T, \tag{8}$$

and the component weights are

$$\lambda_v = \frac{\mathcal{N}(v | \mathbf{0}, K_2)}{\mathbb{P}\{A\}}, \quad \mathbb{P}\{A\} = \sum_{v \in A} \mathcal{N}(v | \mathbf{0}, K_2). \tag{9}$$

Wrapped Gaussian processes. Wrapped Gaussian processes generalize GPs to Riemannian manifolds (Mallasto and Feragen, 2018). A collection f of random points on a Riemannian manifold M indexed over a set Ω is a *wrapped Gaussian process* (WGP), if every finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly WGD on M^N. The functions

$$\begin{aligned} m(\omega) &= \mu_{\mathcal{N}_M}(f(\omega)), \\ k(\omega, \omega') &= \mathrm{Cov}_{\mathcal{N}_M}(f(\omega), f(\omega')), \end{aligned} \tag{10}$$

are called the *basepoint function* and the *tangent space covariance function* of f (also called as kernel of f), respectively. To denote such a WGP, we use the notation $f \sim \mathcal{GP}_M(m, k)$.

Formally, a WGP f can be viewed as a GP f_{Euc} on $T_mM \subset TM$, the family of tangent spaces over the basepoint function m. Then, the resulting GP is pushed forward to M using the Riemannian exponential map Exp_m over m to obtain the WGP, see Fig. 4.

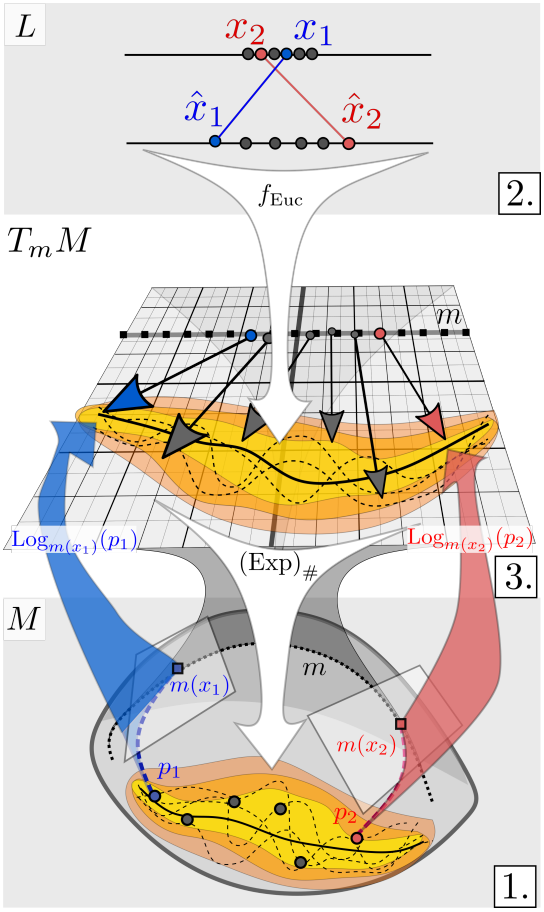


Figure 5: **The WGPLVM pipeline. 1.** The data $p_i \in \mathcal{M}$ (blue and red dots) is transformed to the tangent bundle by $p_i \mapsto \text{Log}_{m(x_i)}(p_i) \in T_{m(x_i)}\mathcal{M} \subset T_m\mathcal{M}$ along the prior basepoint function m (dotted black line) at initial latent variables x_i. **2.** A GPLVM is learned, yielding the latent variables $\hat{x}_i \in L$ and the GP f_{Euc} from L to the tangent bundle. **3.** The GP f_{Euc} is then pushed forward onto $\mathcal{M}$ by $(\text{Exp})_{\#}(f_{\text{Euc}})$, resulting in the predicted data submanifold.

3 Wrapped Gaussian Process Latent Variable Model

We now introduce the *wrapped Gaussian process latent variable model* (WGPLVM) for data $P = \{p_i\}_{i=1}^N$ lying on an n-dimensional ambient Riemannian manifold $\mathcal{M}$. The goal of WGPLVM is to learn a lower-dimensional submanifold $M_{Pred} \subset \mathcal{M}$, where the data is assumed to reside. The WGPLVM model is a straight-forward generalization of the GPLVM model, where instead of GPs, we maximize the likelihood of our data combined with the latent variables under

the WGPs that are suitable for the manifold context. The WGPLVM pipeline is illustrated in Fig. 5.

We consider a family of WGPs $f \sim \mathcal{GP}_{\mathcal{M}}(m, k_\theta)$ from the latent space L onto the ambient manifold $\mathcal{M}$, where $\theta \in \Theta$ are hyperparameters, that will be optimized over. The basepoint function m can be utilized to delocalize the learning process in order to avoid distortions of the metric caused by linearization of the curved $\mathcal{M}$. The kernel k_θ affects how observations in different tangent spaces affect each other. For coherence, the kernel should be adapted to a smooth *frame* (a smoothly changing basis over m). Such a frame can e.g. be constructed by *parallel transporting* a basis along m.

The likelihood assigned by the prior f to a data point p with associated latent variable x is

$$\begin{aligned}\mathbb{P}\{p|x,\theta\} &= \sum_{v \in \mathrm{Exp}_{m(x)}^{-1}(p)} \mathcal{N}(v|\mathbf{0}, K_{x,\theta}) \\ &\approx \mathcal{N}\left(\mathrm{Log}_{m(x)}(p)|\mathbf{0}, K_{x,\theta}\right),\end{aligned} \tag{11}$$

where $(K_{x,\theta})_{ij} = k_\theta(x^i, x^j)$ and $x = (x^1, x^2, ..., x^n)$.

The approximation in Eq. (11) only takes into account the preimage of p with a minimal norm (and thus maximal likelihood), denoted by $\mathrm{Log}_{m(x)}(p)$. The expression gives a lower bound for $\mathbb{P}\{p|x,\theta\}$, thus, maximizing the likelihood of $\mathrm{Log}_{m(x)}(p)$ maximizes the lower bound for $\mathbb{P}\{p|x,\theta\}$. It also enforces the WGPLVM to prefer *local* models over ones that wrap considerably around the manifold. Note that, for manifolds with empty cut-locus (such as ones with non-positive curvature), the approximation in (11) is exact.

The objective function to be maximized is then the approximated log-likelihood

$$\begin{aligned}\ln\left(\mathbb{P}\{p|x,\theta\}\right) \approx &-\frac{dN}{2}\ln(2\pi) - \frac{d}{2}\ln|K_{x,\theta}| \\ &-\frac{1}{2}\mathrm{Log}_{m(x)}(p)^T K_{x,\theta}^{-1}\mathrm{Log}_{m(x)}(p),\end{aligned} \tag{12}$$

for which the gradient with respect to x is given by

$$\begin{aligned}&\frac{\partial}{\partial x_j}\ln\left(\mathbb{P}\{p|x,\theta\}\right) \approx \\ &-\frac{d}{2}\mathrm{Tr}\left(K_{x,\theta}^{-1}\frac{\partial K_{x,\theta}}{\partial x_j}\right) \\ &-\frac{1}{2}\mathrm{Log}_{m(x)}(p)^T K_{x,\theta}^{-1} D_{m(x)}\mathrm{Log}_{m(x)}(p)\frac{\partial m}{\partial x_j}(x) \\ &-\frac{1}{2}\mathrm{Log}_{m(x)}(p)^T \frac{\partial K_{x,\theta}^{-1}}{\partial x_j}\mathrm{Log}_{m(x)}(p),\end{aligned} \tag{13}$$

The differential $D_{m(x)}\mathrm{Log}_{m(x)}(p)$ can be computed using Jacobi fields as explained in expression (3), if no analytical expression exists.

Assuming that the data is i.i.d, the approximate log-likelihood for the data set P can be written using Eq. (12), by considering P as a single element of the

product manifold P^N. This quantitity is then maximized by optimizing over the latent variables and the hyperparameters θ, resulting in the optimal latent variables $\hat{X}$ and hyperparameters $\hat{\theta}$ for the kernel.

The approximate submanifold can then be predicted at arbitrary latent variables X_{Pred}, by conditioning $\hat{f} \sim \mathcal{GP}_{\mathcal{M}}(m, k_{\hat{\theta}})$ on the data P with the associated latent variables $\hat{X}$ (using Eq. (7)). The conditional distribution will then be a non-centered GP $f_{\text{Euc}} \sim \mathcal{GP}(m_{\text{Euc}}, k_{\text{Euc}})$ defined on $T_m\mathcal{M}$ pushed forward by the exponential map (see Fig. 5), resulting in the predictive distribution $\varphi_{\text{pred}} \sim (\text{Exp}_{m(x)})_{\#}(f_{\text{Euc}})$. Then, the *mean prediction* is given by $\bar{\varphi}_{\text{pred}}(x) = (\text{Exp}_{m(x)})_{\#}(m_{\text{Euc}})(x))$.

In Eq. (7), if the preimage $\text{Exp}_{\mu_2}^{-1}(p_2)$ is not uniquely defined, the conditional distribution is approximated by using a preimage with minimal norm, as previously. This approximation is justified as the weights λ_v of the components of the Gaussian mixture decrease exponentially as $\|v\|_{p_2}$ increases.

The initial latent variables $X = \{x_i\}_{i=1}^N$ can be chosen strategically to aid optimization. We use *principal geodesic analysis* (PGA) (Fletcher et al., 2004) and *principal curves* (Hauberg, 2016). PGA is appropriate when the data expresses a geodesic trend (analogy of linearity on Riemannian manifolds), which is not the case for the femur dataset, see Fig. 6 in Section 4.

The Computational complexity for the method is $\mathcal{O}(NL + N^3)$, where L is the cost of computing the Riemannian logarithm. This varies from manifold to manifold, but for example, in Section 4, the most expensive is $\mathcal{O}(d^3)$ for the Log-Euclidean metric on $d \times d$ symmetric, positive-definite matrices.

We provide a pseudo-algorithm for the method in the supplementary material.

4 Experiments

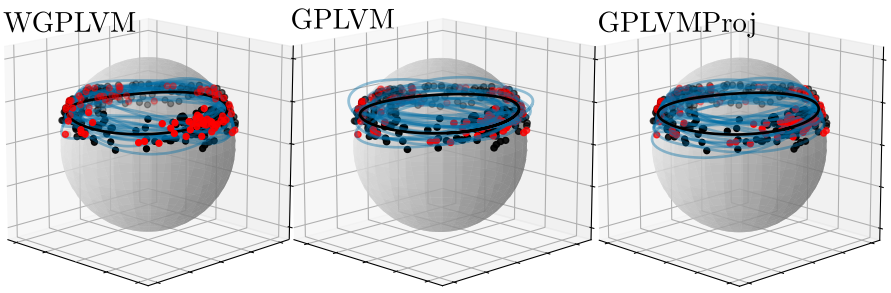


Figure 6: WGPLVM, GPLVM and GPLVMProj submanifold predictions for the *femur* data set. Mean predictions are in black, with 20 samples from the noise models (in blue). Training data in black, with test points in red.

The WGPLVM is demonstrated on three different manifolds, arising from three different applications: The sphere, Kendall's shape space (Kendall, 1984), and the space of symmetric, positive definite (SPD) matrices. Furthermore, the WGPLVM is compared with the Euclidean GPLVM, whose predictive distribution is expected not to lie on the manifold. This effect is clearly visible in Fig. 6. A third model, also shown in Fig. 6, is a modification of the Euclidean GPLVM, where the GP predictions are projected onto the manifold in order to make them satisfy the desired constraints.

We first introduce the datasets and their associated tasks, along with dataset-specific details related to training the models. In each case, we train the model assuming independent coordinates, applying the same kernel to each coordinate.

Femur dataset on S^2. A set of directions $P = \{p_i\}_{i=1}^N \in S^2$ of the left *femur* bone of a person walking in a circular pattern (CMU Graphics Lab, 2003; Hauberg, 2016) is measured at $N = 338$ time points. The movement is expected to be one dimensional and periodic, and thus we learn a 1-dimensional submanifold homeomorphic to a circle to approximate the data manifold. The latent variable optimization is initialized using principal curves (Hauberg, 2016), and the prior WGP and GP had kernel

$$k(t, t') = \sigma^2 \exp\left(-\frac{2\sin^2(|t - t'|)/2}{l^2}\right), \tag{14}$$

and mean $m(t) = \mu_{S^2}$ and $m(t) = 0$, respectively, where μ_{S^2} is the *Fréchet mean* of the training set and σ^2, l^2 are hyperparameters optimized to maximize the likelihood of the dataset P with the latent variables X. The trained models are visualized in Fig. 6.

Diatom shapes in Kendall's shape space. Diatoms are unicellular algae, whose species are related to their shapes. In Kendall's shape space M_K we analyze a set of outline shapes of 780 *diatoms* (du Buf and Bayer, 2002; Jalba et al., 2006) from 37 different species. For visualization, a two dimensional latent space is learned, using the prior $f \sim \mathcal{GP}_{M_K}(m, k)$, with constant basepoint function $m(t) = \mu_{M_K}$ set to be the Fréchet mean of the population and k given by the *radial basis function* (RBF) kernel

$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|_2^2}{2l^2}\right). \tag{15}$$

We initialize the GPLVM and WGPLVM models with PGA and PCA, respectively.

Diffusion tensors in $SPD(3)$. In the space of 3×3 SPD matrices with the Log-Euclidean metric (Arsigny et al., 2006), we collect a set of 750 diffusion tensors from a diffusion MRI dataset, sampled with approximately uniform fractional anisotropy (FA) values. The FA is a well-known tensor shape descriptor; see the supplementary material for the definition. As SPD matrices form an open subset of the Euclidean space of symmetric matrices, *we do not get a "for*

free" dimensionality reduction by restricting to SPD matrices. Instead, the data is transformed nonlinearly according to the Log-Euclidean metric, which is commonly used for diffusion tensors (Arsigny et al., 2006). The diffusion MRI image was a single subject from the Human Connectome Project (Glasser et al., 2013; Sotiropoulos et al., 2013; Van Essen et al., 2013). In diffusion MRI, low-dimensional encoding with uncertainty estimates may speed up image acquisition and processing.

Riemannian	Femur	Diatoms	Diffusion tensors	Crypto-tensors
GPLVMProj	$(9.22 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25) \times 10^{-2}$	0.582 ± 0.025	21.91 ± 2.26
WGPLVM	$\mathbf{(9.20 \pm 0.53) \times 10^{-2}}$	$\mathbf{(2.39 \pm 0.15) \times 10^{-2}}$	$\mathbf{0.391 \pm 0.035}$	$\mathbf{3.04 \pm 0.26}$

Euclidean	Femur	Diatoms	Diffusion tensors	Crypto-tensors
GPLVM	$(9.21 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25) \times 10^{-2}$	$\mathbf{(6.03 \pm 0.34) \times 10^{-2}}$	$(7.36 \pm 5.27) \times 10^{5}$
GPLVMProj	$(9.21 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25 \times 10^{-2}$	$\mathbf{(6.03 \pm 0.34) \times 10^{-2}}$	$(5.49 \pm 3.17) \times 10^{5}$
WGPLVM	$\mathbf{(9.19 \pm 0.53) \times 10^{-2}}$	$\mathbf{(2.39 \pm 0.15) \times 10^{-2}}$	$(7.54 \pm 0.36) \times 10^{-2}$	$(8.69 \pm 7.12) \times 10^{5}$

Table 1: Mean ± standard error of mean reconstruction errors, measured in RMSE, over 10 repetitions of the experiment. **Top table:** Deviations measured in the intrinsic distance on the manifold. **Bottom table:** Deviations measured in the Euclidean distance.

Crypto-tensors in $SPD(10)$. On $SPD(10)$ we collect the price of 10 popular crypto-currencies[1] in the time 2.12.2014-15.5.2018. The crypto-currency intra-relationship at a given time is encoded in the covariance matrix between the prices in the past 20 days; we include every 7th day in the period, resulting in 126 10×10 covariance matrices. Wilson and Ghahramani (2011) provide a discussion of covariance descriptors in economy. As the acquired covariance matrices in $SPD(10)$ have eigenvalues in different orders of magnitude, we use the Log-Euclidean metric (Arsigny et al., 2006), capturing this trend better.

For both $SPD(n)$ datasets, the basepoint function, the kernel and the latent variable initialization are chosen as for Kendall's shape space. The latent spaces are chosen to be 2-dimensional for visualization purposes.

Application 1: Encoding. The datasets are divided into training and test sets (consisting of $^8/_{10}$ and $^2/_{10}$ of the data, respectively), and we learn the models φ_{pred} on the training set. Each test element p is "encoded" by the projection $\pi\colon p \mapsto \operatorname{argmax}_{x \in L} \mathbb{P}\{\varphi_{\text{pred}}(x) = p\}$. We assess the quality of this encoding by measuring the root-mean-square error (RMSE) of the reconstruction, where the error is measured both by the Euclidean metric and the intrinsic metric. Each experiment was repeated 10 times with different training and test sets; the results are reported in Table 1.

Under the intrinsic metric, the WGPLVM performs significantly better on the tensor datasets, and marginally better in the two other cases. Under the Euclidean metric the WGPLVM encoding is better in two cases, worse in one, and inconclusive for the crypto-tensors where no model is significantly better than the others.

Application 2: Uncertainty quantification. Importantly, GPLVM learns a probabilistic model, producing an estimate of uncertainty. We evaluate these

[1] Bitcoin, Dash, Digibyte, Dogecoin, Litecoin, Vertcoin, Stellar, Monero, Ripple, and Verge.

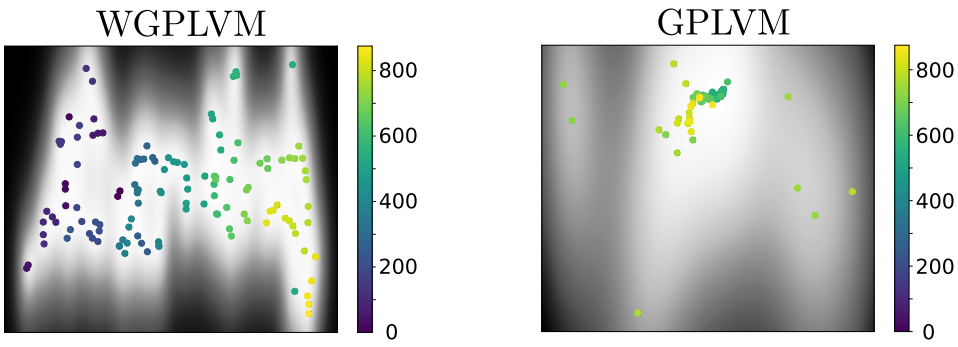


Figure 7: The latent space for the crypto-tensor dataset, with days visualized by color. Note that for GPLVM, the dark blue points corresponding to early times are hidden underneath the green points.

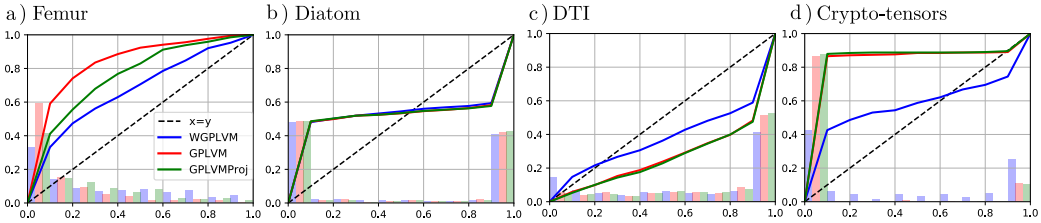


Figure 8: Uncertainty estimates given by the WGPLVM, GPLVM and projected GPLVM models for the four datasets. The bars represent the frequency of occurances, where the fraction of samples, given by the x-value, lie closer to the mean prediction than a test point. The continuous curves represent the cumulative distributions. Whenever the cumulative distribution lies above $x = y$, we are overestimating the corresponding quantile.

uncertainty estimates on all four datasets. Since the predictive distributions live in different spaces, the likelihoods of observed data under the different models are not directly comparable. However, all three models yield confidence intervals, which we compare using 10 resampled training and test sets (⁸⁄₁₀ and ⁸⁄₁₀ of the data). The test set is projected onto the predicted submanifold via π. Then, we sample the respective predictive distributions 50 times, computing the fraction of samples closer to the mean prediction than the test point. The results are visualized in Fig. 8, where the densities of these fractions are shown with corresponding cumulative distributions. For a perfect model fit, we would observe the $x = y$ curve (dashed line) as the cumulative distribution. The experiment shows that all models estimate uncertainty incorrectly, but that WGPLVM obtains the best estimate.

Application 3: Visualization. In Fig. 7, we illustrate the latent spaces of WGPLVM versus GPLVM on the crypto-tensor dataset, which comes with an associated time variable, shown in color. The WGPLVM provides a smoother and more consistent transition in color, while the GPLVM plots all the earlier (dark blue) tensors on top of each other. Similar visualizations for the other

datasets can be found in the supplementary material; in these examples, the two visualizations are not significantly different in quality.

In the supplementary material, we provide a discussion on why our model might perform better in the $SPD(n)$ experiments, including a comparison between the Euclidean and Riemannian geometries.

5 Discussion and Conclusion

We introduced the WGPLVM for non-parametric and probabilistic submanifold learning on Riemannian manifolds. The model encodes known constraints or invariances, and provides model flexibility, as metrics other than the Euclidean one can be incorporated. This is useful if a different metric captures trends in the data better. The model was evaluated on several manifolds and tasks against the GPLVM and a modified GPLVM, which projects predictions onto the manifold.

The experimental results show that the WGPLVM provides a better probabilistic model to fit the data; in particular the uncertainty estimates are superior to the Euclidean models on three out of four datasets, and virtually identical on the fourth. We note that for Euclidean models, the uncertainty is visibly higher. These are strong indications that our model carries out modelling the data distribution better. The mean predictions of the WGPLVM encode the data space significantly better than the GPLVM and projected GPLVM models on two of the datasets, and marginally better on the other two, when measured in the Riemannian metric. Under the Euclidean metricr, the GPVLM performs notably better in one experiment, and WGPLVM marginally better in two. On crypto-tensors, we deem the results inconclusive due to high variance. The aforementioned effects are also seen in the latent space visualizations, e.g. on the cryptotensors the WGPLVM better detects small-scale differences in the early time steps.

One might suspect that the improved performance stems from a "for free" dimensionality reduction through constraints. However, we note that the most significant improvement in both reconstruction error and visualization was obtained on $SPD(n)$, where the Riemannian manifold is a full-dimensional, convex subset of the Euclidean ambient space. This might still be due to the constraints, which forces the distributions to lie in the manifold. The difference could also be caused by the choice of metric. For the crypto-tensors in particular, we observe that some of the eigenvalues are very small; the Log-Euclidean metric essentially acts as a log-transform and therefore converts the data to a scale on which changes in the smaller eigenvalues can be detected.

In three of the experiments, the mean predictions of GPLVM lie essentially on the manifold, thus the projected version does not improve the mean reconstruction error. However, in the femur experiment, the uncertainty estimates are clearly improved, but also notably outperformed by WGPLVM. Due to the metric and

curvature of the manifold, interpolation between two points in the ambient space $\mathbb{R}^n$ does not necessarily project even closely onto the manifold interpolation between the projected points. This distortion affects the statistics relying on interpolation, and explains both the reduced reconstruction capability and the increased variance. Furthermore, the projected model ignores any metric choices imposed on the manifold.

Although the WGPLVM provides flexibility through the prior basepoint function, we fixed this to be the Fréchet mean of the training set in our experiments. The choice is well justified if the data is local enough, and makes the comparison to GPLVM fair. The flexibility to delocalize the learning process through the basepoint function is, however, important for inference on manifolds when the locality assumption fails. The non-trivial optimization of the basepoint function thus provides a venue for future research.

In summary, the WGPLVM is a probabilistic submanifold learning algorithm that respects known Riemannian manifold structure in the data by taking values in the associated Riemannian manifold. We compare the model to its Euclidean counterparts on a number of manifolds, datasets and tasks, and show that it has superior representation capabilities more faithful visualizations and improved uncertainty estimates.

Acknowledgements

AM and AF were supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation, and SH was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757360), as well as a research grant (15334) from VILLUM FONDEN. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The data [in part] used in this project was obtained from `mocap.cs.cmu.edu`. The database was created with funding from NSF EIA-0196217. The authors wish to thank Thomas Hamelryck for helpful comments.

References

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(2):411–421, 2006.

Phillipp G. Batchelor, Maher Moakher, David. Atkinson, Fernando. Calamante,

and Alan Connelly. A rigorous framework for diffusion tensor calculus. *Magnetic Resonance in Medicine*, 53(1):221–225, 2005. ISSN 1522-2594.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

CMU Graphics Lab. CMU Graphics Lab Motion Capture Database . `http://mocap.cs.cmu.edu/`, 2003. The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

Manfredo Perdigao Do Carmo. *Riemannian geometry*. Birkhauser, 1992.

Hans du Buf and Micha Bayer. *Automatic Diatom Identification*. 2002.

P. Thomas Fletcher and Sarang Joshi. *Principal Geodesic Analysis on Symmetric Spaces: Statistics of Diffusion Tensors*, pages 87–98. 2004.

P. Thomas Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

Oren Freifeld and Michael J Black. Lie bodies: A manifold representation of 3D human shape. In A. Fitzgibbon et al. (Eds.), editor, *European Conference on Computer Vision (ECCV)*, Part I, LNCS 7572, pages 1–14. Springer-Verlag, 2012.

Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, et al. The minimal preprocessing pipelines for the Human Connectome project. *Neuroimage*, 80:105–124, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Søren Hauberg. Principal curves on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

Søren Hauberg, Oren Freifeld, and Michael J. Black. A geometric take on metric learning. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2033–2041. MIT Press, 2012.

Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pages 1–58, 2010.

Andrei C. Jalba, Michael HF Wilkinson, and Jos BTM Roerdink. Shape representation and recognition through morphological curvature scale spaces. *IEEE Transactions on Image Processing*, 15(2):331–341, feb 2006.

David G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*, pages 329–336, 2004.

Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005. ISSN 1532-4435.

Neil D. Lawrence and Andrew J. Moore. Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pages 481–488. ACM, 2007.

Lizhen Lin, Mu Niu, Pokman Cheung, and David Dunson. Extrinsic gaussian processes for regression and classification on manifolds. *arXiv preprint arXiv:1706.08757*, 2017.

Anton Mallasto and Aasa Feragen. Wrapped Gaussian process regression on Riemannian manifolds. In *CVPR - IEEE Conference on Computer Vision and Pattern Recognition, to appear*, 2018.

Kanti V. Mardia and Peter E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

Xavier Pennec. Barycentric subspaces and affine spans in manifolds. In *Geometric Science of Information - Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*, pages 12–21, 2015.

Xavier Pennec. Barycentric subspace analysis on manifolds. *arXiv preprint arXiv:1607.02833*, 2016.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

Stamatios N. Sotiropoulos, Steen Moeller, Saad Jbabdi, Jungqian Xu, Jesper Andersson, Edward John Auerbach, Essa Yacoub, David A. Feinberg, Kawin Setsompop, Lawrence L. Wald, et al. Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE. *Magnetic resonance in medicine*, 70(6):1682–1689, 2013.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Michalis Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

Oncel. Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. *European Conference on Computer Vision (ECCV)*, pages 589–600, 2006.

Raquel Urtasun and Trevor Darrell. Discriminative Gaussian process latent variable model for classification. In *Proceedings of the 24th international conference on Machine learning*, pages 927–934. ACM, 2007.

Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with Gaussian process dynamical models. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 238–245. IEEE, 2006.

Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor J. Darrell, and Neil D Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th international conference on Machine learning*, pages 1080–1087. ACM, 2008.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn Human Connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Aydin Varol, Mathieu Salzmann, Pascal Fua, and Raquel Urtasun. A constrained latent variable model. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2248–2255. Ieee, 2012.

Andrew Gordon Wilson and Zoubin Ghahramani. Generalised Wishart processes. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 736–744, 2011.

Miaomiao Zhang and P. Thomas Fletcher. Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1178–1186, 2013.

Supplementary Material

A Pseudo-Algorithm forWGPLVM

A pseudo-code algorithm for training the WGPLVM is provied in Alg. 1.

Algorithm 1 Training WGPLVM. Input: basepoint function m, kernel k_Θ, initial latent variables $x = \{x_i\}_{i=1}^N$, dataset $p = \{p_i\}_{i=1}^N$, learning rate λ. Each logarithmic map should be express with respect to a frame W on the manifold.

while Not converged **do**
  # Compute logarithmic maps and save into a matrix as rows
  $[\mathrm{Log}_{m(x)}(p)]_i \leftarrow \mathrm{Log}_{m(x_i)}(p_i)$
  # Compute prior covariance matrix:
  $[K_{x,\Theta}]_{ij} \leftarrow k_\Theta(x_i, x_j)$
  # Compute objective:
  $L \leftarrow --\frac{dN}{2}\ln(2\pi) - \frac{d}{2}\ln|K_{x,\Theta}| - \frac{1}{2}\mathrm{Log}_{m(x)}(p)^T K_{x,\Theta}^{-1}\mathrm{Log}_{m(x)}(p)$
  # Compute gradients and update parameters
  $x \leftarrow x + \lambda\nabla_x L$
  $\Theta \leftarrow \Theta + \lambda\nabla_\Theta L$
end while

B Details on Manifolds Used

The n-sphere S^n is a Riemannian manifold with exponential and logarithmic maps given by

$$\begin{aligned}\mathrm{Exp}_p(v) &= \cos(\|v\|_2)p + \sin(\|v\|_2)\frac{v}{\|v\|_2},\\ \mathrm{Log}_p(q) &= \arccos\left(\langle p, q\rangle\right)\frac{q - \langle p, q\rangle p}{\|q - \langle p, q\rangle p\|_2},\end{aligned} \tag{16}$$

where $\|\cdot\|_2$ is the 2-norm induced by the standard Euclidean innerproduct $\langle\cdot,\cdot\rangle$.

Kendall's shape space forms a quontient manifold of the sphere, so the operations defined for S^n apply, when working with the right quotient representatives. Kendall's shape space has the additional constraint of representing shapes with respect to an optimal translation between a pair of shapes. Let X, Y be the $2 \times N$ data matrices of two shapes, where N is the amount of landmarks, and each column represents the x, y-coordinates after quontienting away scale and translation. Then, the *Procrustean* distance between the shapes X, Y is given by

$$\min_R \|X - RY\|_2, \tag{17}$$

where R is a rotation matrix. The shapes are aligned by choosing a reference point, and aligning the population elements by minimizing the Procrustean distance.

The space $SPD(n)$ of symmetric, positive definite matrices can be given the structure of a Riemannian manifold, by endowing it with the *Log-Euclidean* metric. The tangent space at each point is the space of n-by-n symmetric matrices, and the affine-invariant metric is given by

$$g_P(V, U) = \mathrm{Trace}[V^T U], \tag{18}$$

and the exponential and logarithmic maps are given by

$$\mathrm{Exp}_P(A) = \exp(\log(p) + v), \quad \mathrm{Log}_P(Q) = \log(Q) - \log(P), \tag{19}$$

where exp stands for the matrix exponential and log for the matrix logarithm.

C Latent Space Visualization

Here we provide the latent space visualizations for the diffusion-tensor and diatom datasets.

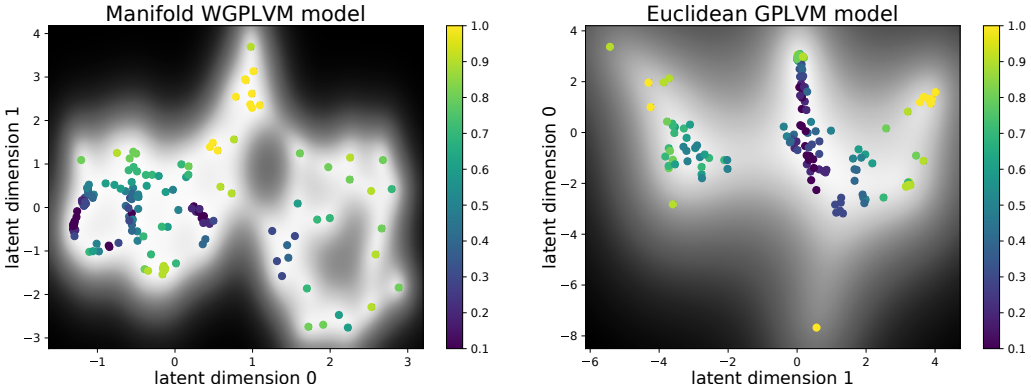


Figure 9: The latent spaces for the diffusion-tensor dataset learned using the WGPLVM and GPLVM models. The colors indicate the FA of the given tensor.

The *fractional anisotropy* (FA) of a 3×3 SPD matrix is a shape descriptor taking values between 0 and 1, where an FA of 0 corresponds to a round tensor, and an FA near 0 corresponds to a very thin one. Given the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ for an SPD matrix, its FA is defined as

$$\sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \hat{\lambda})^2 + (\lambda_2 - \hat{\lambda})^2 + (\lambda_3 - \hat{\lambda})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}},$$

where $\hat{\lambda}$ is the mean of the eigenvalues. In the latent space shown in Fig. 9, the latent variables are colored according to the FA of their associated tensor, and we see that both models provide a smooth transition between different FA values.

The latent space visualization of the diatom dataset is found in Fig. 10; here the latent variables are colored by the species of the corresponding diatom, see Fig. 11 for a visualization of species representatives.

D Comparing the Geometries

In this section, we compare the geometries in Euclidean and Riemannian cases. The aim is to try and understand, when the performance is improved. We do this

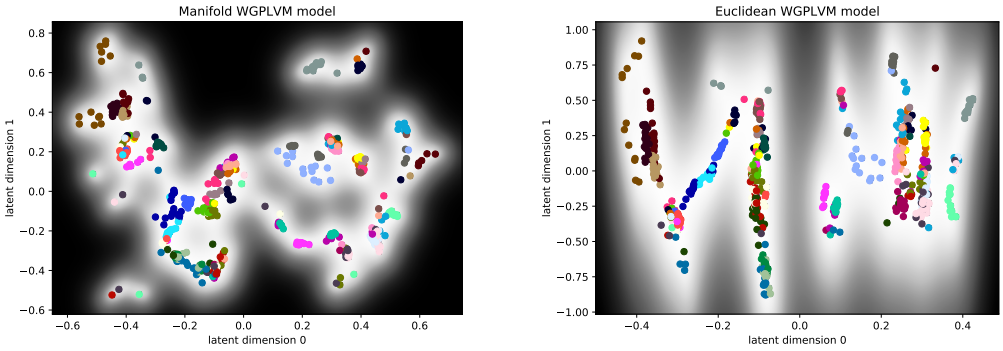


Figure 10: The latent spaces for the diatom dataset learned using the WGPLVM and GPLVM models. The colors indicate the species of the diatom corresponding to the latent variable, see Fig. 11.

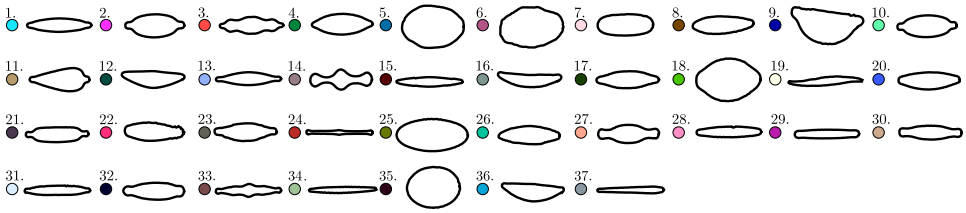


Figure 11: Representatives of each of the 37 diatom classes with corresponding class colors used in Fig. 10. Note that variation inside of each class can be considerable.

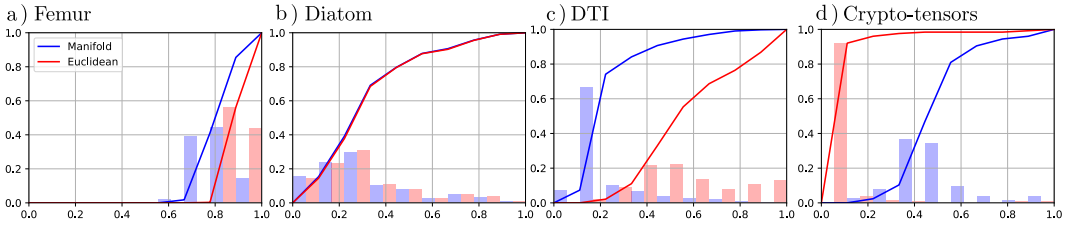


Figure 12: Distributions of distances between the data points and the population means. The bar plots indicate the density of data points that lie x-fraction of the maximum distance away from the mean. The corresponding continuous curves represent the cumulative distributions.

by visualizing the distribution of data point distances to the corresponding population means, the distances and means computed according to the corresponding metrics.

As can be seen in Fig. 12, in the femur (2-sphere) and diatom (Kendall's shape space) cases, the distributions look very similar. In fact, in the diatom case, they are essentially the same. The Kendall's shape space forms a quotient manifold of the sphere, which in this case is high dimensional ($d = 180$). In such high dimension, escaping the manifold becomes increasingly more difficult (most of the volume of the sphere is close to the boundary), and thus both the metrics are

essentially the same. This might explain, why the WGPLVM did not improve notably on the GPLVM.

In the crypto-tensor experiment, the distribution implies the presence of extreme outliers under the Euclidean metric. The Log-Euclidean metric, on the other hand, transforms the metric scale, evening out the distribution. This could very well explain, why we see large improvement with the WGPLVM compared to the GPLVM.

In the DTI experiment, the distribution of Euclidean distances looks more even. This might imply, that in this occasion, the Euclidean distance is better at capturing the trend of the data. However, the improved uncertainty estimates of the WGPLVM could be explained, as the Euclidean models are not confined to $SPD(n)$. Therefore, the distributions do not follow the conic shape of $SPD(n)$.

Appendix D

Optimal Transport Distance between Wrapped Gaussian Distributions

The following chapter presents (up to formatting) the article

> Anton Mallasto, and Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions." 38th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt). 2018.

Akin to Appendix A, this work builds on the work on 2-Wasserstein metric between Gaussian distributions, by considering an optimal transport metric between WGDs. To do this, we consider the *pullback-WGD*, that is, the Gaussian distribution on a tangent space before pushing it forward to the manifold using the exponential map. The distributions live in the tangent bundle, for which a metric has to be introduced before any OT can be carried out. A distance for tangent vectors is then introduced using parallel transport to a reference point, which allows us to compute analytically the 2-Wasserstein distance between pullback-WGDs under the parallel transport distance.

The theory of OT on Riemannian manifolds is well known [8], however, in practice, only OT between discrete measures has been considered. This work is one of the first in its kind, deriving an analytical expression for a 2-Wasserstein distance between continuous distributions on manifolds.

As a little bit of trivia, the image on the cover of this thesis visualizes a non-minimizing geodesic between two WGDs on the 2-sphere under the framework developed in this work.

Optimal Transport Distance between Wrapped Gaussian Distributions

Anton Mallasto and Aasa Feragen

Department of Computer Science, University of Copenhagen

Abstract

Optimal mass transport has recently become increasingly popular in data sciences, where it provides metric distances between probability measures. Mostly, it finds applications in the Euclidean setting, although the theoretical framework is also well studied in the case of Riemannian manifolds. In this work, we study an optimal transport distance between wrapped Gaussian distributions on a complete Riemannian manifold by pulling them back to the tangent bundle. We provide an analytical formula for a 2-Wasserstein distance between pullback wrapped Gaussian distributions, and show that it is induced by a Riemannian structure. We illustrate the framework on the 2-sphere, by plotting geodesics and Fréchet means.

1 Introduction

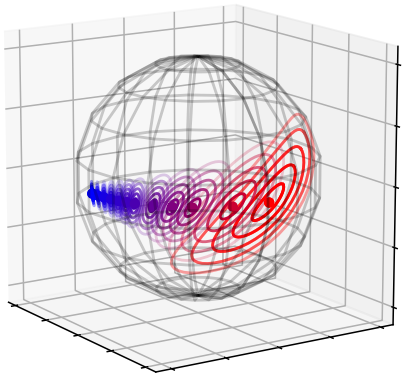


Figure 1: A geodesic interpolation between two WGDs (red and blue) on the 2-sphere.

Optimal mass transport (OMT), originating from the works of Monge [21] and Kantorovich [17], defines a geometric theory for comparing probability distributions, naturally extending the geometry of the underlying space on which the distributions are defined, to the space of the distributions. The idea is to transport *minimally* the mass of two distributions to each other (for intuition, think of transporting goods produced at a factory to outlets) when there exists a known cost of moving a unit mass between two points (such as the cost of the fuel used by a delivery truck). A common cost function is the squared L_2-distance, resulting in the *2-Wasserstein distance* between distributions. OMT naturally finds applications in economics [13], computer vision and medical imaging [16, 26], but also more recently in statistics and machine

learning [2, 3, 22] thanks to appealing theoretical and computational properties [25, 29].

Between multivariate Gaussian distributions, the 2-Wasserstein distance can be expressed analytically [11, 14, 18, 24]. For Gaussian distributions, the 2-Wasserstein distance corresponds to the *Bures metric* originating in quantum information geometry [5], and finds recent applications in machine learning [8, 19, 23]. Gaussian distributions are also theoretically interesting in optimal transport, as the Wasserstein distance between Gaussian distributions formed using the two first moments of the distributions, gives a lower bound for the Wasserstein distance between the distributions [9].

The Wasserstein distance can be defined over any *Polish space* (a separable and complete metric space). It also defines a metric between distributions on complete Riemannian manifolds, where the Wasserstein distance is induced by a *weak Riemannian structure* [1]. Even though the theory of optimal transport on Riemannian manifolds enjoys many nice theoretical properties, due to the nonlinearity of the manifold, explicit formulas and tools for computations between distributions are scarce. The only known examples to the authors consist of optimal transport between discrete measures on manifolds, such as [7].

In this paper, we aim to define a Wasserstein metric on the space of wrapped Gaussian distributions (WGDs) on Riemannian manifolds. This is formalized by pulling the WGDs back to the tangent bundle of the manifold (yielding a *pullback-WGD*), and analytically computing the 2-Wasserstein metric between the pullback distributions.

The 2-Wasserstein distance between pullback distributions requires a metric distance on the tangent bundle. While one could, in theory, use a well-known metric, such as the *Sasaki metric* or the *Cheeger-Gromoll metric* [6, 15, 27], we define a metric using parallel transportation as this allows us to derive analytical formulas for distances and geodesics with desirable properties. For instance, in the case of Sasaki metric, projection of geodesics between tangent vectors onto the manifold does not result in a geodesic between the basepoints, hence, in general, an optimal transport of WGDs would not be a WGD anymore. Another compelling idea, is to use the *complete lift* of the metric on M to TM, under which *Jacobi fields* form geodesics [30]. However, the associated metric is not positive definite, resulting in paths with negative distances.

In summary, we contribute: A Wasserstein distance between pullback-WGDs with an analytical expression, along with a proof that this metric is induced by a Riemannian metric on the space of pullback-WGDs. The resulting framework can be used to metrisize the space of WGDs over a manifold, when the pullback is uniquely defined. We illustrate the computation of geodesics and Fréchet means between WGDs on a 2-sphere for ease of visualization.

2 Preliminaries

2.1 Probabilistic Notions

Denote by $\mathcal{N}(\mu, K)$ the multivariate Gaussian distribution with mean vector $\mu \in \mathbb{R}^n$ and symmetric and positive definite covariance matrix $K \in \mathbb{R}^{n\times n}$, and write the probability density function of a random variable $X \sim \mathcal{N}(\mu, K)$ as $\mathbb{P}\{X = v\} = \mathcal{N}(v|\mu, K)$ for $v \in \mathbb{R}^n$. We write $\mathcal{N}(n)$ for the set of all n-variate Gaussian distributions. For convenience, we drop n when not explicitely required.

Given probability spaces (X, Σ_X), (Y, Σ_Y), and a measurable map $f\colon X \to Y$, the *push-forward* $f_{\#}\nu$ of a measure ν, defined on X, is defined as

$$f_{\#}\nu(A) = \nu(f^{-1}(A)), \tag{1}$$

for any measurable set A in the sigma-algebra Σ_Y.

Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear map, then the push-forward of a Gaussian distribution is given by $T_{\#}\mathcal{N}(\mu, K) = \mathcal{N}(T\mu, TKT^T)$. For convenience, we extend the push-forward to an operation on linear maps by $T_{\#}K := TKT^T$.

2.2 Optimal Transport

The *Wasserstein metric* on probability measures derives from the optimal transport problem introduced by Monge and made rigorous by Kantorovich. The p-Wasserstein distance describes the minimal cost of transporting the mass of a probability measure into the mass of another probability measure, when the cost is given by a L^p distance. For more on optimal transport, see [1, 29].

Let (X, d) be a complete and separable metric space, and denote by $\mathcal{P}_p(X)$ the set of all probability measures ν on X satisfying $\int_X d^p(x, x')d\nu(x) < \infty$ for some $x' \in X$, where $d^p(x, x')$ is used instead of $d(x, x')^p$. The p-Wasserstein distance between two probability measures $\nu, \nu' \in \mathcal{P}_p(X)$ is given by

$$W_p(\nu, \nu') = \left(\inf_{\gamma\in \mathrm{Adm}[\nu,\nu']} \int_{X\times X} d^p(x_0, x_1)d\gamma(x_0, x_1)\right)^{\frac{1}{p}}, \tag{2}$$

where $\mathrm{Adm}[\nu, \nu']$ is the set of joint probability measures on $X\times X$ with marginals ν and ν'. Defined as above, W_p satisfies the properties of a metric. Furhermore, a minimizer in (2) is always achieved.

The 2-Wasserstein metric between two Gaussians $\mathcal{N}(\mu_0, K_0)$ and $\mathcal{N}(\mu_1, K_1)$ is given by [11, 14, 18, 24]

$$W_2^2(\mathcal{N}(\mu_0, K_0), \mathcal{N}(\mu_1, K_1)) = d^2_{\mathrm{Euc}}(\mu_0, \mu_1) + \mathrm{Tr}(K_0 + K_1 - 2(K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}})^{\frac{1}{2}}), \tag{3}$$

where d_{Euc} is the canonical metric on $\mathbb{R}^n$. The metric can then be defined for any random variables $X_i \sim \mathcal{N}(\mu_i, K_i)$, $i = 1, 2$, as the distance between their distributions.

The linear map $T = K_1^{\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_0 K_1^{\frac{1}{2}} \right)^{-\frac{1}{2}} K_1^{\frac{1}{2}}$ pushes $\mathcal{N}(0, K_0)$ forward to $\mathcal{N}(0, K_1)$, and is called the *optimal map* between K_0 and K_1. The metric space $(\mathcal{N}, W_2)$ can be shown to be the induced metric space of the Riemannian manifold $(\mathcal{N}, g^{\mathcal{N}})$, where $g^{\mathcal{N}}_{\mathcal{N}(\mu,K)}(v, u) = v_0^T u_0 + v_1^T K u_1$ [28], and $v = (v_0, v_1), u = (u_0, u_1) \in T_{\mathcal{N}(\mu,K)}\mathcal{N} \simeq \mathbb{R}^n \times \mathrm{Sym}(n)$ are tangent vectors at $\mathcal{N}(\mu, K)$. Here, $\mathrm{Sym}(n)$ denotes the vector space of symmetric n-by-n matrices.

The Riemannian exponential and logarithm maps for the manifold $\mathcal{N}$ are given by

$$\begin{aligned} \mathrm{Log}^{\mathcal{N}}_{\mathcal{N}(\mu_0,K_0)}(\mathcal{N}(\mu_1, K_1)) &= (\mu_1 - \mu_0, T - I), \\ \mathrm{Exp}^{\mathcal{N}}_{\mathcal{N}(\mu_0,K_0)}((v_0, v_1)) &= \mathcal{N}(\mu_0 + v_0, (I + v_1)_{\#} K_0), \end{aligned} \tag{4}$$

where $v = (v_0, v_1)$ is a tangent vector and T is the optimal map between $\mathcal{N}(0, K_0)$ and $\mathcal{N}(0, K_1)$.

2.3 Riemannian Geometry

We briefly recall Riemannian geometry required in this work. For more, see [10]. A smooth manifold M is *Riemannian*, if it admits a smoothly varying inner product $g_p(\cdot, \cdot)$ in the tangent space T_pM for any $p \in M$, called a *Riemannian metric*. The metric g induces a distance function d between two points $p, q \in M$, by measuring the length of curves minimally connecting p and q. These curves are called *geodesics*. Any geodesic γ originating from some $p \in M$, is uniquelly defined by a pair (p, v) in the *tangent bundle* TM of M, where $v = \dot{\gamma}(0)$. The tangent bundle is defined as the disjoint union of the tangent spaces T_pM of all $p \in M$.

The *Riemannian exponential* $\mathrm{Exp} : TM \to M$ takes a point in the tangent bundle, and maps it to the point on the associated geodesic $\gamma_{(p,v)}$ at time one, that is,

$$\mathrm{Exp} : TM \to M, \ (p, v) \mapsto \mathrm{Exp}_p(v) := \gamma_{(p,v)}(1). \tag{5}$$

The exponential at p forms a diffeomorphism between a neighborhood $0 \in U_p \subset T_pM$ and a neighborhood $p \in V_p \subset M$, where V_p is chosen maximally to preserve injectivity. Then, the set V_p is the *area of injectivity* at p. Inside of V_p, the exponential Exp_p admits the inverse $\mathrm{Log}_p : V_p \to T_pM$, called the *Riemannian logarithm* at p, characterized by $\mathrm{Exp}_p(\mathrm{Log}_p(p')) = p'$, for any $p' \in V_p$. The complement of V_p is called the *cut-locus* at p, denoted by $\mathcal{C}_p$. If the cut-locus is empty, then $V_p = M$, and so the logarithmic map is defined everywhere. An important class of such manifolds consists of manifolds with *non-positive curvature*.

Given the *Levi-Civita connection* ∇ on M, we can define the *parallel transport* $P_{p_0,p_1} : T_{p_0}M \to T_{p_1}M$, that intuitively moves a vector in a parallel manner along the geodesic γ connecting p_0 and p_1. More rigorously, fix $v \in T_{p_0}M$. Then, there exists a unique vector field $V : [0, 1] \to TM$, $V_t \in T_{\gamma(t)}M$ along γ with

$$\nabla_{\dot{\gamma}} V = 0, \tag{6}$$

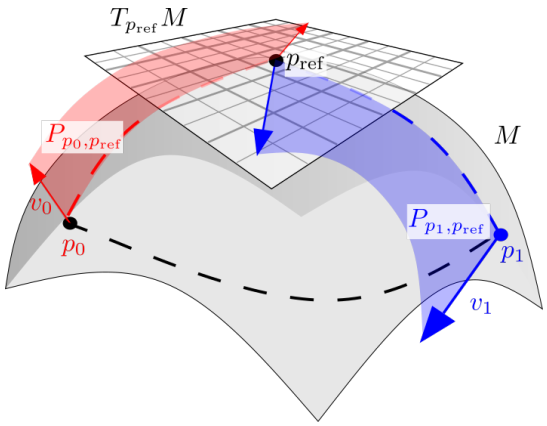


Figure 2: An illustration of the metric $d_{p_{\text{ref}}}$. We measure the distance between the basepoints p_0 and p_1 along the geodesic (dashed line). Then, we parallel transport v_0 and v_1 to the tangent space $T_{p_{\text{ref}}}M$ at the reference point, where the distance between the vectors is computed.

satisfying $V(0) = v$. Then, we define $P_{p_0,\gamma(t)}v := V(t)$, and note that $P_{\gamma(s),\gamma(t)}$ is a linear isometry between $T_{\gamma(s)}M$ and $T_{\gamma(t)}M$. Finally, when we fix bases for T_pM and T_qM, by abuse of notation, we use $P_{p,q}$ to denote the matrix associated with the parallel transport from T_pM to T_qM.

3 Metric on the Tangent Bundle

Let (M, g) be a Riemannian manifold with the induced metric function d_M, and TM its tangent bundle. We wish to define a metric function $d_{p_{\text{ref}}}$ on TM, that allows an analytical expression for the Wasserstein distance between wrapped Gaussian distributions, to be defined in Section 4. The idea is to form a distance using two terms: a basepoint distance given by d_M, and a parallel transport scheme for comparing the tangent vectors in different tangent spaces. For this, we will have to fix a reference point p_{ref}, where the tangent vectors will be transported. If we were to compare the tangent vectors by parallel transporting one to the other's tangent space, the resulting "distance" would not satisfy the triangle inequality.

Due to the reference point p_{ref}, we get a family of distances on the tangent bundle TM. Effectively, the tangent space $T_{p_{\text{ref}}}M$ at p_{ref} gives us a reference orientation system for the tangent vectors in the tangent bundle. In practice, for a given data set, we will choose p_{ref} to be the *Fréchet mean* of the data set (see Section 7 for definition). See Fig. 2 for an illustration of the construction of $d_{p_{\text{ref}}}$.

Proposition 1. *Let (M, g) be a Riemannian manifold with the induced metric function d_M, furthermore let $x_i = (p_i, v_i) \in TM$ for $i = 0, 1$ and fix a reference*

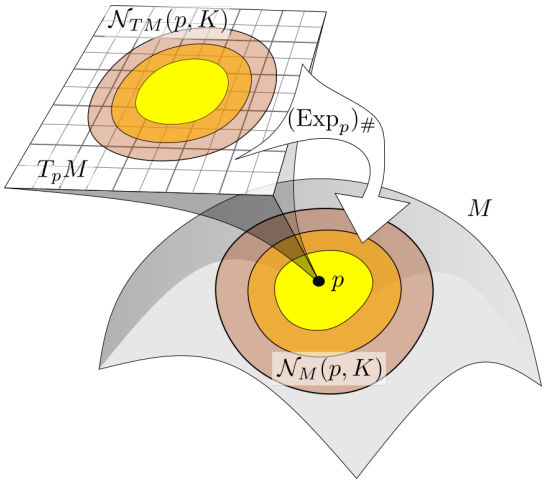


Figure 3: Illustrating a WGD on a manifold M. First, a pullback WGD $\mathcal{N}_{TM}(p, K)$ is defined in the tangent space T_pM, which is then pushed-forwad onto M by the exponential map Exp_p.

point $p_{\mathrm{ref}} \in M$. *Now define* $d_{p_{\mathrm{ref}}} : TM \times TM \to \mathbb{R}$ *by*

$$d_{p_{\mathrm{ref}}}(x_0, x_1)^2 = \begin{cases} d_M(p_0, p_1)^2 + \|P_{p_0,p_{\mathrm{ref}}}(v_0) - P_{p_1,p_{\mathrm{ref}}}(v_1)\|^2_{p_{\mathrm{ref}}}, \ p_0, p_1 \notin \mathcal{C}_{p_{\mathrm{ref}}}, \\ \infty, \ \text{otherwise.} \end{cases} \tag{7}$$

Here, $\|\cdot\|_{p_{\mathrm{ref}}}$ *is the norm in* $T_{p_{\mathrm{ref}}}M$ *induced by the Riemannian metric at* p_{ref}. *The condition* $p_0, p_1 \notin \mathcal{C}_{p_{\mathrm{ref}}}$ *is required for the parallel translations to be well-defined. Then,* $(TM, d_{p_{\mathrm{ref}}})$ *forms a metric space.*

Proof. This follows immediately from the fact, that the parallel transportation between tangent spaces forms an isometry, and that $d_{p_{\mathrm{ref}}}$ can be viewed as a product metric on $M \times T_{p_{\mathrm{ref}}}M$. □

4 Wrapped Gaussian Distributions

We now recall *wrapped Gaussian distributions* (WGDs) on a manifold M. We start by defining *pullback wrapped Gaussian distributions* (pullback-WGD) on the tangent bundle TM, which are then pushed forwad using the Riemannian exponential map, yielding a WGD. The idea of WGDs originated in *directional statistics* [20].

Fix $p \in M$ and let $\mathcal{N}(x|0, K)$ be the probability density function of $\mathcal{N}(0, K)$. A random point X on TM is said to be distributed according to a *pullback wrapped Gaussian distribution* (pullback-WGD) with *basepoint* μ and *tangent space covariance* K, if it follows the distribution

$$\mathbb{P}\{X = (q, v)\} = \chi_p(q)\mathcal{N}(v|0, K), \tag{8}$$

where χ_p is the indicator function satisfying $\chi_p(q) = 1$ if $p = q$, $\chi_p(q) = 0$ otherwise. In other words, X is a Gaussian random vector living in T_pM. We use the notation $X \sim \mathcal{N}_{TM}(p, K)$

A random point Y on M is said to be distributed according to a wrapped Gaussian distribution (WGD) on M, if

$$Y = \mathrm{Exp}(X), \tag{9}$$

for some random point $X \sim \mathcal{N}_{TM}(p, K)$ on TM. Then, the distribution of Y is given by

$$Y \sim (\mathrm{Exp}_p)_{\#}\mathcal{N}_{TM}(p, K), \tag{10}$$

which we denote by $Y \sim \mathcal{N}_M(p, K)$. We call X a *pullback* of Y.This construction is illustrated in Fig. 3.

5 2-Wasserstein Distance between Wrapped Gaussian Distributions

Let $X_i \sim \mathcal{N}_{TM}(p_i, K_i)$ be distributed according to pullback-WGDs on TM, and denote their laws by μ_i for $i = 0, 1$. Then, the 2-Wasserstein distance between them, with the tangent bundle metric given by $d_{p_{\mathrm{ref}}}$, is computed as

$$\begin{aligned} W_2^2(X_0, X_1) &= \inf_{\gamma \in \mathrm{Adm}(\mu_0, \mu_1)} \int_{TM} d^2_{p_{\mathrm{ref}}}((q_0, v_0), (q_1, v_1)) d\gamma((q_0, v_0), (q_1, v_1)) \\ &= \inf_{\gamma \in \mathrm{Adm}(\mu_0, \mu_1)} \int_{T_{p_{\mathrm{ref}}}M} \| P_{p_0, p_{\mathrm{ref}}}(v_0) - P_{p_1, p_{\mathrm{ref}}}(v_1) \|^2_{p_{\mathrm{ref}}} d\gamma \\ &\quad + d^2_M(p_0, p_1), \end{aligned} \tag{11}$$

where the last term is the 2-Wasserstein distance between $(P_{p_0, p_{\mathrm{ref}}})_{\#}\mathcal{N}(0, K_0)$ and $(P_{p_1, p_{\mathrm{ref}}})_{\#}\mathcal{N}(0, K_1)$ in the tangent space $T_{p_{\mathrm{ref}}}M$, which is given by

$$\begin{aligned} & W_2^2([P_{p_0, p_{\mathrm{ref}}}]_{\#}\mathcal{N}(0, K_0), (P_{p_1, p_{\mathrm{ref}}})_{\#}\mathcal{N}(0, K_1)) \\ =& W_2^2(\mathcal{N}(0, P^T_{p_0, p_{\mathrm{ref}}} K_0 P_{p_0, p_{\mathrm{ref}}}), \mathcal{N}(0, P^T_{p_1, p_{\mathrm{ref}}} K_1 P_{p_1, p_{\mathrm{ref}}})) \\ =& \mathrm{Tr}\left(P^T_{p_1, p_{\mathrm{ref}}} K_1 P_{p_1, p_{\mathrm{ref}}} + P^T_{p_0, p_{\mathrm{ref}}} K_0 P_{p_0, p_{\mathrm{ref}}}\right) \\ & - 2\mathrm{Tr}\left((P^T_{p_1, p_{\mathrm{ref}}} K_1 P_{p_1, p_{\mathrm{ref}}})^{\frac{1}{2}} P^T_{p_0, p_{\mathrm{ref}}} K_0 P_{p_0, p_{\mathrm{ref}}} (P^T_{p_1, p_{\mathrm{ref}}} K_1 P_{p_1, p_{\mathrm{ref}}})^{\frac{1}{2}}\right)^{\frac{1}{2}} \\ =& \mathrm{Tr}(K_1) + \mathrm{Tr}(K_2) - 2\mathrm{Tr}\left(K_1^{\frac{1}{2}} P_{p_1, p_{\mathrm{ref}}} P^T_{p_0, p_{\mathrm{ref}}} K_0 P_{p_0, p_{\mathrm{ref}}} P^T_{p_1, p_{\mathrm{ref}}} K_1^{\frac{1}{2}}\right), \end{aligned} \tag{12}$$

where we used the fact that the parallel transport is an isometry, and thus the associated matrix is orthogonal. Recall, that for an orthogonal matrix P and symmetric, positive definite K, we have $\mathrm{Tr}(P^T K P) = \mathrm{Tr}(K)$, and $(P^T K P)^{\frac{1}{2}} = P^T K^{\frac{1}{2}} P$. Combining the above, we get the following proposition.

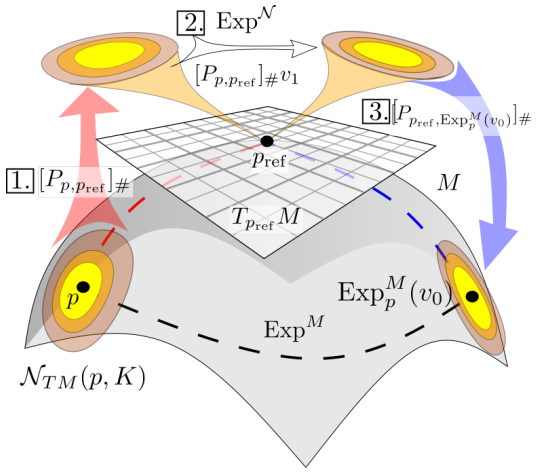


Figure 4: The exponential map $\mathrm{Exp}^{\mathcal{N}_{TM}}$ at $\mathcal{N}_{TM}(p,K)$ acts on a tangent vector (v_0,v_1) by sending the basepoint p to $\mathrm{Exp}_p^M(v_0)$. **1.** The covariance matrix K is parallel transported to $T_{p_{\mathrm{ref}}}M$, **2.** where it functions as the basepoint for the exponential $\mathrm{Exp}^{\mathcal{N}}$, mapping the parallel translation of the tangent vector component v_1. **3.** The result is then parallel transported from $T_{p_{\mathrm{ref}}}M$ to $T_{\mathrm{Exp}_p^M(v_0)}M$.

Proposition 2. *Let $X_i \sim \mathcal{N}_{TM}(p_i,K_i)$ be distributed as pullback-WGDs for $i=1,2$. Their 2-Wasserstein distance is then given by*

$$\begin{aligned} W_2^2(X_0,X_1) &= d_M^2(p_0,p_1) + W_2^2([P_{p_0,p_{\mathrm{ref}}}]_{\#}\mathcal{N}(0,K_0),[P_{p_1,p_{\mathrm{ref}}}]_{\#}\mathcal{N}(0,K_1)) \\ &= d_M^2(p_0,p_1) + \mathrm{Tr}(K_1) + \mathrm{Tr}(K_2) \\ &\quad - 2\mathrm{Tr}\left(K_1^{\frac{1}{2}}P_{p_1,p_{\mathrm{ref}}}P_{p_0,p_{\mathrm{ref}}}^T K_0 P_{p_0,p_{\mathrm{ref}}}P_{p_1,p_{\mathrm{ref}}}^T K_1^{\frac{1}{2}}\right). \end{aligned} \tag{13}$$

Proof. □

Proposition 2 yields a formula for the 2-Wasserstein distance W_2 between pullback-WGDs on TM. As the pullback of a WGD on M might not be unique, we quotient the space of pullback WGDs by definining an equivalence class: the pullbacks are equivalent if their push-forwards are equal. Then, we may define a quotient pseudometric [4] $d_{W,p_{\mathrm{ref}}}$ between the wrapped Gaussian distributions on M. For manifolds with empty cut locus, each WGD on M is the push-forward of exactly one WGD in TM, and hence the quotient pseudometric $d_{W,p_{\mathrm{ref}}}$ is a metric and is given by the following equation:

Definition 1. *Assume that M has an empty cut locus. For any $Y_i \sim \mathcal{N}_M(p_i,K_i)$, $i=0,1$, distributed as WGDs on M, let $X_i \sim \mathcal{N}_{TM}(p_i,K_i)$ have the corresponding unique pullback-WGDs on TM. Then, the quotient metric $d_{W,p_{\mathrm{ref}}}$ between Y_0 and Y_1 on M is given as*

$$d_{W,p_{\mathrm{ref}}}(Y_0,Y_1) := W_2(X_0,X_1). \tag{14}$$

An important class of manifolds with an empty cut-locus is the set complete and connected manifolds with non-positive curvature (*Hadamard manifolds*).

We now turn to study the geodesics between two WGDs on M under the metric $d_{W,p_{\text{ref}}}$.

Proposition 3. *Let $Y_i \sim \mathcal{N}_M(p_i, K_i)$ for $i = 0, 1$. Assuming that $v = \text{Log}_{p_0} p_1$ exists, we define the geodesic $p_t = \text{Exp}_{p_0}(tv)$ between the basepoints. Furthermore, let T be the optimal map between $[P_{p_0,p_{\text{ref}}}]_\# \mathcal{N}(0, K_0)$ and $[P_{p_1,p_{\text{ref}}}]_\# \mathcal{N}(0, K_1)$, so that the geodesic K_t between their covariance matrices is given by*

$$K_t = [(1-t)I + tT]_\# [P_{p_0,p_{\text{ref}}}]_\# K_0. \tag{15}$$

Then, the path

$$\gamma(t) = Y_t \sim \mathcal{N}_M(p_t, [P_{p_{\text{ref}},p_t}]_\# K_t), \tag{16}$$

defines a geodesic between Y_0 and Y_1 under the metric $d_{W,p_{\text{ref}}}$ in $\mathcal{N}_M$, the space of WGDs on M.

Proof. By a straight-forward computation, we get

$$\begin{aligned}
&d^2_{W,p_{\text{ref}}}(Y_{t_0}, Y_{t_1}) \\
=&W_2^2\left(\mathcal{N}_{TM}(p_{t_0}, [P_{p_{\text{ref}},p_{t_0}}]_\# K_{t_0}), \mathcal{N}_{TM}(p_{t_1}, [P_{p_{\text{ref}},p_{t_1}}]_\# K_{t_1})\right) \\
=&d^2_M(p_{t_0}, p_{t_1}) + W_2^2\left(\mathcal{N}(0, [P_{p_{t_0},p_{\text{ref}}}]_\# [P_{p_{\text{ref}},p_{t_0}}]_\# K_{t_0}), \mathcal{N}(0, [P_{p_{t_1},p_{\text{ref}}}]_\# [P_{p_{\text{ref}},p_{t_1}}]_\# K_{t_1})\right) \\
=&d^2(p_{t_0}, p_{t_1}) + W_2^2\left(\mathcal{N}(0, K_{t_0}), \mathcal{N}(0, K_{t_1})\right) \\
=&(t_1 - t_0)^2 d^2(p_0, p_1) + (t_1 - t_0)^2 W_2^2\left([P_{p_0,p_{\text{ref}}}]_\# \mathcal{N}(0, K_0), [P_{p_1,p_{\text{ref}}}]_\# \mathcal{N}(0, K_1)\right) \\
=&(t_1 - t_0)^2 d^2_{W,p_{\text{ref}}}(Y_0, Y_1),
\end{aligned} \tag{17}$$

where the first equation follows from the definition of $d_{W,p_{\text{ref}}}$, the second from Proposition 2, third from the parallel transports forth and back cancelling each other, and fourth follows from p_t and K_t being geodesics on M and $\mathcal{N}$, respectively. The computations shows that Y_t forms a geodesic, as it is unit parametrized with length equal to the distance between Y_0 and Y_1. □

The intuition behind this result, is that the geodesics of WGDs under $d_{W,p_{\text{ref}}}$ are given by following geodesics between the basepoints on M while parallel transporting the geodesic between Gaussian distributions in $(\mathcal{N}_{T_{p_{\text{ref}}}M}, W_2)$ in the reference tangent space. Here $\mathcal{N}_{T_{p_{\text{ref}}}M}$ denotes the space of Gaussian distributions on $T_{p_{\text{ref}}}M$. See Fig. 1 for an illustration on the 2-sphere.

6 Riemannian Structure of $(\mathcal{N}_{TM}, W_2)$

The space $(\mathcal{N}_{TM}, W_2)$ of pullback-WGDs can be viewed as the product manifold $M \times \mathcal{N}_{T_{p_{\text{ref}}}M}$ with the Riemannian metric given by

$$g^{\mathcal{N}_{TM}}_{\mathcal{N}_M(p,K)}((v_0, v_1), (u_0, u_1)) = g^M_p(v_0, u_0) + g^{\mathcal{N}}_{[P_{p,p_{\text{ref}}}]_\# \mathcal{N}(0,K)}(P_{p,p_{\text{ref}}}(v_1), P_{p,p_{\text{ref}}}(u_1)), \tag{18}$$

for tangent vectors (v_0, v_1) and (u_0, u_1) in the tangent space

$$T_{\mathcal{N}_{TM}(p,K)} \simeq T_p M \times T_{\mathcal{N}(0,K)} \mathcal{N}_{T_{p_{\text{ref}}}M}. \tag{19}$$

Here g^M denotes the Riemannian metric on M, and $g^{\mathcal{N}}$ the Riemannian metric on $\mathcal{N}_{T_{p_{\text{ref}}}M}$.

Note, how the Riemannian structure of $M \times \mathcal{N}_{T_{p_{\text{ref}}}M}$ is naturally extended to the whole of $\mathcal{N}_{TM}$ via parallel transportation. Furthermore, as Proposition 3 shows that the geodesics on $\mathcal{N}_{TM}$ are given by combining geodesics on M and $\mathcal{N}_{T_{p_{\text{ref}}}M}$, we conclude that W_2 is induced by $g^{\mathcal{N}_{TM}}$.

Then, the Riemannian exponential and logarithm on $(\mathcal{N}_{TM}, g^{\mathcal{N}_{TM}})$ are given by

$$\begin{aligned}
&\text{Log}^{\mathcal{N}_{TM}}_{\mathcal{N}_M(p_0,K_0)}(\mathcal{N}_M(p_1,K_1)) \\
:=&\left(\text{Log}^M_{p_0}(p_1),\ [P_{p_{\text{ref}},p_0}]_\# \text{Log}^{\mathcal{N}}_{[P_{p_0,p_{\text{ref}}}]_\#\mathcal{N}(0,K_0)}([P_{p_1,p_{\text{ref}}}]_\#\mathcal{N}(0,K_1))\right) \\
&\text{Exp}^{\mathcal{N}_{TM}}_{\mathcal{N}_M(p,K)}((v_0,v_1)) \\
:=&\mathcal{N}_{TM}\left(\text{Exp}^M_p(v_0),\ [P_{p_{\text{ref}},\text{Exp}^M_p(v_0)}]_\# \text{Exp}^{\mathcal{N}}_{[P_{p,p_{\text{ref}}}]_\#\mathcal{N}(0,K)}([P_{p,p_{\text{ref}}}]_\# v_1)\right),
\end{aligned} \tag{20}$$

see Fig. 4 for an illustration of the exponential map. Summarizing the above, we get the following result

Theorem 1. *The metric space $(\mathcal{N}_{TM}, W_2)$ is induced by the Riemannian metric $(\mathcal{N}_{TM}, g^{\mathcal{N}_{TM}})$.*

Proof. □

Remark, that when the pullbacks of WGDs on M are uniquely defined, we can extend this Riemannian structure to $(\mathcal{N}_M, d_{W,p_{\text{ref}}})$.

7 Fréchet Means on $(\mathcal{N}_{TM}, W_2)$

A *Fréchet mean* [12] is a generalization of the Euclidean population mean to metric spaces. Formally, $\bar{x}$ is a Fréchet mean of a population $x_1, x_2, ..., x_N \in X$ on a metric space (X, d), if it satisfies

$$\bar{x} \in \arg\min_x \sum_{i=1}^N d^2(x_i, x). \tag{21}$$

Note that unlike in the Euclidean case, the Fréchet mean is not necessarily unique.

Let $\mathcal{N}_{TM}(p_i, K_i),\ i = 1, 2, ..., N$ be a population of pullback-WGDs on TM. Then, by (21), the Fréchet mean $\mathcal{N}_{TM}(\bar{p}, \bar{K})$ minimizes

$$\begin{aligned}
&\sum_{i=1}^N W_2^2(\mathcal{N}_{TM}(p_i,K_i), \mathcal{N}_{TM}(p,K)) \\
=&\sum_{i=1}^N \left(d_M^2(p_i,p) + W_2^2(\mathcal{N}(0,[P_{p_i,p_{\text{ref}}}]_\# K_i), [P_{p,p_{\text{ref}}}]_\#\mathcal{N}(0,K))\right).
\end{aligned} \tag{22}$$

The two terms on the right-hand side can be minimized independently. Thus, we split the computation into two: first, compute $\bar{p}$ as the Fréchet mean on M and use it as the reference point p_{ref} for the metric $d_{p_{\text{ref}}}$ on TM. Then, $P_{p,p_{\text{ref}}}$ is the identity transformation, so it suffices to compute $\mathcal{N}(0, \bar{K})$ as the Fréchet mean of the population $\mathcal{N}(0, [P_{p_i,p_{\text{ref}}}]_{\#}K_i)$, $i = 1, 2, ..., N$ on the manifold $\mathcal{N}$ of Gaussian distributions. See Fig. 5 for an example of a Fréchet mean on the 2-sphere, where we have computed the Fréchet mean of three pullback-WGDs, and visualized them by their push-forward WGDs on M.

As $\mathcal{N}_{TM}$ is a Riemannian manifold, and the optimization of the two terms are carried on submanifolds, a common strategy for the computation of the mean is to note that on a Riemannian manifold (N, g), we have

$$\frac{d}{dp} d_N^2(p, p_i)\Big|_{p=\bar{p}} = -2\text{Log}_{\bar{p}}^N(p_i), \tag{23}$$

which can then be applied with a gradient based minimization algorithm.

8 Discussion

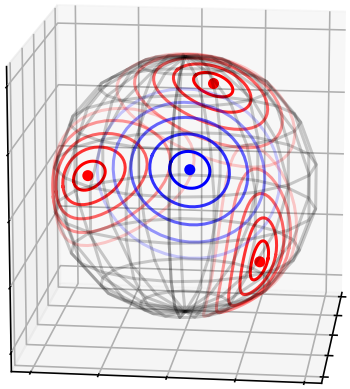


Figure 5: A population of WGDs (in red) on the 2-sphere, visualized by mapping σ confidence intervals from corresponding pullback-WGDs. The Fréchet mean is given in blue. As the 2-sphere has non-empty cut-locus, the pullback distributions are not necessarily uniquely defined. Here, the Fréchet mean is computed between chosen represenative pullback WGDs.

We provided an analytical formula for the 2-Wasserstein metric between pullback-WGDs, and showed that the distance is induced by a Riemannian structure on the space of pullback-WGDs. We then extended the results to a class of WGDs on a Riemannian manifold M, and visualized a geodesic and a Fréchet mean on the 2-sphere.

The extension to WGDs is well defined whenever the base manifold M allows uniqueness of the pullback-WGDs. As we remarked, this follows immediately, if M has an empty cut-locus. Other cases have not been as well studied, which poses an open question; what are the sufficient conditions for a pullback of WGD to be uniquely defined?

An alternative to our approach would be to compute a distance (or a similarity measure) directly between distributions on M. However, tackling the expressions for the *Kullback-Leibler* (KL) *divergence* or the 2-Wasserstein distance seems quite challenging. In contrast to this, the pullback approach seems attractive for its analytical formulas, even though the aforementioned complications are encountered.

Acknowledgements

This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Furthermore, the authors wish to thank Ivan D'Annibale for inspirational conversations and revision of the text.

References

[1] Ambrosio, L., Gigli, N.: A user's guide to optimal transport. In: Modelling and optimisation of flows on networks, pp. 1–155. Springer (2013)

[2] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)

[3] Bernton, E., Jacob, P.E., Gerber, M., Robert, C.P.: Inference in generative models using the wasserstein distance. arXiv preprint arXiv:1701.05146 (2017)

[4] Bridson, M.R., Haefliger, A.: Metric spaces of non-positive curvature, vol. 319. Springer Science & Business Media (2013)

[5] Bures, D.: An extension of kakutani?s theorem on infinite product measures to the tensor product of semifinite ??*-algebras. Transactions of the American Mathematical Society **135**, 199–212 (1969)

[6] Cheeger, J., Gromoll, D.: On the structure of complete manifolds of nonnegative curvature. Annals of Mathematics pp. 413–443 (1972)

[7] Chen, Y., Georgiou, T.T., Tannenbaum, A.: Optimal transport for gaussian mixture models. arXiv preprint arXiv:1710.07876 (2017)

[8] Chen, Y., Ye, J., Li, J.: A distance for hmms based on aggregated wasserstein metric and state registration. In: European Conference on Computer Vision. pp. 451–466. Springer (2016)

[9] Cuesta-Albertos, J., Matrán-Bea, C., Tuero-Diaz, A.: On lower bounds for thel 2-wasserstein metric in a hilbert space. Journal of Theoretical Probability **9**(2), 263–283 (1996)

[10] Do Carmo, M.P.: Riemannian geometry. Birkhauser (1992)

[11] Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. Journal of multivariate analysis **12**(3), 450–455 (1982)

[12] Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Ann. Inst. H. Poincaré **10**(3), 215–310 (1948)

[13] Galichon, A.: Optimal Transport Methods in Economics. Princeton University Press (2016)

[14] Givens, C.R., Shortt, R.M., et al.: A class of wasserstein metrics for probability distributions. The Michigan Mathematical Journal **31**(2), 231–240 (1984)

[15] Gudmundsson, S., Kappos, E.: On the geometry of tangent bundles. Expositiones Mathematicae **20**(1), 1–41 (2002)

[16] Haker, S., Zhu, L., Tannenbaum, A., Angenent, S.: Optimal mass transport for registration and warping. International Journal of computer vision **60**(3), 225–240 (2004)

[17] Kantorovich, L.V.: On the translocation of masses. In: Dokl. Akad. Nauk. USSR (NS). vol. 37, pp. 199–201 (1942)

[18] Knott, M., Smith, C.S.: On the optimal mapping of distributions. Journal of Optimization Theory and Applications **43**(1), 39–49 (1984)

[19] Mallasto, A., Feragen, A.: Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In: Advances in Neural Information Processing Systems. pp. 5665–5674 (2017)

[20] Mardia, K.V., Jupp, P.E.: Directional statistics, vol. 494. John Wiley & Sons (2009)

[21] Monge, G.: Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris (1781)

[22] Montavon, G., Müller, K.R., Cuturi, M.: Wasserstein training of restricted boltzmann machines. In: Advances in Neural Information Processing Systems. pp. 3718–3726 (2016)

[23] Muzellec, B., Cuturi, M.: Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions. ArXiv e-prints (May 2018)

[24] Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra and its Applications **48**, 257–263 (1982)

[25] Peyré, G., Cuturi, M., et al.: Computational optimal transport. Tech. rep. (2017)

[26] Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)

[27] Sasaki, S.: On the differential geometry of tangent bundles of riewannian manifolds. Tohoku Mathematical Journal, Second Series **10**(3), 338–354 (1958)

[28] Takatsu, A., et al.: Wasserstein geometry of gaussian measures. Osaka Journal of Mathematics **48**(4), 1005–1026 (2011)

[29] Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008)

[30] Yano, K., Kobayashi, S.: Prolongations of tensor fields and connections to tangent bundles i. Journal of the Mathematical Society of Japan **18**(2), 194–210 (1966)

Appendix E

A Formalization of the Natural Gradient Method for General Similarity Measures

The following chapter presents (up to formatting) the article

> Anton Mallasto, Tom Dela Haije, and Aasa Feragen. "A Formalization of The Natural Gradient Method for General Similarity Measures." Proceedings of 4$^{\text{th}}$ International Conference on Geometric Science of Information (GSI). 2019.

This work presents a generalization of the natural gradient to arbitrary divergences. Traditionally, the natural gradient [4] has been associated with the Riemannian gradient resulting from pulling back the *Fisher–Rao* metric onto the statistical manifold. The Fisher–Rao metric can also be viewed as the Riemannian metric associated with the Kullback–Leibler divergence [5]. A more modern approach is to define the natural gradient with respect to any chosen Riemannian geometry on the statistical manifold [6, Sec. 12]. However, this poses the question, whether the chosen Riemannian geometry relates reasonably to the geometry induced on the space of probabilities by the divergence.

The article aims at motivating the use of the pullback metric on the statistical manifold associated with the divergence being minimized, which we interpret as a *local Hessian* of the divergence. Furthermore, relations between different optimization methods, such as the *proximal method*, *trust-region method*, and the natural gradient are discussed. The article ends with providing example computations for the natural gradient under different divergences, notably in the p-Wasserstein case, yielding in the $p = 2$ case the Wasserstein natural gradient [17, 40].

A Formalization of the Natural Gradient Method for General Similarity Measures

Anton Mallasto Tom Dela Haije Aasa Feragen

Department of Computer Science, University of Copenhagen

Abstract

In optimization, the natural gradient method is well-known for likelihood maximization. The method uses the Kullback–Leibler (KL) divergence, corresponding infinitesimally to the Fisher–Rao metric, which is pulled back to the parameter space of a family of probability distributions. This way, gradients with respect to the parameters respect the Fisher–Rao geometry of the space of distributions, which might differ vastly from the standard Euclidean geometry of the parameter space, often leading to faster convergence. The concept of natural gradient has in most discussions been restricted to the KL-divergence / Fisher–Rao case, although in information geometry the local C^2 structure of a general divergence has been used for deriving a closely related Riemannian metric analogous to the KL-divergence case. In this work, we wish to cast natural gradients into this more general context and provide example computations, notably in the case of a Finsler metric and the p-Wasserstein metric. We additionally discuss connections between the natural gradient method and multiple other optimization techniques in the literature.

1 Introduction

The natural gradient method [2] in optimization originates from *information geometry* [4], which utilizes the Riemannian geometry of statistical manifolds (the parameter spaces of model families) endowed with the *Fisher–Rao metric*. The natural gradient is used for minimizing the *Kullback–Leibler* (KL) divergence, a *similarity measure* between a model distribution and a target distribution, that can be shown to be equivalent to maximizing model likelihood of given data. The success of natural gradient in optimization stems from accelerating likelihood maximization and providing infinitesimal invariance to reparametrizations of the model, providing robustness towards arbitrary parametrization choices.

In the modern formulation of the natural gradient, a *Riemannian metric* on the statistical manifold is chosen, with respect to which the gradient of the

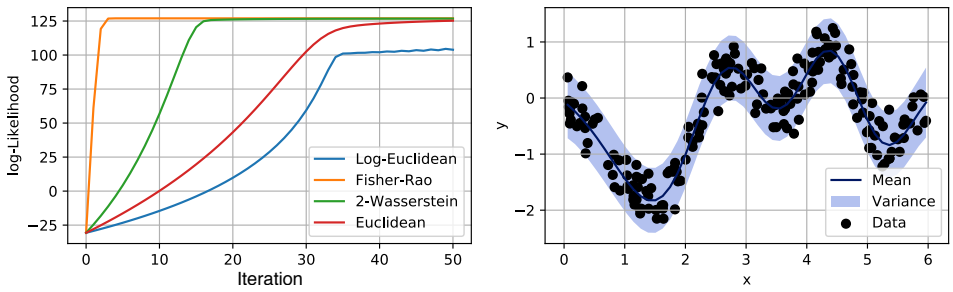


Figure 1: Maximizing prior likelihood for Gaussian process regression using natural gradients under different metrics on Gaussian distributions. Convergence plots on left. Data and model fit, with optimal exponentiated quadratic kernel parameters, on right.

given similarity is computed [4, Sec. 12]. The choice of the Riemannian metric should, however, relate closely to the similarity measure being minimized. We have illustrated this in Fig. 1, where model selection for Gaussian process regression is carried out by maximizing the prior-likelihood of the data with natural gradients stemming form different metrics. Clearly, the Fisher–Rao metric—which infinitesimally corresponds to the KL-divergence—achieves the fastest convergence.

An example of an approach to choose a related Riemannian metric is the classical Newton's method that derives a metric from the Hessian of a convex objective function, or its absolute value in the non-convex case [7]. Unfortunately, evaluating the Hessian is not feasible in some cases. Instead, we can compute a *local Hessian*, which corresponds to a local second order expansion of the similarity measure [3]. This approach generalizes the natural gradient from the KL-divergence case to general similarity measures, and to avoid confusion with the well-known KL-divergence setting, we refer to this approach as the *formal natural gradient*. We furthermore discuss the similarities between the trust region, proximal, and natural gradient methods in Section 3 and provide example computations in Section 4.

2 Useful Metrics via Formalizing the Natural Gradient

The natural gradient is computed with respect to a chosen metric on the statistical manifold, which often results from pulling back a metric between distributions. This way, the gradient takes into account how the metric on distributions penalizes movement into different directions. We will now review how the natural gradient is computed given a Riemannian metric. Then, we introduce the formal natural gradient, which derives this metric from the similarity measure.

Statistical manifold. Let $\mathrm{AC}(X)$ denote the set of absolutely continuous

probability distributions on some manifold X. A *statistical manifold* is defined by a triple (X, Θ, ρ), where X is called the *sample space* and $\Theta \subseteq \mathbb{R}^n$ the *parameter space*. Then, $\rho\colon \Theta \to \mathrm{AC}(X)$ maps a parameter to a density, given by $\rho\colon \theta \mapsto \rho_\theta(\cdot)$, for any $\theta \in \Theta$. Abusing terminology, we also call Θ the statistical manifold.

Cost function. Let a *similarity measure* $c^*\colon \mathrm{AC}(X) \times \mathrm{AC}(X) \to \mathbb{R}_{\geq 0}$ (e.g. a metric or an information divergence) be defined on $\mathrm{AC}(X)$ satisfying $c^*(\rho, \rho') = 0$ if and only if $\rho = \rho'$. Assume c^* to be strictly convex in ρ. Given a target distribution $\rho \in \mathrm{AC}(X)$ and a statistical manifold (X, Θ, ρ), we wish to minimize the *cost function* $c \to \Theta \times \mathrm{AC}(X) \to \mathbb{R}_{\geq 0}$ given by

$$c(\theta, \rho) = c^*(\rho_\theta, \rho). \tag{2.1}$$

If $\rho = \rho_{\theta'}$ for some $\theta' \in \Theta$, then by abuse of notation we write $c(\theta, \theta')$. We finally assume that $\theta \mapsto c(\theta, \theta')$ is C^2 whenever $\theta \neq \theta'$.

Natural gradient. Assume a Riemannian structure (Θ, g^Θ) on the statistical manifold. The *Riemannian metric* g^Θ induces a *metric tensor* G^Θ, given by $g^\Theta_\theta(u, v) = u^T G^\Theta_\theta v$ and a *distance function* which we denote by d_Θ. The vectors u, v belong to the *tangent space* $T_\theta\Theta$ at θ. It is common intuition that the negative gradient $v = -\nabla_\theta c(\theta, \rho)$ gives the direction of maximal descent for c. However, this is only true on a Euclidean manifold. Consider

$$\hat{v} = \operatorname*{arg\,min}_{v \in T_\theta\Theta : d_\Theta(\theta, \theta+v) = \Delta} c(\theta + v, \rho), \tag{2.2}$$

where $\theta + v$ is to be understood in a chart of Θ, and $\Delta > 0$ defines the radius of the trust region. Linearly approximating the objective and quadratically approximating the constraint, this is solved using Lagrangian multipliers, giving the *natural gradient*

$$\hat{v} = -\frac{1}{\lambda} \left[G^\Theta_\theta\right]^{-1} \nabla_\theta c(\theta, \rho), \tag{2.3}$$

for some Lagrangrian multiplier $\lambda > 0$, which we refer to as the *learning rate*. Below, a similar derivation is carried out in more detail.

Formal natural gradient. Traditionally, the natural gradient uses the Fisher–Rao metric when the similarity measure used is the KL-divergence. We will now show, how a trust region formulation with respect to the chosen similarity measure can be used to derive a natural metric under which the natural gradient can be computed, resulting in the *formal natural gradient*. Thus, consider the minimization task

$$\hat{v} := \operatorname*{arg\,min}_{v \in T_\theta\Theta,\ c(\theta+v, \theta) = \Delta} c(\theta + v, \rho). \tag{2.4}$$

We approximate the constraint by the second degree Taylor expansion

$$c(\theta + v, \theta) \approx \frac{1}{2} v^T \left(\nabla^2_{\eta \to \theta} c(\eta, \theta)\right) v, \tag{2.5}$$

where the 0^{th} and 1^{st} degree terms disappear as $c(\theta + v, \theta)$ has a minimum 0 at $v = 0$. We call the symmetric positive definite matrix $H^c_\theta := \nabla^2_{\eta \to \theta} c(\eta, \theta)$ the

local Hessian. Then, we further approximate the objective function

$$c(\theta + v, \rho) \approx c(\theta, \rho) + \nabla_\theta c(\theta, \rho)^T v. \tag{2.6}$$

Writing the approximate Langrangian $\mathcal{L}(v)$ of (2.4) with a multiplier $\lambda > 0$, we get

$$\mathcal{L}(v) \approx c(\theta, \rho) + \nabla_\theta c(\theta, \rho)^T v + \frac{\lambda}{2} v^T \left(\nabla^2_{\eta \to \theta} c(\eta, \theta)\right) v. \tag{2.7}$$

Thus by the method of Langrangian multipliers, (2.4) is solved as

$$\hat{v} = -\frac{1}{\lambda} \left[H^c_\theta\right]^{-1} \nabla_\theta c(\theta, \rho). \tag{2.8}$$

We refer to $\hat{v}$ as the *formal natural gradient* with respect to c.

Remark 1. *We could have just substituted $\eta = \theta$ in the local Hessian if $\nabla^2_\eta c(\eta, \theta)$ was continuous at η. However, when studying Finsler metrics later in this work, the expression has a discontinuity at $\eta = \theta$. Therefore, a direction for a limit has to be chosen, and as a straight-forward candidate we compute the limit from the direction of the gradient.*

Metric interpretation. The local Hessian G^c_θ can be seen as a metric tensor at any $\theta \in \Theta$, inducing an inner product $g^c_\theta \colon T_\theta\Theta \times T_\theta\Theta \to \mathbb{R}$ given by $g^c_\theta(v, u) = v^T H^c_\theta u$. This imposes a *pseudo-Riemannian* structure on Θ, forming the pseudo-Riemannian manifold (Θ, g^c). Therefore, G^c_x provides us a natural metric under which to compute the natural gradient for a general c^*. If ρ has a full rank Jacobian everywhere, then a Riemannian metric is retrieved. Also, there is an obvious *pullback* structure at play. Recall, that the cost is defined by $c(\theta, \theta') = c^*(\rho_\theta, \rho_{\theta'})$. Then, computing the local Hessian yields

$$H^c_\theta = J^T_\theta H^{c^*}_{\rho_\theta} J_\theta, \tag{2.9}$$

where $H^{c^*}_{\rho_\theta} = \nabla^2_{\rho \to \rho_\theta} c^*(\rho, \rho_\theta)$. Thus, H^c results from pulling back the c^* induced metric tensor H^{c^*} on $\mathrm{AC}(X)$ to the statistical manifold Θ. In information geometry, this Riemannian metric is said to be induced by the corresponding divergence (similarity measure) [3]. Therefore, the formal natural gradient is just the Riemannian gradient under the aforementioned induced metric.

Asymptotically Newton's method. We provide a straightforward result, stating that the local Hessian approaches the actual Hessian in the limit, thus the formal natural gradient method approaches Newton's method. This is well known in the Fisher–Rao case, but for completeness we provide the result for the formal natural gradient.

Proposition 1. *Assume $c(\theta, \rho) = c(\theta, \theta')$ for some $\theta' \in \Theta$, and that c is C^2 in θ. Then, the natural gradient yields asymptotically Newton's method.*

Proof. The Hessian at θ is given by $\nabla^2_\theta c(\theta, \theta')$. Then, as c is C^2 in the first argument, passing the limit $\theta \to \theta'$ yields

$$H^c_\theta = \nabla^2_{\eta \to \theta} c(\eta, \theta) \overset{\theta \to \theta'}{\to} \nabla^2_{\eta \to \theta'} c(\eta, \theta') = \nabla^2_{\eta = \theta'} c(\eta, \theta'), \tag{2.10}$$

where the last expression is the Hessian at θ'. □

3 Loved Child Has Many Names – Related Methods

In this section, we discuss connections between seemingly different optimization methods. Some of these connections have already been reported in the literature, some are likely to be known to some extent in the community. However, the authors are unaware of previous work drawing out these connections in their full extent. We provide such a discussion, and then present other related connections.

As discussed in [13], *proximal methods* and *trust region methods* are equivalent up to learning rate. Trust region methods employ an l^2-metric constraint

$$x_{t+1} = \underset{x:\|x-x_t\|_2\leq\Delta}{\arg\min} f(x),\ \Delta > 0, \tag{3.1}$$

whereas proximal methods include a l^2-metric penalization term

$$x_{t+1} = \arg\min_x \left\{ f(x) + \frac{1}{2\lambda}\|x - x_t\|_2^2 \right\},\ \lambda > 0, \tag{3.2}$$

The two can be shown to be equivalent up to learning rate via Lagrangian duality.

Instead of the l^2 metric penalization, *mirror gradient descent* [12] employs a more general *proximity function* $\Psi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{>0}$, that is strictly convex in the first argument. Then, the mirror descent step is given by

$$x_{t+1} = \arg\min_x \left\{ \langle x - x_t, \nabla f(x_t)\rangle + \frac{1}{\lambda}\Psi(x, x_t) \right\}. \tag{3.3}$$

Commonly, Ψ is chosen to be a *Bregman divergence* D_g, defined by choosing a strictly convex C^2 function g and writing

$$D_g(x, x') = g(x) - g(x') - \langle \nabla g(x'), x - x'\rangle. \tag{3.4}$$

To explain how these methods are related to the natural gradient, assume that we are minimizing a general similarity measure $c(x, y)$ with respect to x, as in Sec. 2. Recall, that we first defined the natural gradient as a *trust region step*. In order to derive an analytical expression for the iteration, we approximated the objective function with the first order Taylor polynomial and the constraints by the local Hessian and then used Lagrangian duality to yield a *proximal expression*, which yields the formal natural gradient when solved. In Sec. 4, we will show how this workflow indeed corresponds to known examples of the natural gradient.

Further connections. Raskutti and Mukherjee [15] showed, that Bregman divergence proximal mirror gradient descent is equivalent to the natural gradient method on the *dual manifold* of the Bregman divergence. Khan et al. [8], consider a KL divergence proximal algorithm for learning *conditionally conjugate*

exponential families, which they show to correspond to a natural gradient step. For exponential families, the KL-divergence corresponds to a Bregman divergence, and so the natural gradient step is on the *primal manifold* of the Bregman divergence. Thus the result seems to conflict with the resut in [15]. However, this can be explained, as the gradient is taken with respect to a different argument of the divergence, i.e., they consider $\nabla_x D_g(x', x)$ and not $\nabla_x D_g(x, x')$. It is intriguing how two different geometries are involved in this choice.

Pascanu and Bengio [14] remarked on the connections between the natural gradient method and Hessian-free optimization [10], Krylov Subspace Descent [17], and TONGA [16]. The main connection between Hessian-free optimization and Krylov subspace descent is the use of *extended Gauss–Newton approximation of the Hessian* [18], which gives a similar square form involving the Jacobian as the *pullback* Fisher–Rao metric on a statistical manifold. The connection was further studied by Martens [11], where an equivalence criterion between the Fisher–Rao natural gradient and extended Gauss–Newton was given.

4 Example Computations

We will now provide example computations for the local Hessian H^c of different similarity measures c, as it is the essential object in computing the natural gradient given in (2.8). We first show that in the cases of KL-divergence and a Riemannian metric, the definition of the formal natural gradient matches the classical definition, as expected. Furthermore, we contribute local Hessians for general f-divergences and Finsler metrics, specifically for the p-Wasserstein metrics.

Natural gradient of f-divergences. Let $\rho, \rho' \in \mathrm{AC}(X)$ and $f\colon \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ be a convex function satisfying $f(1) = 0$. Then, the f-divergence from ρ' to ρ is

$$D_f(\rho||\rho') = \int_X \rho(x) f\left(\frac{\rho'(x)}{\rho(x)}\right) dx. \tag{4.1}$$

Now, consider the statistical manifold $(\mathbb{R}^d, \Theta, \rho)$, and compute the local Hessian

$$\left[H_\theta^{D_f}\right]_{ij} = \nabla^2 f(1) \int_X \frac{\partial \log \rho_\theta(x)}{\partial \theta_i} \frac{\partial \log \rho_\theta(x)}{\partial \theta_j} \rho_\theta(x) dx. \tag{4.2}$$

Substituting $f = -\log$ in (4.1) results in the KL-divergence, denoted by $D_{\mathrm{KL}}(\rho||\rho')$. Noticing that $\nabla^2 f(1) = 1$ with this substitution, we can write (4.2) as $H_\theta^{D_f} = \nabla^2 f(1) H_\theta^{D_{\mathrm{KL}}}$, where the local Hessian $H_\theta^{D_{\mathrm{KL}}}$ is also the Fisher–Rao metric tensor at θ, and thus the natural gradient of Amari [2] is retrieved.

Natural gradient of Riemannian distance. Let (M, g) be a Riemannian manifold with the induced distance function d_g and the metric tensor at $\rho \in M$ denoted by G_ρ^M. Finally, denote by ρ_θ a submanifold of M parametrized by

$\theta \in \Theta$. Then, when $c = \frac{1}{2}d^2$, we compute $G_\theta^{\frac{1}{2}d_g}$ as follows

$$\begin{aligned}\left[H_\theta^{\frac{1}{2}d^2}\right]_{ij} =& \frac{1}{2}\left(\frac{\partial}{\partial\theta_j}\rho_\theta\right)^T \left[\nabla^2_{\rho_\eta\to\rho_\theta} d^2(\rho_\eta,\rho_\theta)\right]\left(\frac{\partial}{\partial\theta_i}\rho_\theta\right) \\ &+ \frac{1}{2}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_i}\rho_\theta\right]\left[\nabla_{\rho_\eta\to\rho_\theta} d^2(\rho_\eta,\rho_\theta)\right],\end{aligned} \tag{4.3}$$

as $\theta' \to \theta$, the second term vanishes. Finally, $\nabla^2_{\rho_\eta\to\rho_\theta} d^2(\rho_\eta,\rho_\theta) = 2G^M_{\rho_\theta}$, thus

$$H_\theta^{\frac{1}{2}d_g} = J_\theta^T G^M_{x_\theta} J_\theta, \tag{4.4}$$

where $J_\theta = \frac{\partial}{\partial\theta}\rho_\theta$ denotes the Jacobian. Therefore, the formal natural gradient corresponds to the traditional coordinate-free definition of a gradient on a Riemannian manifold, when the metric is given by the pullback

Natural gradient of Finsler distance. Let (M, F) denote a Finsler manifold, where $F_\rho \colon T_\rho M \to \mathbb{R}_{\geq 0}$, for any $\rho \in M$, is a *Finsler metric*, satisfying the properties of strong convexity, positive 1-homogeneity and positive definiteness. Then, a distance d_F is induced on M by

$$d_F(\rho,\rho') = \inf_\gamma \int_0^1 F_{\gamma(t)}(\dot\gamma(t))dt,\ \rho,\rho' \in M \tag{4.5}$$

where γ is any continuous, unit-parametrized curve with $\gamma(0) = \rho$ and $\gamma(1) = \rho'$.

The *fundamental tensor* G^F of F at (ρ, v) is defined as $G^F_\rho(v) = \frac{1}{2}\nabla^2_v F^2_\rho(v)$. Then, G^F_ρ is 0-homogeneous as the second differential of a 2-homogeneous function. Therefore, $G^F_\rho(\lambda v) = G^F_\rho(v)$ for any $\lambda > 0$. Furthermore, $G^F_\rho(v)$ is positive-definite when $v \neq 0$. Now, let $u = -J_\theta \nabla_\theta d^2_F(\rho_\theta,\rho')$, and as we can locally write $d^2_F(\rho,\rho') = F^2_{\rho\theta}(v)$ for a suitable v, then

$$H_\theta^{\frac{1}{2}d^2_F} = \frac{1}{2}\nabla^2_{\eta\to\theta} d^2_F(\rho_\eta,\rho_\theta) = \frac{1}{2}\lim_{\lambda\to 0}\nabla^2_{v=\lambda u} F^2_{\rho_\theta}(v) = J_\theta^T G^F_{\rho_\theta}(u) J_\theta. \tag{4.6}$$

Coordinate-free gradient descent on Finsler manifolds has been studied by Bercu [5]. The formal natural gradient differs slightly from this, as we use $v = -J_\theta\nabla_\theta d^2_F(\rho_\theta,\rho')$ in the preconditioning matrix $G^F_{(\rho_\theta,v)}$ (see Remark 1), where as in [5], v is chosen to maximize the descent. Thus the natural gradient descent in the Finsler case approximates the geometry in the direction of the gradient quadratically to improve the descent, but fails to take the entire local geometry into account.

***p*-Wasserstein metric.** Let $X = \mathbb{R}^n$ and $\rho \in \mathcal{P}_p(X)$ if

$$\int_X d_2^p(x_0,x)\rho(x)dx,\ \text{for some } x_0 \in X, \tag{4.7}$$

where d_2 is the Euclidean distance. Then, the p-Wasserstein distance W_p between $\rho,\rho' \in \mathcal{P}_p(X)$ is given by

$$W_p(\rho,\rho') = \left(\inf_{\gamma\in \mathrm{ADM}(\rho,\rho')}\int_{X\times X} d_2^p(x,x')d\gamma(x,x')\right)^{\frac{1}{p}}, \tag{4.8}$$

where $\mathrm{ADM}(\rho, \rho')$ is the set of joint measures with marginal densities ρ and ρ'. The p-Wasserstein distance is induced by a Finsler metric [1], given by

$$F_\rho(v) = \left(\int_X \|\nabla \Phi_v\|_2^p d\rho \right)^{\frac{1}{p}}, \tag{4.9}$$

where $v \in T_\rho \mathcal{P}_p(X)$ and Φ_v satisfies $v(x) = -\nabla \cdot (\rho(x) \nabla_x \Phi_v(x))$ for any $x \in X$, where $\nabla \cdot$ is the divergence operator. Now, choose $v = -J_\theta \nabla_\theta W_p^2(\rho_\theta, \rho)$. Then, through a cumbersome computation, we compute how the local Hessian acts on two tangent vectors $d\theta_1, d\theta_2 \in T_\theta \Theta$

$$\begin{aligned}
&H_\theta^{\frac{1}{2} W_p^2}(d\theta_1, d\theta_2) \\
=&(2-p) F_{\rho_\theta}^{2(1-p)}(v) \left(\int_X \|\nabla \Phi_v\|_2^{p-2} \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_v \rangle d\rho_\theta \right) \\
&\times \left(\int_X \|\nabla \Phi_v\|_2^{p-2} \langle \nabla \Phi_{d\theta_2}, \nabla \Phi_v \rangle d\rho_\theta \right) \\
&+ F_{\rho_\theta}^{2-p}(v) \int_X \|\nabla \Phi_v\|_2^{p-2} \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_{d\theta_2} \rangle d\rho_\theta \\
&+ (p-2) F_{\rho_\theta}^{2-p}(v) \int_X \|\nabla \Phi_v\|_2^{p-4} \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_v \rangle \langle \nabla \Phi_{d\theta_2}, \nabla \Phi_v \rangle d\rho_\theta,
\end{aligned} \tag{4.10}$$

where $J_\theta d\theta_i = -\nabla \cdot (\rho_\theta \nabla \Phi_{d\theta_i})$ for $i = 1, 2$. The case $p = 2$ is special, as the 2-Wasserstein metric is induced by a Riemannian metric, whose pullback can be recovered by substituting $p = 2$ in (4.10), yielding

$$H_\theta^{\frac{1}{2} W_2^2}(d\theta_1, d\theta_2) = \int_X \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_{d\theta_2} \rangle d\rho_\theta. \tag{4.11}$$

This yields the natural gradient of W_2^2 as introduced in [6, 9].

Acknowledgements. The authors were supported by Centre for Stochastic Geometry and Advanced Bioimaging, and a block stipendium, both funded by a grant from the Villum Foundation. We furthermore wish to thank the anonymous reviewers for their very useful comments.

References

[1] Agueh, M.: Finsler structure in the p-Wasserstein space and gradient flows. Comptes Rendus Mathematique **350**(1-2), 35–40 (2012)

[2] Amari, S.I.: Natural gradient works efficiently in learning. Neural computation **10**(2), 251–276 (1998)

[3] Amari, S.i.: Divergence function, information monotonicity and information geometry. In: Workshop on Information Theoretic Methods in Science and Engineering (WITMSE). Citeseer (2009)

[4] Amari, S.I.: Information geometry and its applications. Springer (2016)

[5] Bercu, G.: Gradient methods on Finsler manifolds. In: Proceedings of the workshop on global analysis, differential geometry and Lie Algebras. pp. 230–233 (2000)

[6] Chen, Y., Li, W.: Natural gradient in Wasserstein statistical manifold. arXiv preprint arXiv:1805.08380 (2018)

[7] Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in neural information processing systems. pp. 2933–2941 (2014)

[8] Khan, M.E., Baqué, P., Fleuret, F., Fua, P.: Kullback-Leibler proximal variational inference. In: Advances in Neural Information Processing Systems. pp. 3402–3410 (2015)

[9] Li, W., Montúfar, G.: Natural gradient via optimal transport. Information Geometry **1**(2), 181–214 (2018)

[10] Martens, J.: Deep learning via Hessian-free optimization. In: ICML. vol. 27, pp. 735–742 (2010)

[11] Martens, J.: New insights and perspectives on the natural gradient method. arXiv preprint arXiv:1412.1193 (2014)

[12] Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization. (1983)

[13] Parikh, N., Boyd, S., et al.: Proximal algorithms. Foundations and Trends in Optimization **1**(3), 127–239 (2014)

[14] Pascanu, R., Bengio, Y.: Revisiting natural gradient for deep networks. arXiv preprint arXiv:1301.3584 (2013)

[15] Raskutti, G., Mukherjee, S.: The information geometry of mirror descent. IEEE Transactions on Information Theory **61**(3), 1451–1457 (2015)

[16] Roux, N.L., Manzagol, P.A., Bengio, Y.: Topmoumoute online natural gradient algorithm. In: Advances in neural information processing systems. pp. 849–856 (2008)

[17] Saad, Y.: Krylov subspace methods for solving large unsymmetric linear systems. Mathematics of computation **37**(155), 105–126 (1981)

[18] Schraudolph, N.N.: Fast curvature matrix-vector products for second-order gradient descent. Neural computation **14**(7), 1723–1738 (2002)

Appendix F

(q,p)-Wasserstein GANs: Comparing Ground Metrics for Wasserstein GANs

The following chapter presents (up to formatting) the preprint

> Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen "(q, p)-Wasserstein GANs: Comparing Ground Metrics for Wasserstein GANs." arXiv preprint arXiv:1902.03642. 2019.

This work considers incorporating constraints, arising from the dual formulation of OT, in the WGAN setting. The vanilla WGAN [9] is based on the 1-Wasserstein metric, in which case the dual constraints translate into enforcing 1-Lipschitzness over the discriminator neural network. In practice, this enforcement forms the main implementational difficulty of WGANs. To allow for more general cost functions in the WGAN setting, we consider utilizing the discrete c-transform for enforcing the constraints. With this novel methodology, we study the effect of choosing the l^q ground metric for $q \geq 1$ in WGANs, and the effect of using different p-Wasserstein metrics over the different ground metrics for $p \geq 1$.

As the metric chosen on the sample space affects the resulting OT task significantly, it is important to be able to have control over it. In computer vision, it is well known that the l^2 metric does not yield an optimal similarity measure between two images. This has inspired work that changes the sample metric used in WGANs [1, 20]. However, the models specified are still very restricted to specific metrics, which is why studying the c-transform is worthwhile, allowing the inclusion of a general sample metric.

(q,p)-Wasserstein GANs: Comparing Ground Metrics for Wasserstein GANs

Anton Mallasto[1] Jes Frellsen[2] Wouter Boomsma[1]
Aasa Feragen[1]

[1]Department of Computer Science, University of Copenhagen
[2]Department of Computer Science, IT University of Copenhagen

Abstract

Generative Adversial Networks (GANs) have made a major impact in computer vision and machine learning as generative models. Wasserstein GANs (WGANs) brought Optimal Transport (OT) theory into GANs, by minimizing the 1-Wasserstein distance between model and data distributions as their objective function. Since then, WGANs have gained considerable interest due to their stability and theoretical framework. We contribute to the WGAN literature by introducing the family of (q, p)-Wasserstein GANs, which allow the use of more general p-Wasserstein metrics for $p \geq 1$ in the GAN learning procedure. While the method is able to incorporate any cost function as the ground metric, we focus on studying the l^q metrics for $q \geq 1$. This is a notable generalization as in the WGAN literature the OT distances are commonly based on the l^2 ground metric. We demonstrate the effect of different p-Wasserstein distances in two toy examples. Furthermore, we show that the ground metric does make a difference, by comparing different (q, p) pairs on the MNIST and CIFAR-10 datasets. Our experiments demonstrate that changing the ground metric and p can notably improve on the common $(q, p) = (2, 1)$ case.

1 Introduction

Generative modelling considers learning models to generate data, such as images, text or audio. Prominent generative models include the *Variational Auto-Encoders* (VAEs) Kingma & Welling (2013) and *Generative Adversial Networks* (GANs) Goodfellow et al. (2014), the latter of which will be studied in this work. The generative models can be trained on unlabelled data, which is a considerable advantage over supervised models, as data labelling is expensive. The usual approach employs the *manifold assumption*, stating that all meaningful data lies on a low-dimensional manifold of the sample space. Based on this assumption, one is then able to learn a map from a low dimensional distribution to the

true data distribution. In this step, it is essential to quantitatively measure the discrepancy between the two distributions. To this end, one chooses a *metric* or a *divergence* between probability distributions. The metric should reflect modelling choices with respect to which properties of the distributions are deemed similar, or what kind of invariances one wants the metric to respect.

Traditionally, probability measures have been compared using non-metric divergence measures from information geometry, e.g. the *Kullback-Leibler (KL) divergence* and *Bregman divergences*. The KL-divergence has deep connections with Bayesian statistics, where likelihood maximization in model selection can be cast as minimizing the KL-divergence.

Recently, a popular family of metrics has been provided by the theory of *Optimal Transport* (OT), which studies probability distributions through a geometric framework. At its heart lie the *Wasserstein metrics*, which extend the underlying metric between sample points to entire distributions. Consequently, the metrics can be used to e.g. derive statistics between populations of probability distributions, allowing the inclusion of stochastic data objects in statistical pipelines Mallasto & Feragen (2017). Recent algorithmic advances Peyré & Cuturi (2017) have made OT widespread in the fields of machine learning and computer vision, where it has been used for e.g. domain adaption Courty et al. (2017), point embeddings Muzellec & Cuturi (2018) and VAEs Tolstikhin et al. (2017).

Quite notably, OT has impacted GANs. The original formulation of Goodfellow et al. (2014) defines GANs through a minimax game of two neural networks. One of the networks acts as a generator, whereas the other network discriminates samples based on whether they originate from the data population or not. The minimax game results in the minimization of the *Jensen-Shannon divergence* between the generated distribution and the data distribution. Arjovsky et al. (2017) then propose to minimize the 1-Wasserstein distance, instead, demonstrating that the new loss function provides stability to the training. This stability was mainly attributed to the Wasserstein metric being well defined even when the two distributions do not share the same support. This results in the *Wasserstein GAN* (WGAN). Other notable OT inspired variations of the original GAN are discussed below.

Notably, Adler & Lunz (2018) consider *Banach Wasserstein GANs* for minimizing the 1-Wasserstein metric when the ground metric is induced by a norm, by imposing a dual norm penalty of the gradient. The method thus allows for any l^q norm, but is restricted to the $p = 1$ case.

1.1 Related Literature

The original WGAN architecture Arjovsky et al. (2017) enforces k-Lipschitz constraints through *weight clipping*. An alternative to clipping the weights is provided in *Spectral Normalization GANs* (SNGANS) Miyato et al. (2018), which impose Lipschitzness through l^2-normalization of the network weights. A body

of work includes the constraints through gradient penalties, first introduced in Gulrajani et al. (2017), where a penalty term for non-unit-norm gradients of the discriminator is added, resulting in the WGAN-GP. *Consistency Term GANs* (CTGANs), on the other hand, penalize exceeding the Lipschitz constraint directly.

The aforementioned work focuses on training the GAN when the 1-Wasserstein metric with the l^2 ground metric forms the objective function. On top of this, a body of work exists exploring the use of other OT inspired metrics and divergences. Below, we discuss some notable examples.

Deshpande et al. (2018) propose using the sliced Wasserstein distance Bonneel et al. (2015), which computes the expectation of the Wasserstein distance between one dimensional projections of the measures. This approach allows omitting learning a discriminator, but in practice a discriminator is trained for choosing meaningful projections, essential when working with high-dimensional data. The authors report increased stability in training and show that the training objective is an upper bound for the true distance between the generator and target distribution.

Genevay et al. (2017), on the other hand, rely on the favorable computational properties of relaxing the original OT problem with entropic penalization. Instead of relying on the dual Rubinstein-Kantorovich formulation, they compute the *Sinkhorn divergence* Cuturi (2013) between minibatches in the primal formulation. This also allows omitting learning a discriminator, however, the authors do propose learning a cost function, as they argue the l^2 ground metric is not suitable in every application. The hyperparameters of the Sinkhorn divergence allows interpolating between the 2-Wasserstein distance and Maximum Mean Discrepancy (MMD), providing more freedom the in the metric model choice. This method also allows for a general cost function to be used, like our $(q,p)-WGAN$ method, but the experiments are limited to the $p = 2$ and learned distance function cases without comparison.

Wu et al. (2018) introduce the *Wasserstein divergence*, motivated by the gradient penalty approach on the 1-Wasserstein metric. The divergence builds on the dual formulation, by relaxing the Lipschitz constraint. Additionally, a gradient norm penalty is included, that is considered over the support of a fixed test distribution.

1.2 Our Contribution

We wish to add more flexibility to WGANs by using the p-Wasserstein distance on top of more general l^q ground metrics for $p, q \geq 1$. This is achieved through the (q,p)-Wasserstein GAN ((q,p)-WGAN), which generalizes Wasserstein GANs to allow arbitrary cost functions for the OT problem, however, we limit the scope of this paper to the l^q metric case. This generalization broadens the existing WGAN literature, as mostly the 1-Wasserstein distance with l^2 metric is considered. We demonstrate the importance of the resulting flexibility in

our experiements. Moreover, $(2, 1)$-WGAN provides a novel way of taking into account the 1-Lipschitz constraints required in the original WGAN minimizing the 1-Wasserstein distance.

Given our (q, p)-WGAN implementation, we study the effect of p when we fix $q = 2$ in two toy examples. Additionally, we compare p-Wasserstein metrics based on the l^q ground metric between samples for $p = 1, 2$ and $q = 1, 2$ on the MNIST and CIFAR-10 datasets. The (q, p)-WGANs are compared to WGAN and WGAN-GP on the CIFAR-10 dataset to assess the performance of our implementation. The experiments show, that choosing $q = 1$ outperforms $q = 2$ on colored image data, where as $p = 2$ slightly outperforms $p = 1$. Based on the results, it is clear that the metric used for GANs should be tailored to fit the needs of the application.

Finally, the OT theory suggests that the Kantorovich potentials (or discriminators) can also function as generators through their gradients. We try this on the MNIST dataset, and conclude that the generator clearly improves the results.

2 Background

We briefly summarize the prequisites for this work. The methodology is founded on optimal transport, which we will revise first. We finish the section by reviewing the mathematical details of GANs with a focus on WGANs.

2.1 Optimal Transport

The aim in *Optimal Transport* (OT) is to define a geometric framework for the study of probability measures. This is carried out by defining a *cost function* between samples (e.g. the l^2 metric), and then studying *transport plans* that relate two compared probability measures to each other while minimizing the total cost. A common example states the problem as moving a pile of dirt into another with minimal effort, by finding an optimal allocation for each grain of dirt so that the cumulative distance of dirt moved is minimized.

We start with basic definitions, and conclude by discussing the *Wasserstein metric*. The interested reader may refer to Villani (2008) for theoretical and Peyré & Cuturi (2017) for computational aspects of OT.

Optimal Transport Problem. Let μ be a probability measure on a metric space X, denoted by $\mu \in \mathcal{M}(X)$. Let $f \colon X \to Y$ be a measurable map. Then $f_{\#}\mu(A) := \mu(f^{-1}(A))$ denotes the push-forward of μ with respect to f. Here A is any measurable set in another metric space Y. The push-forwad can be also explained from a sampling perspective; assume ξ is a random variable with distribution μ. Then $f(\xi)$ has distribution $f_{\#}\mu$.

Given two probability measures $\mu \in \mathcal{M}(X)$, and $\nu \in M(Y)$, we define the set of *admissable plans* by

$$\begin{aligned} &\mathrm{ADM}(\mu,\nu) \\ =&\{\gamma \in \mathcal{M}(X \times Y) |\ (\pi_1)_{\#}\gamma = \mu,\ (\pi_2)_{\#}\gamma = \nu\}, \end{aligned} \tag{1}$$

where π_i denotes the projection onto the ith coordinate. In layman's terms, a joint measure on $X \times Y$ is admissable, if its marginals are μ and ν.

Now, given a lower semi-continuous *cost function* $c : X \times Y \to \mathbb{R}$ (such as the l^q metric d_q), the task in optimal transport is to compute

$$\mathrm{OT}_c(\mu,\nu) := \min_{\gamma \in \mathrm{ADM}(\mu,\nu)} \mathbb{E}_\gamma[c], \tag{2}$$

where we use $\mathbb{E}_\mu[f]$ to denote the expectation of a function f under the measure μ, that is,

$$\mathbb{E}_\mu[f] = \int_X f(x) d\mu(x). \tag{3}$$

Next, denote by $L^1(\mu) = \{f |\ \mathbb{E}_\mu[f] < \infty\}$ the set of functions that have finite expectations with respect to μ. Let $\varphi \in L^1(\mu)$, $\psi \in L^1(\nu)$. Then, assume φ, ψ satisfy

$$\varphi(x) + \psi(y) \leq c(x,y),\ \forall (x,y) \in X \times Y. \tag{4}$$

We denote the set of all such pairs by $\mathrm{ADM}(c)$. Then, $\mathrm{OT}_c(\mu,\nu)$ can be expressed in the *dual formulation*

$$\mathrm{OT}_c(\mu,\nu) = \max_{(\varphi,\psi) \in \mathrm{ADM}(c)} \left\{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\psi]\right\}. \tag{5}$$

The optimal functions φ, ψ are called *Kantorovich potentials*, and they satisfy

$$\varphi(x) + \psi(y) = c(x,y),\ \forall (x,y) \in \mathrm{Supp}(\gamma). \tag{6}$$

The Kantorovich potentials φ and ψ are intimately related. Define the c-transform of φ as

$$\varphi^c : Y \to \mathbb{R},\ y \mapsto \inf_{x \in X} \left\{c(x,y) - \varphi(x)\right\}, \tag{7}$$

then according to the fundamental theorem of optimal transport, the Kantorovich potentials satisfy $\psi = \varphi^c$, and thus (5) can be written as

$$\mathrm{OT}_c(\mu,\nu) = \max_{(\varphi,\varphi^c) \in \mathrm{ADM}(c)} \left\{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\varphi^c]\right\}, \tag{8}$$

reducing the optimization to be carried out over a single function.

Wasserstein Metric. It turns out that the OT framework can be used to define a distance between probability distributions. Define the set

$$\mathcal{P}_d^p(X) = \left\{\mu \in \mathcal{M}(X) \,\middle|\, \int d^p(x_0,x) d\mu(x) < \infty\right\} \tag{9}$$

for any $x_0 \in X$. Then, $\mathrm{OT}_c(\mu, \nu)$ defines a metric between $\mu, \nu \in \mathcal{P}_d^p(X)$, if we choose the cost c to be related to a metric d on X, called the *ground metric*, in the following way.

The p-Wasserstein metric W_p between $\mu, \nu \in \mathcal{P}_d^p(X)$, where (X, d) is a metric space, is given by

$$W_p(\mu, \nu) := \left(\mathrm{OT}_{d^p/p}(\mu, \nu)\right)^{\frac{1}{p}}. \tag{10}$$

When μ, ν are absolutely continuous measures on $X = Y = \mathbb{R}^n$ with the Euclidean l^2 metric and $p > 1$, the optimal transport plan is induced by a unique *transport map* $T \colon X \mapsto X$, for which $T_{\#}\mu = \nu$, given by

$$T = (I - \|\nabla\varphi\|^{p'-2}\nabla\varphi), \tag{11}$$

where φ stands for the optimal Kantorovich potential in the dual formulation (5), and $p^{-1} + (p')^{-1} = 1$. Therefore, computing the p-Wasserstein distance by the dual formulation yields us a map between the distributions, which we will later employ in the experimental section.

The Ground Metric. When $X = Y = \mathbb{R}^d$, commonly the l^2 metric is chosen as the ground metric for the p-Wasserstein distance. However, depending on the application, any other distance can be also considered, for example any l^q distance d_q for $q \geq 1$, given by

$$d_q(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^q\right)^{\frac{1}{q}}. \tag{12}$$

In the experimental section, we study the effect of the ground metric, when minimizing the p-Wasserstein distance in the context of GANs. To emphasize the ground metric, we introduce the (q, p)-Wasserstein distance notation

$$W_{q,p}(\mu, \nu) = \left(\mathrm{OT}_{d_q^p/p}(\mu, \nu)\right)^{\frac{1}{p}}. \tag{13}$$

To not diverge too far from the standard notation, we assume that $q = 2$ for the p-Wasserstein distance denoted by W_p.

2.2 Generative Adversial Networks

Generative Adversial Networks (GANs) are a popular tool for learning data distributions Goodfellow et al. (2014). The GAN approach consists of a competitive game between two networks, the *generator* g_ω and the *discriminator* $\varphi_{\omega'}$, with parameters ω and ω', respectively. Given the target distribution μ_t of the data, and a low-dimensional *source distribution* μ_s, the GAN minimax objective is given by

$$\begin{aligned} \min_{\omega} \max_{\omega'} \{&\mathbb{E}_{x\sim\mu_t}\left[\log(\varphi_{\omega'}(x))\right] \\ &+ \mathbb{E}_{z\sim\mu_s}\left[\log(1 - \varphi_{\omega'}(g_\omega(z)))\right]\}. \end{aligned} \tag{14}$$

At optimality, this corresponds to minimizing the Jensen-Shannon divergence between μ_t and $(g_\omega)_\# \mu_s$, the push-forward of the source with respect to the generator. The discriminator has range $[0, 1]$, expressing the probability of a sample being from the original data distribution.

The *Wasserstein GAN* introduced by Arjovsky et al. (2017) minimizes the 1-Wasserstein metric instead. The authors argue that divergences such as Jensen-Shannon, or Kullback-Leibler do not behave well with respect to the generator's parameters. This is due to these divergences being singular when the two distributions do not share the same support. They then demonstrate, that the 1-Wasserstein distance behaves in a more continuous way, and provides a meaningful loss, whose decrease corresponds to increased image quality when generating images.

Writing the 1-Wasserstein metric in the dual form, and using that for the $(q, p) = (2, 1)$ case $\varphi^c_{\omega'} = -\varphi_{\omega'}$, and $(\varphi_{\omega'}, \varphi^c_{\omega'}) \in \mathrm{ADM}(c)$ implies that $\varphi_{\omega'}$ is 1-Lipschitz, the minimax objective for WGANs is written as

$$\min_{\omega} \max_{\omega'} \left\{ \mathbb{E}_{x \sim \mu_t} \left[\varphi_{\omega'}(x) \right] - \mathbb{E}_{z \sim \mu_s} \left[\varphi_{\omega'}(g_\omega(z)) \right] \right\}. \tag{15}$$

This time $\varphi_{\omega'}$ is called the *critic* and not the discriminator, as its range is not limited. However, in this paper, we use either name interchangeably, or might also use the name Kantorovich potential.

In the original paper Arjovsky et al. (2017), the Lipschitz constraints are enforced through weight-clipping. This, however, only quarantees k-Lipschitzness for some k, and thus a scalar multiple of the 1-Wasserstein distance is computed. Remarking that a function is 1-Lipschitz if and only if its gradient has norm at most 1 everywhere, a gradient norm penalty was introduced in the WGAN-GP method of Gulrajani et al. (2017). See Subsec. 1.1 for more discussion on imposing the constraints.

3 (q, p)-Wasserstein GAN

Algorithm 1. *(q, p)-WGAN. Batch size $m = 64$, learning rate $\alpha = 10^{-4}$, search space B, and the Adam parameters $\beta_0 = 0.5$ and $\beta_1 = 0.999$.*

for iter $= 1, ..., N_{\text{Iterations}}$ ***do***
  Sample from target $x_i \sim \mu_t$ and source $z_i \sim \mu_s$, $i = 1, 2, ..., m$, where m is the batch-size. Denote $B_x = \{x_i\}_{i=1}^m$.
  $y_i \leftarrow g_\omega(z_i)$, *denote* $B_y = \{y_i\}_{i=1}^m$.
  for $t = 1, ..., N_{\text{critic}}$ ***do***
    #Define $\psi_{\omega'}$:
    $\psi_{\omega'}(y) \leftarrow \min_{x \in B} \left\{ \frac{1}{p} d_q^p(y, x) - \varphi_{\omega'}(x) \right\}$.
    #Compute penalties:
    $P_1 = \frac{1}{m^2} \sum_{i,j=1}^m \xi(x_i, y_j)^2$
    $P_2 = \frac{1}{4m^2} \sum_{x,y \in B_x \cup B_y} \xi(x, y)^2$

#Compute objective:
$L \leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(\varphi_{\omega'}(x_i) + \psi_{\omega'}(y_i)\right) - P_1 - P_2.$
#Update critic:
$\omega' \leftarrow \omega' + \mathrm{Adam}(\nabla_{\omega'} L, \alpha, \beta_0, \beta_1).$
end for
#Compute Wasserstein loss:
$\leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(\varphi_{\omega'}(x_i) + \psi_{\omega'}(y_i)\right)$
#Update generator:
$\omega \leftarrow \omega - \mathrm{Adam}(\nabla_{\omega} W, \alpha, \beta_0, \beta_1).$
end for

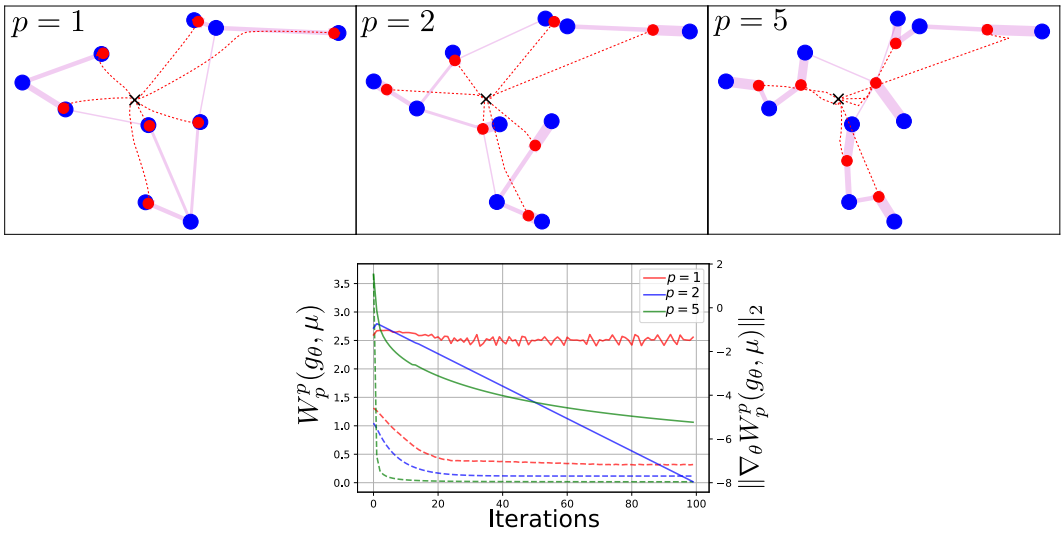


Figure 1: Comparing the p-Wasserstein metrics for $p = 1, 2, 5$ with the Euclidean metric. We minimize W_p^p between a discrete target distribution μ with 10 atoms (blue), by optimizing the support of a model distribution g_θ with 7 atoms (in red). The trail of the changing support is drawn in dashed red, starting from the origo (cross). The magenta lines express the optimal mass transport between the two measures after optimization. Both measures have uniform weights. Plot on the right shows convergence for each case in distance W_p^p (dashed) and gradient norm $\|\nabla_\theta W_p^p(g_\theta, \mu)\|_2$ (solid).

We will now introduce the novel (q, p)-Wasserstein GAN ((q, p)-WGAN) architecture, which minimizes the (q, p)-Wasserstein distance between the target distribution μ_t and the approximation $(g_\omega)_{\#}\mu_s$. That is, the cost function is given by $c = d_q^p/p$. The objective reads

$$\begin{aligned} &\min_{\omega} W_{q,p}^p((g_\omega)_{\#}\mu_s, \mu_t) \\ =&\min_{\omega} \max_{(\varphi_{\omega'}, \varphi_{\omega'}^c)\in \mathrm{ADM}(c)} \{\mathbb{E}_{x\sim\mu_t}[\varphi_{\omega'}(x)] \\ &+ \mathbb{E}_{z\sim\mu_s}[\varphi_{\omega'}^c(g_\omega(z))]\}\,. \end{aligned} \tag{16}$$

This formulation requires one to approximate the c-transform defined in (7) and to enforce the constraint $(\varphi_{\omega'}, \varphi_{\omega'}^c) \in \mathrm{ADM}(c)$.

The c-transform. For computing the c-transform, we choose a *search space* B for the minimization. For example, the learning procedure of the GAN is carried out through mini-batches. Hence, we can compute the discrete c-transform over the mini-batches. That is, given sets of samples $B_x = \{x_i\}_{i=1}^m$ and $B_y = \{y_i\}_{i=1}^m$, from the target μ_t and generator $(g_\omega)_\# \mu_s$, respectively, we compute the approximation $\varphi^c_{\omega'}$ over $B = B_x \cup B_y$

$$\varphi^c_{\omega'}(y_j) \approx \min_{x \in B} \{c(x, y_j) - \varphi_{\omega'}(x)\}. \tag{17}$$

In the experiments, we use both $B = B_x$ and $B = B_x \cup B_y$.

Enforcing the constraints. Define

$$\xi(x, y) = c(x, y) - \varphi_{\omega'}(x) - \varphi^c_{\omega'}(y). \tag{18}$$

Then, when training the discriminator, we add two penalty terms given by

$$\begin{aligned} P_1(\varphi) &= \lambda_1 \sum_{i,j=1}^{m} \xi(x_i, y_j)^2, \\ P_2(\varphi) &= \lambda_2 \sum_{x,y \in B_x \cup B_y} \min(\xi(x, y), 0)^2. \end{aligned} \tag{19}$$

Here P_2 enforces $(\varphi, \varphi^c) \in \mathrm{ADM}(c)$ over all elements in $B_x \cup B_y$, and P_1 encourages pairs (x_i, y_j) to belong in the support of the optimal plan.

The (q, p)-WGAN method is summarized in Algorithm 1.

4 Comparison of p-Wasserstein Metrics

To give some intuition about the differences between different p-Wasserstein metrics W_p, we compare the behavior of W_p for $p = 1, 2, 5$ on two toy examples. The first example consists of approximating a discrete probability measure with another discrete measure with smaller support. This example is intended to give general intuition of the behavior of the p-Wasserstein distance when compromises are required, however, the intuition might not translate directly into the GAN setting. The second example demonstrates fitting a $(2, p)$-WGAN to a 2-dimensional Gaussian mixture. We abbreviate $(2, p)$-WGAN as p-WGAN.

In the first example, the target distribution μ has 10 atoms with uniform weights. We approximate the target with a model distribution ν with 7 atoms and uniform weights. This objective is closely related to k-means clustering Canas & Rosasco (2012); Pollard (1982). In fact, the objective would be equivalent to k-means, if each model distribution atom was assigned the mass of the corresponding cluster of target distribution atoms.

In Fig. 1, it is clearly seen that in the $p = 1$ case, the model distribution prefers to have a support that overlaps with the target. When $p = 2$, the model prefers cluster means as its support, and thus samples from the model are not exactly

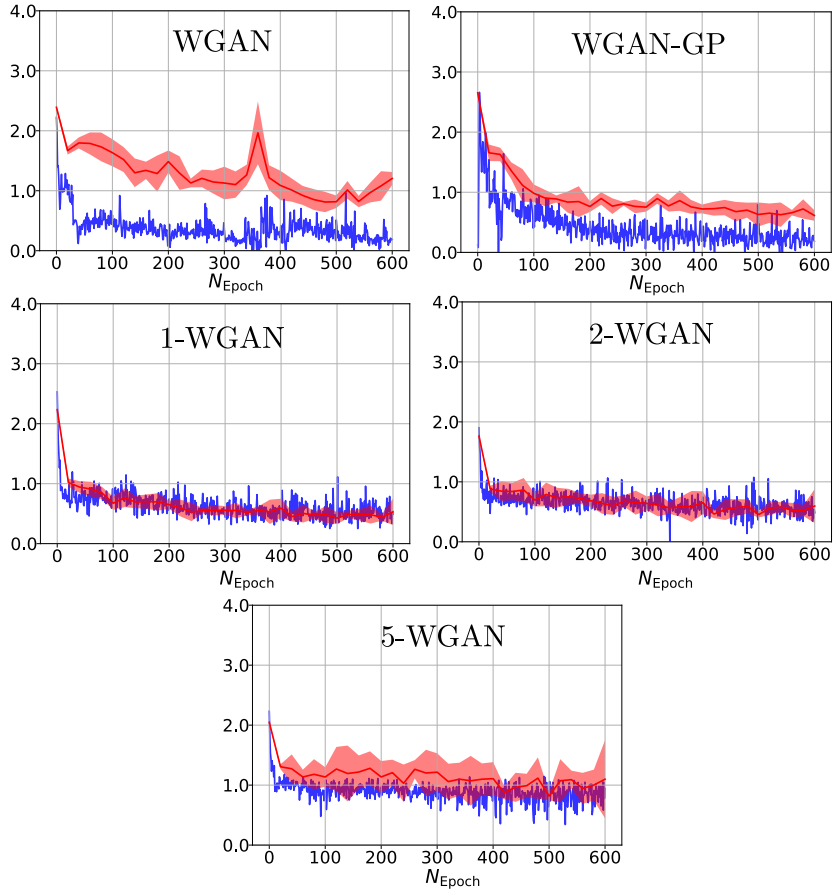


Figure 2: Convergence of model distributions for the Gaussian mixture model. Objective function (approximation of the p-Wasserstein distance) after each epoch (for WGAN, this has been renormalized with the estimated Lipschitz constant) in blue. The true p-Wasserstein distance computed between the data set and the same amount of generator samples in red.

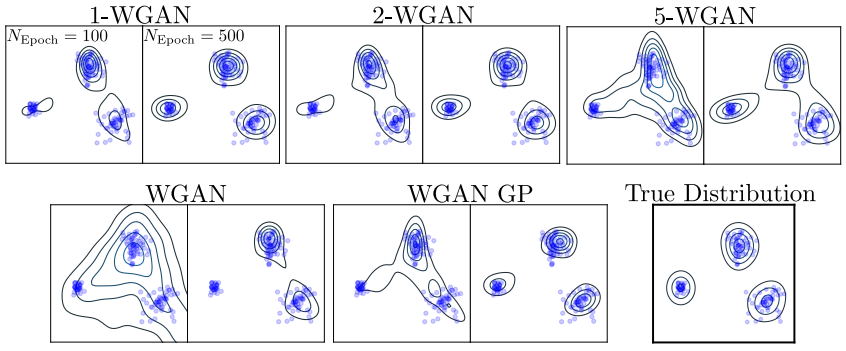


Figure 3: Approximating a Gaussian mixture distribution (samples in blue) with different WGAN architectures. Presented are the results after 100 and 500 epochs for $(2, p)$-WGAN, abbreviated p-WGAN, for $p = 1, 2, 5$. Furthermore, we present the results for the original WGAN and WGAN-GP.

the same as the real samples of the target. Looking at the $p = 5$ case, it seems that the model starts shrinking to the interior of the convex hull of the target's support, reducing the variance of the model distribution. Higher p-value seems to imply faster and more stable optimization, however, we do not witness this in the second example below (Fig. 2).

In the second example, we model a Gaussian mixture model with three clusters (cluster sizes 60, 30 and 50) using a GAN that minimizes W_p^p. The critic and generator are Multi-Layer Perceptrons (MLPs) with ReLU activations (the output is without activation) and two fully connected hidden layers of size 128. In addition to comparing the p-Wasserstein distances for $p = 1, 2, 5$, we also compare the results to WGAN and WGAN-GP architectures.

In Fig. 3, the learned distributions are visualized after 100 and 500 epochs under the original dataset. When comparing to the true distribution, 1-WGAN, 2-WGAN and WGAN-GP seem to converge the fastest and provide qualitatively the best results. 5-WGAN seems to fail separating the clusters from each other, whereas WGAN expresses mode collapse.

The convergence of each model is demonstrated in Fig. 2, where the objective function value and p-Wasserstein distance between the original dataset and the same amount of generator samples are visualized. For 1-WGAN and 2-WGAN, the objective function approximates well the real p-Wasserstein distance, whereas 5-WGAN is more unstable. Note that in the WGAN case, the Lipschitz constant is estimated to normalize the objective function for an approximation of the 1-Wasserstein distance. WGAN convergence is clearly more unstable than the others.

Conclusion. From the toy examples it is obvious, that different p values result in differently behaving optimization problems. If the model is given extreme freedom (but still limited expressive power), as in the first example on discrete probability measures, higher p-values result in stabler optimization, but also reduces the variance. On the other hand, $p = 1$ overfits by trying to overlap with the target distribution. However, this does not directly translate to the GAN example, which might be because of the model being expressive enough to match the data distribution well. Nevertheless, this example demonstrates that the (q, p)-WGAN models the objective Wasserstein distance well.

5 Experiments

We evaluate the performance of the (q, p)-WGAN method on two different datasets; MNIST LeCun et al. (1998) and CIFAR-10 Krizhevsky & Hinton (2009). The effect of ground metric is explored on the MNIST dataset by visually assessing the image quality. We quantify the performance of different (q, p)-WGANs by computing the *Inception Score* (IS) Salimans et al. (2016) and the *Fréchet Inception Distance* (FID) Heusel et al. (2017) on the CIFAR-10 dataset. We use the DCGAN architecture from Radford et al. (2015) for CIFAR-10, and

Model	IS	FID
(1,1)-WGAN	$\mathbf{4.18 \pm 0.08}$	80.7
(1,2-WGAN)	4.11 ± 0.11	**78.7**
(2,1)-WGAN	3.79 ± 0.09	100.0
(2,2)-WGAN	4.09 ± 0.13	82.5
WGAN, $N = 1$	2.87 ± 0.07	152.9
WGAN, $N = 5$	2.33 ± 0.05	164.7
WGAN-GP, $N = 1$	3.65 ± 0.09	117.6
WGAN-GP, $N = 5$	2.85 ± 0.07	162.6

Table 1: The *Inception Score* (IS) and *Fréchet Inception Distance* (FID) for the CIFAR-10 dataset reported for four different (q, p)-WGANs, the original WGAN, and WGAN-GP. Here N implies discriminator iterations per generator iteration. The models are trained for 50K discriminator iterations.

.

Multi-Layer Perceptrons (MLP) for MNIST, which are trained for 50K generator iterations. We use $m = 64$ as the batch-size for every experiment and $N_{\text{critic}} = 1$ for (q, p)-WGANs.

5.1 Effect of Ground Metric on MNIST.

The MNIST dataset consists of 28×28 greyscale images of hand-written digits, grouped into training and validation sets of sizes 60k and 10k, respectively. We train five different (q, p)-WGAN models, listed in Fig. 4, on the training set. We also show the distribution of distances of generated samples to closest training samples for each model, to quantify whether we are creating new digits or just memorizing the ones from the training set. Based on the first toy-example in Fig. 1, the hypothesis is that 1-Wasserstein GAN tends to overfit to the data compared to a higher p value. However, this is not evident in Fig. 4.

The neural networks used are simple MLPs with 3 hidden layers (specifics in the supplementary material), that are trained for 50K generator and discriminator iterations. For the discrete c-transform, the search space for the minimizer is restricted to B_x and $\lambda_1 = \lambda_2 = 0$, as otherwise the model tended to collapse to single point, and $\alpha = 10^{-4}$ was used as the learning rate.

The ground metric clearly affects the sharpness of produced images. When $q = 2$, the generated digits have quite blurry edges. On the other hand, when $q = 1$, the digits are sharp, but also more degenerate samples are produced. The sharpness can be adjusted, as shown by the samples generated by the $(1.2, 1.2)$-WGAN.

5.2 Assessing the Quality on CIFAR-10.

The CIFAR-10 dataset consists of 50K 32×32 color images for training. We train four different (q, p)-WGANs, the original WGAN, and WGAN-GP. The methods are compared by computing the IS and FID after 50K discriminator iterations. As the original WGAN and WGAN-GP propose to use 5 critic iterations per generator iteration, for fair comparison we carry out the training with $N_{\text{critic}} = 1, 5$.

This time, we use the DCGAN architecture for the generator and discriminator, see supplementary for details. We use the hyperparameters proposed for WGAN and WGAN-GP by the original papers, except for the different critic iteration amounts. For (q, p)-WGANs, $\alpha = 10^{-4}$ is the learning rate, $\lambda_1 = 0.1$ and $\lambda_2 = 10$, and we use $B_x \cup B_y$ as the c-transform search space. Restricting the search space to B_x produced very blurry images.

The scores are presented in Table 1, and example samples in Fig. 5. Based on Table 1, the (q, p)-WGANs outperform WGAN and WGAN-GP. The IS and FID scores are notably higher when $q = 1$. In the $q = 2$ case, the $(2, 2)$-Wasserstein metric scores better than the $(2, 1)$-Wasserstein metric, but in the $q = 1$ case the difference is marginal.

5.3 Kantorovich Potentials as Generators

As pointed out earlier, the learned Kantorovich potentials can also be used as generators by computing the optimal transport map using (11). To see if this is applicable in practice, we train the Kantorovich potentials for the $(2, 2)$-WGAN for 100K iterations on MNIST. Although the samples clearly look like digits, we conclude that the quality of the samples in Fig. 6 is much worse than with a generator.

6 Conclusion

We introduced the (q, p)-WGAN to allow the study of different p-Wasserstein metrics and l^q ground metrics in the GAN setting. We show that these parameters do have a definite effect on GAN training; 1-Wasserstein metric encourages models to overfit, whereas too high p causes too low variance in the model. The FID scores from the CIFAR-10 dataset indicate that $p = 2$ performs better compared to $p = 1$. We also demonstrate that the l^1 metric outperforms l^2 when learning the distribution of colored images of the CIFAR-10 dataset. Moreover, the experiments show that our implementation is competitive with the literature, outperforming the WGAN and WGAN-GP implementations.

The (q, p)-WGAN incorporates the ADM(c) constraints directly on the neural network modelling the Kantorovich potential φ. The other WGAN implementations, on the other hand, seem to focus on enforcing Lipschitzness and using

the knowledge $\psi = \varphi^c$, which are implications of the ADM(d_2) constraints. Working with the general constraint allows for more flexibility in the modelling choices, resulting in improved performance, as we demonstrated. However, our implementation of taking the constraints into account leaves room for improvement, as we had to use considerably different hyperparameters on MNIST and CIFAR-10 to achieve stable training. We hope that our results on the importance of the ground metric and the p parameter inspire research into more efficient implementations to incorporate general cost functions.

Although the generative properties of the Kantorovich potentials did not perform well in our experiment, this might be implementation dependant. We learned the Kantorovich potential field, but in some applications, learning the gradient field directly can be more fruitful Chmiela et al. (2017).

Finally, from the theoretical perspective, choosing $p = 2$ and a Riemannian ground metric d results in a Riemannian structure over the manifold of probability measures, shown by Otto (2001). Thus Riemannian geometry can be used to study the probability distributions. When $p \neq 2$, a Finslerian structure is induced instead Agueh (2012). In layman's terms, Riemannian structure allows the study of lengths and comparison of directions through local inner-products, whereas Finslerian structures provide only direction dependant length-structures. Thus the Riemannian structure results in a more powerful framework for studying the geometry of probability distributions, and possibly GANs.

Acknowledgements

AM and AF were supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation.

References

Adler, J. and Lunz, S. Banach wasserstein gan. In *Advances in Neural Information Processing Systems*, pp. 6755–6764, 2018.

Agueh, M. Finsler structure in the p-Wasserstein space and gradient flows. *Comptes Rendus Mathematique*, 350(1-2):35–40, 2012.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1): 22–45, 2015.

Canas, G. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pp. 2492–2500, 2012.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Deshpande, I., Zhang, Z., and Schwing, A. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3483–3491, 2018.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mallasto, A. and Feragen, A. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 5660–5670, 2017.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Muzellec, B. and Cuturi, M. Generalizing point embeddings using the Wasserstein space of elliptical distributions. *arXiv preprint arXiv:1805.07594*, 2018.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Journal Communications in Partial Differential Equations*, 26: 101–174, 2001.

Peyré, G. and Cuturi, M. Computational optimal transport. Technical report, 2017.

Pollard, D. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205, 1982.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wu, J., Huang, Z., Thoma, J., Acharya, D., and Van Gool, L. Wasserstein divergence for gans. In *Computer Vision – ECCV 2018*, pp. 673–688, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01228-1.

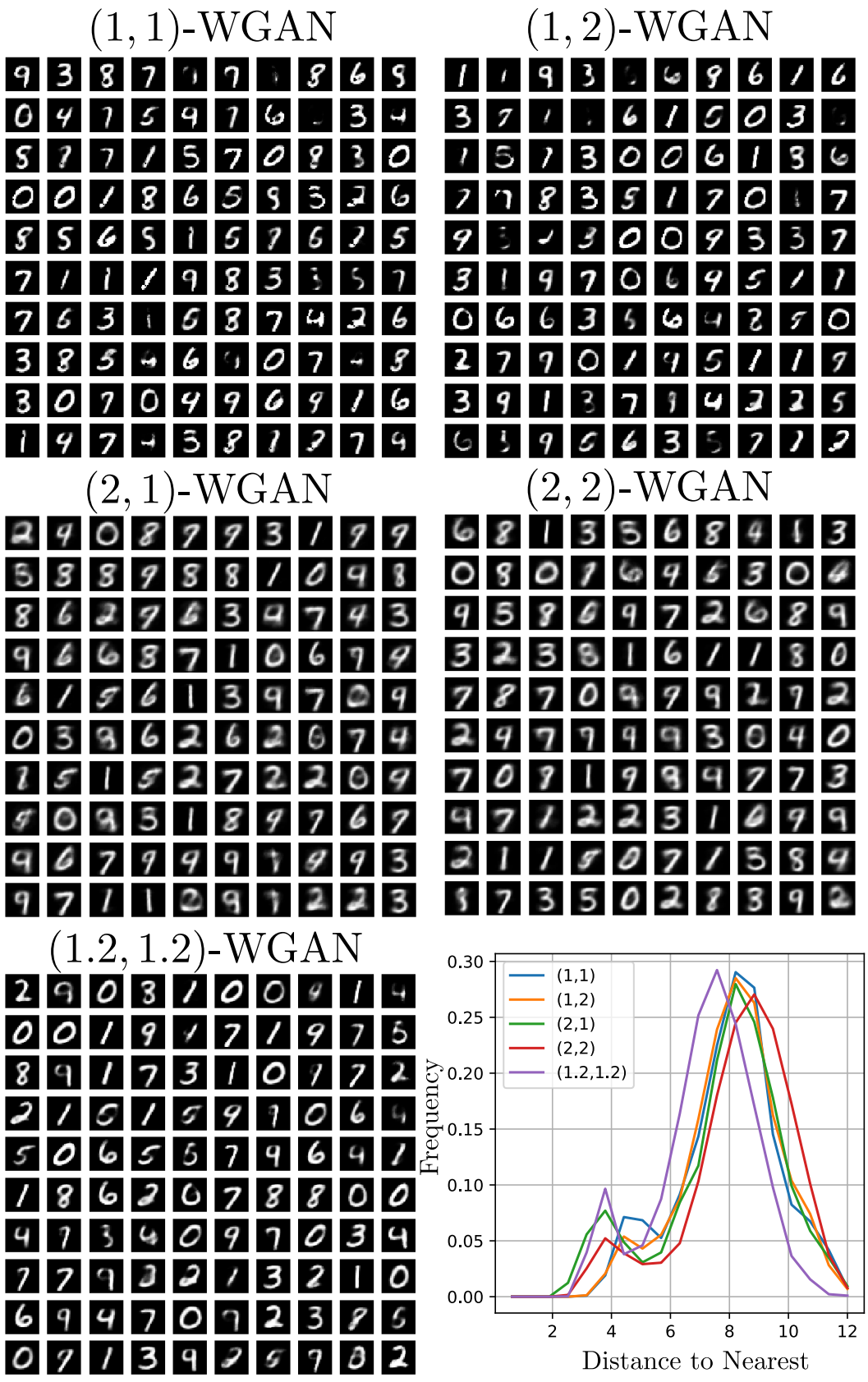


Figure 4: Generated samples from different (q, p)-WGANS trained on the MNIST training set. Furthermore, plotted is the distribution of l^2 distances to closest training points of 5000 generated samples from each model.

Figure 5: Generated samples from different (q, p)-WGANS, the original WGAN and WGAN-GP trained on the CIFAR-10 training set. Here N refers to the amount of critic iterations, for (q, p)-WGANS, this is 1. The IS and FID scores are reported in Table 1.

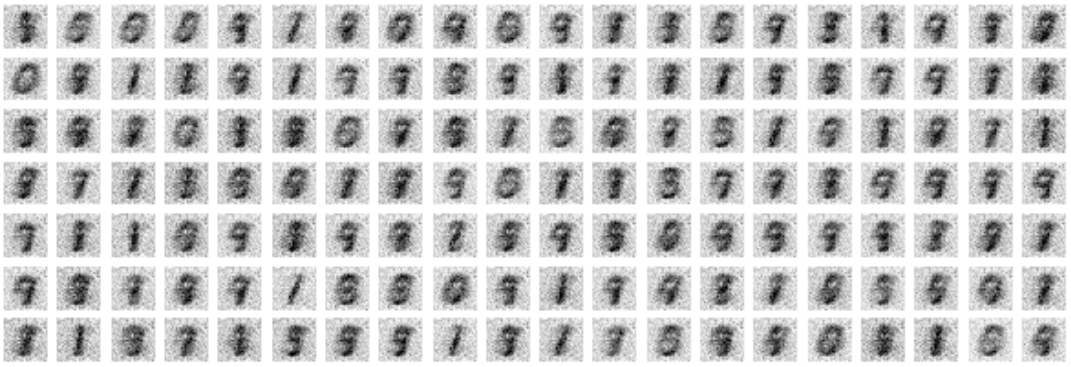


Figure 6: Samples generated by *only* the discriminator on the MNIST dataset.

Appendix G

How Well Do WGANs Estimate the Wasserstein Metric?

The following chapter presents (up to formatting) the preprint

> Anton Mallasto, Guido Montúfar, and Augusto Gerolin. "How Well Do WGANs Estimate the Wasserstein Metric?" arXiv preprint arXiv:1910.03875 2019.

Different heuristics have been introduced in WGANs to incorporate the constraints of the dual formulation of optimal transport induced on the discriminator neural networks. For example, in the vanilla WGAN [9], this was carried out through clipping the weights of the discriminator network, whereas the gradient penalty WGAN (WGAN-GP) [32] penalizes the deviation of the gradient norm from 1, and the (q, p)-WGANs presented in Appendix F through the c-transform.

This work studies how well these different heuristics approximate and estimate the 1-Wasserstein distance. In addition, we consider computing the entropic relaxation of the 1-Wasserstein distance [18], instead, which is achieved by utilizing the (c, ε)-transform for $\varepsilon > 0$, determining the strength of the relaxation. An important question relevant to the results of this work, is that how exactly do we actually want to compute the optimal transport quantity in GANs? The results indicate that the method best at approximation does not produce the best quality images.

How Well Do WGANs Estimate the Wasserstein Metric?

Anton Mallasto[1] Guido Montúfar[2,3,4] Augusto Gerolin[5]

[1]Department of Computer Science, University of Copenhagen
[2]Department of Mathematics, University of California, Los Angeles
[3]Department of Statistics, University of California, Los Angeles
[4]Max Planck Institute for Mathematics in the Sciences
[5]Department of Theoretical Chemistry, Vrije Universiteit Amsterdam

Abstract

Generative modelling is often cast as minimizing a similarity measure between a data distribution and a model distribution. Recently, a popular choice for the similarity measure has been the Wasserstein metric, which can be expressed in the Kantorovich duality formulation as the optimum difference of the expected values of a potential function under the real data distribution and the model hypothesis. In practice, the potential is approximated with a neural network and is called the discriminator. Duality constraints on the function class of the discriminator are enforced approximately, and the expectations are estimated from samples. This gives at least three sources of errors: the approximated discriminator and constraints, the estimation of the expectation value, and the optimization required to find the optimal potential. In this work, we study how well the methods, that are used in generative adversarial networks to approximate the Wasserstein metric, perform. We consider, in particular, the c-transform formulation, which eliminates the need to enforce the constraints explicitly. We demonstrate that the c-transform allows for a more accurate estimation of the true Wasserstein metric from samples, but surprisingly, does not perform the best in the generative setting.

1 Introduction

Recently, optimal transport (OT) has become increasingly prevalent in computer vision and machine learning, as it allows for robust comparison of structured data that can be cast as probability measures, e.g., images, point clouds and empirical distributions (Lai and Zhao, 2014; Rubner et al., 2000), or more general measures (Gangbo et al., 2019). Key properties of OT include its non-singular behavior, when comparing measures with disjoint supports, and the fact that OT inspired objectives can be seen as lifting similarity measures between samples to similarity measures between probability measures. This is in stark

contrast to the more traditional information theoretical divergences, which rely on only comparing the difference in mass assignment. Additionally, when the *cost function* c is related to a distance function d by $c(x, y) = d^p(x, y)$, $p \geq 1$, the OT formulation defines the so called *Wasserstein metric*, which is a distance on the space of probability measures, i.e. a symmetric and positive definite function that satisfies the triangle inequality. Despite its success, scaling OT to big data applications has not been without challenges, since it suffers from the curse of dimensionality (Dudley, 1969; Weed and Bach, 2017). However, significant computational advancements have been made recently, for which a summary is given by Peyré et al. (2019). Notably, the *entropic regularization* of OT introduced by Cuturi (2013) preserves the geometrical structure endowed by the Wasserstein spaces and provides an efficient way to approximate optimal transport between measures.

Generative modelling, where *generators* are trained for sampling from given data distributions, is a popular application of OT. In this field, *generative adversarial networks* (GANs) by Goodfellow et al. (2014) have attracted substantial interest, particularly due to their success in generating photo-realistic images (Karras et al., 2017). The original GAN formulation minimizes the *Jensen-Shannon divergence* between a model distribution and the data distribution, which suffers from unstable training. *Wasserstein GANs* (WGANs) (Arjovsky et al., 2017) minimize the 1-Wasserstein distance over the l^2-metric, instead, resulting in more robust training.

The main challenge in WGANs is estimating the Wasserstein metric, consisting of estimating expected values of the *discriminator* from samples (drawn from a model distribution and a given data distribution), and optimizing the discriminator to maximize an expression of these expected values. The discriminators are functions from the sample space to the real line, that have different interpretations in different GAN variations. The main technical issue is that the discriminators have to satisfy specific conditions, such as being 1-Lipschitz in the 1-Wasserstein case. In the original paper, this was enforced by clipping the weights of the discriminator to lie inside some small box, which, however, proved to be inefficient. The *Gradient penalty WGAN* (WGAN-GP) (Gulrajani et al., 2017) was more successful at this, by enforcing the constraint through a gradient norm penalization. Another notable improvement was given by the *consistency term WGAN* (CT-WGAN) (Wei et al., 2018), which penalizes diverging from 1-Lipschitzness directly. Other derivative work of the WGAN include different OT inspired similarity measures between distributions, such as the *sliced Wasserstein distance* (Deshpande et al., 2018), the *Sinkhorn divergence* (Genevay et al., 2018) and the *Wasserstein divergence* (Wu et al., 2018). Another line of work studies how to incorporate more general ground cost functions than the l^2-metric (Adler and Lunz, 2018; Dukler et al., 2019; Mallasto et al., 2019).

Recent works have studied the convergence of estimates of the Wasserstein distance between two probability distributions, both in the case of continuous (Klein et al., 2017) and finite (Sommerfeld, 2017; Sommerfeld and Munk, 2018) sample spaces. The decay rate of the approximation error of estimating the true

distance with minibatches of size N is of order $O(N^{-1/d})$ for the Wasserstein distances, where d is the dimension of the sample space (Weed and Bach, 2017). Entropic regularized optimal transport has more favorable sample complexity of order $O(1/\sqrt{N})$ for suitable choices of regularization strength (see Genevay et al. 2019 and also Feydy et al. 2018; Mena and Weed 2019). For this reason, entropic relaxation of the 1-Wasserstein distance is also considered in this work.

Contribution. In this work, we study the efficiency and stability of computing the Wasserstein metric through its dual formulation under different schemes presented in the WGAN literature. We present a detailed discussion on how the different approaches arise and differ from each other qualitatively, and finally measure the differences in performance quantitatively. This is done by quantifying how much the estimates differ from accurately computed ground truth values between subsets of commonly used datasets. Finally, we measure how well the distance is approximated during the training of a generative model. This results in a surprising observation; the method best approximating the Wasserstein distance does not produce the best looking images in the generative setting.

2 Optimal Transport

In this section, we recall essential formulations of optimal transport to fix notation.

Probabilistic Notions. Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be Polish spaces, i.e., complete and separable metric spaces, denote by $\mathcal{P}(\mathcal{X})$ the set of probability measures on $\mathcal{X}$, and let $f\colon \mathcal{X} \to \mathcal{Y}$ be a measurable map. Then, given a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we write $f_{\#}\mu$ for the *push-forward* of μ with respect to f, given by $f_{\#}\mu(A) = \mu(f^{-1}(A))$ for any measurable $A \subseteq \mathcal{Y}$. Intuitively speaking, if ξ is a random variable with law μ, then $f(\xi)$ has law $f_{\#}\mu$. Then, given $\nu \in \mathcal{P}(\mathcal{Y})$, we define

$$\mathrm{ADM}(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) |\ (\pi_1)_{\#}\gamma = \mu,\ (\pi_2)_{\#}\gamma = \nu\}, \tag{1}$$

where π_i denotes the projection onto the i^{th} marginal. An element $\gamma \in \mathrm{ADM}(\mu, \nu)$ is called an *admissible plan* and defines a joint probability between μ and ν.

Optimal Transport Problem. Given a continuous and lower-bounded *cost function* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the optimal transport problem between probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ is defined as

$$\mathrm{OT}_c(\mu, \nu) := \min_{\gamma \in \mathrm{ADM}(\mu,\nu)} \mathbb{E}_\gamma[c], \tag{2}$$

where $\mathbb{E}_\mu[f] = \int_{\mathcal{X}} f(x) d\mu(x)$ is the expectation of a measurable function f with respect to μ.

Note that (2) defines a constrained linear program, and thus admits a *dual formulation*. From the perspective of WGANs, the dual is more important than

the primal formulation, as it can be approximated using *discriminator* neural networks. Denote by $L^1(\mu) = \{f\colon \mathcal{X} \to \mathbb{R} \mid \mathbb{E}_\mu[f] < \infty\}$ the set of measurable functions of μ that have finite expectations under μ, and by $\mathrm{ADM}(c)$ the set of *admissible pairs* (φ, ψ) that satisfy

$$\varphi(x) + \psi(y) \leq c(x,y), \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad \varphi \in L^1(\mu), \psi \in L^1(\nu). \tag{3}$$

Then, the following duality holds (Peyré et al., 2019, Sec. 4)

$$\mathrm{OT}_c(\mu,\nu) = \sup_{(\varphi,\psi)\in \mathrm{ADM}(c)} \left\{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\psi]\right\}. \tag{4}$$

When the supremum is attained, the optimal φ^*, ψ^* in (4) are called *Kantorovich potentials*, which, in particular, satisfy $\varphi^*(x) + \psi^*(y) = c(x,y)$ for any $(x,y) \in \mathrm{Supp}(\gamma^*)$, where γ^* solves (2). Given φ, we can obtain an admissable ψ satisfying (4) through the *c-transform* of φ,

$$\varphi^c : \mathcal{Y} \to \mathbb{R}, \quad y \mapsto \inf_{x\in\mathcal{X}} \left\{c(x,y) - \varphi(x)\right\}, \tag{5}$$

so that $(\varphi, \varphi^c) \in \mathrm{ADM}(c)$ for any $\varphi \in L^1(\mu)$. Moreover, the Kantorovich potentials satisfy $\psi = \varphi^c$, and therefore (4) can be written as (Villani, 2008, Thm. 5.9)

$$\mathrm{OT}_c(\mu,\nu) = \max_{(\varphi,\varphi^c)\in \mathrm{ADM}(c)} \left\{\mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\varphi^c]\right\}. \tag{6}$$

In other words, the $\mathrm{ADM}(c)$ constraint can be enforced with the c-transform, and reduces the optimization in (6) to be carried out over a single function.

Wasserstein Metric. Consider the case when $\mathcal{X} = \mathcal{Y}$ and $c(x,y) = d^p_{\mathcal{X}}(x,y)$, $p \geq 1$, where we refer to $d_{\mathcal{X}}$ as the *ground metric*. Then, the optimal transport problem (2) defines the *p-Wasserstein metric* $W_p(\mu,\nu) := \mathrm{OT}_{d^p_{\mathcal{X}}}(\mu,\nu)^{\frac{1}{p}}$ on the space

$$\mathcal{P}^p_{d_{\mathcal{X}}}(X) = \left\{\mu \in \mathcal{P}(X) \;\middle|\; \int d^p_{\mathcal{X}}(x_0,x) d\mu(x) < \infty\right\}, \quad \text{for some } x_0 \in \mathcal{X}, \tag{7}$$

of probability measures with finite p-moments. It can be shown that $(\mathcal{P}^p_{d_{\mathcal{X}}}(\mathcal{X}), W_p)$ forms a complete, separable metric space (Villani, 2008, Sec. 6, Thm 6.16).

Entropy Relaxed Optimal Transport. We can relax (2) by imposing entropic penalization introduced by Cuturi (2013). Recall the definition of the *Kullback-Leibler (KL) divergence* from ν to μ as

$$\mathrm{KL}(\mu\|\nu) = \int_{\mathcal{X}} \log\left(\frac{p_\mu}{p_\nu}\right) p_\mu d\chi, \tag{8}$$

where we assume that μ, ν are absolutely continuous with respect to the Lebesgue measure χ on $\mathcal{X}$ with densities p_μ, p_ν, respectively. Using the KL-divergence as penalization, the *entropy relaxed optimal transport* is defined as

$$\mathrm{OT}^\epsilon_c(\mu,\nu) := \min_{\gamma\in\mathrm{ADM}(\mu,\nu)} \left\{\mathbb{E}_\gamma[c] + \epsilon \mathrm{KL}(\gamma\|\mu\otimes\nu)\right\}, \tag{9}$$

where $\epsilon > 0$ defines the magnitude of the penalization, and $\mu \otimes \nu$ denotes the independent joint distribution of μ and ν. We remark that when $\epsilon \to 0$, any minimizing sequence $(\gamma^\epsilon)_{\epsilon>0}$ solving (9) converges to a minimizer of (2), and in particular, $\mathrm{OT}_c^\epsilon(\mu,\nu) \to \mathrm{OT}_c(\mu,\nu)$.

Analogously to (4), the entropy relaxed optimal transport admits the following dual formulation (Di Marino and Gerolin, 2019; Feydy et al., 2018; Peyré et al., 2019)

$$\mathrm{OT}_c^\epsilon(\mu,\nu) = \sup_{\varphi\in L^1(\mu),\psi\in L^1(\nu)} \left\{ \mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\psi] - \epsilon \mathbb{E}_{\mu\otimes\nu}\left[\exp\left(\frac{-c + (\varphi\oplus\psi)}{\epsilon} \right) - 1 \right] \right\}, \tag{10}$$

where $(\varphi\oplus\psi)(x,y) = \varphi(x) + \psi(y)$. In contrast to (4), this is an *unconstrained* optimization problem, where the entropic penalization can be seen as a smooth relaxation of the constraint.

As shown by Di Marino and Gerolin (2019); Feydy et al. (2018), a similar approach to (6) for computing the Kantorovich potentials can be taken in the entropic case, by generalizing the c-transform. Let $L_\epsilon^{\exp}(\mu) := \{g : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}_\mu[\exp(g/\epsilon)] < \infty\}$, and consider the (c,ϵ)-transform of $\varphi \in L_\epsilon^{\exp}(\mu)$,

$$\varphi^{(c,\epsilon)}(y) = -\epsilon \log\left(\int_{\mathcal{X}} \exp\left(-\frac{c(x,y) - \varphi(x)}{\epsilon} \right) d\mu(x) \right). \tag{11}$$

As $\epsilon \to 0$, $\varphi^{(c,\epsilon)}(y) \to \varphi^c(y)$, making the (c,ϵ)-transform consistent with the c-transform. Analogously to (6), one can show under mild assumptions on the cost c (Di Marino and Gerolin, 2019), that

$$\mathrm{OT}_c^\epsilon(\mu,\nu) = \max_{\varphi\in L_\epsilon^{\exp}(\mu)} \left\{ \mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\varphi^{(c,\epsilon)}] \right\}. \tag{12}$$

Sinkhorn Divergence. Since the functional OT_c^ϵ fails to be positive-definite (e.g. $\mathrm{OT}_c^\epsilon(\mu,\mu) \neq 0$), it is convenient to introduce the (p,ϵ)-*Sinkhorn divergence* S_p^ϵ with parameter $\epsilon > 0$, given by

$$S_p^\epsilon(\mu,\nu) = \mathrm{OT}_{d_\mathcal{X}^p}^\epsilon(\mu,\nu)^{\frac{1}{p}} - \frac{1}{2}\left(\mathrm{OT}_{d_\mathcal{X}^p}^\epsilon(\mu,\mu)^{\frac{1}{p}} + \mathrm{OT}_{d_\mathcal{X}^p}^\epsilon(\nu,\nu)^{\frac{1}{p}} \right), \tag{13}$$

where the terms $\mathrm{OT}_{d_\mathcal{X}^p}^\epsilon(\mu,\mu)^{\frac{1}{p}}$ and $\mathrm{OT}_{d_\mathcal{X}^p}^\epsilon(\nu,\nu)^{\frac{1}{p}}$ are added to avoid bias, as in general $\mathrm{OT}_{d_\mathcal{X}^p}^\epsilon(\mu,\mu)^{\frac{1}{p}} \neq 0$. The Sinkhorn divergence was introduced by Genevay et al. (2018), and has the following properties: (i) it metrizes weak convergence in the space of probability measures; (ii) it interpolates between *maximum mean discrepancy* (MMD), as $\epsilon \to \infty$, and the p-Wasserstein metric, as $\epsilon \to 0$. For more about the Sinkhorn divergence, see Feydy et al. (2018).

3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are popular for learning to sample from distributions. The idea behind GANs can be summarized as follows: given

a *source distribution* $\mu_s \in \mathcal{P}(\mathbb{R}^{n_s})$, we want to push it forward by a parametrized *generator* $g_{\omega'}\colon \mathbb{R}^{n_s} \to \mathbb{R}^{n_t}$, so that a chosen similarity measure ρ between the pushforward distribution and the *target distribution* $\mu_t \in \mathcal{P}(\mathbb{R}^{n_t})$ is minimized. Usually the target distribution is only accessible in form of a dataset of samples, and one considers an 'empirical' version of the distributions. Note that $n_s \ll n_t$ is chosen, which is justified by the *manifold hypothesis*. This objective can be expressed as

$$\min_{\omega'} \rho((g_{\omega'})_{\#}\mu_s, \mu_t). \tag{14}$$

The similarity ρ is commonly estimated with a *discriminator* $\varphi_\omega\colon \mathbb{R}^{n_t} \to \mathbb{R}$, parametrized by ω, whose role will become apparent below.

The vanilla GAN (Goodfellow et al., 2014) minimizes an approximation to the *Jensen-Shannon (JS) divergence* between the push-forward and target, given by

$$\mathrm{JS}(\nu\|\mu) \approx \max_{\omega} \left\{\mathbb{E}_{x\sim\mu}\left[\log(\varphi_\omega(x))\right] + \mathbb{E}_{y\sim\nu}\left[\log(1-\varphi_\omega(y))\right]\right\}, \tag{15}$$

for probability measures μ and ν. The discriminator φ_ω is restricted to take values between 0 and 1, assigning a probability to whether a point lies in μ or ν. It can be shown (Goodfellow et al., 2014), that at optimality the JS-divergence is recovered in (15), if optimized over all possible functions. Substituting $\mu = \mu_t$ and $\nu = (g_{\omega'})_{\#}\mu_s$ in (15) yields the minimax objective for the vanilla GAN

$$\min_{\omega'}\max_{\omega} \left\{\mathbb{E}_{x\sim\mu_t}\left[\log(\varphi_\omega(x))\right] + \mathbb{E}_{z\sim\mu_s}\left[\log(1-\varphi_\omega(g_{\omega'}(z)))\right]\right\}. \tag{16}$$

As mentioned above, in practice one considers empirical versions of the distributions so that the expectations are replaced by sample averages.

The Wasserstein GANs (WGANs) (Arjovsky et al., 2017) minimize an approximation to the 1-Wasserstein metric over the l^2 ground metric, instead. The reason why the $p = 1$ Wasserstein case is considered is motivated by a special property of the c-transform of 1-Lipschitz functions, when $c = d$ for any metric d: if f is 1-Lipschitz, then $f^c = -f$ (Villani, 2008, Sec. 5). It can also be shown, that a Kantorovich potential φ^* solving the dual problem (6) is 1-Lipschitz when $c = d$, and therefore the WGAN minimax objective can be written as

$$\min_{\omega'}\max_{\omega} \left\{\mathbb{E}_{x\sim\mu_t}\left[\varphi_\omega(x)\right] - \mathbb{E}_{z\sim\mu_s}\left[\varphi_\omega(g_{\omega'}(z))\right]\right\}. \tag{17}$$

In the WGAN case, there is no restriction on the range of φ_ω as opposed to the GAN case above. The assumptions above require enforcing φ_ω to be 1-Lipschitz. This poses a main implementational difficulty in the WGAN formulation, and has been subjected to a considerable amount of research.

In this work, we will investigate the original approach by weight clipping (Arjovsky et al., 2017) and the popular approach by gradient norm penalties for the discriminator (Gulrajani et al., 2017). We furthermore consider a more direct approach that computes the c-transform over minibatches (Mallasto et al., 2019), avoiding the need to ensure Lipschitzness. We also discuss an entropic relaxation approach through (c, ϵ)-transforms over minibatches, introduced by Genevay et al. (2016). In the original work, the discriminator φ_ω is expressed

as a sum of kernel functions, however, in this work we will consider multi-layer perceptrons (MLPs), as we do with the other methods we consider.

3.1 Estimating the 1-Wasserstein Metric

In the experimental section, we will consider four ways to estimate the 1-Wasserstein distance between two measures μ and ν, these being the weight clipping (WC), gradient penalty (GP), c-transform and (c, ϵ)-transform methods. To this end, we now discuss how these estimates are computed in practice by sampling minibatches of size N from μ and ν, yielding $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, respectively, at each training iteration. Then, with each method, the distance is estimated by maximizing a model specific expression that relates to the dual formulation of the 1-Wasserstein distance in (6) over the minibatches. In practice, this maximization is carried out via gradient ascent or one of its variants, such as Adam (Kingma and Ba, 2014) or RMSprop (Tieleman and Hinton, 2012).

Weight clipping (WC). The vanilla WGAN enforces K-Lipschitzness of the discriminator at each iteration by forcing the weights W^k of the neural network to lie inside some box $-\xi \leq W^k \leq \xi$, considered coordinate-wise, for some small $\xi > 0$ ($\xi = 0.01$ in the original work). Here k stands for the k^{th} layer in the neural network. Then, the identity for the c-transform (with $c = d$) of 1-Lipschitz maps is used, and so (17) can be written as

$$\max_{\omega} \left\{ \frac{1}{N} \sum_{i=1}^{N} \varphi_\omega(x_i) - \frac{1}{N} \sum_{i=1}^{N} \varphi_\omega(y_i) \right\}. \tag{18}$$

Gradient penalty (GP). The weight clipping is omitted in WGAN-GP, by noticing that the 1-Lipschitz condition implies that $\|\nabla_x \varphi_\omega(x)\| \leq 1$ holds for x almost surely under μ and ν. This condition can be enforced through the penalization term $\mathbb{E}_{x\sim\chi}\big[\max\left(0, 1 - \|\nabla_x \varphi_\omega(x)\|\right)^2\big]$, where χ is some reference measure, proposed to be the uniform distribution between pairs of points of the minibatches by Gulrajani et al. (2017). The authors remarked that in practice it suffices to enforce $\|\nabla_x \varphi_\omega(x)\| = 1$, and thus the objective can be written as

$$\max_{\omega} \left\{ \frac{1}{N} \sum_{i=1}^{N} \varphi_\omega(x_i) - \frac{1}{N} \sum_{i=1}^{N} \varphi_\omega(y_i) - \frac{\lambda}{M} \sum_{i=1}^{M} \left(1 - \|\nabla_{z=z_i} \varphi_\omega(z)\|\right)^2 \right\}, \tag{19}$$

where λ is the magnitude of the penalization, which was chosen to be $\lambda = 10$ in the original paper, and $\{z_i\}_{i=1}^M$ are samples from χ.

***c*-transform.** Enforcing 1-Lipschitzness has the benefit of reducing the computational cost of the c-transform, but the enforcement introduces an additional cost, which in the gradient penalty case is substantial. The ADM(c) constraint can be taken into account directly, as done in (q, p)-WGANs (Mallasto et al., 2019), by directly computing the c-transform over the minibatches as

$$\varphi_\omega^c(y_i) \approx \widehat{\varphi_\omega^c}(y_i) = \min_j \left\{ c(x_j, y_i) - \varphi_\omega(x_j) \right\}, \tag{20}$$

where $c = d_2$ in the 1-Wasserstein case. This amounts to the relatively cheap operation of computing the row minima of the matrix $A_{ij} = c(x_j, y_i) - \varphi_\omega(x_j)$. The original paper proposes to include penalization terms to enforce the discriminator constraints, however, this is unnecessary as the c-transform enforces the constraints. Therefore, the objective can be written as

$$\max_{\omega} \left\{ \frac{1}{N} \sum_{i=1}^{N} \varphi_\omega(x_i) + \frac{1}{N} \sum_{i=1}^{N} \widehat{\varphi_\omega^c}(y_i) \right\}. \tag{21}$$

(c, ϵ)-transform. As discussed in Section 2, entropic relaxation applied to W_1 results in the $(1, \epsilon)$-Sinkhorn divergence S_1^ϵ, which satisfies $S_1^\epsilon \to W_1$ as $\epsilon \to 0$. Then, S_1^ϵ can be viewed as a smooth approximation to W_1. The benefits of this approach are that φ_ω is not required to satisfy the ADM(c) constraint, and the resulting transport plan is smoother, providing robustness towards noisy samples.

The expression (13) for S_1^ϵ consists of three terms, where each results from solving an entropy relaxed optimal transport problem. As stated by Feydy et al. (2018, Sec. 3.1), the terms $\mathrm{OT}_1^\epsilon(\mu, \mu)$ and $\mathrm{OT}_1^\epsilon(\nu, \nu)$ are straight-forward to compute, and tend to converge within couple of iterations of the symmetric Sinkhorn-Knopp algorithm. For efficiency, we approximate these terms with one Sinkhorn-Knopp iteration. The discriminator is employed in approximating $\mathrm{OT}_1^\epsilon(\mu, \nu)$, which is done by computing the (c, ϵ)-transform defined in (11) over the minibatches

$$\varphi_\omega^c(y_i) \approx \widehat{\varphi_\omega^{(c,\epsilon)}}(y_i) = -\epsilon \log \left(\frac{1}{N} \sum_{j=1}^{N} \exp\left(-\frac{1}{\epsilon} \left(\varphi_\omega(x_j) - c(x_j, y_i) \right) \right) \right), \tag{22}$$

and so we write the objective (12) for the discriminator as

$$\max_{\omega} \left\{ \frac{1}{N} \sum_{i=1}^{N} \varphi_\omega(x_i) + \frac{1}{N} \sum_{j=1}^{N} \widehat{\varphi_\omega^{(c,\epsilon)}}(y_j) \right\}. \tag{23}$$

4 Experiments

We now study how efficiently the four methods presented in Section 3.1 estimate the 1-Wasserstein metric. The experiments use the MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2015) datasets. On these datasets, we focus on two tasks: approximation and stability. By approximation we mean how well the *minibatch-wise* distance between two measures can be computed, and by stability how well these minibatch-wise distances relate to the 1-Wasserstein distance between the two full measures.

Implementation. In the approximation task, we model the discriminator as either (i) a simple multilayer perceptron (MLP) with two hidden layers of width 128 and ReLU activations, and a linear output, or (ii) a convolutional neural network architecture (CNN) used in DCGAN (Radford et al., 2015). In the stability

task we use the simpler MLPs for computational efficiency. The discriminator is trained by optimizing the objective function using stochastic or batch gradient ascent. For the gradient penalty method, we use the Adam optimizer with learning rate 10^{-4} and beta values $(0, 0.9)$. For weight clipping we use RMSprop with learning rate 5×10^{-5} as specified in the original paper by Arjovsky et al. (2017). Finally, for the c-transform and the (c, ϵ)-transform, we use RMSprop with learning rate 10^{-4}.

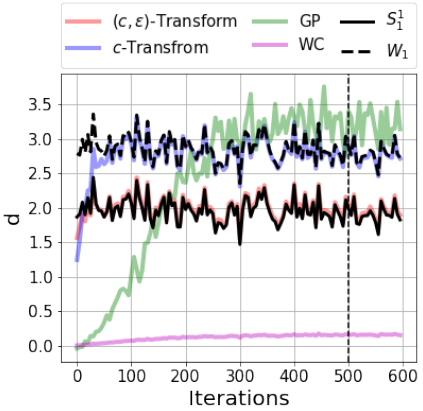


Figure 1: Estimating the distance between two standard 2-dimensional Gaussian distributions that have been shifted by $\pm[1, 1]$.

The estimated distances d_{est} obtained from the optimization are compared to ground truth values d_{ground} computed by POT[1]. The (c, ϵ)-transform might improve on the POT estimates when $d_{\text{est}} > d_{\text{ground}}$, as both values result from maximizing the same unconstrained quantity. This discrepancy cannot be viewed as error, which we quantify as

$$\text{err}(d_{\text{est}}, d_{\text{ground}}) = \max(0, d_{\text{ground}} - d_{\text{est}}). \quad (24)$$

Note that this is a subjective error based on the POT estimate that can also err, and not absolute error. In practice POT is rather accurate, and we found this to make only a small difference (see Figure 2 with the ground truth and estimated distances visualized). The weight clipping and the gradient penalty methods might also return a higher value than POT, but in this case it is not guaranteed that the discriminators are admissible, meaning that the constraints of the maximization objective would not be satisfied. In the case of the c-transform, the discriminators are always admissible. However, the POT package's *ot.emd* (used to compute the 1-Wasserstein distance ground truth) does not utilize the dual formulation for computing the optimal transport distance, and therefore we cannot argue in the same way as in the (c, ϵ)-transform case. As the Sinkhorn divergence (13) consists of three terms that are each maximized, we measure the error as the sum of the errors given in (24) of each term.

Approximation. We divide the datasets into two, forming the two measures μ and ν, between which the distance is approximated, and train the discriminators on 500 *training minibatches* $\mu_k \subset \mu$ and $\nu_k \subset \nu$, $k = 1, ..., 500$, of size 64. See Section 3.1 for how the discriminator objectives. Then, without training the discriminators further, we sample another 100 *evaluation minibatches* $\mu'_l \subset \mu$ and $\nu'_l \subset \nu$, $l = 1, ..., 100$, and use the discriminators to approximate the *minibatch-wise* distance between each μ'_l and ν'_l. This approximation will then be compared to the ground-truth minibatch-wise distance computed by POT. We run this experiment 20 times, initializing the networks again each time, and report the

[1]Python Optimal Transport, `https://pot.readthedocs.io/`.

MLP	MNIST	CIFAR10	CelebA
WC	14.98 ± 0.32	27.26 ± 0.61	48.65 ± 1.29
GP	14.89 ± 0.38	27.14 ± 0.87	48.00 ± 2.88
c-transform	0.82 ± 0.16	1.53 ± 0.29	2.84 ± 0.49
$(c, 0.1)$-transform	0.43 ± 0.17	1.29 ± 0.48	2.52 ± 1.28
$(c, 1)$-transform	$(1.12 \pm 4.76) \times 10^{-10}$	$(0.26 \pm 5.97) \times 10^{-4}$	0.04 ± 0.26
ConvNet	MNIST	CIFAR10	CelebA
WC	20.73 ± 18.59	27.28 ± 0.63	48.72 ± 1.33
GP	14.78 ± 0.54	25.20 ± 24.32	96.19 ± 77.90
c-transform	0.79 ± 0.16	1.00 ± 0.26	2.11 ± 0.46
$(c, 0.1)$-transform	0.42 ± 0.17	0.60 ± 0.41	1.74 ± 1.13
$(c, 1)$-transform	$(0.23 \pm 1.90) \times 10^{-8}$	$(0.40 \pm 3.60) \times 10^{-13}$	0.02 ± 0.17

Table 1: **Approximation**. For each method, the discriminators are trained 20 times for 500 iterations on minibatches of size 64 drawn without replacement, after which training is stopped and the error between the ground truth and the estimate are computed.

average error in Table 1. Note that the discriminators are not updated for the last 100 iterations. Results for a toy example between two Gaussians are presented in Fig. 1.

As Table 1 shows, the c-transform approximates the minibatch-wise 1-Wasserstein distance far better than weight clipping or gradient penalty, and (c, ϵ)-transform does even better at approximating the 1-Sinkhorn divergence. The low errors in this case are due to the (c, ϵ)-transform outperforming the POT library, which results in an error of 0 on many iterations, which also explains why the error variance is so high in the $\epsilon = 1$ case.

Stability. We train the discriminators for 500 iterations on small datasets of size 512, that form subsets of the datasets mentioned above. We train with two minibatch sizes $N = 64$ and $N = 512$. We then compare how the resulting discriminators estimate the minibatch-wise and total distances, that is, the evaluation minibatches are of size $M = 64$ and $M = 512$. Letting Ψ_{Method} be the objective presented in 3.1 for a given method, we train the discriminator by maximizing $\Psi_{\text{Method}}(\phi, \mu_N, \nu_N)$, and finally compare $\Psi_{\text{Method}}(\phi, \mu_M, \nu_M)$ for different M. The results are presented in Table 2. An experiment on CIFAR10 was carried out to illustrate how the distance estimate between the full measures behaves when trained minibatch-wise, which is included in Appendix A.

The ground truth values computed using POT are also included, but are not of the main interest in this experiment. The focus is on observing how different batch-sizes on training and evaluation affect the resulting distance. For the c-transform and (c, ϵ)-transform, the results varies depending on whether the distances are evaluated minibatch-wise or on the full datasets. On the other hand, for the gradient penalty and weight clip methods, the training batch-size has more effect on the result, but the minibatch-wise and full evaluations are comparable.

WGAN training. Finally, we measure how the different methods fare when

MNIST	$(c,1)$-transform	c-transform	GP	WC
$N=512$, $M=64$	17.20 ± 0.16	13.87 ± 0.23	4.25 ± 0.49	2.10 ± 0.26
$N=512$, $M=512$	16.95	12.64	4.21	2.03
$N=64$, $M=64$	17.45 ± 0.06	14.12 ± 0.13	1.54 ± 0.25	1.12 ± 0.13
$N=64$, $M=512$	16.76	11.4	1.49	1.08
Ground truth	14.22	12.65	12.65	12.65
CIFAR10	$(c,1)$-transform	c-transform	GP	WC
$N=512$, $M=64$	29.98 ± 0.28	26.44 ± 0.25	11.04 ± 1.16	3.85 ± 0.67
$N=512$, $M=512$	29.41	24.77	11.10	4.00
$N=64$, $M=64$	29.67 ± 0.41	26.21 ± 0.40	3.25 ± 0.53	2.19 ± 0.23
$N=64$, $M=512$	29.18	24.16	3.59	2.34
Ground truth	26.10	24.78	24.78	24.78
CelebA	$(c,1)$-transform	c-transform	GP	WC
$N=512$, $M=64$	50.55 ± 0.86	46.56 ± 0.89	28.07 ± 10.61	19.18 ± 73.86
$N=512$, $M=512$	48.42	43.06	28.17	20.93
$N=64$, $M=64$	50.80 ± 0.91	46.83 ± 0.86	10.24 ± 7.31	13.98 ± 39.54
$N=64$, $M=512$	47.60	41.80	10.10	15.20
Ground truth	43.74	43.07	43.07	43.07

Table 2: **Stability**. The discriminators are trained using two training batch sizes, $N=64$ and $N=512$. Then, the distances between the measures are estimated, by evaluating the discriminators on the full measures (of size $M=512$), or by evaluating minibatch-wise with batch size $M=64$. Presented here are the distances approximated by each way of training the discriminators.

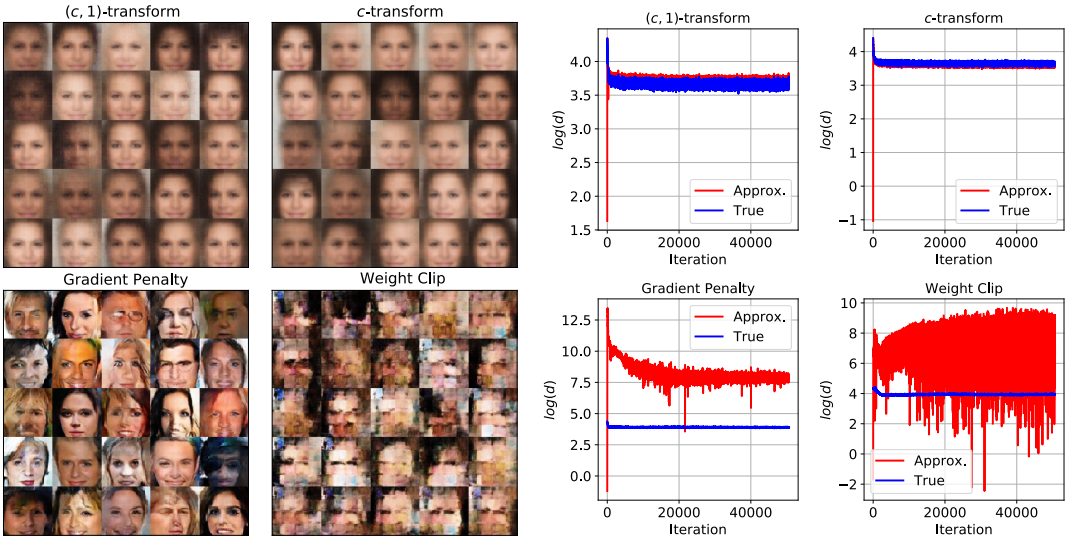


Figure 2: Approximating the minibatch-wise distances while training a generator. Left: generated faces after 5×10^4 generator iterations. Right: True and estimated log-distances between sampled batches. The large values of GP and WC can be attributed to the fact that they are susceptible to failure in enforcing the Lipschitz constraints on the discriminator.

training a Wasserstein GAN. For this, we use the architecture from DCGAN (Radford et al., 2015) to learn a generative model on the CelebA dataset. During this training, we compute the POT ground truth distance between presented minibatches, and compare these to the estimated distances given by the discriminators. This is carried out for a total of 5×10^4 generator iterations, each of which is accompanied by 5 discriminator iterations. For (c, ϵ)- and c-transform the discriminators are evaluated on the fake samples, which gave similar results to evaluating on the real samples. The results are presented in Fig. 2. The results clearly show how the (c, ϵ)-transform and c-transform estimate the Wasserstein distances at each iteration better than gradient penalty or weight clipping. However, the resulting images are blurry and look like averages of clusters of faces. The best quality images are produced by the gradient penalty method, whereas weight clipping does not yet converge to anything meaningful. We include the same experiment ran with the simpler MLP architecture for the discriminator, while the generator still is based on DCAN, in Appendix B.

5 Discussion

Based on the experiments, (c, ϵ)-transform and c-transform are more accurate at computing the minibatch distance and estimating the batch distance than the gradient penalty and weight clipping methods. However, despite the lower performance of the latter methods, in the generative setting they produce more unique and compelling samples than the former. This raises the question, whether the exact 1-Wasserstein distance between batches is the quantity that should be considered in generative modelling, or not. On the other hand, an interesting direction is to study regularization strategies in conjunction with the (c, ϵ)- and c-transforms to improve generative modelling with less training.

The results of Table 2 indicate that the entropic relaxation provides stability under different training schemes, endorsing theoretical results implying more favorable sample complexity in the entropic case. In contrast to what one could hypothetise, the blurriness in Fig. 2 seems not to be produced by the entropic relaxation, but the c-transform scheme.

Finally, it is interesting to see how the gradient penalty method performs so well in the generative setting, when based on our experiments, it is not able to approximate the 1-Wasserstein distance so well. What is it, then, that makes it such a good objective in the generative case?

References

Jonas Adler and Sebastian Lunz. Banach Wasserstein GAN. In *Advances in Neural Information Processing Systems*, pages 6755–6764, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Ishan Deshpande, Ziyu Zhang, and Alexander Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

Simone Di Marino and Augusto Gerolin. An Optimal Transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *in preparation*, 2019.

Richard Mansfield Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Yonatan Dukler, Wuchen Li, Alex Lin, and Guido Montúfar. Wasserstein of Wasserstein loss for learning generative models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1716–1725, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

Wilfrid Gangbo, Wuchen Li, Stanley Osher, and Michael Puthawala. Unnormalized optimal transport. *arXiv preprint arXiv:1902.03367*, 2019.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thierry Klein, Jean-Claude Fort, and Philippe Berthet. Convergence of an estimator of the wasserstein distance between two continuous probability distributions. 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Rongjie Lai and Hongkai Zhao. Multi-scale non-rigid point cloud registration using robust sliced-Wasserstein distance via Laplace-Beltrami eigenmap. *arXiv preprint arXiv:1406.3758*, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q, p)-Wasserstein GANs: Comparing ground metrics for Wasserstein GANs. *arXiv preprint arXiv:1902.03642*, 2019.

Gonzalo Mena and Jonathan Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *arXiv preprint arXiv:1905.11882*, 2019.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40 (2):99–121, 2000.

Martin Sommerfeld. Wasserstein distance on finite spaces: Statistical inference and algorithms. 2017.

Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient

by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *International Conference on Learning Representation (ICLR)*, 2018.

Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. Wasserstein divergence for GANs. In *Computer Vision – ECCV 2018*, pages 673–688, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01228-1.

A Appendix: Stability during Training

Related to the **Stability** experiment, we train discriminators modelled by the simple MLPs (see Section 4) to approximate the distance between two measures μ and ν, which both are subsets of the CIFAR10 dataset of size 512. We train for 15000 iterations with training minibatch size of 64, and report the estimated minibatch-wise distances with the estimated distances between the full measures in Fig. 3.

The experiments demonstrate, how (c, ϵ)-transform and c-transform converge rapidly compared to gradient penalty and weight clipping method, which have not entirely converged after 15000 iterations. Also visible is the bias resulting form minibatch-wise computing of the distances compared to the distance between the full measures.

B Appendix: WGAN Training with MLP

We repeat the **WGAN training** experiment with the simpler MLP architecture (see Section 4) for the discriminators. The distance estimates at each iteration are given in Fig. 4, and generated samples in Fig. 5.

The training process for (c, ϵ)- and c-transforms is more unstable with the MLP, as notable in the sudden jumps in the true distance between minibatches. This seems to be caused when the discriminator underestimates the distance. The fluctuation between estimated distances is much higher in the gradient penalty and weight clipping cases, but the decrease in the true distance between minibatches is still consistent. Notice how the fluctuation decreases considerably when we use a ConvNet architecture for the gradient penalty method in Fig. 2.

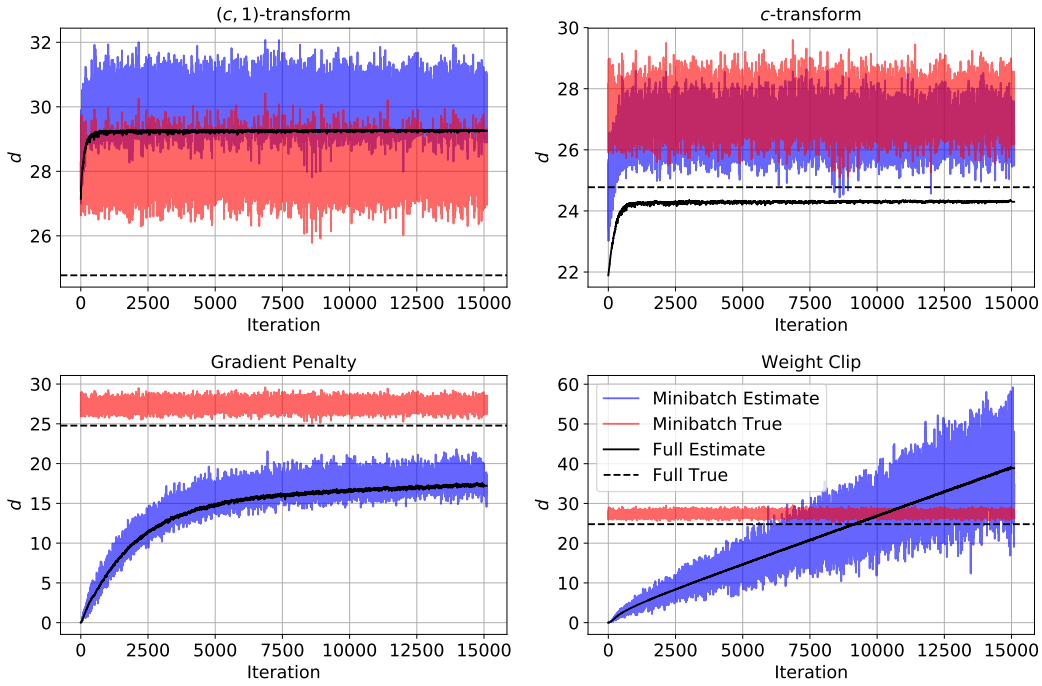


Figure 3: **Stability**. While training the discriminators on minibatches, taken from two measures consisting of 512 samples from CIFAR10, we also report the estimated distance between the full measures.

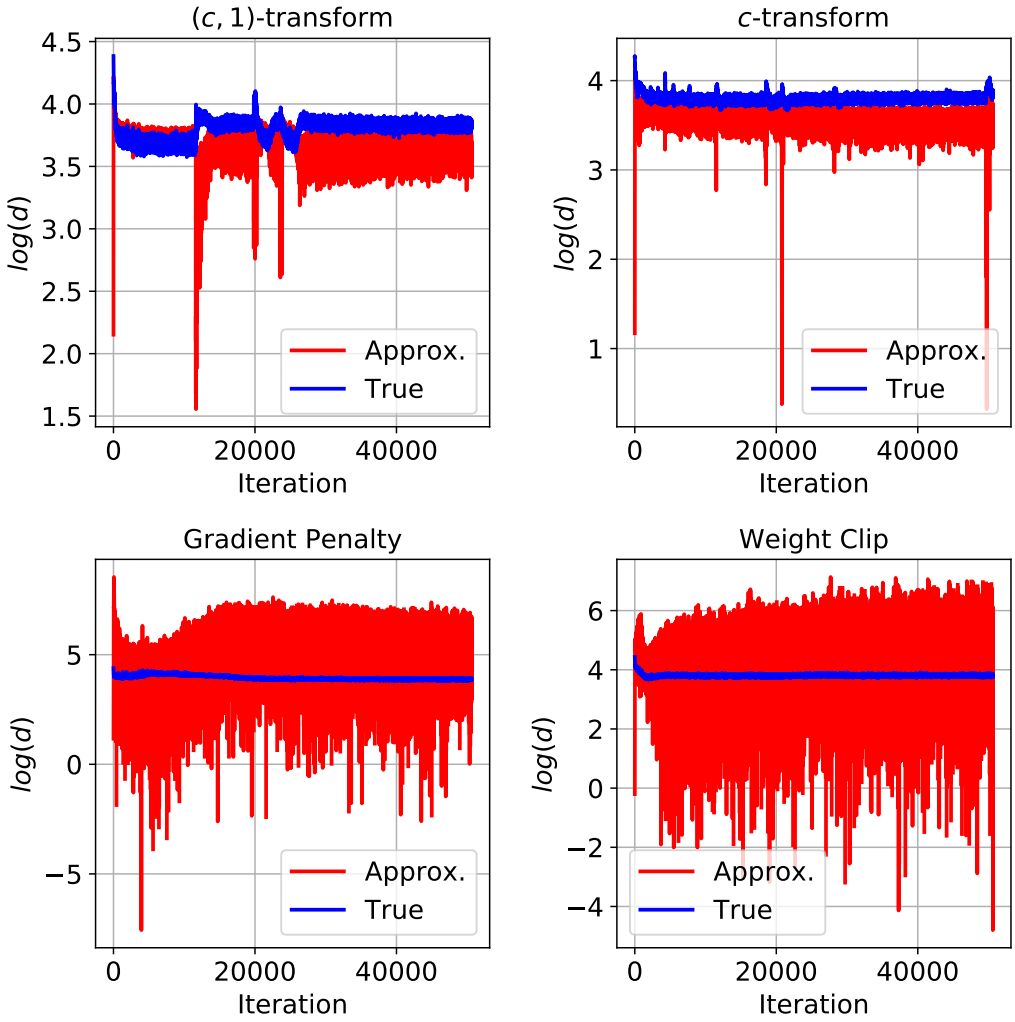


Figure 4: Repeating the experiment presented in Fig. 2, but with the simpler MLP architecture for the discriminator. Presented here are the estimated batchwise distances at each iteration against the ground truth.

Figure 5: Repeating the experiment presented in Fig. 2, but with the simpler MLP architecture for the discriminator. Presented here are generated samples after 5×10^4 iterations.