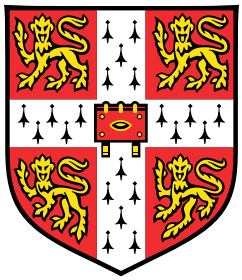
Efficient Gaussian Processes for data-driven decision making



Vincent Dutordoir

Department of Engineering
University of Cambridge

Report submitted to be registered for the PhD Degree
First-Year-Report

Jesus College

August 2021

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Vincent Dutordoir
August 2021

Acknowledgements

I would like to thank my advisor Dr. Carl Henrik Ek, and my supervisor Professor Zoubin Ghahramani for their guidance and support during the first year of my PhD degree. Their advice has greatly shaped the form of this thesis, and at the same time it has also been a pleasure to work with them.

Abstract

This is where you write your abstract ...

Table of contents

Chapter 1

Introduction

As the world we live in grows ever more interconnected and complex, making good decisions becomes increasingly difficult. Heterogenous transportation systems in large cities need to be optimised, global supply chains must be operated such thay they are as efficient and reliable as possible, and connected smart agents, such as self-driving cars, need to adhere to a policy that benefits the overall cause. Humans can typically make good decisions when the number of covariates is small, but suffer quickly when they have to consider thousands of influencing factors, or have to take into account the impact of their earlier decisions on future ones. Data-driven decision-making is a general framework that studies this problem. The aim of data-driven decision-making is to use past and current information to build a model of the environment that can be used to reason about future decisions and their impact.

A crucial building block for data-driven decision-making is a *model*, which tries to explain the given data. uncertainty many explanations for the data aleatoric epistemic

1. Data-driven Decision-making
2. The data can be explained by many models
3. Models that represent uncertainty
4. Statistical learning theory: Emperical risk minimisation

   Supervised machine learning setting $x \in \mathcal{X}$ and $y \in \mathbb{R}$, and a dataset $\mathcal{D} = \{x_i, y_i\}_i^N$

   General problem:

   $$\operatorname*{argmin}_f \sum_i L(f(x_i), y_i) + ||f|| \tag{1.1}$$

5. No Free Lunch Theorem
6.
7. Bayesian Linear Regression: $f(x) = w^\top \phi(x)$

8. Parametric models
9. Probabilistic machine learning: Bayes Rule
10. Kernel methods

1.1 Contributions and Layout of this Report

This report represents my learning and the research that I conducted during the first year of my PhD degree. Most notably, we developed a novel sparse approximation for (deep) Gaussian processes based on the decomposition of the kernel in Spherical harmonics. In chapter 2 we cover the necessary theoretical background.

Chapter 3 In this chapter we introduce a new class of inter-domain variational GPs where data is mapped onto the unit hypersphere in order to use spherical harmonic representations. The inference scheme is comparable to Variational Fourier Features, but it does not suffer from the curse of dimensionality, and leads to diagonal covariance matrices between inducing variables. This enables a speed-up in inference, because it bypasses the need to invert large covariance matrices. The experiments show that our model is able to fit a regression model for a dataset with 6 million entries two orders of magnitude faster compared to standard sparse GPs, while retaining state of the art accuracy.

The content of this chapter is largely based on:

Vincent Dutordoir, Nicolas Durrande, and James Hensman [2020]. "Sparse Gaussian Processes with Spherical Harmonic Features". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*,

with the exception of the algorithm for computing the spherical harmonics in high dimensions.

Chapter 3 Following up on the previous chapter, we use the decomposition of zonal kernels to design an interdomain inducing variable that mimics the behaviour of activation functions is neural network layers.

The content of this chapter is largely based on:

Vincent Dutordoir, James Hensman, Mark van der Wilk, Carl Henrik Ek, Zoubin Ghahramani, and Nicolas Durrande [2021a]. "Deep Neural Networks as Point Estimate for Deep Gaussian Processes". In: *submission to NeurIPS*.

Chapter 4

Chapter 2

Theoretical Framework

This chapter discusses Gaussian processes (GP) and deep Gaussian processes (DGPs), the composite model obtained by stacking multiple GP models on top of each other. We also review how to perform approximate Bayesian inference in these models, with a particular attention to Variational Inference. We also cover the theory of postive definite kernels and the Reproducing Kernel Hilbert Spaces (RKHS) they characterise.

2.1 Gaussian Processes

Gaussian processes (GPs) [Rasmussen and Williams, 2006] are non-parametric distributions over functions similar to Bayesian Neural Networks (BNNs). The core difference is that neural networks represent distributions over functions through distributions on weights, while a Gaussian process specifies a distribution on function values at a collection of input locations. This representation allows us to use an infinite number of basis functions, while still enables Bayesian inference [Neal, 1995].

Following from the Kolomogrov extension theorem, we can construct a real-valued stochastic process (i.e. function) on a non-empty set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, if there exists on all finite subsets $\{x_1, \dots x_N\} \subset \mathcal{X}$, a *consistent* collection of finite-dimensional marginal distributions over $f(\{x_1, \dots, x_n\})$. For a Gaussian process, in particular, the marginal distribution over every finite-dimensional subset is given by a multivariate normal distribution. This implies that, in order to fully specify a Gaussian process, it suffice to only define the mean and covariance (kernel) function because they are the sufficient statistics for every finite-dimensional marginal distribution. We can therefore denote the GP as

$$f \sim \mathcal{GP}(\mu, k), \tag{2.1}$$

where $\mu : \mathcal{X} \to \mathbb{R}$ is the mean function, which encodes the expected value of f at every x, $\mu(x) = \mathbb{E}_f[f(x)]$, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance (kernel) function that describes the

covariance between function values, $k(x, x') = \mathrm{Cov}(f(x), f(x'))$. The covariance function has to adhere to certain properties which we will specify later.

Given a GP as defined in eq. (2.1), for every finite collection of points $\{x_1, \dots x_N\} \subset \mathcal{X}$ and for all $N \in \mathbb{N}$, the distribution over the function values follows

$$f(\{x_i\}_{i=1}^N) \sim \mathcal{N}(\boldsymbol{\mu}_f, \mathbf{K}_{ff}), \quad \text{where } \boldsymbol{\mu}_f \in \mathbb{R}^N \text{ and } \mathbf{K}_{ff} \in \mathbb{R}^{N \times N}, \tag{2.2}$$

with entries $[\boldsymbol{\mu}_f]_i = \mu(x_i)$ and $[\mathbf{K}_{ff}]_{i,j} = k(x_i, x_j)$. Throughout this report we will assume, without loss of generality, a zero *prior* mean function. The Gaussianity, and the fact that we can manipulate function values at some finite points of interest without taking the behaviour at any other points into account (the marginalisation property) make GPs particularly convenient to manipulate and use as priors over functions in Bayesian models.

2.1.1 The Beauty of Gaussian Processes: Exact Bayesian Inference

1. Bayesian Machine Learning
2. Bayes rule
3. Gaussian likelihood
4. posterior, derivate and marginal likelihood
5. Plot: Prior, Data, Posterior
6. occam's razor

Problems A common criticism for GPs is that any modification to this approach breaks the Gaussian assumption.

1. Classifaction
2. Large datasets
3. Transformations

Solutions

1. Laplace
2. Expectation Propagation
3. Sparse Variational Inference

2.2 Approximate Inference in Sparse Gaussian Processes

1. General introduction to Variational inference [Blei et al., 2017] variational inference (VI), where the problem of Bayesian inference is cast as an optimization problem—namely, to maximize a lower bound of the logarithm of the marginal likelihood.

2. Sparse approximations [Snelson and Ghahramani, 2005; Quiñonero-Candela and Rasmussen, 2005]

2.3 Interdomain Inducing Variables

2.3.1 Example: heavyside inducing variable

2.4 Deep Gaussian Processes

Vincent Dutordoir, Hugh Salimbeni, Eric Hambro, John McLeod, Felix Leibfried, Artem Artemev, Mark van der Wilk, James Hensman, Marc P Deisenroth, and ST John [2021b]. "GPflux: A Library for Deep Gaussian Processes". In: *arXiv preprint arXiv:2003.01115*

2.5 Covariance Functions

1. Positive Definite and Symmetry

2. RKHS

3. Bochner's theorem

4. Mercer Decomposition

5. Examples of RKHS

6. RKHS through Spectral Decomposition

7. Representer Theorem

8. Show how sparse approximation links anchor points

Chapter 3

Chapter 4

Sparse Gaussian Processes with Spherical Harmonic Features

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association*.

Vincent Dutordoir, Nicolas Durrande, and James Hensman (2020). "Sparse Gaussian Processes with Spherical Harmonic Features". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Vincent Dutordoir, James Hensman, Mark van der Wilk, Carl Henrik Ek, Zoubin Ghahramani, and Nicolas Durrande (2021a). "Deep Neural Networks as Point Estimate for Deep Gaussian Processes". In: *submission to NeurIPS*.

Vincent Dutordoir, Hugh Salimbeni, Eric Hambro, John McLeod, Felix Leibfried, Artem Artemev, Mark van der Wilk, James Hensman, Marc P Deisenroth, and ST John (2021b). "GPflux: A Library for Deep Gaussian Processes". In: *arXiv preprint arXiv:2003.01115*.

Radford M. Neal (1995). *Bayesian Learning for Neural Networks*. Springer.

Joaquin Quiñonero-Candela and Carl E. Rasmussen (2005). "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *Journal of Machine Learning Research*.

Carl E. Rasmussen and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Edward Snelson and Zoubin Ghahramani (2005). "Sparse Gaussian Processes using Pseudo-inputs". In: *Advances in Neural Information Processing Systems 4 (NIPS 2005)*.