UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

CONTRIBUTIONS TO BAYESIAN MACHINE LEARNING VIA TRANSPORT MAPS

TESIS PARA OPTAR AL GRADO DE DOCTOR EN
CIENCIAS DE LA INGENIERÍA, MENCIÓN MODELACIÓN MATEMÁTICA

GONZALO ANDRÉS RÍOS DÍAZ

PROFESOR GUÍA:
JOAQUÍN FONTBONA TORRES
PROFESOR GUÍA 2:
FELIPE TOBAR HENRÍQUEZ

MIEMBROS DE LA COMISIÓN:
ALEJANDRO JOFRÉ CÁCERES
JAIME SAN MARTÍN ARISTEGUI
ELSA CAZELLES

Este trabajo ha sido parcialmente financiado por CMM Conicyt PIA AFB170001 y
Conicyt-PCHA Doctorado Nacional 2016-21161789

SANTIAGO DE CHILE
2020

ABSTRACT OF THE TESIS TO QUALIFY TO THE DEGREE OF
DOCTOR OF SCIENCE IN ENGINEERING, MENTION MATHEMATICAL MODELING
BY: GONZALO ANDRÉS RÍOS DÍAZ
DATE: 2020
GUIDES: JOAQUÍN FONTBONA TORRES AND FELIPE TOBAR HENRÍQUEZ

CONTRIBUTIONS TO BAYESIAN MACHINE LEARNING VIA TRANSPORT MAPS

The uncertainty is intrinsic in machine learning since it is present in data, models, parameters, and prediction. The Bayesian approach to machine learning considers all the uncertainty under the same point of view, and thanks to Bayes law, it applies the probabilistic reasoning on all levels, including the inference of the parameters of statistical models. In this work, we develop two lines of research, using results of transport maps on two Bayesian contexts, each of them under a unifying approach of previous works from the literature. After an introduction to the Bayesian paradigm for modelling, the first part of this work reviews Gaussian processes (GP), to then propose generalisations of these Bayesian non-parametric models for regression. The second part focuses on the study of novel estimators and practical methods for training models from data. We develop both topics in a fundamental way, in the sense that we present general models and techniques that can be applied, potentially, in any context of natural science, social science or engineering. In each chapter, we provide illustrative numerical examples, using synthetic and real-world datasets, in order to experimentally validate the proposed models and methods, to finally confirm their applicability, accuracy and robustness.

On the first half of this thesis, we introduce GPs, non-parametric prior distributions over functions, used as generative models with appealing modelling properties for Bayesian inference: they can model non-linear relationships with noisy observations, have closed-form expressions for training and inference, and are governed by interpretable hyperparameters. However, GP models rely on Gaussianity, an assumption that is not true in several real-world scenarios, e.g., when observations are bounded or have extreme-value dependencies, a natural phenomenon in physics, finance and social sciences. First, to model non-Gaussian data, we propose the compositionally-warped GP, a computationally efficient non-Gaussian generative model. After that, we extend this model via different layers based on transport maps, which allows us to isolate marginals, correlations and copula of the induced stochastic process. Our proposal encompasses GPs, warped GPs, Student-t processes and other models under a single unified approach. We also provide analytical expressions and algorithms for training and inference of the proposed models in the regression problem.

On the second half, we introduce a novel paradigm for Bayesian learning based on optimal transport theory. Namely, we propose to use the Wasserstein barycenter of the posterior law on models as model selection criterion, thus introducing an alternative to classical choices like maximum a posteriori estimator or Bayesian model average. We exhibit general conditions granting the existence and statistical consistency of this estimator, discuss some of its broad and specific properties, and provide insight into its theoretical advantages. Finally, we introduce a novel method which is ideally suited for the computation of our estimator, explicitly presenting its implementation for expressive families of models. This method corresponds to a stochastic gradient descent algorithm in the Wasserstein space, so it is of general interest and applicability for the computation of populations Wasserstein barycenters.

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
DOCTORADO EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MODELACIÓN MATEMÁTICA
POR: GONZALO ANDRÉS RÍOS DÍAZ
FECHA: 2020
PROF. GUÍA: JOAQUÍN FONTBONA TORRES Y FELIPE TOBAR HENRÍQUEZ

CONTRIBUTIONS TO BAYESIAN MACHINE LEARNING VIA TRANSPORT MAPS

La incertidumbre es intrínseca en el aprendizaje automático ya que está presente en los datos, modelos, parámetros y predicciones. El enfoque Bayesiano del aprendizaje automático considera toda la incertidumbre bajo un mismo punto de vista y, gracias a la ley de Bayes, aplica el razonamiento probabilístico en todos los niveles, incluida la inferencia de los parámetros de los modelos estadísticos. En este trabajo desarrollamos dos líneas de investigación, utilizando resultados de mapas de transporte en dos contextos Bayesianos, cada uno de ellos bajo un enfoque unificador de trabajos anteriores en la literatura. Después de una introducción al paradigma Bayesiano para el modelado, la primera parte de este trabajo revisa los procesos Gaussianos (GP), para luego proponer generalizaciones de estos modelos Bayesianos no paramétricos de regresión. La segunda parte se centra en el estudio de estimadores novedosos y métodos prácticos para entrenar modelos a partir de datos. Desarrollamos ambos temas de manera fundamental, en el sentido de que presentamos modelos y técnicas generales que pueden aplicarse, potencialmente, en cualquier contexto de ciencias naturales, ciencias sociales o ingeniería. En cada capítulo proporcionamos ejemplos numéricos ilustrativos, utilizando conjuntos de datos sintéticos y del mundo real, para validar experimentalmente los modelos y métodos propuestos, para finalmente confirmar su aplicabilidad, precisión y robustez.

En la primera mitad de esta tesis, presentamos GP, distribuciones *a priori* no paramétricas sobre funciones, utilizadas como modelos generativos con propiedades de modelado atractivas para la inferencia Bayesiana: pueden modelar relaciones no lineales con observaciones ruidosas, tienen expresiones de forma cerrada para el entrenamiento e inferencia, y se rigen por hiperparámetros interpretables. Sin embargo, los GP se basan en la Gaussianidad, una suposición que no es cierta en varios escenarios del mundo real, por ejemplo, cuando las observaciones están limitadas o tienen dependencias de valor extremo, un fenómeno natural en física, finanzas y ciencias sociales. Primero, para modelar datos no Gaussianos, proponemos el *compositionally-warped* GP, un modelo generativo no Gaussiano computacionalmente eficiente. Después de eso, extendemos este modelo a través de diferentes capas basadas en mapas de transporte, lo que nos permite aislar marginales, correlaciones y cópulas del proceso estocástico modelado. Nuestra propuesta abarca GP, *warped* GP, procesos de Student-t y otros modelos bajo un único enfoque unificado. También proporcionamos expresiones analíticas y algoritmos para el entrenamiento e inferencia de los modelos de regresión propuestos.

En la segunda mitad, presentamos un paradigma novedoso para el aprendizaje Bayesiano basado en la teoría de transporte óptimo. Es decir, proponemos utilizar el baricentro de Wasserstein de la ley posterior sobre modelos como criterio de selección, introduciendo así una alternativa a las elecciones clásicas como estimador máximo a posteriori o *Bayesian model average*. Exhibimos condiciones generales que garantizan la existencia y la consistencia estadística de este estimador, discutimos algunas de sus propiedades, y proporcionamos información sobre sus ventajas teóricas. Finalmente, presentamos un método novedoso que es

ideal para el cálculo de nuestro estimador, presentando explícitamente su implementación para familias expresivas de modelos. Este método corresponde a un algoritmo de descenso de gradiente estocástico en el espacio de Wasserstein, por lo que es de interés general y de aplicabilidad para el cálculo de baricentros de Wasserstein.

Gracias a mis padres por darme todo lo necesario para empezar, gracias a León por darme la motivación para continuar, gracias a Tracy por darme la fuerza para terminar.

Acknowledgements

"It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment."

– Carl Friedrich Gauss

This thesis was supported by Conicyt PIA AFB170001 Center for Mathematical Modeling and Conicyt-Pcha Doctorado Nacional 2016-21161789. Part of this work was carried out during a visit to Julio Backhoff, which was partially funded by the FWF-grant Y00782. I thank Mathias Beiglböck and the Vienna University of Technology for the support and their hospitality. I appreciate the valuable feedback obtained from discussions with Felipe Tobar, Joaquín Jontbona and Julio Backhoff. Thanks to all the reviewers of my work who guided me on how to improve this work, especially Leonardo Zepeda and Ignacio Correa. Finally thanks to Andrés Groisman, Andrés Gottlieb and Daniel Remenik for their patience and understanding during the development of this work.

Contents

List of Tables

List of Figures

Introduction

> *"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed."*
>
> – Arthur Samuel, 1959

Due to the tremendous technological development in the last decade, the amount of data generated and collected has reached dimensions not known to humanity until now. By this fact, the field of machine learning has attracted attention in a cross-disciplinary way in various areas of natural sciences, social sciences, engineering and medicine, to name a few. Besides, development has not only occurred in the Academy but large companies such as Google, Facebook and Amazon have strengthened the area with their own research and development teams. Society is experiencing the fourth industrial revolution, the artificial intelligence revolution, where machine learning occupies a central role, the role of making machines learn from data.

There are multiple classifications and divisions of machine learning methods and models, but we can highlight a characteristic that divides them into two broad groups: those with a probabilistic approach and those without it. It is a fact that most real-world data is noisy observations of latent phenomena, so it is necessary to model these random sources, and the Bayesian approach allows us to infer models and make predictions naturally. This work adopts the point of view that the best way to make machines that can learn from data is through the tools of probability theory, which has been the mainstay of statistics and engineering for centuries, so we can proudly say that this work follows the Bayesian approach.

What initially motivated us to study the Bayesian approach are the so-called non-parametric models: those models that, despite their name, have infinite parameters, or an unbounded number of parameters that increases as we observe more data. Some common examples of non-parametric models are histograms and spline functions, but those based on Bayesian statistics have more elegant and formal mathematics, coinciding in many respects with the stochastic processes. An example of this is the so-called Dirichlet process, a distribution on discrete distributions, so it is very useful in clustering problems.

A widely non-parametric Bayesian model used for regression tasks is the Gaussian process, a distribution over functions where the well-known Brownian motion is a particular case of them. Once we understood the elegance of Gaussian processes as regression models, this made us wonder, are there more general stochastic processes that maintain the same grace

and beauty as the Gaussian processes? This question initiated an investigation that tries to answer it in this thesis, but also motivated the development of interactions between Bayesian statistics and optimal transport theory, a relation not very explored until now. This field, also known as mass transportation, is a universal and transversal mathematical field that has vast applications in probability, physics, finance, and has had a growing interest in recent years given its use in machine learning, mainly by the good results in problems where the data has an intrinsic geometry, as images, or geospatial data.

This work develops two main topics related to Bayesian statistics and transport maps. The first topic, as mentioned above, is based on proposing non-parametric Bayesian models more general than Gaussian processes, but maintaining the suitable properties for their application as regression models. We recommend to build models based on layers, following the same paradigm as deep learning models, where the first layer is a fixed Gaussian process, then we apply different transformations, or transport maps, and thus we generate non-Gaussian stochastic processes. In our development, we prove the existence of proposed models, provides methods for training and inference with these models, and expose properties for each of the defined transports. The second topic is devoted to introducing novel methods and selection criteria for the Bayesian learning of probabilistic models. Our proposal uses elements from optimal transport theory, more precisely the Wasserstein barycenter as an estimator for predictive models. In our development, we provide conditions for existence, uniqueness and consistency, general and particular characteristics of this estimator, and conclude our contributions with the derivation of novel practical methods to calculate it.

This thesis is organised as follow. Chapter 1 is an introduction to the Bayesian approach to learning models from data, describing and exemplifying parametric and non-parametric models. This Chapter allows us to continue to Chapters 2, 3 and 4, where the Gaussian models for regression are presented and extended, or we can go directly to Chapters 5 and 6 where we deepen the Bayesian approach to model learning and propose a novel alternative to classics estimators. In Chapter 2, we present the problem of regression and the solution based on Gaussian processes from different interpretations and points of view. In Chapter 3 we introduce a model called Compositionally-Warped Gaussian processes, based on transporting a Gaussian process by coordinates, or diagonally, through the composition of elementary functions, its interpretation, the closed-form formulas for training and inference, showing its advantages experimentally. In Chapter 4, we extend the previous results analysing the diagonal transports with a more theoretical approach, incorporating new transport families, that allow us to isolate marginals, correlations and copula of induced stochastic processes. We study their properties and derive the formulas for their training and inference. In Chapter 5, our motivation is to find an alternative, non-parametric learning strategy which can cope with some of the drawbacks of standard approaches such as maximum a posteriori (MAP) or Bayesian model average (BMA). We present a novel Bayesian model estimator based on Wasserstein barycenter, named Bayesian Wasserstein barycenter (BWB), studying its existence, uniqueness, consistency and general properties. Finally, in Chapter 6, we introduce methods to calculate Wasserstein barycenter in general, including BWB, where we highlight one that can be interpreted as a stochastic gradient method in the Wasserstein space. We study its properties, convergence and advantages in a theoretical and experimental way.

Results overview

During the development of this thesis work, we obtained many thought-provoking results, so we will proceed to give a general overview of them. It is worth mentioning that we separated our results into two groups; those, under the name of *Transport Gaussian Processes,* devoted in the generalisation of Gaussian processes as non-parametric Bayesian models for regression; and those results that deepen the Bayesian approach to model selection using the Wasserstein distance of the optimal transport theory, referred to as as *Wasserstein Bayesian Learning.*

Transport Gaussian Processes

In Chapter 3, our main motivation is to extend the Gaussian processes framework [102] presented in Chapter 2, to include non-Gaussian processes and to be more accurate in the assumptions concerning the modelled data. To achieve our goal, first, in Section 3.2 we review a generative model for non-Gaussian processes, named warped Gaussian processes (WGP) [125], where a latent Gaussian process is *passed through* (in a coordinate-wise manner) an expressive and *invertible* non-linear transformation, called *warping*. The main contribution in this chapter is the proposed model in Section 3.3, termed compositionally-warped GP (CWGP) [108], that is a WGP where the warping function is the composition of elementary functions. By choosing elementary functions with derivatives and inverses known in closed-form, this model requires minimal numerical approximations, achieving an appealing computational complexity for prediction and learning. In Section 3.4, we describe an ad-hoc set of elementary functions, with explicit formulas for their derivatives and inverses, and we highlight their properties together with a recommendation on how to use them, described in Section 3.5. To conclude this chapter, in Section 3.6, we give illustrative examples using synthetic and real-world data that validate the proposed method against WGP in terms of replicability, computational efficiency, and predictive ability.

In Chapter 4, we desire to explore the theoretical limits of expressiveness of CWGP, and exploit the composition-based principle to unify other non-Gaussian models under the same point of view. In Section 4.1, we review some results on copulas to towards the study of the expressiveness of CWGP. The main contribution is the proposed novel procedure to construct stochastic process, in Section 4.2, named *transport Gaussian processes* (TGP), by the composition of transformations or transport maps [78]. We introduce three different types of transports that allow us to isolate specific characteristics of the stochastic process; the marginals coordinates (in Section 4.3), the covariance and correlation (in Section 4.4) and the intrinsic copula [144] (in Section 4.5), thereby setting the strength of dependence between the

coordinates. In each Section we determine the way to compose these transports to generate distributions that satisfy the Kolmogorov consistent conditions [134], besides the derivation of their formulas and methods for prediction and learning. In Section 4.6 we describe some computational aspects for the implementation of the families of stochastic processes that TGP approach allows expressing, including GP, WGP, Student-t processes [119], encompassing general elliptical [91] and Archimedean [81] processes. Finally, in Section 4.7, we validate our proposed model with real-world data examples.

Wasserstein Bayesian Learning

In Chapter 5, we continue with the general framework for Bayesian estimation based on loss functions over probability measures presented in Chapter 1. The first result shows that this framework covers, besides classical parametric selection criteria as MAP, non-parametric model-selection alternatives as Bayesian model average estimators and generalisations thereof, as particular instances of *Fréchet means* [93] with respect to suitable metrics/divergences on the space of probability measures. The main conceptual contribution of this section is the *Bayesian Wasserstein barycenter estimator* (BWB), a novel model-selection criterion based on optimal transport theory. In Section 5.2 we recall the notions of the celebrated p-Wasserstein distance [138, 139] and, relying on the previously developed framework, we rigorously introduce the proposed BWB estimator in Section 5.3. There we explore the existence of BWB on the Bayesian context, uniqueness, absolute continuity, and prove that our estimator has less variance than the Bayesian model average.

The second main contribution of this chapter, carried out in Section 5.4 and culminating in Theorem 5.4.10, provides sufficient conditions guaranteeing the statistical consistency for the BWB estimator, under the Wasserstein distances. This behaviour is a highly desirable feature of our estimator, both from a semi-frequentist perspective as well as from the "merging of opinions" point of view in the Bayesian framework (cf. [52, Chapter 6]). The main mathematical difficulty in our analysis comes from the fact that the data space is, in general, an unbounded metric space. The underlying tools that we employ are the celebrated Schwartz theorem ([118], [52, Proposition 6.16]) on the one hand and the concentration of measure phenomenon for averages of unbounded random variables (e.g., [79, Corollary 2.10]) on the other. We refer the reader to the works [88, 89] for a previous study of posterior consistency in a Wasserstein topology, though these works focus on discrete-measures under assumptions incomparable to ours, and do not discuss the convergence of barycenters. At a practical level, Section 5.5 provides illustrative examples and experimental evidence supporting the potential of the proposed estimator, highlighting the computationally-appealing Gaussian case and their use for real-world data.

In Chapter 6, our main aim is to provide an implementable methodology to calculate the proposed BWB estimator in practice. Current numerical methods allowing to compute minimisers of integral functionals like eq. (5.1), and therefore to calculate the BWB estimator, in particular, are mostly conceived for the case when the prior measure over models has finite support. Among these methods we stress the contributions [6, 93]. This leads us to find a method which can directly deal with the general case when the support of measures over models is possibly infinite. In Section 6.1, we present a result that allows us to approximate

the BWB estimator via an empirical version, which in turn can be calculated by the method from [93]. The main contribution of this chapter is in Section 6.2, the development of a novel algorithm which can be seen as a *stochastic gradient descent on Wasserstein space*. The proposed method is ideally suited for the computation of the BWB estimator, and more generally, for Wasserstein barycenters of measures with infinite support. Crucially, we will establish the almost sure convergence of our stochastic algorithm under given conditions in Theorem 6.2.4, and for a useful generalisation in Proposition 6.2.8.

Our stochastic gradient descent method, just like all other algorithms for the computation of Wasserstein barycenters, assumes the availability of optimal transport maps between regular probability measures. For this reason, we shall present in Section 6.3 examples of model-families for which these optimal maps are explicitly given. These families also serve to illustrate how the iterations of our stochastic descent algorithm simplify. We close the work with a comprehensive numerical experiment in Section 6.4. On the one hand, this serves to illustrate the advantages of the Bayesian Wasserstein barycenter estimator over the Bayesian model average. On the other hand, this experiment suggests as well that the stochastic gradient descent method is a superior alternative for the computation of the Bayesian Wasserstein barycenter estimator, when compared to the empirical barycenter estimator described in Section 6.1.

Publications

Throughout this investigation, the following papers and posters were submitted, published and presented:

1. *Gonzalo Rios and Felipe Tobar. Box-Cox Gaussian Processes. In 2016 Escuela de Verano Latino-Americana en Inteligencia Computacional (EVIC), Santiago, poster.*

2. *Gonzalo Rios and Felipe Tobar. Learning non-Gaussian time series using the Box-Cox Gaussian process. In 2018 IEEE International Joint Conference on Neural Networks (IJCNN), pages 1–8, July 2018* [107].

3. *Gonzalo Rios and Felipe Tobar. Compositionally-warped Gaussian processes. Neural Network, 118:235-246, 2019.* [108].

4. *Gonzalo Rios. Wasserstein Barycenters for Bayesian Learning: Application to Gaussian Process. In 2018 The Machine Learning Summer School (MLSS), Buenos Aires, poster.*

5. *Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. ArXiv preprint arXiv:1805.10833 (2018).* [12].

Chapter 1

Bayesian Approach for Model Learning

"Inside every Non-Bayesian, there is a Bayesian struggling to get out."

– Dennis Lindley

Consider samples $D = \{x_1, \ldots, x_n\}$ in a data space $\mathcal{X}$ (e.g. $\mathcal{X} \subset \mathbb{R}^q$) and a set of feasible models or probability measures $\mathcal{M} \subseteq \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$. Learning a model, also know as *model selection*, from D consists in choosing an element $m \in \mathcal{M}$ that *best* explains the data as generated by m, under some given criterion [26].

We adopt the Bayesian viewpoint, which provides a probabilistic framework to deal with model uncertainty, in terms of a *prior distribution* Π on the space $\mathcal{M}$ of models; we refer the reader to [52, 85] and references therein for mathematical background on Bayesian statistics and methods. A critical challenge in the Bayesian perspective is that of calculating a predictive law on $\mathcal{X}$, usually referred to as the *predictive posterior* [48], from the posterior distribution on $\mathcal{M}$. This learning task shall be to which this work is devoted. Let us introduce the adopted notation.

Consider a fixed *prior* probability measure Π on the model space $\mathcal{M}$, namely $\Pi \in \mathcal{P}(\mathcal{M})$. By virtue of the Bayes rule, the *posterior* measure $\Pi(\mathrm{d}m|x_1, \ldots, x_n)$ on models given the data, which is denoted for simplicity $\Pi_n(\mathrm{d}m)$, is given by

$$\Pi_n(\mathrm{d}m) := \frac{\Pi\left(x_1, \ldots, x_n|m\right)\Pi\left(\mathrm{d}m\right)}{\Pi\left(x_1, \ldots, x_n\right)}, \tag{1.1}$$

where $\Pi\left(x_1, \ldots, x_n\right) = \int_{\mathcal{M}} \Pi\left(x_1, \ldots, x_n|m\right)\Pi(\mathrm{d}m)$ is the marginal likelihood or *evidence*. The Radon–Nikodym derivative [19] of $\Pi_n(\mathrm{d}m)$ w.r.t. the prior $\Pi(\mathrm{d}m)$ is the normalized likelihood function $\Lambda_n(m) = \frac{\Pi(x_1,\ldots,x_n|m)}{\Pi(x_1,\ldots,x_n)}$, while $\mathcal{L}_n(m) = \Pi\left(x_1, \ldots, x_n|m\right)$ is just the *likelihood function*.

We assume throughout that $\mathcal{M} \subseteq \mathcal{P}_{ac}(\mathcal{X})$, where $\mathcal{P}_{ac}(\mathcal{X})$ is the subset of absolutely continuous measures with respect to a common reference σ-finite measure λ on $\mathcal{X}$ (e.g. Lebesgue). As a convention, we use the same notation for an element $m(\mathrm{d}x) \in \mathcal{M}$ and its density $m(x)$ w.r.t. λ. Assuming as customary that, conditionally on the choice of model m, the data

$x_1, \ldots, x_n \in \mathcal{X}$ are distributed as i.i.d. observations from the common law m, we can write

$$\Pi(\mathrm{d}x_1, \ldots, \mathrm{d}x_n | m) = m(x_1) \cdots m(x_n)\, \lambda(\mathrm{d}x_1) \cdots \lambda(\mathrm{d}x_n). \tag{1.2}$$

In what follows, we briefly describe how this general framework includes model spaces which are finitely parameterised, and discuss standard choices in that setting, together with their appealing features and drawbacks. This scenario could be helpful for readers who are used to parametrically-defined models.

1.1 Parametric Setting

We say that $\mathcal{M}$ is finitely parametrized if there is a number $k \in \mathbb{N}$, a set $\Theta \subseteq \mathbb{R}^k$ termed parameter space, and a (measurable) function $\mathcal{T} : \Theta \mapsto \mathcal{P}_{ac}(\mathcal{X})$, called parametrisation mapping, s.t. $\mathcal{M} = \mathcal{T}(\Theta)$; in such case we denote the model as $m_\theta := \mathcal{T}(\theta)$. If the model space $\mathcal{M}$ is finitely parametrized, learning a model boils down to finding the *best* parameters $\theta \in \Theta$. This is often done in a frequentist fashion through the *maximum likelihood estimator* (MLE) given by

$$\hat{\theta}_{MLE} \in \operatorname*{argmin}_{\theta \in \Theta} \mathcal{L}_n(\theta),$$

where $\mathcal{L}_n(\theta) = p(x_1, ..., x_n|\theta)$ is the likelihood function. The frequentist approach disregard the prior over models. In some particular cases the extreme value is unique, but in general, there are many local and global extrema. A numerical trick is to consider maximising the *log likelihood* function, usually denoted as $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$, since the solutions that interest us have positive likelihood, and the logarithm is a strictly increasing function, so maximising the likelihood is equivalent to maximising the log-likelihood. Furthermore, under the common i.i.d. assumption, the functional to optimise is a sum of terms instead of a product, which is more stable numerically, as well as its evaluation and its derivative: $\ell_n(\theta) = \sum_{\mathrm{i}=1}^n \log m_\theta(x_\mathrm{i})$ and $\partial_\theta \ell_n(\theta) = \sum_{\mathrm{i}=1}^n \frac{\partial_\theta m_\theta(x_\mathrm{i})}{m_\theta(x_\mathrm{i})}$. We next illustrate the role of the above objects in a standard machine learning application.

1.1.1 Regression problem

Given $n \in \mathbb{N}$ observations, where data consist of input z_i and output y_i pairs, that is, $x_\mathrm{i} = (z_\mathrm{i}, y_\mathrm{i}) \in \mathbb{R}^q \times \mathbb{R}$ for $\mathrm{i} = 1, \ldots, n$, the regression problem aims to find the *best* function $f : \mathbb{R}^q \to \mathbb{R}$, such that $f(z_\mathrm{i})$ is *close* to y_i for $\mathrm{i} = 1, ..., n$. Under the frequentist fashion, the terms *close* and *best* are determined only by the likelihood function.

A model $m \in \mathcal{M}$ is given by a joint distributions $p(z, y)$, named the generative model since $p(z, y) = p(y|z)p(z)$ generate outputs and inputs together. Though in regression one often needs only to deal with the conditional distribution $p(y|z)$, named as the discriminative model that generates outputs given inputs. For this reason, we can fix $p_0 \in \mathcal{P}_{ac}(\mathbb{R}^q)$ and consider parametric discriminative models as $p_\theta(z, y) = p_\theta(y|z)p_0(z)$.

If we assume a linear relationship between y and z, and that $y|z$ is normally distributed, then $p_\theta(y|z) = \mathcal{N}(y; z^\top\beta, \sigma^2)$ for $\theta = (\beta, \sigma) \in \Theta = \mathbb{R}^q \times \mathbb{R}^+$. If the data is i.i.d. then the likelihood function is given by $p(x_1, ..., x_n|\theta) = \prod_{\mathrm{i}=1}^n p_\theta(y_\mathrm{i}|z_\mathrm{i})p_0(z_\mathrm{i})$, so by denoting $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $Z = (z_1, \ldots, z_n)^\top \in \mathbb{R}^{n\times q}$, then $\hat{\beta}_{MLE} = (Z^\top Z)^{-1}Z^\top\mathbf{y}$ and $\hat{\sigma}^2_{MLE} = \frac{1}{n}(\mathbf{y} - Z\hat{\beta})^\top(\mathbf{y} - Z\hat{\beta})$.

1.1.2 Maximum A Posteriori estimator

Given $p \in \mathcal{P}(\Theta)$ a prior distribution over a parameter space Θ, its *push-forward* through the map $\mathcal{T}$ is the probability measure $\Pi = \mathcal{T}(p)$ over the space of parametrised models $\mathcal{M} = \mathcal{T}(\Theta)$, given by $\Pi(A) = p(\mathcal{T}^{-1}(A))$ for $A \in \mathcal{B}(\mathcal{M})$. Expressing the likelihood function $\Lambda_n(m)$ in terms of the parameter θ s.t. $\mathcal{T}(\theta) = m$, we then easily recover from eq. (1.1) the standard posterior density over the parameter space,

$$p_n(\theta) := p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)}.$$

Analogously to MLE, the *maximum a posteriori* (MAP) estimate is defined by

$$\hat{\theta}_{MAP} \in \operatorname*{argmax}_{\theta\in\Theta} p_n(\theta).$$

Since the marginal likelihood $p(x_1, ..., x_n)$ is constant for θ, the MAP can be calculated via

$$\hat{\theta}_{MAP} \in \operatorname*{argmax}_{\theta\in\Theta} \ell_n(\theta) + \log p(\theta).$$

Under a frequentist point of view, the *log prior* term $\log p(\theta)$ can be interpreted as a regularisation term, so in turn the MAP is a regularised estimator of MLE. In the case that $p(\theta)$ is *uninformative*, i.e. $p(\theta) \propto 1$, the MAP estimator coincides with the MLE.

The MAP approach is computationally appealing as it reduces to an optimisation problem in a finite dimensional space. The performance of this method might, however, be highly sensitive to the choice of the initial condition used in the optimisation algorithm [90]. This issue is a critical drawback, since likelihood functions over parameters may be populated with numerous local optima. The second drawback of this method is that it fails to capture global information of the model space, which might result in an overfit of the predictive distribution. Indeed, the mode can often be a very poor summary or atypical choice of the posterior distribution (e.g. the mode of an exponential density is 0, irrespective of its parameter).

Another serious failure of the MAP estimation is its dependence on the parametrisation, in other words, the estimated model we get depends on the choice of the mapping $\mathcal{T} : \Theta \to \mathcal{M}$ [85]. For instance, let $\mathcal{X} = \{0, 1\}$ so $\mathcal{M} = \{m_\mu \in \mathcal{P}(\mathcal{X})|m_\mu(\{0\}) = \mu, m_\mu(\{1\}) = 1 - \mu$, for $\mu \in [0, 1]\}$ is the space of Bernoulli distributions. Under this natural parametrisation, we can define an uniform prior Π over $\mathcal{M}$ as $\Pi(\{m_\mu|\mu \in I\}) = \lambda(I)$, where λ denotes the Lebesgue measure. Given data $x_1, ..., x_n \in \mathcal{X}$, and denoting $\bar{x} = \frac{1}{n}\sum_{\mathrm{i}=1}^n x_\mathrm{i}$, the *log likelihood* function is given by

$$\ell_n(\mu) = n\bar{x}\log\mu + n(1 - \bar{x})\log(1 - \mu).$$

Denoting $m_\mu = \mathrm{Be}(\mu)$, consider the parameter space $\Theta = [0,1]$ and three bijective parametrisation maps: $\mathcal{T}_0(\theta) = \mathrm{Be}(\theta)$, $\mathcal{T}_1(\theta) = \mathrm{Be}(\theta^{1/2})$ and $\mathcal{T}_2(\theta) = \mathrm{Be}(\theta^2)$. Under the natural parametrisation $\mathcal{T}_0$, the prior over Θ is given by $p_0(\theta) = \mathbf{1}_{\{\theta\in[0,1]\}}(\theta)$, and the respective maximisation functional and their derivative, in function of $\mu = \theta$, are given by

$$\begin{aligned} J_0(\mu) &= n\bar{x}\log\mu + n(1-\bar{x})\log(1-\mu), \\ \partial_\mu J_0(\mu) &= \frac{n\bar{x}}{\mu} - \frac{n(1-\bar{x})}{1-\mu}, \end{aligned}$$

so the MAP estimator coincides with the MLE given by $\hat{m}_0 = \mathrm{Be}(\bar{x})$. By the other hand, under $\mathcal{T}_1$ the induced prior over Θ is given by $p_1(\theta) = \frac{1}{2\theta^{1/2}}\mathbf{1}_{\{\theta\in[0,1]\}}(\theta)$. Analogously, as $\mu = \theta^{1/2}$, then

$$\begin{aligned} J_1(\mu) &= c + (n\bar{x}-1)\log\mu + n(1-\bar{x})\log(1-\mu), \\ \partial_\mu J_1(\mu) &= \frac{n\bar{x}-1}{\mu} - \frac{n(1-\bar{x})}{1-\mu}, \\ \hat{m}_1 &= \mathrm{Be}\left(\frac{n\bar{x}-1}{n-1}\right). \end{aligned}$$

Finally, under $\mathcal{T}_2$ we have that $\mu = \theta^2$, the induced prior over Θ is $p_2(\theta) = 2\theta\mathbf{1}_{\{\theta\in[0,1]\}}(\theta)$ and

$$\begin{aligned} J_2(\mu) &= c + (n\bar{x}+1/2)\log\mu + n(1-\bar{x})\log(1-\mu), \\ \partial_\mu J_2(\mu) &= \frac{n\bar{x}+1/2}{\mu} - \frac{n(1-\bar{x})}{1-\mu}, \\ \hat{m}_2 &= \mathrm{Be}\left(\frac{2n\bar{x}+1}{2n+1}\right). \end{aligned}$$

Thus the MAP estimate depends on the parametrisation, but the MLE does not suffer from these issues since the likelihood is a function, not a probability density, and satisfies the invariance property [83, Theorem 7.2.1]. As discussed below, Bayes estimators do not suffer from these problems either, since the change of measure is taken into account when integrating over the parameter space.

1.2 Bayes Estimators

Going back to the general case, given the model space $\mathcal{M}$, a loss function $L : \mathcal{M}\times\mathcal{M} \to \mathbb{R}$ is a non-negative functional. We interpret $L(m_0, \bar{m})$ as the cost of selecting model $\bar{m} \in \mathcal{M}$ when the true model is $m_0 \in \mathcal{M}$. With a loss function and the posterior distribution over models, we define the Bayes risk (or expected loss[1]) $R(\bar{m}|D)$ and the Bayes estimator $\hat{m}_L$ as follows:

$$R_L(\bar{m}|D) \;:=\; \int_{\mathcal{M}} L(m,\bar{m})\Pi_n(\mathrm{d}m)\,, \tag{1.3}$$

[1]In the literature, Bayes risk refers to the expected loss w.r.t. a fixed measure, but in our context, it is implicitly that the expectations and estimators are w.r.t. the posterior measure Π_n.

$$\hat{m}_L \quad \in \quad \underset{\bar{m}\in\mathcal{M}}{\operatorname{argmin}} R_L(\bar{m}|D). \tag{1.4}$$

In the parametric setting, any loss function L induces a functional l defined on $\Theta \times \Theta$ (and vice versa) by $l(\theta_0, \bar{\theta}) = L(m_{\theta_0}, m_{\bar{\theta}})$, interpreted as the cost of choosing parameter $\bar{\theta}$ when the true parameter is θ_0. The Bayes risk [15] of $\bar{\theta} \in \Theta$ and its Bayes estimator $\hat{\theta}_l$ are defined by

$$R_l(\bar{\theta}|D) \quad := \quad \int_\Theta l(\theta, \bar{\theta}) p_n(\mathrm{d}\theta) = \int_\mathcal{M} L(m, \bar{m}) \Pi_n(\mathrm{d}m), \tag{1.5}$$

$$\hat{\theta}_l \quad \in \quad \underset{\bar{\theta}\in\Theta}{\operatorname{argmin}} R_l(\bar{\theta}|D), \tag{1.6}$$

where $\Pi_n(\mathrm{d}m) = \Lambda_n(m)\Pi(\mathrm{d}m)$, with the prior distribution $\Pi = \mathcal{T}(p)$.

For illustration, consider the 0-1 loss defined as $l_{0-1}(\theta, \bar{\theta}) = 1 - \delta_{\bar{\theta}}(\theta)$. It yields $R_{l_{0-1}}(\bar{\theta}|D) = 1 - p(\bar{\theta}|D)$, that is, the corresponding Bayes estimator is the posterior mode, i.e. $\hat{\theta}_{l_{0-1}} = \hat{\theta}_{MAP}$. For continuous-valued quantities the use of a quadratic loss $l_2(\theta, \bar{\theta}) = \|\theta - \bar{\theta}\|^2$ is often preferred, and its Bayes estimator is the posterior mean $\hat{\theta}_{l_2} = \int_\Theta \theta p(\mathrm{d}\theta|D)$. In one dimensional parameter space, the absolute loss $l_1(\theta, \bar{\theta}) = |\theta - \bar{\theta}|$ yields the posterior median estimator [131].

Using general Bayes estimators on parametrised models enables for a richer choice of criteria for model selection by integrating global information of the parameter space while providing a measure of uncertainty through the Bayes risk value. However, this approach might also neglect parametrisation related issues, such as overparametrisation of the model space (we say that $\mathcal{T}$ overparametrises $\mathcal{M}$ if $m_\theta : \Theta \to \mathcal{M}$ is not one-to-one). The latter might result in a multimodal posterior distribution over parameters. For example, take $\mathcal{X} = \Theta = \mathbb{R}$, $m_0 = \mathcal{N}(x; \mu, 1)$ and $\mathcal{T}(\theta) = \mathcal{N}(x|\theta^2, 1)$. If we choose a symmetric prior $p(\theta)$, e.g. $p(\theta) = \mathcal{N}(\theta|0, 1)$, then with enough data, the posterior distribution is symmetric with modes near $\{\mu, -\mu\}$, so both l_1 and l_2 estimators are close to 0.

1.3 Fréchet Means

To address the above issues, we propose using parameter-free selection criteria via loss functions that compare directly distributions instead of their parameters. Since both L and Π_n operate directly on the model space, model learning according to the above equations does not depend on geometric aspects of parameter spaces. Moreover, this point of view allows us to define loss functions in terms of various metrics/divergences directly on the space $\mathcal{P}(\mathcal{X})$, and therefore to enhance the classical Bayesian estimation framework.

The next result, proved and extended in Chapter 5, illustrates the fact that many Bayesian estimators, including the *model average estimator*, correspond to finding a so-called Fréchet mean or barycenter [93] under a suitable metric/divergence on probability measures. Let $\mathcal{M} = \mathcal{P}_{ac}(\mathcal{X})$ and consider the L_2 loss function $L_2(m, \bar{m}) = \frac{1}{2}\int_\mathcal{X} (m(x) - \bar{m}(x))^2 \lambda(\mathrm{d}x)$, then

the corresponding Bayes estimator coincides with the *Bayesian model average*:

$$\bar{m}(x) := \mathbb{E}_{\Pi_n}[m] = \int_{\mathcal{M}} m(x)\Pi_n(\mathrm{d}m).$$

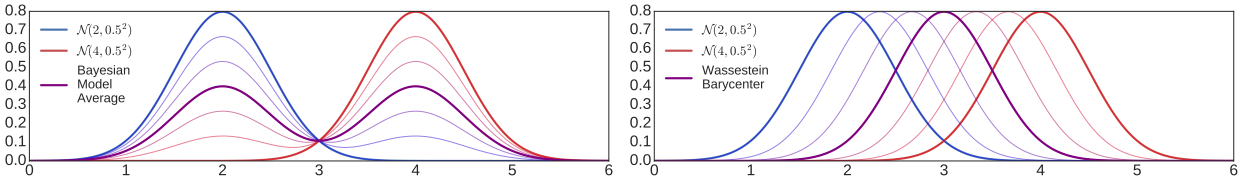

Figure 1.1: Model average (left) and Wasserstein barycenter (right) of two Gaussian densities.

An inconvenient about model average is that it does not always preserve properties of the original model space. E.g. if the posterior distribution is equally concentrated on two different models $m_0 = \mathcal{N}(\mu_0, 1)$ and $m_1 = \mathcal{N}(\mu_1, 1)$ with $\mu_0 \neq \mu_1$, i.e. both models are unimodal (Gaussian) with unit variance, the Bayesian model average is in turn a bimodal (non-Gaussian) distribution with variance strictly greater than 1. More generally, the model average might yield intractable representations or be hardly interpretable in terms of the prior and parameters.

An alternative is to consider different loss functions for eq. (1.3), e.g. the well-known Wasserstein distance, arising in optimal transport theory (see [138, 139] for delve in this field). In Chapter 5 of this work, we will develop the theory of the corresponding Bayes estimators, which coincides with the *Wasserstein barycenters* (see [2, 96, 65, 74]). For now, for the reader's convenience, we illustrate this estimator applying a simple result to the above Gaussian example: for $m_0 = \mathcal{N}(\mu_0, 1)$ and $m_1 = \mathcal{N}(\mu_1, 1)$, the so-called 2-Wasserstein barycenter distribution is the Gaussian distribution with unitary variance given by $\hat{m} = \mathcal{N}(\frac{\mu_0+\mu_1}{2}, 1)$. In Fig. 1.1 we illustrate the Bayes estimators, and interpolations, between two Gaussian densities using L_2 and W_2 loss functions, studied with more detail in Chapter 5.

In Chapter 2, we will introduce the model known as *Gaussian process*, to then delve into more general models. Otherwise, if the reader wishes to delve directly into the theory of Bayesian estimation of models, he can go directly to Chapter 5.

Chapter 2

Gaussian Processes for Regression

> *"Experimentalists think that it is a mathematical theorem while the mathematicians believe it to be an experimental fact."*
>
> – Gabriel Lippmann to Henri Poincaré, about Gaussian distribution

The Gaussian distribution is one of the most studied mathematical objects in probabilities and statistics, if not the most, where its application is universal and multidisciplinary, both in natural and social sciences. The multivariate distribution of a jointly-Gaussian random vector $\mathbf{x} \in \mathbb{R}^n$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ has a density function given by

$$\mathcal{N}_n\left(\mathbf{x}|\mu, \Sigma\right) \quad = \quad \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)},$$

where $|\Sigma|$ denotes the determinant of Σ. In Fig. 2.1 we show a bivariate example of this density.

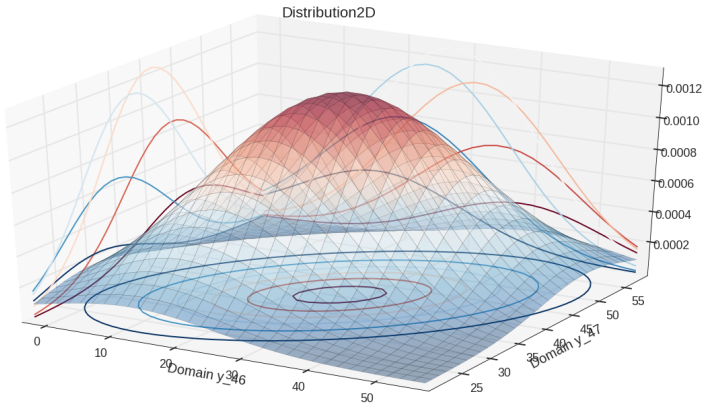

Figure 2.1: An example of a multivariate Gaussian density in $\mathbb{R}^2$.

Several models rely on the Gaussian distribution, even when the data are known to be non-Gaussian but Gaussianity is assumed to avoid the computational complexity related to more realistic models—see, e.g. the use of the Kalman filter in the Apollo missions [13].

Two main reasons for the extensive use of the Gaussian distribution in science can be identified: one is conjectural, and the other one is practical. The first reason obeys the simplifying assumptions in mathematical modelling since observed data comprise multiple error-corrupted phenomena, an exact description of these real-world data-generating engines is challenging—if not impossible. Therefore, we partially model the data using first principles to then describe the remaining components as several sources of uncertainty added together, i.e., the *noise*. Then, based on the central limit theorem [9], we can define this so-called noise in statistical terms by a Gaussian distribution.

The second reason is the appealing mathematical properties of the Gaussian distribution, in particular, for Bayesian inference and learning [102]. Gaussian random variables (RVs) are closed under *conditioning* and *marginalisation*, i.e., all marginal and conditional distributions of a set of jointly-Gaussian RVs are Gaussian; this allows for tractable inference. Additionally, Gaussian distributions are conjugate for themselves, meaning that a Gaussian prior and a Gaussian likelihood result in a Gaussian posterior distribution. This closed-form posterior allows for (i) efficient gradient-based learning via optimisation, and (ii) exact Bayesian inference.

In next section we will introduce the regression problem, one of the main tasks in machine learning, to then construct a Gaussian process, i.e. a model for infinitely-many jointly-Gaussian random variables, under different viewpoints and interpretations.

2.1 The Regression Problem

In several fields, such as finance, physics and engineering, we can find settings where the observations are indexed by time or space and convey some hidden dependence structure that we aim to discover. This setting corresponds to a regression problem, previously introduced in Section 1.1.1, that can be summarised as follow: given $N \in \mathbb{N}$ observations $(\mathbf{t}, \mathbf{x}) = \{(t_{\mathrm{i}}, x_{\mathrm{i}})\}_{\mathrm{i}=1}^{N}$ where $t_{\mathrm{i}} \in \mathcal{T} \subseteq \mathbb{R}^T$, $T \in \mathbb{N}$ and $x_{\mathrm{i}} \in \mathcal{X} \subseteq \mathbb{R}$ for $\mathrm{i} = 1, \ldots, n$ the regression problem aims to estimate some predictor $f : \mathcal{T} \to \mathcal{X}$, such that $f(t_{\mathrm{i}})$ is *close* to x_{i}, where the terms *best* and *close* are given by the chosen criterion of optimality. For solving this regression problem, we desire a model to be able to interpolate and extrapolate, calculate point estimations, error bars and generate plausible functions, as in Fig. 2.2. A widely used solution to this regression problem is the Gaussian process [102], also know as kriging [129, 29], which is a case of Bayesian nonparametric model. On following section we introduce a general Bayesian nonparametric framework for regression.

2.2 Bayesian Nonparametric Models

An important aspect of Bayesian modelling is the useful concept of hierarchies. Given a parametrised model space $\mathcal{M} = \mathcal{T}(\Theta)$, consider $p(\theta) \in \mathcal{P}(\Theta)$ a prior distribution over parameters. If this prior is, in turn, parametrised by $\omega \in \Omega$, named *hyperparameters*, then we can also set a *hyperprior* $p(\omega) \in \mathcal{P}(\Omega)$ over these. In principle, one can iterate this process: if

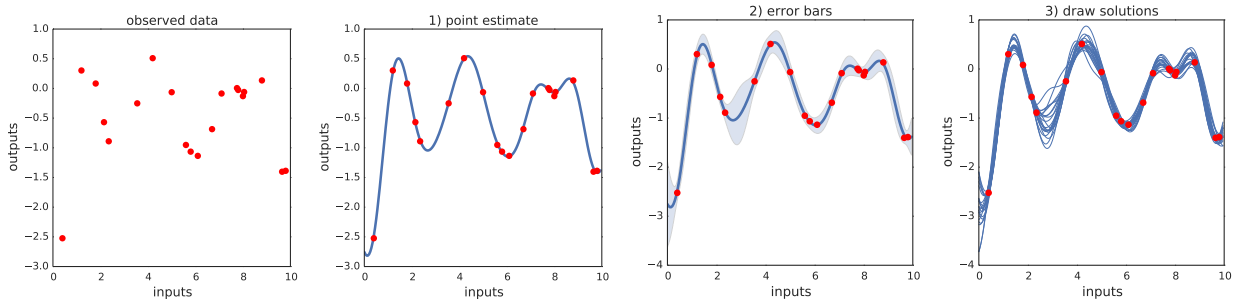

Figure 2.2: Data, point estimations, error bars and draw solution from a regression problem.

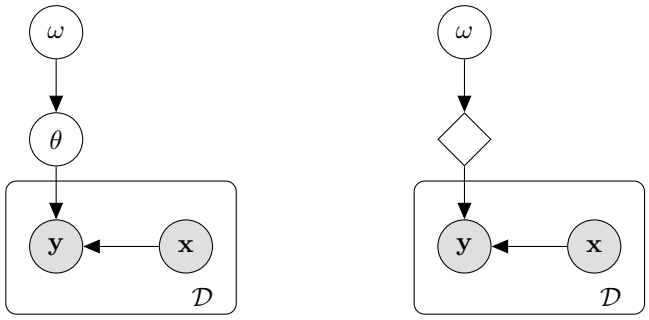

Figure 2.3: Left: The graphical representation of a Bayesian hierarchical model for regression, with *hyperparameters* ω, *parameters* θ and input/output data $\mathbf{x}, \mathbf{y}$. Right: The graphical representation of the same Bayesian hierarchical model, but where we integrate out the parameter θ.

the hyperprior itself has parameters, these may be called hyper-hyperparameters, and so forth. However, at some point, we must stop. A *Bayesian hierarchical model* is written in multiple stages or levels, where all uncertainty is modelled in probabilistic terms and is allowed to use the Bayes rule between stages.

In Fig. 2.3 (left), a regression scheme of 2-stages is presented, where $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ is input/output data and the joint distribution is $p(\mathbf{y}, \mathbf{x}, \theta, \omega) = p(\mathbf{y}, \mathbf{x}|\theta, \omega)\, p(\theta|\omega)\, p(\omega)$. As we mentioned earlier in Section 1.1.1, $p(\mathbf{y}, \mathbf{x}|\theta, \omega)$ is the generative model, and since in the regression context we often only need the discriminative model $p(\mathbf{y}|\mathbf{x}, \theta, \omega)$, we set an *uninformative* prior over $\mathbf{x}$, i.e. $p(\mathbf{x}|\theta, \omega) \propto 1$. With this setting, the posterior distribution of parameters θ given $\mathbf{x}, \mathbf{y}, \omega$ is

$$p(\theta|\mathbf{y}, \mathbf{x}, \omega) = \frac{p(\mathbf{y}|\mathbf{x}, \theta, \omega)\, p(\theta|\omega)}{p(\mathbf{y}|\mathbf{x}, \omega)},$$

where $p(\mathbf{y}|\mathbf{x}, \theta, \omega)$ is the likelihood of θ, $p(\theta|\omega)$ is the prior of θ and the marginal likelihood is

$$p(\mathbf{y}|\mathbf{x}, \omega) = \int p(\mathbf{y}|\mathbf{x}, \theta, \omega)\, p(\mathrm{d}\theta|\omega)\,.$$

For a fixed hyperparameter ω, we can train the θ-parametrised model with any Bayes estimator, as maximum a posteriori $\hat{\theta}_{MAP} \in \operatorname{argmax}_{\theta\in\Theta} p(\theta|\mathbf{y}, \mathbf{x}, \omega)$ or posterior mean $\hat{\theta}_{l_2} = \int \theta p(\mathrm{d}\theta|\mathbf{y}, \mathbf{x}, \omega)$. Additionally, in the Bayesian hierarchical models context we calculate the so-called *posterior predictive distribution* of $\bar{\mathbf{y}}$ for new inputs $\bar{\mathbf{x}}$, that is given by

$$p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{y}, \mathbf{x}, \omega) = \int p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \theta, \omega)\, p(\mathrm{d}\theta|\mathbf{y}, \mathbf{x}, \omega)\,,$$

that coincides with the Bayesian model average (on θ) of the discriminative model, i.e.

$$\hat{m}(\bar{\mathbf{y}}|\bar{\mathbf{x}},\omega) = \int m_\theta(\bar{\mathbf{y}}|\bar{\mathbf{x}},\omega)p(\mathrm{d}\theta|\mathbf{x},\mathbf{y},\omega) = \int p\left(\bar{\mathbf{y}}|\bar{\mathbf{x}},\theta,\omega\right)p(\mathrm{d}\theta|\mathbf{y},\mathbf{x},\omega) = p\left(\bar{\mathbf{y}}|\bar{\mathbf{x}},\mathbf{y},\mathbf{x},\omega\right). \tag{2.1}$$

Note that the *Bayesian model average* $\hat{m}(\bar{\mathbf{y}}|\bar{\mathbf{x}},\omega)$ depends only on ω, since we *integrate out* the parameter θ, so we also must *choose* the *hyperparameter* ω. For this, following the Bayesian paradigm illustrated in Figure 2.3 (right), we calculate the posterior distribution of ω given $\mathbf{x},\mathbf{y}$ as

$$p\left(\omega|\mathbf{y},\mathbf{x}\right) = \frac{p\left(\mathbf{y}|\mathbf{x},\omega\right)p\left(\omega\right)}{p\left(\mathbf{y}|\mathbf{x}\right)},$$

where $p\left(\mathbf{y}|\mathbf{x},\omega\right)$ is the *likelihood* of ω (matching with the marginal likelihood related to θ), $p\left(\omega\right)$ the prior of ω and $p\left(\mathbf{y}|\mathbf{x}\right)$ the marginal likelihood related to ω. Given $p\left(\omega|\mathbf{y},\mathbf{x}\right)$, we also can train the ω-parametrised model with any Bayes estimator. In *hyperparameter* stage, it is usual to use a maximum a posteriori estimator $\hat{\omega}_{MAP}$, denoted MAP-II to differentiate it from the parameter MAP estimator $\hat{\theta}_{MAP}$, which is denoted MAP-I. If we want a different Bayes estimator, like a *hyperparameter model average*, usually the integrals with respect to $p\left(\omega|\mathbf{y},\mathbf{x}\right)$ are intractable, but can be approximated by sampling the distribution using Markov Chain Monte Carlo (MCMC) methods [24].

The introduced framework is widely used to define *nonparametric* models, i.e. these models without a fixed number of parameters that grow up with the data. We can highlight one nonparametric model that has many mathematical properties that make it very versatile and flexible, especially for regression tasks, which is based on the *Gaussian* distribution.

2.3 Constructing a Gaussian Process

Consider the following Gaussian-based Bayesian hierarchical linear model:

$$\begin{aligned} f_\theta(t) &= \langle t,\theta\rangle, \text{ for } \theta\in\mathbb{R}^T \\ \omega &= \Sigma_\theta \in \mathbb{R}^{T\times T} \\ p(\theta|\omega) &= \mathcal{N}_T\left(0,\Sigma_\theta\right) \\ p\left(\mathbf{x}|\mathbf{t},\omega,\theta\right) &= \mathcal{N}_n\left(\mathbf{t}^\top\theta,\sigma^2 I_n\right), \end{aligned}$$

where we assume an uninformative prior over ω, i.e. $p(\omega)\propto 1$. Given n observations denoted as $(\mathbf{t},\mathbf{x})\in\mathcal{T}^n\times\mathcal{X}^n\subset\mathbb{R}^{T\times n}\times\mathbb{R}^n$, the posterior density of θ is the closed-form Gaussian

$$p(\theta|\mathbf{t},\mathbf{x},\omega) = \mathcal{N}_{\bar{n}}\left(\Sigma_\theta^{\mathbf{t}}\mathbf{t}\mathbf{x},\sigma^2\Sigma_\theta^{\mathbf{t}}\right)$$

where $\Sigma_\theta^{\mathbf{t}} := \left[\mathbf{t}\mathbf{t}^\top+\sigma^2\Sigma_\theta^{-1}\right]^{-1}$. It is straightforward that, given new inputs $\bar{\mathbf{t}}\in\mathcal{T}^{\bar{n}}$, the posterior predictive distribution of $\bar{\mathbf{f}}=\bar{\mathbf{t}}^\top\theta\in\mathcal{X}^{\bar{n}}$ is also a closed-form Gaussian given by

$$p(\bar{\mathbf{f}}|\bar{\mathbf{t}},\mathbf{t},\mathbf{x},\omega) = \mathcal{N}_T\left(\bar{\mathbf{t}}^\top\Sigma_\theta^{\mathbf{t}}\mathbf{t}\mathbf{x},\sigma^2\bar{\mathbf{t}}^\top\Sigma_\theta^{\mathbf{t}}\bar{\mathbf{t}}\right).$$

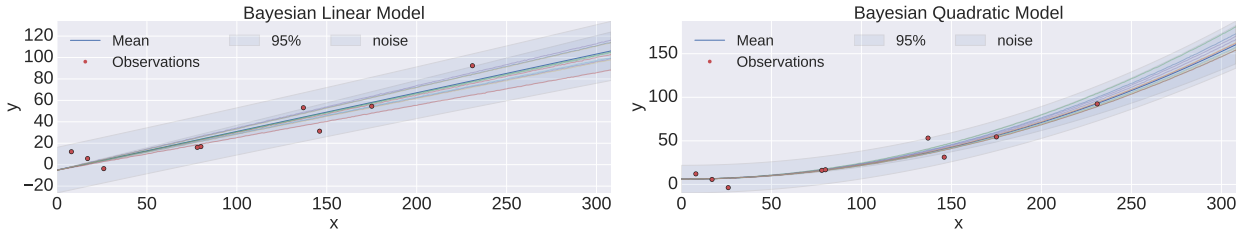

Figure 2.4: Left: Posterior distribution of a Bayesian linear model. Right: Posterior distribution of a Bayesian quadratic model. Both posteriors are given the same observations

The above results show an interesting case of a hierarchical model with closed-form *model average*. Although the model is linear, it is possible to extend it as a non-linear model. Given a function $\phi : \mathcal{T} \to \mathcal{S} \subseteq \mathbb{R}^S$, where $\mathcal{S}$ is known as the *feature* space, consider the model

$$\begin{aligned} f_\theta(t) &= \langle \phi(t), \theta \rangle, \text{ for } \theta \in \mathbb{R}^S \\ \omega &= \Sigma_\theta \in \mathbb{R}^{S \times S} \\ p(\theta|\omega) &= \mathcal{N}_S(0, \Sigma_\theta) \\ p(\mathbf{x}|\mathbf{t}, \omega, \theta) &= \mathcal{N}_n\left(\phi(\mathbf{t})^\top \theta, \sigma^2 I_n\right). \end{aligned}$$

Note the similarity of this model with the linear case, where we supersede $\mathbf{t} \in \mathcal{T}^n$ by $\phi(\mathbf{t}) \in \mathcal{S}^n$, so the posterior of θ and the respective posterior predictive of $\bar{\mathbf{f}} = \phi(\bar{\mathbf{t}})^\top \theta$ are analogous:

$$\begin{aligned} p(\theta|\mathbf{t}, \mathbf{x}, \omega) &= \mathcal{N}_S\left(\Sigma_\theta^{\phi(\mathbf{t})} \phi(\mathbf{t})\mathbf{x}, \sigma^2 \Sigma_\theta^{\phi(\mathbf{t})}\right) \\ p(\bar{\mathbf{f}}|\bar{\mathbf{t}}, \mathbf{t}, \mathbf{x}, \omega) &= \mathcal{N}_{\bar{n}}\left(\phi(\bar{\mathbf{t}})^\top \Sigma_\theta^{\phi(\mathbf{t})} \phi(\mathbf{t})\mathbf{x}, \sigma^2 \phi(\bar{\mathbf{t}})^\top \Sigma_\theta^{\phi(\mathbf{t})} \phi(\bar{\mathbf{t}})\right), \end{aligned}$$

where $\Sigma_\theta^{\phi(\mathbf{t})} := \left[\phi(\mathbf{t})\phi(\mathbf{t})^\top + \sigma^2 \Sigma_\theta^{-1}\right]^{-1}$. In Fig. 2.4 we plot Bayesian linear and quadratic models using the above framework, given the same observation. In each case we plot observations, mean, 0.95 confidence interval for $\bar{\mathbf{f}}$ and $\bar{\mathbf{x}}$, also 10 samples of plausible functions.

To compute predictions using this model, it is necessary to calculate the matrix $\Sigma_\theta^{\phi(\mathbf{t})}$ via inverting an $S \times S$ dimensional matrix, so the complexity grows up with respect to the dimension of the feature space, becoming intractable if S is very large. However, the model can be rewritten in an equivalent way but with a nonparametric interpretation. If we denote $k_\omega(\mathbf{t}, \mathbf{s}) = \phi(\mathbf{t})^\top \Sigma_\theta \phi(\mathbf{s})$, through the *Woodbury matrix inversion lemma* [1], we can write the posterior predictive in terms of function k_ω as

$$\begin{aligned} p(\bar{\mathbf{f}}|\bar{\mathbf{t}}, \mathbf{t}, \mathbf{x}, \omega) &= \mathcal{N}_{\bar{n}}\left(\bar{\mu}, \bar{\Sigma}\right) \\ \bar{\mu} &= k_\omega(\bar{\mathbf{t}}, \mathbf{t})\left[k_\omega(\mathbf{t}, \mathbf{t}) + \sigma^2 I_n\right]^{-1} \mathbf{x} \\ \bar{\Sigma} &= k_\omega(\bar{\mathbf{t}}, \bar{\mathbf{t}}) - k_\omega(\bar{\mathbf{t}}, \mathbf{t})\left[k_\omega(\mathbf{t}, \mathbf{t}) + \sigma^2 I_n\right]^{-1} k_\omega(\mathbf{t}, \bar{\mathbf{t}}). \end{aligned}$$

Unlike the previous formula, to compute predictions with this version it is necessary to calculate and invert the $n \times n$ dimensional matrix $[k_\omega(\mathbf{t}, \mathbf{t}) + \sigma^2 I_n]$, so the complexity grows up with respect to the number of observations, independent of the features space dimension.

[1] $\left[Z + UWV^\top\right]^{-1} = Z^{-1} - Z^{-1}U^\top \left(W^{-1} + V^\top Z^{-1} U\right)^{-1} V Z^{-1}$

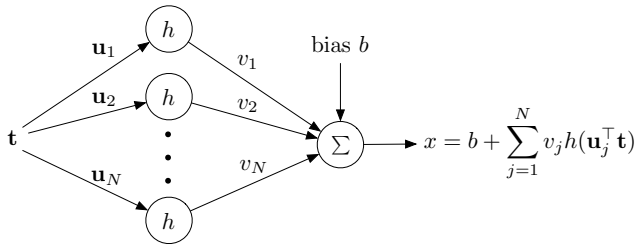

Figure 2.5: Single-layer feedforward neural network: $\mathbf{t}$ is the input, x is the output, $h(\cdot)$ is the activation function, b is the bias, $\mathbf{u}_{j=1:N}$ are the input weights, $v_{j=1:N}$ are the output weights.

Kernel trick is the technique of writing the model only in terms of kernel k_ω avoiding the computation of the map ϕ, allowing us to consider an implicit features space of high dimension, even infinite.

2.4 From Neural Networks to Gaussian Processes

Among neural network practitioners, it is widely believed that the number of neurons should be determined based on the amount of available data. However, as pointed out by C. Williams in [142], this makes little sense from a Bayesian standpoint, where the complexity of the model should be dictated by the complexity of the problem and not by the amount of available data. In this regard, R. Neal demonstrated that the output of a single-layer neural network with random weights converges to Gaussian process when the number of neurons approaches infinity [86].

Following [102, 142, 86], let us consider a single-layer N-neuron neural network as shown in Fig. 2.5. By modelling the bias and weights as independent random variables, the outputs $x_1, x_2, \ldots, x_N$ are also random for any choice of inputs $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N$, with a distribution that is not necessarily tractable due to the nonlinear activation function $h(\cdot)$. Nevertheless, notice that the network in Fig. 2.5 is defined by a sum of i.i.d. terms, therefore, by virtue of the multidimensional central limit theorem (CLT [9]), taking the number of neurons $N \to \infty$ results in the outputs $x_1, x_2, \ldots, x_N$ being jointly Gaussian[2]. This construction can be further extended to the case of an infinite number of outputs, thus yielding the *Gaussian process* [102]. In the following section, we deepen the properties of this model as a stochastic process.

2.5 Stochastic Process Characterisation

We can interpret the model from a *probabilistic* point of view. A $\mathcal{X}$-valued stochastic process $f = \{f_t\}_{t \in \mathcal{T}}$ is a collection of random variables, indexed by $\mathcal{T}$, that takes values in $\mathcal{X}$. While the

[2]The motivation for taking the number of neurons to infinity follows [59], which states that the network in Fig. 2.5 is a universal approximator. Furthermore, the CLT can in fact be applied since the bounded activation function h results in finite variance for the outputs $x_1, x_2, \ldots, x_N$. Notice that scaling the output weights variance $\propto 1/N$ is required for the CLT to hold.

measure-theoretic approach to stochastic processes starts with a probability space, in machine learning the starting point is a collection of finite-dimensional distributions. Given any finite collection of points $t_1, ..., t_n \in \mathcal{T}$, the distribution function of $f_{t_1}, ..., f_{t_n}$ is denoted as $F_{t_1,...,t_n}$. The set $\mathcal{F} = \{F_{t_1,...,t_n} | t_1, ..., t_n \in \mathcal{T}, n \in \mathbb{N}\}$ correspond to their family of finite-dimensional distributions, that satisfy the well-known Kolmogorov consistency conditions:

1. Permutation condition: $F_{t_1,...,t_n}(x_1, ..., x_n) = F_{t_{\pi(1)},...,t_{\pi(n)}}\left(x_{\pi(1)}, ..., x_{\pi(n)}\right)$ for all $t_1, ..., t_n \in \mathcal{T}$, all $x_1, ..., x_n \in \mathcal{X}$ and any n-permutation π.

2. Marginalisation condition: $F_{t_1,...,t_{n+m}}(x_1, ..., x_n, +\infty, ..., +\infty) = F_{t_1,...,t_n}(x_1, ..., x_n)$ for all $t_1, ..., t_{n+m} \in \mathcal{T}$ and all $x_1, ..., x_n \in \mathcal{X}$.

If a family of finite-dimensional distributions $\mathcal{F}$ satisfies the conditions of consistency, then the *Kolmogorov's consistency theorem* [134] allows us to construct a stochastic process $\hat{f} = \left\{\hat{f}_t\right\}_{t\in\mathcal{T}}$ in which the associated family of finite-dimensional distributions $\hat{\mathcal{F}}$ coincides with $\mathcal{F}$. As the law of a stochastic process is completely determined by the associated family of finite dimensional distribution [110], for abuse of notation we refer to $\mathcal{F}$ as its law.

As we will show, the Gaussian distribution satisfies useful appealing properties for our purposes. Let $\mathbf{x} \in \mathcal{X}^n, \bar{\mathbf{x}} \in \mathcal{X}^{\bar{n}}$ be jointly Gaussian distributed random variables as

$$\eta_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{x}, \bar{\mathbf{x}}) = \mathcal{N}_{n+\bar{n}}\left(\begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\bar{\mathbf{x}}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{x}\bar{\mathbf{x}}} \\ \Sigma_{\bar{\mathbf{x}}\mathbf{x}} & \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}} \end{bmatrix}\right).$$

The marginalisation condition is satisfied due to

$$\int_{\mathcal{X}^n} \eta_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{x}, \bar{\mathbf{x}})\, \mathrm{d}\mathbf{x} = \mathcal{N}_{\bar{n}}(\mu_{\bar{\mathbf{x}}}, \Sigma_{\bar{\mathbf{x}},\bar{\mathbf{x}}}) = \eta_{\bar{\mathbf{t}}}(\bar{\mathbf{x}}),$$

and the permutation condition is fulfilled because, given a n-permutation π, there is a permutation matrix P, and since $P^{-1} = P^\top$ it satisfies

$$\eta_{\pi(\mathbf{t})}(\pi(\mathbf{x})) = \frac{1}{(2\pi)^{\frac{n}{2}}\left|P\Sigma P^\top\right|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu)^\top P^\top (P\Sigma P^\top)^{-1} P(\mathbf{x}-\mu)} = \eta_{\mathbf{t}}(\mathbf{x}).$$

Due to its consistency under both marginalisation and permutation, we can extend the finite-dimensional multivariate Gaussian distribution to the infinite-dimensional case through Kolmogorov's consistency theorem. This construction is referred to as the Gaussian process (GP) [102], a prior probability distribution over functions that defines non-linear nonparametric regression models by assuming joint Gaussianity of the observed data.

Definition 2.5.1 A stochastic process $f = \{x_t\}_{t\in\mathcal{T}}$ is a Gaussian process (GP) with mean function $m(\cdot)$ and covariance kernel[3] $k(\cdot,\cdot)$, denoted by $f \sim \mathcal{GP}(m, k)$, if, for any finite collection of points in their domain $\mathbf{t} = [t_1, \ldots, t_n]^\top \in \mathcal{T}^n$, the distribution $\eta_{\mathbf{t}}$ of the vector[4] $\mathbf{x} := f(\mathbf{t}) = [x_{t_1}, \ldots, x_{t_n}]^\top \in \mathcal{X}^n$ follows a multivariate Gaussian distribution with mean vector $\mu_{\mathbf{x}} = [m(t_1), \ldots, m(t_n)]^\top$ and covariance matrix $[\Sigma_{\mathbf{xx}}]_{ij} = k(t_i, t_j)$, i.e. $\eta_{\mathbf{t}} = \mathcal{N}_n(\mu_{\mathbf{x}}, \Sigma_{\mathbf{xx}})$.

[3]Common covariance functions are square exponential, rational quadratic, Matérn, and polynomial [102].
[4]By abuse of notation, we identify the random vector $f(\mathbf{t})$ as $\mathbf{x}$, which denote the observations on $\mathbf{t}$.

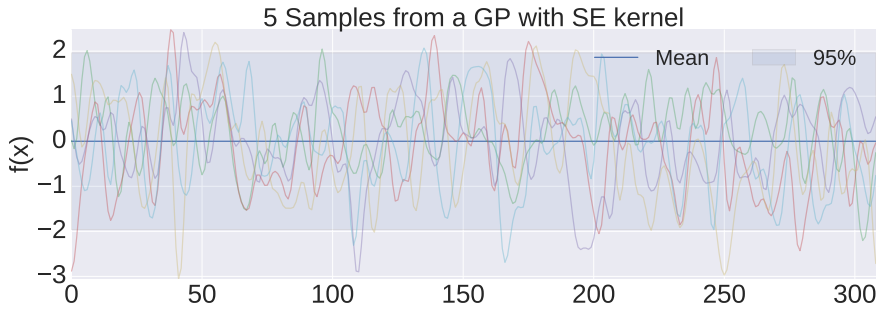

Figure 2.6: Example of a GP with zero mean and SE kernel as prior over function. In this plot, we show the prior mean, the 0.95 confidence interval and 5 samples.

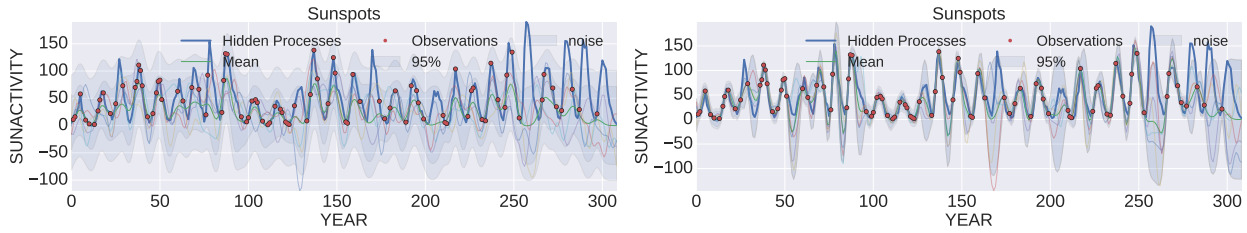

Figure 2.7: The posterior distribution of the GP. Left: Non-trained GP. Right: Trained GP.

A Gaussian process is completely determined by its mean m and covariance k functions and it is used on *machine learning* as an *a priori* distribution over functions. The parameters of m and k are referred to as *hyperparameters* of the GP. In Fig. 2.6 we plot an example of a GP with the commonly used covariance function, named square exponential (SE) kernel given by

$$k_{SE}(x, \bar{x}) = \sigma^2 \exp\left(-\frac{(x-\bar{x})^2}{l^2}\right), \text{with } \sigma^2 > 0, l > 0 \text{ the } \textit{hyperparameters} \text{ of the GP.}$$

Performing inference on new inputs[5] $\bar{\mathbf{t}}$ rests on calculating the posterior distribution of $\bar{\mathbf{x}}$ given observations $\mathbf{x}$, which is also Gaussian and has distribution

$$\eta_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{x}}|\mathbf{x}) = \mathcal{N}\left(\bar{\mathbf{x}}|\mu_{\bar{\mathbf{x}}|\mathbf{x}}, \Sigma_{\bar{\mathbf{x}}|\mathbf{x}}\right),$$

where $\mu_{\bar{\mathbf{x}}|\mathbf{x}} = \mu_{\bar{\mathbf{x}}} + \Sigma_{\bar{\mathbf{x}}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}})$ and $\Sigma_{\bar{\mathbf{x}}|\mathbf{x}} = \Sigma_{\bar{\mathbf{x}}\bar{\mathbf{x}}} - \Sigma_{\bar{\mathbf{x}}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}\Sigma_{\mathbf{x}\bar{\mathbf{x}}}$ are referred to as conditional mean and variance respectively; these statistics allow for computing point estimates, confidence bands and sample functions directly. In Fig. 2.7 (Left) we show the posterior distribution of a GP with SE kernel, given observations from sunset activity data.

The kernel is usually chosen heuristically based on expertise and the prior know-how of modelled phenomenon. In Fig. 2.8 we consider perform inference with the over same 4 observations and three different kernels:

- Ornstein-Uhlenbeck: $k_{OU}(x, \bar{x}) = \sigma^2 \exp\left(-\frac{|x-\bar{x}|}{2l^2}\right)$

[5] As long as there is no ambiguity in the choice of points $\mathbf{t}$, we will denote $x(\mathbf{t})$ as $\mathbf{x}$, $m(\mathbf{t})$ as $\mu_{\mathbf{x}}$ and $k(\mathbf{t}, \mathbf{t})$ as $\Sigma_{\mathbf{x}}$. For a second collection of input points $\bar{\mathbf{t}}$ the notation is analogue: the process evaluation is $\bar{\mathbf{x}} = x(\bar{\mathbf{t}})$, the mean is $\mu_{\bar{\mathbf{x}}} = m(\bar{\mathbf{t}})$ and the cross-covariance between $\mathbf{x}$ and $\bar{\mathbf{x}}$ is $\Sigma_{\mathbf{x}\bar{\mathbf{x}}} = k(\mathbf{t}, \bar{\mathbf{t}})$.

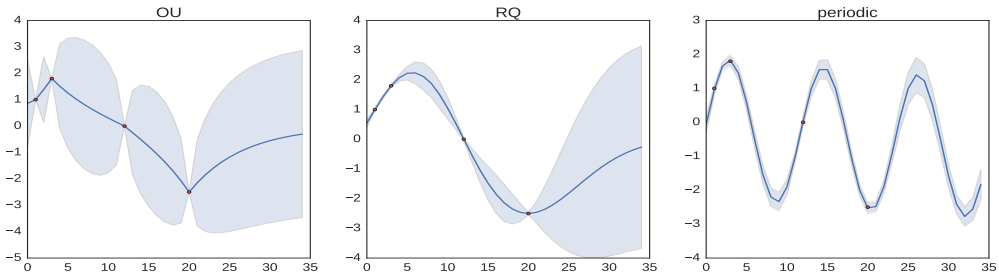

Figure 2.8: The posterior distribution of GPs with different kernels and same observations. Left: Ornstein-Uhlenbeck. Center: Rational Quadratic. Right: Locally Periodic.

- Rational Quadratic: $k_{RQ}(x, \bar{x}) = \sigma^2 \left(1 + \frac{|x-\bar{x}|^2}{2\alpha l^2}\right)^{-\alpha}$
- Locally Periodic: $k_{per}(x, \bar{x}) = \sigma^2 \exp\left(-\frac{|x-\bar{x}|^2}{2l^2}\right) \exp\left(-\frac{2\sin^2(\pi|x-\bar{x}|/p)}{l^2}\right)$

Learning, given observations $(\mathbf{t}, \mathbf{x})$, is equivalent to finding $k(\cdot,\cdot)$ and $m(\cdot)$, usually finitely-parameterised by $\theta = (\theta_k, \theta_m) \in \mathbb{R}^p$, which is usually achieved through of minimisation of the negative logarithm of their marginal likelihood (NLL) given by

$$-\log \eta_{\mathbf{t}}(\mathbf{x}|\theta) = \frac{n}{2}\log(2\pi) + \frac{1}{2}(\mathbf{x} - \mu_{\mathbf{x}})^\top \Sigma_{\mathbf{xx}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}) + \frac{1}{2}\log|\Sigma_{\mathbf{xx}}|, \tag{2.2}$$

where $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{xx}}$ are the mean and covariance of $\mathbf{x}$ given parameters $\theta = (\theta_k, \theta_m)$. The most used optimisation methods are the gradient-based quasi-Newton BFGS method and free-derivative Powell's method. In Fig. 2.7 we show a GP with SE kernel, given observations from sunset activity data, where the left plot have default hyperparameters while the right plot has NLL-based trained hyperparameters. In the trained case, the mean is closer to the real (hidden) signal, and the confidence interval is tighter, so the prediction has less uncertainty.

Chapter 3

Compositionally-Warped Gaussian Processes

"...all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind."

– George Box

The results presented in this Chapter correspond to them in two published papers [107, 108]: 1) *Gonzalo Rios and Felipe Tobar. Learning non-Gaussian time series using the Box-Cox Gaussian process. In International Joint Conference on Neural Networks, 2018*, and 2) *Gonzalo Rios and Felipe Tobar. Compositionally-warped Gaussian processes. Neural Networks, 118:235-246, 2019.*

Despite the facts in favour of the Gaussian distribution presented in Chapter 2, the assumption of joint Gaussianity is far from reality in several settings. In practice, one deals with observations that are non-symmetric, heavy-tailed, or bounded by a physical or economic restriction; all of these properties are contradictory with the Gaussian framework. For instance, under the presence of strictly-positive observations, e.g. prices of a currency or the streamflow of a river, assuming Gaussianity is a mistake; since the Gaussian distribution is supported on the entirety of the real line. This fact motivates us to study models that have the appeal properties that the Gaussian processes, but that are more flexible in the hypotheses over modelled phenomena.

To model non-Gaussian data while still making use of the advantages of Gaussian models, one can transform the observed data $\mathbf{y} \in \mathcal{Y}^N$ via a non-linear differentiable bijection $\varphi : \mathcal{Y} \to \mathcal{X}$,

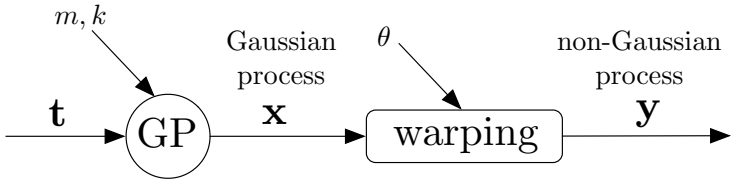

Figure 3.1: General structure of warped Gaussian processes where a GP is nonlinearly transformed to model non-Gaussian observations.

referred to as *warping*, such that $\mathbf{x} = \Phi(\mathbf{y}) = [\varphi(y_1), ..., \varphi(y_N)]^\top$ is *more Gaussian* and thus can modelled as a GP–see Fig. 3.1. This approach is standard in statistics, where a common choice for such a map is $\varphi(y) = \log(y)$, where the implicit assumption is that the observed process has log-normal marginals, so the modelled phenomenon take positive values.

As the transform Φ is diagonal, i.e. in a coordinate-wise manner, the transformed distributions satisfy the conditions of Kolmogorov's consistency theorem [134] (introduced in Section 2.5), such a generative model is a non-Gaussian process named warped Gaussian process [125]. In Section 4.3 we will prove this proposition in a more general approach, so we take it for granted for the rest of this chapter.

We aim to construct a novel warping for Gaussian processes that inherits the expressiveness of deep structures but at the same time require minimal numerical approximations for prediction; this will be attained by constructing warpings with known closed-form inverse.

3.1 The Change of Variables Theorem

A standard approach to model non-Gaussian observations is to transform the data using, e.g., the logarithmic [17] or hyperbolic tangent [60] functions, so that the transformed data are (closer to being) normally distributed. This transformation results in a change of probability measure [134], where the distribution of the transformed variable is known explicitly given the transformation. However, this result and its theoretical implications in the construction of expressive non-Gaussian models are usually neglected. We will now formally present the change of probability measure resulting from transforming a random variable via the following theorem and then study the Gaussian case.

Theorem 3.1.1 (Probability change of variables [58]) Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ be a random vector with a probability density function given by $p_{\mathbf{x}}(\mathbf{x})$, and let $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^n$ be a random vector such that $\varphi(\mathbf{y}) = \mathbf{x}$, where the function $\varphi : \mathcal{Y} \to \mathcal{X}$ is bijective of class $\mathcal{C}^1$ and $|\nabla\varphi(\mathbf{y})| > 0$ $\forall \mathbf{y} \in \mathcal{Y}$. Then, the probability density function $p_{\mathbf{y}}(\cdot)$ induced in $\mathcal{Y}$ is given by

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\varphi(\mathbf{y}))\,|\nabla\varphi(\mathbf{y})|\,,$$

where $\nabla\varphi(\cdot)$ denotes the Jacobian of $\varphi(\cdot)$, and $|\cdot|$ denotes the determinant operator.

We refer to $\mathbf{x} = [x_1, ..., x_n]^\top$ as the *base* variables and to $\mathbf{y} = [y_1, ..., y_n]^\top$ as the *transformed* variables. The change of variables theorem gives a principled methodology to express the probability density function (pdf) of the transformed variables in terms of (i) the pdf of the base variables and (ii) the applied transformation.

As our aim is to use the change of variables theorem to construct non-Gaussian tractable models, let us consider a multivariate normal random vector $\mathbf{x} \in \mathbb{R}^n$ with mean $\mu_{\mathbf{x}}$ and covariance $\Sigma_{\mathbf{x}}$, denoted by $\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, and a coordinate-wise[1] mapping from the transformed space to the base space given by

$$\mathbf{y} \mapsto \mathbf{x} = \varphi(\mathbf{y}) = [\varphi(y_1), ..., \varphi(y_n)]^\top.$$

[1]To simplify the notation we refer to both the vector or scalar maps indistinctly as φ.

Notice that the Jacobian of $\varphi(\mathbf{y})$ is diagonal and therefore its determinant factorises as

$$|\nabla\varphi(\mathbf{y})| = \prod_{\mathrm{i}=1}^{n} \frac{\mathrm{d}\varphi(y_{\mathrm{i}})}{\mathrm{d}y} > 0.$$

In this setting, the pdf of $\mathbf{y} = [y_1, \ldots, y_N]^\top$ can be obtained explicitly through Theorem 3.1.1 and takes the form

$$p(\mathbf{y}) = \prod_{\mathrm{i}=1}^{n} \frac{\mathrm{d}\varphi(y_{\mathrm{i}})}{\mathrm{d}y} \mathcal{N}(\varphi(\mathbf{y})|\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}),$$

where the function φ is affine if and only if the distribution $p(\mathbf{y})$ is Gaussian. Crucially, the distribution $p(\mathbf{y})$ is not Gaussian in general, but it is parametrised by the base mean $\mu_{\mathbf{x}}$, the base variance $\Sigma_{\mathbf{x}}$ and the transformation φ.

Theorem 3.1.1 can also be used to calculate conditional densities of transformed Gaussian random vectors: For two jointly-Gaussian vectors $\mathbf{x}, \mathbf{x}'$ with conditional density $p(\mathbf{x}|\mathbf{x}') = \mathcal{N}\left(\mu_{\mathbf{x}|\mathbf{x}'}, \Sigma_{\mathbf{x}|\mathbf{x}'}\right)$, and a pair of vectors $\mathbf{y}, \mathbf{y}'$ such that $\mathbf{x} = \varphi(\mathbf{y})$ and $\mathbf{x}' = \varphi(\mathbf{y}')$, the conditional density $p(\mathbf{y}|\mathbf{y}')$ is given by

$$\begin{aligned} p(\mathbf{y}|\mathbf{y}') &= \prod_{\mathrm{i}=1}^{n} \frac{\mathrm{d}\varphi(y_{\mathrm{i}})}{\mathrm{d}y} \mathcal{N}\left(\varphi(\mathbf{y})|\mu_{\mathbf{x}|\mathbf{x}'}, \Sigma_{\mathbf{x}|\mathbf{x}'}\right) \\ \mu_{\mathbf{x}|\mathbf{x}'} &= \mu_{\mathbf{x}} + \Sigma_{\mathbf{x}\mathbf{x}'}\Sigma_{\mathbf{x}'\mathbf{x}'}^{-1}\left(\varphi(\mathbf{y}') - \mu_{\mathbf{x}'}\right) \\ \Sigma_{\mathbf{x}|\mathbf{x}'} &= \Sigma_{\mathbf{x}\mathbf{x}} - \Sigma_{\mathbf{x}\mathbf{x}'}\Sigma_{\mathbf{x}'\mathbf{x}'}^{-1}\Sigma_{\mathbf{x}'\mathbf{x}}, \end{aligned}$$

where recall that $\Sigma_{\mathbf{x}\mathbf{x}'}$ denotes the covariance between $\mathbf{x}$ and $\mathbf{x}'$, and $\mu_{\mathbf{x}}$ denotes the marginal mean of $\mathbf{x}$.

Observe that the posterior density of the transformed element $p(\mathbf{y}|\mathbf{y}')$ belongs to the same family as the unconditional density $p(\mathbf{y})$. This property of closure under conditioning is inherited from the (base) Gaussian pdf, and it is preserved by the coordinate-wise transformation φ. Furthermore, the non-Gaussian multivariate distribution $p(\mathbf{y})$ is also closed under marginalisation and permutation, again since φ is defined coordinate-wise.

Therefore, we can construct a non-Gaussian process by *transforming* (or *warping*) a GP in the following manner: (i) choose a base GP x and a coordinate-wise transformation φ, (ii) compute the finite-dimensional marginal densities of y s.t. $x = \varphi(y)$ via the change of variable theorem, and (iii) apply the Kolmogorov consistency theorem [134]. This construction guarantees the existence of such non-Gaussian process with known hyperparameters: the mean and covariance of the base GP and the transformation φ.

3.2 Warped Gaussian Processes

Warped Gaussian processes (WGP) [125] follow the rationale explained in the previous section. WGP considers a GP with zero mean and square-exponential (SE) covariance function, as well as a monotonic (and thus invertible) parametric coordinate-wise transformation.

The transformation $\varphi : \mathbb{R} \to \mathbb{R}$ considered by WGP [125] is given by

$$\varphi(y) = y + \sum_{j=1}^{\mathrm{d}} a_j \tanh\left(b_j\left(y + c_j\right)\right), \tag{3.1}$$

where $a_j, b_j \geq 0,\ j = 1, \ldots, \mathrm{d}$. The mixture of the identity and hyperbolic tangent functions in eq. (3.1) acts as a parametric warping of the identity function, meaning that standard transformations such as the logarithm are not allowed by WGP. Observe that since $\varphi(y)$ in eq. (3.1) is a sum of monotonic terms, its inverse does exists. However, as this inverse is not known explicitly, computing the predictive posterior WGP requires approximating φ^{-1} using, e.g., the Newton-Raphson method (NRM) [11]. This iterative procedure requires several evaluations of φ and $\frac{\mathrm{d}\varphi}{\mathrm{d}y}$, thus increasing computational complexity, in addition to being sensitive to the initial condition. In practice, the use of NRM is the computational bottleneck of WGP: the original model proposed in [125] considered a naive NRM approach that resulted in inference being one or two orders of magnitude more expensive than that of standard GPs. For computational efficiency, the implementation of [125] considered a bisection search to find appropriate initial conditions for NRM. We emphasise that although the implementation of WGP can be made more efficient by using sophisticated numerical tools for approximating inverse functions, e.g., to train a surrogate model for the inverse using splines or neural networks, WGP always requires numerical approximations when performing predictions due to the lack of the explicit inverse of a sum of hyperbolic tangents. In [144] the authors propose the alternative warped function

$$\varphi(x) \;=\; \sum_{j=1}^{\mathrm{d}} a_j \log\left[1 + \exp[b_j\left(x + c_j\right)]\right]$$

where $a_j, b_j \geq 0, j = 1, ..., \mathrm{d}$, however this warping inherit the same issues described above. On the contrary, the model proposed in Sec. 3.3.1 does not suffer from this drawback.

3.2.1 Bayesian warped Gaussian processes

A non-parametric version of WGP is the Bayesian WGP [72], denoted BWGP, which models the transformation itself as a GP with the identity function as mean. This transformation ϕ in BWGP corresponds to the inverse of the transformation φ in WGP and can be expressed as

$$y(t) = \phi\left(x(t)\right) + \varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$ and both x and ϕ are GPs, that is,

$$x(t) \sim \mathcal{GP}\left(m(t), k\left(t, \bar{t}\right)\right) \tag{3.2}$$

$$\phi(f) \sim \mathcal{GP}\left(f, c\left(f, \bar{f}\right)\right), \tag{3.3}$$

where f denotes the input (function) to the warping ϕ and c is its covariance kernel. Furthermore, [33] proposes a deep version of BWGP termed Deep GP (DGP), where the warping function is a composition of multiple GPs.

DGP, which has been proposed primarily as a hierarchical extension of the Bayesian Gaussian process latent variable model (GP-LVM) [136], which, in turn, is a deep belief network based on Gaussian process mappings, and it focuses initially on unsupervised problems (unobserved hidden inputs) about discovering structure in high-dimensional data [71, 76, 34]. However, by replacing the latent inputs with observed input, a one-hidden-layer model coincides with BWGP, so DGP for regression is also a generalisation of BWGP [32]. DGP is one GP feeding another GP, so it is a flexible model that can capture highly-nonlinear functions for complex data sets. However, the network structure of a DGP makes inference computationally expensive; even the inner layers has an identified pathology [40]. To use DGP in regression scenarios, some authors propose making inference via variational approximations [25, 114] or using sequential sampling approach [140]. Finally, DGP loses its interpretability, so, like other deep models, it is difficult to understand the properties of each layer and component.

Training and inference are intractable both for BWGP and DGP; therefore, both methods rely on a variational approach to perform inference using a sparse representation [135]. Due to their considerable computational complexity, comparisons of the proposed method against BWGP and DGP are beyond the scope of this article, since we focus on expressive warping functions that provide computationally-efficient closed-form formulas for training and prediction. Therefore, the experimental validation of the proposed method will be performed against WGP [125] only.

3.3 A Novel Warping for WGPs

Inspired by deep architectures, we propose a generative model for non-Gaussian processes by transforming a latent GP through a composition of *elementary functions* φ_{i} with two main objectives. The first objective is that the class of transformations has to be general enough to replicate a broad class of data using few parameters to avoid overfitting, while the second objective is that the approximations required for learning and inference should be minimal to maintain high numerical precision and low computational complexity.

3.3.1 Model description

Let us consider a family of parametric functions $\{\varphi_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{d}}$, $\mathrm{d} \in \mathcal{N}$, that are differentiable and invertible with closed-form inverse, hereinafter referred to as *elementary functions*. Then, we can construct warping functions $\varphi(\cdot)$ as a composition of such elementary functions, that is,

$$\varphi(\cdot) = \varphi_{\mathrm{d}}(\varphi_{\mathrm{d}-1}(\cdots(\varphi_2(\varphi_1(\cdot)))\cdots)). \tag{3.4}$$

This construction is motivated by the fact that the inverse and derivatives of function compositions are given by the inverses and derivatives of their component functions. For instance, for a two-elementary-function composition $\varphi(y) = \varphi_2(\varphi_1(y)) = x$, the inverse and the derivative are given respectively by

$$\varphi^{-1}(x) \;=\; \varphi_1^{-1}(\varphi_2^{-1}(x))$$

$$\frac{\mathrm{d}\varphi\left(y\right)}{\mathrm{d}y} = \frac{\mathrm{d}\varphi_2\left(\varphi_1\left(y\right)\right)}{\mathrm{d}y}\frac{\mathrm{d}\varphi_1\left(y\right)}{\mathrm{d}y}.$$

Notice that this class of warping functions goes one step further compared to WGP: WGP requires invertibility but then deals with finding the inverse numerically, whereas the compositional warping proposed here requires invertibility and closed-form inverses, meaning that the evaluation of the inverse is straightforward.

We then propose the compositionally-warped Gaussian process (CWGP) given by $y(t)$ s.t.

$$\begin{aligned}\varphi(y(t)) &= x(t),\\ x(t) &\sim \mathcal{GP}(m(t), k(t,\bar{t})),\\ \varphi(\cdot) &= \varphi_{\mathrm{d}}(\cdots(\varphi_2(\varphi_1(\cdot)))\cdots),\end{aligned}$$

where $\{\varphi_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{d}}$ are elementary functions. Additionally, as the inverse of φ is known, CWGP can also be interpreted as a generative model that transforms $x(t)$ into $y(t)$ using the transformation φ^{-1}. For notational clarity we emphasise that φ is defined from the non-Gaussian process y to the Gaussian process x.

Finally, we also clarify that the model described above differs radically from the concept of Normalising Flows (NF) [133, 132, 103]. NF focuses on approximating the posterior density of an intractable model, whereas we construct a non-Gaussian generative model directly.

3.3.2 Learning: robust, interpretable and efficient

Learning under CWGP means finding the hyperparameters of the GP x (parameters of the kernel and mean functions denoted by θ_x) in addition to the parameters of the compositional transformation φ, denoted by θ_φ. Thanks to the change of variables theorem, learning these parameters is tractable and can be achieved via minimisation of the negative logarithm of the marginal likelihood (NLL).

Robustness. Just as standard GPs, warped GPs are protected from overfitting, since they directly parametrise a prior distribution over functions and not the specific trajectories of the function. Additionally, recall that the warping considered is component-wise and given by the same scalar-valued map for all the coordinates. Thus the warping can be understood as a parametrisation of the marginal histogram. Therefore, the resulting generative model has non-Gaussian marginals with Gaussian copulas, known as *Gaussian copula process* [144], meaning that in the broad sense of modelling the law of stochastic process, the proposed model is regularised by design.

Interpretability. The NLL is given by

$$\begin{aligned}\mathrm{NLL} &= -\log p(\mathbf{y}|\theta_x, \theta_\varphi) \\ &= \underbrace{\frac{n\log(2\pi)}{2}}_{\text{constant term}} + \underbrace{\frac{1}{2}\left(\varphi(\mathbf{y}) - \mu_{\mathbf{x}}\right)^\top \Sigma_{\mathbf{xx}}^{-1}\left(\varphi(\mathbf{y}) - \mu_{\mathbf{x}}\right)}_{\text{data-fit term}}\end{aligned} \tag{3.5}$$

$$+ \quad \underbrace{\frac{1}{2}\log|\Sigma_{\mathbf{xx}}|}_{\text{kernel-complexity term}} \quad - \quad \underbrace{\sum_{\mathrm{i}=1}^{n}\log\left(\frac{\mathrm{d}\varphi(y_{\mathrm{i}})}{\mathrm{d}y}\right)}_{\text{warping-complexity term}},$$

where $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{xx}}$ are the mean and covariance of $\mathbf{x} = \varphi(\mathbf{y})$.

Akin to standard GPs, for which the NLL reveals automatic penalty of model complexity, WGP features the *warping-complexity term*. Therefore, the NLL is minimised balancing the Gaussianity of the base GP x via the first three terms in eq. (3.5) and the regularity of the warping via the warping-complexity term. The first criterion prioritises solutions such that $||\varphi(\mathbf{y}) - \mu_{\mathbf{x}}||$ is small wrt to the norm induced by $\Sigma_{\mathbf{xx}}^{-1}$, where the extreme solution is given by $\varphi(\mathbf{y}) = \mu_{\mathbf{x}} = \text{constant } \forall \mathbf{y}, t$, since $\varphi(\mathbf{y}) : \mathbf{y} \mapsto x$ and $\mu_{\mathbf{x}} : t \mapsto x$. However, notice that the warping-complexity term $\sum_{\mathrm{i}=1}^{n}\log\left(\frac{\mathrm{d}\varphi(y_{\mathrm{i}})}{\mathrm{d}y}\right)$ forces solutions $\varphi(\mathbf{y})$ that have large derivatives (i.e., which grow steeply), thus ruling out the constant case. These terms offer a clear interpretation of the likelihood function of WGP: the warping-penalty term promotes the preservation of the data variability by choosing warpings with large derivatives, while the remaining terms ensure that this variability remains as Gaussian as possible.

Computational complexity. Notice that minimising the NLL does not require the inverse of φ but only its log-derivatives, which are known in closed form, therefore, the cost of training CWGP is only dominated by the matrix inversion: $\mathcal{O}(n^3)$ for n observations. Recall that this is the same order of complexity of training standard GPs. Intuitively, learning is then achieved by transforming the non-Gaussian observations to then maximise the (Gaussian) probability of the transformed samples wrt to the parameters of (i) the Gaussian distribution and (ii) those of the transformation. Although the complexity of evaluating the NLL is the same for CWGP and standard GPs, our model is more expressive so the NLL could have more local minima due to having more parameters to train. For further details, we recommend [107], where multiple local minima are explored with derivative-free and Monte Carlo based optimisation.

3.3.3 Closed-form inference

Inference follows from a corollary of the change of variables theorem that states that the probability (measure) of a set E under the density of $\mathbf{y}$, is equal to the probability of the image of E, $\varphi(E)$, under the density of $\mathbf{x}$. Conditioning on observed data $\mathbf{y}$, we can express the corollary as

$$\int_{E} p_y\left(y|\mathbf{y}\right)\mathrm{d}y = \int_{\varphi(E)} p_x\left(x|\mathbf{y}\right)\mathrm{d}x = \int_{\varphi(E)} p_x\left(x|\mathbf{x}\right)\mathrm{d}x,$$

Inference follows from a corollary of the change of variables theorem that states that the probability (measure) of a set E under the density of $\mathbf{y}$ is equal to the probability of the image of E, $\varphi(E)$, under the density of $\mathbf{x}$. Conditioning on observed data $\mathbf{y}$, we can express the corollary as

$$\text{median}(y(t)) = \varphi^{-1}\left(\text{median}(x(t))\right) = \varphi^{-1}\left(m(t)\right)$$

$$I^p_{y(t)} = \left[\phi^{-1}\left(m(t) - z_p\sigma(t)\right), \phi^{-1}\left(m(t) + z_p\sigma(t)\right)\right],$$

where $\sigma(t) = \sqrt{k(t,t)}$ is the base GP standard deviation, z_p is the p-quantile of a standard Gaussian (ex. $z_{0.975} \approx 1.96$) and we used the fact that for a Gaussian median(x) = mean(x).

Sampling the non-Gaussian process is also direct: it is only required to simulate a realisation of the GP and then apply the inverse of the transformation in a coordinate-wise way, that is,

$$\begin{aligned} x(\mathbf{t}) &\sim \mathcal{GP}(m(\mathbf{t}), k(\mathbf{t},\mathbf{t})) \\ y(\mathbf{t}) &= \varphi^{-1}\left(x(\mathbf{t})\right). \end{aligned}$$

3.3.4 Complexity analysis of inference

Relying on the change of variables theorem once again, the expectation of a measurable function $h : \mathcal{Y} \to \mathbb{R}$ under the non-Gaussian law $p(\mathbf{y})$ is given by

$$\mathbb{E}_{\mathbf{y}}\left[h\left(\mathbf{y}\right)\right] = \mathbb{E}_{\mathbf{x}}\left[h\left(\varphi^{-1}\left(\mathbf{x}\right)\right)\right].$$

Additionally, since the distribution of $\mathbf{x}$ is Gaussian, we can efficiently compute the above integral numerically using the Gauss-Hermite quadrature [1], for which k-point approximations are exact when the integrand $h\left(\varphi^{-1}\left(\cdot\right)\right)$ is a polynomial of order $2k-1$. Choosing $h(y) = y$, we have the approximation of the mean of y given by

$$\begin{aligned} \mathbb{E}_y\left[y\right] &= \int \varphi^{-1}\left(x\right) p_x\left(x\right) \mathrm{d}x \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{\mathrm{i}=1}^{k} w_{\mathrm{i}} \varphi^{-1}\left(\sqrt{2}\sigma_x x_{\mathrm{i}} + m_x\right), \end{aligned} \tag{3.6}$$

where the weights $\{w_{\mathrm{i}}\}_{\mathrm{i}=1}^k$ and locations $\{x_{\mathrm{i}}\}_{\mathrm{i}=1}^k$ are given by the Gauss-Hermite quadrature method [1].

Finally, observe that evaluating φ^{-1} is required to compute expectations, the median and confidence intervals of the non-Gaussian model. Since for CWGP φ^{-1} is known, the cost of evaluating it is $\mathcal{O}(\mathrm{d})$, where d is the number of elementary components of φ. Therefore, the cost of evaluating $\mathbb{E}_y\left[y\right]$ in eq. (3.6) using the k-point Gauss-Hermite quadrature is $\mathcal{O}(k\mathrm{d})$ for CWGP. Conversely, WGP approximates φ^{-1} using the Newton-Raphson method (NRM) [11] (with the bisection method to find the initial point), meaning that the cost of evaluating $\mathbb{E}_y\left[y\right]$ for WGP is $\mathcal{O}(k\mathrm{d}t)$, where t is the number of iterations of NRM (and bisection). In practice, the explicit expression for φ^{-1} is key in computational terms: even using efficient numerical methods, WGP always requires numerical approximations of φ^{-1}, whereas CWGP does not and can evaluate φ^{-1} directly.

3.4 Elementary Transformations

As a companion to the CWGP proposed in the previous section, we now present a set of elementary transformations with explicit inverse and derivative to be used as building blocks

Table 3.1: Elementary transformations: functional forms with derivatives and inverses

Function	$\varphi(y)$	$\frac{\mathrm{d}\varphi(y)}{\mathrm{d}y}$	$\varphi^{-1}(x)$
Affine	$a + by$	b	$\frac{x-a}{b}$
Logarithmic	$\log(y)$	y^{-1}	$\exp(x)$
Arcsinh	$a + b\,\mathrm{asinh}\left(\frac{y-c}{\mathrm{d}}\right)$	$\frac{b}{\sqrt{\mathrm{d}^2+(y-c)^2}}$	$c + \mathrm{d}\,\sinh\left(\frac{x-a}{b}\right)$
Box-Cox	$\frac{\mathrm{sgn}(y)\lvert y\rvert^{\lambda}-1}{\lambda}$	$\lvert y\rvert^{\lambda-1}$	$\mathrm{sgn}\,(\lambda x+1)\,\lvert\lambda x+1\rvert^{\frac{1}{\lambda}}$
Sinh-Arcsinh	$\sinh\left(b\,\mathrm{asinh}(y)-a\right)$	$\frac{b\cosh(b\,\mathrm{asinh}(y)-a)}{\sqrt{1+y^2}}$	$\sinh\left(\frac{1}{b}\left(\mathrm{asinh}(x)+a\right)\right)$

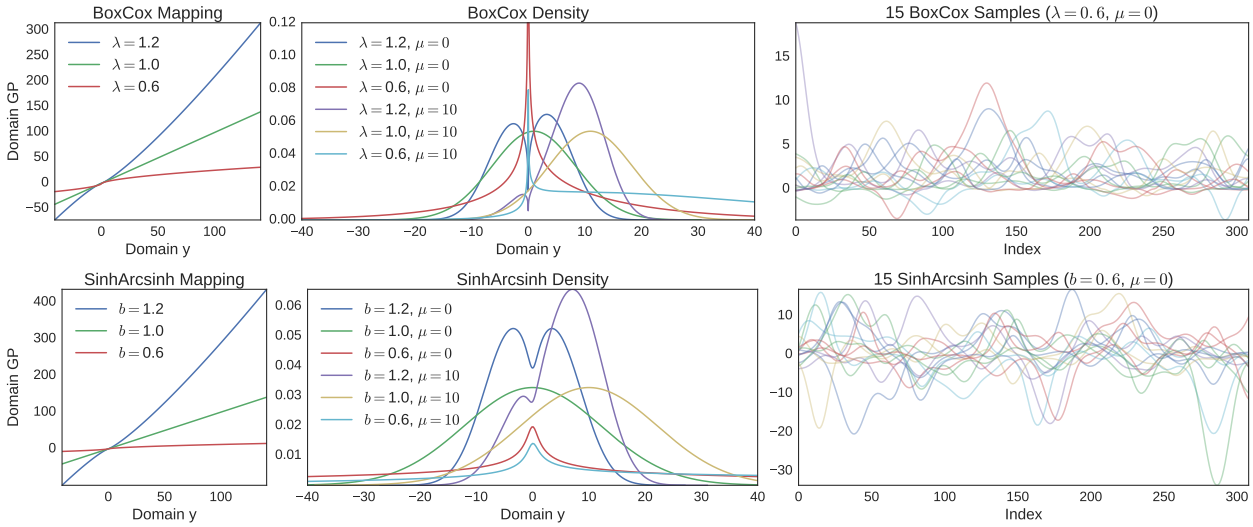

Figure 3.2: Proposed Box-Cox and SinhArcsinh elementary transformations. For all plots, μ denotes the mean of the base GP x. Top: Box-Cox transformation in eq.(3.9). Bottom: SinhArcsinh transformation in eq. (3.11). Left: transformations (or warpings). Middle: induced marginal densities. Right: samples of the warped GP.

of CWGP's compositional transformation. Furthermore, for consistency with Theorem 3.1.1, we present the transformations from the non-Gaussian process y to the GP x. Table 3.1 gives a summary of these transformations together with their inverses and derivatives, and Fig. 3.2 shows two different families of transformations together with their induced marginal densities and sample trajectories.

3.4.1 Affine transformation

The affine transformation is given by

$$\varphi_{\text{affine}}(y) = a + by, \quad a, b \in \mathbb{R}, \tag{3.7}$$

and is referred to as *shift* when $b = 1$ and as *scale* when $a = 0$. The affine transformation does not provide enhanced modelling ability over standard GPs, since an affine-transformed GP is still a GP with a shifted mean and scaled variance. However, the affine warping will be composed with other elementary functions to produce expressive transformations.

3.4.2 Box-Cox transformations

A standard strategy in Statistics to transform non-Gaussian positive observations into *closer-to-Gaussian* ones is to apply the logarithmic function $\varphi_{\log}(y) = \log(y)$; this is the case for positive-valued heavy-tailed stochastic processes [3]. Notice that with the logarithmic transformation, both the mean m_x and variance σ_x^2 of the original GP x affect all moments of the transformed process y. Explicitly, the n-th moment of y is given by

$$\mathbb{E}_y\left[y^n\right] = \exp\left(nm_x + \tfrac{1}{2}n^2\sigma_x\right), \tag{3.8}$$

meaning that a heavy-tailed distribution for y is obtained through only modifying the mean and variance of the original process x.

A generalisation of the logarithmic transformation is the Box-Cox transformation [17, 113], a single-parameter power function given by

$$\varphi_\lambda(y) \;=\; \frac{\operatorname{sgn}(y)\,|y|^\lambda - 1}{\lambda},\ \lambda \in \mathbb{R}_0^+, \tag{3.9}$$

where φ_λ becomes a power function for $\lambda > 0$, an affine transformation for $\lambda = 1$, and the logarithmic transformation for $\lambda = 0$ since $\lim_{\lambda\to 0}\varphi_\lambda(y) = \log(y)$.

The Box-Cox transformation has two useful properties: Firstly, its mode is known to be [46]

$$\operatorname{mode}_y = \left[\frac{1}{2}\left(1 + \lambda m_x + \sqrt{\left(1+\lambda m_x\right)^2 + 4\sigma_x^2\lambda\left(\lambda - 1\right)}\right)\right]^{\frac{1}{\lambda}},$$

where m_x and σ_x^2 are the mean and variance of the GP x respectively. This formula is particularly useful for skewed distributions where the mode is usually considered as a point estimate instead of the mean or the median. Secondly, the computation of moments using numerical methods, e.g., the Gauss-Hermite quadrature [1], can be performed with high precision due to the polynomial nature of the Box-Cox transformation. Fig. 3.2 (top) shows different Box-Cox transformations with their induced marginal densities.

3.4.3 Hyperbolic transformations

The distribution resulting from passing a $\mathcal{N}(0,1)$-distributed random variable through the inverse hyperbolic sine transformation

$$\varphi_{\operatorname{arcsinh}}(y) = a + b\operatorname{arcsinh}\left(\frac{y-c}{\mathrm{d}}\right), \tag{3.10}$$

where $a, c \in \mathbb{R}$ and $b, \mathrm{d} \in \mathbb{R}^+$, is known as the Johnson's SU-distribution [60] and has closed-form expressions for the mean and variance, given respectively by

$$\mu_{\mathrm{SU}} = c - \mathrm{d}\exp\left(\frac{b^{-2}}{2}\right)\sinh\left(\frac{a}{b}\right)$$

$$\sigma^2_{\mathrm{SU}} = \frac{\mathrm{d}^2}{2}\left[\exp\left(b^{-2}\right) - 1\right]\left[\exp\left(b^{-2}\right)\cosh\left(\frac{2a}{b}\right) + 1\right],$$

and also for the skewness and kurtosis [60].

Another transformation based on hyperbolic functions is the Sinh-Arcsinh [61], where arcsinh, affine and sinh are composed together, that is,

$$\varphi_{\mathrm{SinhArcsinh}}(y) = \sinh\left(b\,\mathrm{arcsinh}(y) - a\right), \tag{3.11}$$

where $a, b \in \mathbb{R}$. This distribution admits explicit expressions for all moments of y, using the modified Bessel function, and it induces a distribution where the third and fourth moments can be controlled via parameters a and b. This distribution is symmetric if $a = 0$; positively-skewed (cf. negatively-skewed) if $a > 0$ (cf. $a < 0$); mesokurtic if $b = 1$; and leptokurtic (cf. platykurtic) is $b > 1$ (cf. $b < 1$). Additionally, this distribution meets $0 < |\mathrm{mode}(y)| < \sinh(|a|/b)$ and $\mathrm{sgn}(\mathrm{mode}(y)) = \mathrm{sgn}(a)$.

Fig. 3.2 (bottom) shows the Sinh-Arcsinh transformations with the induced marginals and samples for a skewness parameter set to $a = 0$ and different values of the kurtosis parameter b. Observe that the mean of the base GP, μ, can also change the skewness of the induced marginal distribution.

3.5 How to Choose the Elementary Transformations?

As in the vast majority of deep structures, the number of layers and the type of neurons are defined by experts or by trial and error, where interpretability is a desired property [141]. This expertise is also needed in the case when choosing the kernel in support vector machines or Gaussian processes (as studied in [39]). Recall that in standard mixture models (such as WGP) the user only defines the number of components, whereas within the proposed CWGP one also needs to choose the types and order of the elements (in our case, elementary functions). This section guides the choice of the elementary transformations under two scenarios, the first one being the case when expert knowledge about the data is available. The second scenario, that is, when no prior information of the data is available, we show that CWGP can be implemented by concatenating multiple instances of a particular sequence of elementary transformations (referred to as the *SinhArcsinh-Affine* layer), and we show that this construction has appealing experimental performance. This way, CWGP can be regarded as a black-box where, akin to deep structures, the user only needs to choose the number of layers. We illustrate this concept based on the NLL (Sec. 3.5.2) and via a toy example (Sec. 3.5.3), as well as its robustness to overfitting and through a real-world data in Section 3.6.2.

3.5.1 When prior knowledge of the data is available

As mentioned in Sec. 3.4.2, when the data are strictly positive, standard practice is to apply the logarithmic transformation. Critically, if the data is known to be lower-bounded

by an unknown quantity, one can compose the logarithmic transformation with the shift transformation in eq. (3.7) to find the shift parameter during training. An upper bound to the data can be found analogously by replacing the shift by an affine transformation, thus allowing for a negative scaling. In this sense, composing two affine-logarithmic transformations enables us to find the upper and lower bounds simultaneously.

To further relax the strict (lower) bound condition of the logarithmic transformation to a more permissive one, we can also replace the logarithm by the Box-Cox transformation in eq. (3.9), where the permissiveness of the bound is controlled by the parameter λ. Additionally, if the data is such that their range is not bounded but instead have a large dispersion, then the data follows a heavy-tailed distribution. This phenomenon can be modelled using the Arcsinh or Sinh-Arcsinh transformations in eqs. (3.10) and (3.11) respectively since such transformations allow to control the mean and variance of the distribution, as well their asymmetry and kurtosis. All these transformations can be composed with one another to construct more complex distributions, as in the case of multimodal distributions.

3.5.2 Sparse compositional transformations

As in any model that involves choosing a finite order (such as layers, neurons, components), it is required that the addition of more elementary functions in CWGP results in a monotonically-increasing performance. In particular, if one considers an unnecessarily-large number of elementary transformations, it is desired that some of these transformations *revert* to the identity function (and thus can be removed). If, after training, some of the transformations considered revert to the identity, we will say that the compositional transformation is sparse.

When insight into the statistical properties of the data is scarce, or even non-existent, a recommended procedure is to sequentially add transformations that can revert to the identity when needed. Notice that if a transformation is not able to improve performance and at the same time can revert to the identity, it will indeed do this. This fact can be justified based on the NLL in eq. (3.5): where the data-fit term remains invariant, and the warping-complexity term contributes to a lower NLL. Additionally, one can always choose a prior distribution over the warping parameters to promote further warpings that are close to the identity. Lastly, recall that from the proposed transformations, the Box-Cox, the Sinh-Arcsinh and the affine transformations can revert to the identity, therefore, under limited knowledge about the underlying properties of the data, we recommend adding these components iteratively until the performance of the model reaches a plateau. We next implement this concept based only on the Sinh-Arcsinh and affine transformations on synthetic data and, in Section 3.6.2, on real-world data.

3.5.3 Structure discovery via deep compositional transformations

For the cases when expert knowledge about the nature of the data is scarce, the proposed CWGP can be implemented just concatenating multiple instances of the proposed elementary transformations, this procedure is usual and widely accepted in general deep architectures

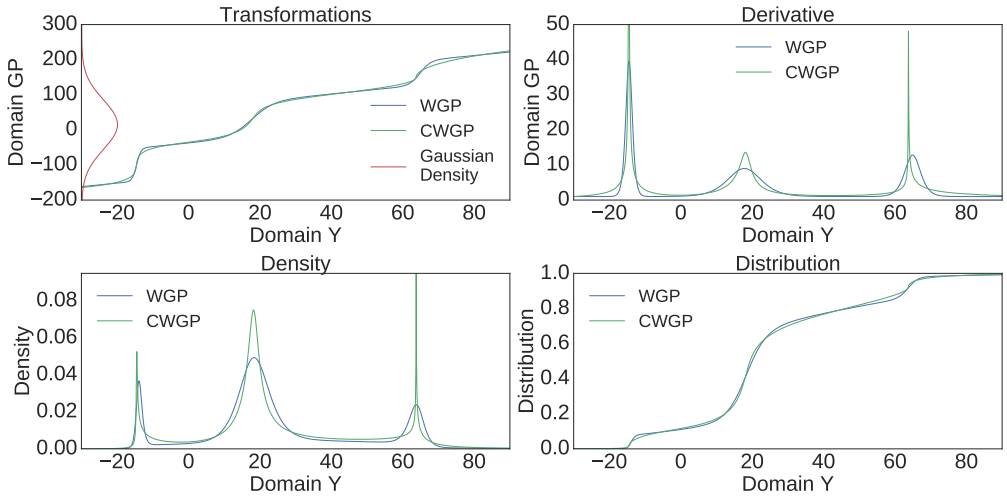

Figure 3.3: Approximation of a WGP warping (sum of three hyperbolic tangents, blue) using the proposed compositional method (three SAL layers, green).

[14, 55, 116]. To illustrate this, let us first define the composition of a Sinh-Arcsinh and Affine transformations, in eqs. (3.10) and (3.7) respectively, as the *SAL layer*[2] given by

$$l(y) = a + b\sinh(c\,\mathrm{arcsinh}(y) - \mathrm{d}), \tag{3.12}$$

where $a, b, c, \mathrm{d} \in \mathbb{R}$ are the only four parameters of the so-defined layer. We next show that, by only concatenating SAL layers, we can replicate the sum-of-hyperbolic-tangent warping implemented by WGP [125], in eq. (3.1). The reason to assess the proposed model in the approximation of the WGP is that the sum of hyperbolic tangents is known to be *universal*, meaning that it can approximate continuous functions to any desired degree of accuracy in a closed interval.

Intending to gain an intuitive understanding about the modelling ability of the compositional approach, the first illustrative example is to train a three-SAL-layer compositional transformation, via least squares, to replicate a mixture of three hyperbolic tangents. Fig. 3.3 shows the transformations, derivatives, densities and distributions of the ground truth (WGP, blue) and the three-SAL-layer compositional approximation (CWGP, green). Observe the point-wise similarity of the warpings and that the probability mass is concentrated around the three common modes in the domain of y.

Regarding the expressiveness of the proposed compositional approach as a function of the number of considered SAL layers, Fig. 3.4 shows the induced distributions for a five-hyperbolic-tangent WGP warping (blue) and those of the compositional approximations using one to six layers (green) fitted by least squares. Notice how the distributions learnt by the compositional transformation becomes indistinguishable from the ground truth as the number of SAL layers increases. Table 3.2 reports the approximation errors both for the transformation and the resulting (warped) distribution, using the L_1, L_2 and L_∞ norms given respectively by

$$\mathrm{e}_1 = ||f_{\mathrm{SoT}} - f_{\mathrm{CT}}||_1 \;= \int_{\mathbb{R}} |f_{\mathrm{SoT}}(x) - f_{\mathrm{CT}}(x)|\mathrm{d}x$$

[2]The acronym SAL comes from SinhArcsinh and Affine, where the use of "L" stems from "linear". This terminology has been chosen to be consistent with the experimental part in the next section.

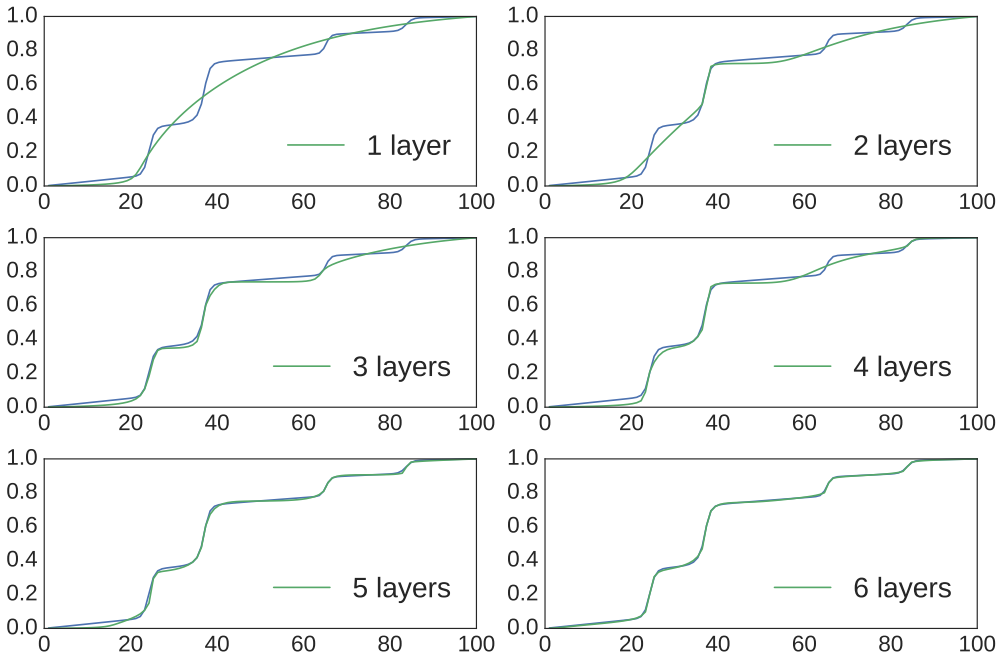

Figure 3.4: CWGP approximation of the distribution of a WGP with five hyperbolic tangents: Ground truth (blue) and CWGP approximations (green) using a variable number of SAL layers in eq. (3.12).

$$\mathrm{e}_2 = ||f_{\mathrm{SoT}} - f_{\mathrm{CT}}||_2 = \sqrt{\int_{\mathbb{R}} |f_{\mathrm{SoT}}(x) - f_{\mathrm{CT}}(x)|^2 \mathrm{d}x}$$
$$\mathrm{e}_\infty = ||f_{\mathrm{SoT}} - f_{\mathrm{CT}}||_\infty = \sup_{x \in \mathbb{R}} |f_{\mathrm{SoT}}(x) - f_{\mathrm{CT}}(x)|,$$

where f_{SoT} denotes de transformation (or distribution) of WGP's *sum of hyperbolic tangents*, and f_{CT} those of the proposed *compositional transformation*. Fig. 3.5 also shows the above error measures normalised wrt the to the single-layer case—observe the monotonic performance of the approximation as the number of SAL layers increases.

3.6 Experimental Validation

We evaluated CWGP experimentally in three real-world scenarios. The first one has an illustrative purpose and demonstrates the robustness of CWGP wrt the number of chosen elementary functions using an astronomical time series. The second experiment validates the ability of the proposed CWGP to identify critical statistical properties of a real-world financial time series. Lastly, the third experiment tests CWGP on the three datasets used initially in [125, 72], where we aim to assess the proposed model in terms of predictive performance and experimental, computational complexity.

We compared the proposed CWGP against GP and WGP only and left BWGP and DGP out of this study due to several reasons. First, we aim to offer a computationally efficient method with exact inference and minimal numerical approximations for prediction, BWGP and DGP

Layers	Trans L1	Trans L2	Trans L∞	Dist L1	Dist L2	Dist L∞
1	1878.2	243.11	60.57	3.342	0.464	0.147
2	1147.7	151.86	33.77	2.232	0.292	0.107
3	845.14	124.80	37.19	1.628	0.192	0.047
4	582.53	83.71	27.34	1.464	0.184	0.041
5	319.64	41.33	15.28	0.793	0.115	0.057
6	147.78	19.95	6.48	0.316	0.042	0.015
7	91.64	15.71	8.32	0.174	0.025	0.011

Table 3.2: Black-box approximation of a WGP warping with five hyperbolic tangents: L_1, L_2 and L_∞ error measures for transformations and induced distributions for different number of layers.

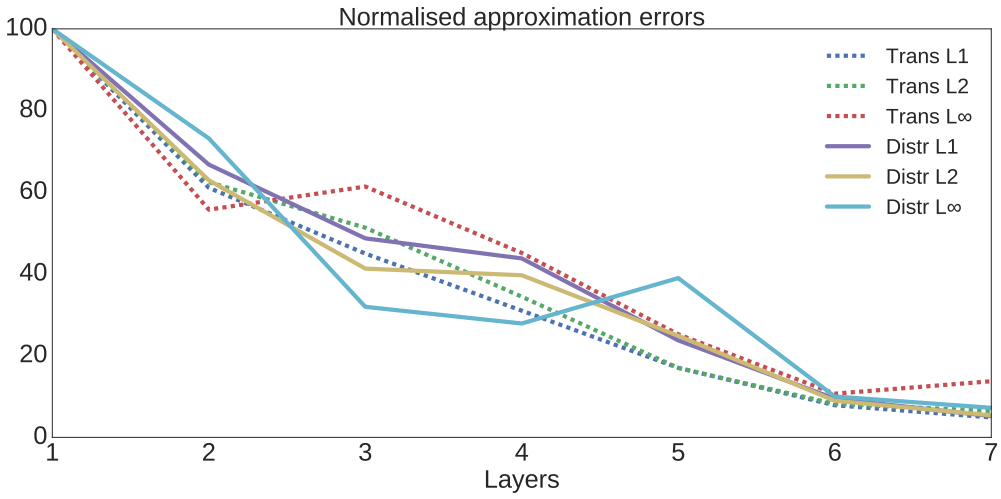

Figure 3.5: Representation of error measures in Table 3.2 normalised wrt to the error of the single-layer case.

fall well outside this aim due to their intractable inference. Second, both BWGP and DGP rely on variational inference (VI) methods. Therefore, the performance of BWGP/DGP depends on the considered approximation. Consequently, a comparison using off-the-shelf VI methods might be misleading; in fact, notice that DGP [33] did not compare against BWGP. Third, according to [72], the standard WGP performed better that BWGP in five out of six performance indices for the same datasets; as we consider those datasets in Sec. 3.6.4, we are also indirectly comparing against BWGP. Finally, we believe that the availability of an invertible warping is vital for interpreting the relationship between the base GP and the transformed (non-Gaussian) process, as this leads to discovering statistical properties of the data; this is an advantage of the CWGP that neither BWGP nor DGP can provide.

We next define performance indices to be used in our experimental evaluation to then proceed to the simulations.

3.6.1 Performance indices

For consistency with the existing literature on warped GPs [125, 72] and to give a thorough evaluation of the model proposed, we considered four performance indices: the negative log-likelihood (NLL), the root mean squared error (RMSE), the mean absolute error (MAE), and the negative log predictive distribution (NLPD). These indices are described below and should be interpreted as *the lower the better*.

Firstly, the NLL in eq. (3.5) is a measure of the probability of the observed data under the chosen model. Model selection and fitting will be achieved by minimising the NLL wrt to the model parameters and hyperparameters.

Let us now denote a test set $\{y_{\mathrm{i}}\}_{\mathrm{i}=1}^{n}$ and the reported predictive means $\{y_{\mathrm{i}}^{*}\}_{\mathrm{i}=1}^{n}$, and define the RMSE and the MAE respectively by

$$\mathrm{RMSE} = \left(\frac{1}{n}\sum_{\mathrm{i}=1}^{n}(y_{\mathrm{i}} - y_{\mathrm{i}}^{*})^{2}\right)^{\frac{1}{2}} \tag{3.13}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}|y_{\mathrm{i}} - y_{\mathrm{i}}^{*}|. \tag{3.14}$$

These two indices are representative of point prediction errors.

Lastly, the NLPD, a measure of the (not necessarily Gaussian) distribution prediction error is defined by

$$\mathrm{NLPD} = -\frac{1}{n}\sum_{\mathrm{i}=1}^{n}\log(p_{\mathrm{i}}(y_{\mathrm{i}})), \tag{3.15}$$

where $\{p_{\mathrm{i}}(\cdot)\}_{\mathrm{i}=1}^{n}$ are the learnt predictive densities.

In addition to the above performance indices, the models considered are also evaluated in terms of their training and evaluation times in the second set of experiments.

3.6.2 Testing for robustness with the Sunspots time series

This example aims to show that adding more elementary functions to the CWGP only improves performance and does not overfit to the training set. Using the Sunspot time series [122] corresponding to the yearly number of sunspots between 1700 and 2008 (309 data points), we randomly selected half of the data between 1700 and 1961 (131 observations) as the training set. The remaining data points were used for evaluation: the data between 1700 and 1961 not used from training (131 test points) were used for a *reconstruction* experiment, whereas the data after 1961 (47 test points) were used for a *forescasting* experiment.

As the Sunspot series is positive valued and semiperiodic, we used the CWGP with a 2-component spectral mixture (SM) kernel [143, 95] and different quantities of Box-Cox and

Sinh-Arcsinh elementary functions. Each model was trained to minimise the NLL in eq. (3.5) using both the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) [90] and the derivative-free global optimisation Powell [100]; this choice was due to the large number of local minima that characterises the spectral-based kernels [137] and follows [107].

Fig. 3.6 shows the performance (NLL and NLPD) as a function of the number of elementary functions of both models, where zero elementary functions mean standard GP. Notice how these experiments confirm the robustness-to-overfitting ability of the CWGP, where despite the unnecessary addition of elementary functions, the validation performance does not degrade—even for forecasting. Also, Fig. 3.7 shows the trained models with zero elementary functions (standard GP, top) and 6 elementary functions for the Sinh-ArcSinh (middle) and Box-Cox (bottom) compositions.

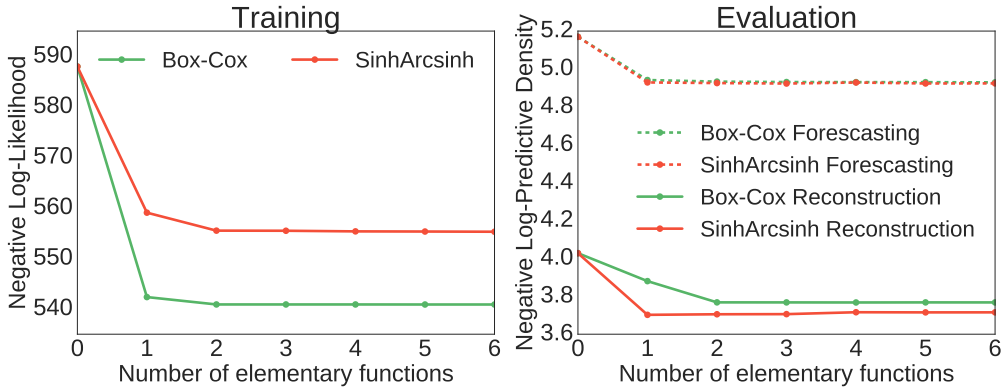

Figure 3.6: Training (left, NLL) and evaluation (right, NLPD) performance of Box-Cox and Sinh-ArcSinh compositions as a function of the number of elementary transformations. Evaluation is assessed over the reconstruction and forecasting experiments.

3.6.3 Learning a macroeconomic time series

We then implemented CWGP alongside a standard GP to learn the quarterly average *3-Month Treasury Bill: Secondary Market Rate* [42] between the first quarter of 1959 and the third quarter of 2009, that is, 203 observations. We know beforehand that this macroeconomic signal is the price of U.S. government risk-free bonds, which cannot take negative values and can have large positive deviations. Therefore, we implemented CWGP with a warping consisting of one affine and one Box-Cox elementary transformations in eqs. (3.7) and (3.9) respectively. This experiment reveals the ability of CWGP to identify the complex statistical properties of the data—where the standard GP fails.

Fig. 3.8 shows both GP (top) and CWGP (bottom) posterior distributions with only 40 observations for the time series, together with their means, error bars and sample trajectories, while Table 3.3 shows the performance metrics. Notice the evident non-Gaussianity of the posterior revealed by the asymmetry of the error bars. From this experiment, we identified four key points that illustrate the superiority of the CWGP against GP: First, the proposed CWGP performed better than GP under all metrics considered (see Section 3.6.1). Second, the error bars and the noise variance are much tighter under CWGP, particularly around

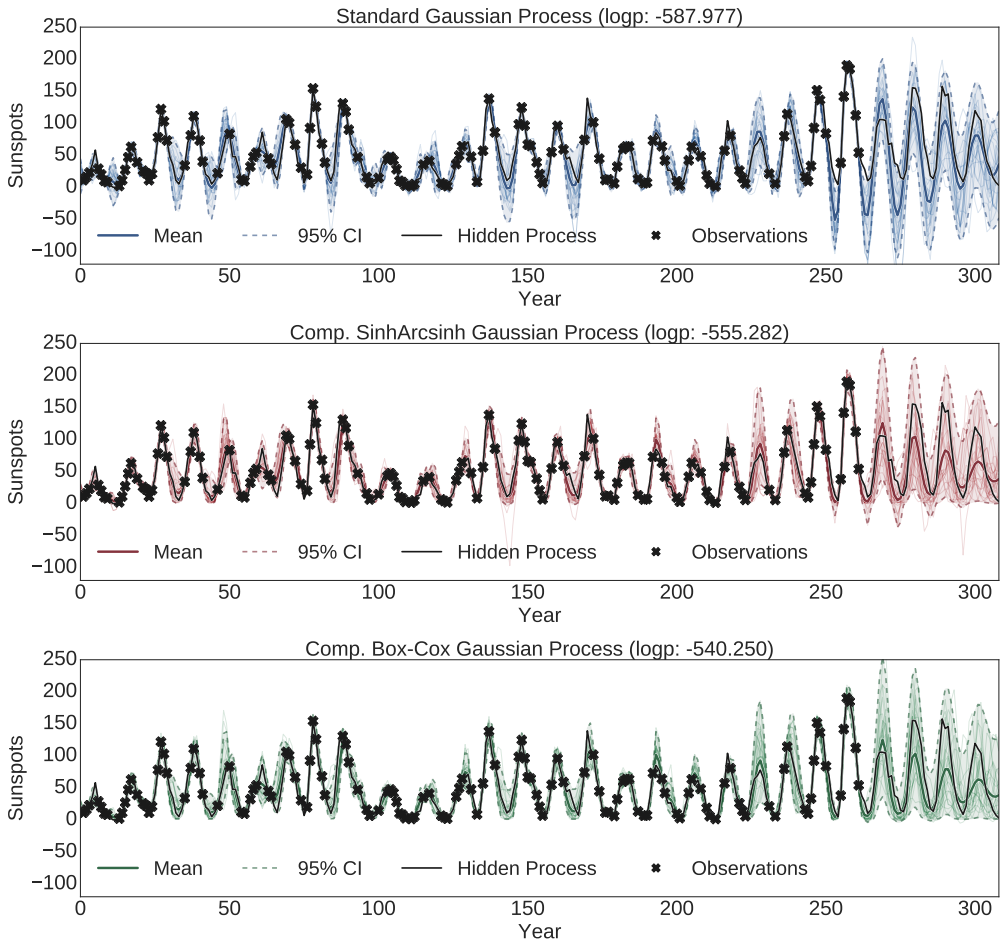

Figure 3.7: Posterior distribution over sunspots trajectories: GP (top), 6-component SinhArcsinh (middle), and 6-component Box-Cox (bottom). Notice the tighter error bars of the CWGP models and the skewed marginal posteriors that are concentrated on positive values.

quarters number 50 and 200. Third, the proposed CWGP was able to successfully identify that the distribution of the signal cannot have negative support: even for ranges with missing data (see between quarters 150 and 200) the error bars did not reach zero. Fourth, CWGP was able to model positive deviations (see the peak around quarter 125) that fully contain the true process.

	MAE	MSE	NLPD	NLL
GP	0.95	1.69	1.74	64.96
CWGP	**0.88**	**1.75**	**1.42**	**57.36**

Table 3.3: Macroeconomic data: Performance of GP and CWGP.

3.6.4 The Abalone, Ailerons and Creep datasets

In this experiment, we considered the three datasets used initially by WGP in [125] and then by BWGP in [72] to validate CWGP. We regard the original WGP model with up to 3

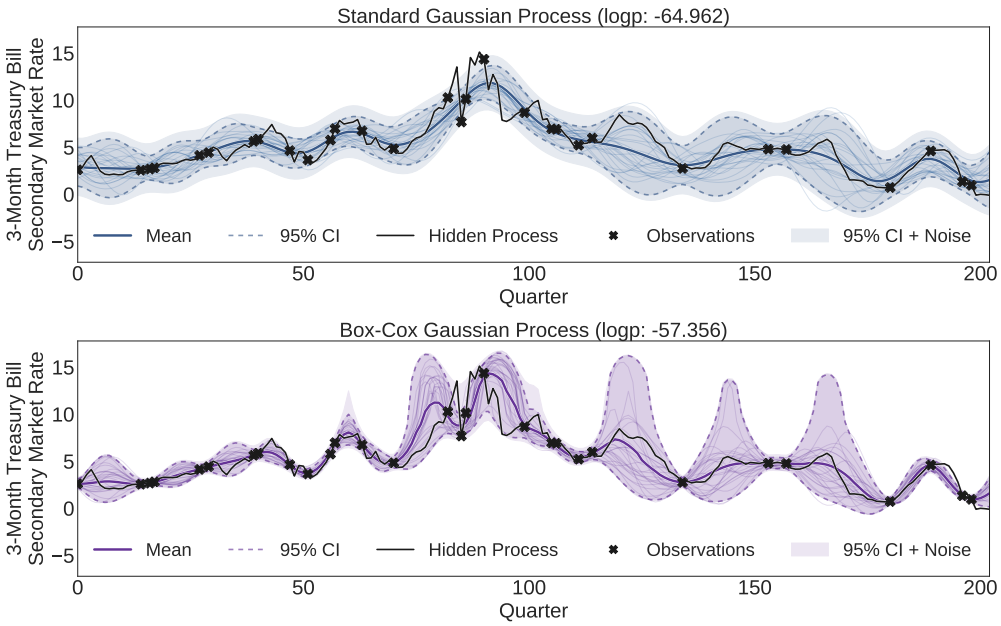

Figure 3.8: Posterior distribution of the *Quarterly Average 3-Month Treasury Bill: Secondary Market Rate* between 1959 and 2009 using 40 observations (203 datapoints in total). Top: Standard GP. Bottom: Proposed CWGP. Both models used a constant mean and a SE kernel. The CWGP warping comprised an Affine and a Sinh-arcsinh transformation.

non-linear components, and the proposed CWGP model maximum of 2 nonlinear components only. Notice that this follows the idea of compositional kernel search presented in [39].

Datasets and models considered

The regression problem associated with the Abalone dataset is to predict the age of an abalone (a type of sea snail) from 8-dimensional physical features. The Ailerons dataset is a simulated control problem designed to predict the control action on the ailerons of an F16 aircraft from a 40-dimensional feature. In the Creep dataset, the objective is to predict creep rupture stress (in MPa) for steel given the chemical composition and other 30-dimensional features. Following [125, 72], the training set sizes were chosen to be 1000 out of 4177, 1000 out of 7154 and 800 out of 2066 for Abalone, Ailerons and Creep datasets respectively.

The models implemented were: (i) a standard GP, (ii) three variants of warped GP with one, two and three $\tanh(\cdot)$ components, and (iii) ten variants of the CWGP constructed by composing the elementary transformations presented in Section 3.4. In total, 14 models were trained and evaluated; all of these used automatic relevance determination squared-exponential kernels [86] and a constant mean function for the base (latent) Gaussian process. The motivation to implement ten variants of CWGP was to show the robustness of the proposed model to the choice of warpings in terms of both predictive performance and computational efficiency. All the experiments were implemented in Python using *g3py* [105], an open-source library for stochastic process modelling.

Abalone	TimeT	TimeE	RMSE	MAE	NLPD
GP	19.927	1.362	**2.158**	1.543	2.287
WGP1	103.55	82.94	2.164	1.534	2.189
WGP2	124.57	72.48	2.174	1.526	2.079
WGP3	127.93	84.98	2.181	1.539	2.200
SA	15.112	1.374	2.191	1.516	2.190
BC-L	17.226	1.383	2.201	1.525	2.181
A-L	23.552	1.385	2.223	**1.512**	2.073
BC-S	**10.811**	1.382	2.211	1.530	2.225
BC-L-SA	16.456	1.395	2.465	1.561	5.272
BC-S-SA	11.354	1.380	3.980	1.525	2.295
A-L-BC-L	23.101	1.373	2.576	1.514	**2.042**
BC-L-A-L	16.236	1.396	2.295	1.547	2.263
A-L-BC-S	24.731	**1.361**	2.302	1.516	2.076
BC-S-A-L	19.215	1.375	2.490	1.517	2.115

Table 3.4: Performance of non-Gaussian models for the Abalone dataset: Training time (TimeT), evaluation time (TimeE), RMSE, MAE and NLPD. The first model is a GP; WP1, WGP2 and WGP3 are WGP models with one, two and three components respectively; and the remaining models are different variants of the proposed CWGP composed by the following elementary transformations: SA:SinhArcsinh, BC:Box-Cox, A:Arcsinh, L:affine, S:shifted. Times are measured in seconds and recall that the lower the score, the better the model.

Ailerons	TimeT	TimeE	RMSE	MAE	NLPD
GP	23.880	8.189	1.814	1.268	1.941
WGP1	151.571	239.947	1.800	1.264	1.935
WGP2	160.557	229.789	1.739	1.231	1.881
WGP3	179.417	245.485	1.765	1.247	1.903
SA	**11.523**	10.274	1.876	1.258	1.821
BC-L	22.708	7.948	1.741	1.228	1.810
A-L	24.892	9.447	1.959	1.385	1.919
BC-S	20.001	**6.992**	**1.702**	**1.210**	1.815
BC-L-SA	12.427	10.472	1.909	1.296	1.820
BC-S-SA	14.587	8.752	2.009	1.334	1.866
A-L-BC-L	19.113	8.266	1.733	1.224	1.793
BC-L-A-L	17.277	7.299	1.727	1.223	**1.791**
A-L-BC-S	18.417	7.023	1.707	1.212	**1.791**
BC-S-A-L	20.225	8.223	1.725	1.223	1.816

Table 3.5: Performance of non-Gaussian models for the Ailerons datasets. Notation follows that of Table 3.4.

Learning the latent GPs and the transformations

For each model, training was as follows. We randomly split the training set in two: An evaluation set and a validation set, both of the same size. We minimised the NLL in eq. (3.5) concerning the evaluation set using the BFGS method starting from 6 initial values of the (hyper)parameters: A default value independent of observations, a value calculated from the observations, a *prelearning* value computed using the trained standard GP, and three random values. We then selected the best model among the 6 trained models according to their RMSE in eq. (3.13) over the validation set. This procedure was repeated 65 times for each model and dataset to obtain an empirical distribution of the performance indices for each considered model.

Creep	TimeT	TimeE	RMSE	MAE	NLPD
GP	12.711	1.312	3.163	2.123	2.462
WGP1	58.281	19.060	**2.750**	1.813	2.162
WGP2	73.323	29.419	2.758	**1.808**	2.166
WGP3	82.402	30.223	2.777	1.822	2.167
SA	14.325	0.918	2.813	1.826	**2.148**
BC-L	9.058	1.426	3.222	2.092	2.268
A-L	14.157	1.024	2.909	1.907	2.218
BC-S	8.139	1.582	3.076	2.055	2.325
BC-L-SA	10.401	0.828	3.592	2.374	2.378
BC-S-SA	**6.759**	1.269	3.879	2.416	2.434
A-L-BC-L	12.845	**0.912**	3.281	2.088	2.269
BC-L-A-L	11.161	1.002	3.252	2.103	2.296
A-L-BC-S	11.191	1.503	3.207	2.026	2.236
BC-S-A-L	8.117	0.917	4.231	2.447	2.399

Table 3.6: Performance of non-Gaussian models for the Creep datasets. Notation follows that of Table 3.4.

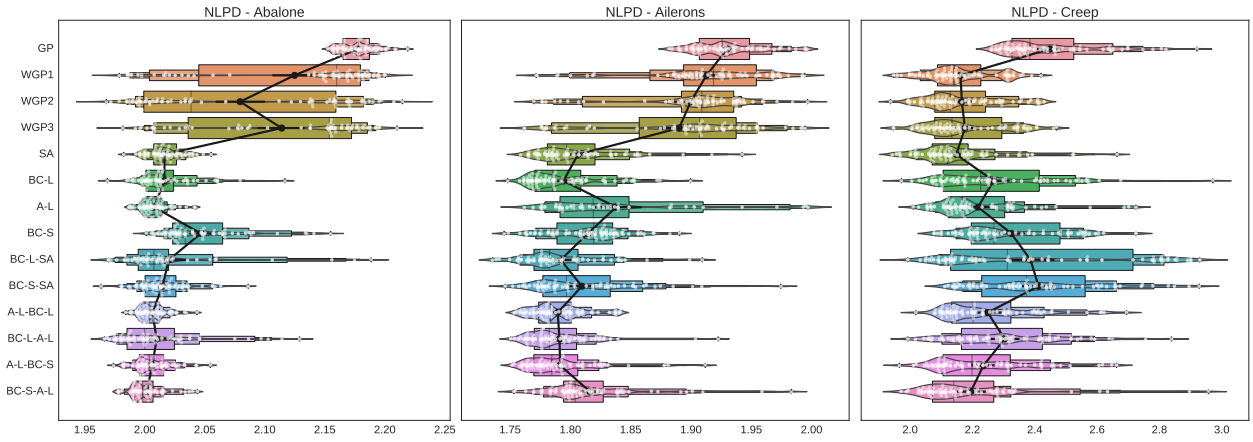

Figure 3.9: NLPD histograms (65 runs) for all models considered and the Abalone, Ailerons and Creep datasets. The white points are the scores, the black marks are the average scores per model, and the boxes denote the quantiles. The models with more white dots to the left-hand side of the plot are the better ones.

Model evaluation

We evaluated each selected model using the NLL, RMSE, MAE and NLPD indices in Section 3.6.1 over the evaluation set. Tables 3.4-3.6 show the training and evaluation average times (TimeT and TimeE respectively) and the average values of all the performance indices considered for the 65 runs for all models and datasets. The proposed CWGP outperformed all models according to the NLPD, a non-Gaussian performance indicator, whereas GP and WGP performed better than CWGP in three cases according to RMSE/MAE. We attribute this to the Gaussian nature of RMSE/MAE that neglects asymmetry or kurtosis.

Observe the appealing training and evaluation times of CWGP. Notice that CWGP's training time was in the same order as that of the standard GP and sometimes even lower, this is because fitting a Gaussian model to non-Gaussian data might yield a flat NLL and therefore minimisation requires several steps of BFGS. Fig. 3.9 shows a histogram of the 65 NLPD scores for each model and dataset, where the white points are the scores, the black marks are

the average score per model and the boxes denote the quantiles. All non-Gaussian models outperform the standard GP in average, and we can see that WGP scores have two modes (especially in the Abalone and Aileron datasets): one closer to the standard GP and another one closer to the scores of the proposed CWGP. This result is due to the difficulty of training WGP, wherein several cases the combination of the Newton-Raphson approximation and the BFGS optimiser fail to find an appropriate nonlinear map. Therefore, the sum-of-tanh warping reduces to the identity, and thus WGP collapses to the standard GP.

Chapter 4

Transport Gaussian Processes

> *"In stochastic processes the future is not uniquely determined, but we have at least probability relations enabling us to make predictions."*
>
> – William Feller, in *An Introduction To Probability Theory And Its Applications*

Following the work developed in Chapter 3, our primary motivation is to extend the Gaussian process methods to other stochastic processes that are more accurate in their assumptions concerning the modelled data, maintaining the elegance and interpretability of its elements. Some authors have defined other models much more expressive than GPs [145], providing methods and approximation techniques, since their exact inference is intractable [70]. In addition to the models discussed previously (WGP [125], BWGP [73] and DGP [33]), a related model is the Student-t process [119] (SP), an extension of the GP with appealing closed-form formulas for training and prediction. It is strictly more flexible due to heavier tails, stability against outliers and stronger dependencies structures, thanks to its non-Gaussian copula. In practice, it has better performance than GPs on Bayesian optimisation [120] and state-space model regression [126]. However, SPs are viewed differently from the models discussed previously, and to date, we do not know of any work that relates them in any way.

The main difficulty of generalising the idea of transform a reference stochastic process is that the transformation must be evaluated over the paths of the process, and except for specific cases such as coordinate transformations, it cannot be implemented as practical models. While the measure-theoretic approach to stochastic processes starts with a probability space, in machine learning the starting point is a collection of finite-dimensional distributions.

The well-know *Kolmogorov's consistency theorem* [134] guarantees that a suitably *consistent* collection of these distributions $\mathcal{F} = \{\eta_{t_1,...,t_n} | t_1, ..., t_n \in \mathcal{T}, n \in \mathbb{N}\}$ will define a stochastic process $f = \{x_t\}_{t\in\mathcal{T}}$, with finite-dimensional laws $\mathcal{F}$. By abuse of notation, their law is denoted as η. Denoting by $F_{t_1,...,t_n}(x_1, ..., x_n)$ the cumulative distribution function of $\eta_{t_1,...,t_n}$, the *consistency* conditions over $\mathcal{F}$ are:

1. Permutation condition: $F_{t_1,...,t_n}(x_1, ..., x_n) = F_{t_{\tau(1)},...,t_{\tau(n)}}\left(x_{\tau(1)}, ..., x_{\tau(n)}\right)$ for all $t_1, ..., t_n \in \mathcal{T}$, all $x_1, ..., x_n \in \mathcal{X}$ and any n-permutation τ.

2. Marginalisation condition: $F_{t_1,...,t_{n+m}}(x_1,...,x_n,+\infty,...,+\infty) = F_{t_1,...,t_n}(x_1,...,x_n)$ for all $t_1,...,t_{n+m} \in \mathcal{T}$ and all $x_1,...,x_n \in \mathcal{X}$.

The main idea that we develop in this Chapter is, for a given and fixed reference stochastic process f, *push-forwarding*[1] each of its finite-dimensional laws $\eta_{\mathbf{t}} \in \mathcal{F}$ by some measurable maps[2] $T_{\mathbf{t}} \in T$, to generate a new set of finite-dimensional distributions $\hat{\mathcal{F}}$ and thus a stochastic process. The main difficulty of this approach is that, in general, $\hat{\mathcal{F}}$ can be inconsistent, in the sense that it can violate some consistency conditions; however, it is possible to choose the maps that induce a consistent set of finite-dimensional laws and therefore a stochastic process.

The main idea is to construct stochastic processes, composed of different *layers*, following the same guidelines as deep architectures, but where each layer has an interpretation defining a feature of the process. In this Chapter we define four types of finite-dimensional transports, that can be seen as elementary layers for our proposed regression model. Our approach starts from a reference Gaussian process noise, since it is a well-know process with explicit density and efficient sampling methods, to generate more expressive stochastic processes. The proposed approach can model non-Gaussian copula and marginals, beyond the known WGP [125, 107, 108] and SP [119], but including all of them from a unifying point of view. The first layer determines the *copula* of the induced process, that can be elliptical or Archimedian via elliptical or Archimedian transports. In the elliptical case, it is possible to compose it with a covariance transport in order to determine the correlation on the induced stochastic process. Finally, in any case, we can compose any number of marginal transports to define an expressive marginal distribution over the induced stochastic process, as it is shown in the previous work [108]. As we saw in the previous sections, these compositions are consistent and expressive enough to include GPs, WGPs, SPs, Archimedean processes, elliptical processes, and those that we could call *warped Archimedean processes* and *warped elliptical processes*.

Our main contribution is to understand the consistency in compositions, to derive general analytic expressions for their posterior distributions and likelihoods functions, and to develop practical methods for the inference and training of our model, given data. The remainder of this Chapter is organised as follows. In Section 4.1, we introduce the notation and necessary mathematical background to develop our work. Our main definition is in Section 4.2, where we propose the transport process (TP) and the inference approach. On Section 4.3, we study the marginal transport that isolates all properties over the univariate marginals of the TP. Similarly, in Section 4.4, we develop the covariance transport, that determines the correlation over the TP. Finally, the main contribution is in Section 4.5, where we introduce the radial transports, that allow us to define the dependency structure (a.k.a copula) over the TP. On Section 4.6, we deepen in details over the computational and algorithmic implementation, and on Section 4.7 we validate our approach with real-world data.

[1] Given a measure η and a measurable map φ, the *push-forward* of η by φ is the measure defined as $[\varphi\#\eta](\cdot) = \eta(\varphi^{-1}(\cdot))$.

[2] Since the set of all indexed measurable maps $T_{\mathbf{t}}$ contains information on all coordinates, by abuse of notation it is denoted as T.

4.1 Introduction

As we reviewed in Chapter 3, WGP define non-Gaussian models with appealing mathematical properties akin to GPs, such as having closed-form expressions for inference and learning. However, they inherit an unwanted Gaussian drawback: the dependence structure, known as copula, remains purely Gaussian. To understand the implications of this issue, we need to formalise the concept of dependence. Let us fix some notation and conventions.

Given a multivariate distribution η, we denote its cumulative distribution function by $F_\eta(\cdot)$. As long as there is no ambiguity, the cumulative distribution function of their i-th marginal distribution η_{i} is denoted as $F_{\mathrm{i}}(x) := F_{\eta_{\mathrm{i}}}(x)$, as well as its right-continuous quantile function, $Q_{\mathrm{i}}(u) := F_{\mathrm{i}}^{-1}(u) = \inf\{x|F_{\mathrm{i}}(x) \geq u\}$. If a multivariate cumulative distribution function C has uniform univariate marginals, that is, $C_{\mathrm{i}}(u) = \max(0, u \wedge 1)$ for $\mathrm{i} = 1, ..., n$, then we say that C is a *copula*. The next result, known as Sklar's theorem [123], shows that any distribution has a related copula.

Theorem 4.1.1 Given a multivariate distribution η, there exists a copula C such that $F_\eta(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$. If the F_{i} are continuous, for $\mathrm{i} = 1, ..., n$, then the copula is unique and given by $C_\eta(u_1, ..., u_n) = F_\eta(F_1^{-1}(u_1), ..., F_n^{-1}(u_n))$.

If η is a Gaussian distribution, its unique copula has a density determined entirely by its correlation matrix R, and it is given by $c_\eta(\mathbf{u}) = \det(R)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top[R^{-1} - I]\mathbf{x}\right)$, where $x_{\mathrm{i}} = F_s^{-1}(u_{\mathrm{i}})$ with F_s the standard normal cumulative distribution function. Note that if their coordinates are uncorrelated, then C_η coincides with the independence copula. For Gaussian models, correlation and dependence are equivalent; however, beyond the realm of Gaussianity, this is not the case. Some variables can be uncorrelated but can show dependence on unusually events, as exhibited in financial crises or natural disasters. Unfortunately, as outlined below, the Gaussian copula is not suitable for these kinds of structural dependences.

Dependence between random variables is more complex than just correlation, highlighting an extreme value theory concept: tail dependence [27]. The coefficients of lower and upper tail dependence between two r.v. x_1 and x_2 are defined as $\lambda_l = \lim_{q\to 0} \mathbb{P}\left(x_2 \leq F_2^{-1}(q)|x_1 \leq F_1^{-1}(q)\right)$ and $\lambda_u = \lim_{q\to 1} \mathbb{P}\left(x_2 > F_2^{-1}(q)|x_1 > F_1^{-1}(q)\right)$ [117], where F_{i} denote the cumulative distribution function of x_{i} for $\mathrm{i} = 1, 2$. These coefficients provide asymptotic measures of the dependence in the tails (extreme values), which are isolates of their marginals distributions. For independent continuous r.v. we have that $\lambda_l = \lambda_u = 0$, whereas for variables with correlation $\rho = 1$ we have that $\lambda_l = \lambda_u = 1$. For Gaussian distributions, however, the result is surprising: for $\rho < 1$ we have that $\lambda_l = \lambda_u = 0$.

The above result implies that Gaussian variables are *asymptotically independent*, meaning that the Gaussian assumption does not allow for modelling extreme values dependence. This inability, inherited by any diagonal transformation such as Φ aforementioned, can result in misleading calculations of probabilities over extreme cases. This issue was observed mainly in the 2008 subprime crisis, where the Gaussian dependence structure is pointed out as one of the leading causes, thus evidencing that *the devil is in the tails* [38]. Constructing stochastic processes that account for tail dependence is challenging since, in general, distributions satisfying the consistency conditions are scarce.

4.2 Transport Process

The following definition is one of our main contributions as it allows us to construct non-Gaussian processes as non-parametric regression models.

Definition 4.2.1 Let $T = \{T_{\mathbf{t}} : \mathcal{X}^n \to \mathcal{Y}^n \subseteq \mathbb{R}^n | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ be a collection of measurable maps and $f = \{x_t\}_{t\in\mathcal{T}}$ a stochastic process with law η. We say that T is a f-transport if the push-forward finite-dimensional distributions $\hat{\mathcal{F}} = \{\pi_{\mathbf{t}} := T_{\mathbf{t}}\#\eta_{\mathbf{t}} | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ are consistent and define a stochastic process $g = \{y_t\}_{t\in\mathcal{T}}$ with law π. In this case we say that the maps $T_{\mathbf{t}}$ are f-consistent, and that $T(f) := g$ is a transport process (TP) with law denoted as $T\#\eta := \pi$.

The main idea of the previous definition is to start from a simple stochastic process, one that is easy to simulate, and then to generate another stochastic process that is more complex and more expressive. Since our purpose is to model data through their finite-dimensional laws, our definition implies a correspondence between the laws of the reference process and those of the objective process; for this reason, it is important that the mappings retain the size of the distributions and the respective indexes.

It is straightforward that are many collection of measurable maps that are inconsistent, even in some simple cases. For example, consider the swap maps given by $T_1(x_1) = x_1$, $T_{12}(x_1, x_2) = (x_2, x_1)$ and so on. If f is a heteroscedastic Gaussian process, then we have $F_1(x_1) = \mathcal{N}_1(x_1|0, \sigma_1^2)$ and $F_{12}(x_1, x_2) = \mathcal{N}_2\left((x_1, x_2)|0, \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\right)$. The push-forward distributions are given by $G_1(y_1) = \mathcal{N}_1(x_1|0, \sigma_1^2)$ and $G_{12}(y_1, y_2) = \mathcal{N}_2\left((y_1, y_2)|0, \begin{bmatrix}\sigma_2^2 & \sigma_{12}\\ \sigma_{12} & \sigma_1^2\end{bmatrix}\right)$, and since $\lim_{y_2\to\infty} G_{12}(y_1, y_2) = \mathcal{N}_1(x_1|0, \sigma_2^2) \neq \mathcal{N}_1(x_1|0, \sigma_1^2) = G_1(y_1)$, so we have that T is inconsistent for f. Note that if f is a trivial *i.i.d.* stochastic process, then T is f-consistent.

To be able to use transport processes as regression models, we must be able to define a finitely-parameterised transport T^θ with $\theta \in \Theta \subset \mathbb{R}^{\mathrm{d}}$, where the finite-dimensional maps $(T^\theta)_{\mathbf{t}}$ are consistent and invertible. For example, given $\theta \in \Theta = \mathcal{X}$ the *shift* transport is $T^\theta = \{T_{\mathbf{t}}(\mathbf{x}) = \mathbf{x} + \theta | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$, or simply $(T^\theta)_{\mathbf{t}}(\mathbf{x}) = \mathbf{x} + \theta$. For simplicity, if there is no ambiguity, we will denote $(T^\theta)_{\mathbf{t}}$ as $T_{\mathbf{t}}$. In the next sections, we will show more sophisticated examples of finitely-parameterised transports T^θ, so in what follows we concentrate on explaining the general approach of using TP as regression models.

4.2.1 Learning transport process

As in the GP approach, given observations, the learning task corresponds to finding the *best* transport T^θ, determined by the parameters θ that minimises the negative logarithm of their marginal likelihood (NLL), given below.

Proposition 4.2.2 Let $g = T^\theta(f)$ be a transport process with law $\pi = T^\theta\#\eta$, where η has

finite-dimensional distributions with density denoted $\eta_{\mathbf{t}}$. Given observations $(\mathbf{t}, \mathbf{y})$, if the map $T_{\mathbf{t}}$ is invertible on $\mathbf{y}$ (for simplicity we denote $T_{\mathbf{t}}^{-1}$ as $S_{\mathbf{t}}$) and differentiable on $\mathbf{x} = S_{\mathbf{t}}(\mathbf{y})$, its NLL is given by

$$\begin{aligned} -\log \pi_{\mathbf{t}}(\mathbf{y}|\theta) &= -\log \eta_{\mathbf{t}}(S_{\mathbf{t}}(\mathbf{y})) - \log |\nabla S_{\mathbf{t}}(\mathbf{y})| \\ &= -\log \eta_{\mathbf{t}}(S_{\mathbf{t}}(\mathbf{y})) + \log |\nabla T_{\mathbf{t}}(S_{\mathbf{t}}(\mathbf{y}))|. \end{aligned} \tag{4.1}$$

The first equality is due to the change of variables formula [58]. For the second identity, via the inverse function theorem [112] we have that $\nabla S_{\mathbf{t}}(\mathbf{y}) = \nabla T_{\mathbf{t}}(\mathbf{x})^{-1}$, and by the determinant of the inverse property [99] we get $|\nabla T_{\mathbf{t}}(\mathbf{x})^{-1}| = |\nabla T_{\mathbf{t}}(\mathbf{x})|^{-1}$. To calculate eq. (4.1) we need to be able to compute the log-density of $\eta_{\mathbf{t}}$, the inverse $S_{\mathbf{t}}$, and the gradient $\nabla T_{\mathbf{t}}$ (or $\nabla S_{\mathbf{t}}$).

It is important to note that the reference process is fixed and the trainable object corresponds to transport. In other words, following the principle known as *reparametrisation trick* [66], the model is defined so that random sources have no parameters, so that optimization algorithms can be applied over deterministic parametric functions. Akin to the GP approach, the NLL for transport process (eq. (4.1)) follows an elegant interpretation of how to avoid overfitting:

- The first term $-\log \eta_{\mathbf{t}}(S_{\mathbf{t}}(\mathbf{y}))$ is the *goodness of fit* score between the model and the data, privileging those θ that make $S_{\mathbf{t}}(\mathbf{y})$ to be close to the mode of $\eta_{\mathbf{t}}$. E.g., if $\eta_{\mathbf{t}}$ is a standard Gaussian, this term (omitting a constant) is $\frac{1}{2}\|S_{\mathbf{t}}(\mathbf{y})\|_2^2$, and with enough observations it results in overfitting: $S_{\mathbf{t}}$ is the null function.
- On the other hand, the second term $-\log |\nabla(S_{\mathbf{t}}(\mathbf{y})|$ is the *model complexity penalty*, and it prioritises those θ that make $|\nabla S_{\mathbf{t}}(\mathbf{y})|$ to be large, i.e. $S_{\mathbf{t}}$ has large deviations around $\mathbf{y}$, thus avoiding the null function and, in turn, the overfitting. Note that a valid map satisfies $|\nabla S_{\mathbf{t}}(\mathbf{y})| > 0$.

4.2.2 Inference with transport process

Once the transport T^θ is trained, via minimising the NLL, inference is performed via calculating the posterior distribution of $(\bar{\mathbf{t}}, \bar{\mathbf{y}})$ given observations $(\mathbf{t}, \mathbf{y})$ under the law π: for any inputs $\bar{\mathbf{t}}$ we compute the posterior distributions $\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|\mathbf{y})$. As our goal is to generate stochastic processes more expressive than GPs, the mean and variance are not sufficient to compute (e.g. we need expectations associated with extreme values). For this reason, our approach is based on generating efficiently independent samples from $\pi_{\bar{\mathbf{t}}|\mathbf{t}}$, to then perform calculations via Monte Carlo methods [111].

Since we assume that we can easily obtain samples from $\eta_{\bar{\mathbf{t}}}$ (and $\eta_{\bar{\mathbf{t}}|\mathbf{t}}$ if necessary), we will show how to use these samples and the transport T^θ to efficiently generate samples from $\pi_{\bar{\mathbf{t}}|\mathbf{t}}$. The principle behind this idea is that if $\pi_{\bar{\mathbf{t}}|\mathbf{t}} = \varphi \# \eta_{\bar{\mathbf{t}}}$ and $\mathbf{x} \sim \eta_{\bar{\mathbf{t}}}$ then $\varphi(\mathbf{x}) \sim \pi_{\bar{\mathbf{t}}|\mathbf{t}}$. In cases where this principle can not be applied, we can alternatively obtain samples using methods based on MCMC, which need to be able to evaluate the density of the posterior distribution.

4.3 Marginal Transport

In this section, we present a family of transports named *marginal transports*, given that they can change the marginals distributions of a stochastic process, extending in this way the mean function from GPs, as well as the warping function from WGPs, including the model CWGP presented previously on Chapter 3. We prove their consistency, deliver the formulas for training, and give a general method to sampling.

Definition 4.3.1 $T = \{T_{\mathbf{t}} | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ is a marginal transport if there exists a measurable function $h : \mathcal{T} \times \mathcal{X} \to \mathcal{X}$, so that $[T_{\mathbf{t}}(\mathbf{x})]_{\mathrm{i}} = h(t_{\mathrm{i}}, x_{\mathrm{i}})$ for $\mathbf{t} \in \mathcal{T}^n, \mathbf{x} \in \mathcal{X}^n, n \in \mathbb{N}$. Additionally, if $h(t, \cdot) : \mathcal{X} \to \mathcal{X}$ is increasing (so differentiable a.e.) for all $t \in \mathcal{T}$, then we said that T is a increasing marginal transport.

A *marginal* transport is defined in a *coordinate-wise* manner via the function h. For example, given a *location* function $m : \mathcal{I} \to \mathcal{X}$, then $h(t, x) = m(t) + x$ induces a marginal transport T^h such that if $\eta = \mathcal{GP}(0, k)$ then $T^h \# \eta = \mathcal{GP}(m, k)$. As T^h determinates the mean on the induced stochastic process, usual choices for m are elementary functions like polynomial, exponential, trigonometric and additive/multiplicative combinations.

However, this family of transports is more expressive than just determining the mean, being able to define higher moments such as variance, skewness and kurtosis. This expressiveness can be achieved, beside the *location* function m, by considering a *warping* $\varphi : \mathcal{Y} \to \mathcal{X}$ to define the transport T^h induced by the composite function $h(t, x) = \varphi^{-1}(m(t) + x)$, such that if $\eta = \mathcal{GP}(0, k)$ then we have that $T^h \# \eta = \mathcal{WGP}(\varphi, m, k)$. The most common *warping* functions are affine, logarithm, Box-Cox [107], and sinh-arcsinh [61], which can be composed to generate more expressive warpings. This layers-based model, named compositionally WGP, has been thoroughly studied in previous works [107, 108]. However, the expressiveness of marginal transport is more general since the warping function can change across the coordinates.

4.3.1 Consistency of the marginal transport

Marginal transports are well-defined with a GP reference, in the sense that it always defines a set of consistent finite-dimensional distributions, and thus it induces a stochastic process. The following proposition shows that this family of transports is compatible with any stochastic process, a property which we refer to as *universally consistent*.

Proposition 4.3.2 Given any stochastic process $f = \{x_t\}_{t \in \mathcal{T}}$ and any increasing marginal transport T, then T is an f-transport.

Proof. Given $\eta_{\mathbf{t}} \in \mathcal{F}$ a finite-dimensional distribution, the transported cumulative distribution function is given by $F_{\pi_{\mathbf{t}}}(\mathbf{y}) = F_{\eta_{\mathbf{t}}}((h^{-1}(t_{\mathrm{i}}, y_{\mathrm{i}}))_{\mathrm{i}=1}^n)$, where $h^{-1}(t, \cdot)$ denotes the inverse on the $\mathcal{X}$-coordinate of h, which is also increasing.

The marginalisation condition is fulfilled since $F_{\eta_{\mathbf{t},t_{n+1}}}(\mathbf{x},\infty) = F_{\eta_{\mathbf{t}}}(\mathbf{x})$, so we have

$$\begin{aligned} F_{\pi_{\mathbf{t},t_{n+1}}}(\mathbf{y},\infty) &= F_{\eta_{\mathbf{t},t_{n+1}}}((h^{-1}(t_{\mathrm{i}},y_{\mathrm{i}}))_{\mathrm{i}=1}^{n}, h^{-1}(t_{n+1},\infty)), \\ &= F_{\eta_{\mathbf{t},t_{n+1}}}((h^{-1}(t_{\mathrm{i}},y_{\mathrm{i}}))_{\mathrm{i}=1}^{n},\infty) = F_{\eta_{\mathbf{t}}}((h^{-1}(t_{\mathrm{i}},y_{\mathrm{i}}))_{\mathrm{i}=1}^{n}) = F_{\pi_{\mathbf{t}}}(\mathbf{y}). \end{aligned}$$

Given an n-permutation τ, we denote $\tau(\mathbf{t}) = t_{\tau(1)},...,t_{\tau(n)}$ and $\tau(\mathbf{y}) = y_{\tau(1)},...,y_{\tau(n)}$. Since $F_{\eta_{\tau(\mathbf{t})}}(\tau(\mathbf{x})) = F_{\eta_{\mathbf{t}}}(\mathbf{x})$ then $F_{\pi_{\tau(\mathbf{t})}}(\tau(\mathbf{y})) = F_{\eta_{\tau(\mathbf{t})}}((h^{-1}(t_{\tau(\mathrm{i})},y_{\tau(\mathrm{i})}))_{\mathrm{i}=1}^{n}) = F_{\eta_{\mathbf{t}}}((h^{-1}(t_{\mathrm{i}},y_{\mathrm{i}}))_{\mathrm{i}=1}^{n}) = F_{\pi_{\mathbf{t}}}(\mathbf{y})$, satisfying the conditions. □

Remark 4.3.3. In general we will assume that marginal transports are increasing, due to for any fixed stochastic process f and any marginal transport T, exist an increasing marginal transport T^h such that $T\#f$ and $T^h\#f$ have the same distributions (i.e. all their finite-dimensional distributions agree [121]). The increasing function h is defined via the unique monotone transport maps from η_t to π_t given by $h(t,x) = F_{\pi_t}^{-1}(F_{\eta_t}(x))$ for each $t \in \mathcal{T}$ [30].

Marginal transports T^h satisfy straightforwardly the consistency condition since there are coordinate-wise maps. This *diagonality* is an appealing mathematical property, but it has a high cost: the transport process inherits the same copula from the reference process. This fact implies that independent marginals, such as white noise, remain independent with the marginal transport. The following proposition shows the benefits and limitations of diagonality [144].

Proposition 4.3.4 Let $f = \{x_t\}_{t\in\mathcal{T}}$ be a stochastic process with marginal cumulative distribution functions F_t for $t \in \mathcal{T}$, and copula process C. Given any sequence of cumulative distribution functions $\{G_t\}_{t\in\mathcal{I}}$, the function $h(t,x) = G_t^{-1}(F_t(x))$ induces a marginal transport T^h where $g = T^h\#f$ is a transport process with marginals G_t and copula process C.

Proof. The copula of f is the stochastic process $C = \{C_t\}_{t\in\mathcal{T}}$ where $C_t := F_t(x_t)$ follows a uniform distribution. The transport process $g = T^h\#f = \{y_t\}_{t\in\mathcal{T}}$ satisfies $y_t = G_t^{-1}(F_t(x_t)) = G_t^{-1}(C_t)$, so its copula process $D = \{D_t\}_{t\in\mathcal{T}}$ is given by $D_t = G_t(y_t) = G_t(G_t^{-1}(C_t)) = C_t$. Thus, f and g have the same copula. □

4.3.2 Learning of the marginal transport

For learning we have to calculate the NLL given by eq. (4.1). The inverse map is given by $S_{\mathbf{t}}(\mathbf{y})_{\mathrm{i}} = h^{-1}(t_{\mathrm{i}},y_{\mathrm{i}}) = x_{\mathrm{i}}$ and the *model complexity penalty* is given by

$$\log|\nabla S_{\mathbf{t}}(\mathbf{y})| = \sum_{\mathrm{i}} \log\frac{\partial h^{-1}}{\partial y}(t_{\mathrm{i}},y_{\mathrm{i}}) = -\sum_{\mathrm{i}} \log\frac{\partial h}{\partial y}(t_{\mathrm{i}},x_{\mathrm{i}}). \tag{4.2}$$

E.g., if $h(t,x) = \varphi^{-1}(m(t)+\sigma(t)x)$, then $h(t,y)^{-1} = \frac{\varphi(y)-m(t)}{\sigma(t)}$ and $\log|\nabla S_{\mathbf{t}}(\mathbf{y})| = \sum_{\mathrm{i}} \log\frac{\varphi'(y_{\mathrm{i}})}{\sigma(t_{\mathrm{i}})}$.

4.3.3 Inference with marginal transport

For inference on new inputs $\bar{\mathbf{t}}$, the posterior distribution $\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|\mathbf{y})$ is the push-forward of $\eta_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|S_{\mathbf{t}}(\mathbf{y}))$ by $T_{\bar{\mathbf{t}}}$, so if $\bar{\mathbf{x}} \sim \eta_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|S_{\mathbf{t}}(\mathbf{y}))$ then $\bar{\mathbf{y}} = T_{\bar{\mathbf{t}}}(\bar{\mathbf{x}}) \sim \pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y})$. Note that the probability of a set E under the density of π_t is equal to the probability of the image $h_t^{-1}(E)$ under the density of η_t, where $h_t(\cdot) := h_\theta(t, \cdot)$. Thus, if we can compute marginals quantiles under $\eta_{\mathbf{t}}$, such as the median and confidence intervals, we can do the same under $\pi_{\mathbf{t}}$. Even more, the expectation of any measurable function $v : \mathcal{Y} \to \mathbb{R}$ under the law $\pi_{\mathbf{t}}(\mathbf{y})$ is given by $\mathbb{E}_{\pi_{\mathbf{t}}}[v(\mathbf{y})] = \mathbb{E}_{\eta_{\mathbf{t}}}[v(h_{\mathbf{t}}(\mathbf{x}))]$.

4.4 Covariance Transport

From the results of the previous section, the only way to induce a different copula under our transport-based approach is to consider non-diagonal maps. The problem with these maps is that we lose the property of *universally consistent*, but it is possible to find conditions over the reference stochastic processes so that the transport is consistent.

In this section, we present a family of transports named *covariance transports*, that allows us to change the covariance, and therefore the correlation, over the induced stochastic process. These transports are based on covariance kernels, e.g. the squared exponential given by $k(t, s) = \sigma^2 \exp(-r|t - s|^2)$ with parameters $\theta = (\sigma, r)$.

Definition 4.4.1 $T^k = \{T_{\mathbf{t}} | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ is a covariance transport if there exists a covariance kernel $k : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, so that $T_{\mathbf{t}}(\mathbf{x}) = L_{\mathbf{t}}\mathbf{x}$, where $L_{\mathbf{t}}$ is a square root of $\Sigma_{\mathbf{tt}} = k(\mathbf{t}, \mathbf{t})$, i.e. $L_{\mathbf{t}} L_{\mathbf{t}}^\top = \Sigma_{\mathbf{tt}}$.

Since $\Sigma_{\mathbf{tt}}$ is a definite positive matrix, always exist an unique definite positive square root denoted $\Sigma_{\mathbf{tt}}^{1/2}$ and named the *principal square root* of $\Sigma_{\mathbf{tt}}$. Additionally, always exist an unique lower triangular square root denoted $\text{chol}(\Sigma)$ and named as the *lower Cholesky decomposition* of $\Sigma_{\mathbf{tt}}$, where later we will show his importance to getting practical transports.

If T^k is a covariance transport induced by k and $f \sim \mathcal{GP}(0, \delta(t, \bar{t}))$ is a Gaussian white noise process, then we have that T^k is a f-transport where $T^k(f) \sim \mathcal{GP}(0, k)$, i.e. T^k fully defines the covariance over the transport process. This fact is true due to the maps $T_{\mathbf{t}}(\mathbf{x})$ being linear (given by $T_{\mathbf{t}}(\mathbf{x})_{\mathrm{i}} = \sum_{j=1}^n l_{\mathrm{i}j} x_j$ where $[L_{\mathbf{t}}]_{\mathrm{i}j} = l_{\mathrm{i}j}$), so given a finite-dimensional law $\eta_{\mathbf{t}} =\sim \mathcal{N}_n(0, I)$, by the linear closure of Gaussian distributions we have that $T_{\mathbf{t}} \# \eta_{\mathbf{t}} = \mathcal{N}_n(0, \Sigma_{\mathbf{tt}})$ where $L_{\mathbf{t}} L_{\mathbf{t}}^\top = \Sigma_{\mathbf{tt}} = k_\theta(\mathbf{t}, \mathbf{t})$. We assume for now the consistency of the covariance transport, but we will study it at the end of this section, once we have revised the concept of triangularity.

4.4.1 Learning of the covariance transport

We say that a finite-dimensional map $T_{\mathbf{t}} : \mathbb{R}^n \to \mathbb{R}^n$ is *triangular* if it structure is triangular, in the sense $T_{\mathbf{t}}(\mathbf{x})_{\mathrm{i}} = T_{\mathrm{i}}(x_1, ..., x_{\mathrm{i}})$ for $\mathrm{i} = 1, ..., n$. If $T_{\mathbf{t}}$ is differentiable, then it is triangular if and only if its Jacobian $\nabla T_{\mathbf{t}}$ is a lower triangular matrix. We say that a transport T is *triangular* if its finite-dimensional maps are triangular. While a marginal transport is *diagonal*, a covariance transport with lower Cholesky decomposition is *triangular*. Note that diagonal maps are also triangular maps, and the composition of triangular maps remains triangular. Triangularity is an appealing property for maps, since it allows us to perform calculations more efficiently that in the general case. The following result shows the similarity between triangular and diagonal maps for the learning task.

Proposition 4.4.2 Let $T_{\mathbf{t}}$ be an invertible and differentiable triangular map on $\mathbf{x}$. If we denote $T_{\mathbf{t}}(\mathbf{x}) = \mathbf{y}$ then:

- the inverse map $S_{\mathbf{t}}$ is also triangular that fulfills that $S_{\mathbf{t}}(\mathbf{y}) = \mathbf{x}$,
- the model complexity penalty is given by

$$\log |\nabla S_{\mathbf{t}}(\mathbf{y})| = \sum_{\mathrm{i}} \log \frac{\partial S_{\mathrm{i}}}{\partial y_{\mathrm{i}}}(y_1, ..., y_{\mathrm{i}}) = -\sum_{\mathrm{i}} \log \frac{\partial T_{\mathrm{i}}}{\partial x_{\mathrm{i}}}(x_1, ..., x_{\mathrm{i}}).$$

Proof. The first coordinate satisfies $T_1(x_1) = y_1$ so $S_1(y_1) = x_1$. By induction, we have $S_k(y_1, ..., y_k) = x_k$, and since $T_{k+1}(x_1, ..., x_{k+1}) = y_{k+1}$, then we have the equation

$$T_{k+1}(S_1(y_1), ..., S_k(y_1, ..., y_k), x_{k+1}) = y_{k+1},$$

so we can express x_{k+1} in function of $y_1, ..., y_{k+1}$, i.e. $S_{k+1}(y_1, ..., y_{k+1}) = x_{k+1}$ so $S_{\mathbf{t}}$ is triangular. With this we have that $\nabla S_{\mathbf{t}}(\mathbf{y})$ is a lower triangular matrix, so its determinant is equal to the product of all the elements on the diagonal. The complexity penalty, then, is analogous to the diagonal case. □

For triangular covariance transports we have that $S_{\mathbf{t}}(\mathbf{y}) = L_{\mathbf{t}}^{-1}\mathbf{y}$, which can be computed straightforwardly via forward substitution [36], and $\log |\nabla S_{\mathbf{t}}(\mathbf{y})| = -\sum_{\mathrm{i}} \log l_{\mathrm{ii}}$, where l_{ii} are the diagonal values of $L_{\mathbf{t}}$.

4.4.2 Inference with the covariance transport

Triangular maps allow efficient inference since posterior distributions can be calculated as a push-forward from the reference.

Proposition 4.4.3 Given observations $\mathbf{y} \sim \pi_{\mathbf{t}}$, denote $\mathbf{x} = T_{\mathbf{t}}^{-1}(\mathbf{y})$ and by $\eta_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{x}}|\mathbf{x})$ the posterior distribution of η. Assume that the transports $T_{\mathbf{t}}$ are triangular, then the posterior distribution of π is given by

$$\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y}) = \left[P_{\bar{\mathbf{t}}} \circ T^{\mathbf{x}}_{\mathbf{t},\bar{\mathbf{t}}}\right] \# \eta_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|\mathbf{x}), \tag{4.3}$$

where $T^{\mathbf{x}}_{\mathbf{t},\bar{\mathbf{t}}}(\cdot) = T_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{x}, \cdot)$, and $P_{\bar{\mathbf{t}}}(\cdot)$ is the projection on $\bar{\mathbf{t}}$, i.e. $P_{\bar{\mathbf{t}}}(\mathbf{x}, \bar{\mathbf{x}}) = \bar{\mathbf{x}}$.

Proof. Since the maps are triangular, their inverses also are triangular:

$$T^{-1}_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{y},\bar{\mathbf{y}}) = [T^{-1}_{\mathbf{t}}(\mathbf{y}), T^{-1}_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|T^{-1}_{\mathbf{t}}(\mathbf{y}))],$$

and as its gradient it is also triangular, then their determinants satisfy

$$|\nabla T^{-1}_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{y},\bar{\mathbf{y}})| = |\nabla T^{-1}_{\mathbf{t}}(\mathbf{y})||\nabla_{\bar{\mathbf{y}}} T^{-1}_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|T^{-1}_{\mathbf{t}}(\mathbf{y}))|.$$

With these identities, the posterior density of $\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y})$ is given by

$$\begin{aligned}
\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y}) =& \frac{\pi_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{y},\bar{\mathbf{y}})}{\pi_{\mathbf{t}}(\mathbf{y})} = \frac{\eta_{\mathbf{t},\bar{\mathbf{t}}}(T^{-1}_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{y},\bar{\mathbf{y}}))|\nabla T^{-1}_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{y},\bar{\mathbf{y}})|}{\eta_{\mathbf{t}}(T^{-1}_{\mathbf{t}}(\mathbf{y}))|\nabla T^{-1}_{\mathbf{t}}(\mathbf{y})|}, \\
=& \frac{\eta_{\mathbf{t},\bar{\mathbf{t}}}(T^{-1}_{\mathbf{t}}(\mathbf{y}), T^{-1}_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|T^{-1}_{\mathbf{t}}(\mathbf{y})))}{\eta_{\mathbf{t}}(T^{-1}_{\mathbf{t}}(\mathbf{y}))} \frac{|\nabla T^{-1}_{\mathbf{t}}(\mathbf{y})||\nabla_{\bar{\mathbf{y}}} T^{-1}_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|T^{-1}_{\mathbf{t}}(\mathbf{y}))|}{|\nabla T^{-1}_{\mathbf{t}}(\mathbf{y})|}, \\
=& \eta_{\bar{\mathbf{t}}|\mathbf{t}}(T^{-1}_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|T^{-1}_{\mathbf{t}}(\mathbf{y}))|T^{-1}_{\mathbf{t}}(\mathbf{y}))|\nabla_{\bar{\mathbf{y}}} T^{-1}_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|T^{-1}_{\mathbf{t}}(\mathbf{y}))|, \\
=& T_{\mathbf{t},\bar{\mathbf{t}}}(T^{-1}_{\mathbf{t}}(\mathbf{y}),\cdot)|_{\bar{\mathbf{t}}} \# \eta_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|T^{-1}_{\mathbf{t}}(\mathbf{y})) = \left[P_{\bar{\mathbf{t}}} \circ T^{\mathbf{x}}_{\mathbf{t},\bar{\mathbf{t}}}\right] \# \eta_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|\mathbf{x}).
\end{aligned}$$

□

For the covariance transport, and given new inputs $\bar{\mathbf{t}}$, the posterior distribution $\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y})$ is the push-forward of $\eta_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|L^{-1}_{\mathbf{t}}\mathbf{y})$ by the affine map $T(\mathbf{u}) = A_{\mathbf{t}}L^{-1}_{\mathbf{t}}\mathbf{y} + A_{\bar{\mathbf{t}}}\mathbf{u}$, where $L_{\mathbf{t},\bar{\mathbf{t}}} = \begin{bmatrix} L_{\mathbf{t}} & 0 \\ A_{\mathbf{t}} & A_{\bar{\mathbf{t}}} \end{bmatrix}$. Note that $A_{\mathbf{t}}L^{-1}_{\mathbf{t}} = \Sigma_{\bar{\mathbf{t}}\mathbf{t}}\Sigma^{-1}_{\mathbf{t}\mathbf{t}}$ and $A_{\bar{\mathbf{t}}}A^{\top}_{\bar{\mathbf{t}}} = \Sigma_{\bar{\mathbf{t}}\bar{\mathbf{t}}} - \Sigma_{\bar{\mathbf{t}}\mathbf{t}}\Sigma^{-1}_{\mathbf{t}\mathbf{t}}\Sigma_{\bar{\mathbf{t}}\mathbf{t}}$, so the map agrees with $T(\mathbf{u}) = \Sigma_{\bar{\mathbf{t}}\mathbf{t}}\Sigma^{-1}_{\mathbf{t}\mathbf{t}}\mathbf{y} + L_{\bar{\mathbf{t}}|\mathbf{t}}\mathbf{u}$, where $L_{\bar{\mathbf{t}}|\mathbf{t}} = \text{chol}(\Sigma_{\bar{\mathbf{t}}|\mathbf{t}})$ with $\Sigma_{\bar{\mathbf{t}}|\mathbf{t}} = \Sigma_{\bar{\mathbf{t}}\bar{\mathbf{t}}} - \Sigma_{\bar{\mathbf{t}}\mathbf{t}}\Sigma^{-1}_{\mathbf{t}\mathbf{t}}\Sigma_{\bar{\mathbf{t}}\mathbf{t}}$.

4.4.3 Consistency of the covariance transport

Going back to the issue of consistency, the following proposition gives us a condition over triangular maps that imply consistency under marginalisation.

Proposition 4.4.4 Let $T = \{T_{\mathbf{t}} : \mathcal{X}^n \to \mathcal{X}^n | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ be a collection of triangular measurable maps that satisfy $P_{\mathbf{t}} \circ T_{\mathbf{t},t_{n+1}}(\mathbf{y}, y_{n+1}) = T_{\mathbf{t}}(\mathbf{y})$, with $P_{\mathbf{t}}$ the projection on $\mathbf{t}$. Then T is universally consistent under marginalisation.

Proof. The push-forward finite-dimensional distribution function is $F_{\pi_{\mathbf{t}}}(\mathbf{y}) = F_{\eta_{\mathbf{t}}}(S_{\mathbf{t}}(\mathbf{y}))$. Since a valid map satisfies $\frac{\partial S_{\mathrm{i}}}{\partial y_{\mathrm{i}}}(y_1, ..., y_{\mathrm{i}}) > 0$ for all $\mathrm{i} \geq 1$, then $S_{t_{n+1}}$ is increasing on y_{n+1} so $S_{t_{n+1}}(\mathbf{y}, \infty) = \infty$. With this, if $P_{\mathbf{t}} \circ T_{\mathbf{t},t_{n+1}}(\mathbf{y}, y_{n+1}) = T_{\mathbf{t}}(\mathbf{y})$ then the inverse also satisfies this. Finally, the marginalisation condition is fulfilled becauses $F_{\pi_{\mathbf{t},t_{n+1}}}(\mathbf{y}, \infty) = F_{\eta_{\mathbf{t},t_{n+1}}}(S_{\mathbf{t},t_{n+1}}(\mathbf{y}, \infty)) = F_{\eta_{\mathbf{t},t_{n+1}}}(S_{\mathbf{t}}(\mathbf{y}), S_{t_{n+1}}(\mathbf{y}, \infty)) = F_{\eta_{\mathbf{t},t_{n+1}}}(S_{\mathbf{t}}(\mathbf{y}), \infty) = F_{\eta_{\mathbf{t}}}(S_{\mathbf{t}}(\mathbf{y})) = F_{\pi_{\mathbf{t}}}(\mathbf{y})$. □

Note that diagonal and covariance transports satisfy the above condition, that can be interpreted like an *order* between their finite-dimensional triangular maps. The consistency under permutations means that, given any n-permutation τ, it satisfies $F_{\pi_{\tau(\mathbf{t})}}(\tau(\mathbf{y})) = F_{\pi_{\mathbf{t}}}(\mathbf{y})$,

or equivalently, $F_{\eta_{\tau(\mathbf{t})}}(S_{\tau(\mathbf{t})}(\tau(\mathbf{y}))) = F_{\eta_\mathbf{t}}(S_\mathbf{t}(\mathbf{y}))$. Since η is consistent under permutations, we have the following condition over $\eta_\mathbf{t}$ and $S_\mathbf{t}$:

$$F_{\eta_\mathbf{t}}(\tau^{-1}(S_{\tau(\mathbf{t})}(\tau(\mathbf{y})))) = F_{\eta_\mathbf{t}}(S_\mathbf{t}(\mathbf{y})). \tag{4.4}$$

The above equality can be written in terms of the density function as

$$\eta_\mathbf{t}(\tau^{-1}(S_{\tau(\mathbf{t})}(\tau(\mathbf{y})))) \left|\nabla(\tau^{-1}(S_{\tau(\mathbf{t})}(\tau(\mathbf{y}))))\right| = \eta_\mathbf{t}(S_\mathbf{t}(\mathbf{y})) \left|\nabla S_\mathbf{t}(\mathbf{y})\right|. \tag{4.5}$$

Note that if T is *universally* consistent under permutations, then it has to satisfy $\tau(S_\mathbf{t}(\mathbf{y})) = S_{\tau(\mathbf{t})}(\tau(\mathbf{y}))$, so T must be diagonal. This mean that strictly triangular transports can be consistent only for some families of distributions. The following proposition shows one condition over η for consistency of covariance transports.

Proposition 4.4.5 Let $f = \{x_t\}_{t\in\mathcal{T}}$ be a stochastic process where its finite-dimensional laws have densities with the form $\eta_\mathbf{t}(\mathbf{x}) = \beta_n(\|\mathbf{x}\|_2)$, for some functions β_n with $n = |\mathbf{t}|$. Then any triangular covariance transport T^k is an f-transport.

Proof. We just need to check consistency under permutations. We have that $S_\mathbf{t}(\mathbf{y}) = L_\mathbf{t}^{-1}\mathbf{y}$, so $|\nabla S_\mathbf{t}(\mathbf{y})| = |L_\mathbf{t}|^{-1} = \prod_\mathrm{i} l_{\mathrm{ii}}^{-1}$, where l_{ii} are the diagonal values of $L_\mathbf{t}$. Note that this calculation is independent of $\mathbf{y}$ and it only depends on the values of the diagonal, so $\left|\nabla(\tau^{-1}(S_{\tau(\mathbf{t})}(\tau(\mathbf{y}))))\right| = |L_{\tau(\mathbf{t})}|^{-1} = \prod_\mathrm{i} \mathrm{d}_{\mathrm{ii}}^{-1}$, where d_{ii} are the diagonal values of $L_{\tau(\mathbf{t})}$. Since $|\Sigma_{\mathbf{tt}}| = |L_\mathbf{t}|^2$ and $|\Sigma_{\tau(\mathbf{t})\tau(\mathbf{t})}| = |P_\tau \Sigma_{\mathbf{tt}} P_\tau| = |\Sigma_{\mathbf{tt}}|$ then we have that $|L_{\tau(\mathbf{t})}| = |L_\mathbf{t}|$. With this identity, we need that $\eta_\mathbf{t}(\tau^{-1}(L_{\tau(\mathbf{t})}^{-1}\tau(\mathbf{y}))) = \eta_\mathbf{t}(L_\mathbf{t}^{-1}\mathbf{y})$, but this is fulfilled under the hypothesis over $\eta_\mathbf{t}$, since

$$\begin{aligned}\eta_\mathbf{t}(\tau^{-1}(S_{\tau(\mathbf{t})}(\tau(\mathbf{y})))) = \beta_n\left(\left\|\tau^{-1}(L_{\tau(\mathbf{t})}^{-1}\tau(\mathbf{y}))\right\|_2\right) &= \beta_n\left(\tau(\mathbf{y})^\top \Sigma_{\tau(\mathbf{t})\tau(\mathbf{t})}^{-1}\tau(\mathbf{y}))\right) \\ &= \beta_n\left(\mathbf{y}\Sigma_{\mathbf{tt}}^{-1}\mathbf{y}\right) = \eta_\mathbf{t}(L_\mathbf{t}^{-1}\mathbf{y}).\end{aligned}$$

□

Note that the standard Gaussian distribution satisfies the hypothesis with $\beta_n(r) = c_n \exp(-r^2/2)$ where $c_n = (2\pi)^{-n/2}$. This family of distributions is known in the literature as spherical distributions, and their generalisation with covariance is known as elliptical distributions [91]. In the next section, we will study these distributions via a new type of transports.

4.5 Radial Transports

While covariance and marginal transports can model correlation and marginals, they inherit the base copula from the reference. For example, if the reference process is a GP, through covariance and marginal transports we can only generate WGP with Gaussian copulas. Our proposal to construct other copulas relies on radial transformations that are capable of modifying the norm of a random vector, changing its copula in this way.

Definition 4.5.1 $T = \{T_{\mathbf{t}} | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ is a radial transport if there exists a radial function $\phi(r) = \frac{\alpha(r)}{r}$, with $\alpha : \mathbb{R}^+ \to \mathbb{R}^+$ monotonically non-decreasing, and $\|\cdot\|$ a norm over $\mathcal{X}^n$ so that $T_{\mathbf{t}}(\mathbf{x}) = \phi(\|\mathbf{x}\|)\mathbf{x}$.

According to the chosen norm $\|\cdot\|$, the copula family generated by our approach is different. The Euclidean ℓ_2 norm, $\|\cdot\|_2$, allows us to define elliptical processes; the Manhattan ℓ_1 norm, $\|\cdot\|_1$, allows us to define Archimedean processes. In the following sections we will study these respective *elliptical transports* and *Archimedean transports*.

4.5.1 Elliptical processes

In the previous section, we introduced a particular family of distributions known as spherical distributions that are consistent with covariance transport. We now introduce a generalisation called elliptical distributions [91].

Definition 4.5.2 $\mathbf{x} \in \mathbb{R}^n$ is elliptically distributed iff there exists a vector $\mu \in \mathbb{R}^n$, a (symmetric) full rank scale matrix $A \in \mathbb{R}^{n \times n}$, a uniform random variable $U^{(n)}$ on the unit sphere in $\mathbb{R}^n$, i.e. $\left\|U^{(n)}\right\|_2 = 1$, and a real non-negative random variable $R \in \mathbb{R}^+$, independent of $U^{(n)}$, such that $\mathbf{x} \overset{\mathrm{d}}{=} \mu + RAU^{(n)}$, where $\overset{\mathrm{d}}{=}$ denotes equality in distribution.

Remark 4.5.3. If $\mathbf{x}$ is elliptically distributed and has density $\eta(\mathbf{x})$, then for some positive function β_n, it has the form $\eta(\mathbf{x}) = |\Sigma|^{-1/2} \beta_n((\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu))$, where $\Sigma = A^\top A$ and R has density $p_R(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1} \beta_n(r^2)$ [91].

Gaussian distributions are members of elliptical distributions: if $\mathbf{x} \sim \mathcal{N}_n(0, \Sigma_{\mathbf{xx}})$ then $\mathbf{x} \overset{\mathrm{d}}{=} R_n L_{\mathbf{t}} U^{(n)}$ with $R_n \sim \sqrt{\chi^2(n)}$ (i.e. follow a Rayleigh distribution) and $\Sigma_{\mathbf{xx}} = L_{\mathbf{t}}^\top L_{\mathbf{t}}$. However elliptical distributions include other distributions like the Student-t [35], a widely-used alternative due to its heavy-tail behaviour. Elliptical processes have a useful characterisation as follows:

Theorem 4.5.4 (Kelker's theorem [62]) f is an elliptical process where the finite-dimensional marginals $\mathbf{x}$ have density if and only if there exists a positive random variable R such that $\mathbf{x}|R \sim \mathcal{N}_n(\mu_{\mathbf{x}}, R\Sigma_{\mathbf{xx}})$.

The above result can be summarised in that elliptic processes are mixtures of Gaussian processes. This characterisation gives us a direction to achieve our goal through radial transports.

Elliptical transport

Our goal is to define stochastic processes via our transport approach where their copula is elliptical, beyond the Gaussian case. Let us set some notation. Given a r.v. R, its cumulative distribution function is denoted F_R. The square-root of a chi-squared (a.k.a. Rayleigh) distributed r.v. will be denoted $R_n \sim \sqrt{\chi^2(n)}$. Our idea to transport a Gaussian copula to

another elliptical copula is based on the following optimal transport result [30, 50].

Proposition 4.5.5 Let $\mathbf{x} \overset{\mathrm{d}}{=} RAU^{(n)}$ be an elliptically distributed r.v. Given a positive r.v. S, consider the radial map $T^{\alpha}(\mathbf{x}) = \phi(\|\mathbf{x}\|_2)\mathbf{x} = \frac{\alpha(\|A^{-1}\mathbf{x}\|_2)}{\|A^{-1}\mathbf{x}\|_2}\mathbf{x}$ where $\alpha(r) = F_S^{-1}(F_R(r))$. Then we have that $T^{\alpha}(\mathbf{x}) \overset{\mathrm{d}}{=} SAU^{(n)}$.

A useful property of this type of transports is that we can generate distributions with different elliptical copulas by changing the norm without altering the correlation.

Lemma 4.5.6 The radial transport T^{α} does not modify the correlation.

Proof. Let $\mathbf{x} \overset{\mathrm{d}}{=} RAU^{(n)}$. Then, $Cov(\mathbf{x}) = \frac{\mathbb{E}(R^2)}{rank(A)}A^{\top}A = c\Sigma$. As $\mathbf{y} =: T_{\mathbf{t}}(\mathbf{x}) \overset{\mathrm{d}}{=} \alpha(R)AU^{(n)}$ then $Cov(\mathbf{y}) = \frac{\mathbb{E}(\alpha(R)^2)}{rank(A)}A^{\top}A = \mathrm{d}\Sigma$. As $Cov(\mathbf{y}) = \frac{\mathrm{d}}{c}Cov(\mathbf{x})$, we have $Corr(\mathbf{y}) = Corr(\mathbf{x})$. □

Note that if $\mathbf{x} \overset{\mathrm{d}}{=} RU^{(n)}$ then $T^{\alpha}(\mathbf{x}) = \phi(\|\mathbf{x}\|_2)A\mathbf{x} \overset{\mathrm{d}}{=} \alpha(R)AU^{(n)}$. Since we can decompose $T^{\alpha}(\mathbf{x}) = A(\phi(\|\mathbf{x}\|_2)\mathbf{x})$ in a covariance transport, we merely consider the elliptical transport as $T_{\mathbf{t}}(\mathbf{x}) = \phi(\|\mathbf{x}\|_2)\mathbf{x}$. The next result characterises a family of transports based on radial functions that generate elliptical processes from Gaussian white noise processes.

Theorem 4.5.7 Let p_{θ} be a density function supported on positive real line. Define $F_{R_{n,\theta}}(r) := \int_0^{\infty} p_{\theta}(s)F_{R_n}(r/s)\mathrm{d}s$ and $\alpha_{n,\theta}(r) = F^{-1}_{R_{n,\theta}} \circ F_{R_n}(r)$. Then the elliptical radial transport defined by $T_{\mathbf{t}}(\mathbf{x}) := \frac{\alpha_{n,\theta}(\|\mathbf{x}\|_2)}{\|\mathbf{x}\|_2}\mathbf{x}$ is an f-transport with $f \sim \mathcal{GP}(0, \delta(t, \bar{t}))$, where the transport process $g := T(f)$ has finite-dimensional elliptical distributions.

Proof. Let R_{θ} be a positive r.v. with density function p_{θ}. Since $R_n \sim \sqrt{\chi^2(n)}$ is also a positive r.v., by the product distribution formula [109] we have that the r.v. $R_{n,\theta} := R_{\theta}R_n$ has a cumulative distribution function given by $F_{R_{n,\theta}}(r) := \int_0^{\infty} p_{\theta}(s)F_{R_n}(r/s)\mathrm{d}s$. Given that the finite-dimensional laws of f are $\eta_{\mathbf{t}} = \mathcal{N}_n(0, I)$, if $\mathbf{x} \sim \eta_{\mathbf{t}}$, then $\|\mathbf{x}\|_2 \overset{\mathrm{d}}{=} R_n$, so $\alpha_{n,\theta}(\|\mathbf{x}\|_2) \overset{\mathrm{d}}{=} R_{n,\theta} \overset{\mathrm{d}}{=} R_{\theta}R_n$ and $\frac{\mathbf{x}}{\|\mathbf{x}\|_2} \overset{\mathrm{d}}{=} U^{(n)}$ are independent, having thus that $T_{\mathbf{t}}(\mathbf{x}) \overset{\mathrm{d}}{=} R_{\theta}R_nU^{(n)}$ is elliptically distributed. Since $T_{\mathbf{t}}(\mathbf{x})|R_{\theta} \sim \mathcal{N}_n(0, R_{\theta}^2 I)$ and R_{θ} is independent of $\mathbf{x}$, by Kelker's theorem the push-forward finite-dimensional distributions $\hat{\mathcal{F}} = \{T_{\mathbf{t}}\#\eta_{\mathbf{t}}|\mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$ are consistent and define an elliptical process. □

Learning of the elliptical transport

The following proposition allow us to calculate the determinant of the gradient of this radial transport.

Proposition 4.5.8 Let $T_{\mathbf{t}}(\mathbf{x}) = \phi(\|\mathbf{x}\|_2)\mathbf{x} = \frac{\alpha(\|\mathbf{x}\|_2)}{\|\mathbf{x}\|_2}\mathbf{x}$. Then $|\nabla T_{\mathbf{t}}(\mathbf{x})| = \phi(\|\mathbf{x}\|_2)^{n-1}\alpha'(\|\mathbf{x}\|_2)$.

Proof.

$$\frac{\partial T_{\mathbf{t}}(\mathbf{x})_{\mathrm{i}}}{\partial x_{\mathrm{i}}} = \phi(\|\mathbf{x}\|_2) + \phi'(\|\mathbf{x}\|_2)\frac{x_{\mathrm{i}}^2}{\|\mathbf{x}\|_2},$$
$$\frac{\partial T_{\mathbf{t}}(\mathbf{x})_{\mathrm{i}}}{\partial x_j} = \phi'(\|\mathbf{x}\|_2)\frac{x_{\mathrm{i}}x_j}{\|\mathbf{x}\|_2}, \text{if i} \neq j,$$
$$\nabla T_{\mathbf{t}}(\mathbf{x}) = \frac{\phi'(\|\mathbf{x}\|_2)}{\|\mathbf{x}\|_2}\left[\mathbf{x}\mathbf{x}^\top + I\frac{\phi(\|\mathbf{x}\|_2)\,\|\mathbf{x}\|_2}{\phi'(\|\mathbf{x}\|_2)}\right], \text{and,}$$
$$|\nabla T_{\mathbf{t}}(\mathbf{x})| = \left(\frac{\phi'(\|\mathbf{x}\|_2)}{\|\mathbf{x}\|_2}\right)^n \left|\mathbf{x}\mathbf{x}^\top + I\frac{\phi(\|\mathbf{x}\|_2)\,\|\mathbf{x}\|_2}{\phi'(\|\mathbf{x}\|_2)}\right|.$$

By Sylvester's determinant theorem we have

$$\left|\mathbf{x}\mathbf{x}^\top + I\frac{\phi(\|\mathbf{x}\|_2)\,\|\mathbf{x}\|_2}{\phi'(\|\mathbf{x}\|_2)}\right| = \left(1 + \frac{\phi'(\|\mathbf{x}\|_2)}{\phi(\|\mathbf{x}\|_2)\,\|\mathbf{x}\|_2}\,\|\mathbf{x}\|_2^2\right)\left(\frac{\phi(\|\mathbf{x}\|_2)\,\|\mathbf{x}\|_2}{\phi'(\|\mathbf{x}\|_2)}\right)^n$$
$$|\nabla T_{\mathbf{t}}(\mathbf{x})| = \phi(\|\mathbf{x}\|_2)^{n-1}\left(\phi(\|\mathbf{x}\|_2) + \phi'(\|\mathbf{x}\|_2)\|\mathbf{x}\|_2\right)$$

and since $\alpha(r) = \phi(r)r$ and $\alpha'(r) = \phi(r)+\phi'(r)r$, we have $|\nabla T_{\mathbf{t}}(\mathbf{x})| = \phi(\|\mathbf{x}\|_2)^{n-1}\alpha'(\|\mathbf{x}\|_2)$. □

For the learning task, since $|\nabla T_{\mathbf{t}}(\mathbf{x})| = \phi_{n,\theta}(\|\mathbf{x}\|_2)^{n-1}\alpha'_{n,\theta}(\|\mathbf{x}\|_2)$ and $T_{\mathbf{t}}^{-1}(\mathbf{y}) = \psi_{n,\theta}(\|\mathbf{y}\|_2)\mathbf{y} = \frac{\alpha_{n,\theta}^{-1}(\|\mathbf{y}\|_2)}{\|\mathbf{y}\|_2}\mathbf{y}$, we have that the complexity term is given by

$$\log|\nabla S_{\mathbf{t}}(\mathbf{y})| = (n-1)\log(\alpha_{n,\theta}^{-1}(\|\mathbf{y}\|_2)) - \log\left(\alpha'_{n,\theta}(\alpha_{n,\theta}^{-1}(\|\mathbf{y}\|_2))\right).$$

Inference on elliptical transport

Since the reference distribution $\eta_{\mathbf{t}}$ is spherical, then $\eta_{\mathbf{t}}(\mathbf{x}) = \beta_n(\mathbf{x}^\top\mathbf{x})$ for some positive function β_n. The transported distribution is also spherical with density $\pi_{\mathbf{t}}(\mathbf{y}) = h_n(\mathbf{y}^\top\mathbf{y}) := \beta_n(\psi_{n,\theta}^2(\|\mathbf{y}\|_2)\mathbf{y}^\top\mathbf{y})\psi_{n,\theta}(\|\mathbf{y}\|_2)^{(n-1)}(\alpha_{n,\theta}^{-1})'(\|\mathbf{y}\|_2)$.

Given observations $(\mathbf{t}, \mathbf{y})$, for inference on new inputs $\bar{\mathbf{t}}$ we have that the posterior distribution is also a spherical distribution, with density given by $\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y}) = \frac{h_{n+\bar{n}}(\bar{\mathbf{y}}^\top\bar{\mathbf{y}}+\|\mathbf{y}\|_2^2)}{h_n(\|\mathbf{y}\|_2^2)}$.

Since $\bar{\mathbf{x}} \sim \eta_{\bar{\mathbf{t}}}$ is spherical then $\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|_2} \overset{\mathrm{d}}{=} U^{(\bar{n})}$, so if $\beta \sim p(\|\bar{\mathbf{y}}\|_2|\|\mathbf{y}\|_2)$ is independent of $\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|_2}$ then we have

$$\bar{\mathbf{y}}|\mathbf{y} \overset{\mathrm{d}}{=} \frac{\beta}{\|\bar{\mathbf{x}}\|_2}\bar{\mathbf{x}},$$

where β is the positive r.v. of the norm of $\bar{\mathbf{y}}|\mathbf{y}$, that has density

$$p(\|\bar{\mathbf{y}}\|_2|\|\mathbf{y}\|_2) = \frac{2\pi^{\bar{n}/2}}{\Gamma(\bar{n}/2)}\|\bar{\mathbf{y}}\|_2^{\bar{n}-1}\frac{h_{n+\bar{n}}(\|\bar{\mathbf{y}}\|_2^2 + \|\mathbf{y}\|_2^2)}{h_{2,n}(\|\mathbf{y}\|_2^2)},$$

where $h_{2,n}$ is the marginal distribution of $\mathbf{y}$ from $(\mathbf{y}, \bar{\mathbf{y}})$. We can generate samples efficiently: sampling $\bar{\mathbf{x}}$ is straightforward from η, and β is an independent one dimensional positive random variable with explicit density. Note that $h_{2,n}(\|\mathbf{y}\|_2^2)$ is the normalisation constant, so we can avoid its computation via MCMC methods like slice sampling or emcee sampling [24, 87, 45].

Student-t case

The approach above includes the special case of the Student-t[3] process as follows: Consider $R_\theta \sim \sqrt{\Gamma^{-1}(\frac{\theta}{2}, \frac{\theta}{2})}$ with Γ^{-1} the inverse-gamma. Then $R_{n,\theta} := R_n R_\theta \sim \sqrt{nF_{n,\theta}}$, where $F_{n,\theta}$ denote the Fisher–Snedecor distribution, and we have that $\pi_{\mathbf{t}} = \mathcal{T}_n(\theta, 0, I_n)$ is a uncorrelated Student-t distribution with $\theta > 2$ degrees of freedom. Given observations $\mathbf{y}$, the distribution has closed-form posteriors: $R_\theta|\mathbf{y} \sim \sqrt{\Gamma^{-1}(\frac{\theta+n}{2}, \frac{\theta+\|\mathbf{y}\|_2^2}{2})}$ and $R_{\bar{n},\theta}|\mathbf{y} \sim \sqrt{\frac{\bar{n}(\theta+\|\mathbf{y}\|_2^2)}{\theta+n} F_{\bar{n},\theta+n}}$. Also, for a bivariate Student-t distribution with correlation ρ and degrees of freedom θ, its copula has coefficients of tail dependence given by $\lambda_u = \lambda_l = 2t_{\theta+1}\left(-\frac{\sqrt{\theta+1}\sqrt{1-\rho}}{\sqrt{1+\rho}}\right) > 0$, strictly heavier that the Gaussian case.

As an illustrative example, in Fig. 4.1 we can see the mean (solid line), the 95% confidence interval (dashed line) and 1000 samples (blurred lines) from 4 TGPs. All of them use a Brownian kernel $k(t, s) = min(t, s)$ for covariance transport, beside the second and fourth have an affine margin transport and the third and fourth have a Student-t elliptical transport. On the left column we plot the priors and on the right column we plot the posterior. The given observations are denoted with black dots. In this example we can see the difference between the Gaussian and Student-t copulas, although the priors look similar, the posteriors are quite different, where the Student-t copulas have more mass at the extrema.

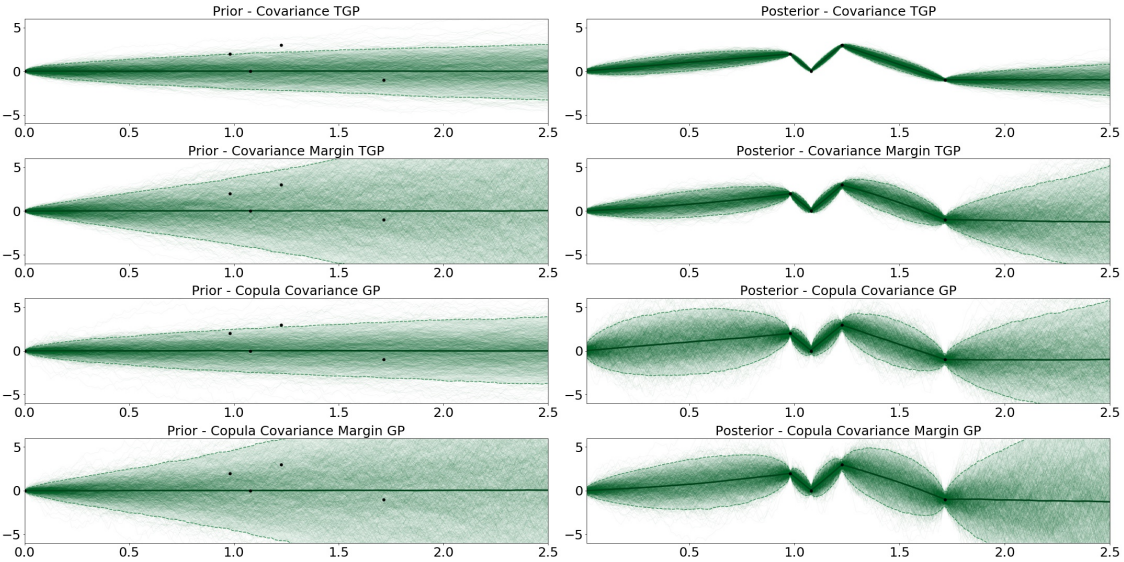

Figure 4.1: Samples from 4 TGP: the first and second examples have Gaussian copula, while third and fourth examples have Student-t copula.

4.5.2 Archimedean processes

From a Gaussian reference, the previous transport allows the generation of any elliptical copula. However, our approach is more general, and it is possible to obtain non-elliptical

[3]The Student-t distribution, and Gaussian as its limit, is the unique elliptical distribution with positive density over all $\mathbb{R}^n$ that is closed under conditioning [130].

copulas, specifically the so-called Archimedean copulas.

Definition 4.5.9 A copula $C(\mathbf{u})$ is called Archimedean if it can be written in the form $C(\mathbf{u}) = \psi\left(\sum_{\mathrm{i}=1}^{n} \psi^{-1}(u_{\mathrm{i}})\right)$ where $\psi : \mathbb{R}^+ \to [0,1]$ is continuous, with $\psi(0) = 1$, $\psi(\infty) = 0$ and its generalized inverse $\psi^{-1}(x) = \inf\{u : \psi(u) \leq x\}$.

Archimedean copulas have explicit form for tail dependency: $\lambda_l = 2 \lim_{x\to 0^+} \frac{\psi'(x)-\psi'(2x)}{\psi'(x)}$ and $\lambda_u = 2 \lim_{x\to\infty} \frac{\psi'(2x)}{\psi'(x)}$.

For example, if we consider the generator $\psi(u) = \exp(-u)$ then their Archimedean copula coincides with the independence copula $C(\mathbf{u}) = \prod_{\mathrm{i}=1}^{n} u_{\mathrm{i}}$ and $\lambda_l = \lambda_u = 0$. Some Archimedean copulas, like the independent one, can be extended as stochastic processes, which are characterised by the following proposition.

Proposition 4.5.10 Let $\psi : \mathbb{R}^+ \to [0,1]$ completely monotone, i.e. $\psi \in \mathcal{C}^\infty(\mathbb{R}^+, [0,1])$ and $(-1)^k \psi^{(k)}(x) \geq 0$ for $k \geq 1$. Then there exists a stochastic process where there finite-dimensional laws are $C_n(\mathbf{u}) = \psi\left(\sum_{\mathrm{i}=1}^{n} \psi^{-1}(u_{\mathrm{i}})\right)$.

Proof. By Kimberling's Theorem[80] ψ generates an Archimedean copula in any dimension iff ψ is completely monotone. Note that Archimedean copulas are exchangeable, i.e. for any n-permutation τ we have that $\mathbf{u} \overset{\mathrm{d}}{=} \tau(\mathbf{u})$, so in particular they are consistent under permutation, so we have that $F_{\eta_{\tau(\mathbf{t})}}(\tau(\mathbf{u})) = C_n(\tau(\mathbf{u})) = C_n(\mathbf{u}) = F_{\eta_{\mathbf{t}}}(\mathbf{u})$. The consistency under marginalisation is straightforward since $C_{n+1}(\mathbf{u}, 1) = \psi\left(\sum_{\mathrm{i}=1}^{n} \psi^{-1}(u_{\mathrm{i}}) + \psi^{-1}(1)\right) = C_n(\mathbf{u})$, and we conclude. □

Any Archimedean copula process has a completely monotone generator ψ associated that, by Bernstein's Theorem[80], is the Laplace transform [4] of a positive distribution F, i.e. $\psi = \mathcal{L}[F]$ and $F = \mathcal{L}^{-1}[\psi]$. The following proposition shows the relation between Archimedean copulas and simplicial contoured distributions [50, 81]..

Proposition 4.5.11 Let $S_n \sim \Gamma(n, 1)$, W a real positive r.v. and $U^{[n]}$ a uniform r.v. on the unit simplex in $\mathbb{R}^n$ (i.e. $\left\|U^{[n]}\right\|_1 = 1$), where S_n, W and $U^{[n]}$ are independent. Then $\mathbf{x} = (S_n/W)U^{[n]}$ follows a simplicial contoured distribution with an Archimedean survival copula generated by $\psi = \mathcal{L}[F_W]$, and each x_{i} has marginal distribution $F_{x_{\mathrm{i}}}(x) = 1 - \psi(x)$.

Proof. We have that $S_n U^{[n]} \overset{\mathrm{d}}{=} (E_1, ..., E_n)$ where $E_{\mathrm{i}} \sim Exp(1)$ are independent. By Marshall and Olkin algorithm [80], if $W \sim \mathcal{L}^{-1}[\psi]$ then $\mathbf{v} \sim C(\mathbf{v}) = \psi\left(\sum_{\mathrm{i}=1}^{n} \psi^{-1}(v_{\mathrm{i}})\right)$ where $v_{\mathrm{i}} = \psi(x_{\mathrm{i}})$. Since the transport from $\mathbf{x}$ to $\mathbf{v}$ is diagonal, they share the same copula, so $\mathbf{x}$ also has copula $C(\mathbf{v})$. Finally, since $\psi(x_{\mathrm{i}}) = v_{\mathrm{i}} \overset{\mathrm{d}}{=} 1 - v_{\mathrm{i}} \sim \mathbf{U}[0,1]$ then $1 - \psi(x_{\mathrm{i}})$ is the marginal distribution of each x_{i} for $\mathrm{i} = 1, ..., n$. □

Simplicial distributions $\mathbf{x} \overset{\mathrm{d}}{=} RU^{[n]}$, also know as ℓ_1-norm symmetric distributions, satisfy

[4] The Laplace transform of a random variable $Z > 0$ is defined as $\mathcal{L}(Z)(s) = \mathbb{E}(\exp(-sZ)) = \int_0^\infty \mathrm{e}^{-sz} \mathrm{d}F_Z(z)$ for $s \in [0, \infty]$.

$\|\mathbf{x}\|_1 = \sum_{\mathrm{i}=1}^n x_\mathrm{i} \overset{\mathrm{d}}{=} R$ and $\frac{\mathbf{x}}{\|\mathbf{x}\|_1} \overset{\mathrm{d}}{=} U^{[n]}$. If R has density p_R then $\mathbf{x}$ has density $p_\mathbf{x}(\mathbf{x}) = \Gamma(n)\|\mathbf{x}\|_1^{1-n} p_R(\|\mathbf{x}\|_1)$. For example, if the independence copula has generator $\psi(x) = \exp(-x)$ then W is degenerate on 1, so $R \overset{\mathrm{d}}{=} S_n/W \sim \Gamma(n,1)$ and marginals distribute as $x_\mathrm{i} \sim Exp(1)$. In another example, if $W \sim \Gamma(\frac{1}{\theta}, 1)$ then $\psi_\theta(s) = (1+s)^{-1/\theta}$ and $C(\mathbf{u}) = (\sum_{\mathrm{i}=1}^n u_\mathrm{i}^{-\theta} - n + 1)^{-1/\theta}$, the so-called Clayton copula. We have that $R \overset{\mathrm{d}}{=} S_n/W \sim \theta n F(2n, 2/\theta)$ and marginals distribute as $F(x_\mathrm{i}) = 1 - (1+x_\mathrm{i})^{-1/\theta}$, a shifted Pareto distribution.

Archimedean transport

Note the similitude between spherical and simplicial distributions, changing the role of the ℓ_2-norm by the ℓ_1-norm. If $\mathbf{y} \overset{\mathrm{d}}{=} SU^{[n]}$ for another real non-negative r.v. $S \in \mathbb{R}^+$, then the radial map $T^\alpha(\mathbf{x}) = \frac{F_S^{-1}(F_R(\|\mathbf{x}\|_1))}{\|\mathbf{x}\|_1}\mathbf{x} \overset{\mathrm{d}}{=} \frac{S}{R}\mathbf{x} \overset{\mathrm{d}}{=} SU^{[n]} \overset{\mathrm{d}}{=} \mathbf{y}$ is a transport map from $\mathbf{x}$ to $\mathbf{y}$. The next proposition shows how to transport a normal distribution into a simplicial distribution.

Proposition 4.5.12 Let $\mathbf{x} \sim \mathcal{N}_n(0, I_n)$. Denote Φ the distribution function of standard normal and consider the marginal transport T^h defined by $h(t,x) = -\log\Phi(x)$, i.e. $T^h(\mathbf{x})_\mathrm{i} = -\log(\Phi(x_\mathrm{i}))$. Given $S_n \overset{\mathrm{d}}{=} R_n/W$ for a positive r.v. W independent of $R_n \sim \Gamma(n,1)$, then the Archimedean transport $T_n^\alpha(\mathbf{y}) = \phi(\|\mathbf{y}\|_1)\mathbf{y} = \frac{F_{S_n}^{-1}(F_{R_n}(\|\mathbf{y}\|_1))}{\|\mathbf{y}\|_1}\mathbf{y}$ satisfies that $T_n^\alpha \circ T^h(\mathbf{x})$ has an Archimedean copula with generator $\psi = \mathcal{L}^{-1}(W)$.

Proof. If $x_\mathrm{i} \sim \mathcal{N}(0,1)$ then $y_\mathrm{i} = -\log(\Phi(x_\mathrm{i})) \sim Exp(1)$, so the sum satisfies that $\|\mathbf{y}\|_1 = \sum_{\mathrm{i}=1}^n y_\mathrm{i} \sim \Gamma(n,1)$ so $\|\mathbf{y}\|_1 \overset{\mathrm{d}}{=} R_n$. It is know that $\left(\frac{y_1}{\|\mathbf{y}\|_1}, ..., \frac{y_n}{\|\mathbf{y}\|_1}\right) \overset{\mathrm{d}}{=} U^{[n]}$ is independent from $\|\mathbf{y}\|_1$, so $T^h(\mathbf{x}) = \mathbf{y} = \|\mathbf{y}\|_1 \frac{\mathbf{y}}{\|\mathbf{y}\|_1} \overset{\mathrm{d}}{=} R_n U^{[n]}$. As T_n^α is a radial transport, then $T_n^\alpha \circ T^h$ transports $\mathbf{x}$ into a simplicial distribution, and by the prop. 4.5.11, we conclude. □

The last proposition implies that the transport $T = \{T_\mathbf{t} | \mathbf{t} \in \mathcal{T}^n, n \in \mathbb{N}\}$, where $T_\mathbf{t}(\mathbf{x}) = T_n^\alpha \circ T^h(\mathbf{x})$, is an f-transport with $f \sim \mathcal{GP}(0, \delta(t,\bar{t}))$, where the transport process $g := T(f)$ has a finite-dimensional Archimedean copula.

Learning an Archimedean transport

As the marginal transport was studied previously, we only need the *model complexity penalty* for this radial map.

Proposition 4.5.13 Given the map $T(\mathbf{y}) = \phi(\|\mathbf{y}\|_1)\mathbf{y} = \frac{F_S^{-1}(F_R(\|\mathbf{y}\|_1))}{\|\mathbf{y}\|_1}\mathbf{y}$, then $|\nabla T_\mathbf{t}(\mathbf{x})| = \phi(\|\mathbf{x}\|_1)^{n-1}\alpha'(\|\mathbf{x}\|_1)$.

Proof. Note that

$$\frac{\partial T_\mathbf{t}(\mathbf{x})_\mathrm{i}}{\partial x_\mathrm{i}} = \phi(\|\mathbf{x}\|_1) + \phi'(\|\mathbf{x}\|_1)x_\mathrm{i},$$

$$\frac{\partial T_{\mathbf{t}}(\mathbf{x})_{\mathrm{i}}}{\partial x_j} = \phi'(\|\mathbf{x}\|_1)x_{\mathrm{i}}, \text{if i} \neq j,$$
$$\nabla T_{\mathbf{t}}(\mathbf{x}) = \phi(\|\mathbf{x}\|_1)I + \phi'(\|\mathbf{x}\|_1)\mathbf{x}\mathbf{1}^\top = \phi'(\|\mathbf{x}\|_1)\left[\frac{\phi(\|\mathbf{x}\|_1)}{\phi'(\|\mathbf{x}\|_1)}I + \mathbf{x}\mathbf{1}^\top\right].$$

By Sylvester's determinant theorem we have

$$\begin{aligned}|\nabla T_{\mathbf{t}}(\mathbf{x})| &= \phi'(\|\mathbf{x}\|_1)^n \left(\frac{\phi(\|\mathbf{x}\|_1)}{\phi'(\|\mathbf{x}\|_1)}\right)^n \left(1 + \mathbf{1}^\top \left(\frac{\phi'(\|\mathbf{x}\|_1)}{\phi(\|\mathbf{x}\|_1)}I\right)\mathbf{x}\right), \\ &= \phi(\|\mathbf{x}\|_1)^{n-1}\left(\phi(\|\mathbf{x}\|_1) + \phi'(\|\mathbf{x}\|_1)\|\mathbf{x}\|_1\right), \\ &= \phi(\|\mathbf{x}\|_1)^{n-1}\alpha'(\|\mathbf{x}\|_1).\end{aligned}$$

thus concluding the proposed. □

With the above result, we have that the *model complexity penalty* is given by

$$\begin{aligned}\log|\nabla S_{\mathbf{t}}(\mathbf{y})| &= -\log|\nabla T_{\mathbf{t}}(S_{\mathbf{t}}(\mathbf{y}))|, \\ &= -(n-1)\log\left(\frac{\|\mathbf{y}\|_2}{\alpha^{-1}(\|\mathbf{y}\|_2)}\right) - \log\left(\alpha'(\alpha^{-1}(\|\mathbf{y}\|_2))\right), \\ &= -(n-1)\log\left(\frac{\|\mathbf{y}\|_2}{\alpha^{-1}(\|\mathbf{y}\|_2)}\right) + \log\left(\alpha^{-1}(\|\mathbf{y}\|_2)'\right).\end{aligned}$$

Inference with Archimedean transport

For an Archimedean copula, the conditional distribution given k observations $o_1, ..., o_k$ is given by $C(\mathbf{u}|o_1, ..., o_k) = \frac{\psi^{(k)}\left(\sum_{\mathrm{i}=1}^n \psi^{-1}(u_{\mathrm{i}})+a\right)}{\psi^{(k)}(a)}$ where $a = \sum_{j=1}^k \psi^{-1}(o_j)$ and $\psi^{(k)}$ is the k-th derivative of the generator ψ. We can then use methods for sampling the conditional Archimedean $\mathbf{u}$,to then apply the diagonal push-forward via $F^{-1}(u_{\mathrm{i}})$ where $F(x) = 1 - \psi(x)$.

4.6 Deep Transport Process

Both the generality and the feasible calculation of the presented transport-based approach to non-parametric regression motivate us to define complex models inspired on recent advances from the deep learning community. Via the composition of elementary transports (or *layers*) we can generate more expressive (or *deep*) transports. In this section, we will explain how to build such an architecture, describe the properties that are inherited through the composition, to finally propose families of transports that can be composed together and study their properties in the regression problem.

4.6.1 Learning deep transport process

Assume $T\#\eta = \pi$, where T is the composition of k transports, i.e. $T = T^{(k)} \circ ... \circ T^{(1)}$. Denote $\eta^{(0)} = \eta$ and assume that each $\eta^{(j)} = T^{(j)}\#\eta^{(j-1)}$ is a transport process with finite-

dimensional transports $\{T_{\mathbf{t}}^{(j)}\}_{j=1}^{k}$. Note that $\eta^{(k)} = T\#\eta = \pi$, where $T_{\mathbf{t}} = T_{\mathbf{t}}^{(k)} \circ ... \circ T_{\mathbf{t}}^{(1)}$ are finite-dimensional transports with $S_{\mathbf{t}} = S_{\mathbf{t}}^{(1)} \circ ... \circ S_{\mathbf{t}}^{(k)}$. As a consequence, the composition of transport processes is a transport process. Consequently, the NLL can be calculated as

$$-\log \pi_{\mathbf{t}}(\mathbf{y}|\theta) = -\log \eta_{\mathbf{t}}(S_{\mathbf{t}}(\mathbf{y})) - \sum\nolimits_{j=1}^{k} \log |\nabla S_{\mathbf{t}}^{(j)}(S_{\mathbf{t}}^{[(j+1):k]}(\mathbf{y}))|, \tag{4.6}$$

where $S_{\mathbf{t}}^{[j:k]}(\mathbf{y}) = S_{\mathbf{t}}^{(j)} \circ ... \circ S_{\mathbf{t}}^{(k)}(\mathbf{y})$, with the convention $S_{\mathbf{t}}^{[(k+1):k]}(\mathbf{y}) = \mathbf{y}$. The formula above is based on calculating each $F_{\mathbf{t}}^{(j)}(\mathbf{z}) = \log |\nabla S_{\mathbf{t}}^{(j)}(\mathbf{z})|$, which can be computed alternatively as $F_{\mathbf{t}}^{(j)}(\mathbf{z}) = -\log |\nabla T_{\mathbf{t}}^{(j)}(S_{\mathbf{t}}^{(j)}(\mathbf{z}))|$, or, for the triangular case, as $F_{\mathbf{t}}^{(j)}(\mathbf{z}) = \sum_{\mathrm{i}} \log \frac{\partial (S_{\mathbf{t}})_{\mathrm{i}}}{\partial y_{\mathrm{i}}}(\mathbf{z})$. The following algorithm computes the NLL, subject to being able to evaluate each function $F_{\mathbf{t}}^{(j)}$ and $S_{\mathbf{t}}^{(j)}$.

Algorithm 1 Calculate the NLL of a deep transport process

Require: Data $(\mathbf{t}, \mathbf{y})$, inverse transports $T_{\mathbf{t}}^{-1}(\mathbf{z}) = S_{\mathbf{t}}^{(1)} \circ ... \circ S_{\mathbf{t}}^{(k)}(\mathbf{z})$ and $F_{\mathbf{t}}^{(j)}(\mathbf{z}) = \log |\nabla S_{\mathbf{t}}^{(j)}(\mathbf{z})|$.
Ensure: $\mathcal{L} = -\log \pi_{\mathbf{t}}(\mathbf{y}|\theta)$
  $\mathbf{z} \leftarrow \mathbf{y}$, $\mathcal{L} \leftarrow 0$
  for $j \in k, ..., 1$ **do**
    $\mathcal{L} \leftarrow \mathcal{L} - F_{\mathbf{t}}^{(j)}(\mathbf{z})$
    $\mathbf{z} \leftarrow S_{\mathbf{t}}^{(j)}(\mathbf{z})$
  end for
  $\mathcal{L} \leftarrow \mathcal{L} - \log \eta_{\mathbf{t}}(\mathbf{z})$
  return $\mathcal{L}$

Remark 4.6.1. Algorithm 1 is based in applying the chain rule and the inverse function theorem over the composited inverse $S_{\mathbf{t}} = S_{\mathbf{t}}^{(1)} \circ ... \circ S_{\mathbf{t}}^{(k)}$, so

$$\begin{aligned} \nabla S_{\mathbf{t}}(\mathbf{y}) &= \nabla S_{\mathbf{t}}^{(1)}(S_{\mathbf{t}}^{(2)} \circ ... \circ S_{\mathbf{t}}^{(k)}) \nabla S_{\mathbf{t}}^{(2)}(S_{\mathbf{t}}^{(3)} \circ ... \circ S_{\mathbf{t}}^{(k)}) .... \nabla S_{\mathbf{t}}^{(k-1)}(S_{\mathbf{t}}^{(k)}(\mathbf{y})) \nabla S_{\mathbf{t}}^{(k)}(\mathbf{y}), & (4.7) \\ &= \nabla T_{\mathbf{t}}^{(1)}(S_{\mathbf{t}}^{(1)} \circ ... \circ S_{\mathbf{t}}^{(k)})^{-1} \nabla T_{\mathbf{t}}^{(2)}(S_{\mathbf{t}}^{(2)} \circ ... \circ S_{\mathbf{t}}^{(k)})^{-1} .... \nabla T_{\mathbf{t}}^{(k)}(S_{\mathbf{t}}^{(k)}(\mathbf{y}))^{-1}. & (4.8) \end{aligned}$$

Algorithm 1 is computationally efficient in terms of minimal use of memory (even the variable $\mathbf{z}$ can use the same memory as $\mathbf{y}$), and it can be executed in the shortest possible time by calling each function $F_{\mathbf{t}}^{(j)}$ and $S_{\mathbf{t}}^{(j)}$ only once. By implementing the calculations of NLL in any modern tensor framework, such as PyTorch, it is possible to apply automatic differentiation [97] to calculating the derivative of NLL with respect to parameters. Additionally, this algorithm is parallelizable in θ, thus allowing an efficient evaluation of NLL for multiple values for θ simultaneously in architectures such as GPUs. This is a desired property for derivative-free optimization methods such as particle swarm optimization [63], or MCMC ensemble samplers [56]. In stochastic gradient descent methods [22], given that in each step we use a subsampling from the data, we can take advantage of the GPU-based architectures running in parallel multiple executions, in order to better navigate the space of models.

4.6.2 Inference deep transport process

As the composition operation preserves triangularity, we assume $T^{(j)}$ are triangular for $j > l$, in addition to being able to calculate the posterior of $\eta^{(l)}$, i.e. compute $\eta^{(l)}_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|\mathbf{x})$ for any input $\bar{\mathbf{t}}$. Without loss of generality, it can be assumed that $l = 1$, since it is possible to collapse by composition the l transports in only one. The following algorithm generates samples from the posterior distribution $\pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y})$ under the above assumptions.

Algorithm 2 Generate samples from the posterior

Require: Observations $\mathbf{y} \sim \pi_{\mathbf{t}}$, new inputs $\bar{\mathbf{t}} \in \mathcal{I}^{\mathrm{d}}, \mathrm{d} \in \mathbb{N}$, number of samples $N \in \mathbb{N}$.
Ensure: $\bar{\mathbf{y}}_{\mathrm{i}} \sim \pi_{\bar{\mathbf{t}}|\mathbf{t}}(\bar{\mathbf{y}}|\mathbf{y})$ for $\mathrm{i} = 1, ..., N$
  $\mathbf{x} \leftarrow S^{[l+1:k]}_{\mathbf{t}}(\mathbf{y})$
  $R(\cdot) \leftarrow P_{\bar{\mathbf{t}}} \circ T^{[l+1:k]}_{\mathbf{t},\bar{\mathbf{t}}}(\mathbf{x}, \cdot)$
  for $\mathrm{i} \in 1, ..., N$ **do**
    $\bar{\mathbf{x}}_{\mathrm{i}} \sim \eta^{(l)}_{\bar{\mathbf{t}}|\mathbf{t}}(\cdot|\mathbf{x})$
    $\bar{\mathbf{y}}_{\mathrm{i}} \leftarrow R(\bar{\mathbf{x}}_{\mathrm{i}})$
  end for
  return $\{\bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_N\}$

Algorithm 2 is parallelisable in N, since the function $R(\cdot)$ is the same for all samples, and thus allows us to obtain multiple samples simultaneously in an efficient manner. This can be used in turn to calculate moments, quantiles or other statistics in an empirical way through Monte Carlo.

4.6.3 Noise layer

Under the presence of noisy observations, following the same rationale as GPs, warped GPs [125] and Student-t processes [119], we consider that the covariance transport has a special behavior. Let $k(t, s) = r(t, s) + \sigma_0 \delta_{t,s}$, where δ is Kronecker delta, σ_0 is the parameter that controls the intensity of noise and $r(t, s)$ is the noise-free covariance function. We consider that the observations have uncorrelated noise. While for training we use $k(t, s)$ in the formula for NLL, in inference we use $k(t, s)$ on the backward-step (i.e. for the inverse map $\mathbf{x} = T^{-1}_{\mathbf{t}}(\mathbf{y})$), and on the forward-step (i.e. for push-forward the reference distribution) we use $r(t, s)$, instead of $k(t, s)$, to perform a free-noise prediction.

4.6.4 Sparse layer

While marginal and copula transports can be evaluated efficiently without needing training data, the covariance transports needs all the data $\mathbf{y}$ to performance inference. The computational complexity of evaluation is $\mathcal{O}(n^2)$ in memory and $\mathcal{O}(n^3)$ in time, where $n = |\mathbf{y}|$. Sparse approximations are widely used to solve this issue on GPs [101, 124, 135], and it is natural to define a *sparse* transport as $T_{\bar{\mathbf{t}}}(\mathbf{u}) = \Sigma_{\bar{\mathbf{t}}\mathbf{s}}\Sigma^{-1}_{\mathbf{s}\mathbf{s}}\mathbf{z} + \mathrm{chol}(\Sigma_{\bar{\mathbf{t}}\bar{\mathbf{t}}} - \Sigma_{\bar{\mathbf{t}}\mathbf{s}}\Sigma^{-1}_{\mathbf{s}\mathbf{s}}\Sigma_{\bar{\mathbf{t}}\mathbf{s}})\mathbf{u}$, where $(\mathbf{s}, \mathbf{z})$ are

	Sunspots		Heart		Economic	
	WGP	TGP	WGP	TGP	WGP	TGP
MAE	25.266 ± 4.607	24.710 ± 4.271	2.965 ± 0.827	2.907 ± 0.715	1.132 ± 0.260	1.111 ± 0.215
EAE	30.166 ± 4.374	29.649 ± 4.168	3.431 ± 0.732	3.388 ± 0.660	1.392 ± 0.235	1.380 ± 0.206
MSE	1,306.253 ± 560.496	1,223.257 ± 421.385	16.405 ± 8.809	15.740 ± 7.619	3.002 ± 1.643	2.860 ± 1.311
ESE	1,889.318 ± 633.325	1,796.989 ± 514.193	21.963 ± 8.524	21.554 ± 8.213	4.376 ± 1.725	4.272 ± 1.424

Table 4.1: WGP and TGP results over Sunspots, Heart and Economic datasets.

trainable pseudo-data with $|\mathbf{s}| = m < n$. The training of pseudo-data follows the same ideas that sparse GPs, like SoD and SoR approximations [101], where the computational cost drops to $\mathcal{O}(nm)$ in space and $\mathcal{O}(nm^2)$ in time.

4.7 Experimental validation

We validate our approach with three real-world time series, described as follows:

1. **Sunspots Data**: The Sunspot time series [122] corresponds to the yearly number of sunspots between 1700 and 2008, resulting in 309 data points, one per year. These measures are positive and semi-periodic, with a cycle period of around 11-years.

2. **Heart Data**: This is a heart-rate time series from the MIT-BIH Database (ecg.mit.edu) [54]. This series contains 1800 evenly-spaced positive measurements of instantaneous heart rate (in units of beats per minute) from a single subject, happening at 0.5 second intervals, and showing a semi-periodic pattern. For performance issues, we take a subsample of 450 measures at 2.0 seconds intervals.

3. **Economic Data**: This time series corresponds to the quarterly average *3-Month Treasury Bill: Secondary Market Rate* [42] between the first quarter of 1959 and the third quarter of 2009, that is, 203 observations, one per quarter. We know beforehand that this macroeconomic signal is the price of U.S. government risk-free bonds, which cannot take negative values and can have large positive deviations.

Due to the semi-periodic nature of the time series, we consider a noisy spectral mixture with two components kernel k_{SM} [143] for the covariance transport. Since the time series are positive, we use a shifted Box-Cox warping ϕ_{BC} [107] for marginal transport. We compare two models: a warped GP, with k_{SM} kernel and ϕ_{BC} warping; and a TGP with a Student-t copula transport, besides the above-described covariance and marginal transports.

We leave the standard GPs out of the experiment since the assumption of Gaussianity violates the nature of the datasets, having a lower predictive power than the WGP, as shown in [107, 108]. To illustrate this fact, in Fig. 4.2 we show the posterior of three trained models: GP in blue, WGP in green and TGP in purple. We plot the observations (black dots), the mean (solid line), the 95% confidence interval (dashed line) and 25 samples (blurred lines). Notice how the GP fails to model the positivity and the correct amplitude of the phenomena.

The experiment was implemented in a Python-based library named *tpy: Transport processes in Python*[106], with a PyTorch backend for GPU-support and automatic differentiation

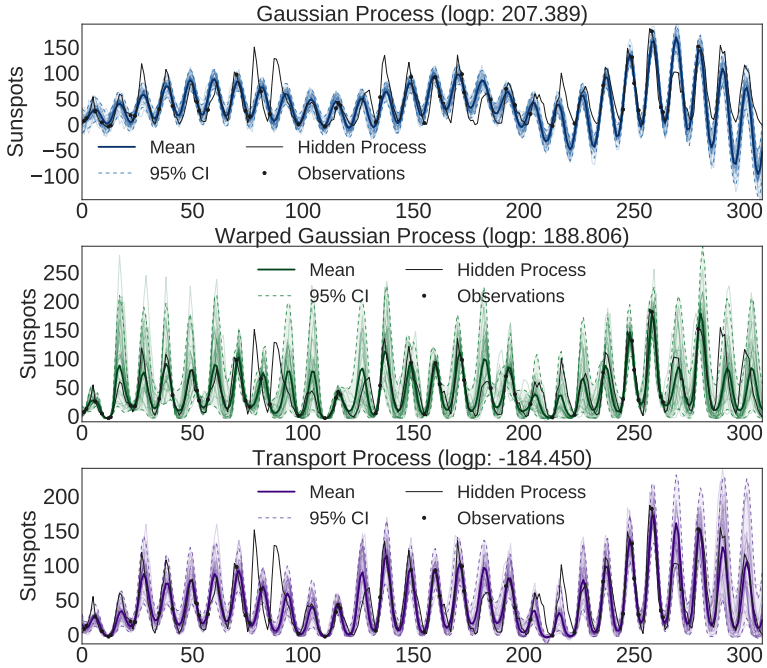

Figure 4.2: GP (blue), WGP (green) and TGP (purple) over Sunspots data.

[97]. The training was performed by minimising the NLL from eq. (4.6), via a stochastic mini-batches rprop method [104], to then end with non-stochastic iterations.

In each experiment, we randomly (uniformly) select 15% of the data for training and the remaining 85% for validation. Given the validation data points $\{y_{\mathrm{i}}\}_{\mathrm{i}=1}^{n}$, for each model we generate S samples $\{y_{\mathrm{i}}^{(k)}\}_{\mathrm{i}=1}^{n}$ for $k = 1, ..., S$, and then we calculate four performance indices: the mean square error as $\mathrm{MSE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\left(y_{\mathrm{i}} - \frac{1}{S}\sum_{k=1}^{S} y_{\mathrm{i}}^{(k)}\right)^2$, the mean absolute error as $\mathrm{MAE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}|y_{\mathrm{i}} - \frac{1}{S}\sum_{k=1}^{S} y_{\mathrm{i}}^{(k)}|$, the expected square error as $\mathrm{ESE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\frac{1}{S}\sum_{k=1}^{S}(y_{\mathrm{i}} - y_{\mathrm{i}}^{(k)})^2$, and the expected absolute error as $\mathrm{EAE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\frac{1}{S}\sum_{k=1}^{S}|y_{\mathrm{i}} - y_{\mathrm{i}}^{(k)}|$. We repeat each experiment 100 times. The results for all of these experiments are summarized in Table 1, showing each mean and standard deviation. Consistently, the proposed TGP has better performance that the warped GP alternative, for each dataset and evaluation index.

Chapter 5

Bayesian Learning with Wasserstein Barycenters

"...optimal transport is a simple, meaningful, natural and therefore universal concept."

– Cédric Villani, in *Optimal transport, old and new*

The main results presented in this Chapter are included in the preprint paper [12]: *Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. arXiv preprint arXiv:1805.10833, 2018.*

Going back to the general framework for Bayesian estimation based on loss functions over probability measures consider in Chapter 1, our motivation in this chapter is to find an alternative, non-parametric learning strategy which can cope with some of the drawbacks of standard approaches such as *Bayesian model average* (BMA). The main conceptual contribution of this chapter is the introduction of the *Bayesian Wasserstein barycenter estimator* (BWB) as a novel model-selection criterion based on optimal transport theory. In a nutshell, given a prior on models Π and observations $D = \{x_1, \dots, x_n\} \subset \mathcal{X}$, a BWB estimator is any minimizer $\hat{m}_p^n \in \mathcal{M}$ of the loss function

$$\mathcal{M} \ni \bar{m} \mapsto \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Pi_n(\mathrm{d}m), \tag{5.1}$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$, Π_n is the posterior distribution on models given the data D, and W_p is the celebrated p-Wasserstein distance ([138, 139]). The minimization of functionals akin to (5.1) is an active field of current research in machine learning [31, 69]. For instance, if the model space $\mathcal{M}$ equals the set of all probability measures on $\mathcal{X}$, then our estimator $\hat{m}_p^n$ coincides with the *(population) Wasserstein barycenter* of Π_n. The study of Wasserstein barycenters was introduced by [2], but see also [74, 18] for more recent developments and references to the literature, or our own. In Section 5.2, we recall the notions of Wasserstein distances and, relying on the previously developed framework, we rigorously introduce the Bayesian Wasserstein barycenter estimator in Section 5.3. We explore its existence, uniqueness, absolute continuity and illustrate the advantage of this estimator by comparing it to the Bayesian model average: it turns out that our estimator is less dispersed, and in particular, it has less variance than the model average. In Section 5.4

we state conditions for the statistical *consistency* of our estimator, which is a basic desirable property: briefly put, this means that as more data becomes available, the estimator converges to the *true* model. The main result in this regard is Theorem 5.4.10.

We remark that the use of Wasserstein barycenters in Bayesian statistics was initiated, to the best of our knowledge, by the works [127, 75, 82, 128]. There the authors consider the problem of how to stitch together posteriors computed on different data sets; their answer is to do it by calculating the barycenter *between* the posteriors. In contrast to this, we take the availability of a posterior for granted and instead compute the barycenter *of* the posterior.

Let us fix some notation and conventions. As we mentioned in Chapter 1, we assume throughout that $\mathcal{M} \subseteq \mathcal{P}_{ac}(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$, and $\mathcal{P}_{ac}(\mathcal{X})$ is the subset of absolutely continuous measures with respect to a common reference σ-finite measure λ on $\mathcal{X}$. As a convention, we use the same notation for an element $m(\mathrm{d}x) \in \mathcal{M}$ and its density $m(x)$ w.r.t. λ. Given a measurable map $T : \mathcal{Y} \to \mathcal{Z}$ and a measure ν on $\mathcal{Y}$ we denote by $T(\nu)$ the image measure (push-forward), which is the measure on $\mathcal{Z}$ given by $T(\nu)(\cdot) = \nu(T^{-1}(\cdot))$. We denote by $\mathrm{supp}(\nu)$ the support of a measure ν and by $|\mathrm{supp}(\nu)|$ its cardinality. Moreover, we assume that the *true model* $m_0 \in \mathcal{P}_{ac}(\mathcal{X})$ —such that $x_1, ...x_n$ are i.i.d. according to m_0— does exist, although in general m_0 may not be an element of $\mathcal{M}$.

5.1 Bayesian Posterior Averages Estimators

The next result illustrates the fact that many Bayesian estimators, including the *model average estimator*, correspond to finding a so-called Fréchet mean or barycenter [93] under a suitable metric/divergence on probability measures.

Proposition 5.1.1 Let $\mathcal{M} = \mathcal{P}_{ac}(\mathcal{X})$ and consider the loss functions:

i) The L_2-distance: $L_2(m, \bar{m}) = \frac{1}{2} \int_{\mathcal{X}} (m(x) - \bar{m}(x))^2 \, \lambda(\mathrm{d}x)$,

ii) The reverse KL divergence: $D_{KL}(m||\bar{m}) = \int_{\mathcal{X}} m(x) \ln \frac{m(x)}{\bar{m}(x)} \lambda(\mathrm{d}x)$,

iii) The forward KL divergence $D_{KL}(\bar{m}||m) = \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{m(x)} \lambda(\mathrm{d}x)$,

iv) The squared Hellinger distance $H^2(m, \bar{m}) = \frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{m(x)} - \sqrt{\bar{m}(x)} \right)^2 \lambda(\mathrm{d}x)$.

Then, in cases i) and ii) the corresponding Bayes estimators of Equation (1.3) coincide with the *Bayesian model average*:

$$\bar{m}(x) := \mathbb{E}_{\Pi_n}[m] = \int_{\mathcal{M}} m(x) \Pi_n(\mathrm{d}m).$$

Furthermore, with Z_exp and Z_2 denoting normalizing constants, the Bayes estimators corresponding to the cases iii) and iv) are given by the *exponential model average* and the *square model average*, respectively:

$$\hat{m}_{exp}(x) = \frac{1}{Z_{exp}} \exp \int_{\mathcal{M}} \ln m(x) \Pi_n(\mathrm{d}m) \ , \ \hat{m}_2(x) = \frac{1}{Z_2} \left(\int_{\mathcal{M}} \sqrt{m(x)} \Pi_n(\mathrm{d}m) \right)^2.$$

Proof of Proposition 5.1.1. Consider the squared L_2-distance between densities $L_2(m, \bar{m}) = \frac{1}{2}\int_{\mathcal{X}} (m(x) - \bar{m}(x))^2 \lambda(\mathrm{d}x)$. By Fubini we have

$$R_L(\bar{m}|D) = \tfrac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{M}} (m(x) - \bar{m}(x))^2 \Pi(\mathrm{d}m|D)\lambda(\mathrm{d}x).$$

By the fundamental lemma of calculus of variations, denoting

$$\mathcal{L}(x, \bar{m}, \bar{m}') = \tfrac{1}{2}\int_{\mathcal{M}} (m(x) - \bar{m}(x))^2 \Pi(\mathrm{d}m|D)$$

the extrema of $R_L(\bar{m}|D)$ are weak solutions of the Euler-Lagrange equation

$$\begin{aligned} \tfrac{\partial \mathcal{L}(x,\bar{m},\bar{m}')}{\partial \bar{m}} &= \tfrac{\mathrm{d}}{\mathrm{d}x}\tfrac{\partial \mathcal{L}(x,\bar{m},\bar{m}')}{\partial \bar{m}'} \\ \int_{\mathcal{M}} (m(x) - \bar{m}(x))\,\Pi(\mathrm{d}m|D) &= 0, \end{aligned}$$

so we have that the optimal is reached on the Bayesian model average

$$\int_{\mathcal{M}} m(x)\Pi(\mathrm{d}m|D).$$

If we take the loss function as the reverse Kullback-Leibler divergence

$$D_{KL}(m||\bar{m}) = \int_{\mathcal{X}} m(x) \ln \tfrac{m(x)}{\bar{m}(x)}\lambda(\mathrm{d}x),$$

we have that the associate Bayes risk can be written as

$$\begin{aligned} &R_{D_{RKL}}(\bar{m}|D) \\ =&\int_{\mathcal{M}}\int_{\mathcal{X}} m(x) \ln \tfrac{m(x)}{\bar{m}(x)}\lambda(\mathrm{d}x)\Pi(\mathrm{d}m|D) \\ =&\int_{\mathcal{X}}\int_{\mathcal{M}} m(x) \ln m(x)\Pi(\mathrm{d}m|D)\lambda(\mathrm{d}x) - \int_{\mathcal{X}}\int_{\mathcal{M}} m(x)\Pi(\mathrm{d}m|D) \ln \bar{m}(x)\lambda(\mathrm{d}x) \\ =&C - \int_{\mathcal{X}} \mathbb{E}[m](x) \ln \bar{m}(x)\lambda(\mathrm{d}x) \end{aligned}$$

and changing the constant C by the entropy of $\mathbb{E}[m]$ we have that

$$\begin{aligned} &R_{D_{RKL}}(\bar{m}|D) \\ =&C' + \int_{\mathcal{X}} \mathbb{E}[m](x) \ln \mathbb{E}[m](x)\lambda(\mathrm{d}x) - \int_{\mathcal{X}} \mathbb{E}[m](x) \ln \bar{m}(x)\lambda(\mathrm{d}x) \\ =&C' + D_{RKL}(\mathbb{E}[m], \bar{m}), \end{aligned}$$

so the extremum of $R_{D_{RKL}}(\bar{m}|D)$ is given by the Bayesian model average. Instead if we take the forward Kullback-Leibler divergence as loss function

$$D_{KL}(\bar{m}||m) = \int_{\mathcal{X}} \bar{m}(x) \ln \tfrac{\bar{m}(x)}{m(x)}\lambda(\mathrm{d}x),$$

we have

$$\begin{aligned} &R_{D_{KL}}(\bar{m}|D) \\ =&\int_{\mathcal{M}}\int_{\mathcal{X}} \bar{m}(x) \ln \tfrac{\bar{m}(x)}{m(x)}\lambda(\mathrm{d}x)\Pi(\mathrm{d}m|x_1, \ldots, x_n) \\ =&\int_{\mathcal{X}} \bar{m}(x) \ln \bar{m}(x)\lambda(\mathrm{d}x) - \int_{\mathcal{X}} \bar{m}(x) \int_{\mathcal{M}} \ln m(x)\Pi(\mathrm{d}m|x_1, \ldots, x_n)\lambda(\mathrm{d}x) \\ =&\int_{\mathcal{X}} \bar{m}(x) \ln \bar{m}(x)\lambda(\mathrm{d}x) - \int_{\mathcal{X}} \bar{m}(x) \ln \exp \mathbb{E}[\ln m]\lambda(\mathrm{d}x) \\ =&\int_{\mathcal{X}} \bar{m}(x) \ln \tfrac{\bar{m}(x)}{\exp \mathbb{E}[\ln m]}\lambda(\mathrm{d}x). \end{aligned}$$

Denoting by Z the normalization constant, we have

$$\begin{aligned} R_{D_{KL}}(\bar{m}|D) + \ln Z &= \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{\exp \mathbb{E}[\ln m]} \lambda(\mathrm{d}x) + \int_{\mathcal{X}} \bar{m}(x) \ln Z \lambda(\mathrm{d}x) \\ &= \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{\frac{1}{Z} \exp \mathbb{E}[\ln m]} \lambda(\mathrm{d}x) \\ &= D_{KL}\left(\tfrac{1}{Z} \exp \mathbb{E}[\ln m], \bar{m}\right). \end{aligned}$$

So the extremum of $R_{D_{KL}}(\bar{m}|D)$ is the Bayesian *exponential* model average given by

$$\hat{m}(x) = \tfrac{1}{Z} \exp \int_{\mathcal{M}} \ln m(x) \Pi(\mathrm{d}m).$$

Finally, if we take the squared Hellinger distance as loss function

$$H^2(m, \bar{m}) = \tfrac{1}{2} \int_{\mathcal{X}} \left(\sqrt{m(x)} - \sqrt{\bar{m}(x)}\right)^2 \lambda(\mathrm{d}x) = 1 - \int_{\mathcal{X}} \sqrt{m(x)\bar{m}(x)} \lambda(\mathrm{d}x),$$

we easily check that the extremum of $R_{H^2}(\bar{m}|D)$ is the Bayesian *square* model average:

$$\hat{m}(x) = \tfrac{1}{Z} \left(\int_{\mathcal{M}} \sqrt{m(x)} \Pi(\mathrm{d}m|x_1, \ldots, x_n)\right)^2.$$

□

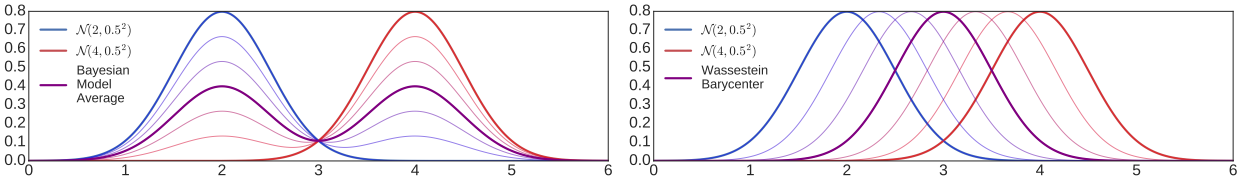

Figure 5.1: Model average (left) and Wasserstein barycenter (right) of two Gaussian densities.

The Bayesian estimators $\bar{m}, \hat{m}_{exp}, \hat{m}_2$ share a common feature: their values at each point $x \in \mathcal{X}$ are computed in terms of some posterior *average* of the values of certain functions evaluated at x. This is due to the fact that all the above distances are *vertical* [115], in the sense that computing the distance between m and $\bar{m}$ involves the integral of vertical displacements between the graphs of these two densities. An undesirable fact about *vertical averages* is that they do not preserve properties of the original model space. E.g. if the posterior distribution is equally concentrated on two different models $m_0 = \mathcal{N}(\mu_0, 1)$ and $m_1 = \mathcal{N}(\mu_1, 1)$ with $\mu_0 \neq \mu_1$, i.e. both models are unimodal (Gaussian) with unit variance, the model average is in turn a bimodal (non-Gaussian) distribution with variance strictly greater than 1. More generally, model averages might yield intractable representations or be hardly interpretable in terms of the prior and parameters.

We shall next introduce the analogous objects in the case of Wasserstein distances, which are *horizontal* distances [115], in the sense that they involve integrating horizontal displacements between the graphs of the densities. We will further develop the theory of the corresponding Bayes estimators, which will correspond to *Wasserstein barycenters* arising in optimal transport theory (see [2, 96, 65, 74]). Going back to the Gaussian example, say for two models given by the univariate Gaussian distributions $m_0 = \mathcal{N}(\mu_0, \sigma_0^2)$ and $m_1 = \mathcal{N}(\mu_1, \sigma_1^2)$, it turns out that the so-called 2-Wasserstein barycenter distribution is given by $\hat{m} = m_{\frac{1}{2}} = \mathcal{N}(\frac{\mu_0+\mu_1}{2}, (\frac{\sigma_0+\sigma_1}{2})^2)$. In Fig. 5.1 we illustrate a vertical (with L_2) and a horizontal (with W_2) Fréchet mean, and interpolations, between two Gaussian densities.

5.2 Wasserstein Space

We propose a novel Bayesian estimator obtained by using the Wasserstein distance as loss function. This estimator is thus a Fréchet mean in the Wasserstein metric and is usually referred to as Wasserstein barycenter [2].

From now until the end of this work, unless otherwise stated, we assume:

Assumption 5.2.1 $(\mathcal{X}, \mathrm{d})$ is a separable locally-compact geodesic space and $p \geq 1$.

In this context, geodesic means complete and that any pair of points admit a mid-point with respect to d. The reader can think of $\mathcal{X}$ as a Euclidean space with d the Euclidean distance. On the other hand, d^p controls the tails of the models to be considered. We now recall some elements of optimal transport.

5.2.1 Wasserstein distance

A thorough introduction of optimal transport and some of its applications can be found in the books by Villani [138, 139]. It is difficult to overstate the impact that the field has had in mathematics as a whole. In particular, regarding statistical applications, we refer to the recent survey [92] and the many references therein. In parallel, optimal transport has become increasingly popular within the machine learning community [69], though most of the published works have focused on the discrete setting (e.g., comparing histograms in [31], classification in [47] and images in [28, 10], among others). Let us briefly review the definitions and results needed to present our approach.

Given measures μ, υ over $\mathcal{X}$ we denote by $\Gamma(\mu, \upsilon)$ the set of couplings with marginals μ and υ, i.e. $\gamma \in \Gamma(\mu, \upsilon)$ if $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ and $\gamma(\mathrm{d}x, \mathcal{X}) = \mu(\mathrm{d}x)$ and $\gamma(\mathcal{X}, \mathrm{d}y) = \upsilon(\mathrm{d}y)$. Given a real number $p \geq 1$ we define the p-Wasserstein space $\mathcal{W}_p(\mathcal{X})$ by

$$\mathcal{W}_p(\mathcal{X}) := \left\{ \eta \in \mathcal{P}(\mathcal{X}) : \textstyle\int_{\mathcal{X}} \mathrm{d}(x_0, x)^p \eta(\mathrm{d}x) < \infty, \text{ some } x_0 \right\}.$$

The p-Wasserstein between measures μ and υ is given by

$$W_p(\mu, \upsilon) = \left(\inf_{\gamma \in \Gamma(\mu,\upsilon)} \textstyle\int_{\mathcal{X}\times\mathcal{X}} \mathrm{d}(x, y)^p \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{p}}. \tag{5.2}$$

An optimizer of the r.h.s. of (5.2) is called an optimal transport. The quantity W_p defines a distance turning $\mathcal{W}_p(\mathcal{X})$ into a complete metric space. In the Euclidean case, there often exist explicit formulae for optimal transports and the Wasserstein distance, e.g. for the generic one-dimensional case, and the multivariate Gaussian case ($p = 2$); see [30]. If in (5.2) we assume $p = 2$, $\mathcal{X}$ is Euclidean space, and μ is absolutely continuous, then Brenier's theorem [138, Theorem 2.12(ii)] establishes the uniqueness of a minimizer. Furthermore, this optimiser is supported on the graph of the gradient of a convex function.

5.2.2 Wasserstein barycenter

We start with the definition of Wasserstein population barycenter:

Definition 5.2.2 Given $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$, the p-Wasserstein risk of $\bar{m} \in \mathcal{P}(\mathcal{X})$ is

$$V_p(\bar{m}) := \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Gamma(\mathrm{d}m).$$

Any measure $\hat{m}_p \in \mathcal{M}$ which is a minimizer of the problem

$$\inf_{\bar{m} \in \mathcal{M}} V_p(\bar{m}),$$

is called a p-Wasserstein population barycenter of Γ over $\mathcal{M}$.

In the case $\mathcal{M} = \mathcal{W}_p(\mathcal{X})$, the above is nothing but the p-Wasserstein population barycenter of Γ introduced in [18]. The term *population* emphasizes that the support of Γ might be infinite.

Let us introduce some required notation. For $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ we write $\Gamma \in \mathcal{P}(\mathcal{W}_p(\mathcal{X}))$ if Γ is concentrated on a set of measures with finite moments of order p. We can now consider $\mathcal{W}_p(\mathcal{X})$ with the complete metric W_p as a base Polish space, and define $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ analogously, with an associated Wasserstein distance of order p which for simplicity we still call W_p. We have that $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ if $\Gamma \in \mathcal{P}(\mathcal{W}_p(\mathcal{X}))$, and for some (and then all) $\tilde{m} \in \mathcal{W}_p(\mathcal{X})$ it satisfies

$$\int_{\mathcal{P}(\mathcal{X})} W_p(m, \tilde{m})^p \Gamma(\mathrm{d}m) < \infty.$$

If Γ is concentrated on measures with finite moments of order p and with density with respect to λ, then we rather write $\Gamma \in \mathcal{P}(\mathcal{W}_{p,ac}(\mathcal{X}))$, with the notation $\Gamma \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$ if as before $\int_{\mathcal{P}(\mathcal{X})} W_p(m, \tilde{m})^p \Gamma(\mathrm{d}m) < \infty$ for some $\tilde{m}$.

Let $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$. By definition its *model average* belongs to $\mathcal{W}_p(X)$, since

$$\infty > \int W_p(m, \delta_x)^p \Gamma(\mathrm{d}m) = \int\int \mathrm{d}(x,y)^p m(\mathrm{d}y) \Gamma(\mathrm{d}m) = \int \mathrm{d}(x,y)^p \int m(\mathrm{d}y) \Gamma(\mathrm{d}m).$$

We state an existence result of p-Wasserstein barycenter, first obtained in [74, Theorem 2]; our argument here seems more elementary.

Lemma 5.2.3 If $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(X))$, there exists a p-Wasserstein barycenter, i.e. exist a minimizer for the positive functional

$$V(\Gamma) := \inf \left\{ \int_{\mathcal{W}_p(\mathcal{X})} W_p(\nu, m)^p \Gamma(\mathrm{d}m) : \nu \in \mathcal{W}_p(\mathcal{X}) \right\}.$$

Proof. Taking $\nu = \delta_x$ we get that $V(\Gamma)$ is finite. Now, let $\{\nu_n\} \subset \mathcal{W}_p(\mathcal{X})$ such that

$$\int_{\mathcal{W}_p(\mathcal{X})} W_p(\nu_n, m)^p \Gamma(\mathrm{d}m) \searrow V(\Gamma).$$

For n large enough we have

$$W_p\left(\nu_n\,,\, \int_{\mathcal{W}_p(\mathcal{X})} m \Gamma(\mathrm{d}m)\right)^p \le \int_{\mathcal{W}_p(\mathcal{X})} W_p(\nu_n\,,\, m)^p \Gamma(\mathrm{d}m) \le V(\Gamma) + 1 =: K,$$

by convexity of optimal transport costs. From this we derive that (for every x)

$$\sup_n \int_{\mathcal{X}} \mathrm{d}(x,y)^p \nu_n(\mathrm{d}y) < \infty.$$

By Markov inequality this shows, for each $\varepsilon > 0$, that there is ℓ large enough such that $\sup_n \nu_n(\{y \in \mathcal{X} : \mathrm{d}(x,y) > \ell\}) \leq \varepsilon$. As explained in [74], the assumptions made on $\mathcal{X}$ imply that $\{y \in \mathcal{X} : \mathrm{d}(x,y) \leq \ell\}$ is compact (Hopf-Rinow theorem), and so we deduce the tightness of $\{\nu_n\}$. By Prokhorov theorem, up to selection of a subsequence, there exists $\nu \in \mathcal{W}_p(\mathcal{X})$ which is its weak limit. We can conclude by Fatou's lemma:

$$V(\Gamma) = \lim \int W_p(\nu_n, m)^p \Gamma(\mathrm{d}m) \geq \int W_p(\nu, m)^p \Gamma(\mathrm{d}m).$$

□

It is plain from the above proof that if $\mathcal{M} \subset \mathcal{W}_p(\mathcal{X})$ is weakly closed, then there also exists a minimizer in $\mathcal{M}$ of

$$\inf \left\{ \int_{\mathcal{W}_p(\mathcal{X})} W_p^p(\nu, m) \Gamma(\mathrm{d}m) : \nu \in \mathcal{M} \right\}.$$

Let us now consider the relevant case of $p = 2$, $\mathcal{X} = \mathbb{R}^q$ and d = Euclidean distance. We take $\Gamma \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^q))$, observing that in such situation the previous lemma applies. We recall now the uniqueness result stated in [74, Proposition 6]:

Lemma 5.2.4 Assume that there exists a set $A \subset \mathcal{W}_2(\mathbb{R}^q)$ of measures with

$$\mu \in A,\ B \in \mathcal{B}(\mathbb{R}^q),\ \dim(B) \leq q-1 \implies \mu(B) = 0\ ,$$

and $\Pi(A) > 0$. Then Π admits a unique 2-Wasserstein population barycenter.

Note that Lebesgue measure λ satisfy above condition, so all measures absolutely continuous with respect to λ also fulfil it, in particular distributions with density.

5.3 Bayesian Wasserstein Barycenter Estimator

We come to the most important definition (and conceptual contribution) of the chapter. A Bayesian Wasserstein barycenter estimator is nothing but a p-Wasserstein population barycenter of the posteriors Π_n over the model space $\mathcal{M}$:

Definition 5.3.1 Given $\Pi \in \mathcal{P}(\mathcal{M}) \subset \mathcal{P}(\mathcal{P}(\mathcal{X}))$ and data $D = \{x_1, \ldots, x_n\}$ determining Π_n as in (1.1), the p-Wasserstein Bayes risk of $\bar{m} \in \mathcal{W}_{p,ac}(\mathcal{X})$, and a Bayes Wasserstein barycenter estimator $\hat{m}_p^n$ over the model space $\mathcal{M}$, are defined respectively by:

$$V_p^n(\bar{m}|D) := \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Pi_n(\mathrm{d}m), \tag{5.3}$$

$$\hat{m}_p^n \in \operatorname*{argmin}_{\bar{m} \in \mathcal{M}} V_p^n(\bar{m}|D). \tag{5.4}$$

Remark 5.3.2. Under the standing assumption that $\mathcal{X}$ is a locally compact separable geodesic space, the existence of a population barycenter is granted if $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$, see [74, Theorem

2] or Lemma 5.2.3 for our own argument. The latter condition is equivalent to the model average $\bar{m}(\mathrm{d}x) := \mathbb{E}_P[m](\mathrm{d}x)$ having a finite p-moment, since

$$\begin{aligned} \int_{\mathcal{W}_p(\mathcal{X})} W_p(\delta_y, m)^p \Gamma(\mathrm{d}m) &= \int_{\mathcal{W}_p(\mathcal{X})} \int_{\mathcal{X}} \mathrm{d}(y,x)^p m(\mathrm{d}x)\Gamma(\mathrm{d}m) \quad (5.5) \\ &= \int_{\mathcal{X}} \mathrm{d}(y,x)^p \int_{\mathcal{W}_p(\mathcal{X})} m(\mathrm{d}x)\Gamma(\mathrm{d}m), \quad (5.6) \end{aligned}$$

for any $y \in \mathcal{X}$. If $\mathcal{M}$ is weakly closed the same reasoning gives the existence of a p-Wasserstein population barycenter of Γ over $\mathcal{M}$.

We summarize this discussion, for the case $\Gamma = \Pi_n$, in a simple statement:

Lemma 5.3.3 If $\mathcal{X}$ is a locally compact separable geodesic space, $\mathcal{M}$ is weakly closed, and the model average $\bar{m}^n(\mathrm{d}x) = \mathbb{E}_{\Pi_n}[m](\mathrm{d}x)$ has a.s. finite p-moment, then a.s. a p-Wasserstein barycenter estimator $\hat{m}_p^n$ over $\mathcal{M}$ exists.

We remark that even if $\Pi \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$, it may still happen that $\Pi_n \notin \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$. We provide a general condition on the prior prior Π ensuring that

$$a.s.: \ \Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X})) \ \text{ for all n},$$

and therefore the existence of a barycenter estimator.

Definition 5.3.4 We say that $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ is *integrable after updates* if it satisfies the conditions

1. For all $x \in \mathcal{X}, \ell > 1$:
$$\int_{\mathcal{M}} m(x)^\ell \Pi(\mathrm{d}m) < \infty.$$

2. For some $y \in \mathcal{X}, \varepsilon > 0$:
$$\int_{\mathcal{M}} \left(\int_{\mathcal{X}} \mathrm{d}(y,z)^p m(\mathrm{d}z) \right)^{1+\varepsilon} \Pi(\mathrm{d}m) < \infty.$$

Condition (2) above could be intuitively summarized with the notation $\Pi \in \mathcal{W}_{p+}(\mathcal{W}_p(\mathcal{X}))$.
Remark 5.3.5. If $\Pi \in \mathcal{P}(\mathcal{W}_{p,ac}(\mathcal{X}))$ has finite support, then Conditions (1) and (2) are satisfied. On the other hand, if Π is supported on a scatter-location family (see Section 6.3.4) containing one element with a bounded density and a finite p-moment, then Conditions (1) and (2) are fulfilled if for example supp(Π) is tight.

Lemma 5.3.6 Suppose that Π is integrable after updates. Then, for each $x \in \mathcal{X}$, the measure

$$\tilde{\Pi}(\mathrm{d}m) := \tfrac{m(x)\Pi(\mathrm{d}m)}{\int_{\mathcal{M}} \bar{m}(x)\Pi(\mathrm{d}\bar{m})},$$

is also integrable after updates.

Proof. We verify Property (1) first. Let $\ell > 1$ and $\bar{x} \in \mathcal{X}$ given. Then

$$\int_{\mathcal{M}} m(\bar{x})^\ell m(x)\Pi(\mathrm{d}m) \leq \left(\int_{\mathcal{M}} m(x)^s \Pi(\mathrm{d}m) \right)^{1/s} \left(\int_{\mathcal{M}} m(\bar{x})^{t\ell} \Pi(\mathrm{d}m) \right)^{1/t},$$

with s, t conjugate Hölder exponents. This is finite since Π fulfils Property (1).
We now establish Property (2). Let $y \in \mathcal{X}, \varepsilon > 0$. Then

$$\begin{aligned} &\int_{\mathcal{M}} \left(\int_{\mathcal{X}} \mathrm{d}(y,z)^p m(\mathrm{d}z)\right)^{1+\varepsilon} m(x)\Pi(\mathrm{d}m) \\ \leq\ &\left(\int_{\mathcal{M}} m(x)^s \Pi(\mathrm{d}m)\right)^{1/s} \left(\int_{\mathcal{M}} \left(\int_{\mathcal{X}} \mathrm{d}(y,z)^p m(\mathrm{d}z)\right)^{(1+\varepsilon)t} \Pi(\mathrm{d}m)\right)^{1/t}. \end{aligned}$$

The first term in the r.h.s. is finite by Property (1). The second term in the r.h.s. is finite by Property (2), if we take ε small enough and t close enough to 1. We conclude. □

Lemma 5.3.7 Suppose that Π is integrable after updates. Then for all $n \in \mathbb{N}$ and $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$, the posterior Π_n is also integrable after updates.

Proof. By Lemma 5.3.6, we obtain that Π_1 is integrable after updates. By induction, suppose Π_{n-1} has this property. Then as

$$\Pi_n(\mathrm{d}m) = \tfrac{m(x_n)\Pi_{n-1}(\mathrm{d}m)}{\int_{\mathcal{M}} \bar{m}(x_n)\Pi_{n-1}(\mathrm{d}\bar{m})},$$

we likewise conclude that Π_n is integrable after updates. □

We now make a set of simplifying assumptions which are supposed to hold from now:

Assumption 5.3.8 $\mathcal{M} = \mathcal{W}_{p,ac}(\mathcal{X})$, $\Pi \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$, $\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ ($\forall n$, a.s.).

5.3.1 On uniqueness of 2-Wasserstein barycenter

We now briefly consider the special case of $\mathcal{M} = \mathcal{W}_{2,ac}(\mathbb{R}^q)$, $\Pi \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathbb{R}^q))$, $\lambda =$ Lebesgue, and $\mathrm{d} =$ Euclidean distance until the end of this section. By Lemma 5.2.4 it is straightforward that the barycenter of Π_n is unique. We make an important observation regarding the absolute continuity of the barycenter, which is relevant since the model space $\mathcal{M} = \mathcal{W}_{2,ac}(\mathcal{X})$ is not weakly closed. The next remark states that in spite of Lemma 5.3.3 not being applicable, the existence of a barycenter belonging to the model space can still be guaranteed.

Remark 5.3.9. If $p = 2$, $\mathcal{X} = \mathbb{R}^q$, $\mathrm{d} =$ Euclidean distance, $\lambda =$Lebesgue measure, and

$$\Pi\left(\left\{m : \left\|\tfrac{\mathrm{d}m}{\mathrm{d}\lambda}\right\|_\infty < \infty\right\}\right) > 0, \tag{5.7}$$

then the population barycenter of Π_n exists, is unique, and is absolutely continuous. The only delicate point is the absolute continuity. This was proven in [65, Theorem 6.2] for compact finite-dimensional manifolds with lower-bounded Ricci curvature equipped with the volume measure, but one can read-off the non-compact but flat Euclidean case $\mathcal{X} = \mathbb{R}^q$ from the proof therein. If $|\mathrm{supp}(\Pi)| < \infty$ then (5.7) can be dropped, as shown in [2] or [65, Theorem 5.1].

We provide a useful characterization of barycenters, which is a generalization of the corresponding result in [6] where only the case $|\mathrm{supp}(\Pi)| < \infty$ is covered.

Lemma 5.3.10 Assume $p = 2$, $\mathcal{X} = \mathbb{R}^q$, d = Euclidean distance, λ =Lebesgue measure. Let $\hat{m}$ be the unique barycenter of Π. Then there exists a jointly measurable function $(m, x) \mapsto T^m(x)$ which is $\lambda(\mathrm{d}x)\Pi(\mathrm{d}m)$-a.s. equal to the unique optimal transport map from $\hat{m}$ to $m \in \mathcal{W}_2(\mathcal{X})$. Furthermore we have $x = \int T^m(x)\Pi(\mathrm{d}m)$, $\hat{m}(\mathrm{d}x)$-a.s.

Proof of Lemma 5.3.10. The existence of a jointly measurable version of the unique optimal maps is proved in [44]. Now assume that the last assertion is not true, so in particular

$$\begin{aligned} 0 &< \int \left(x - \int T^m(x)\Pi(\mathrm{d}m)\right)^2 \hat{m}(\mathrm{d}x) \\ &= \int |x|^2 \hat{m}(\mathrm{d}x) - 2\int\int x T^m(x)\Pi(\mathrm{d}m)\hat{m}(\mathrm{d}x) + \int \left(\int T^m(x)\Pi(\mathrm{d}m)\right)^2 \hat{m}(\mathrm{d}x). \end{aligned}$$

On the other hand, we have

$$\begin{aligned} &\int W_2\left(\left(\int T^m \Pi(\mathrm{d}m)\right)(\hat{m}), \bar{m}\right)^2 \Pi(\mathrm{d}\bar{m}) \\ \leq &\int\int \left[T^{\bar{m}}(x) - \int T^m(x)\Pi(\mathrm{d}m)\right]^2 \hat{m}(\mathrm{d}x)\Pi(\mathrm{d}\bar{m}) \\ = &\int\int \left[T^m(x)\right]^2 \hat{m}(\mathrm{d}x)\Pi(\mathrm{d}m) - \int \left(\int T^m(x)\Pi(\mathrm{d}m)\right)^2 \hat{m}(\mathrm{d}x), \end{aligned}$$

after a few computations. But, by Brenier's theorem [138, Theorem 2.12(ii)] we know that

$$\int\int (x - T^m(x))^2 \hat{m}(\mathrm{d}x)\Pi(\mathrm{d}m) = \int W_2(\hat{m}, m)^2 \Pi(\mathrm{d}m).$$

Bringing together these three observations, we deduce

$$\int W_2\left(\left(\int T^m \Pi(\mathrm{d}m)\right)(\hat{m}), \bar{m}\right)^2 \Pi(\mathrm{d}\bar{m}) < \int W_2(\hat{m}, m)^2 \Pi(\mathrm{d}m),$$

and in particular $\hat{m}$ cannot be the barycenter. □

5.3.2 Comparation with Bayesian model average

Let $\hat{m}$ be its unique population barycenter, and denote by $(m, x) \mapsto T^m(x)$ a measurable function equal $\lambda(\mathrm{d}x)\Pi(\mathrm{d}m)$ a.e. to the unique optimal transport map from $\hat{m}$ to $m \in \mathcal{W}_2(\mathcal{X})$. As a consequence of Lemma 5.3.10 we have $\hat{m} = (\int T^m \Pi(\mathrm{d}m))(\hat{m})$. Thanks to this fixed-point property, for all convex functions ϕ with at most quadratic growth, we have

$$\begin{aligned} \mathbb{E}_{\hat{m}}[\phi(x)] &= \int_{\mathcal{X}} \phi(x)\hat{m}(\mathrm{d}x) = \int_{\mathcal{X}} \phi\left(\int_{\mathcal{M}} T^m(x)\Pi(\mathrm{d}m)\right)\hat{m}(\mathrm{d}x) \\ &\leq \int_{\mathcal{X}}\int_{\mathcal{M}} \phi(T^m(x))\Pi(\mathrm{d}m)\hat{m}(\mathrm{d}x) = \int_{\mathcal{M}}\int_{\mathcal{X}} \phi(T^m(x))\hat{m}(\mathrm{d}x)\Pi(\mathrm{d}m) \\ &= \int_{\mathcal{M}}\int_{\mathcal{X}} \phi(x)m(\mathrm{d}x)\Pi(\mathrm{d}m) = \int_{\mathcal{X}} \phi(x)\int_{\mathcal{M}} m(\mathrm{d}x)\Pi(\mathrm{d}m) \\ &= \mathbb{E}_{\bar{m}}[\phi(x)], \end{aligned}$$

where $\bar{m} = \mathbb{E}_\Pi[m]$ is the Bayesian model average. We have used here Jensen's inequality and Fubini. Since we can replace Π by Π_n in this discussion, we have established that the 2-Wasserstein barycenter estimator is less dispersed than the Bayesian model average: namely, in the convex-order sense. In particular, we have established:

Lemma 5.3.11 Let $\bar{m}^n$ be the Bayesian model average and $\hat{m}^n$ the 2-Wasserstein barycenter of the posterior Π_n. Then $\mathbb{E}_{\bar{m}^n}[x] = \mathbb{E}_{\hat{m}^n}[x]$ and $\mathbb{E}_{\bar{m}^n}[\|x\|^2] \geq \mathbb{E}_{\hat{m}^n}[\|x\|^2]$, so the 2-Wasserstein barycenter estimator has less variance than the model average estimator.

5.4 Statistical Consistency

A natural question is whether our estimator is *consistent* in the statistical sense (see [118, 37, 51, 52], and references therein, for a detailed treatment on consistency). In short, consistency corresponds to the convergence of our estimator $\hat{m}_p^n$ towards the *true* model m_0, as we observe more i.i.d. data distributed like m_0. In Bayesian language this is a desirable *convergence of opinions* phenomenon [52].

Here and in the sequel $m_0^{(\infty)}$ denotes the product probability measure corresponding to the infinite sample $\{x_n\}_n$ of i.i.d. data distributed according to m_0. In the setting that concerns us, the correct notion of consistency at the level of the prior is given by:

Definition 5.4.1 A prior Π is said to be strongly consistent at m_0 for some topology $\mathcal{T}$, denoted $\mathcal{T}$-strongly consistent, if for each $\mathcal{T}$ open neighbourhood U of m_0 of $\mathcal{M}$, we have

$$\Pi_n(U^c) \to 0\,, \;\; m_0^{(\infty)} - a.s.$$

We are interested in the important question, of whether our Wasserstein barycenter estimator converges to the model m_0, i.e. we are after conditions which guarantee that

$$W_p(\hat{m}_p^n, m_0) \to 0, \;\; m_0^{(\infty)} a.s.$$

This is evidently linked to the question of strong consistency of the prior. The definition of Wasserstein consistent is a bit redundant, but we leave it explicitly given the importance for whole Section 5.4.

Definition 5.4.2 A prior Π is said to be p-Wasserstein strongly consistent at m_0 if for each open p-Wasserstein neighbourhood U of m_0, we have $\Pi_n(U^c) \to 0\,, \;\; m_0^{(\infty)} - a.s.$

We use W_p to denote throughout the Wasserstein distance both on $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ and on $\mathcal{W}_p(\mathcal{X})$, not to make the notation heavier. It is straightforward that p-Wasserstein convergence implies p-Wasserstein strongly consistency.

Proposition 5.4.3 If $W_p(\Pi_n, \delta_{m_0}) \to 0, m_0^{(\infty)}$-a.s. then Π is p-Wasserstein strongly consistent at m_0.

Proof. Since p-Wasserstein convergence implies weak convergence, we have $\Pi_n \to \delta_{m_0}$ weakly, so for any neighbourhood U of m_0 by Portmanteau's Theorem [52, Thm. A.2] the closed set U^c satisfies $\limsup \Pi_n(U^c) \leq \delta_{m_0}(U^c) = 0$. □

Indeed, p-Wasserstein convergence implies that the Wasserstein barycenter estimator converges.

Proposition 5.4.4 $W_p(\Pi_n, \delta_{m_0}) \to 0\,(m_0^{(\infty)}\text{-a.s.}) \Rightarrow W_p(\hat{m}_p^n, m_0) \to 0\,(m_0^{(\infty)}\text{-a.s.})$.

Proof of Proposition 5.4.4. We have, by minimality of the barycenter

$$W_p(\Pi_n, \delta_{m_0})^p = \int_{\mathcal{M}} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) \geq \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \, \Pi_n(\mathrm{d}m).$$

On the other hand,

$$W_p(m_0, \hat{m}_p^n)^p \ \leq c\, W_p(m, \hat{m}_p^n)^p + c\, W_p(m, m_0)^p\,,\ \forall m,$$

where the constant c only depends on p. We conclude by

$$\begin{aligned} W_p(m_0, \hat{m}_p^n)^p \ &\leq c \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \, \Pi_n(\mathrm{d}m) + c \int_{\mathcal{M}} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) \\ &= c \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \, \Pi_n(\mathrm{d}m) + c\, W_p(\Pi_n, \delta_{m_0})^p \\ &\leq 2c\, W_p(\Pi_n, \delta_{m_0})^p. \end{aligned}$$

□

The celebrated Schwartz's theorem [118] provides sufficient conditions for strong consistency. See [52, Proposition 6.16] for a more modern treatment. A key ingredient in Schwartz' approach is the notion of Kullback-Leibler support:

Definition 5.4.5 A measure m_0 belongs to the Kullback-Leibler support of Π, denoted $m_0 \in \mathrm{KL}(\Pi)$, if $\Pi\left(m : D_{KL}\left(m_0 || m\right) < \varepsilon\right) > 0$ for every $\varepsilon > 0$, where $D_{KL}\left(m_0 || m\right) = \int \log \frac{m_0}{m} \mathrm{d}m_0$.

Schwartz's theorem is the basic result on posterior consistency for dominated models: the true density m_0 should belong to the Kullback-Leibler support of the prior and the hypothesis $m = m_0$ should be testable against complements of neighborhoods of m_0.

Theorem 5.4.6 A test ϕ_n is a measurable function $\phi_n : \mathcal{X}^n \to [0, 1]$ and for a density m we denote $M^n \phi_n := \mathbb{E}_M\left[\phi_n\left(X_1, ..., X_n\right)\right] = \int \phi_n(x_1, \ldots, x_n) \prod_{\mathrm{i}=1}^n m(x_\mathrm{i}) \lambda(\mathrm{d}x_1) \ldots \lambda(\mathrm{d}x_n)$. If $m_0 \in KL\left(\Pi\right)$ and for every $\mathcal{T}$-neighborhood U of m_0, there exist tests ϕ_n such that $M_0^n \phi_n \to 0$ and $\sup_{M \in U^c} M^n\left(1 - \phi_n\right) \to 0$, then the prior Π is $\mathcal{T}$-strongly consistent at m_0.

If the model space $\mathcal{M}$ is smoothly parameterised by a finite-dimensional compact parameter space Θ and the parametrization map $\mathcal{T}$ is injective and continuous, then consistency tests exist for $m_0 \in KL(\Pi)$ (see [52]). We desire to specialize in the consistency for p-Wasserstein spaces. The following proposition show a a hierarchy within consistency on p-Wasserstein, and also that are stronger that consistency on weak topology.

Proposition 5.4.7 If Π is p-Wasserstein strongly consistent at m_0 with $p \geq 1$, then Π is q-Wasserstein strongly consistent at m_0 with $q < p$. Besides, Π is strongly consistent at m_0 for the weak topology.

Proof. If $q < p$ by Hölder's inequality we have that $W_q \leq W_p$, so $U_q = \{m : W_q(m, m_0) < \varepsilon\} \supseteq \{m : W_p(m, m_0) < \varepsilon\} = U_p$ so $\Pi_n\left(U_q^c\right) \leq \Pi_n\left(U_p^c\right) \to 0$, $m_0^{(\infty)} - a.s.$ Also, as the Prokhorov metric d_W metrizes weak convergence and $\mathrm{d}_W^2 \leq W_1$ [see [53]], then we conclude that Π is strongly consistent at m_0 under the weak topology. □

Remark 5.4.8. As mentioned in [52, Proposition 6.2], strong consistency with respect to the weak topology is equivalent to the $m_0^{(\infty)}$-almost sure weak convergence of Π_n to δ_{m_0}.

Remark 5.4.9. As can be derived from [52, Example 6.20], in our particular setting, we have

$$\Pi \text{ is strongly consistent at } m_0 \text{ w.r.t. the weak topology } \iff m_0 \in KL(\Pi).$$

We assume throughout Section 5.4 that

$$m_0 \in KL(\Pi) \quad \text{and} \quad m_0 \in \mathcal{M}.$$

This implies that the model is *correct* or *well-specified* as discussed in [16, 57, 67, 68]. This setting could be slightly relaxed in the *misspecified* framework dealt with in those works by considering the reverse Kullback–Leibler projection on $\mathcal{M}$ instead of the true model m_0, i.e. the unique model $\hat{m}_0 \in \mathcal{M}$ that minimizes $D_{KL}(m_0||\hat{m}_0)$ over $\mathcal{M}$.

We can now state our main result concerning consistency of the barycenter estimator:

Theorem 5.4.10 Suppose that Π fulfils the following condition:

(A) There exist $\lambda_0 > 0, x_0 \in \mathcal{X}$ such that $\sup_{m \in \mathrm{supp}(\Pi)} \int_{\mathcal{X}} \mathrm{e}^{\lambda_0 \mathrm{d}^p(x,x_0)} \mathrm{d}m(x) < +\infty$.

Then under our standing assumptions (in particular, $m_0 \in KL(\Pi)$) we have that Π is p-Wasserstein strongly consistent at m_0, $W_p(\Pi_n, \delta_{m_0}) \to 0$ ($m_0^{(\infty)}$-a.s.), and the barycenter estimator is consistent in the sense that

$$W_p(\hat{m}_p^n, m_0) \to 0, \quad m_0^{(\infty)} - a.s.$$

Remark 5.4.11. Notice that Assumption (A) implies that $\mathrm{diam}(\Pi) := \sup\{ W_p(m, \bar{m}) : m, \bar{m} \in \mathrm{supp}(\Pi)\} < \infty$. A typical example where this holds is in the finitely parametrized case, when the parameter space is compact and the parametrization function continuous. We stress that $\mathcal{X}$ may be unbounded but $\mathrm{diam}(\Pi)$ still be finite.

The proof of Theorem 5.4.10 is given at the end of this part. Towards this goal, we start with a direct sufficient condition for the convergence of $\hat{m}_p^n$ to m_0.

By Remark 5.4.8, if Π is p-Wasserstein strongly consistent at m_0, then $m_0^{(\infty)}$-a.s. weak convergence of Π_n to δ_{m_0}. It is known that if their p-moments also converge then $W_p(\Pi_n, \delta_{m_0}) \to 0, m_0^{(\infty)}$-a.s. The following proposition gives conditions for the convergence of their moments.

Proposition 5.4.12 If Π is p-Wasserstein strongly consistent at m_0 and $\mathrm{diam}(\Pi) < \infty$, then $W_p(\Pi_n, \delta_{m_0}) \to 0$ and in particular $W_p(\hat{m}_p^n, m_0) \to 0$ ($m_0^{(\infty)} - a.s.$).

Proof of Proposition 5.4.12. Let $B = \{m : W_p(m, m_0) < \varepsilon\}$ and ε arbitrary, then

$$\begin{aligned} W_p(\Pi_n, \delta_{m_0})^p &= \int_{\mathcal{M}} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) \\ &\leq \int_B W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) + \int_{B^c} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) \\ &\leq \varepsilon^p + \int_{B^c} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m). \end{aligned}$$

Since ε is arbitrary, we only need to check that the second term goes to zero. Strong consistency implies $\Pi_n(B^c) \to 0$ $(m_0^{(\infty)} - a.s.)$, and since $\mathrm{supp}(\Pi_n) \subset \mathrm{supp}(\Pi)$, we have

$$\int_{B^c} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) \leq \mathrm{diam}(\Pi)^p \Pi_n(B^c) \to 0 \ (m_0^{(\infty)} - a.s.).$$

□

We now provide the proof of Theorem 5.4.10. If the Wasserstein metric was bounded, the argument would be as in [52, Example 6.20], where the main tool is Hoeffding's inequality. In general Wasserstein metrics are unbounded if $\mathcal{X}$ is itself unbounded, and this forces us to assume Condition (A) in Theorem 5.4.10. The argument still rests on the concentration of measure phenomenon:

Proof of Theorem 5.4.10. We will apply Proposition 5.4.12. First we show that if U is any $\mathcal{W}_p(\mathcal{X})$-neighbourhood of m_0 then $\liminf_n \Pi_n(U) \geq 1$ ($m_0^{(\infty)}$-a.s.). According to Schwartz Theorem (see [52, Theorem 6.17]), under the assumption that $m_0 \in KL(\Pi)$, it suffices to find for each such U a sequence of measurable functions $\phi_n : \mathcal{X}^n \to [0, 1]$ s.t.

1. $\phi_n(x_1, \ldots, x_n) \to 0$, $m_0^{(\infty)} - a.s$, and
2. $\limsup_n \frac{1}{n} \log \left(\int_{U^c} m^n (1 - \phi_n) \Pi(\mathrm{d}m) \right) < 0$.

For this purpose, first we will construct tests $\{\phi_n\}_n$ that satisfy the above conditions (Point 1 and Point 2) over an appropriate subbase of neighbourhood, to finally extend it to general neighborhoods. It is known that $\mu_k \to \mu$ on W_p iff for all continuous functions ψ with $|\psi(x)| \leq K(1 + \mathrm{d}^p(x, x_0))$, $K \in \mathbb{R}$ it holds that $\int_{\mathcal{X}} \psi(x) \mathrm{d}\mu_n(x) \to \int_{\mathcal{X}} \psi(x) \mathrm{d}\mu(x)$; see [139]. Given such ψ and $\varepsilon > 0$ we define the open set

$$U_{\psi,\varepsilon} := \left\{ m : \int_{\mathcal{X}} \psi(x) \mathrm{d}m(x) < \int_{\mathcal{X}} \psi(x) \mathrm{d}m_0(x) + \varepsilon \right\}.$$

These sets form a subbase for the p-Wasserstein neighborhood system at the distribution m_0, and w.l.o.g. we can assume that $K = 1$ by otherwise considering $U_{\psi/K,\varepsilon/K}$ instead. Given a neighborhood $U := U_{\psi,\varepsilon}$ as above, we define the test functions

$$\phi_n(x_1, \ldots, x_n) = \begin{cases} 1 & \frac{1}{n} \sum_{\mathrm{i}=1}^n \psi(x_\mathrm{i}) > \int_{\mathcal{X}} \psi(x) \mathrm{d}m_0(x) + \frac{\varepsilon}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

By law of large numbers, $m_0^{(\infty)}$-a.s: $\phi_n(x_1, \ldots, x_n) \to 0$, so Point 1 is verified. Point 2 is trivial if $r := \Pi(U^c) = 0$, so assume from now on that $r > 0$. Finite p-exponential moments of $m \in \mathrm{supp}(\Pi)$ imply that the random variable $Z = 1 + \mathrm{d}^p(X, x_0)$ with $X \sim m$ has a moment-generating function $\mathcal{L}_m(t)$ which is finite for all $\lambda_0 \geq t \geq 0$, namely

$$\mathcal{L}_m(t) := \mathbb{E}_m \left[\mathrm{e}^{tZ} \right] = \mathrm{e}^t \int_{\mathcal{X}} \mathrm{e}^{t\mathrm{d}^p(x,x_0)} \mathrm{d}m(x) < +\infty.$$

Since all the moments of Z are non-negative, we can bound all the k-moments by

$$\mathbb{E}_m \left[Z^k \right] \leq k! \mathcal{L}_m(t) t^{-k}, \ \forall \lambda_0 \geq t > 0.$$

Thanks to the above bound, we have

$$\int_{\mathcal{X}} |\psi(x)|^k \mathrm{d}m(x) \leq \int_{\mathcal{X}} (1 + \mathrm{d}^p(x, x_0))^k \mathrm{d}m(x) \leq k! \mathcal{L}_m(t) t^{-k}.$$

We may apply Bernstein's inequality in the form of [79, Corollary 2.10] to the random variables $\{-\psi(x_{\mathrm{i}})\}_{\mathrm{i}}$ under the measure $m^{(\infty)}$ on $\mathcal{X}^{\mathbb{N}}$, obtaining for any $\alpha < 0$ that

$$m^{(\infty)}\left(\sum_{\mathrm{i}=1}^{n}\left[\psi(x_{\mathrm{i}}) - \int_{\mathcal{X}}\psi(x)\mathrm{d}m(x)\right] \leq \alpha\right) \leq \mathrm{e}^{-\frac{\alpha^2}{2(v-c\alpha)}},$$

where $v := 2n\mathcal{L}_m(t)t^{-2}$, $c := t^{-1}$, and $0 < t \leq \lambda_0$. Using the definition of U^c we deduce

$$\begin{aligned}
\int_{U^c} m^n(1-\phi_n)\Pi(\mathrm{d}m) &= \int_{U^c} m^n\left(\tfrac{1}{n}\sum_{\mathrm{i}=1}^{n}\psi(x_{\mathrm{i}}) \leq \int_{\mathcal{X}}\psi(x)\mathrm{d}m_0(x) + \tfrac{\varepsilon}{2}\right)\Pi(\mathrm{d}m)\\
&\leq \int_{U^c} m^n\left(\tfrac{1}{n}\sum_{\mathrm{i}=1}^{n}\psi(x_{\mathrm{i}}) \leq \int_{\mathcal{X}}\psi(x)\mathrm{d}m(x) - \tfrac{\varepsilon}{2}\right)\Pi(\mathrm{d}m)\\
&= \int_{U^c} m^n\left(\sum_{\mathrm{i}=1}^{n}\left[\psi(x_{\mathrm{i}}) - \int_{\mathcal{X}}\psi(x)\mathrm{d}m(x)\right] \leq -\tfrac{n\varepsilon}{2}\right)\Pi(\mathrm{d}m)\\
&\leq \int_{U^c}\exp\left\{-\tfrac{n\varepsilon^2}{2}\tfrac{t^2}{8\mathcal{L}_m(t)+t\varepsilon}\right\}\Pi(\mathrm{d}m)\\
&\leq r\,\exp\left\{-\tfrac{n\varepsilon^2}{2}\tfrac{t^2}{8\sup_{m\in U^c\cap\mathrm{supp}(\Pi)}\mathcal{L}_m(t)+t\varepsilon}\right\}.
\end{aligned}$$

Under our assumption (A) we conclude as desired that

$$\limsup_n \tfrac{1}{n}\log\left(\int_{U^c} m^n(1-\phi_n)\Pi(\mathrm{d}m)\right) \leq -\tfrac{t^2\varepsilon^2}{16\sup_{m\in U^c\cap\mathrm{supp}(\Pi)}\mathcal{L}_m(t)+2t\varepsilon} < 0.$$

Now, a general neighborhood U contains a finite intersection of $N \in \mathbb{N}$ neighborhoods from the subbase, i.e. $\bigcap_{\mathrm{i}=1}^{N} U_{\psi_{\mathrm{i}},\varepsilon_{\mathrm{i}}} \subset U$, so

$$\int_{U^c} m^n(1-\phi_n)\Pi(\mathrm{d}m) \leq \sum_{\mathrm{i}=1}^{N}\int_{U^c_{\psi_{\mathrm{i}},\varepsilon_{\mathrm{i}}}} m^n(1-\phi_n)\Pi(\mathrm{d}m),$$

and therefore we may conclude as in the subbase case that Point 2 is verified. All in all we have established that Π is p-Wasserstein strongly consistent at m_0, so we conclude by Proposition 5.4.12 thanks to our Assumption (A). □

Remark 5.4.13. The above is a self-contained proof for consistency in Wasserstein topologies. An alternative argument could be as follows: Under Assumption (A), and if $p \geq 2$, the measures in the support of Π enjoy the Talagrand T_1 inequality (cf. [139, Theorem 22.10]) from which the Wasserstein distance is controlled by a relative entropy. By [146, Theorem 5] it is possible, under additional integrability assumptions on the densities in the support of Π, to control relative entropies by Hellinger distances. Hence one may leverage existing results on consistency (plausibly with convergence rates) for the Hellinger distance in order to obtain respective results for Wasserstein distances.

The next result states that if the prior is consistent in the p-Wasserstein sense, then under some alternative conditions we have the p-Wasserstein convergence of the posterior Π_n to δ_{m_0} for models in the Kullback-Leibler support $KL(\Pi)$ of Π, thus our p-Wasserstein barycenter estimator $\hat{m}_p^n$ converge to the true model m_0.

Proposition 5.4.14 If Π is p-Wasserstein strongly consistent at $m_0 \in KL(\Pi)$ then

$$W_p(\Pi_n, \delta_{m_0}) \to 0 (m_0^{(\infty)}\text{-a.s.})$$

if any of the following conditions is fulfilled:

1. $\mathrm{supp}(\Pi)$ is bounded,

2. $\int W_p^q(m, m_0)\Pi_n(\mathrm{d}m) < C$ for some $q > p$ and $C > 0$,

3. the likelihood function $\Lambda_n(m)$ converge $m_0^{(\infty)}$-a.s. to 0 with L^∞_Π-norm as $n \to \infty$, on the sets $B^c(m_0, \varepsilon) = \{\nu | W_p(\nu, m_0) > \varepsilon\}$ for every $\varepsilon > 0$.

In either case, the barycenter is consistent at m_0 in the sense that

$$W_p(\hat{m}_p^n, m_0) \to 0 \; m_0^{(\infty)} - a.s.$$

Proof. For condition (1) it was proved on 5.4.12. Under condition (2) and applying Hölder inequality choosing $\frac{1}{s} + \frac{1}{r} = 1$ with $\frac{q}{p} = r$ we have

$$\begin{aligned}\int_{B^c} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) &= \int_{\mathcal{M}} \mathbf{1}_{B^c} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) \\ &\leq \left[\int_{\mathcal{M}} W_p(m, m_0)^{pr} \, \Pi_n(\mathrm{d}m)\right]^{\frac{1}{r}} \Pi_n(B^c)^{\frac{1}{s}} \\ &\leq C^{\frac{1}{r}} \Pi_n(B^c)^{\frac{1}{s}} \to 0 \,, \;\; m_0^{(\infty)} - a.s.\end{aligned}$$

Finally, under condition (3) over Λ_n we have that

$$\begin{aligned}\int_{B^c} W_p(m, m_0)^p \, \Pi_n(\mathrm{d}m) &= \int_{\mathcal{M}} \mathbf{1}_{B^c} W_p(m, m_0)^p \Lambda_n(m) \Pi(\mathrm{d}m) \\ &\leq \left[\int_{\mathcal{M}} W_p(m, m_0)^p \Pi(\mathrm{d}m)\right] \|\Lambda_n(m) \mathbf{1}_{B^c}\|_\infty \\ &\leq C \|\Lambda_n(m) \mathbf{1}_{B^c}\|_\infty \to 0 \,, \;\; m_0^{(\infty)} - a.s.\end{aligned}$$

By Prop. 5.4.4 the barycenter is consistent at m_0. □

The last result about consistency of Wasserstein barycenter estimator show that, if the Bayesian model converges in $\mathcal{W}_p$ to the true model m_0, then our estimator converges too. Recall that the model average is given by $\bar{m}^n(\mathrm{d}x) = \mathbb{E}_{\Pi_n}[m](\mathrm{d}x)$.

Lemma 5.4.15 If $m_0^{(\infty)}$-a.s. the p-moments of the model average converge to those of $m_0 \in \mathrm{KL}(\Pi)$, then $W_p(\Pi_n, \delta_{m_0}) \to 0$ ($m_0^{(\infty)}$-a.s.). Also $W_p(\hat{m}_p^n, m_0) \to 0$ ($m_0^{(\infty)}$-a.s.).

Proof of Lemma 5.4.15. By [52, Example 6.20] we already know that the prior is strongly consistent at m_0 with respect to the weak topology (rather than the p-Wasserstein topology). Notice that

$$\int W_p(m, \delta_x)^p \Pi_n(\mathrm{d}m) = \int \int \mathrm{d}(x, z)^p m(\mathrm{d}z) \Pi_n(\mathrm{d}m) = \int \mathrm{d}(x, z)^p \bar{m}^n(\mathrm{d}z),$$

so a.s. $\Pi_n \to \delta_{m_0}$ not only weakly but in W_p. Conclude by Proposition 5.4.4. □

Remark 5.4.16. Since by Remark 5.4.9, the prior is strongly consistent at $m_0 \in KL(\Pi)$ with respect to the weak topology, [52, Theorem 6.8] and the discussion thereafter imply that the model average is consistent at m_0 too.

5.5 Examples of Bayesian Wasserstein Barycenter

In this section we present two examples in order to show and validate our proposal. The first case is a didactic example where it is possible to calculate the estimators explicitly, to understand the similarities and differences of each selection criterion. The second case is an example with real data, to show evidence of the utility of this estimator vs classical estimators.

5.5.1 The conjugate prior over Gaussian distributions

In this example we show that the conjugate prior for Gaussian distributions is a consistent continuous prior which allows us to calculate the model average and the 2-Wassestein barycenter in closed form.

Consider the observations $D = \{x_1, ..., x_n\}$ generated by the true model $m_0 = \mathcal{N}(\bar{\mu}, \bar{\sigma}^2) \in \mathcal{M}$, where $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) | \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$. Let us also choose the prior over models by placing a Normal-inverse-gamma distribution (NIG) over the parameters (μ, σ^2), given by $\mathcal{NIG}(\mu, \sigma^2 | \mu_0, \lambda_0, \alpha_0, \beta_0) = \mathcal{N}(\mu | \mu_0, \sigma^2/\lambda_0)\mathcal{IG}(\sigma^2 | \alpha_0, \beta_0)$, $\mu_0 \in \mathbb{R}$ and $\lambda_0, \alpha_0, \beta_0 \in \mathbb{R}^+$, which induces a prior Π over models $\mathcal{M}$. As the NIG distribution is conjugate to the Gaussian likelihood, the posterior distribution of the model parameters is given by $(\mu, \sigma^2 | x_1, ..., x_n) \sim \mathcal{NIG}(\mu_n, \lambda_n, \alpha_n, \beta_n)$ with $\mu_n = \frac{\lambda_0 \mu_0 + n\bar{x}_n}{\lambda_0 + n}$, $\lambda_n = \lambda_0 + n$, $\alpha_n = \alpha_0 + \frac{n}{2}$ and $\beta_n = \beta_0 + \frac{1}{2}\left(n\bar{s}_n + \frac{n\lambda_0(\bar{x}_n - \mu_0)^2}{\lambda_0 + n}\right)$, where $\bar{x}_n = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} x_{\mathrm{i}}$ and $\bar{s}_n = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x}_n)^2$. See more details in [84].

We will show that the above prior is strongly consistent at m_0. The mean of the posterior is $(\mu_n, \frac{\beta_n}{\alpha_n - 1/2})$, which converges to $(\bar{x}, \bar{s}) = \lim_{n\to\infty}(\bar{x}_n, \bar{s}_n)$ and are respectively the mean and variance of m_0, due to the strong law of large numbers. Since the variance of the posterior is $\mathcal{O}(\frac{1}{n})$ in both variables (μ, σ^2), the posterior converges a.s. in the weak topology to the point mass at $(\bar{\mu}, \bar{\sigma}^2)$, therefore, NIG prior is strongly consistent at m_0 in the weak topology.

Additionally, we know [85] that the MAP estimator is $\mathcal{N}(x | \mu_n, \frac{\beta_n}{\alpha_n + \frac{3}{2}})$, while the model average is the Student's $t-$distribution $t_{2\alpha_n}(x | \mu_n, \frac{\beta_n(1+\lambda_n)}{\alpha_n \lambda_n})$ with variance is $\frac{\beta_n(1+\lambda_n)}{(\alpha_n - 1)\lambda_n}$. This reveals the non-Gaussianity of the model average, despite the prior (and all posteriors) being Gaussian.

The second moment of the model average is given by $\mu_n^2 + \frac{\beta_n(1+\lambda_n)}{(\alpha_n - 1)\lambda_n} = \mu_n^2 + \frac{\bar{s}_n}{1 + \mathcal{O}(\frac{1}{n})} + \mathcal{O}(\frac{1}{n})$, which converges to the second moment of m_0. By Lemma 5.4.15 the 2-Wasserstein barycenter of the posterior (which exists) converges a.s. to m_0 and is given by $\mathcal{N}(x | \mu_n, \hat{\sigma}^2)$. From [7, Thm. 3.10], denoting σ_m^2 the variance for a model $m \in \mathcal{M}$, the barycenter variance $\hat{\sigma}^2$ satisfies

$$\hat{\sigma}^2 = \int \left(\hat{\sigma}\sigma_m^2\hat{\sigma}\right)^{1/2} \Pi_n(\mathrm{d}m) = \hat{\sigma} \int \sigma_m \Pi_n(\mathrm{d}m).$$

Furthermore, using the variance posterior $\mathcal{IG}(\sigma^2 | \alpha_n, \beta_n)$ and the change of variable $z = \sigma^2$ we have

$$\hat{\sigma} = \int \sigma_m \Pi_n(\mathrm{d}m) = \int z^{1/2} \mathcal{IG}(z | \alpha_n, \beta_n)\mathrm{d}z = \frac{\beta_n^{1/2}\Gamma(\alpha_n - \frac{1}{2})}{\Gamma(\alpha_n)}.$$

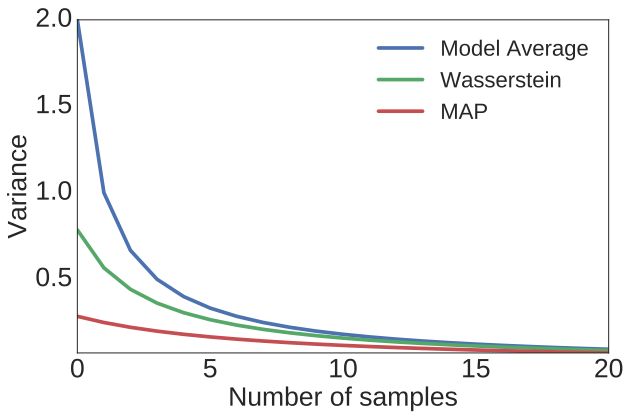

Figure 5.2: Variance of the selected model under three criterion.

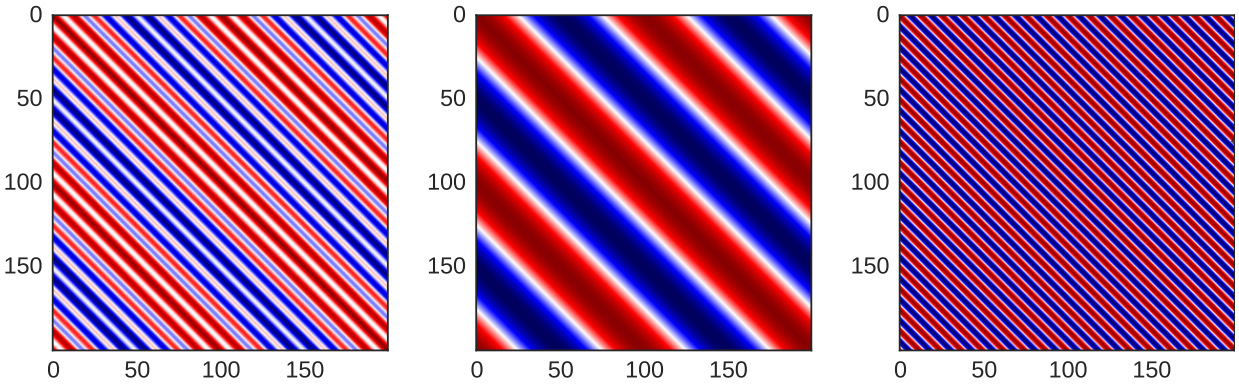

Figure 5.3: Barycenter (first) of two covariance matrices (second, third).

Thus, the MAP, average and barycenter models have the same mean μ_n but different variance. Fig. 5.2 (left) compares the variances in a numerical example with $a_0 = 2$, $\lambda_0 = 1$ and $\beta_n = 1$. Note that the Wasserstein barycenter estimator has higher variance than MAP, but less than the model average.

5.5.2 Bayesian Wasserstein learning for Gaussian processes using a real-world data

In this examples, we train a Gaussian process (see Chapter 2) using the proposed Bayesian Wasserstein barycenter estimator. Although we have not explained how to calculate barycenter in practice, all this is detailed in Chapter 6, so now we will make a simple description. Given the posterior distribution over hyperparameters (see Section 2.2), using MCMC, we generated k independent mean vectors and covariance matrices. We then found the barycenter GP by averaging the mean vectors and applying a fixed-point algorithm for Gaussian case [5, 6]. According to Prop. 6.1.3, the number of sampled models k is data-dependent, thus we searched for k based on empirical convergence of the barycenter. In Fig. 5.3 shows the covariance matrix of the 2-Wasserstein barycenter between two Gaussians distribution with cosine-based covariance matrices of dimension 200×200.

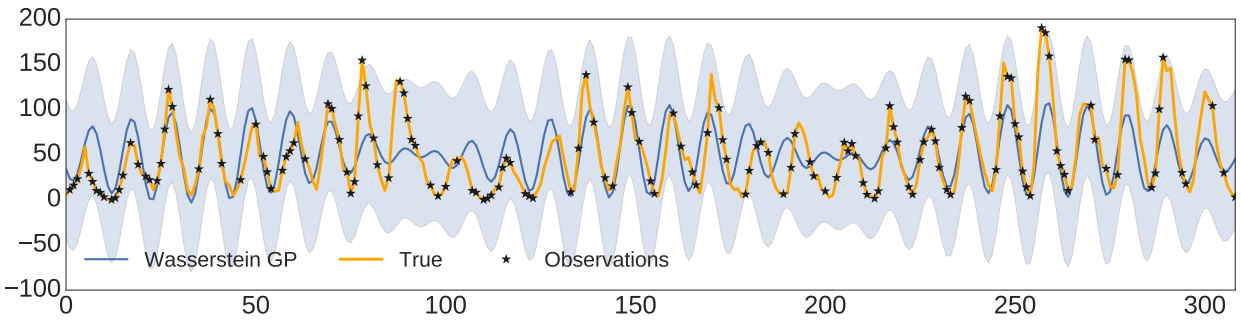

Figure 5.4: A Gaussian process with a cosine kernel, learned with Wasserstein barycenter.

Score	MAE			RMSE		
Model / Dataset	Obs	Test	Total	Obs	Test	Total
MAP	29.380	29.067	29.223	37.515	36.483	37.001
Model Average	27.631	25.057	26.340	35.460	31.648	33.602
Wasserstein	23.143	22.552	22.846	30.874	28.977	29.937

Table 5.1: Result of model selection with Sunspots dataset.

We considered the Sunspots time series (available from [98]) between 1700 and 2008 and used half of the data (154 points) for training and the rest for testing. Setting a non-informative prior [49] over the hyperparameters, we define a GP with constant mean function and cosine covariance kernel. We remind the reader that in this case, $\mathcal{M}$ is the space of all Gaussian processes [77] and that the true model m_0 is unknown.

Fig. 5.4 shows the posterior predictive mean and the 95%-confidence interval of the Wasserstein barycenter model. Note that our model was able to recover a varying-waveform, close-to-periodic, signal using a prior with support only for perfectly-periodic time series. This result validates the proposed methodology to handle model mismatch, and in this case, recover the signal frequency. Table 5.1 shows that the model selected with Wasserstein barycenter has a better performance than MAP and model average in mean absolute error (MAE) and square root mean error (RMSE) on observed and test data.

Chapter 6

Computing the Wasserstein Barycenter

"The most important part of learning is actually forgetting."

– Naftali Tishby

The main results presented in this Chapter are included in the preprint paper [12]: *Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. arXiv preprint arXiv:1805.10833, 2018.*

In this chapter, we discuss possible ways to compute and approximate the population Wasserstein barycenter. This calculation is a crucial step in constructing our Bayesian Wasserstein barycenter estimator, eq. (5.4). We begin this development in Section 6.1 with a straightforward Monte-Carlo method to approximate our estimator with an empiric version. This method motivates us to summarise in Section 6.1.1 the essentials of the gradient descent method in Wasserstein space, developed in [93, 6], where we can fix necessary notation and ideas for our main contribution. We introduce a novel algorithm for computation of barycenters in Section 6.2, which can be seen as a *stochastic gradient descent* method on Wasserstein space. This algorithm is the last main contribution of this work, followed by Section 6.2.1, where we present a generalisation of this method, named *batch stochastic gradient descent*, which it is a mixed idea between empirical and stochastic estimators.

To illustrate the applicability of our proposed approach, and several methods, in Section 6.3, we give explicit formulas for the proposed method for several useful families of distributions. To close this chapter, in Section 6.4 we provide a comprehensive numerical experiment to illustrate the advantages of the Bayesian Wasserstein barycenter over the Bayesian model average, besides to show that the stochastic gradient descent method is a superior alternative for their computation versus a Monte Carlo approximation.

For our results, we assume we are capable of generating independent models m_{i} from the posteriors Π_n and the prior Π for $\mathrm{i} = 1, ..., k$. In the parametric setting, we can use efficient Markov Chain Monte Carlo (MCMC) techniques [56] or transport-based sampling procedures [41, 94, 64, 78] to generate samples of parameters θ_{i}, for then via the parametrisation function get the models $m_{\mathrm{i}} = m_{\theta_{\mathrm{i}}}$ for $\mathrm{i} = 1, ..., k$.

6.1 Empirical Wasserstein barycenter

In general, we cannot calculate integrals over the model space $\mathcal{M}$, so we must approximate such integrals by, e.g. Monte Carlo methods. For this reason, we discuss the *empirical Wasserstein barycenter*. When $|\mathrm{supp}(\Pi)| < \infty$ this is related to [20, Theorem 3.1].

Definition 6.1.1 Given $m_{\mathrm{i}} \overset{\mathrm{iid}}{\sim} \Pi_n$ for $\mathrm{i} \leq k$, the empirical measure $\Pi_n^{(k)}$ over models is

$$\Pi_n^{(k)} := \tfrac{1}{k} \textstyle\sum_{\mathrm{i}=1}^{k} \delta_{m_{\mathrm{i}}} \ \in \mathcal{P}(\mathcal{M}) \,.$$

Note that if a.s. $\Pi_n \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$ then a.s. $\Pi_n^{(k)} \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$, so all hypothesis about Π_n stand on $\Pi_n^{(k)}$. Using $\Pi_n^{(k)}$ instead of Π_n, we define the p-Wasserstein empirical Bayes risk $V_p^{(n,k)}(\bar{m}|D)$, as well as a corresponding empirical Bayes estimator $\hat{m}_p^{(n,k)}$. If $\mathcal{M} = \mathcal{W}_p$ then $\hat{m}_p^{(n,k)}$ is referred to as a p-Wasserstein empirical barycenter of Π_n ([18]).

Remark 6.1.2. It is known that a.s. $\Pi_n^{(k)}$ converges weakly to Π_n as $k \to \infty$. If Π_n has finite p-th moments, by strong law of large numbers we have p-th moments convergence:

$$\textstyle\int W_p(m, m_0)^p \Pi_n^{(k)}(\mathrm{d}m) = \frac{1}{k} \sum_{\mathrm{i}=1}^{k} W_p(m_{\mathrm{i}}, m_0)^p \to \int W_p(m, m_0)^p \Pi_n(\mathrm{d}m) \text{ a.s.}$$

Thus a.s. $\Pi_n^{(k)} \to \Pi_n$ in $\mathcal{W}_p$. By [74, Theorem 3], any sequence of empirical barycenters $(\hat{m}_n^k)_{k\geq 1}$ of $(\Pi_n^k)_{k\geq 1}$ converges (up to subsequence) in p-Wasserstein distance to a (population) barycenter $\hat{m}_n$ of Π_n. Combining these facts, the following is immediate:

Lemma 6.1.3 If $W_p(\Pi_n, \delta_{m_0}) \to 0$, $m_0^{(\infty)}$-a.s., there exists a data-dependent sequence $k_n := k_n(x_1, \ldots, x_n)$ such that $(\hat{m}_n^{k_n})_{n\geq 1}$ satisfy $W_p(\hat{m}_n^{k_n}, m_0) \to 0$, $m_0^{(\infty)}$-a.s.

Proof of Lemma 6.1.3. Since W_p is a metric we have that $W_p(\hat{m}_n^k, m_0) \leq W_p(\hat{m}_n^k, \hat{m}_n) + W_p(\hat{m}_n, m_0)$ for all $k, n \geq 0$, and thanks to Proposition 5.4.4 the last term tends to zero $m_0^{(\infty)}$-a.s. as $n \to \infty$. Using a diagonal argument, for each $\hat{m}_n$ exists k_n (determined by the data-dependent Π_n) s.t. the empirical barycenter $\hat{m}_n^k$ satisfies $W_p(\hat{m}_n^{k_n}, \hat{m}_n) \leq \frac{1}{n}$, thus obtaining the convergence. □

6.1.1 Gradient descent on Wasserstein space

We first survey the gradient descent method for the computation of 2-Wasserstein empirical barycenters. This method will serve as a motivation for the subsequent development of the stochastic gradient descent in Sections 6.2 and 6.2.1.

From now until the end of the article we strengthen Assumption 5.3.8 by further assuming (cf. Remark 5.3.9) that

Assumption 6.1.4 $p = 2$, $\mathcal{X} = \mathbb{R}^q$, d = Euclidean metric, λ = Lebesgue measure.

Let us consider $m_1, \ldots, m_k \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$, weights $\lambda_1, \ldots, \lambda_k \in \mathbb{R}^+$ with $\sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} = 1$ and the respective discrete measure[1] $\Pi^{(k)} = \sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} \delta_{m_\mathrm{i}}$. Given some measure $m \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$, we denote the optimal transport map from m to m_i as $T_m^{m_\mathrm{i}}$ for $\mathrm{i} = 1, \ldots, k$. The uniqueness and existence of this map is guaranteed by Brenier's Theorem. With this notation one can define the operator $G_k : \mathcal{W}_{2,ac}(\mathbb{R}^q) \to \mathcal{W}_{2,ac}(\mathbb{R}^q)$ as

$$G_k(m) := \left(\sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} T_m^{m_\mathrm{i}} \right)(m). \tag{6.1}$$

Owing to [6] the operator G_k is continuous for the W_2 distance. Also, if at least one of the m_i has a bounded density, then the unique Wasserstein barycenter $\hat{m}$ of $\Pi^{(k)}$ has a bounded density and satisfies $G_k(\hat{m}) = \hat{m}$. Thanks to this, starting from $\mu_0 \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$ one can define the sequence

$$\mu_{n+1} := G_k(\mu_n), \text{ for } n \geq 0. \tag{6.2}$$

The next result was proven by Álvarez-Esteban, Barrio, Cuesta-Albertos, Matrán in [6, Theorem 3.6] and independently by Zemel and Panaretos in [93, Theorem 3,Corollary 2]:

Proposition 6.1.5 The sequence $\{\mu_n\}_{n\geq 0}$ defined in (eq. 6.2) is tight and every weakly convergent subsequence of $\{\mu_n\}_{n\geq 0}$ must converge in W_2 distance to a measure in $\mathcal{W}_{2,ac}(\mathbb{R}^q)$ which is a fixed point of G_k. If some m_i has a bounded density, and if G_k has a unique fixed point $\hat{m}$, then $\hat{m}$ is the Wasserstein barycenter of $\Pi^{(k)}$ and we have that $W_2(\mu_n, \hat{m}) \to 0$.

The previous result allows one to estimate the barycenter of any discrete measure (i.e. any prior/posterior with finite support), as long as one is able to construct the optimal transports $T_m^{m_\mathrm{i}}$. Thanks to the *Riemannian*-like geometry of $\mathcal{W}_2(\mathbb{R}^q)$ (see [8, Chapter 8]) one can reinterpret the iterations in (eq. 6.2) as a gradient descent step. This was discovered by Panaretos and Zemel in [93, 92]. In fact, in [93, Theorem 1] the authors prove the following: Letting $\Pi^{(k)} = \sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} \delta_{m_\mathrm{i}}$ as above, then the (half) Wasserstein Bayes risk of $m \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$ and its Fréchet derivative are given respectively by

$$F_k(m) := \tfrac{1}{2} \sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} W_2^2(m_\mathrm{i}, m), \tag{6.3}$$

$$F_k'(m) = -\sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} (T_m^{m_\mathrm{i}} - I) = I - \sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} T_m^{m_\mathrm{i}}, \tag{6.4}$$

where I is the identity map on $\mathbb{R}^q$. It follows by Brenier's theorem [138, Theorem 2.12(ii)] that $\hat{m}$ is a fixed point of G_k defined in (eq. 6.1) if and only if $F_k'(\hat{m}) = 0$ (one says that $\hat{m}$ is a *Karcher mean* of $\Pi^{(k)}$). The gradient descent sequence with step γ starting from $\mu_0 \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$ is defined as

$$\mu_{n+1} := G_{k,\gamma}(\mu_n), \text{ for } n \geq 0, \tag{6.5}$$

where $G_{k,\gamma}(m) := \left[I + \gamma F_k'(m) \right](m) = \left[(1-\gamma) I + \gamma \sum_{\mathrm{i}=1}^k \lambda_\mathrm{i} T_m^{m_\mathrm{i}} \right](m)$. These ideas serve us as inspiration for the stochastic gradient descent iteration in the next part. We finally remark that if $\gamma = 1$ the aforementioned gradient descent sequence equals the sequence in (eq. 6.2), i.e. $G_{k,1} = G_k$. In fact in [93] the authors prove that the choice $\gamma = 1$ is optimal.

[1] One can think of $\Pi^{(k)}$ as an empirical approximation of the posterior Π_n or the prior Π.

6.2 Stochastic Gradient Descent on Wasserstein Space

The method in Section 6.1.1 works perfectly well for calculating the empirical barycenter. For the estimation of a population barycenter (i.e. when the prior does not have finite support) we would need to construct a convergent sequence of empirical barycenters, as in Section 6.1, and then apply the method in Section 6.1.1. Altogether this can be computationally expensive. To remedy this, we follow the ideas in [23] and define a *stochastic* version of the gradient descent sequence for the barycenter of $\Pi \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathbb{R}^q))$. Needless to say that Π could represent the posterior or prior distribution.

Definition 6.2.1 Let $\mu_0 \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$, $m_k \overset{\text{iid}}{\sim} \Pi$, and $\gamma_k > 0$ for $k \geq 0$. Then we define the stochastic gradient descent (SGD) sequence as

$$\mu_{k+1} := \left[(1-\gamma_k)I + \gamma_k T_{\mu_k}^{m_k}\right](\mu_k) \text{ , for } k \geq 0. \tag{6.6}$$

By Remark 5.3.9 and an induction argument, we clearly have

$$\{\mu_k\}_k \subset \mathcal{W}_{2,ac}(\mathbb{R}^q), \ a.s. \tag{6.7}$$

The key ingredients for the convergence analysis of the stochastic gradient iterations are:

$$F(\mu) :=\tfrac{1}{2}\textstyle\int_{\mathcal{W}_{2,ac}(\mathcal{X})} W_2^2(\mu, m)\Pi(\mathrm{d}m) \tag{6.8}$$

$$F'(\mu) := -\textstyle\int_{\mathcal{W}_{2,ac}(\mathcal{X})}(T_\mu^m - I))\Pi(\mathrm{d}m). \tag{6.9}$$

Observe that the population barycenter $\hat{\mu}$ is the minimizer of F and that by Lemma 5.3.10 also $\|F'(\hat{\mu})\|_{L^2(\hat{\mu})} = 0$. The next proposition (cf. [93, Lemma 2]) indicates us that, in expectation, the sequence $\{F(\mu_k)\}_k$ is essentially decreasing for a sufficiently small step γ_k. This is a first indication of the behaviour of the sequence $\{\mu_k\}_k$. We denote by $\{\mathcal{F}_k\}_k$ the filtration of the i.i.d. sample $m_k \sim \Pi$, namely $\mathcal{F}_{-1}$ is the trivial sigma-algebra and $\mathcal{F}_{k+1}$ is the sigma-algebra generated by $m_0, \ldots, m_k$. In this way μ_k is $\mathcal{F}_k$-measurable.

Proposition 6.2.2 For the stochastic gradient descent sequence in (6.6), we have

$$\mathbb{E}\left[F(\mu_{k+1}) - F(\mu_k)|\mathcal{F}_k\right] \leq \gamma_k^2 F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2. \tag{6.10}$$

If we set $\gamma_0 = \gamma \frac{\|F'(\mu_0)\|_{L^2(\mu_0)}^2}{F(\mu_0)}$, then for $k = 0$ the inequality (6.10) becomes

$$\mathbb{E}\left[F(\mu_1) - F(\mu_0)\right] \leq -\frac{\|F'(\mu_0)\|_{L^2(\mu_0)}^4}{F(\mu_0)}\left(\gamma - \gamma^2\right)$$

which is optimal with $\gamma = \frac{1}{2}$.

Proof of Proposition 6.2.2. Let $\nu \in \mathrm{supp}(\Pi)$. By (6.7) we know that

$$\left(\left[(1-\gamma_k)I + \gamma_k T_{\mu_k}^{m_k}\right] \ , \ T_{\mu_k}^{\nu}\right)(\mu_k),$$

is a feasible (not necessarily optimal) coupling with first and second marginals μ_{k+1} and ν respectively. Denoting $O_m := T^m_{\mu_k} - I$, we have

$$\begin{aligned} W_2^2(\mu_{k+1},\nu) &\leq \|(1-\gamma_k)I + \gamma_k T^{m_k}_{\mu_k} - T^{\nu}_{\mu_k}\|^2_{L^2(\mu_k)} \\ &= \|-O_\nu + \gamma_k O_{m_k}\|^2_{L^2(\mu_k)} \\ &= \|O_\nu\|^2_{L^2(\mu_k)} - 2\gamma_k \langle O_\nu, O_{m_k}\rangle_{L^2(\mu_k)} + \gamma_k^2 \|O_{m_k}\|^2_{L^2(\mu_k)}. \end{aligned}$$

Evaluating μ_{k+1} on the functional F and thanks to the previous inequality, we have

$$\begin{aligned} F(\mu_{k+1}) =& \tfrac{1}{2}\textstyle\int W_2^2(\mu_{k+1},\nu)\Pi(\mathrm{d}\nu) \\ \leq& \tfrac{1}{2}\textstyle\int \|O_\nu\|^2_{L^2(\mu_k)}\Pi(\mathrm{d}\nu) - \gamma_k \left\langle \int O_\nu \Pi(\mathrm{d}\nu), O_{m_k}\right\rangle_{L^2(\mu_k)} + \frac{\gamma_k^2}{2}\|O_{m_k}\|^2_{L^2(\mu_k)} \\ =& F(\mu_k) + \gamma_k \left\langle F'(\mu_k), O_{m_k}\right\rangle_{L^2(\mu_k)} + \frac{\gamma_k^2}{2}\|O_{m_k}\|^2_{L^2(\mu_k)}. \end{aligned}$$

Taking conditional expectation with respect to $\mathcal{F}_k$, and as m_k is independently sampled from this sigma-algebra, we conclude

$$\begin{aligned} &\mathbb{E}\left[F(\mu_{k+1})|\mathcal{F}_k\right] \\ \leq& F(\mu_k) + \gamma_k \left\langle F'(\mu_k), \textstyle\int O_m \Pi(\mathrm{d}m)\right\rangle_{L^2(\mu_k)} + \frac{\gamma_k^2}{2}\textstyle\int \|O_m\|^2_{L^2(\mu_k)}\Pi(\mathrm{d}m) \\ =& (1+\gamma_k^2)F(\mu_k) - \gamma_k \|F'(\mu_k)\|^2_{L^2(\mu_k)}. \end{aligned}$$

□

Now we show that under reasonable assumptions the sequence $\{F(\mu_k)\}_k$ converges a.s. to the unique minimizer of F. As mentioned above, this minimiser is the 2-Wasserstein population barycenter of Π, denoted $\hat{\mu}$. We will need the following convergence result recalled in [21]:

Theorem 6.2.3 (Quasi-martingale convergence theorem) Given a random sequence $\{h_t\}_{t\geq 0}$ adapted to the filtration $\{\mathcal{F}_t\}$, define $\delta_t := 1$ if $\mathbb{E}\left[h_{t+1} - h_t|\mathcal{F}_t\right] > 0$ and $\delta_t := 0$ otherwise. If $h_t \geq 0$ for all $t \geq 0$, and the infinite sum of the positive expected variations is finite, i.e. $\sum_{t=1}^{\infty} \mathbb{E}\left[\delta_t(h_{t+1} - h_t)\right] < \infty$, then the sequence $\{h_t\}$ converges almost surely to some $h_\infty \geq 0$.

We will assume the following conditions on the steps γ_t appearing in eq. (6.6):

$$\textstyle\sum_{t=1}^{\infty} \gamma_t^2 < \infty \tag{6.11}$$

$$\textstyle\sum_{t=1}^{\infty} \gamma_t = \infty. \tag{6.12}$$

For example, the above conditions are satisfy straightforward with $\gamma_t = 1/t$. Additionally the following conditions will be useful to finish the arguments:

$$\mathcal{W}_{2,ac}(\mathcal{X}) \ni \mu \mapsto \|F'(\mu)\|^2_{L^2(\mu)} \text{ is lower semicontinuous w.r.t. } W_q \text{ some } q < 2, \tag{6.13}$$

$$\mathcal{W}_{2,ac}(\mathcal{X}) \ni \mu \mapsto \|F'(\mu)\|^2_{L^2(\mu)} \text{ has a unique zero.} \tag{6.14}$$

We examine these conditions in Remark 6.2.5. We can state the main result of this part:

Theorem 6.2.4 Under conditions (6.11) and (6.12) the stochastic gradient descent sequence $\{\mu_t\}_t$ is a.s. relatively compact in $\mathcal{W}_q$ for all $q < 2$ (in particular it is tight). If furthermore (6.13) and (6.14) hold, then a.s. $\{\mu_t\}_{t\geq 0}$ converges to the $\mathcal{W}_2$-population barycenter $\hat{\mu}$ of Π in the $\mathcal{W}_q$ topology (in particular it weakly converges).

PROOF. Denote $\hat{F} := F(\hat{\mu})$ and introduce the sequences

$$h_t := F(\mu_t) - \hat{F}, \qquad \alpha_t := \textstyle\prod_{\mathrm{i}=1}^{t-1} \frac{1}{1+\gamma_{\mathrm{i}}^2}.$$

Observe that $h_t \geq 0$ for all t. Thanks to condition (6.11) the sequence α_t converges to some $\alpha_\infty > 0$, as can be checked by taking logarithm. By Proposition 6.2.2 we have

$$\mathbb{E}\left[h_{t+1} - (1+\gamma_t^2)h_t|\mathcal{F}_t\right] \leq \gamma_t^2\hat{F} - \gamma_t\|F'(\mu_t)\|^2_{L^2(\mu_t)} \leq \gamma_t^2\hat{F}, \tag{6.15}$$

so after multiplying by α_{t+1} we derive the bound

$$\mathbb{E}\left[\alpha_{t+1}h_{t+1} - \alpha_t h_t|\mathcal{F}_t\right] \leq \alpha_{t+1}\gamma_t^2\hat{F} - \alpha_{t+1}\gamma_t\|F'(\mu_t)\|^2_{L^2(\mu_t)} \leq \alpha_{t+1}\gamma_t^2\hat{F}. \tag{6.16}$$

We define $\delta_t := 1$ if $\mathbb{E}\left[\alpha_{t+1}h_{t+1} - \alpha_t h_t|\mathcal{F}_t\right] > 0$ and $\delta_t := 0$ otherwise. Then

$$\begin{aligned}\textstyle\sum_{t=1}^\infty \mathbb{E}\left[\delta_t(\alpha_{t+1}h_{t+1} - \alpha_t h_t)\right] &= \textstyle\sum_{t=1}^\infty \mathbb{E}\left[\delta_t \mathbb{E}\left[\alpha_{t+1}h_{t+1} - \alpha_t h_t|\mathcal{F}_t\right]\right] \\ &\leq \hat{F}\textstyle\sum_{t=1}^\infty \alpha_{t+1}\gamma_t^2 \leq \hat{F}\sum_{t=1}^\infty \gamma_t^2 < \infty.\end{aligned}$$

Since $\alpha_t h_t \geq 0$, by quasi-martingale convergence theorem $\{\alpha_t h_t\}_t$ is a.s. convergent, but as α_t converges to $\alpha_\infty > 0$, then h_t also converges almost surely to some $h_\infty \geq 0$. Taking expectations is (6.16), summing in t so that a telescopic sum forms, we have

$$\mathbb{E}[\alpha_{t+1}h_{t+1}] \leq \alpha_0 h_0 + \hat{F}\textstyle\sum_{s=1}^t \alpha_{s+1}\gamma_s^2 \leq C.$$

Taking limit inferior, applying Fatou's lemma, and since $\alpha_\infty > 0$, we conclude $\mathbb{E}[h_\infty] < \infty$. In particular h_∞ is a.s. finite. This means that $F(\mu_t)$ has a finite a.s. limit, which we call L. By convexity of transport costs ([139, Theorem 4.8]) we have

$$\tfrac{1}{2}W_2^2\left(\mu_t, \textstyle\int m\Pi(\mathrm{d}m)\right) \leq F(\mu_t) \leq L + 1,$$

for t eventually large enough. Since $\Pi \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^q))$ we have $\int m\Pi(\mathrm{d}m) \in \mathcal{W}_2(\mathbb{R}^q)$, so the second moments of $\{\mu_t\}_t$ are a.s. bounded by a finite (random) constant M. By Markov's inequality the sequence $\{\mu_t\}_t$ is a.s. tight, since closed balls in $\mathbb{R}^q$ are compact. Further, for $q < 2$, by Hölder and Chebyshev inequalities we have that

$$\textstyle\int_{\|x\|>R}\|x\|^q\mathrm{d}\mu_t \leq \frac{1}{R^{1-q/2}}\int\|x\|^2\mathrm{d}\mu_t \leq \frac{M}{R^{1-q/2}},$$

so $\{\mu_t\}_{t\geq 0}$ is a.s. relatively compact in $\mathcal{W}_q$ thanks to [138, Theorem 7.12] and

$$\textstyle\lim_{R\to\infty} \limsup_{t\to\infty} \int_{\|x\|>R}\|x\|^q\mathrm{d}\mu_t \leq \lim_{R\to\infty} \limsup_{t\to\infty} \frac{M}{R^{1-q/2}} = 0.$$

Back to (6.16), taking expectations, summing in t to obtain a telescopic sum, we get

$$\mathbb{E}[\alpha_{t+1}h_{t+1}] - h_0\alpha_0 \leq \hat{F}\textstyle\sum_{s=1}^t \alpha_{s+1}\gamma_s^2 - \sum_{s=1}^t \alpha_{s+1}\gamma_s\|F'(\mu_s)\|^2_{L^2(\mu_s)}.$$

Taking liminf, by Fatou on the l.h.s. and monotone convergence on the r.h.s. we get

$$-\infty < \mathbb{E}[\alpha_\infty h_\infty] - h_0\alpha_0 \leq C - \mathbb{E}\left[\textstyle\sum_{s=1}^\infty \alpha_{s+1}\gamma_s\|F'(\mu_s)\|^2_{L^2(\mu_s)}\right].$$

In particular, we have

$$\textstyle\sum_{t=1}^\infty \gamma_t\|F'(\mu_t)\|^2_{L^2(\mu_t)} < \infty, \text{ a.s.} \tag{6.17}$$

Observe that $\liminf \|F'(\mu_t)\|^2_{L^2(\mu_t)} > 0$ would be at odds with (6.17) and (6.12), so further

$$\liminf \|F'(\mu_t)\|^2_{L^2(\mu_t)} = 0, \text{ a.s.}$$

Assume conditions (6.13) and (6.14). If a subsequence of $\{\mu_t\}_t$ W_q-converges to some $\mu \neq \hat{\mu}$, then along this subsequence we have $\liminf \|F'(\mu_t)\|^2_{L^2(\mu_t)} > 0$: indeed, otherwise by (6.13) we would get $\|F'(\mu)\|^2_{L^2(\mu)} = 0$, contradicting (6.14) since already $\|F'(\hat{\mu})\|^2_{L^2(\hat{\mu})} = 0$. Since we do know that $\liminf \|F'(\mu_t)\|^2_{L^2(\mu_t)} = 0$ a.s., it follows that realizations where $\{\mu_t\}_t$ accumulates into a limit different than $\hat{\mu}$ have zero measure. Thus a.s. the only possible accumulation point of $\{\mu_t\}_t$ is $\hat{\mu}$. In particular, by a.s. relative compactness of $\{\mu_t\}_t$, this sequence must W_q-converge a.s. to the population barycenter $\hat{\mu}$, concluding the proof. □

Remark 6.2.5. The validity of eq. (6.14) is equivalent to the uniqueness of an (absolutely continuous) fixed point for the functional

$$\bar{m} \mapsto \left(\int T^m_{\bar{m}} \Pi(\mathrm{d}m)\right)(\bar{m}), \tag{6.18}$$

which is in general unsettled. In the finite-support case [2, Remark 3.9] and specially [93, Theorem 2] provide reasonable sufficient conditions. For the infinite-support case the uniqueness of fixed-points, as far as we know, has only been explored in [18, Theorem 5.1] under strong assumptions. It is imaginable that the arguments in [93] can be generalized to the infinite-support case, but we do not explore this in the present work.

On the other hand it seems plausible that (6.13) holds in full generality. In this direction we refer to [93, Proposition 3] for a continuity statement when, again, Π has finite support. We give next a sufficient/alternative condition for (6.13) of our own, which does work for the infinite-support case.

Proposition 6.2.6 Assumption (6.13) is fulfilled if

(i) $\mathcal{X} = \mathbb{R}$.

Alternatively, assume that

(ii) $\mu_0 \in \mathrm{supp}(\Pi) \subset \mathcal{H} \subset \mathcal{W}_{2,ac}(\mathbb{R}^q)$, where $\mathcal{H}$ is geodesically closed and closed under composition of optimal maps, meaning respectively[2]

$$\forall m, \tilde{m} \in \mathcal{H}, \forall \alpha \in [0,1] : \ ([1-\alpha]I + \alpha T^{\tilde{m}}_m)(m) \in \mathcal{H}, \tag{6.19}$$

$$\forall \mu, m, \tilde{m} \in \mathcal{H} : \ T^{\bar{m}}_m = T^{\bar{m}}_\mu \circ \left(T^m_\mu\right)^{-1}. \tag{6.20}$$

Then for the stochastic gradient descent sequence we have a.s. $\{\mu_k\}_{k\geq 1} \subset \mathcal{H}$. Further the functional $\mathcal{H} \ni \mu \mapsto \|F'(\mu)\|^2_{L^2(\mu)}$ is W_2-continuous and weakly lower semicontinuous, and the conclusions of Theorem 6.2.4 remain valid if Condition (6.13) is dropped.

Proof of Proposition 6.2.6. We first settle the case of Condition (ii). It is immediate from (6.19) that $\mu_1 \in \mathcal{H}$, and by induction it follows similarly that a.s. $\{\mu_k\}_{k\geq 1} \subset \mathcal{H}$. We now

[2]Since μ, m are absolutely continuous, we have by [138, Theorem 2.12(iv)] $\left(T^m_\mu\right)^{-1} = T^\mu_m, (m-a.s.)$

establish the continuity statement, decomposing the functional as follows

$$\begin{aligned}\|F'(\mu)\|^2_{L^2(\mu)} &= \int_{\mathcal{X}} \left|\int T^m_\mu(y)\Pi(\mathrm{d}m) - y\right|^2 \mu(\mathrm{d}y) \\ &= \int_{\mathcal{X}} \left|\int T^m_\mu(y)\Pi(\mathrm{d}m)\right|^2 \mu(\mathrm{d}y) - 2\int\int_{\mathcal{X}} y \cdot T^m_\mu(y)\mu(\mathrm{d}y)\Pi(\mathrm{d}m) + \int_{\mathcal{X}}\|y\|^2\mu(\mathrm{d}y).\end{aligned}$$

The term $\mu \mapsto \int_{\mathcal{X}}\|y\|^2\mu(\mathrm{d}y)$ is continuous in $\mathcal{W}_2$ and weakly lower semicontinuous. As Brenier maps are optimal, we have

$$\int_{\mathcal{X}} y \cdot T^m_\mu(y)\mu(\mathrm{d}y) = \sup_{y\sim\mu,\, z\sim m} \mathbb{E}\left[y \cdot z\right] := \rho(\mu, m).$$

Thus $\rho(\cdot, m)$ is continuous in $\mathcal{W}_2$ and weakly upper semicontinuous, so under the standing assumption that $\Pi \in \mathcal{W}_2(\mathcal{W}_{2,ac})$ the term $\int \rho(\mu, m)\Pi(\mathrm{d}m)$ is continuous in $\mathcal{W}_2$ and weakly upper semicontinuous too. Finally we only have to check that the first term is continuous:

$$\begin{aligned}\int_{\mathcal{X}} \left|\int T^m_\mu(y)\Pi(\mathrm{d}m)\right|^2 \mu(\mathrm{d}y) &= \int_{\mathcal{X}} \left[\int T^m_\mu(y)\Pi(\mathrm{d}m)\right] \cdot \left[\int T^{\tilde{m}}_\mu(y)\Pi(\mathrm{d}\tilde{m})\right] \mu(\mathrm{d}y) \\ &= \int\int \left[\int_{\mathcal{X}} T^m_\mu(y) \cdot T^{\tilde{m}}_\mu(y)\mu(\mathrm{d}y)\right] \Pi(\mathrm{d}\tilde{m})\Pi(\mathrm{d}m) \\ &= \int\int G(\mu, m, \tilde{m})\Pi(\mathrm{d}\tilde{m})\Pi(\mathrm{d}m)\end{aligned}$$

where $G(\mu, m, \tilde{m}) = \int_{\mathcal{X}} T^m_\mu(y) \cdot T^{\tilde{m}}_\mu(y)\mu(\mathrm{d}y)$. For $\mu, m, \tilde{m} \in \mathcal{H}$ we have that

$$\begin{aligned}G(\mu, m, \tilde{m}) &= \int_{\mathcal{X}} T^m_\mu(y) \cdot T^{\tilde{m}}_\mu(y)\mu(\mathrm{d}y) \\ &= \int_{\mathcal{X}} T^m_\mu(y) \cdot \left[T^{\tilde{m}}_\mu \circ \left(T^m_\mu\right)^{-1} \circ T^m_\mu(y)\right] \mu(\mathrm{d}y) \\ &= \int_{\mathcal{X}} z \cdot \left[T^{\tilde{m}}_\mu \circ \left(T^m_\mu\right)^{-1}(z)\right] m(\mathrm{d}z) \\ &= \int_{\mathcal{X}} z \cdot T^{\tilde{m}}_m(z) m(\mathrm{d}z),\end{aligned}$$

thanks to the Condition (6.20). Since $G(\mu, m, \tilde{m})$ is independent of μ, we conclude that the functional $\mu \mapsto \|F'(\mu)\|^2_{L^2(\mu)}$ is $\mathcal{W}_2$-continuous and weakly lower semicontinuous on $\mathcal{H}$ as desired. With this at hand we can go back to the arguments in the proof of Theorem 6.2.4, checking their validity without Condition (6.13). Finally let us consider Condition (i). In this case (6.20) is true for all $\mu, m, \tilde{m}$ absolutely continuous, since the composition of increasing functions on the line is increasing. The above arguments verbatim prove the validity of (6.13). □

See [20, Proposition 4.1] for examples where eq. (6.20) is fulfilled, including the case of radial or component-wise transformations of a base measure. Eq. (6.20) is rather restrictive, since the composition of gradients of convex functions need not be of the same kind.

6.2.1 Batch Stochastic gradient descent on Wasserstein space

To generate the sequence (6.6) in the k-step, we sampled $m_k \overset{\text{iid}}{\sim} \Pi$, chose $\gamma_k > 0$ and updated μ_k via the map $T_k := I + \gamma_k(T^{m_k}_{\mu_k} - I)$. The expected transport map is

$$\mathbb{E}\left[T_k\right] = I + \gamma_k \int (T^{m_k}_{\mu_k} - I)\Pi(\mathrm{d}m_k) = I - \gamma_k F'(\mu_k).$$

Notice that $-(T_\mu^{m_k} - I)$ is an unbiased estimator for $F'(\mu)$, but in many cases it can have a high variance so the learning rates γ must be very small for convergence. This motivates us to propose alternative estimators for $F'(\mu)$ with less variance:

Definition 6.2.7 Let $\mu_0 \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$, $m_k^{\mathrm{i}} \overset{\text{iid}}{\sim} \Pi$, and $\gamma_k > 0$ for $k \geq 0$ and $\mathrm{i} = 1, \ldots, S_k$. The batch stochastic gradient descent (BSGD) sequence is given by

$$\mu_{k+1} := \left[(1-\gamma_k)I + \gamma_k \tfrac{1}{S_k} \textstyle\sum_{\mathrm{i}=1}^{S_k} T_{\mu_k}^{m_k^{\mathrm{i}}}\right](\mu_k). \tag{6.21}$$

Denote this time $\mathcal{F}_{k+1}$ the sigma-algebra generated by $\{m_\ell^{\mathrm{i}} : \ell \leq k, \mathrm{i} \leq S_k\}$. Notice that $D := \frac{1}{S_k} \sum_{\mathrm{i}=1}^{S_k} T_{\mu_k}^{m_k^{\mathrm{i}}} - I$ is an unbiased estimator of $-F'(\mu_k)$. Then, much as in Proposition 6.2.2, we have

$$\begin{aligned}
\mathbb{E}\left[F(\mu_{k+1})|\mathcal{F}_k\right] =& F(\mu_k) + \gamma_k \langle F'(\mu_k), \textstyle\int D\, \Pi(\mathrm{d}m_k^1 \cdots \mathrm{d}m_k^{S_k})\rangle_{L^2(\mu_k)} \\
&+ \frac{\gamma_k^2}{2} \int \|D\|^2_{L^2(\mu_k)} \Pi(\mathrm{d}m_k^1 \cdots \mathrm{d}m_k^{S_k}) \\
=& F(\mu_k) - \gamma_k \|F'(\mu_k)\|^2_{L^2(\mu_k)} + \tfrac{\gamma_k^2}{2} \textstyle\int \|\frac{1}{S_k} \sum_{\mathrm{i}=1}^{S_k} T_{\mu_k}^{m_k^{\mathrm{i}}} - I\|^2_{L^2(\mu_k)} \Pi(\mathrm{d}m_k^1 \cdots \mathrm{d}m_k^{S_k}) \\
\leq& F(\mu_k) - \gamma_k \|F'(\mu_k)\|^2_{L^2(\mu_k)} + \tfrac{\gamma_k^2}{2} \tfrac{1}{S_k} \textstyle\sum_{\mathrm{i}=1}^{S_k} \int \|T_{\mu_k}^{m_k^{\mathrm{i}}} - I\|^2_{L^2(\mu_k)} \Pi(\mathrm{d}m_k^{\mathrm{i}}) \\
=& (1+\gamma_k^2) F(\mu_k) - \gamma_k \|F'(\mu_k)\|^2_{L^2(\mu_k)}.
\end{aligned}$$

From here on it is routine to follow the arguments in the proof of Theorem 6.2.4, obtaining the following result, whose proof we omit:

Proposition 6.2.8 Under conditions (6.11) and (6.12) the BSGD sequence $\{\mu_t\}_{t\geq 0}$ defined in (6.21) is a.s. relatively compact in $\mathcal{W}_q$ for all $q < 2$. If also (6.13) and (6.14) hold, then a.s. $\{\mu_t\}_{t\geq 0}$ converges to the $\mathcal{W}_2$-population barycenter $\hat{\mu}$ of Π in the $\mathcal{W}_q$-topology.

The main idea of using mini-batch is *noise reduction* for the estimator of $F'(\mu)$.

Proposition 6.2.9 The variance of the mini batch estimator for $F'(\mu)$, given namely by $-\frac{1}{S}\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_{\mathrm{i}}} - I)$, decreases linearly in the sample size, ie.

$$\mathbb{V}[-\tfrac{1}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_{\mathrm{i}}} - I)] = \mathcal{O}(\tfrac{1}{S}).$$

Proof of Proposition 6.2.9. The variance of the estimator where $m \sim \Pi$ is

$$\begin{aligned}
\mathbb{V}[-(T_\mu^m - I)] &= \mathbb{E}\left[\|-(T_\mu^m - I)\|^2_{L^2(\mu)}\right] - \left\|\mathbb{E}\left[-(T_\mu^m - I)\right]\right\|^2_{L^2(\mu)} \\
&= \mathbb{E}\left[W_2^2(\mu, m)\right] - \|F'(\mu)\|^2_{L^2(\mu)} = 2F(\mu) - \|F'(\mu)\|^2_{L^2(\mu)}.
\end{aligned}$$

On the other hand, the variance of the mini-batch estimator where $m_{\mathrm{i}} \sim \Pi$ for $\mathrm{i} \leq S$ is

$$\begin{aligned}
&\mathbb{V}\left[-\tfrac{1}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_{\mathrm{i}}} - I)\right] \\
=&\mathbb{E}\left[\left\|-\tfrac{1}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_{\mathrm{i}}} - I)\right\|^2_{L^2(\mu)}\right] - \left\|\mathbb{E}\left[-\tfrac{1}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_{\mathrm{i}}} - I)\right]\right\|^2_{L^2(\mu)}
\end{aligned}$$

$$= \mathbb{E}\left[\left\| -\tfrac{1}{S}\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_\mathrm{i}} - I)\right\|^2_{L^2(\mu)}\right] - \|F'(\mu)\|^2_{L^2(\mu)} \ .$$

For the first term we can expand it as

$$\begin{aligned}
&\left\| -\tfrac{1}{S}\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_\mathrm{i}} - I)\right\|^2_{L^2(\mu)} \\
=&\tfrac{1}{S^2}\langle \textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_\mathrm{i}} - I), \sum_{j=1}^{S}(T_\mu^{m_j} - I)\rangle_{L^2(\mu)} \\
=&\tfrac{1}{S^2}\textstyle\sum_{\mathrm{i}=1}^{S}\sum_{j=1}^{S}\langle T_\mu^{m_\mathrm{i}} - I, T_\mu^{m_j} - I\rangle_{L^2(\mu)} \\
=&\tfrac{1}{S^2}\textstyle\sum_{\mathrm{i}=1}^{S}\|-(T_\mu^{m_\mathrm{i}} - I)\|^2_{L^2(\mu)} + \tfrac{1}{S^2}\sum_{j\neq \mathrm{i}}^{S}\langle T_\mu^{m_\mathrm{i}} - I, T_\mu^{m_j} - I\rangle_{L^2(\mu)},
\end{aligned}$$

so if we take expectation, as the samples $m_\mathrm{i} \sim \Pi$ are independent, we have

$$\begin{aligned}
&\mathbb{E}\left[\|-\tfrac{1}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_\mathrm{i}} - I)\|^2_{L^2(\mu)}\right] \\
=&\tfrac{1}{S^2}\textstyle\sum_{\mathrm{i}=1}^{S}\mathbb{E}\left[W_2^2(\mu, m_\mathrm{i})\right] + \tfrac{1}{S^2}\sum_{j\neq \mathrm{i}}^{S}\langle \mathbb{E}\left[T_\mu^{m_\mathrm{i}} - I\right], \mathbb{E}\left[T_\mu^{m_j} - I\right]\rangle_{L^2(\mu)} \\
=&\tfrac{2}{S^2}\textstyle\sum_{\mathrm{i}=1}^{S}F(\mu) + \tfrac{1}{S^2}\sum_{j\neq \mathrm{i}}^{S}\langle F'(\mu), F'(\mu)\rangle_{L^2(\mu)} = \tfrac{2}{S}F(\mu) + \tfrac{S-1}{S}\|F'(\mu)\|^2_{L^2(\mu)}.
\end{aligned}$$

Finally the variance of the mini-bath estimator is given by

$$\mathbb{V}\left[-\tfrac{1}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(T_\mu^{m_\mathrm{i}} - I)\right] = \tfrac{1}{S}\left[2F(\mu) - \|F'(\mu)\|^2_{L^2(\mu)}\right].$$

□

6.3 On Closed-Form Wasserstein Barycenters

In Chapter 6 we presented some methods to compute the population Wasserstein barycenter, which assume that we are capable of getting samples from the distributions Π and Π_n, and that we can calculate the optimal transports between measures. While sampling is solved by MCMC techniques [56] or transport-based sampling procedures [41, 94, 64, 78], computing optimal transports is not achievable in a general way. For this reason, we exhibit in this section some families of distributions for which it is possible to calculate these optimal transports [30]. Furthermore, we will examine their barycenter, establishing some properties which are conserved under the operation of taking barycenter.

6.3.1 Univariate distributions

For a continuous distribution m in $\mathbb{R}$ we denote its cumulative distribution function by $F_m(x)$ and its right-continuous quantile function by $Q_m(\cdot) = F_m^{-1}(\cdot)$. The p-Wasserstein optimal transport map from some continuous m_0 to m is independent of p and given by the monotone rearrangement (see [138, Remark 2.19(iv)]):

$$T_0^m(x) = Q_m(F_{m_0}(x)).$$

Note that this class of functions is closed under composition, convex combination, and contains the identity. Given Π the barycenter $\hat{m}$ is also independent of p and characterized by the *averaged quantile function*, i.e.

$$Q_{\hat{m}}(\cdot) = \int Q_m(\cdot)\Pi(\mathrm{d}m).$$

A stochastic gradient descent iteration, starting from a distribution function $F_\mu(x)$, sampling some $m \sim \Pi$, and with step γ, produces the measure

$$\nu = ((1-\gamma)I + \gamma T_\mu^m)(\mu),$$

which is characterized by its quantile function

$$Q_\nu(\cdot) = (1-\gamma)Q_\mu(\cdot) + \gamma Q_m(\cdot).$$

The batch stochastic gradient descent iteration is given by $Q_\nu(\cdot) = (1-\gamma)Q_\mu(\cdot) + \frac{\gamma}{S}\sum_{\mathrm{i}=1}^{S} Q_{m^{\mathrm{i}}}(\cdot)$.

Interestingly the model average $\bar{m}$ is characterized by the *averaged cumulative distribution function*, i.e. $F_{\bar{m}}(\cdot) = \int F_m(\cdot)\Pi(\mathrm{d}m)$. As we mentioned earlier, the model average does not preserve intrinsic *shape* properties from the distributions such as symmetry or unimodality. For example if $\Pi = 0.3 * \delta_{m_1} + 0.7 * \delta_{m_2}$ with $m_1 = \mathcal{N}(1,1)$ and $m_2 = \mathcal{N}(3,1)$, the *model average* is an asymmetric bimodal distribution with modes on 1 and 3, while the Wasserstein barycenter is the Gaussian distribution $\hat{m} = \mathcal{N}(2,1)$. The following reasoning formalises the fact that Wasserstein barycenters preserve some *geometric properties*.

A continuous distribution m on $\mathbb{R}$ is called unimodal with a mode on $\tilde{x} \in \mathbb{R}$ if its cumulative distribution function $F(x)$ is convex for $x < \tilde{x}$ and concave for $x > \tilde{x}$. One says that m is symmetric around $x_m \in \mathbb{R}$ if $F(x_m + x) = 1 - F(x_m - x)$ for $x \in \mathbb{R}$. One can also characterize unimodality and symmetry by quantile function. A continuous distribution m on $\mathbb{R}$ is unimodal with a mode on $\tilde{x}$ if its quantile function $Q(y)$ is concave for $y < \tilde{y}$ and convex for $y > \tilde{y}$, where $Q(\tilde{y}) = \tilde{x}$. Likewise, m is symmetric around $x_m \in \mathbb{R}$ if $Q(\frac{1}{2} + y) = 2x_m - Q(\frac{1}{2} - y)$ for $y \in [0, \frac{1}{2}]$. Thanks to this characterization we conclude that the barycenter preserves unimodality/symmetry:

Proposition 6.3.1 If $\Pi \in \mathcal{W}_p(\mathcal{P}_{ac}(\mathbb{R}))$ is concentrated on symmetric (resp. symmetric unimodal) univariate distributions, then the barycenter $\hat{m}$ is symmetric (resp. symmetric unimodal).

Proof of Proposition 6.3.1. Using the quantile function characterization, we have that

$$Q_{\hat{m}}\left(\tfrac{1}{2} + y\right) = \int Q_m\left(\tfrac{1}{2} + y\right)\Pi(\mathrm{d}m) = 2x_{\hat{m}} - Q_{\hat{m}}\left(\tfrac{1}{2} - y\right),$$

where $x_{\hat{m}} := \int x_m \Pi(\mathrm{d}m)$ is the symmetric point, that coincides with the median and the mean of the barycenter. If some symmetric distribution is unimodal, then its mode coincides with the median and mean, i.e $Q_m(\frac{1}{2}) = x_m$. Since the average of convex (concave) functions is convex (concave), it is clear that the barycenter of symmetric unimodal distributions is also symmetric unimodal. □

Although the unimodality is not preserved in general non-symmetric cases, there are still many families of distributions in which the unimodality is maintained after taking barycenter, as we show in the next result.

Proposition 6.3.2 If $\Pi \in \mathcal{W}_p(\mathcal{P}_{ac}(\mathbb{R}))$ is concentrated on log-concave univariate distributions, then the barycenter $\hat{m}$ is unimodal.

Proof of Proposition 6.3.2. Let $f(x)$ be a log-concave density, then $-\log(f(x))$ is convex so $\exp(-\log(f(x)) = \frac{1}{f(x)}$ is convex. Necessarily f must be unimodal for some $\tilde{x} \in \mathbb{R}$, so quantile function $Q(y)$ is concave for $y < \tilde{y}$ and convex for $y > \tilde{y}$ where $Q(\tilde{y}) = \tilde{x}$. Since $\frac{1}{f(x)}$ is convex decreasing for $x < \tilde{x}$ and convex increasing for $x > \tilde{x}$, then $\frac{1}{f(Q(y))}$ is convex. Hence $\frac{\mathrm{d}Q}{\mathrm{d}y}(y) = \frac{1}{f(Q(y))}$ is convex positive with minima on $\tilde{y}$. Given Π, its barycenter $\hat{m}$ satisfies

$$\frac{\mathrm{d}Q_{\hat{m}}}{\mathrm{d}y} = \int \frac{\mathrm{d}Q_m}{\mathrm{d}y} \Pi(\mathrm{d}m),$$

so if all $\frac{\mathrm{d}Q_m}{\mathrm{d}y}$ are convex, then $\frac{\mathrm{d}Q_{\hat{m}}}{\mathrm{d}y}$ is convex positive with minima on some $\hat{y}$ so $Q_{\hat{m}}(y)$ is concave for $y < \hat{y}$ and convex for $y > \hat{y}$ and $\hat{m}$ is unimodal with a mode on $\hat{x} = Q_{\hat{m}}(\hat{y})$. □

There are many useful typical log-concave distribution families like the normal one, the exponential, logistic, Gumbel, chi-squared, chi and Laplace. Other examples include the Weibull, power, gamma and beta families when the shape parameters are equal or greater than 1. It is interesting to note that some of these families are closed under taking barycenter. For example, the barycenter of normal distributions is normal, and this remains true for the exponential, logistic, Gumbel and Laplace families.

6.3.2 Distributions sharing a common copula

If two multivariate distributions P and Q over $\mathbb{R}^q$ share the same copula, then their $W_p(\mathbb{R}^q)$ distance to the p-th power is the sum of the $W_p(\mathbb{R})$ distances between their marginals raised to the p-power. Furthermore, if the marginals of P are continuous, then an optimal map is given by the coordinate-wise transformation $T(x) = (T^1(x_1), \ldots, T^q(x_q))$ where $T^{\mathrm{i}}(x_{\mathrm{i}})$ is the monotone rearrangement between the marginals P^{i} and Q^{i} for $\mathrm{i} = 1, \ldots, q$. Note that these kinds of transports are closed under composition, convex combination, and contain the identity. This setting allows us to easily extend the results from the univariate case to the multidimensional case.

Lemma 6.3.3 If $\Pi \in \mathcal{W}_p(\mathcal{P}_{ac}(\mathbb{R}^q))$ is concentrated on a set of measures sharing the same copula C, then the p-Wasserstein barycenter $\hat{m}$ of Π has copula C as well, and its i-th marginal $\hat{m}^{\mathrm{i}}$ is the barycenter of the i-th marginal measures of Π. In particular the barycenter does not depend on p.

Proof of Lemma 6.3.3. It is know [30, 4] that for two distributions m and μ with i-th marginals m^{i} and μ^{i} for $\mathrm{i} = 1, ..., q$ respectively, the p-Wasserstein metric satisfies

$$W_p^p(m, \mu) \geq \textstyle\sum_{\mathrm{i}=1}^n W_p^p(m^{\mathrm{i}}, \mu^{\mathrm{i}}),$$

where equality is reached if m and μ share the same copula C. (We abuse notation denoting W_p the p-Wasserstein distance on $\mathbb{R}^q$ as well as on $\mathbb{R}$.) Thus

$$\int W_p^p(m,\mu)\Pi(\mathrm{d}m) \geq \int \sum_{\mathrm{i}=1}^q W_p^p(m^{\mathrm{i}},\mu^{\mathrm{i}})\Pi(\mathrm{d}m) = \sum_{\mathrm{i}=1}^q \int W_p^p(\nu,\mu^{\mathrm{i}})\Pi^{\mathrm{i}}(\mathrm{d}\nu),$$

where Π^{i} is defined via the identity $\int_{\mathcal{P}(\mathbb{R})} f(\nu)\Pi^{\mathrm{i}}(\mathrm{d}\nu) = \int_{\mathcal{P}(\mathbb{R}^q)} f(m^{\mathrm{i}})\Pi(\mathrm{d}m)$. The infimum for the lower bound is reached on the univariate measures $\hat{m}^1, ..., \hat{m}^q$ where $\hat{m}^{\mathrm{i}}$ is the p-barycenter of Π^{i}, which means that $\hat{m}^{\mathrm{i}} = \operatorname{argmin} \int W_p^p(\nu,\mu^{\mathrm{i}})\Pi^{\mathrm{i}}(\mathrm{d}\nu)$. It is plain that the infimum is reached on the distribution $\hat{m}$ with copula C and i-th marginal $\hat{m}^{\mathrm{i}}$ for $\mathrm{i} = 1, ..., q$, which then has to be the barycenter of Π and is independent of p. □

A Wasserstein SGD iteration, starting from a distribution μ, sampling $m \sim \Pi$, and with step γ, both μ and m having copula C, produces the measure $\nu = ((1-\gamma)I + \gamma T_\mu^m)(\mu)$ characterized by having copula C and the i-th marginal quantile functions

$$Q_{\nu^{\mathrm{i}}}(\cdot) = (1-\gamma)Q_{\mu^{\mathrm{i}}}(\cdot) + \gamma Q_{m^{\mathrm{i}}}(\cdot),$$

for $\mathrm{i} = 1, \ldots, q$. The batch stochastic gradient descent iteration works analogously. Alternatively, one can perform (batch) stochastic gradient descent component-wise (with respect to the marginals Π^{i} of Π) and then make use of the copula C.

6.3.3 Spherically equivalent distributions

Following [30], another multidimensional case is constructed as follows: Given a fixed measure $\tilde{m} \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$, its associated family of spherically equivalent distributions is

$$\mathcal{S}_0 := \mathcal{S}(\tilde{m}) = \left\{ \mathcal{L}\left(\tfrac{\alpha(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x} \right) | \alpha \in \mathcal{ND}(\mathbb{R}), \tilde{x} \sim \tilde{m} \right\},$$

where $\|\,\|_2$ is the Euclidean norm and $\mathcal{ND}(\mathbb{R})$ is the set of non-decreasing non-negative functions of $\mathbb{R}_+$. These type of distributions include the simplicially contoured distributions, and also elliptical distributions with the same correlation structure. We denote by $\mathcal{L}(\cdot)$ the law of a random vector, so $m = \mathcal{L}(x)$ and $x \sim m$ are synonyms.

If $y \sim m \in \mathcal{S}_0$, then we have that $\alpha(r) = Q_{\|y\|_2}(F_{\|\tilde{x}\|_2}(r))$, where $Q_{\|y\|_2}$ is the quantile function of the norm of y, $F_{\|\tilde{x}\|_2}$ is the distribution function of the norm of $\tilde{x}$, and $y \sim \frac{\alpha(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}$. More generally, if $m_1 = \mathcal{L}\left(\frac{\alpha_1(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}\right)$ and $m_2 = \mathcal{L}\left(\frac{\alpha_2(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}\right)$, then the optimal transport from m_1 to m_2 is given by $T_{m_1}^{m_2}(x) = \frac{\alpha(\|x\|_2)}{\|x\|_2}x$ where $\alpha(r) = Q_{\|x_2\|_2}(F_{\|x_1\|_2}(r))$. Since $F_{\|x_1\|_2}(r) = F_{\|\tilde{x}\|_2}(\alpha_1^{-1}(r))$ and $Q_{\|x_2\|_2}(r) = \alpha_2(Q_{\|\tilde{x}\|_2}(r))$, we can conclude that $\alpha(r) = \alpha_2(Q_{\|\tilde{x}\|_2}(F_{\|\tilde{x}\|_2}(\alpha_1^{-1}(r)))) = \alpha_2(\alpha_1^{-1}(r))$, so finally

$$T_{m_1}^{m_2}(x) = \tfrac{\alpha_2(\alpha_1^{-1}(\|x\|_2))}{\|x\|_2}x.$$

Note that these kind of transports are closed under composition, convex combination, and contain the identity.

A stochastic gradient descent iteration, starting from a distribution $\mu = \mathcal{L}\left(\frac{\alpha_0(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}\right)$, sampling $m = \mathcal{L}\left(\frac{\alpha(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}\right) \sim \Pi$, with step γ, produces $m_1 = T_0^{\gamma,m}(\mu) := ((1-\gamma)I + \gamma T_\mu^m)(\mu)$. Since $T_0^{\gamma,m}(x) = \frac{(\gamma\alpha+(1-\gamma)\alpha_0)(\alpha_0^{-1}(\|x\|_2))}{\|x\|_2}x$, we have that $m_1 = \mathcal{L}\left(\frac{\alpha_1(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}\right)$ with $\alpha_1 = \gamma\alpha + (1-\gamma)\alpha_0$. Analogously, the batch stochastic gradient iteration produces

$$\alpha_1 = (1-\gamma)\alpha_0 + \tfrac{\gamma}{S}\textstyle\sum_{\mathrm{i}=1}^{S}\alpha_{m^{\mathrm{i}}}.$$

Note that these iterations live in $\mathcal{S}_0$, thus, so does the barycenter $\hat{m} \in \mathcal{S}_0$.

For the barycenter $\hat{m} = \mathcal{L}\left(\frac{\hat{\alpha}(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2}\tilde{x}\right)$, the equation $\int T_{\hat{m}}^m(x)\Pi(\mathrm{d}m) = x$ can be expressed as $\hat{\alpha}(r) = \int \alpha_m(r)\Pi(\mathrm{d}m)$, or equiv. $Q^{\hat{m}}_{\|\hat{y}\|_2}(p) = \int Q^m_{\|y\|_2}(p)\Pi(\mathrm{d}m)$, where $Q^m_{\|y\|_2}$ is the quantile function of the norm of $y \sim m$. This is similar to univariate case.

6.3.4 Scatter-location family

We borrow here the setting of [7], where another useful multidimensional case is defined as follows: Given a fixed distribution $\tilde{m} \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$, referred to as *generator*, the generated scatter-location family is given by

$$\mathcal{F}_0 := \mathcal{F}(\tilde{m}) = \{\mathcal{L}(A\tilde{x} + b) | A \in \mathcal{M}_+^{q\times q}, b \in \mathbb{R}^q, \tilde{x} \sim \tilde{m}\},$$

where $\mathcal{M}_+^{q\times q}$ is the set of symmetric positive definite matrices of size $q \times q$. Without loss of generality we can assume that $\tilde{m}$ has zero mean and identity covariance. If $\tilde{m}$ is the standard multivariate normal distribution, then $\mathcal{F}(\tilde{m})$ is the multivariate normal distribution family.

The optimal map between two members of $\mathcal{F}_0$ is explicit. If $m_1 = \mathcal{L}(A_1\tilde{x} + b_1)$ and $m_2 = \mathcal{L}(A_2\tilde{x} + b_2)$ then the optimal map from m_1 to m_2 is given by $T_{m_1}^{m_2}(x) = A(x - b_1) + b_2$ where $A = A_1^{-1}(A_1A_2^2A_1)^{1/2}A_1^{-1} \in \mathcal{M}_+^{q\times q}$. Observe that this family of optimal transports contains the identity map and is closed under convex combination.

If Π is supported on $\mathcal{F}_0$, then its 2-Wasserstein barycenter $\hat{m}$ belongs to $\mathcal{F}_0$. Call its mean $\hat{b}$ and its covariance matrix $\hat{\Sigma}$. Since the optimal map from $\hat{m}$ to m is $T_{\hat{m}}^m(x) = A_{\hat{m}}^m(x - \hat{b}) + b_m$ where $A_{\hat{m}}^m = \hat{\Sigma}^{-1/2}(\hat{\Sigma}^{1/2}\Sigma_m\hat{\Sigma}^{1/2})^{1/2}\hat{\Sigma}^{-1/2}$ and we know that $\hat{m}$-almost surely $\int T_{\hat{m}}^m(x)\Pi(\mathrm{d}m) = x$. Then we must have that $\int A_{\hat{m}}^m\Pi(\mathrm{d}m) = I$, since clearly $\hat{b} = \int b_m\Pi(\mathrm{d}m)$, and as a consequence $\hat{\Sigma} = \int(\hat{\Sigma}^{1/2}\Sigma_m\hat{\Sigma}^{1/2})^{1/2}\Pi(\mathrm{d}m)$.

A stochastic gradient descent iteration, starting from a distribution $\mu = \mathcal{L}(A_0\tilde{x} + b_0)$, sampling some $m = \mathcal{L}(A_m\tilde{x} + b_m) \sim \Pi$, and with step γ, produces the measure $\nu = T_0^{\gamma,m}(\mu) := ((1-\gamma)I + \gamma T_\mu^m)(\mu)$. If $\tilde{x}$ has a multivariate distribution $\tilde{F}(x)$, then μ has distribution $F_0(x) = \tilde{F}(A_0^{-1}(x - b_0))$ with mean b_0 and covariance $\Sigma_0 = A_0^2$. We have that $T_0^{\gamma,m}(x) = ((1-\gamma)I + \gamma A_\mu^m)(x - b_0) + \gamma b_m + (1-\gamma)b_0$ with $A_\mu^m := A_0^{-1}(A_0A_m^2A_0)^{1/2}A_0^{-1}$. Then ν has distribution

$$F_1(x) = F_0([T_0^{\gamma\cdot m}]^{-1}(x)) = \tilde{F}([(1-\gamma)A_0 + \gamma A_\mu^m A_0]^{-1}(x - \gamma b_m - (1-\gamma)b_0)),$$

with mean $b_1 = (1-\gamma)b_0 + \gamma b_m$ and covariance

$$\begin{aligned}\Sigma_1 = A_1^2 &= [(1-\gamma)A_0 + \gamma A_0^{-1}(A_0A_m^2A_0)^{1/2}][(1-\gamma)A_0 + \gamma(A_0A_m^2A_0)^{1/2}A_0^{-1}] \\ &= A_0^{-1}[(1-\gamma)A_0^2 + \gamma(A_0A_m^2A_0)^{1/2}][(1-\gamma)A_0^2 + \gamma(A_0A_m^2A_0)^{1/2}]A_0^{-1} \\ &= A_0^{-1}[(1-\gamma)A_0^2 + \gamma(A_0A_m^2A_0)^{1/2}]^2A_0^{-1}\end{aligned}$$

The batch stochastic gradient descent iteration is characterized by

$$\begin{aligned}b_1 &= (1-\gamma)b_0 + \tfrac{\gamma}{S}\textstyle\sum_{\mathrm{i}=1}^{S} b_{m^{\mathrm{i}}} \\ A_1^2 &= A_0^{-1}[(1-\gamma)A_0^2 + \tfrac{\gamma}{S}\textstyle\sum_{\mathrm{i}=1}^{S}(A_0A_{m^{\mathrm{i}}}^2A_0)^{1/2}]^2A_0^{-1}.\end{aligned}$$

6.4 Numerical Experiments

We next present experimental validation for our theoretical contribution. This simulation experiment aims to provide practical evidence for the implementation of the proposed approach to Wasserstein Bayesian learning and its relationship to the true model. Precisely, the following experiment consists in: i) defining a true model, ii) sampling from such model to yield a set of data points, iii) sampling from the posterior measures, iv) computing the proposed Bayesian 2-Wasserstein barycenter via empirical approximation, v) analysing our estimator with respect to both the true model and the standard Bayesian model average, and lastly, vi) comparing the empirical estimate versus the proposed stochastic gradient methods for computing population barycenters.

6.4.1 Choice of the true model, prior and posterior samples

Following the discussion in Sec. 6.3.4, we considered models within the location-scatter family (LS), since optimal transports between them can be computed in closed form but are not reduced to the well-known univariate case. We chose the generator of the LS family, denoted $\tilde{m}$, as a distribution on $\mathbb{R}^{15}$ with independent coordinates, where:

- coordinates 1 to 5 are standard Normal distributions
- coordinates 6 to 10 are standard Laplace distributions, and
- coordinates 11 to 15 are standard Student's t-distributions (3 degrees of freedom).

Fig. 6.1 shows uni- and bi-variate marginals for 6 coordinates of $\tilde{m}$.

Within the LS family constructed upon $\tilde{m}$, we chose the true model m_0 to be generated by the location vector $b \in \mathbb{R}^{15}$ defined as $b_{\mathrm{i}} = \mathrm{i} - 1$ for $\mathrm{i} = 1, \ldots, 15$, and the scatter matrix $A = \Sigma^{1/2}$. The covariance matrix Σ was defined as $\Sigma_{\mathrm{i},j} = K\left(\left(\frac{\mathrm{i}-1}{14}\right)^{1.1}, \left(\frac{j-1}{14}\right)^{1.1}\right)$ for $\mathrm{i}, j = 1, \ldots, 15$ [3], with the kernel function $K(\mathrm{i}, j) = \varepsilon\delta_{\mathrm{i}j} + \sigma\cos(\omega(\mathrm{i} - j))$. Given the parameters ε, σ and ω, the constructed covariance matrix is denoted $\Sigma_{\varepsilon,\sigma,\omega}$. We chose parameters $\varepsilon = 0.01$,

[3] We chose $\left(\frac{j-1}{14}\right)^{1.1}$ for $j = 1, \ldots, 15$ because this defines a non-uniform grid over $[0, 1]$.

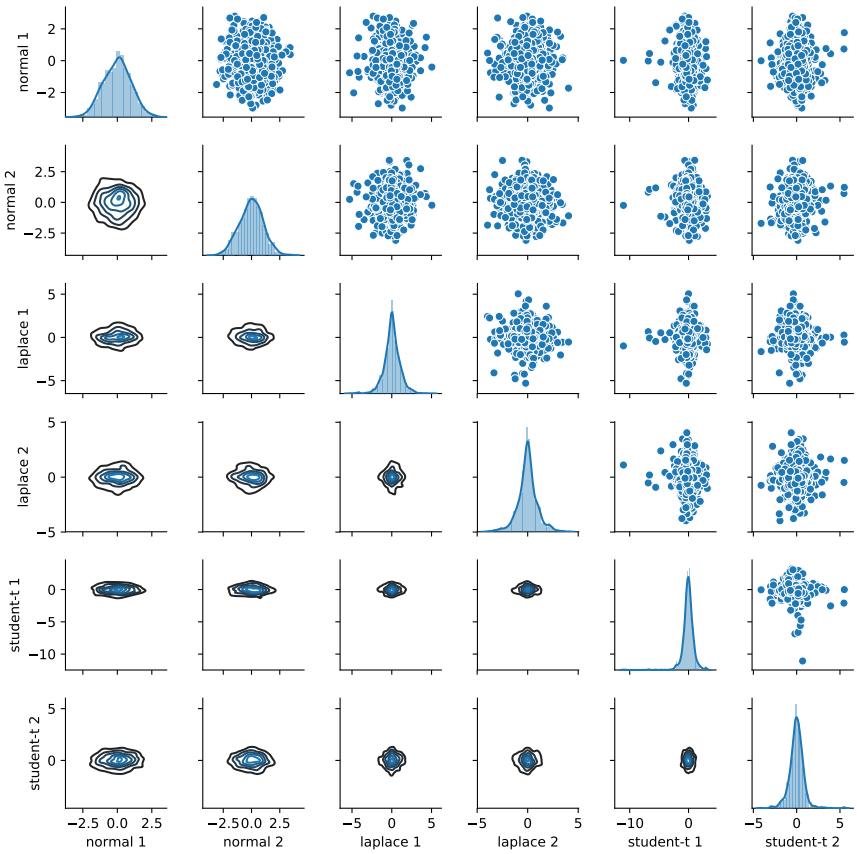

Figure 6.1: Univariate (diagonal) and bivariate (off-diagonal) marginals for 6 coordinates from the generator distribution $\tilde{m}$. The diagonal and lower triangular plots are smoothed histograms, whereas the upper-diagonal ones are collections of samples.

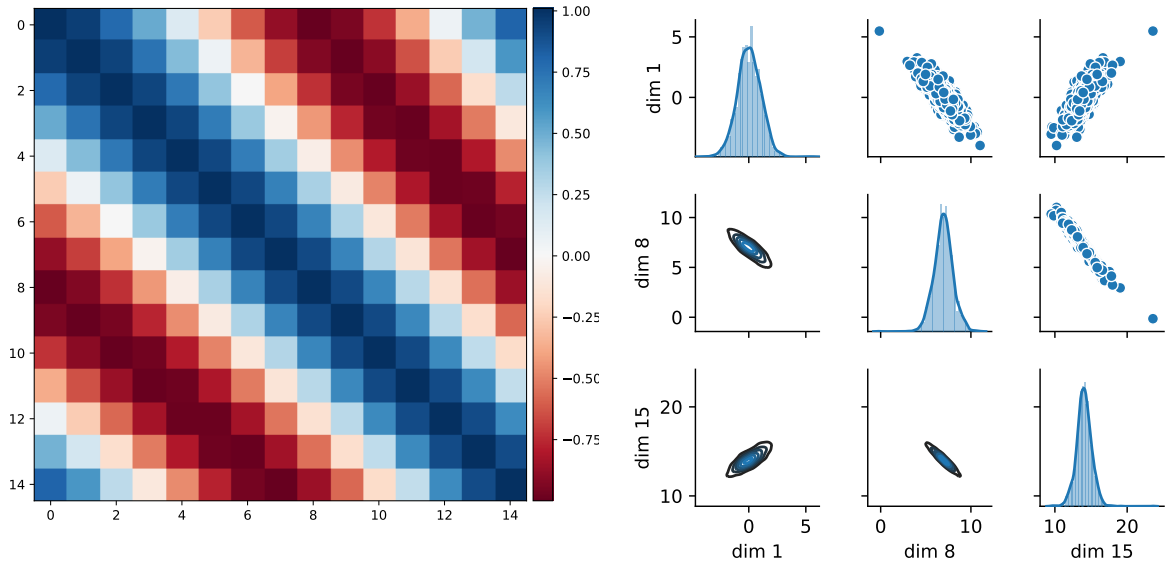

Figure 6.2: True model m_0: covariance matrix (left), and univariate and bivariate marginals for dimensions 1, 8 and 15 (right). Notice that some coordinates are positively or negatively correlated, and some are even close to be uncorrelated.

$\sigma = 1$ and $\omega = 5.652 \approx 1.8\pi$ for m_0. Thus under the true model m_0 the coordinates can be negatively/positively correlated due to the cosine term and there is also a coordinate-independent noise component due to the Kronecker delta δ_{ij}. Fig. 6.2 shows the covariance matrix and three coordinates of the *true* model m_0.

The model prior Π is the push-forward induced by the chosen prior over the mean vector b and the parameters of the covariance $\Sigma_{\varepsilon,\sigma,\omega}$. We chose all these priors to be independent and given by

$$p(b, \Sigma_{\varepsilon,\sigma,\omega}) = \mathcal{N}(b|0, \mathrm{I}) \operatorname{Exp}(\varepsilon|20) \operatorname{Exp}(\sigma|1) \operatorname{Exp}(\omega^{-1}|15), \tag{6.22}$$

where $\operatorname{Exp}(\cdot|\lambda)$ is a exponential distribution with rate λ. Given n samples from the true model m_0 (also referred to as *observations* or *data points*), we generated k samples from the posterior measure Π_n using Markov chain Monte Carlo (MCMC), all to obtain the empirical measure $\Pi_n^{(k)}$. The remaining part of our numerical analysis focuses on the behavior of the Bayesian Wasserstein barycenter as a function of both the number of samples k and the number of data points n.

6.4.2 Numerical consistency of the empirical posterior under the Wasserstein distance

We first validated the empirical measure $\Pi_n^{(k)}$, as a consistent sample version of the true posterior under the W_2 distance, that is, we would like to confirm that $W_2(\Pi_n^{(k)}, \delta_{m_0}) \to W_2(\Pi_n, \delta_{m_0})$

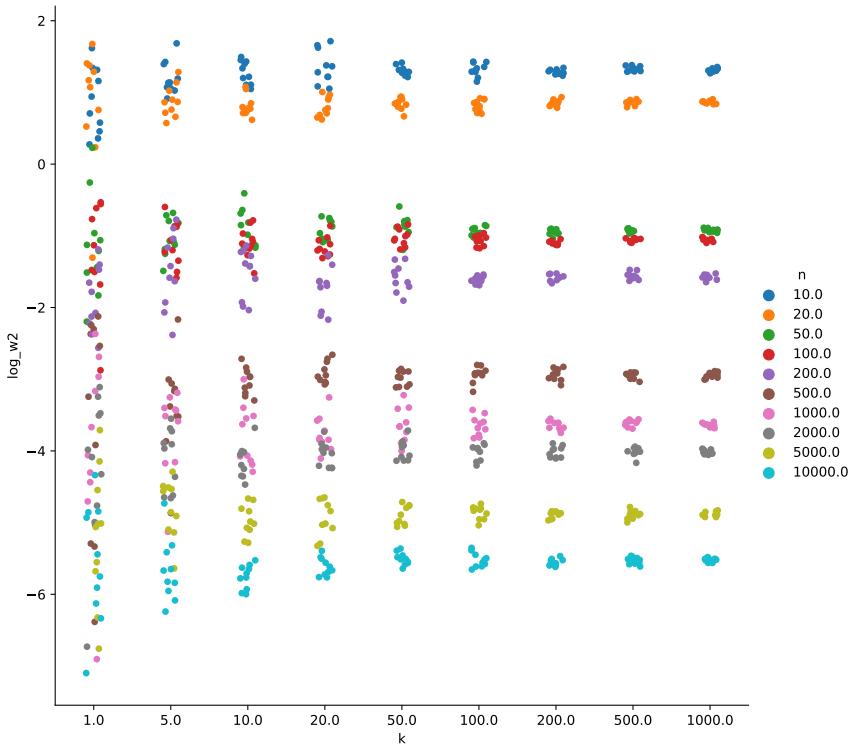

Figure 6.3: Wasserstein distance between the empirical measure $\Pi_n^{(k)}$ and δ_{m_0} in logarithmic scale for different number of observations n (color coded) and samples k (x-axis). For each pair (n, k), 10 estimates of $W_2(\Pi_n^{(k)}, \delta_{m_0})$ are shown.

n / k	1	5	10	20	50	100	200	500	1000
10	1.2506	0.8681	0.5880	0.9690	0.2354	0.3440	0.1253	0.1330	0.0972
20	1.5168	0.5691	0.3524	0.3182	0.1850	0.1841	0.1049	0.0811	0.0509
50	0.3479	0.0948	0.1275	0.0572	0.0623	0.0229	0.0157	0.0085	0.0092
100	0.2003	0.1092	0.0712	0.0469	0.0431	0.0254	0.0087	0.0079	0.0084
200	0.0749	0.1249	0.0717	0.0533	0.0393	0.0101	0.0092	0.0109	0.0072
500	0.0478	0.0285	0.0093	0.0086	0.0053	0.0056	0.0045	0.0023	0.0022
1000	0.0299	0.0113	0.0113	0.0064	0.0067	0.0036	0.0016	0.0012	0.0007
2000	0.0145	0.0071	0.0040	0.0031	0.0027	0.0019	0.0014	0.0011	0.0006
5000	0.0072	0.0031	0.0015	0.0018	0.0010	0.0007	0.0004	0.0005	0.0002
10000	0.0038	0.0020	0.0005	0.0005	0.0004	0.0004	0.0002	0.0002	0.0001

Table 6.1: Standard deviation of $W_2^2(\Pi_n^{(k)}, \delta_{m_0})$, using 10 simulations, for different values of observations n and samples k.

n / k	10	20	50	100	200	500	1000	2000
10	2.1294	2.0139	2.0384	1.9396	1.9608	1.9411	1.9699	1.9548
20	1.4382	1.4498	1.4826	1.4973	1.4785	1.4953	1.4955	1.4914
50	0.2455	0.2759	0.2639	0.2468	0.2499	0.2483	0.2443	0.2454
100	0.1211	0.1387	0.1509	0.1458	0.1379	0.1328	0.1318	0.1349
200	0.1116	0.0922	0.0859	0.0817	0.0777	0.0824	0.0820	0.0819
500	0.0094	0.0077	0.0043	0.0047	0.0041	0.0038	0.0037	0.0039
1000	0.0068	0.0039	0.0031	0.0025	0.0023	0.0022	0.0021	0.0021
2000	0.0072	0.0066	0.0063	0.0062	0.0063	0.0060	0.0062	0.0062
5000	0.0037	0.0037	0.0028	0.0029	0.0031	0.0031	0.0028	0.0030
10000	0.0023	0.0017	0.0017	0.0015	0.0016	0.0017	0.0016	0.0017

Table 6.2: Sample average of $W_2^2(\hat{m}_n^{(k)}, m_0)$, using 10 simulations, for different values of observations n and samples k.

for large k. In this sense, we estimated $W_2(\Pi_n^{(k)}, \delta_{m_0})$ 10 times for each combination of (number of) observations n and samples k in the following sets

- $k \in \{1, 5, 10, 20, 50, 100, 200, 500, 1000\}$
- $n \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000\}$

Fig. 6.3 shows the 10 estimates of $W_2(\Pi_n^{(k)}, \delta_{m_0})$ for different values of k (in the x-axis) and of n (color coded). Notice how the estimates become more concentrated for larger k and that the Wasserstein distance between the empirical measure $\Pi_n^{(k)}$ and the true model m_0 decreases for larger n. Additionally, Table 6.1 shows that the standard deviation of the 10 estimates of $W_2(\Pi_n^{(k)}, \delta_{m_0})$ decreases as either n or k increases.

6.4.3 Distance between empirical barycenter and the true model

For each empirical posterior $\Pi_n^{(k)}$ we intend to compute their Wasserstein barycenter $\hat{m}_n^{(k)}$ as suggested in Section 6.1. We call $\hat{m}_n^{(k)}$ the empirical barycenter. For this purpose, we use the iterative procedure defined in (6.2), namely the (deterministic) gradient descent method, and repeated this calculation 10 times. As a stopping criterion for the gradient descent method, we considered the relative variation of the W_2 cost, terminating the computation if this quantity was less than 10^{-4}. Fig. 6.4 shows all the W_2 distances between the so computed barycenters and the true model, while Table 6.2 shows the average across all these distances for each pair (n, k). Notice that, in general, both the average and standard deviation of the barycenters decrease as either n or k increases, yet for large values (e.g., $n = 2000, 5000$) numerical issues appear.

6.4.4 Distance between the empirical barycenter and the Bayesian model average

Our aim was then to compare the computed empirical Wasserstein barycenters $\hat{m}_n^{(k)}$ to the standard Bayesian model averages $\bar{m}_n^{(k)}$, in terms of their distance to the true model m_0, for

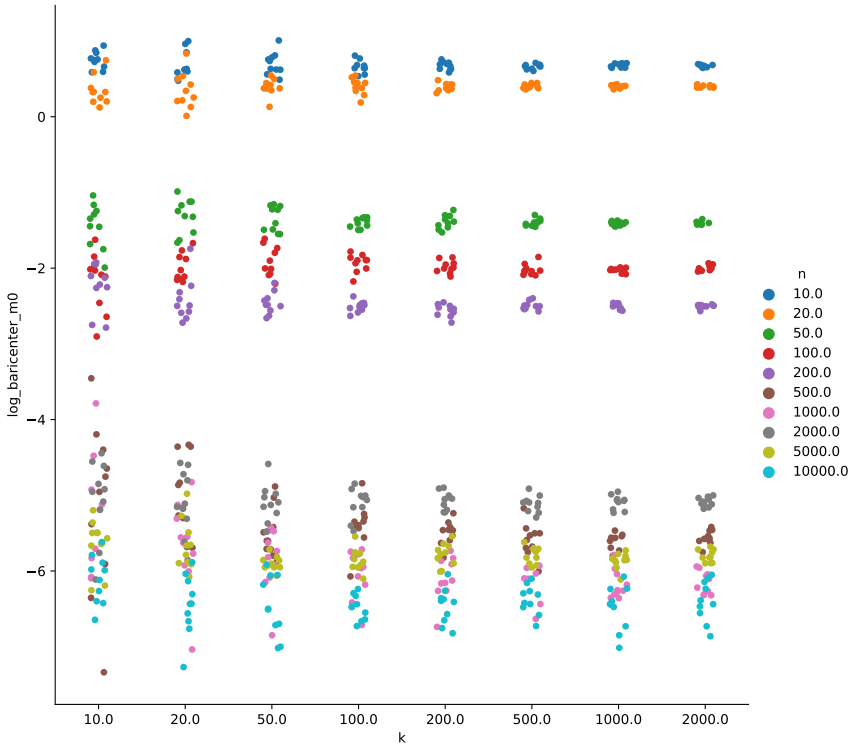

Figure 6.4: W_2 distance between the empirical barycenters $\hat{m}_n^{(k)}$ and the true model m_0 in logarithmic scale for different number of observations n (color coded) and samples k (x-axis). For each pair (n, k), 10 estimates of $W_2(\hat{m}_n^{(k)}, m_0)$ are shown.

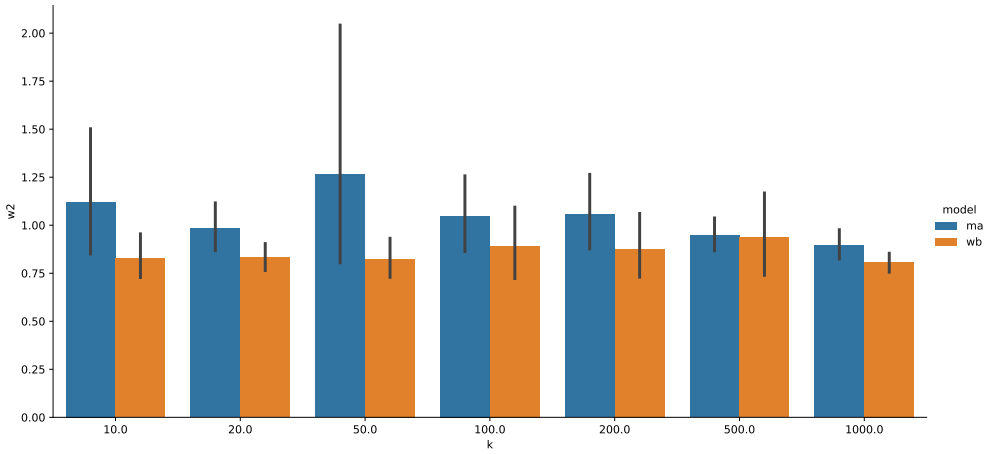

Figure 6.5: Averages (bars) and standard deviations (vertical lines) of $W_2^2(\hat{m}_n^{(k)}, m_0)$ denoted as WB in orange, and $W_2^2(\bar{m}_n^{(k)}, m_0)$ denoted as MA in blue, for $n = 1000$ and different numbers of samples k. We considered 10 simulations for each k.

n / s	1	2	5	10	15	20	empirical
10	2.0421	2.0091	1.9549	1.9721	1.9732	1.9712	1.9532
20	1.4819	1.4868	1.5100	1.4852	1.4840	1.4891	1.4916
50	0.2406	0.2512	0.2465	0.2427	0.2444	0.2460	0.2469
100	0.1340	0.1392	0.1340	0.1349	0.1334	0.1338	0.1366
200	0.0843	0.0811	0.0819	0.0807	0.0820	0.0819	0.0811
500	0.0044	0.0042	0.0039	0.0039	0.0041	0.0040	0.0041

Table 6.3: Means of W_2^2 of the stochastic gradient estimations (using the sequences with $t \geq 100$) and that of the empirical estimator (using the simulations with $k \geq 100$), across different combinations of observations n and batch size s.

n / s	1	2	5	10	15	20	empirical
10	0.1836	0.1071	0.0526	0.0474	0.0397	0.0232	0.0916
20	0.0751	0.0565	0.0553	0.0189	0.0253	0.0186	0.0790
50	0.0210	0.0174	0.0072	0.0084	0.0050	0.0039	0.0138
100	0.0102	0.0076	0.0049	0.0048	0.0035	0.0023	0.0112
200	0.0074	0.0045	0.0021	0.0035	0.0013	0.0017	0.0047
500	0.0016	0.0007	0.0005	0.0004	0.0004	0.0004	0.0009
1000	0.0005	0.0006	0.0004	0.0004	0.0003	0.0003	0.0005

Table 6.4: Std. deviations of W_2^2 of the stochastic gradient estimations (using the sequences with $t \geq 100$) and that of empirical estimator (using the simulations with $k \geq 100$), across different combinations of observations n and batch size s.

$n = 1000$ observations. To do so, we estimated the W_2 distances via empirical approximations with 1000 samples for each model based on [43]. We simulated this procedure 10 times for $k \in \{10, 20, 50, 100, 200, 500, 1000\}$. Fig. 6.5 shows the sample average and variance of the W_2 distances of the Wasserstein barycenters and Bayesian model averages, where it can be seen that the empirical barycenter is closer to the true model than the model average regardless of the number of MCMC samples k.

6.4.5 Computation of the barycenter using batches

Lastly, we compared the empirical barycenters $\hat{m}_n^{(k)}$ against the barycenter obtained by batch stochastic gradient descent method $\hat{m}_{n,s}$. Fig. 6.6 shows the evolution of the W_2^2 distance between the stochastic gradient descent sequences and the true model m_0 for $n \in \{10, 20, 50, 100, 200, 500, 1000\}$ observations and batches of sizes $s \in \{1, 15\}$, with step-size $\gamma_t = \frac{1}{t}$ for $t = 1, \ldots, 200$. Hence, for batch size s and n number of observations, we carry out 200 iterations of the batch stochastic gradient method (6.21) with these explicit step-sizes $\{\gamma_t\}_t$: the resulting estimator is $\hat{m}_{n,s}$. Notice from Fig. 6.6 that the larger the batch, the more concentrated the trajectories of $\hat{m}_{n,s}$ become, and that the estimates exhibit fluctuations when the batch size is small. Table 6.3 summarizes the means of the distance W_2^2 to the true model m_0, using the sequences after $t = 100$ against the empirical estimator using all the simulations with $k \geq 100$. Table 6.4 shows the standard deviation of the distance W_2^2 to the true model m_0, where we notice that the standard deviation decreases as the batch size grows. Observe that for batch sizes $s \geq 5$ the stochastic estimation is *better* than its empirical counterpart, i.e. it has lower variance with similar (or less) bias. This is noteworthy given the fact that

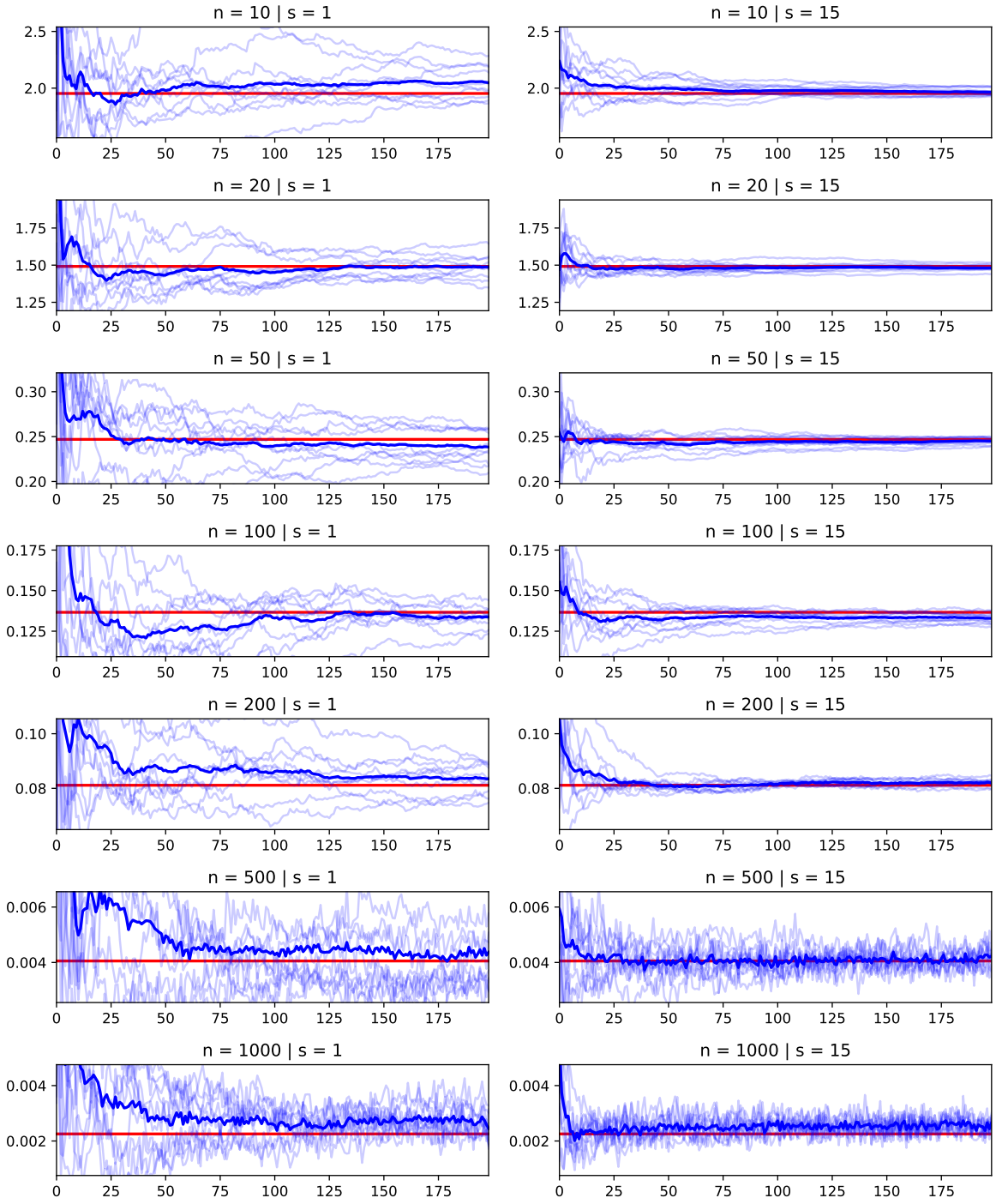

Figure 6.6: Evolution of the W_2^2 cost for 10 realizations of the stochastic barycenter and their mean (blue) versus an empirical barycenter estimator (red), for $n = 10, 20, 50, 100, 200, 500, 1000$ and batches sizes $s = 1, 15$.

computing our Wasserstein barycenter estimator via the batch stochastic gradient descent method is computationally less demanding than computing it via the empirical method.

Based on this illustrative numerical example, we can conclude that:

- the empirical posterior constructed using MCMC sampling is consistent under the W_2 distance and therefore can be relied upon to compute Wasserstein barycenters,
- the empirical Wasserstein barycenter estimator tends to converge faster (and with lower variance) to the true model than the empirical Bayesian model average,
- computing the population Wasserstein barycenter estimator via batch stochastic gradient descent seems to be a superior alternative to calculating the empirical barycenter (i.e., to applying the deterministic gradient descent method to a finitely sampled posterior).

Conclusion

In Chapter 3 we have provided a theoretically-grounded presentation of non-Gaussian processes resulting from nonlinear transformations of GPs using the change of variables theorem, thus complementing existing approaches such as WGP [125], Bayesian WGP [72] and deep GP [33]. Although the warping functions considered by the models mentioned above can be arbitrarily complex, their inverse and derivative require expensive numerical approximations. This fact motivated us to propose the compositionally-warped GP (CWGP), a variant of WGP that uses transformations given by compositions of multiple analytically-invertible and differentiable functions. Due to the expressiveness of the deep composition of elementary functions, the proposed CWGP model represents an improvement in terms of modelling ability with minimal numerical approximations, thus being a competitive alternative to existing methods.

Modelling with copulas [144] is an excellent approach to construct non-Gaussian dependency structures, like heavier-tail Student-t Process introduced in [119] as the most-general elliptical processes with a closed-form density. In Chapter 4 we have proposed a regression model from a unifying point of view with other approaches found in literature, like GP, WGP, Student-t processes, copula processes and a generalised model denoted *Warped Student-t Processes*. We deliver the standard methods of training, inference, and additionally we prove our approach's consistency. We hope to extend the proposed methodology in the future with more expressive models.

In Chapter 5 we have proposed an unifying framework for the Bayesian model selection, covering standard selection criteria, to then introduce the novel *Bayesian Wasserstein barycenter estimator*. We have also illustrated the appealing statistical properties of the proposed estimator, and shown implementation examples in parametric and nonparametric cases, where the desired performance of the proposed method was validated experimentally.

Finally, in Chapter 6 we develop different ways to compute Wasserstein barycenters, where our main contribution is a *stochastic gradient descent* method on the Wasserstein space, showing the convergence under mild conditions. Based on numerical examples, we can conclude that computing the population Wasserstein barycenter estimator via a batch version of the stochastic gradient descent seems to be a superior alternative to calculating the empirical barycenter. This topic has a lot of potential for further development; for example, extending the method, studying its convergence properties and generalizing the kind of problems to which we can apply it.

Bibliography

[1] Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Courier Corporation, 1964.

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

[3] John Aitchison and James Alexander Campbell Brown. *The Lognormal Distribution*, volume 5. CUP Archive, 1976.

[4] Aurélien Alfonsi and Benjamin Jourdain. A remark on the optimal transport between two probability measures sharing the same copula. *Statistics & Probability Letters*, 84:131–134, 2014.

[5] Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A note on the computation of Wasserstein barycenters. *Preprint*, 2015.

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

[7] Pedro C Álvarez-Esteban, Eustasio del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. Wide consensus aggregation in the Wasserstein space. application to location-scatter families. *Bernoulli*, 24(4A):3147–3179, 2018.

[8] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[9] Aloisio Araujo and Evarist Giné. *The Central Limit Theorem for Real and Banach Valued Random Variables*, volume 431. Wiley New York, 1980.

[10] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

[11] Kendall E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, 2008.

[12] Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. *arXiv preprint arXiv:1805.10833*, 2018.

[13] Alan Bain and Dan Crisan. *Fundamentals of Stochastic Filtering*. Springer, 2009.

[14] Yoshua Bengio et al. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[15] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[16] Robert H Berk et al. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

[17] Peter J. Bickel and Kjell A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76:296–311, 1981.

[18] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *arXiv preprint arXiv:1212.2562*, 2012.

[19] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[20] Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes, et al. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[21] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

[22] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[23] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[24] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[25] Thang Bui, Daniel Hernández-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*, pages 1472–1481, 2016.

[26] Hugh Chipman, Edward I George, Robert E McCulloch, Merlise Clyde, Dean P Foster, and Robert A Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

[27] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[28] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

[29] Noel Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

[30] Juan Cuesta-Albertos, L Ruschendorf, and Araceli Tuero-Diaz. Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis*, 46(2):335–361, 1993.

[31] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300. Curran Associates, Inc., 2013.

[32] Andreas Damianou. *Deep Gaussian processes and variational propagation of uncertainty*. PhD thesis, University of Sheffield, 2015.

[33] Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

[34] Andreas C Damianou, Michalis K Titsias, and Neil D Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *The Journal of Machine Learning Research*, 17(1):1425–1486, 2016.

[35] Stefano Demarta and Alexander J McNeil. The t copula and related copulas. *International Statistical Review/Revue Internationale de Statistique*, pages 111–129, 2005.

[36] James W Demmel. *Applied numerical linear algebra*, volume 56. Siam, 1997.

[37] Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

[38] Catherine Donnelly and Paul Embrechts. The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *ASTIN Bulletin: The Journal of the IAA*, 40(1):1–33, 2010.

[39] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, pages 1166–1174, 2013.

[40] David Duvenaud, Oren Rippel, Ryan Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pages 202–210, 2014.

[41] Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[42] Federal Reserve Bank of St. Louis. Federal reserve economic data, 2009.

[43] Rémi Flamary and Nicolas Courty. POT Python Optimal Transport library, 2017.

[44] Joaquin Fontbona, Hélène Guérin, and Sylvie Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electron. Commun. Probab.*, 15:124–133, 2010.

[45] Daniel Foreman-Mackey, David W Hogg, Dustin Lang, and Jonathan Goodman. emcee: the mcmc hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306, 2013.

[46] J. Freeman and R. Modarres. Properties of the power normal distribution. *Department of Statistics. George Washington University*, 2002.

[47] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[48] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[49] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[50] Novin Ghaffari and Stephen Walker. On multivariate optimal transportation. *arXiv preprint arXiv:1801.03516*, 2018.

[51] Subhashis Ghosal, Jayanta K Ghosh, and Aad Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

[52] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[53] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[54] Leon Glass, Peter Hunter, and Andrew McCulloch. *Theory of heart: biomechanics, biophysics, and nonlinear dynamics of cardiac function*. Springer Science & Business Media, 2012.

[55] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[56] Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

[57] Marian Grendár and George Judge. Asymptotic equivalence of empirical likelihood and Bayesian map. *Ann. Statist.*, 37(5A):2445–2457, 10 2009.

[58] Robert V. Hogg and Allen T. Craig. *Introduction to Mathematical Statistics*. Upper Saddle River, New Jersey: Prentice Hall, fifth edition, 1995.

[59] K. Hornik. Some new results on neural network approximation. *Neural Networks*, 6(8):1069–1072, 1993.

[60] Norman L. Johnson. Systems of frequency curves generated by methods of translation.

Biometrika, 36(1/2):149–176, 1949.

[61] M. Chris Jones and Arthur Pewsey. Sinh-Arcsinh distributions. *Biometrika*, 96(4):761, 2009.

[62] Douglas Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 419–430, 1970.

[63] James Kennedy. Particle swarm optimization. *Encyclopedia of machine learning*, pages 760–766, 2010.

[64] Sanggyun Kim, Diego Mesa, Rui Ma, and Todd P Coleman. Tractable fully Bayesian inference via convex optimization and optimal transport theory. *arXiv preprint arXiv:1509.08582*, 2015.

[65] Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.

[66] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[67] Bastiaan Jan Korneel Kleijn et al. *Bayesian asymptotics under misspecification*. PhD thesis, Vrije Universiteit Amsterdam, 2004.

[68] Bastiaan Jan Korneel Kleijn, Adrianus Willem Van der Vaart, et al. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

[69] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[70] Karl Krauth, Edwin V Bonilla, Kurt Cutajar, and Maurizio Filippone. Autogp: Exploring the capabilities and limitations of gaussian process models. *arXiv preprint arXiv:1610.05392*, 2016.

[71] 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.

[72] Miguel Lázaro-Gredilla. Bayesian warped Gaussian processes. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, volume 25, pages 1619–1627. Curran Associates, Inc., 2012.

[73] Miguel Lázaro-Gredilla. Bayesian warped gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1619–1627, 2012.

[74] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein

barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[75] Cheng Li, Sanvesh Srivastava, and David Dunson. Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104(3):665–680, 2017.

[76] Ping Li and Songcan Chen. A review on gaussian process latent variable models. *CAAI Transactions on Intelligence Technology*, 1(4):366–376, 2016.

[77] Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5665–5674, 2017.

[78] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.

[79] Pascal Massart. *Concentration inequalities and model selection*. Springer, 2007.

[80] Scherer Matthias and Mai Jan-frederik. *Simulating copulas: stochastic models, sampling algorithms, and applications*, volume 6. # N/A, 2017.

[81] Alexander J McNeil, Johanna Nešlehová, et al. Multivariate archimedean copulas, d-monotone functions and ℓ1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097, 2009.

[82] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Robust and scalable bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18(1):4488–4527, 2017.

[83] Nitis Mukhopadhyay. *Probability and statistical inference*. CRC Press, 2000.

[84] Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution, 11 2007.

[85] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, Cambridge, MA, 2012.

[86] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[87] Radford M Neal et al. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

[88] XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[89] XuanLong Nguyen. Borrowing strengh in hierarchical bayes: Posterior concentration of the dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016.

[90] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006.

[91] Joel Owen and Ramon Rabinovitch. On the class of elliptical distributions and their

applications to the theory of portfolio choice. *The Journal of Finance*, 38(3):745–752, 1983.

[92] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *arXiv preprint arXiv:1806.05500*, 2018.

[93] VM Panaretos and Y Zemel. Fréchet means and procrustes analysis in Wasserstein space. *Bernoulli*, 2017.

[94] Matthew Parno. *Transport maps for accelerated Bayesian computation.* PhD thesis, Massachusetts Institute of Technology, 2015.

[95] Gabriel Parra and Felipe Tobar. Spectral mixture kernels for multi-output Gaussian processes. In *Advances in Neural Information Processing Systems 30*, pages 6681–6690. Curran Associates, Inc., 2017.

[96] Brendan Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.

[97] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[98] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[99] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[100] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[101] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

[102] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT, 2006.

[103] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1530–1538. JMLR. org, 2015.

[104] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Proceedings of the IEEE international conference on neural networks*, volume 1993, pages 586–591. San Francisco, 1993.

[105] Gonzalo Rios. G3py: Generalized graphical Gaussian processes, github.com/griosd/g3py, 2017.

[106] Gonzalo Rios. Tpy: Transport processes in python, github.com/griosd/tpy, 2017.

[107] Gonzalo Rios and Felipe Tobar. Learning non-Gaussian time series using the Box-Cox Gaussian process. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[108] Gonzalo Rios and Felipe Tobar. Compositionally-warped Gaussian processes. *Neural Networks*, 118:235–246, 2019.

[109] VK Rohatgi. An introduction to probability theory and mathematical statistics. 1976.

[110] Sheldon M Ross, John J Kelly, Roger J Sullivan, William James Perry, Donald Mercer, Ruth M Davis, Thomas Dell Washburn, Earl V Sager, Joseph B Boyce, and Vincent L Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

[111] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

[112] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[113] Remi M. Sakia. The Box-Cox transformation technique: A review. *The Statistician*, pages 169–178, 1992.

[114] Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

[115] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

[116] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[117] Rafael Schmidt. Tail dependence. In *Statistical Tools for Finance and Insurance*, pages 65–91. Springer, 2005.

[118] Lorraine Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

[119] Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to Gaussian processes. In *AISTATS*, pages 877–885, 2014.

[120] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[121] Cosma Rohilla Shalizi and Aryeh Kontorovich. Almost none of the theory of stochastic processes. *Lecture Notes*, 2010.

[122] SILSO World Data Center. The International Sunspot Number. *International Sunspot Number Monthly Bulletin and online catalogue*, 1700-2008.

[123] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[124] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

[125] Edward Snelson, Zoubin Ghahramani, and Carl E Rasmussen. Warped gaussian processes. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in neural information processing systems*, volume 16, pages 337–344. MIT Press, 2004.

[126] Arno Solin and Simo Särkkä. State space methods for efficient inference in student-t process regression. In *Artificial Intelligence and Statistics*, pages 885–893, 2015.

[127] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

[128] Sanvesh Srivastava, Cheng Li, and David Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

[129] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[130] Jakob Stoeber, Harry Joe, and Claudia Czado. Simplified pair copula constructions–limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118, 2013.

[131] Daniel W Stroock. *Probability theory: an analytic view*. Cambridge university press, 2010.

[132] E. G. Tabak and C. V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[133] E. G Tabak and E Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[134] Terence Tao. *An Introduction to Measure Theory*, volume 126. American Mathematical Society, 2011.

[135] Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In David van Dyk and Max Welling, editors, *Proc. of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574, 2009.

[136] 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.

[137] Felipe Tobar. Bayesian nonparametric spectral estimation. In *Advances in Neural Information Processing Systems 31*, pages 10148–10158. Curran Associates, Inc., 2018.

[138] Cédric Villani. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Soc., 2003.

[139] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[140] Yali Wang, Marcus Brubaker, Brahim Chaib-Draa, and Raquel Urtasun. Sequential inference for deep gaussian process. In *Artificial Intelligence and Statistics*, pages 694–703, 2016.

[141] Chihiro Watanabe, Kaoru Hiramatsu, and Kunio Kashino. Modular representation of layered neural networks. *Neural Networks*, 97:62–73, 2018.

[142] Christopher K. I. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems 9*, pages 295–301. MIT Press, 1997.

[143] Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

[144] Andrew Wilson and Zoubin Ghahramani. Copula processes. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2460–2468. Curran Associates, Inc., 2010.

[145] Andrew Gordon Wilson, David A Knowles, and Zoubin Ghahramani. Gaussian process regression networks. *arXiv preprint arXiv:1110.4411*, 2011.

[146] Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362, 1995.